PROBABILITY AND STATISTICS IN PARTICLE PHYSICS

A. G. Frodesen, O. Skjeggestad
DEPARTMENT OF PHYSICS
UNIVERSITY OF BERGEN

H. Tøfte
DEPARTMENT OF COMPUTING SCIENCE
AGDER REGIONAL COLLEGE, KRISTIANSAND

UNIVERSITETSFORLAGET
BERGEN – OSLO – TROMSØ

ISBN 82-00-01906-3

Distribution offices:

NORWAY
Universitetsforlaget
Box 2977 Tøyen
Oslo 6

UNITED KINGDOM
Global Books Resources Ltd.
109 Great Reffill Street
London WC1B 3ND

UNITED STATES and CANADA
Columbia University Press
136 South Broadway
Irvington-on-Hudson
New York 10533

Printed in Norway by
Reklametrykk A.s, Bergen

Preface

The present book on probability theory and statistics is intended for graduate students and research workers in experimental high energy and elementary particle physics. The book has originated from the authors' attempts during many years to provide themselves and their students working for a degree in experimental particle physics with practical knowledge of statistical analysis methods and some further insight required for research in this field.

The first drafting of notes started more than ten years ago when the authors were colleagues at the University of Oslo, and working with bubble chamber experiments. At that time no textbook in statistics was known to us which took its examples and applications from high energy physics and could serve as a reference book in daily work and a suitable curriculum for our students. Several advanced books were available in the library which discussed the fundamentals of probability theory and mathematical statistics *per se* [*e.g.* Cramér, Kendall and Stuart]. Other, less demanding textbooks incorporated useful examples from many fields, including physics, and had the virtue of acquainting a wider scientific community with the universal methods devised by the science of statistics [Fisher, Johnson and Leone, Ostle, Sverdrup, Wine]. Also available were articles and lecture notes discussing statistical estimation in physics in general and in high energy physics in particular [Annis *et al.*, Böck, Hudson, Jauneau and Morellet, Orear, Solmitz]. The need for more coherent presentations apparently was latent and brought on the market a systematic, relatively theoretical account of statistics as used by physicists [Martin, 1971], as well as two treatises written by experimental particle physicists. The latter authors, however, either intended their book "for students and research workers in science, medicine, engineering and economics" [Brandt, 1970], or addressed their course "to physicists (and experimenters in related sciences) in their task of extracting information from experimental data" [Eadie *et al.*, 1971].

The present text has been written for readers who are supposed to have

their main interest in elementary particle physics and who have a need for statistical methods as a tool in their work. This, of course, does not mean that the book can only be comprehended by people whose background is particle physics. However, it is only fair to state that it is a rather specialized book, in which the topics discussed, the disposition and style reflect the need of an experimental particle physicist, and in which examples and applications have been almost exclusively chosen from this field. This fact undoubtedly limits the usefulness of the book to readers from other branches of physics. On the other hand, with the high degree of specialization within the physical and other sciences today, it is, in the opinion of the authors, well worth-while to aim at a more restricted group of readers and to tailor the presentation to meet the specific demands of this group. After all, the statistical methods needed in many disciplines are more or less standard and available in excellent general presentations of mathematical statistics. Often, however, the task of extracting the relevant information from these books can be both time-consuming and troublesome for the non-specialist. Stories are told about physicists who have spent months of their time developing new methods for data analysis, only to find out later that such methods were already described in the statistical literature. A dedicated book like the present can hopefully serve to reduce such instances of vasted time and effort.

The book assumes no previous knowledge beyond basic calculus. The subject of probability theory is entered on an elementary level and given a rather simple and detailed exposition; this is thought to be to the benefit of the student who starts a new course and should get well acquainted with the most common probability concepts and distributions before entering the domain of statistics. The book has been written so that it need not be worked through as a regular course, with extensive reading from the very beginning, but can be studied chapter- or section-wise, should the reader prefer so. The material has, in fact, been organized with an eye to the experimental physicist's practical need, which is likely to be statistical methods for estimation or decision-making. Since an established physicist will probably possess a sufficient background on the fundamentals of probability theory, he can be recommended to start his reading directly at the chapters he is interested in. Cross references are given to indicate where definitions and developed formulae can be found in earlier chapters of the

book. It is hoped that the book will prove useful in everyday work. For this purpose the list of contents and the subject index have been made to include physics key words as well as statistical terms to facilitate the use of the book as a reference manual. To make it selfcontained, the book has also been supplied with a set of statistical tables in an appendix.

With its emphasis on the various practical aspects of statistical methods and techniques the presentation in this book differs from the general, more theoretical points of view shared by statisticians. As non-professionals in statistics the authors make no claim to originality on the subject. We have felt free to borrow material which, over the years, have acquired a status of "common property" among particle physicists, without mentioning originators or written sources. Our reference policy is otherwise to give only the names of authors in the text where examples have been taken from articles in physics publications, lecture notes *etc.*, and to give the full reference in the bibliography at the end of the book. The bibliography also contains references to textbooks which can be suggested as alternative and further reading.

We would like to express explicitly our indebtedness to M.G. Kendall and A. Stuart, the authors of the three-volume work "The Advanced Theory of Statistics" which we have constantly consulted and found to contain answers to any question.

We are also indebted to Addison-Wesley Publishing Company, Inc., for permission to use material from Table 15.1 in *Handbook of Statistical Tables* by D. Owen, and to the Biometrika Trustees for permission to reproduce Table 1 from L.R. Verdooren's paper "Extended tables of critical values for Wilcoxon's test statistic", printed in *Biometrika*.

It is a pleasure to acknowledge the useful help of our many students who over the years contributed their comments on the course. Finally, we wish to thank Mrs. Laila Nøst for her patient and careful cooperation with the typing of the manuscript.

December 1978 A.G.F., O.S., H.T.

I often say
that when you can measure
what you are speaking about,
and express it in numbers,
you know something about it;
but when you cannot measure it,
when you cannot express it in numbers,
your knowledge is
of a meagre and unsatisfactory kind.

Lord Kelvin 1883

Contents

1. Introduction

The term *statistics* is given several precise definitions in the dictionaries and is also used with different meanings in everyday language. It can be used as synonymous with *data*, or taken to mean the entire *scientific discipline* concerned with the methodology of extracting information from data. Often the word is given a meaning between these two extremes, and stands for a compilation of figures pertaining to various interests, or for specified methods or computational procedures applied to sets of numerical observations of similar kind. Used in different contexts the word statistics is associated with the collecting and summarizing of observations in experiments, with measuring the variation in such data, and with more elaborate investigations; these may include comparison with other data or model predictions, and parameter estimation and hypothesis testing on the basis of observed data.

The ultimate goal of the physical sciences is to uncover the fundamental laws governing the phenomena in the material world. In pursuing this ambitious task it is widely recognized that statistical reasoning and statistical analysis methods are becoming increasingly important, even indispensable in many fields. Accordingly, training in statistics has become more or less mandatory for students aiming for an academic career on the experimental side of these fields. Many supervisors recommend a course in mathematical statistics to their students. However, as each discipline has developed its own character and methodology for experimentation, registration, and interrelation of observed facts, as well as its own conceptual and theoretical framework, it is by now usually agreed that the training in data handling and statistics should be made an integrated part of the discipline for practical and pedagogical reasons. Thorough knowledge of means and methods is essential in all scientific research. The education in statistics should therefore go beyond the teaching of cook-book recipes to be applied in a more or less automatic manner to standard problems, and aim at providing some understanding of the general principles involved as well as of the specific assumptions underlying

the various methods, since these determine their applicability and hence their implicit limitations.

An experiment is often motivated by the need to confront current theoretical ideas with new experimental facts to reduce speculative model-building and confine intellectual and other effort in a certain direction. In experimental high energy physics utilizing the facilities of man-made particle accelerators a typical experiment is an impressive undertaking which may represent the combined effort of large teams of technicians, engineers and physicists from several universities or other research institutions, often from many countries, and which may last over a period of many years. Probability arguments and statistical reasoning are employed throughout the experiment, and may in fact enter already at the initial planning stage with estimating the size of the experiment, in terms of the number of events needed to attain a specified precision. The proposal for the experiment will include other estimates for costs in money and man-power, the proposed experimental layout and design. During the long period when the experiment is assembled on the floor, extensive Monte Carlo simulations are carried out to follow the paths of produced particles through the experimental set-up and estimate the acceptances and efficiencies of the different detectors. The actual running period with beam on target may last for a few weeks time up to several months. In electronic experiments the interesting event candidates are, during this period, directly selected through triggering systems, checked and initially analyzed by on-line computers before the relevant information from all parts of the detection system is output event-by-event on magnetic tape. In experiments with bubble chambers as detection device the films from the exposure must first be scanned for events with the selected track topologies, measured, kinematically analyzed, and checked again on the scanning table before the event is ready for the data summary tape. Following the data acquisition comes next a phase in which the collected data are examined for internal consistency, possible error sources located, and biases corrected for. When the observed data have passed all consistency checks and are well understood the final stage involves the interpretation of the experimental findings in the light of theoretical models. At times, when the data can not be explained by the existing models, the experimental outcome may call for revision of current ideas, in exceptional and historic cases even lead to the discovery of fundamental laws.

The interplay between theory and experiment through probability and statistics can be sketched as follows:

A theoretical model predicts a certain correspondence between an observable quantity x and some parameter θ which is not known experimentally and which is not directly accessible to measurement; the value of θ may, or may not, be predicted by the model. The purpose of the experiment is to "determine" the value of θ by performing measurements of the observable x. From the set of observed numbers $x_1, x_2, \cdots, x_n$ statistical methods will tell us how to obtain an estimate for the parameter, as well as a measure of uncertainty in this quantity. *Parameter estimation* on the basis of observations is the most important application of statistics in physics. The second main application is that of *hypothesis testing*, which can consist in, for example, finding out whether the parameter value as predicted by the model is consistent with the value inferred from the experimental observations.

Theoretical models produce answers to questions of the type: "For a given value of the parameter θ, what is the expected distribution for the observable x ?". The experiment has to do with the inverse problem, corresponding to answering questions like: "Given the observations $x_1, x_2, \cdots, x_n$, what is the value of θ ?". Fundamentally different as these questions are, they are nevertheless intimately connected, illustrating the complementary relationship between probability theory and statistics, represented by, respectively, the theoretician and the experimentalist.

The subsequent four chapters of the book are concerned with probability theory, since an account of this subject must be considered an essential prerequisite for a meaningful introduction to statistics. We begin in Chapter 2 by defining different probability concepts and establish rules for the combination of probabilities. In Chapter 3 we consider the properties of probability distributions for random variables in general; we introduce some useful concepts to characterize distributions of a single variable (among them: mean value, variance, and other expectation values or moments) and several variables (in particular: the covariance matrix), develop various formulae pertaining to functions of random variables (rules for error propagation), and give some remarks on sampling. In Chapter 4 we focus on special probability distributions which often serve as mathematical models in experimental situations, notably the

binomial, Poisson, and exponential distributions. Particular attention is given to the normal, or Gaussian, distribution, because of its key role, theoretically (the Central Limit Theorem) as well as practically (describing outcome of measurments). Chapter 5 deals with a class of sampling distributions which are all related to the normal distribution; the most important of these is the chi-square distribution.

The real world seldom fits exactly into the scheme provided by the ideal mathematical models. In Chapter 6 we indicate how different situations can be handled by truncation of the probability distribution, by folding-in of the experimental resolution, and by correcting for inefficiencies in the detecting apparatus.

Passing to the domain of statistics, we begin in Chapter 7 by introducing the important concept of a confidence interval, applying it to the common practical problem of estimating the parameters of the normal distribution. The general aspects of parameter estimation are discussed in Chapter 8, which gives the formal background for the specific estimation methods described in the three subsequent chapters. The Maximum-Likelihood method (Chapter 9), the Least-Squares method (Chapter 10), and the method of moments (Chapter 11) all produce point estimates of the unknown parameters, and we discuss how measures can be obtained for the uncertainty - or error - in these estimates. A point estimate and its error is equivalent to a particular interval estimate of the parameter. Interval estimation in general is also discussed, based on the simplifying assumptions of infinite sample sizes (in the likelihood approach) and linear models (for the Least-Squares estimation), since these facilitate comparisons with the normal and the chi-square distributions. In Chapter 10 we also consider the important issue of constrained parameter estimation, using the technique with Lagrangian multipliers, and discuss how the Least-Squares estimation can provide measures of goodness-of-fit. Chapter 12 describes a simple case study with application of the different estimation methods on simulated polarization experiments.

The Maximum-Likelihood and the Least-Squares estimation methods both require searching the extremum of a function with respect to the unknown parameters. In Chapter 13 we sketch the principles behind commonly used numerical procedures for locating the minima of general functions of many variables.

The second main application of statistics, the testing of statistical hypotheses, is taken up in Chapter 14. In this area of statistical inference the observations are used for decision-making. With a test we mean a given rule or criterion for arriving at a decision of acceptance or rejection of some formulated hypothesis. After a brief survey of the general principles involved, we concentrate on parametric tests for normally distributed variables, which have considerable practical importance. Distribution-free tests of goodness-of-fit between model prediction and experimental observation, or between different sets of observations, include the common χ^2-test and the Kolmogorov-Smirnov tests; the χ^2-test can also be used to test independence between variables and consistency between sets of observations (histograms). Other simple, less well-known prescriptions are given to test randomness within a sample, and consistency between two or more samples.

The Appendix contains tabulations of the most common probability distributions which are referred to throughout the book, as well as a set of tables with percentage points and critical values for some of the test statistics used exclusively in Chapter 14.

There are two important subjects of wide application in particle physics, which are mutually related like probability theory and statistics but not covered in this book. The first of these concerns the simulation of processes and generation of artificial N-particle reactions in the 3N-4 dimensional Lorentz invariant phase space by the Monte Carlo technique. The second deals with methods for analyzing and finding structures in this multi-dimensional space, given a sample of observed N-particle reactions. Readers who are interested in these topics should consult references on Monte Carlo methods and multi-dimensional data analysis given in the bibliography at the end of the book.

2. Probability and statistics

The *theory of probability* is a branch of pure mathematics. From a certain set of axioms and definitions one builds up the theory by deductions. In contrast, *statistics* is a branch of applied mathematics which is essentially inductive. Nevertheless statistics is intimately connected with probability theory as the following considerations may show.

Suppose it is known that when tossing a coin it has an *a priori* probability $p(p=\frac{1}{2})$ of landing "heads" and a probability $1-p$ of landing "tails". We ask: What is the probability of observing r heads out of n tosses? This is a question in probability theory, and an answer is provided by the binomial distribution law, which states that the probability to obtain r heads and $(n-r)$ tails is given by the number

$$B(r;n,p) = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r} .$$

A completely different situation exists if one has no *a priori* knowledge on the probability p and decides to perform an experiment to "determine" this parameter. A simple experiment would consist in tossing the coin repeatedly and counting how many times the outcome heads would occur. It would then be a question of statistics to ask what the parameter p is like, given that in n tosses, r heads were observed. A reasonable answer to this question is to say that "the most likely value" of the parameter is given by the ratio of the number of heads observed to the total number of tosses,

$$p = \hat{p} = \frac{r}{n} .$$

The experiment has then given a *point estimate* $p=\hat{p}$ for the unknown parameter*). We could also say that the value of p was inferred from the observations made by the experiment. However, from the very nature of the experiment, we could not be completely sure that the value $\hat{p}$ thus obtained would be identical to the *true* value of the parameter p. Indeed we would feel that if the experiment were repeated, with new sequences of tosses, then presumably different estimates $\hat{p}$ would be obtained.

Instead of stating the result of the experiment in terms of a single number $\hat{p}$ we could give an *interval estimate* for the parameter p. We would then finalize our experiment by giving two numbers p_1, p_2, hoping that

$$p_1 \leq p \leq p_2$$

represents a true statement about p. The faith we attach to the statement could be expressed by assigning a *confidence level* to it. Given the observation of r heads in n tosses it is again a case of *statistical inference* to determine an interval $[p_1,p_2]$ which is such that it has a certain probability of including the true value of p. In general, the larger we take the interval the more certain we would be that this interval really includes the true value of p, but at the same time a large interval means less precise knowledge on p. On the other hand, a small interval corresponds to a better precision in the determination of p, but the statement that $[p_1,p_2]$ includes p then has a greater chance of not being true.

Since the actual calculation of the limits p_1 and p_2 of a certain confidence interval requires some assumption about the probability for getting just r heads out of n tosses we realize that in order to make this statistical inference it is necessary to know the functional form of the binomial distribution law.

In the following sections we shall state, but not prove, the rules that govern the calculus of probabilities, and illustrate these with some simple examples. Two of the examples, Sects.2.3.10 and 2.3.11, also include statistical inference.

2.1 DEFINITION OF PROBABILITY

Statisticians do not seem to agree about the best way to define probability. We will adopt a rather simple approach, customary among physicists, and define probability in terms of the *limit of relative frequency of*

*) The symbol ^ over a quantity is used to denote an estimate.

occurrence. Thus, if in a sequence of n trials of an experiment the outcome of a specified class, or the event E, occurs r times, then the probability of E is operationally defined as the limiting value

$$P(E) = \frac{r}{n} \quad \text{when } n \to \infty . \tag{2.1}$$

From this definition the probability of the event E is some number satisfying

$$0 \leq P(E) \leq 1 , \tag{2.2}$$

where $P(E) = 0$ if the event never occurs when the experiment is performed, and $P(E) = 1$ if it always occurs.

2.2 RANDOM VARIABLES. SAMPLE SPACE

To illustrate what is meant by a *random variable* consider the simple experiment of rolling a die. Since prior to a throw, its outcome can not be predicted with complete certainty, the number of dots that is observed is called a random variable. In this case the outcome will be one number in the sequence 1,2,...,6, hence the *sample space* consists of the collection of integer numbers between 1 and 6. Since the occurrence of one particular outcome, 1 say, excludes the other possibilities 2,3,...,6 these events are said to be *exclusive*.

If the random variable X can only take on a finite number of values, as in the example above, we call X a discrete random variable. We associate with each possible outcome x_i of the experiment a probability P_i,

$$P(X = x_i) = P_i .$$

Since there must be some outcome of every trial the sum of all P_i for all conceivable outcomes must be equal to one,

$$\sum_i P_i = 1 . \tag{2.3}$$

Exclusive events, for which eq.(2.3) is satisfied, are said to be *exhaustive*.

If the random variable X can have a continuum of values within any finite interval it is called a *continuous* random variable. An example is the

length X obtained from measurements on a bar. We would then associate a probability $P(x \leq X \leq x+dx)$ to the event of getting a measured length in the interval $[x, x+dx]$, and define the *probability density function* $f(x)$ for the continuous random variable X by the equation

$$f(x)dx = P(x \leq X \leq x + dx) .$$

The requirement that all probabilities should add up to one is now formulated by

$$\int_\Omega f(x)dx = 1 , \tag{2.4}$$

where the integration goes over all possible outcomes x defining the sample space Ω.

2.3 CALCULUS OF PROBABILITIES

We shall now introduce some useful concepts and state some basic rules for calculation of probabilities according to *set theory*.

2.3.1 Definitions

The concept of a *set* is used to denote a collection of objects with some common properties. An object that belongs to a set A is said to be an *element* of A. If every element of the set B is also an element of the set A we say that B is a *subset* of A.

Let A be an arbitrary set of elements in the sample space Ω. The *complement* $\bar{A}$ is then defined as the set of all elements in Ω that do not belong to the set A.

The *union* $A \cup B$ of two sets A and B is defined as the set of elements that belong to A or B, or both.

The *intersection* $A \cap B$ is defined as the set of elements that belong to both sets A and B.

Two sets A and B are said to be *exhaustive* if any element of Ω belongs to the union $A \cup B$,

$$A \cup B = \Omega . \qquad \text{(exhaustive sets)} \tag{2.5}$$

Two sets A and B are mutually *exclusive* if they have no elements in common, that is

$$A \cap B = 0. \qquad \text{(exclusive sets)} \qquad (2.6)$$

According to these definitions the set A and its complement $\bar{A}$ are two exclusive and exhaustive sets.

A convenient way of visualizing the concepts introduced above, and also of illustrating the algebra of sets, is by means of *Venn diagrams*.

Let the square of Fig. 2.1 represent the sample space Ω and the areas enclosed by the circles represent the sets A and B. Then the complement $\bar{A}$, the union $A \cup B$, and the intersection $A \cap B$ are given by the different shaded regions in the diagram.

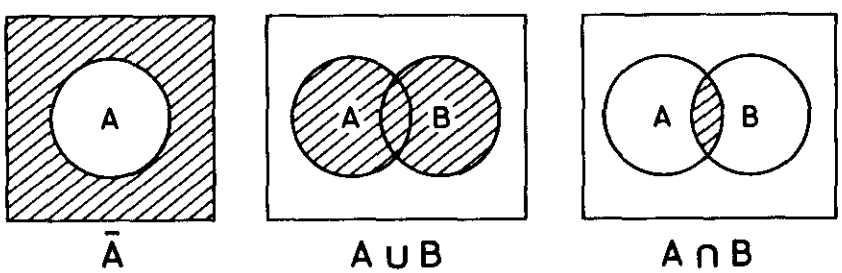

Fig. 2.1. Venn diagram for the complement $\bar{A}$, the union $A \cup B$, and the intersection $A \cap B$.

More complicated combinations of sets can also be conveniently depictured in Venn diagrams.

Exercise 2.1: Let A,B,C be three non-exclusive sets in the sample space Ω. Find representations of the combinations $(A \cup B) \cap C$ and $(A \cap B) \cap C$.

Exercise 2.2: Show, by using the technique of the Venn diagram, that $\overline{(A \cup B) \cap C}$ in general is different from $A \cup (B \cap C)$.

2.3.2 Example: Topologies of bubble chamber events (1)

To illustrate the definitions introduced in the previous section, let us consider the topologies of high-energy proton-proton interactions in a bubble chamber. The reactions will have final states with 2,4,6,... charged particles (prongs) and 0,1,2,... associated neutral strange particles (V^0's). The sample space Ω will be the collection of all conceivable topologies, that is, all combinations of number of prongs and V^0's that can occur at this particular energy.

For definiteness, let the set A be the collection of all events associated with at least one V^0. The complement $\bar{A}$ will then be the collection of events which are not associated with V^0 signals. Similarly, the set B can represent all 2-prong events; the complementary set $\bar{B}$ will then represent all events with more than 2 prongs. The union $A \cup B$ will be the collection of all events that have at least one V^0, or 2-prongs, or both. The intersection $A \cap B$ represents events with 2-prongs and at least one V^0.

Exercise 2.3: Draw Venn diagrams for this example.

2.3.3 Addition rule

From the Venn diagrams in Fig. 2.1 one may write

$$P(A \cup B) = P(A) + P(B) - P(A \cap B), \qquad (2.7)$$

which is known as the *rule of addition* for probabilities. If, in particular, the sets A and B are mutually exclusive, *i.e.* the circles in Fig. 2.1 do not overlap, the addition rule takes the simple form

$$P(A \cup B) = P(A) + P(B); \qquad \text{(exclusive sets)}. \qquad (2.8)$$

If the sets $A_1, A_2, \ldots, A_n$ are exclusive and exhaustive,

$$P(A_1 \cup A_2 \cup \ldots \cup A_n) = \sum_{i=1}^{n} P(A_i) = 1; \quad \text{(exclusive and exhaustive)}. \qquad (2.9)$$

2.3.4 Conditional probability

Suppose A and B are subsets of the sample space Ω and represent the probabilities P(A) and P(B), respectively. Suppose further that we for some reason will be interested only in the elements of A and that we therefore want

to redefine the sample space to the subset A only. How can we then express the probability of the subset B relatively to the "new" sample space A? This new probability is called the *conditional probability* of B relative to A; it is written $P(B|A)$, which should be read "probability of B given A". Another way of expressing the conditional probability, with reference to many applications, is to say that $P(B|A)$ gives the probability that the event B occurs under the condition that the event A has already occurred.

It seems intuitively clear that the conditional probability must have something to do with the "overlap" of A and B, or the intersection $A \cap B$. In fact, the conditional probability $P(B|A)$ is defined through the equation

$$P(A \cap B) = P(B|A)\, P(A). \qquad (2.10)$$

To illustrate the meaning of conditional probability, we look at the Venn diagram of Fig. 2.2. The sample space Ω has a total of N elements and the number of elements in the subsets A and B is N_A and N_B, respectively, while A and B have N_C elements in common.

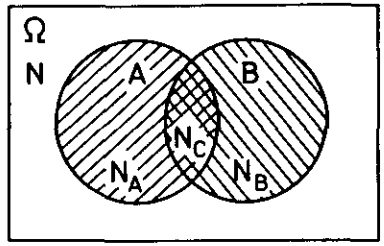

Fig. 2.2. Venn diagram to illustrate conditional probability.

Inspecting the different areas in the diagram we write

$$P(A) = \frac{N_A}{N}, \quad P(B) = \frac{N_B}{N},$$

and

$$P(A \cap B) = \frac{N_C}{N}, \qquad \text{(overlap region relatively to square)}$$

$$P(B|A) = \frac{N_C}{N_A}, \qquad \text{(overlap region relatively to circle A)}$$

$$P(A|B) = \frac{N_C}{N_B}. \qquad \text{(overlap region relatively to circle B)}$$

It is seen that these probabilities satisfy eq.(2.10) as well as the similar equation for the conditional probability $P(A|B)$,

$$P(A \cap B) = P(A|B)\, P(B). \qquad (2.11)$$

It may be worth stressing that <u>all</u> probabilities, in fact, are conditional probabilities, because the occurrence of any event always depends on some conditions. But since these conditions often are the same for all events constituting the sample they are considered to be trivial and therefore not stated explicitly.

2.3.5 Example: $K^o p$ scattering cross section

As an illustration on the implicit use of conditional probability, consider an experiment to determine the sign of the mass difference between the longlived and the shortlived K^o, which can be deduced from the $K^o p$ and $\bar{K}^o p$ scattering cross sections. To evaluate these quantities one can measure the probability that a neutral kaon produced in a hydrogen bubble chamber will interact with a proton within the chamber. The calculated probability for $K^o p$ scattering will, however, depend on the probability for observing the decay of a K^o.

Assume the K^o's to be produced *via* the reaction

$$K^+ + p \to K^o + \pi^+ + p, \qquad \text{(production)},$$

and detected through the decay

$$K^o \to \pi^+ + \pi^-, \qquad \text{(decay, event B)}.$$

The interesting reaction is

$$K^o + p \to K^o + p, \qquad \text{(scatter, event A)},$$

for which we search the probability $P(A)$. We are only able to identify event A if event B is also observed, because the decaying kaon as well as the recoiling proton must be measured to obtain a kinematical fit to the scattering reaction. Writing (eq.(2.10))

$$P(A \cap B) = P(B|A)\, P(A)$$

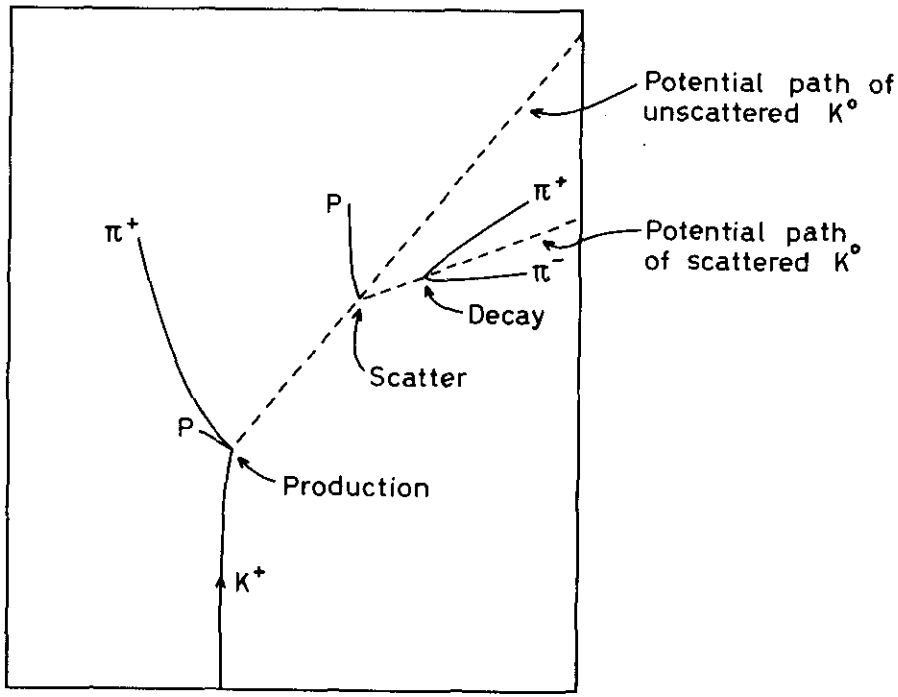

Fig. 2.3. Sketch of sequential events observed in a hydrogen bubble chamber. An incoming K^+ meson gives rise to a K^o *via* the reaction $K^+p \to K^o\pi^+p$. The K^o scatters off a proton, $K^op \to K^op$, and decays into a pair of pions, $K^o \to \pi^+\pi^-$.

the intersection $P(A \cap B)$ represents the fraction of completely identified sequential events,

$$K^+p \to K^o\pi^+p$$
$$\hookrightarrow K^op \to K^op$$
$$\hookrightarrow \pi^+\pi^-$$

relatively to the total number of K^o producing reactions $K^+p \to K^o\pi^+p$ with K^o seen or unseen in the bubble chamber. Thus the probability $P(A \cap B)$ can be found by counting events.

On the other hand $P(B|A)$, the conditional probability for a seen decay of the K^o, given that a scattering has occurred, can be calculated from the identified, complete events taking into account the geometric detection efficiency of the bubble chamber, the mean lifetime of the K^o, and the branching ratio for the charged decay mode. Note that $P(B|A)$ is different from the probability $P(B)$ to observe the decay of an unscattered K^o, due to the change in the K^o momentum and direction in the scattering process. Compare the illustration, Fig. 2.3, where the sequential events are shown, together with indications of the potential paths for the scattered K^o as well as for an unscattered K^o.

2.3.6 Independence; multiplication rule

Two sets A and B are said to be *independent* if the conditional probability of B relative to A is equal to the probability of B,

$$P(B|A) = P(B), \qquad \text{(independence)}. \tag{2.12}$$

This means that the occurrence of the event B is not dependent on the (previous) occurrence of event A. From our earlier definition of conditional probability, eq.(2.10), we see that an alternative formulation of independence for two sets is

$$P(A \cap B) = P(A)\cdot P(B), \qquad \text{(independence)}. \qquad (2.13)$$

In other words, the probability for the occurrence of both events A and B is equal to the product of the probabilities for the two separate events.

Eq.(2.13) provides a necessary and sufficient condition for independence of the two sets A and B. It is often referred to as the *rule of multiplication* for probabilities of independent events.

2.3.7 Example: Relay networks

In Fig. 2.4 the probability for the closing of each relay in the circuit is some given number α. Assuming that all relays act independently we want to find the probability for the flow of a current between the terminals.

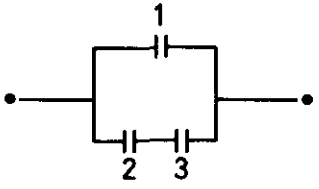

Fig. 2.4. Relay network

Let the event of closing relay i be denoted by E_i, i=1,2,3. Then $P(E_1) = P(E_2) = P(E_3) = \alpha$. Having a current between the terminals corresponds to the event $E = E_1 \cup (E_2 \cap E_3)$, for which we shall find

$$P(E) = P\big(E_1 \cup (E_2 \cap E_3)\big).$$

Applying the addition rule for probabilities, eq.(2.7), leads to

$$P(E) = P(E_1) + P(E_2 \cap E_3) - P\big(E_1 \cap (E_2 \cap E_3)\big),$$

and using the multiplication rule eq.(2.13) for the independent events E_i we obtain

$$P(E) = P(E_1) + P(E_2)P(E_3) - P(E_1)P(E_2)P(E_3) = \alpha + \alpha^2 - \alpha^3.$$

2.3.8 Example: Efficiency of a Čerenkov counter

Suppose the Čerenkov light from particles traversing a Čerenkov counter along its axis is detected by a concentric arrangement of phototubes as illustrated in Fig. 2.5.

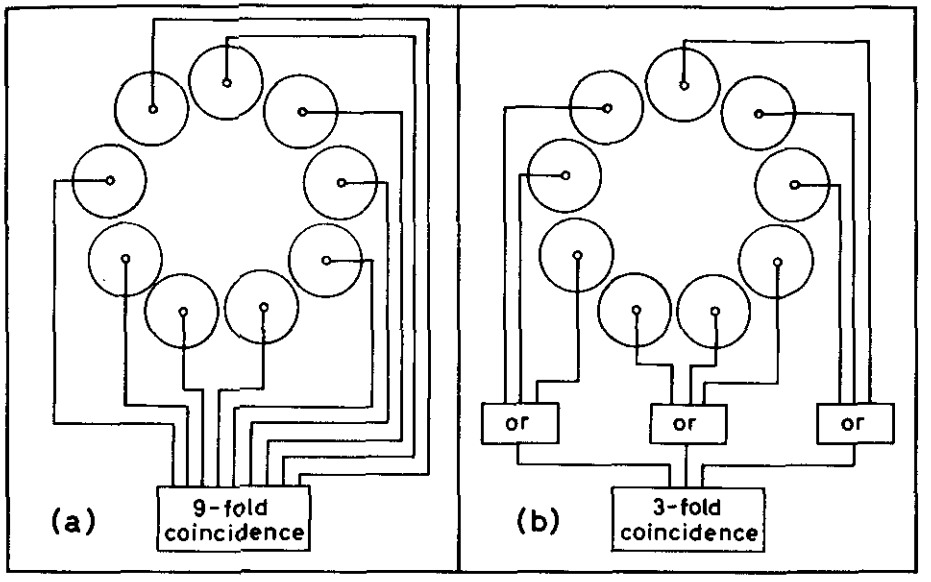

Fig. 2.5. Arrangements of 9 phototubes in a Čerenkov counter.

In order to discriminate against accidental triggering of the system it is desirable to observe coincidences between the signals from several phototubes. We assume all phototubes to act independently.

If the event E of having a signal from one phototube corresponds to a probability $P(E) = \varepsilon = 0.93$, the probability for the detection of the Čerenkov light by all 9 independent phototubes in Fig. 2.5(a) is

$$P_a = \big(P(E)\big)^9 = \varepsilon^9 = 0.52.$$

The efficiency of the detector can be largely improved by a different arrangement of the phototubes. In Fig.2.5(b) the tubes are grouped together three

by three, and each group is activated if at least one of the tubes in the group has a signal. The observation of a particle by the detector then requires a coincidence between the signals from the three groups, for which the probability becomes

$$P_b = \left(P(E \cup E \cup E)\right)^3 = (3\varepsilon - 3\varepsilon^2 + \varepsilon^3)^3 = 0.999 .$$

2.3.9 Example: π^o detection

The π^o meson is known to decay electromagnetically into two γ-quanta. Suppose that we study π^o decays in some detector, *e.g.* a heavy liquid bubble chamber, and that the average probability is α for the conversion of a γ into an electron-positron pair within the detector. We want first to find the probabilities for seeing two, one, or none of the decay products of a single π^o.

Clearly the conversion of different γ-rays can be assumed to occur independently of each other. Therefore, in a simple application of the multiplication rule, eq.(2.13), we can write down the probability to detect both γ's from the decaying π^o as

$$P(2\gamma) = \alpha^2$$

and, similarly, the probability for seeing none of them,

$$P(0\gamma) = (1-\alpha)^2 .$$

The probability that only one of the γ's is visible is also easily written down,

$$P(1\gamma) = 2\alpha(1-\alpha) .$$

The three probabilities can be added to give

$$P(2\gamma) + P(1\gamma) + P(0\gamma) = 1 ,$$

as they should. We also note that the probability to see at least one γ from a decaying π^o is

$$P(\geq 1\gamma) = P(2\gamma) + P(1\gamma) = 2\alpha - \alpha^2 ,$$

which can be rewritten as

$$P(\gamma_1 \cup \gamma_2) = P(\gamma_1) + P(\gamma_2) - P(\gamma_1)\cdot P(\gamma_2) ,$$

to show an application of the addition rule, eq.(2.7).

The same line of thought may be applied to more complex situations where several π^o's are involved. For instance, an ω^o meson, with a decay into $3\pi^o$ will give rise to up to six detected γ-rays. However, unless the detector is very efficient, the probability to see many decay products soon becomes very small.

2.3.10 Example: Beam contamination and delta rays

It is quite often a problem in kaon and antiproton exposures in bubble chambers to find the contamination in the beam of lighter particles, pions and muons. Since a light particle can impact more of its energy to an electron than a heavy particle of the same momentum, it is possible to estimate the contamination from a count of beam tracks having delta ray electrons with an energy E exceeding the maximum possible, E_{max}, for a delta ray produced by the heavier particle. Hence the presence of a delta ray with $E > E_{max}$ ("large δ") indicates a "light" beam track.

We introduce the following notation:

N = total number of beam tracks observed,

N_1 = number of beam tracks observed with one "large δ",

N_2 = number of beam tracks observed with two "large δ",

N_π = number of "light" beam tracks (unknown),

$P(1\delta)$ = probability that a light particle produces one "large δ",

$P(2\delta)$ = probability that a light particle produces two "large δ".

Then, by definition, in the limit of large numbers,

$$P(1\delta) = \frac{N_1}{N_\pi} , \qquad P(2\delta) = \frac{N_2}{N_\pi} .$$

The law of independence, eq.(2.13), implies that

$$P(2\delta) = \left(P(1\delta)\right)^2 .$$

Thus, combining the expressions above one sees that it is possible to estimate the number of "light" beam tracks as

$$N_\pi = \hat{N}_\pi = \frac{N_1^2}{N_2} .$$

Hence the contamination is estimated as

$$F = \hat{F} = \frac{\hat{N}_\pi}{N} = \frac{N_1^2}{NN_2} .$$

2.3.11 Example: Scanning efficiency (1)

In experimental high energy physics employing track detection methods (bubble chamber, nuclear emulsion, spark chamber) the experimenter must rely on scanners who search for specified types of events. The interesting events occur at random among large quantities of extraneous information, and it is rather unlikely that the scanner will be able to register all events of the specified types. The scanning process is therefore probably less than 100% effective. To estimate the probability that the interesting events are recorded one can perform two, or more, completely independent scans of the material.

Suppose that from two independent scans of a given sample of bubble chamber pictures one divides the interesting events into four exclusive classes, where

$N_{12} + N_1$ events were found in the first scan,

$N_{12} + N_2$ events were found in the second scan,

N_{12} events were found in both scans,

$N - (N_1+N_2+N_{12})$ events are undetected by the two scans.

N is the unknown number of interesting events in the film. N_1 is the number of events that are found in the first scan but not in the second, and similarly for N_2. A diagrammatic representation of the situation is given in Fig. 2.6.

From our definition of probability, eq.(2.1), the efficiencies of the individual scans are, respectively, assuming large numbers,

$$P(1) = \frac{N_{12} + N_1}{N} , \qquad P(2) = \frac{N_{12} + N_2}{N} . \tag{2.14}$$

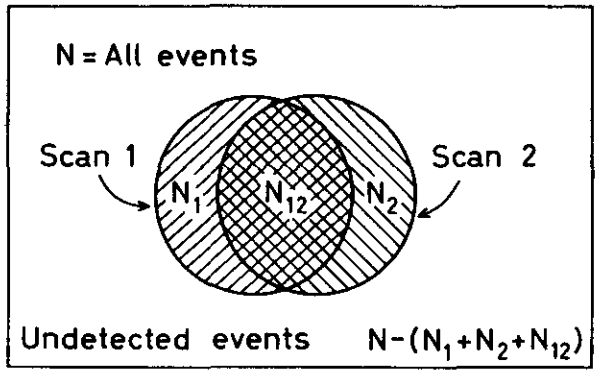

Fig. 2.6. Venn diagram for illustration of scanning efficiency.

The probability that an event has been found in both scans is given by the intersection (see Fig. 2.6)

$$P(1 \cap 2) = \frac{N_{12}}{N} . \tag{2.15}$$

Since the scans were assumed to be independent, the probability that an event will be observed in both scans is also given by the multiplication rule, eq. (2.13),

$$P(1 \cap 2) = P(1)\cdot P(2) . \tag{2.16}$$

Combining the relations (2.14) - (2.16) we find an estimate for the total number of events in the film, by taking

$$N = \hat{N} = \frac{(N_{12} + N_1)(N_{12} + N_2)}{N_{12}} . \tag{2.17}$$

Hence, inserting this expression in eq.(2.14) we obtain the estimated efficiencies of the individual scans as

$$\hat{P}(1) = \frac{N_{12}}{N_{12} + N_2} , \qquad \hat{P}(2) = \frac{N_{12}}{N_{12} + N_1} . \tag{2.18}$$

It is worth noting that because N has to be estimated using the information from both scans the estimated efficiency for the first scan depends on the result of the second scan, and *vice versa*.

From the total number of events found in the two scans, the *overall* scanning efficiency is given by the union,

$$P(1 \cup 2) = \frac{N_{12} + N_1 + N_2}{N} . \tag{2.19}$$

Thus, taking the estimated value for N from eq.(2.17) we arrive at the following expression for the estimated combined efficiency of the two scans,

$$\hat{P}(1 \cup 2) = \frac{N_{12}(N_{12} + N_1 + N_2)}{(N_{12} + N_1)(N_{12} + N_2)} . \tag{2.20}$$

The overall efficiency can alternatively be expressed in terms of the individual efficiencies; applying eq.(2.7) we have

$$P(1 \cup 2) = P(1) + P(2) - P(1 \cap 2)$$

which, using eq.(2.16), gives

$$P(1 \cup 2) = P(1) + P(2) - P(1) \cdot P(2) . \tag{2.21}$$

In developing the formulae above several assumptions were involved, some of which have not been stated explicitly and may hardly be fulfilled in practice. For instance, it has been assumed that

(i) the different scans are performed independently (an assumption that is perhaps best met by having different scanners),

(ii) all events have the same probability of being detected (an ideal requirement which is not achieved in practice when complicated topologies occur),

(iii) the number of events is sufficiently large to warrant the use of the concept "probability".

We briefly touch the question of the statistical uncertainties in the scanning efficiencies in connection with our discussion of the binomial distribution (Sect.4.1.3). The whole problem of estimating scanning efficiencies is taken up again in Sect.9.6.2 where we adopt the Maximum-Likelihood approach to the estimation problem.

2.3.12 Marginal probability

The elements of a set may frequently be classified according to more than one criterion. The term *marginal probability* is then used whenever some of the criteria are being ignored in the classification. Thus if the classifications according to two criteria A and B are $A_1, A_2, \ldots, A_m$ and $B_1, B_2, \ldots, B_n$, respectively, where $\sum_{i=1}^{m} P(A_i) = \sum_{j=1}^{n} P(B_j) = 1$, then the marginal probability of A_i is

$$P(A_i) = \sum_{j=1}^{n} P(A_i \cap B_j) \tag{2.22}$$

and, similarly, the marginal probability of B_j,

$$P(B_j) = \sum_{i=1}^{m} P(A_i \cap B_j) . \tag{2.23}$$

In particle physics the concept of marginal probability is inherent in the notion of *inclusive reactions*. For example, writing

$$a + b \rightarrow c + \text{anything}$$

implies that in the reaction of particles a and b one is only interested in observations on the properties of particle c, ignoring all the other particles.

A more specific example follows.

2.3.13 Example: Topologies of bubble chamber events (2)

An experiment on strange particle production in proton-proton reactions in a bubble chamber has classified the events according to the identified neutral strange particle (V^0), criterion A, and the number of charged particles (prongs) seen in the primary reaction, criterion B. Criterion A gives four exclusive possibilities, with observation of one K_s^0, one Λ, two K_s^0, or one K_s^0 plus one Λ, respectively, while criterion B gives five (exclusive) possible prong number assignments for the primary reaction. Suppose that the probabilities $P(A_i \cap B_j)$ are given by the observed relative frequencies for the various topologies displayed in the following table:

		B				
		2-prong	4-prong	6-prong	8-prong	≥10-prong
A	K_s^o	0.117	0.169	0.055	0.006	0.
	Λ	0.180	0.220	0.092	0.012	0.001
	$K_s^o K_s^o$	0.017	0.014	0.002	0.	0.
	$K_s^o \Lambda$	0.055	0.045	0.014	0.001	0.

The marginal probability to have a reaction with, say, only one K_s^o, irrespective of the number of prongs, is here found by adding the entries in the first row of the table,

$$P(A_1) = P(K_s^o) = 0.347,$$

whereas, for example, the marginal probability for having a 4-prong reaction with any V^o signal is found by adding the numbers in the second column,

$$P(B_2) = P(4\text{-prong}) = 0.448.$$

It will be seen that the marginal probabilities add to unity,

$$\sum_{i=1}^{4} P(A_i) = \sum_{j=1}^{5} P(B_j) = 1.000 .$$

2.4 BAYES' THEOREM

We shall give a brief account of *Bayes' Theorem*, which, although mathematically simple, has a controversial status among the specialists. We first state the theorem, next prove it (Sect.2.4.1), consider an example (Sect. 2.4.2), give further comments (Sect.2.4.3) and a second example (Sect.2.4.4).

2.4.1 Statement and proof

Let the sample space Ω be spanned by the n mutually exclusive and exhaustive subsets B_i,

$$\sum_{i=1}^{n} P(B_i) = 1 .$$

If A is also a set belonging to Ω, Bayes' Theorem states that

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{n} P(A|B_j)P(B_j)} . \qquad (2.24)$$

The proof of this theorem is nothing but an application of the definitions introduced in the previous sections. From the definition of conditional probability we can write down two expressions for the intersection $P(A \cap B_i)$,

$$P(A \cap B_i) = P(B_i|A)P(A) = P(A|B_i)P(B_i) ,$$

(compare eqs.(2.10), (2.11)), hence

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A)} .$$

For P(A) we get an expression from the definition of marginal probability, eq. (2.22), which can be rewritten using eq.(2.11),

$$P(A) = \sum_{j=1}^{n} P(A \cap B_j) = \sum_{j=1}^{n} P(A|B_j)P(B_j) .$$

Substituting for $P(A)$ in $P(B_i|A)$ above is seen to give eq.(2.24).

2.4.2 Example

Let each of three drawers B_1, B_2, B_3 contain two coins; B_1 has two gold coins, B_2 one gold and one silver, B_3 two silver coins. We are to select one drawer at random and pick a coin from it. Supposing that this first coin turns out to be one of gold, what is then the probability that the second coin in the same drawer is also a gold coin?

If A denotes the event of first picking a gold coin we want to calculate the conditional probability $P(B_1|A)$. Obviously the conditional probabilities $P(A|B_i)$ of getting a gold coin from drawer B_i, are

$$P(A|B_1) = 1, \quad P(A|B_2) = \tfrac{1}{2}, \quad P(A|B_3) = 0,$$

respectively. Also, since a drawer is selected at random

$$P(B_1) = P(B_2) = P(B_3) = 1/3 .$$

Hence, from eq.(2.24),

$$P(B_1|A) = \frac{P(A|B_1)P(B_1)}{\sum_{j=1}^{3} P(A|B_j)P(B_j)} = \frac{1\cdot 1/3}{1\cdot 1/3 + \frac{1}{2}\cdot 1/3 + 0\cdot 1/3} = 2/3 .$$

Thus, although the probability was only 1/3 for selecting drawer B_1, the observation that the first coin drawn was a gold coin effectively doubles the probability that the drawer B_1 had been selected.

This example may serve to justify the following notation: The $P(B_i)$ are the *prior probabilities* for the drawers, while $P(B_1|A)$ is the *posterior probability* for the drawer B_1. The probability $P(A|B_i)$ is called the *likelihood* of the event A, given B_i.

2.4.3 Comments

Bayes' Theorem as it has been stated by eq.(2.24) is nothing but a simple and logical consequence of the rules for calculation of probabilities. When it comes to the application of Bayes' Theorem to estimation problems, however, different schools of statisticians have different interpretations of the theorem with far-reaching philosophical implications, and the apparently unconciliatory opinions among specialists have led to a vast literature on the subject.

Let us, in eq.(2.24), make a change in notation and write x instead of A, θ_i instead of B_i. We may think of x as a random variable, to be measured by an experiment; the θ_i may represent different hypotheses about a parameter θ, whose actual numerical value is not known. The set of hypotheses θ_i satisfies the condition for exclusive and exhaustive sets,

$$\sum_{i=1}^{n} P(\theta_i) = 1 . \qquad (2.25)$$

Bayes' Theorem relates these prior probabilities for the hypotheses to their posterior probabilities $P(\theta_i|x)$, given the measurement x,

$$P(\theta_i|x) = \frac{P(x|\theta_i)P(\theta_i)}{\sum_{j=1}^{n} P(x|\theta_j)P(\theta_j)} , \qquad (2.26)$$

where the conditional probability $P(x|\theta_i)$ is the likelihood of obtaining the measurement x for a specified hypothesis θ_i.

From eq.(2.26) any posterior probability $P(\theta_i|x)$ can only be evaluated if <u>all</u> prior probabilities $P(\theta_i)$ are specified. The numerical value of $P(\theta_i)$ is a measure of our prior degree of belief that θ_i is a true hypothesis about the unknown θ. The measurement of x implies a change in our belief that θ_i is true, $P(\theta_i|x)$ giving a numerical value for our new faith in the statement that θ_i is true. If, on the basis of the observation x, we had to choose one hypothesis from the set $\theta_1,\theta_2,\ldots,\theta_n$ we would probably choose the one with the largest posterior probability $P(\theta_i|x)$.

A consequence of Bayes' Theorem is that two physicists, from the same observation x, may find different distributions of posterior probabilities, depending on their formulation of the prior probabilities. From their individual experience and prior knowledge they may arrive at different posterior knowledge, and hence they may, quite legitimately, reach different conclusions about the parameter θ.

To avoid the element of subjectivism in the inference about θ it would be necessary to agree on the prior probabilities $P(\theta_i)$. Thus all physicists would have to combine their prior knowledge, deciding on an "objective" prior distribution of the $P(\theta_i)$. This distribution would, however, still be conditioned in the sense that every $P(\theta_i)$ would be dependent on all knowledge accumulated up to the present *).

The preceding considerations reflect the attitude of Bayesian statisticians. The anti-Bayesians, on the other hand, prefer to abstain from posterior probabilities, being satisfied with destilling and presenting their data as indicative as possible, but leaving the (subjective) conclusions to be drawn by the reader.

In many cases the prior probabilities $P(\theta_i)$ are only incompletely

*) Recall our earlier remark that all probabilities indeed are conditional.

known. If the denominator of eq.(2.26) can not be found, Bayes' Theorem takes the weaker form

$$P(\theta_i|x) \propto P(x|\theta_i)P(\theta_i). \tag{2.27}$$

Although one can not in this case completely determine the posterior probabilities, eq.(2.27) can nevertheless be useful for calculation of relative posterior probabilities. With the observation x, the relative posterior probabilities for θ_i and θ_j define the *betting odds* for the hypothesis θ_i against the hypothesis θ_j,

$$\text{Betting odds of } \theta_i \text{ against } \theta_j = \frac{P(\theta_i|x)}{P(\theta_j|x)} = \frac{P(x|\theta_i)P(\theta_i)}{P(x|\theta_j)P(\theta_j)} . \tag{2.28}$$

An example follows.

2.4.4 Example: Betting odds

Let us go back to the experiment described in Sect.2.3.13 and assume that the numbers in the last two rows of the table indicate *a priori* probabilities for the observation of $K^o_s K^o_s$ and $K^o_s \Lambda$ pairs, respectively.

Suppose further that a new event with two V^o's has been found in the film, and that the subsequent measurement and kinematical analysis of this event shows that while one of the V^o's is uniquely identified as a K^o_s, the second V^o is ambiguous between a K^o_s and a Λ assignment. The probabilities following from the identification and measurements x have been found to be $P(x|K^o_s) = 0.10$ and $P(x|\Lambda) = 0.50$ for the two possibilities. We want to find the betting odds for the hypothesis that the $2V^o$ event is a $K^o_s K^o_s$ pair against the hypothesis that the event is a $K^o_s \Lambda$ pair. From the table in Sect.2.3.13, ignoring the information on the charged tracks of the primary reaction, we find, using eq.(2.28),

$$\frac{P(K^o_s|x)}{P(\Lambda|x)} = \frac{P(x|K^o_s)P(K^o_s K^o_s)}{P(x|\Lambda)P(K^o_s \Lambda)} = \frac{0.10 \cdot 0.033}{0.50 \cdot 0.115} = 0.057 .$$

2.4.5 Bayes' Postulate

From our previous discussion of Bayes' Theorem it is clear that no statement can be made about the parameter θ on the basis of the observation x

if the prior knowledge $P(\theta_i)$ is completely missing. In practice, however, meaningful posterior statements about θ can frequently be made also when the prior knowledge is scanty; this is particularly so if the $P(\theta_i)$ are roughly of the same magnitude and the likelihoods $P(x|\theta_i)$ are very dissimilar for the different hypotheses θ_i.

Bayes' Postulate says that if the distribution of prior probabilities is completely unknown, one may take

$$P(\theta_1) = P(\theta_2) = \ldots = P(\theta_n). \tag{2.29}$$

Then Bayes' Theorem will express the posterior probabilities by the likelihoods alone,

$$P(\theta_i|x) = \frac{P(x|\theta_i)}{\sum_{j=1}^{n} P(x|\theta_j)} . \tag{2.30}$$

As we have argued before, given the choice between different hypotheses θ_i, we would choose the one with the largest $P(\theta_i|x)$. The resemblance to the Maximum-Likelihood method at this point is, however, only artificial, because in the present context, θ is a parameter, whereas in the Maximum-Likelihood approach to the estimation problem θ is regarded as a variable.

Choosing all prior probabilities equal according to Bayes' Postulate eq.(2.29) looks clearly quite arbitrary, and may lead to logical inconsistencies under some circumstances. Consider, for instance, the decay of unstable particles, which may be described in terms of the mean lifetime τ of the particles, or the decay constant λ, where $\lambda = \frac{1}{\tau}$. If we wish to determine τ, having no prior information on this quantity, Bayes' Postulate suggests that we take the prior distribution $P(\tau) = \text{constant}$. If instead λ was chosen to describe the decay, and we had no prior information on this quantity, then Bayes' Postulate would suggest that we use the prior distribution $P(\lambda) = \text{constant}$. But this is quite different from the previous suggestion, because

$$P(\tau) = P(\lambda) \cdot \left|\frac{d\lambda}{d\tau}\right| = \lambda^2 \cdot P(\lambda).$$

As we shall see later in our discussion of the Maximum-Likelihood method the prior probabilities need not represent a problem; see Sect. 9.4.1.

3. General properties of probability distributions

In this chapter we shall be concerned with the formulation of some basic concepts and definitions from probability theory which are also important to applications in statistics. Without specializing our assumptions on the actual form of the probability distribution, we shall discuss a set of general features which are common to most types of probability distributions to be studied later. A number of specific distributions are treated in more detail in the subsequent chapters, which also include examples of practical interest. The present chapter, with its minimum of applications, should be regarded as a rather complete reference list of the general definitions and properties which will be applied and developed further in the remainder of the book.

3.1 THE PROBABILITY DENSITY FUNCTION

From the very outset we will assume that the random variables to be considered are of the continuous type.

For the moment it will suffice to consider a single, continuous random variable X. We assume that X can have any value over a certain domain Ω. The probability that the random variable comes out with a value in the particular interval [x,x+dx] is written

$$P(x < X < x + dx) = f(x)dx\,. \tag{3.1}$$

Here f(x) must represent "probability per unit length" and is called the *probability density function* for x. Since the total probability that the variable will have some value in Ω must be equal to one, the normalization condition is

$$\int_{\Omega} f(x)dx = 1\,. \tag{3.2}$$

The way f(x) is introduced guarantees it to have a non-negative dependence on x. We shall also assume that the functional dependence on x is unique,

that is, f(x) is a single-valued function of x. In our further applications f(x) is anticipated to be sufficiently regular, so as to allow differentiations with respect to x. In short, f(x) is assumed to be a well-behaved function.

The probability density function will be our main topic in the following. We shall often use the notion "p.d.f.", or just "distribution", when no confusion is possible.

From this point we shall also simplify notation and denote the random variable itself as well as its specific values (observations) by lower-case letters.

3.2 THE CUMULATIVE DISTRIBUTION FUNCTION

Instead of characterizing the random variable by its probability density function f(x) one may use the *cumulative distribution* F(x), defined by *)

$$F(x) \equiv \int_{x_{min}}^{x} f(x')dx'\,, \tag{3.3}$$

where x_{min} is the lower limit value of x. Since f(x) is always non-negative F(x) is clearly a monotonic increasing function of x over the interval $x_{min} \leq x \leq x_{max}$. One has

$$F(x_{min}) = 0, \qquad F(x_{max}) = 1. \tag{3.4}$$

3.3 PROPERTIES OF THE PROBABILITY DENSITY FUNCTION

The probability density function f(x) contains all information about the random variable x. Various properties of f(x) serve to characterize the distribution. For instance, a *mode* of the distribution is a value of x which maximizes the p.d.f.. If there is only one mode the function is *unimodular*. The *median* of the p.d.f. is defined as the value of x for which

$$F(x_{median}) = \frac{1}{2}\,. \tag{3.5}$$

---

*) Statisticians often prefer the inverse definition and put

$$f(x) \equiv \frac{dF}{dx}$$

referring to F(x) as the "distribution function".

Particularly useful are the *mean* as a measure of the central value and the *variance* as a measure of the spread of the distribution. These concepts are introduced below.

Fig. 3.1 shows the relative position of the location parameters (mode, median, mean) for a unimodular p.d.f..

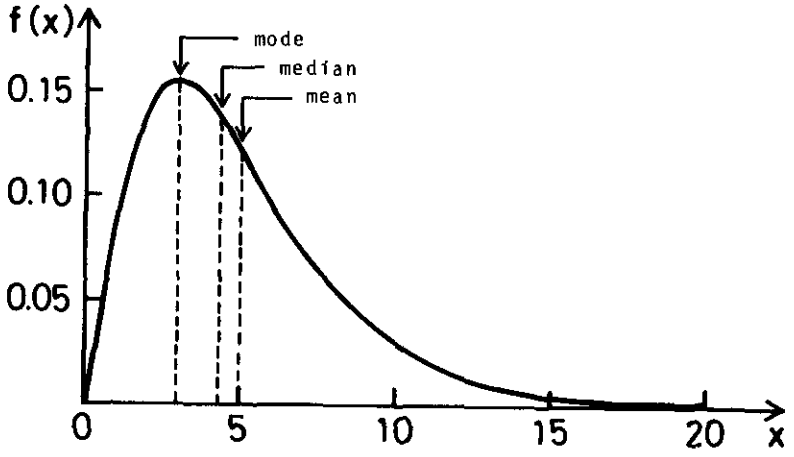

Fig. 3.1. Location parameters for a unimodular probability density function. (The curve corresponds to a chi-square p.d.f. for 5 degrees of freedom.)

3.3.1 Expectation values of a function

Let g(x) be some function of x. We define the mathematical *expectation value* (or *expected value*, or *expectation*) of the function g(x) for the p.d.f. f(x) by

$$E\big(g(x)\big) \equiv \int_{\Omega} g(x)f(x)dx\,, \tag{3.6}$$

where the integration is over the entire domain of the variable. Thus f(x) serves as a weighting function for g(x), and the resulting quantity $E\big(g(x)\big)$ is a *number*, *i.e.* a constant, independent of x. We see that the expectation $E\big(g(x)\big)$ is a measure of the mean or central value of the function g(x).

From the definition of the expectation value one easily derives the following relations:

$$E(a) = a, \qquad \text{where a is a constant}$$

$$E\big(ag(x)\big) = a\,E\big(g(x)\big),$$

$$E\big(a_1g_1(x) + a_2g_2(x)\big) = a_1E\big(g_1(x)\big) + a_2E\big(g_2(x)\big).$$

Thus E has the properties of a linear operator.

An important application of the expectation operator is to derive the expected value of the square of the difference between g(x) and its expectation $E\big(g(x)\big)$. We will then get a measure of the spread or dispersion of g(x) about its central value, which is called the *variance* of g(x) for the p.d.f. f(x):

$$V\big(g(x)\big) \equiv E\Big(g(x) - E\big(g(x)\big)\Big)^2. \tag{3.7}$$

We next specify some definite forms of g(x) which are particularly useful.

3.3.2 Mean value and variance of a random variable

As a simple application we put g(x) = x in the general definition eq. (3.6). We then have the expectation of the random variable itself, which is called the *mean value* of x for the p.d.f. f(x); this number we denote by μ:

$$\mu \equiv E(x) = \int_{\Omega} xf(x)dx\,. \tag{3.8}$$

For the spread of x about its mean value we take the *variance* V(x), or *dispersion*, of x for the p.d.f. f(x); this number is denoted by σ^2:

$$\sigma^2 \equiv V(x) = E(x-\mu)^2 = \int_{\Omega} (x-\mu)^2 f(x)dx\,. \tag{3.9}$$

Clearly σ^2 is a non-negative quantity; σ is called the *standard deviation* of x for the p.d.f. f(x).

Using the linearity property of E we can establish a useful relation between the expectation of x^2 and the parameters σ^2, μ. From eq.(3.9) we find

$$\sigma^2 = E(x-\mu)^2 = E(x^2-2\mu x+\mu^2) = E(x^2) - \mu^2, \tag{3.10}$$

or

$$E(x^2) = \sigma^2 + \mu^2. \tag{3.11}$$

It may be instructive to note that one has

$$E(x-\mu)^2 = E(x^2) - \mu^2 = E(x^2) - \big(E(x)\big)^2. \tag{3.12}$$

Exercise 3.1: Show that, if a is a constant, $V(ax) = a^2V(x)$.

Exercise 3.2: (The Bienaymé-Tshebycheff inequality)
Let g(x) be a non-negative function of the random variable x with p.d.f. f(x) and variance σ^2. Prove that the probability for g(x) to be at least as large as any constant value c is limited if $E\big(g(x)\big)$ exists,

$$P\big(g(x) \geq c\big) \leq \frac{1}{c} E\big(g(x)\big) .$$

In particular, with $g(x)=(x-E(x))^2$, this is equivalent to

$$P\big(|x-E(x)| \geq \lambda\sigma\big) \leq \frac{1}{\lambda^2} ,$$

which is called the *Bienaymé-Tshebycheff inequality*.
This general result on the probability for $|x-E(x)|$ to exceed a given number of standard deviations turns out to be very useful for proofs of limiting properties and convergence theorems; see Sect.3.10.4.

3.3.3 General moments

Because of their simple interpretation $\mu = E(x)$ and $\sigma^2 = E(x-\mu)^2$ are extensively used as parameters in the probability density function.

For theoretical and practical purposes it is also convenient to define expectation values of other powers of x and $(x-\mu)$. With a general definition we shall call the expectation of x^k *the k-th moment of f(x) about the origin*, or *the k-th algebraic moment*,

$$\mu_k' \equiv E(x^k) = \int_\Omega x^k f(x)dx . \tag{3.13}$$

In a similar way the expectation of $(x-\mu)^k$ is called *the k-th moment of f(x) about the mean* or *the k-th central moment*,

$$\mu_k \equiv E(x-\mu)^k = \int_\Omega (x-\mu)^k f(x)dx . \tag{3.14}$$

The mean value and the variance are therefore special examples of more generally defined moments. In particular, the mean is equal to the first moment of f(x) about the origin, whereas the variance is equal to the second moment of f(x) about the mean,

$$\mu = \mu_1' , \qquad \text{(first algebraic moment)}$$

$$\sigma^2 = \mu_2 . \qquad \text{(second central moment)}$$

The moments of lowest order are easily derived from the definitions:

$$\left.\begin{aligned} \mu_0' &= 1 \\ \mu_1' &= \mu \\ \mu_2' &= \sigma^2 + \mu^2 \end{aligned}\right\} \qquad \text{(algebraic moments)}, \tag{3.15}$$

and

$$\left.\begin{aligned} \mu_0 &= 1 \\ \mu_1 &= 0 \\ \mu_2 &= \sigma^2 \end{aligned}\right\} \qquad \text{(central moments)}. \tag{3.16}$$

Note in particular the relation between the second moments:

$$\mu_2 = \mu_2' - \mu^2. \tag{3.17}$$

This is the same result as eq.(3.12).

The general relations between central moments and algebraic moments are as follows:

$$\mu_k = \sum_{r=0}^{k} \binom{k}{r} \mu_{k-r}'(-\mu_1')^r \qquad \text{(algebraic moments known)}, \tag{3.18}$$

$$\mu_k' = \sum_{r=0}^{k} \binom{k}{r} \mu_{k-r}(\mu_1')^r \qquad \text{(central moments known)}. \tag{3.19}$$

Here $\mu_1' = \mu$, in accordance with eq.(3.15).

The moments of higher order become of interest when one wants to study the behaviour of f(x) for large $|x-\mu|$. For a symmetric distribution all odd central moments vanish. Any odd central moment which is not zero may therefore be taken as a measure of the asymmetry or skewness of the distribution. The simplest of these measures is the third central moment μ_3, but its numerical value will depend to the third order on the units of the variable x. To have an absolute and dimensionless measure of the asymmetry one, therefore, defines the *asymmetry coefficient*, or *skewness*, by

$$\gamma_1 \equiv \frac{\mu_3}{(\mu_2)^{3/2}} = \frac{E(x-\mu)^3}{\sigma^3} , \tag{3.20}$$

Clearly a positive value of γ_1 implies that the distribution f(x) has a tail to the right of the mean value, whereas a negative γ_1 indicates a tail to the left.

The *coefficient of kurtosis*, or *peakedness*, of f(x) is defined by the dimensionless quantity

$$\gamma_2 \equiv \frac{\mu_4}{(\mu_2)^2} - 3 = \frac{E(x-\mu)^4}{\sigma^4} - 3 . \tag{3.21}$$

This definition implies a comparison with the normal or Gaussian p.d.f., which has $\gamma_2 = 0$, (see Sect. 4.8.4). A positive (negative) value of γ_2 indicates that the distribution is more (less) peaked about the mean than a normal distribution of the same mean and variance.

3.4 THE CHARACTERISTIC FUNCTION

The various moments introduced in the preceding section serve to characterize the distribution under study. For instance, the first algebraic moment defines a mean or "center of gravity" for the distribution, and the second central moment measures the spread of the distribution about this mean.

In addition to the usefulness of the individual moments there is considerable theoretical interest attached to the *complete set of moments* μ_k (or equivalently of μ_k') since this set determines the probability density function completely.

We shall now introduce the *characteristic function* $\Phi(t)$ for the probability density function f(x). By definition $\Phi(t)$ is the Fourier transform of f(x),

$$\Phi(t) \equiv \int_{-\infty}^{+\infty} e^{itx} f(x)dx = E(e^{itx}) . \tag{3.22}$$

This function then contains the complete set of algebraic moments for the distribution f(x). By taking the Taylor expansion of e^{itx} about the origin and using the linearity property of the expectation operator, we have

$$\Phi(t) = E\left(1 + itx + 1/2!(itx)^2 + 1/3!(itx)^3 + \ldots\right)$$

$$= 1 + (it)E(x) + 1/2!(it)^2E(x^2) + 1/3!(it)^3E(x^3) + \ldots$$

or

$$\Phi(t) = \sum_{k=0}^{\infty} \frac{1}{k!}(it)^k \mu_k' . \tag{3.23}$$

Since the algebraic moments μ_k' appear as coefficients in a series expansion of $\Phi(t)$ they can be expressed as

$$\mu_k' = \left.\frac{\partial^k \Phi(t)}{\partial(it)^k}\right|_{t=0} , \qquad k = 1,2,\ldots \tag{3.24}$$

This relation can be used for the evaluation of the algebraic moments of any order when $\Phi(t)$ is known.

If instead we want the central moments we should use the characteristic function in the form

$$\Phi_\mu(t) \equiv \int_{-\infty}^{+\infty} e^{it(x-\mu)} f(x)dx = E\left(e^{it(x-\mu)}\right) \tag{3.25}$$

and perform an expansion in a power series about μ. Then, by analogy with eq. (3.23),

$$\Phi_\mu(t) = \sum_{k=0}^{\infty} \frac{1}{k!}(it)^k \mu_k \tag{3.26}$$

and the central moments are obtained as

$$\mu_k = \left.\frac{\partial^k \Phi_\mu(t)}{\partial(it)^k}\right|_{t=0} , \qquad k = 1,2,\ldots \tag{3.27}$$

The relation between the two forms of the characteristic function defining the two different sets of moments, is simply

$$\Phi_\mu(t) = e^{-\mu it}\Phi(t) . \tag{3.28}$$

Clearly, if one is interested in a general set of moments about an arbitrary point a, then one should use instead

$$\Phi_a(t) \equiv E\left(e^{it(x-a)}\right) = e^{-ait}\Phi(t) \tag{3.29}$$

and derive the desired moments from this function.

With the knowledge of $\Phi(t)$ the moments can be evaluated to any order; hence the properties of the parent distribution are derived. Furthermore, the probability density function itself can be found explicitly, since, from eq. (3.22)

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{-itx}\phi(t)dt\,. \tag{3.30}$$

Remark: In the literature one frequently finds reference to the *moment-generating function* M(t), which is used to evaluate moments of distributions. Using this function is in many respects equivalent to our use of the characteristic function. Thus, the moment-generating function about the mean $M_\mu(t)$, defined by

$$M_\mu(t) \equiv \int_{-\infty}^{+\infty} e^{t(x-\mu)}f(x)dx = E\left(e^{t(x-\mu)}\right), \tag{3.31}$$

provides the central moments through the formula

$$\mu_k = \left.\frac{\partial^k M_\mu(t)}{\partial t^k}\right|_{t=0}, \qquad k = 1,2,\ldots, \tag{3.32}$$

which is of course equivalent to eq.(3.27). However, from a theoretical viewpoint it may be advantageous to work with the characteristic function rather than with the moment-generating function. For instance, when $\Phi(t)$ is known, f(x) is explicitly given by eq.(3.30), whereas a similar formula can not be written with M(t).

3.5 DISTRIBUTIONS OF MORE THAN ONE RANDOM VARIABLE

3.5.1 The joint probability density function

Up to now we have assumed the probability density function to depend on a single random variable. The extension to several variables $x_1,x_2,\ldots,x_n$ consists in considering a *joint probability density function* $f(x_1,x_2,\ldots,x_n)$. We assume this function to be positive and single-valued at every point $x_1,x_2,\ldots,x_n$ in the n-dimensional space, and that it is properly normalized,

$$\int_\Omega f(x_1,x_2,\ldots,x_n)dx_1dx_2\ldots dx_n = 1$$

when the integration is over the entire domain of all x_i. For short we write

$$\int_\Omega f(\underline{x})d\underline{x} = 1\,. \tag{3.33}$$

3.5.2 Expectation values

The generalization of eq.(3.6) gives the expectation value of a general function $g(x_1,x_2,\ldots,x_n) = g(\underline{x})$ as

$$E\left(g(\underline{x})\right) \equiv \int_\Omega g(\underline{x})f(\underline{x})d\underline{x}\,. \tag{3.34}$$

We can generally define the variance of the function $g(\underline{x})$ by

$$V\left(g(\underline{x})\right) \equiv E\left(g(\underline{x}) - E\left(g(\underline{x})\right)\right)^2 = \int_\Omega \left(g(\underline{x}) - E\left(g(\underline{x})\right)\right)^2 f(\underline{x})d\underline{x}\,. \tag{3.35}$$

3.5.3 The covariance matrix; correlation coefficients

Specializing $g(\underline{x}) = x_i$ we get the expectation or *mean value* of x_i as

$$\mu_i \equiv E(x_i) = \int_\Omega x_i f(\underline{x})d\underline{x}\,. \tag{3.36}$$

We are next lead to handle the extension of the definition of the variance by eq.(3.9) to the case of several variables. The *covariance matrix* $V(\underline{x})$ of $\underline{x}$ is defined by its elements

$$V_{ij} \equiv E\left((x_i-\mu_i)(x_j-\mu_j)\right) = \int_\Omega (x_i-\mu_i)(x_j-\mu_j)f(\underline{x})d\underline{x}\,. \tag{3.37}$$

Here μ_i and μ_j are the expectations of the variables x_i and x_j, respectively, in accordance with eq.(3.36) above.

The covariance matrix is of great importance to physicists. Some of its properties may be stated as follows:

(i) $V(\underline{x})$ is symmetric.

(ii) A diagonal element V_{ii} is called the *variance* σ_i^2 of the variable x_i. σ_i^2 is a non-negative quantity,

$$\sigma_i^2 \equiv V_{ii} = E(x_i-\mu_i)^2 = \int_\Omega (x_i-\mu_i)^2 f(\underline{x})d\underline{x}\,,$$

and, analogously to eqs.(3.10) - (3.12) for a single random variable, we have

$$\sigma_i^2 = V_{ii} = E(x_i^2) - \big(E(x_i)\big)^2. \tag{3.38}$$

(iii) An off-diagonal element V_{ij}, where $i \neq j$, is called the *covariance* of x_i and x_j, and is denoted by $\mathrm{cov}(x_i,x_j)$,

$$\mathrm{cov}(x_i,x_j) \equiv V_{ij} = E(x_ix_j) - E(x_i)E(x_j). \tag{3.39}$$

The covariance may be a positive or negative quantity.

A frequently used measure on the correlation between two variables x_i,x_j is the *correlation coefficient* $\rho(x_i,x_j)$ defined by

$$\rho(x_i,x_j) \equiv \frac{V_{ij}}{(V_{ii}V_{jj})^{\frac{1}{2}}} = \frac{\mathrm{cov}(x_i,x_j)}{\sigma_i\sigma_j}, \tag{3.40}$$

which satisfies

$$-1 \leq \rho(x_i,x_j) \leq +1. \tag{3.41}$$

Two random variables having $\rho(x_i,x_j) = +1(-1)$ are said to be completely positively (negatively) correlated. When $\rho(x_i,x_j) = 0$ the variables are *uncorrelated.*

To prove that the correlation coefficient is a number between -1 and +1 we make use of a theorem on the variance of a linear corbination of variables, which is proved for a more general case in Sect.3.6. The theorem (eq.(3.55)) implies that for the sum x_1+ax_2 the variance is given by

$$V(x_1+ax_2) = V(x_1) + a^2V(x_2) + 2a\ \mathrm{cov}(x_1,x_2).$$

By definition the variance is a non-negative quantity, hence

$$V(x_1) + a^2V(x_2) + 2a\ \mathrm{cov}(x_1,x_2) \geq 0.$$

Dividing by $V(x_1)$, putting $a^2V(x_2)/V(x_1) = \alpha^2$ and using the definition of the correlation coefficient ρ, the condition can be written as

$$1 + \alpha^2 + 2\rho\alpha \geq 0.$$

Since this condition must be satisfied for any value of α, there follows a restriction on ρ, $\rho^2 \leq 1$, which leads to the inequality (3.41).

3.5.4 Independent variables

The random variables $x_1,x_2,\ldots,x_n$ are said to be *mutually independent* if their joint probability density function is completely factorizable as

$$f(x_1,x_2,\ldots,x_n) = f_1(x_1)f_2(x_2)\ldots f_n(x_n), \qquad \text{(independence)}. \tag{3.42}$$

This is just a new formulation of the definition of independence by eq.(2.13) applied to the case of continuous variables.

Independent variables have the property that their covariance, and hence their correlation, vanish. To see this, consider the expectation of the product of two mutually independent variables x_i,x_j:

$$\begin{aligned} E(x_ix_j) &= \int_\Omega x_ix_jf(x_i,x_j)dx_idx_j = \int_\Omega x_ix_jf_i(x_i)f_j(x_j)dx_idx_j \\ &= \int x_if_i(x_i)dx_i \cdot \int x_jf_j(x_j)dx_j = E(x_i)\cdot E(x_j) \end{aligned} \tag{3.43}$$

provided that f_i and f_j are both properly normalized. From the definition of the covariance of two variables and of the correlation coefficient, eqs.(3.39) and (3.40), respectively, it follows that

$$\mathrm{cov}(x_i,x_j) = 0, \qquad \rho(x_i,x_j) = 0, \qquad \text{(independent variables)}. \tag{3.44}$$

It may be noted that although independent variables necessarily are uncorrelated, the opposite statement is not generally true.

The result derived above by eq.(3.43) means that the expectation of the product of two mutually independent variables is equal to the product of the expectations of the individual variables. This is in fact just a special case of a more general statement about the expectation value of a function which is factorizable in the two independent variables. To see this, let us write

$$g(x_i,x_j) = u(x_i)v(x_j) \tag{3.45}$$

where x_i and x_j are assumed to be mutually independent. Using the definition of independence one readily shows that

$$E\{g(x_i,x_j)\} = \int u(x_i)f_i(x_i)dx_i \cdot \int v(x_j)f_j(x_j)dx_j$$

$$= E\{u(x_i)\}\cdot E\{v(x_j)\}, \qquad \text{(independence)}. \qquad (3.46)$$

again assuming f_i and f_j to be separately normalized. This general property frequently simplifies the calculation of various expectation values.

Finally, one may note the following fact about two independent variables x_i and x_j: If $u=u(x_i)$ and $v=v(x_j)$, then u and v are also independent. The proof of this statement is suggested as an exercise for the reader in Sect. 3.7, (Exercise 3.7).

Exercise 3.3: Given $f(x_1,x_2) = \frac{4}{\pi}x_1^2 + \frac{8}{\pi}x_2$ and Ω defined by $x_1^2 + x_2^2 \leq 1$ (dependent variables), show that $\rho(x_1,x_2) = 0$ (no correlation).

3.5.5 Marginal and conditional distributions

A projection of the probability density function $f(\underline{x})$ onto a subspace is called a marginal density function or a *marginal distribution* for $f(\underline{x})$. Consider the n-dimensional p.d.f. $f(x_1,x_2,\ldots,x_n)$. Integrating over all variables except one, say x_1, gives the marginal distribution in this variable; we write

$$h_1(x_1) \equiv \int_{x_2(\min)}^{x_2(\max)} \cdots \int_{x_n(\min)}^{x_n(\max)} f(x_1,x_2,\ldots,x_n)dx_2\ldots dx_n , \qquad (3.47)$$

and similarly for the other n-1 variables. It will be seen that eq.(3.47) represents an application of the definition of marginal probability, eq.(2.22).

In the case of *mutually independent* variables for which the p.d.f. factorizes according to eq.(3.42), the marginal distribution becomes

$$h_1(x_1) = f_1(x_1) \int_{x_2(\min)}^{x_2(\max)} f_2(x_2)dx_2 \ldots \int_{x_n(\min)}^{x_n(\max)} f_n(x_n)dx_n$$

with corresponding expressions for $h_2(x_2)$ etc. Because of the overall normalization condition for $f(\underline{x})$, one must have

$$h_1(x_1) = f_1(x_1), \quad h_2(x_2) = f(x_2), \quad \ldots$$

Thus the concept of independence as stated in Sect.3.5.4 can be formulated by saying that mutually independent random variables have a joint probability density function which is factorizable into its marginal density functions.

We next introduce the concept of a conditional density function or a *conditional distribution* for $f(\underline{x})$. Consider again the n-dimensional p.d.f. $f(x_1,x_2,\ldots,x_n)$. The conditional probability density in all variables except x_1 is then defined as the ratio between $f(x_1,x_2,\ldots,x_n)$ and the marginal density function for x_1, thus

$$f(x_2,\ldots,x_n|x_1) \equiv \frac{f(x_1,x_2,\ldots,x_n)}{h_1(x_1)} . \qquad (3.48)$$

Here it is understood that x_1 is kept fixed, and $f(x_2,\ldots,x_n|x_1)$ is then a function of the remaining variables $x_2,\ldots,x_n$. Similar definitions apply for the other variables. It will be seen that eq.(3.48) corresponds to the previous definition of conditional probability by eq.(2.10).

With the conditional p.d.f. of eq.(3.48) one may define *conditional expectation values* of functions and variables in a manner which is quite analogous to what has appeared before. For example, the conditional expectation of $u(x_2,\ldots,x_n|x_1)$, given x_1, is

$$E\{u(x_2,\ldots,x_n|x_1)\} =$$

$$= \int_{x_2(\min)}^{x_2(\max)} \cdots \int_{x_n(\min)}^{x_n(\max)} u(x_2,\ldots,x_n|x_1)f(x_2,\ldots,x_n|x_1)dx_2\ldots dx_n . \qquad (3.49)$$

3.5.6 Example: Scatterplots of kinematic variables

In particle physics probability density functions of two variables are encountered in studies of scatterplots, or two-dimensional displays of kinematic variables. The most common of these are presumably the Dalitz plot and the Chew-Low plot.

For definiteness let us think of a reaction

$$a + b \to 1 + 2 + 3$$

at a total centre-of-mass energy $\sqrt{s}$, and let s_{ij}, t_{ai} denote, respectively, the squared effective-mass of particles i,j and the squared four-momentum transfer between particles a,i. Then the kinematically allowed region in a Chew-Low display of t_{a3} *versus* s_{12} is a closed area bounded by a straight line and a hyperbola, see Fig. 3.2(a). The marginal distribution in s_{12} is the projection on the s_{12} axis, giving the one-dimensional distribution in the squared effective-mass of particles 1 and 2. The other marginal distribution for the Chew-Low plot gives the one-dimensional distribution in the squared four-momentum transfer between the initial particle a and the final particle 3. For the Dalitz plot of say, s_{12} *versus* s_{13}, both marginal distributions are squared effective-mass spectra, see Fig. 3.2(b).

Lorentz-invariant phase space predicts the density in the Chew-Low plot to be given by the formula

$$\frac{d^2R_3}{ds_{12}dt_{a3}} = \frac{\pi^2}{4s_{12}} \lambda^{-\frac{1}{2}}(s,m_a^2,m_b^2)\lambda^{\frac{1}{2}}(s_{12},m_1^2,m_2^2)$$

where the kinematic function λ is defined by

$$\lambda(x^2,y^2,z^2) \equiv x^4 + y^4 + z^4 - 2x^2y^2 - 2x^2z^2 - 2y^2z^2$$

$$= \left(x^2 - (y-z)^2\right)\left(x^2 - (y+z)^2\right).$$

(The quantity $\frac{1}{2x}\lambda^{\frac{1}{2}}(x^2,y^2,z^2)$ corresponds to the magnitude of the momenta when two particles of masses y and z share the centre-of-mass energy x.) Thus according to the phase space prescription the conditional density, given s_{12}, is a constant,

$$f(t_{a3}|s_{12}) \propto \frac{d^2R_3(t_{a3}|s_{12})}{ds_{12}dt_{a3}} \qquad \text{(independent of } t_{a3}\text{)}.$$

In Fig. 3.2(a) this implies that the density is uniform within the kinematic boundary along lines of constant s_{12}. The marginal distribution in s_{12} is obtained by integrating over t_{a3}, thus

$$h_1(s_{12}) = \int_{t_{a3}(\min)}^{t_{a3}(\max)} f(t_{a3}|s_{12})dt_{a3} = f(t_{a3}|s_{12})\left(t_{a3}(\max) - t_{a3}(\min)\right).$$

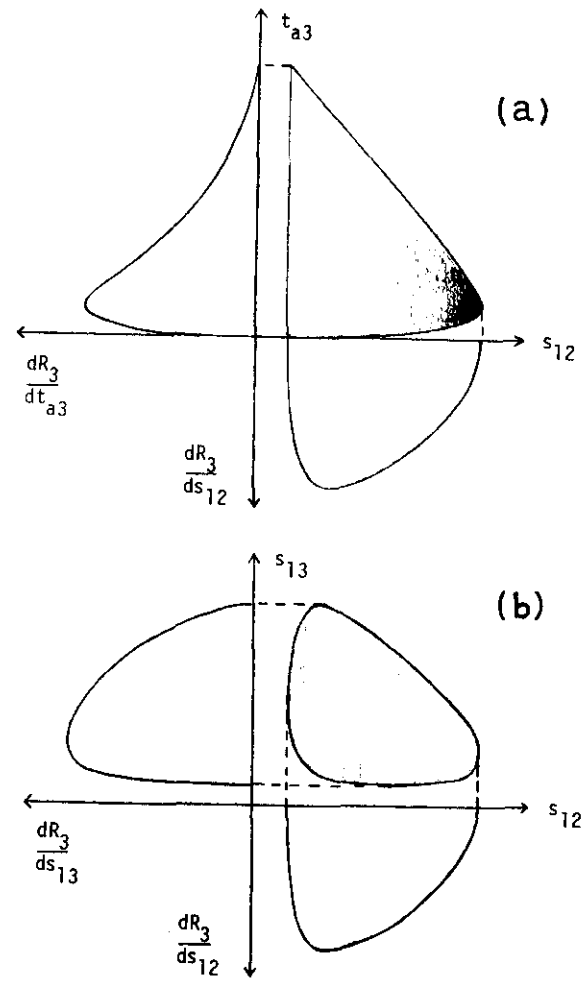

Fig. 3.2. Illustration of joint probability and marginal distributions. The shaded areas correspond to the kinematically allowed physical regions in two dimensions (variables) for the reaction $a + b \to 1 + 2 + 3$; (a) t_{a3} *versus* s_{12} (Chew-Low plot), (b) s_{13} *versus* s_{12} (Dalitz plot). In both diagrams the degree of shading indicates the density expected according to Lorentz-invariant phase space, and the one-dimensional projections on the two axes give the spectral shapes for the variables involved.

Since the boundary corresponds to configurations where the final state particles are collinear, the square bracket can be evaluated quite easily to give

$$\left(t_{a_3}(\max) - t_{a_3}(\min)\right) = \frac{1}{\sqrt{s}}\lambda^{\frac{1}{2}}(s,m_a^2,m_b^2)\cdot\frac{1}{\sqrt{s}}\lambda^{\frac{1}{2}}(s,s_{12},m_3^2).$$

Hence the marginal distribution in s_{12} according to Lorentz-invariant phase space is obtained explicitly as

$$h_1(s_{12}) \propto \frac{\pi^2}{4ss_{12}}\lambda^{\frac{1}{2}}(s,s_{12},m_3^2)\lambda^{\frac{1}{2}}(s_{12},m_1^2,m_2^2);$$

(the proportionality sign is a warning that h_1 is not properly normalized due to the conventional definition adopted for R_3).

If instead we were interested in the marginal distribution in the other variable t_{a_3}, we would have to integrate the conditional density $f(s_{12}|t_{a_3})$ for fixed t_{a_3}, thus

$$h_2(t_{a_3}) \propto \int_{s_{12}(\min)}^{s_{12}(\max)} f(s_{12}|t_{a_3})ds_{12}.$$

It is seen that this leads to an elliptical integral.

Exercise 3.4: Show that the marginal distribution $h_1(s_{12})$ can be normalized using the relation

$$\int_{s_{12}(\min)}^{s_{12}(\max)} h_1(s_{12})ds_{12} = R_3(s;m_1,m_2,m_3)$$
$$= \int_{(m_1+m_2)^2}^{(\sqrt{s}-m_3)^2} R_2(s;s_{12},m_3)R_2(s_{12};m_1,m_2)ds_{12}$$

where the two-particle phase space factor R_2 is related to the kinematic function λ by

$$R_2(x^2,y^2,z^2) = \frac{\pi^2}{2x^2}\lambda^{\frac{1}{2}}(x^2,y^2,z^2).$$

Exercise 3.5: For the three-particle final state Lorentz-invariant phase space predicts the density within the kinematic boundary of the Dalitz plot to be proportional to

$$\frac{d^2R_3}{ds_{12}ds_{13}} = \frac{\pi^2}{4s}$$

i.e. constant. Show that the marginal distribution for s_{12} is given by the expression for $h_1(s_{12})$ in the text.

3.5.7 The joint characteristic function

In analogy with the definition of the characteristic function in the case of a single random variable, we now introduce the *joint characteristic function* $\Phi(t_1,t_2,\ldots,t_n)$ for the joint probability density function $f(x_1,x_2,\ldots,x_n)$, by the definition

$$\Phi(t_1,t_2,\ldots,t_n) \equiv E\left(e^{it_1x_1+it_2x_2+\ldots+it_nx_n}\right)$$
$$= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} e^{it_1x_1+it_2x_2+\ldots+it_nx_n}f(x_1,x_2,\ldots,x_n)dx_1dx_2\ldots dx_n \quad (3.50)$$

Let us examine the case with just two variables, when the characteristic function is

$$\Phi(t_1,t_2) = E\left(e^{it_1x_1+it_2x_2}\right).$$

If x_1 and x_2 are *independent* variables, *i.e.* if $f(x_1,x_2) = f_1(x_1)\cdot f_2(x_2)$, we can write

$$\Phi(t_1,t_2) = E\left(e^{it_1x_1}\cdot e^{it_2x_2}\right) = E\left(e^{it_1x_1}\right)\cdot E\left(e^{it_2x_2}\right)$$
$$= \Phi(t_1)\cdot\Phi(t_2),$$

where $\Phi(t_1)$ is the characteristic function for $f_1(x_1)$, the marginal distribution in x_1, and similarly for $\Phi(t_2)$. Thus, <u>for independent variables the joint characteristic function is factorizable</u> *). In general, with n independent variables,

$$\Phi(t_1,t_2,\ldots,t_n) = \Phi(t_1)\Phi(t_2)\ldots\Phi(t_n), \qquad \text{(independence)}. \quad (3.51)$$

The joint characteristic function may be used to find the general moments of the different variables. The technique is the same as shown before in the case of a single variable. For simplicity, let us again specialize to just two variables, x_1 and x_2. When

*) In fact the inverse statement also holds: If the joint characteristic function can be factorized, the variables are independent. Thus eq.(3.51) represents a necessary and sufficient condition for independence.

$$\Phi(t_1,t_2) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{it_1x_1+it_2x_2} f(x_1,x_2)dx_1dx_2$$

a derivation with respect to (it_1) gives

$$\frac{\partial\Phi}{\partial(it_1)} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} x_1 e^{it_1x_1+it_2x_2} f(x_1,x_2)dx_1dx_2 .$$

Putting $t_1=t_2=0$, the right-hand side is nothing but the expectation of x_1,

$$\left.\frac{\partial\Phi}{\partial(it_1)}\right|_{\underline{t}=0} = E(x_1) .$$

Similarly, after two derivations with respect to (it_1), for $\underline{t}=0$,

$$\left.\frac{\partial^2\Phi}{\partial(it_1)^2}\right|_{\underline{t}=0} = E(x_1^2) ,$$

and so on. Corresponding expressions for the other variable result from derivations with respect to it_2. Also

$$\left.\frac{\partial^2\Phi}{\partial(it_1)\partial(it_2)}\right|_{\underline{t}=0} = E(x_1x_2) .$$

Thus, with two variables,

$$E(x_1^r x_2^s) = \left.\frac{\partial^{r+s}\Phi}{\partial(it_1)^r\partial(it_2)^s}\right|_{\underline{t}=0} . \qquad (3.52)$$

The extension to more variables is clearly straightforward.

3.6 LINEAR FUNCTIONS OF RANDOM VARIABLES

We shall go back to our definitions of the expectation and variance of a general function and investigate the consequences if $g(x_1,x_2,\ldots,x_n)$ is a *linear* function of the random variables. We put

$$g(x_1,x_2,\ldots,x_n) = \sum_{i=1}^{n} a_i x_i . \qquad (3.53)$$

The expectation of this sum is easily found by using the linearity property of E:

$$E\Big(\sum_{i=1}^{n} a_i x_i\Big) = \sum_{i=1}^{n} E(a_i x_i) = \sum_{i=1}^{n} a_i E(x_i) = \sum_{i=1}^{n} a_i \mu_i . \qquad (3.54)$$

Thus the expectation of a linear combination of variables x_i is the same linear combination of the individual mean values.

The variance of the linear function is slightly more troublesome to evaluate. We have

$$\begin{aligned} V\Big(\sum_{i=1}^{n} a_i x_i\Big) &= E\Big\{\sum_{i=1}^{n} a_i x_i - E\Big(\sum_{i=1}^{n} a_i x_i\Big)\Big\}^2 \\ &= E\big(\textstyle\sum a_i x_i - \sum a_i \mu_i\big)^2 = E\big(\sum a_i (x_i-\mu_i)\big)^2 \\ &= E\Big(\sum a_i^2 (x_i-\mu_i)^2 + \sum_{i\neq j} a_i a_j (x_i-\mu_i)(x_j-\mu_j)\Big) . \end{aligned}$$

This can further be written as follows:

$$\begin{aligned} V\Big(\sum_{i=1}^{n} a_i x_i\Big) &= \textstyle\sum a_i^2 E(x_i-\mu_i)^2 + \sum_{i\neq j} a_i a_j E\big((x_i-\mu_i)(x_j-\mu_j)\big) \\ &= \textstyle\sum a_i^2 V(x_i) + \sum_{i\neq j} a_i a_j \mathrm{cov}(x_i,x_j) , \end{aligned}$$

or, finally

$$V\Big(\sum_{i=1}^{n} a_i x_i\Big) = \sum_{i=1}^{n} a_i^2 V_{ii} + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} a_i a_j V_{ij} . \qquad (3.55)$$

Thus the variance of the linear combination consists of two parts: the first part is just the sum of the variances of the individual variables, weighted by the square of their coefficients in the linear combination, the second part is a sum of all the covariance terms that can be made up for the variables.

We note that for the particular case of uncorrelated variables the variance of the linear function reduces to the simple relation

$$V\Big(\sum_{i=1}^{n} a_i x_i\Big) = \sum_{i=1}^{n} a_i^2 V(x_i), \qquad \text{(uncorrelated variables)}. \qquad (3.56)$$

3.6.1 Example: Arithmetic mean of independent variables with the same mean and variance

Let $x_1, x_2, \ldots, x_n$ be n mutually independent random variables having the same mean value $\mu_i = \mu$ and the same variance $\sigma_i^2 = \sigma^2$. We then take as a particular linear combination the *arithmetic mean* $\bar{x}$ or *average* of the x_i,

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i .$$

This is a special case of eq.(3.53) in which all coefficients a_i are equal, $a_i = \frac{1}{n}$, and hence we get from eqs.(3.54) and (3.56) the expectation and variance of $\bar{x}$:

$$E(\bar{x}) = \sum_{i=1}^{n} a_i \mu_i = \sum_{i=1}^{n} \frac{1}{n} \mu = \mu , \tag{3.57}$$

$$V(\bar{x}) = \sum_{i=1}^{n} a_i^2 \sigma_i^2 = \sum_{i=1}^{n} \left(\frac{1}{n}\right)^2 \sigma^2 = \frac{\sigma^2}{n} . \tag{3.58}$$

3.7 CHANGE OF VARIABLES

It often happens that the probability density function is known for a certain set of variables and that one wants to find what the distribution will be like when a transformation is made to a new set of variables. For instance, given a spectrum of particle momenta one may want to have the corresponding energy spectrum.

Suppose first that x is a continuous random variable with p.d.f. f(x) and that we know a functional dependence

$$y = y(x) . \tag{3.59}$$

We ask: What is the p.d.f. g(y) for the new variable y?

For a one-to-one correspondence between the old variable x and the new variable y, an interval [x,x+dx] is mapped onto [y,y+dy] and we require

$$f(x)dx = g(y)dy .$$

To ensure a non-negative dependence, we take

$$g(y) = f(x) \cdot \left|\frac{dx}{dy}\right| , \tag{3.60}$$

and this is the answer to our question.

If the transformation (3.59) is not one-to-one, and several segments [x,x+dx] map onto [y,y+dy], one must sum over all segments,

$$g(y) = \sum f(x) \left|\frac{dx}{dy}\right| . \tag{3.61}$$

Eq.(3.60) can easily be extended to cover a transformation from a set of variables $x_1, x_2, \ldots, x_n$ with p.d.f. $f(x_1, x_2, \ldots, x_n)$ to a second set of variables $y_1, y_2, \ldots, y_n$. The p.d.f. for the new set of variables is

$$g(y_1, y_2, \ldots, y_n) = f(x_1, x_2, \ldots, x_n) \left|\frac{\partial(x_1, x_2, \ldots, x_n)}{\partial(y_1, y_2, \ldots, y_n)}\right| \tag{3.62}$$

or, for short, in obvious vector notation,

$$g(\underline{y}) = f(\underline{x}) \cdot |J| , \tag{3.63}$$

where J is the Jacobian determinant of the transformation, given by

$$|J| \equiv \left|\frac{\partial(x_1, x_2, \ldots, x_n)}{\partial(y_1, y_2, \ldots, y_n)}\right| \equiv \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \cdots & \frac{\partial x_2}{\partial y_n} \\ \vdots & & & \\ \frac{\partial x_n}{\partial y_1} & \frac{\partial x_n}{\partial y_2} & \cdots & \frac{\partial x_n}{\partial y_n} \end{vmatrix} . \tag{3.64}$$

Exercise 3.6: If $f(x) = (2\pi)^{-\frac{1}{2}} \exp(-\frac{1}{2}x^2)$ (the standard normal p.d.f.), show that the variable $y=x^2$ has a p.d.f. given by $g(y) = (2\pi)^{-\frac{1}{2}} y^{-\frac{1}{2}} \exp(-\frac{1}{2}y)$ (the chi-square distribution with one degree of freedom).

Exercise 3.7: Let x_1 and x_2 be two independent variables with p.d.f.'s $f_1(x_1)$ and $f_2(x_2)$, respectively, and let y_1 be a function of x_1 alone, y_2 a function of x_2 alone. Show that y_1 and y_2 are also independent.

3.7.1 Example: Dalitz plot variables

Consider a three-particle final state for which Lorentz-invariant phase space predicts in terms of the squared invariant masses M_{12}^2, M_{13}^2,

$$f(M_{12}^2, M_{13}^2) = \frac{d^2R_3}{dM_{12}^2 dM_{13}^2} = \frac{\pi^2}{4s}$$

that is, a constant density, (compare Exercise 3.5). For new variables choose the linear effective masses M_{12}, M_{13}. Then the Jacobian of the transformation is

$$J = \left|\frac{\partial(M_{12}^2, M_{13}^2)}{\partial(M_{12}, M_{13})}\right| = \begin{vmatrix} \frac{\partial M_{12}^2}{\partial M_{12}} & \frac{\partial M_{12}^2}{\partial M_{13}} \\ \frac{\partial M_{13}^2}{\partial M_{12}} & \frac{\partial M_{13}^2}{\partial M_{13}} \end{vmatrix} = \begin{vmatrix} 2M_{12} & 0 \\ 0 & 2M_{13} \end{vmatrix} = 4M_{12}M_{13}.$$

The density in the new variables is therefore

$$g(M_{12}, M_{13}) = \frac{\pi^2}{4s} \cdot 4M_{12}M_{13} = \frac{\pi^2}{s} M_{12}M_{13},$$

which is not a constant. Thus the nice feature of constant density is lost in the change from squared to linear effective masses.

Exercise 3.8: In the example above, prove that a transformation to the energy variables E_3, E_2 yields a constant probability density.

3.8 PROPAGATION OF ERRORS

We have in the preceding paragraphs studied various functions of random variables. Clearly any such function or new variable which is defined by a functional dependence of random variables is itself a random variable. We shall now study properties of variables of this type.

3.8.1 A single function

Let $x_1, x_2, \ldots, x_n$ be the original random variables and put

$$y = y(x_1, x_2, \ldots, x_n) = y(\underline{x}). \tag{3.65}$$

Let us further assume that the covariance matrix $V(\underline{x})$ of $\underline{x}$ is known. We need not specify whether the x_i's are independent or not, so the form of $V(\underline{x})$

(diagonal or not) is inessential.

To estimate the variance of y we perform a Taylor expansion of y about the mean value $\underline{\mu} = \{\mu_1, \mu_2, \ldots, \mu_n\}$ of $\underline{x}$. Writing out the terms of order zero and one we have

$$y(\underline{x}) = y(\underline{\mu}) + \sum_{i=1}^{n} (x_i - \mu_i) \left.\frac{\partial y}{\partial x_i}\right|_{\underline{x}=\underline{\mu}} + \text{terms of higher order}. \tag{3.66}$$

Taking the expectation value of this expression each first order term will vanish, so that

$$E\{y(\underline{x})\} = y(\underline{\mu}) + \text{terms of higher order}. \tag{3.67}$$

Under the assumption that the quantities $(x_i - \mu_i)$ are small, the remaining terms can be dropped to give the approximate result

$$E\{y(\underline{x})\} \simeq y(\underline{\mu}). \tag{3.68}$$

Introducing this in the formulae for the variance of $y(\underline{x})$, eq.(3.35), we get

$$V\{y(\underline{x})\} = E\left\{y(\underline{x}) - E\{y(\underline{x})\}\right\}^2 \simeq E\{y(\underline{x}) - y(\underline{\mu})\}^2. \tag{3.69}$$

Now we can find an approximate value of the difference $y(\underline{x}) - y(\underline{\mu})$ from eq. (3.66) by dropping all terms of order higher than one,

$$y(\underline{x}) - y(\underline{\mu}) \simeq \sum_{i=1}^{n} (x_i - \mu_i) \left.\frac{\partial y}{\partial x_i}\right|_{\underline{x}=\underline{\mu}}. \tag{3.70}$$

Substituting this back in eq.(3.69) we obtain the following approximation for the variance of $y(\underline{x})$,

$$V\{y(\underline{x})\} \simeq \sum_{i=1}^{n} \sum_{j=1}^{n} \left.\frac{\partial y}{\partial x_i}\right|_{\underline{x}=\underline{\mu}} \left.\frac{\partial y}{\partial x_j}\right|_{\underline{x}=\underline{\mu}} E\{(x_i - \mu_i)(x_j - \mu_j)\}. \tag{3.71}$$

The expectation values here are nothing but the elements of the covariance matrix of $\underline{x}$. So we get the final result

$$V\{y(\underline{x})\} \simeq \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial y}{\partial x_i} \frac{\partial y}{\partial x_j} V_{ij}(\underline{x}), \tag{3.72}$$

where the derivatives are evaluated at $\underline{x}=\underline{\mu}$.

The formula (3.72) is known as the *law of propagation of errors* and is of great importance to physicists. In the general case it is to be regarded as only approximately valid, in view of the assumptions made in deriving it (dropping terms of higher order). We have found an expression for the variance of y valid when $\underline{x}$ is in the neighbourhood of $\underline{\mu}$. Note, however, that for the particular case when y has a linear functional dependence on $\underline{x}$ all derivatives of second and higher order vanish identically; eq.(3.72) is then exact for all $\underline{x}$.

For n mutually independent variables all covariance terms are zero and eq.(3.72) reduces to

$$V\big(y(\underline{x})\big) \simeq \sum_{i=1}^{n} \left(\frac{\partial y}{\partial x_i}\right)^2 V_{ii}(\underline{x}), \qquad \text{(independent variables)}. \tag{3.73}$$

This relation is also exact only for linear functions of $\underline{x}$, and otherwise approximately correct to the extent that higher-order terms can be neglected.

Exercise 3.9: If $y = \frac{x_1}{x_2}$ where x_1 and x_2 are two independent random variables having variances $V(x_1)$ and $V(x_2)$, respectively, show that $V(y) = x_2^{-4}\big(x_2^2 V(x_1) + x_1^2 V(x_2)\big)$, and $V(y)/y^2 = V(x_1)/x_1^2 + V(x_2)/x_2^2$.

3.8.2 Example: Variance of arithmetic mean

Let y be the average of a set of n independent variables $x_1, x_2, \ldots, x_n$, all with the same variance σ^2,

$$y = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i .$$

Then $\frac{\partial y}{\partial x_i} = \frac{1}{n}$ for all i, higher derivatives of y vanish and eq.(3.73) holds strictly, giving

$$V(\bar{x}) = \sum_{i=1}^{n} \left(\frac{\partial y}{\partial x_i}\right)^2 \sigma^2 = \frac{\sigma^2}{n}.$$

This is the same result as eq.(3.58) derived in the example of Sect.3.6.1.

3.8.3 Several functions; matrix notation

We need a generalization of the law of propagation of errors which will cover the important case when there is a *set* of functions $y_1, y_2, \ldots, y_m$

which all depend on the n random variables $x_1, x_2, \ldots, x_n$, thus

$$y_k = y_k(x_1, x_2, \ldots, x_n) = y_k(\underline{x}), \qquad k = 1,2,\ldots,m. \tag{3.74}$$

A Taylor expansion about $\underline{x} = \underline{\mu}$ leads to

$$y_k(\underline{x}) = y_k(\underline{\mu}) + \sum_i (x_i - \mu_i) \left.\frac{\partial y_k}{\partial x_i}\right|_{\underline{x}=\underline{\mu}} + \ldots, \qquad k = 1,2,\ldots,m,$$

by analogy with eq.(3.66). Taking the expectation value each of the first order terms drops out, and we have

$$E\big(y_k(\underline{x})\big) \simeq y_k(\underline{\mu}), \qquad k = 1,2,\ldots,m,$$

which holds exactly in the close neighbourhood of $\underline{\mu}$. Generalizing eqs.(3.69) - (3.71) we find for the covariance between, say, y_k and y_ℓ,

$$V_{k\ell}(\underline{y}) \simeq E\left(\Big(y_k(\underline{x}) - E\big(y_k(\underline{x})\big)\Big)\Big(y_\ell(\underline{x}) - E\big(y_\ell(\underline{x})\big)\Big)\right)$$

or

$$V_{k\ell}(\underline{y}) \simeq \sum_{i=1}^{n}\sum_{j=1}^{n} \left.\frac{\partial y_k}{\partial x_i}\right|_{\underline{x}=\underline{\mu}} \left.\frac{\partial y_\ell}{\partial x_j}\right|_{\underline{x}=\underline{\mu}} E\big((x_i-\mu_i)(x_j-\mu_j)\big).$$

In analogy with eq.(3.72) we write

$$V_{k\ell}(\underline{y}) \simeq \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial y_k}{\partial x_i}\frac{\partial y_\ell}{\partial x_j} V_{ij}(\underline{x}), \tag{3.75}$$

which is the general formulation of the law of error propagation. It is understood that the derivatives should be evaluated for $\underline{x}=\underline{\mu}$, and the formula is only valid to the extent that terms of second order and higher can be neglected.

The covariance terms $V_{k\ell}(\underline{y})$ define the *covariance matrix* $V(\underline{y})$ for the dependent variables $\underline{y}$. Eq.(3.75) in fact provides the basis for error calculation in physics. The errors on the variables $\underline{y}$ are given by the square root of the diagonal terms of $V(\underline{y})$. In general a diagonal term $V_{kk}(\underline{y})$ will contain covariance terms $V_{ij}(\underline{x})$ of the original variables, because, granted a sufficient linearity,

$$V_{kk}(\underline{y}) \simeq \sum_{i=1}^{n}\sum_{j=1}^{n} \left(\frac{\partial y_k}{\partial x_i}\right)\left(\frac{\partial y_k}{\partial x_j}\right) V_{ij}(\underline{x}). \tag{3.76}$$

If, however, the original variables x_i are uncorrelated, the sum reduces to

$$V_{kk}(\underline{y}) = \sum_{i=1}^{n} \left(\frac{\partial y_k}{\partial x_i}\right)^2 V_{ii}(\underline{x}) . \tag{3.77}$$

Hence, in terms of the errors σ,

$$\sigma_k(\underline{y}) = \left(\sum_{i=1}^{n} \left(\frac{\partial y_k}{\partial x_i}\right)^2 \sigma_i^2(\underline{x}) \right)^{\frac{1}{2}}, \qquad \text{(uncorrelated } \underline{x}\text{)}. \tag{3.78}$$

This expression is commonly referred to as the law of error propagation, but as we have emphasized, it represents only a special case.

Note that even if the original variables $\underline{x}$ are uncorrelated the covariance matrix for the new variables $\underline{y}$ may well have off-diagonal elements different from zero.

Finally, let us for convenience summarize in matrix notation. With $\underline{x}$ and $\underline{y}$ as column vectors having, respectively, n and m elements we write

$$\underline{y} = \underline{c} + S\underline{x} + \text{higher order terms} . \tag{3.79}$$

Here $\underline{c}$ is an m-component (column) vector of constants and S an m by n matrix describing the linear part of the transformation $\underline{x} \to \underline{y}$. Then, to first order,

$$\frac{\partial y_k}{\partial x_i} = S_{ki} , \qquad \frac{\partial y_\ell}{\partial x_j} = S_{\ell j} ,$$

and the $k\ell$-th element of the covariance matrix $V(\underline{y})$ can be expressed as (eq. (3.75)),

$$V_{k\ell}(\underline{y}) \simeq \sum_{i=1}^{n} \sum_{j=1}^{n} S_{ki} V_{ij}(\underline{x}) S_{\ell j} .$$

Thus, the law of propagation of errors takes the form

$$V(\underline{y}) = SV(\underline{x})S^T, \tag{3.80}$$

where the superscript denotes the transposed matrix.

Exercise 3.10: If the variables $\frac{1}{p}$, λ, ϕ have been measured with errors $\Delta(\frac{1}{p})$, $\Delta\lambda$, $\Delta\phi$, respectively, and with no correlations, what are the errors associated with the derived quantities $p_x = p\cos\lambda\cos\phi$, $p_y = p\cos\lambda\sin\phi$, $p_z = p\sin\lambda$? What are the correlations in the new variables?

3.9 DISCRETE PROBABILITY DISTRIBUTIONS

3.9.1 Modification of formulae

A random variable which can take on only discrete values we denote by r. Its probability distribution is given by the set of probabilities p_r, normalized such that

$$\sum_r p_r = 1 \tag{3.81}$$

where the summation goes over all r.

For a discrete probability distribution the definitions of expectation values are analogous to the definitions introduced earlier in the case of continuous variables. The expectation and variance of r, for instance, are

$$E(r) = \sum_r r p_r , \tag{3.82}$$

$$V(r) = \sum_r \big(r - E(r)\big)^2 p_r = E(r^2) - \big(E(r)\big)^2 . \tag{3.83}$$

Exercise 3.11: If $p_r = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r}$ (the binomial distribution) and the variable r may take any integer value $0,1,\ldots,n$, show that $E(r) = np$, $V(r) = np(1-p)$.

3.9.2 The probability generating function

Dealing with discrete probabilities one can make use of the *probability generating function*, defined by

$$G(z) \equiv E(z^r) = \sum_r z^r p_r . \tag{3.84}$$

Its usefulness comes from the properties of the derivatives evaluated at the point z=1. Since

$$G'(z) = \sum_r r z^{r-1} p_r ,$$

$$G''(z) = \sum_r r(r-1) z^{r-2} p_r ,$$

etc., we deduce that

$$G'(1) = \sum_r r p_r = E(r),$$

$$G''(1) = \sum_r r(r-1)p_r = \sum_r r^2 p_r - \sum_r r p_r = E(r^2) - E(r).$$

For the most common expectation values one has therefore the convenient expressions

$$E(r) = G'(1), \tag{3.85}$$

$$V(r) = G''(1) + G'(1) - \left(G'(1)\right)^2. \tag{3.86}$$

Exercise 3.12: Given the probability generating function $G(z) = (zp + q)^n$ where $p + q = 1$, show that $E(r) = np$, $V(r) = npq$. (Compare Exercise 3.11.)

3.10 SAMPLING

3.10.1 Universe and sample

A probability density function f(x) for a continuous random variable, or equivalently the set of probabilities in the discrete case, describes the properties of a *population*, or *universe*. In physics one associates random variables with observations on specified physical systems, and the p.d.f. f(x) summarizes the outcome of all conceivable measurements on such a system if the measurements were repeated infinitely many times under the same experimental conditions. Since an infinite number of observations is of course impossible, even on the simplest system, the concept of a population for a physicist represents an idealization which can never be attained in practice.

An actual experiment will consist of a finite number of observations. A sequence of measurements $x_1, x_2, \ldots, x_n$ on some quantity is said to constitute a *sample* of size n. A sample is accordingly a subset of the population or universe; we may say that " sample of size n is drawn from the universe". Physicists would like to think that their measurements are typical, in the sense that repeated experiments with the same number of measurements are likely to give more or less the same result. This corresponds to the notion of *random samples*.

3.10.2 Sample properties

Let the sample of size n be $x_1, x_2, \ldots, x_n$. This set of n independent, random variables possesses certain properties, which we may hope resemble those of the population. Two quantities to characterize the sample are

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{3.87}$$

and

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 . \tag{3.88}$$

Here $\bar{x}$ is the *sample mean*, or *arithmetic mean* (*average*), which we have already met; the quantity s^2 measures the dispersion of the sample about its mean value and is called the *sample variance*.

The two quantities $\bar{x}$ and s^2 defined here as functions*) of the random variables x_i, are themselves random variables. This is clearly so because a repeated drawing of new samples, all of size n, obtained from the same population will produce new $\bar{x}, s^2$. Thus the random variables $\bar{x}$ and s^2 will have their own distributions. Obviously, these distributions must depend on the properties of the parent distribution, and on n. The study of samples by the distributions of $\bar{x}$ and s^2 forms an important part of probability theory.

Particularly interesting are samples drawn from a normal universe. It turns out in this case that the variables $\bar{x}$ and s^2 are independent; this property is unique for the normal distribution. Moreover, the resulting distributions for the two variables become especially simple, $\bar{x}$ being normally distributed, and s^2 related to a chi-square distribution. This will be discussed further in Sects.4.8.6 and 5.1.6.

3.10.3 Inferences from the sample

A physicist's motivation for undertaking an experiment and to perform measurements of physical quantities is that he wants to find out something about "reality"; thus his interest is in some true distribution, or universe. In fact,

*) A function of one or more random variables that does not depend on any unknown parameter is called a *statistic*. In accordance with this definition $\bar{x}$, as well as s^2, may be called a statistic.

he may be prepared to make inferences about this universe on the basis of his restricted number of observations.

Suppose that measurements on the variable x have given the numbers $x_1, x_2, \ldots, x_n$, constituting a sample of size n. Evidently we hope that the sample in some respect is representative of the underlying universe or population. A measure of the population mean value μ which suggests itself is $\bar{x}$, the arithmetic mean of the sample, as given by eq.(3.87). We may therefore call $\bar{x}$ an *estimate of the population mean* μ. Similarly, a measure of the population variance σ^2 is provided by the quantity s^2 describing the dispersion of the sample, eq.(3.88). Hence the notion that s^2 is *the estimate of the population variance* σ^2. We write

$$\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i , \tag{3.89}$$

$$\hat{\sigma^2} = s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 . \tag{3.90}$$

The reason for using (n-1) and not n in the expression for s^2, is to ensure that s^2 is an *unbiased* estimator of σ^2; a discussion on this point is given in Sect.8.4.1. An intuitive explanation why we should take (n-1) instead of n is the following: From the sample alone we do not know exactly what the central value μ of the population is; we only have an estimate, $\hat{\mu}=\bar{x}$, which is subject to uncertainties. As a measure of the dispersion of the population the quantity $\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$ is therefore likely to be too small, and we should be better off replacing n in the denominator by a smaller number.

When n becomes very large the sample properties will approach the properties of the population, hence

$$\hat{\mu} = \bar{x} \to \mu ,$$

$$\hat{\sigma^2} = s^2 \to \sigma^2 .$$

This is the contents of general convergence theorems. In particular, the fact that the sample mean has the population mean as its limiting value is a consequence of the Law of Large Numbers, which will be discussed briefly in the following section. It can be intuitively understood from the observation that the expectation and variance of the variable $\bar{x}$ are, respectively, $E(\bar{x})=\mu$ and $V(\bar{x})=\sigma^2/n$, (eqs.(3.57), (3.58)), implying that the spread of $\hat{\mu}$ about μ will become small when n is large. Similar results apply for the mean and variance of s^2 (Exercise 3.13). Thus, by choosing sufficiently large samples, any desired accuracy can be obtained in the estimates of the population parameters. This property is called *consistency* of the estimators; see further Sect.8.3.

Exercise 3.13: Show that, in terms of the central moments,

$$E(s^2) = \sigma^2 = \mu_2 , \qquad V(s^2) = \frac{1}{n}\left(\mu_4 - \mu_2^2\right) + \frac{2}{(n-1)n}\,\mu_2^2 .$$

3.10.4 The Law of Large Numbers

Convergence theorems play a fundamental role in probability theory and statistics and are thoroughly discussed in treatises on the theoretical foundations of these subjects. Since in this book mathematical rigor is considered of less importance compared to practical implications we shall limit our discussion here to the *Law of Large Numbers* which was mentioned in the preceding section and which will be referred to in later applications.

Let $x_1, x_2, \cdots$ be a set of independent random variables which have identical distributions with mean value μ. For the first n of these variables the arithmetic mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ will also have mean value μ, regardless of the number n. The (weak) Law of Large Numbers states that, given any positive ε, the probability that $\bar{x}$ deviates from μ by an amount more than ε will be zero in the limit of infinite n,

$$\lim_{n \to \infty} P(|\bar{x}-\mu| > \varepsilon) = 0 . \tag{3.91}$$

As stated above the theorem concerns the limiting properties of $\bar{x}$ when n approaches infinity. A stronger version of the theorem says something about the behaviour of $\bar{x}$ for any value of n exceeding some finite value, say N. *Given two positive quantities* ε and δ, an N exists such that

$$P(|\bar{x}-\mu| > \varepsilon) < \delta \tag{3.92}$$

for all $n \geq N$.

The Law of Large Numbers can easily be adapted to the case when the

x's have different mean values. It will be observed that nothing has been said about the variance of the distributions. In fact, the theorem remains true even if the variances do not exist. If we assume that the variance of any x exists and is equal to σ^2 the proof for the theorem follows as an immediate consequence of the Bienaymé-Tshebycheff inequality (Exercise 3.2); when applied to the variable $\bar{x}$ of variance σ^2/n this inequality can be written

$$P(|\bar{x}-\mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2} .$$

Thus, for a given ε, the probability to have $|\bar{x}-\mu|>\varepsilon$ can be made arbitrarily small by choosing n large enough.

In cases where the variances exist a much more precise statement can be made about the behaviour of $\bar{x}$ when n becomes large. It turns out that $\bar{x}$ is to be *normally* distributed in these situations, regardless of the distributional shapes of the individual x's. This is the contents of the *Central Limit Theorem* , which will be discussed in Sect.4.8.8.

4. Special probability distributions

In this chapter we shall be concerned with an examination of those probability distributions which are, probably, most frequently encountered in practice. These distributions represent good approximations to "real life", and/or have particular theoretical importance. Fortunately, they all possess relative mathematical simplicity. In establishing the properties of the different ideal distributions we will make use of the general definitions and theorems from the preceding chapter and, whenever possible, provide illustrations by common, physical examples. The mathematical connection between the different probability distributions is worked out in some detail*); their physical connection is also pointed out in some cases.

Included in this chapter is a number of exercises. Some of these are rather formal and serve to fill a gap in the proof of a statement in the text, which usually requires little more than just a computational effort. Other exercises point out specific properties as well as useful relationships between the different distributions. A few exercises introduce other probability distributions which are related to those discussed in the text. These may be less well-known to particle physicists, but may have a direct, often quite simple mathematical or physical content. Finally, some exercises are included which provide the theoretical background for applications found in the later chapters of the book.

A theoretically and practically important class of probability distributions, the sampling distributions related to the normal p.d.f., is treated separately in Chapter 5.

*) For a graphical illustration of the relationships between the probability distributions, see Fig. 5.4 at the end of Chapter 5.

4.1 THE BINOMIAL DISTRIBUTION

We begin our discussion of probability distributions by considering first a few examples of distributions of random variables of the *discrete* type. The simplest situation one can think of involves a single, discrete variable which describes an experiment with only two possible outcomes.

4.1.1 Definition and properties

Let us denote the two exclusive outcomes of a random experiment by A and $\bar{A}$; A is called a "success" and $\bar{A}$ a "failure". For each experiment or trial let p $(0 \le p \le 1)$ be the probability that a success occurs, and $q = 1-p$ the probability for a failure. Then, in a sequence of n independent trials, the probability to have a total of r successes and $n-r$ failures is

$$B(r;n,p) = \binom{n}{r} p^r (1-p)^{n-r}, \qquad r = 0,1,2,\ldots,n. \tag{4.1}$$

This is the *binomial (or Bernoulli) distribution* for the variable r with the parameters n and p. The binomial coefficient $\binom{n}{r}$ takes account of the fact that the order of the individual r successes and $n-r$ failures is immaterial,

$$\binom{n}{r} \equiv \frac{n!}{r!(n-r)!} = \binom{n}{n-r}. \tag{4.2}$$

It is readily checked that the probabilities of eq.(4.1) are properly normalized, since they add to unity when summed over all r,

$$\sum_{r=0}^{n} B(r;n,p) = \sum_{r=0}^{n} \binom{n}{r} p^r (1-p)^{n-r} = \sum_{r=0}^{n} \binom{n}{r} p^r q^{n-r} = (p+q)^n = 1. \tag{4.3}$$

The binomial probabilities are invariant under the interchange $(r,p) \leftrightarrow (n-r,1-p)$,

$$B(r;n,p) = B(n-r;n,1-p). \tag{4.4}$$

The binomial distribution has been tabulated in Appendix Table A1 for 11 different values of p between 0.01 and 0.50, and for n up to 20. Appendix Table A2 gives a corresponding tabulation of the cumulative binomial distribution,

$$F(x;n,p) = \sum_{r=0}^{x} B(r;n,p), \qquad x = 0,1,\ldots,n. \tag{4.5}$$

The binomial distribution is symmetric when $p=q$ $(= 0.5)$, and otherwise skew. Figure 4.1 shows illustrations of the distribution for two values of the constant p $(= 0.2, 0.5)$ and three different n $(= 5,10,20)$. It is seen that the distribution gets increasingly symmetric for higher values of n. When n becomes large the binomial distribution takes an approximate normal (Gaussian) shape. Compare Exercise 4.3.

The mean value and variance for a variable which is distributed according to the binomial law, eq.(4.1), can be found as follows. First, the mean value of the distribution is obtained from the definition of the expectation value of a discrete variable, eq.(3.82),

$$E(r) = \sum_{r=0}^{n} rB(r;n,p) = \sum_{r=0}^{n} r\binom{n}{r} p^r (1-p)^{n-r}.$$

Since the first term drops out the summation limit can be changed from r=0 to r=1. Writing out the binomial coefficient and extracting the factor np outside the sum we get

$$E(r) = np \cdot \left[\sum_{r=1}^{n} \frac{(n-1)!}{(r-1)!((n-1)-(r-1))!} p^{r-1} (1-p)^{(n-1)-(r-1)} \right].$$

With the substitutions $s=r-1$, $m=n-1$ the sum becomes

$$\sum_{s=0}^{m} \frac{m!}{s!(m-s)!} p^s (1-p)^{m-s} = \sum_{s=0}^{m} \binom{m}{s} p^s (1-p)^{m-s} = 1$$

from eq.(4.3). Hence the mean value of a variable with a binomial distribution is

$$\mu = E(r) = np. \tag{4.6}$$

To find the variance from eq.(3.83) we observe that one can write $E(r^2) = E(r(r-1) + r) = E(r(r-1)) + E(r)$, where

$$E(r(r-1)) = \sum_{r=0}^{n} r(r-1) \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r} = \sum_{r=2}^{n} \frac{n!}{(r-2)!(n-r)!} p^r (1-p)^{n-r}$$

$$= n(n-1)p^2 \sum_{r=2}^{n} \frac{(n-2)!}{(r-2)!((n-2)-(r-2))!} p^{r-2} (1-p)^{(n-2)-(r-2)} = n(n-1)p^2.$$

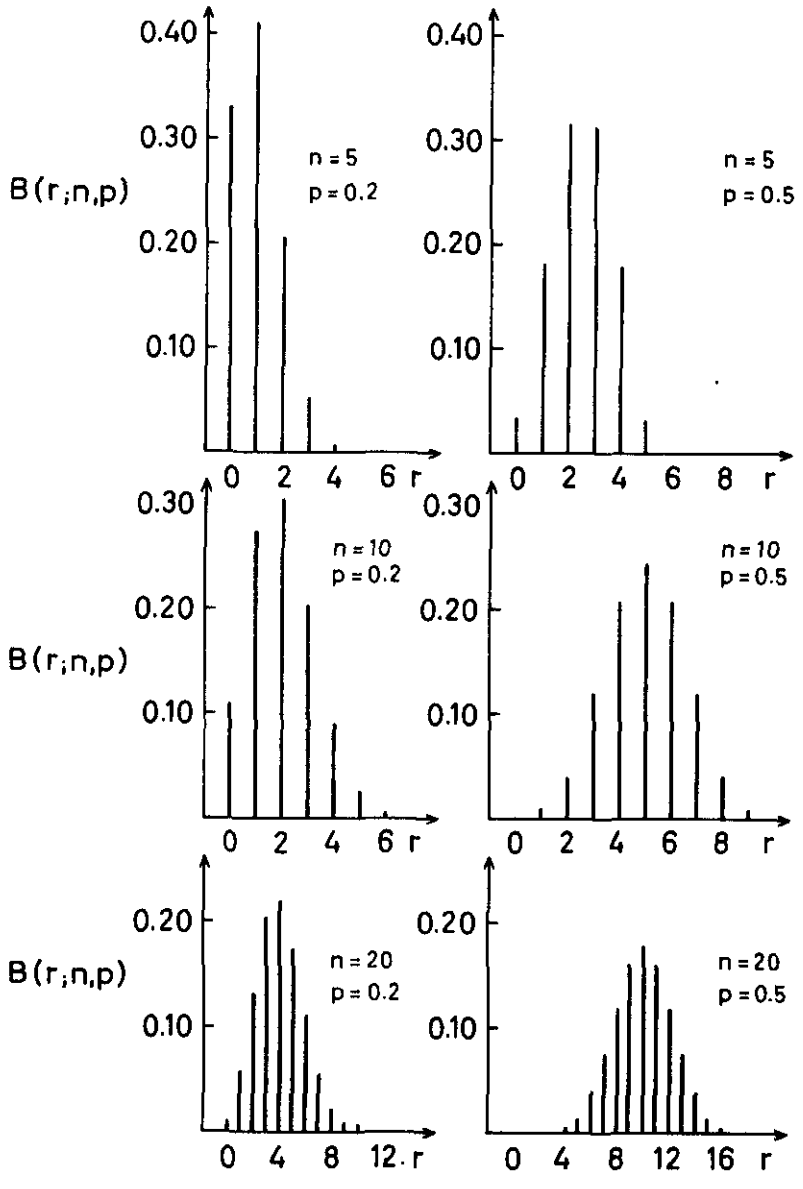

Fig. 4.1. The binomial distribution for indicated values of the parameters n,p.

The variance of the binomial variable is therefore

$$V(r) = E(r^2) - (E(r))^2 = n(n-1)p^2 + np - (np)^2 = np(1-p) = npq. \quad (4.7)$$

One is often interested in the quantity $\frac{r}{n}$, the relative number of successes in n trials. For this variable the mean and variance are given by

$$E\left(\frac{r}{n}\right) = \frac{1}{n} E(r) = p, \quad (4.8)$$

$$V\left(\frac{r}{n}\right) = \left(\frac{1}{n}\right)^2 V(r) = \frac{p(1-p)}{n} = \frac{pq}{n}. \quad (4.9)$$

Exercise 4.1: From the definition of the cumulative binomial distribution by eqs.(4.1),(4.5), show that, for $0 \leq x \leq n-1$,

$$F(x;n,p) = 1 - F(n-x-1;n,1-p).$$

Exercise 4.2: Show from its definition by eq.(3.84) that the probability generating function for the binomial distribution is $G(z) = (zp + q)^n$.

Exercise 4.3: Show from the definitions eqs.(3.20), (3.21) that the asymmetry and kurtosis coefficients of the binomial distribution are given by, respectively,

$$\gamma_1 = (1-2p)/\sqrt{np(1-p)}, \qquad \gamma_2 = \{1-6p(1-p)\}/\{np(1-p)\}.$$

Observe from the expression for γ_1 that, for finite n, $p < 0.5$ ($p > 0.5$) implies that the distribution is positively (negatively) skew and has a tail to the right (left). Note also that both coefficients tend to zero when n becomes large, indicating that the binomial distribution becomes similar to the normal distribution (Sects.3.3.3 and 4.8.4).

4.1.2 Example: Histogramming events (1)

As an application of the binomial distribution suppose that we for some reason are interested in just one particular bin of a compound histogram. Then A (success) may correspond to getting an entry in this particular bin, say bin number i, and $\bar{A}$ (failure) corresponds to an entry in any other bin of the histogram. With a total of n independent events the probability for having just r events in bin i and the remaining n-r events distributed over the other bins is given by the binomial distribution law, eq.(4.1). The expected number of events in the i-th bin is $E(r) = np$ from eq.(4.6), and the variance of this number $V(r) = np(1-p)$, eq.(4.7).

The probability p for a success is some constant whose exact value may not be known prior to the experiment. When the experiment has been performed,

giving r events in the i-th bin out of a total of n events, we may adopt for p its *estimated value*

$$p = \hat{p} = \frac{r}{n}.$$

The number of successes r has an estimated variance

$$V(r) = n\hat{p}(1-\hat{p}) = r\left(1-\frac{r}{n}\right)$$

and a standard deviation

$$\sigma(r) = \sqrt{r\left(1-\frac{r}{n}\right)}\,. \tag{4.10}$$

The last result implies that the error on the number of events r in the i-th bin is <u>not</u> $\sqrt{r}$, but a smaller quantity. Only in the limit when $p \to 0$, usually corresponding to a large number of bins, is $\sigma(r) \simeq \sqrt{r}$. In fact, this asymptotic limit reflects the condition for a variable with a Poisson distribution.

<u>4.1.3 Example: Scanning efficiency (2)</u>

We shall take up again the problem with the scanning efficiencies which was introduced in Sect.2.3.11.

Suppose for the moment that we know the efficiencies ε_1 and ε_2 of two individual, independent scans. The overall scanning efficiency ε is then given by

$$\varepsilon = \varepsilon_1 + \varepsilon_2 - \varepsilon_1\varepsilon_2\,, \tag{4.11}$$

(compare eq.(2.21), which is the same, except for an obvious change in notation). We want to find expressions for the errors to be associated with the quantities $\varepsilon_1, \varepsilon_2, \varepsilon$.

The scanning process itself is of binomial nature, because either an event is registered by the scanner, or it is not. We can therefore for the individual scans apply the formulae of the binomial distribution. Thus the standard deviation of the scanning efficiency for the i-th scan is obtained from eq.(4.9),

$$\sigma(\varepsilon_i) = \sqrt{\frac{\varepsilon_i(1-\varepsilon_i)}{N}}\,, \tag{4.12}$$

where N is the total number of events contained in the film.

For the overall scanning efficiency the error is found by applying the law of propagation of errors to eq.(4.11). If one assumes that ε_1 and ε_2 are independent variables, the variance of ε is given by

$$V(\varepsilon) = \left(\frac{\partial\varepsilon}{\partial\varepsilon_1}\right)^2\sigma^2(\varepsilon_1) + \left(\frac{\partial\varepsilon}{\partial\varepsilon_2}\right)^2\sigma^2(\varepsilon_2). \tag{4.13}$$

Hence eqs.(4.11)-(4.13) lead to the following expression for the standard deviation of the overall scanning efficiency,

$$\sigma(\varepsilon) = \sqrt{\frac{(1-\varepsilon_1)(1-\varepsilon_2)(\varepsilon_1+\varepsilon_2-2\varepsilon_1\varepsilon_2)}{N}}\,. \tag{4.14}$$

In the error formulae (4.12) and (4.14) N is the true number of events contained in the film. Usually N is not known exactly and has to be estimated from the number of events found in the two independent scans.

Strictly speaking, the assumptions stated above are not fulfilled in practice. Specifically, the efficiencies $\varepsilon_1, \varepsilon_2$ are not independently determined, since they, as well as the total number of events N, have to be estimated from the number of events found in the two independent scans, as indicated by the formulae derived in Sect.2.3.11. Thus a better estimate of the errors would require a more elaborate treatment of the error propagation starting from the observed quantities N_1, N_2, N_{12}. This is presumably only seldom tried in practice, since the systematic errors associated with the scanning procedure in most cases are assumed to be more important than the pure statistical error in the overall efficiency.

<u>Exercise 4.4:</u> (The geometric distribution)

(i) Show that, if p is the probability for having a success in each binomial trial, the probability that the first success occurs in the r-th trial is

$$P_1(r;p) = p(1-p)^{r-1},$$

and verify that $P_1(r)$ with r=1,2,... gives a correctly normalized probability distribution for the occurrence of the first success.

(ii) Show that this *geometric distribution* has the probability generating function

$$G(z) = pz/(1-qz), \qquad (q=1-p).$$

(iii) By the use of G(z), show that the mean and variance of this distribution are given by, respectively,

$$E(r) = 1/p, \qquad V(r) = (1-p)/p,$$

and that the asymmetry and curtosis coefficients are

$$\gamma_1 = (2-p)/\sqrt{1-p}, \quad \gamma_2 = (p^2-6p+6)/(1-p).$$

The geometric distribution is illustrated in the upper part of Fig. 4.2 for a few values of the parameter p.

Exercise 4.5: (The negative binomial distribution)

(i) Generalizing the situation from the preceding exercise, show that the probability for obtaining the k-th success in the r-th trial is given by the *negative binomial* (or *Pascal*) *distribution*

$$P_k(r;p) = \binom{r-1}{k-1} p^k (1-p)^{r-k}, \qquad r=k,k+1,....$$

(ii) Show that this probability distribution has the properties

$$E(r) = k/p, \qquad V(r) = k(1-p)/p^2,$$

$$\gamma_1 = (2-p)/\sqrt{k(1-p)}, \qquad \gamma_2 = (p^2-6p+6)/k(1-p).$$

(Hint: The probability generating function is $G(z) = \{pz/(1-qz)\}^k$.)
Note that this distribution is always positively skew.

(iii) Introduce the number of failures, s=r-k, and show that the probability distribution for having s failures when the k-th success occurs can be written

$$P_k(s;k) = \binom{s+k-1}{s} p^k (1-p)^s, \qquad s=1,2,....$$

Verify that this distribution has the mean value shifted (reduced) by the amount k, but that all central moments, and hence $V(s),\gamma_1,\gamma_2$, are as above.

The negative binomial distribution is shown in Fig. 4.2 for a few combinations of the parameters p and k. The probabilities for k=1 in the upper part of the figure correspond to the geometric distribution of Exercise 4.4.

Exercise 4.6: (The hypergeometric distribution (1))

(i) Suppose that of N elements, a have the attribute A and the remaining N-a the attribute $\bar{A}$. Show that, when n elements are picked at random and without replacement from the total N elements, the probability that the random sample of size n will contain r elements with the attribute A and n-r elements with the attribute $\bar{A}$, is given by the *hypergeometric distribution*,

$$P(r;N,n,a) = \binom{N}{n}^{-1} \binom{a}{r} \binom{N-a}{n-r}, \qquad r=0,1,....,\min\{a,n\}.$$

(ii) Show that when N >> n, this distribution reduces to the ordinary binomial distribution of eq.(4.1) with p=a/N, in accordance with common sense expectation.

The conditions above can be generalized to a classification in more than two categories; see Exercise 4.8.

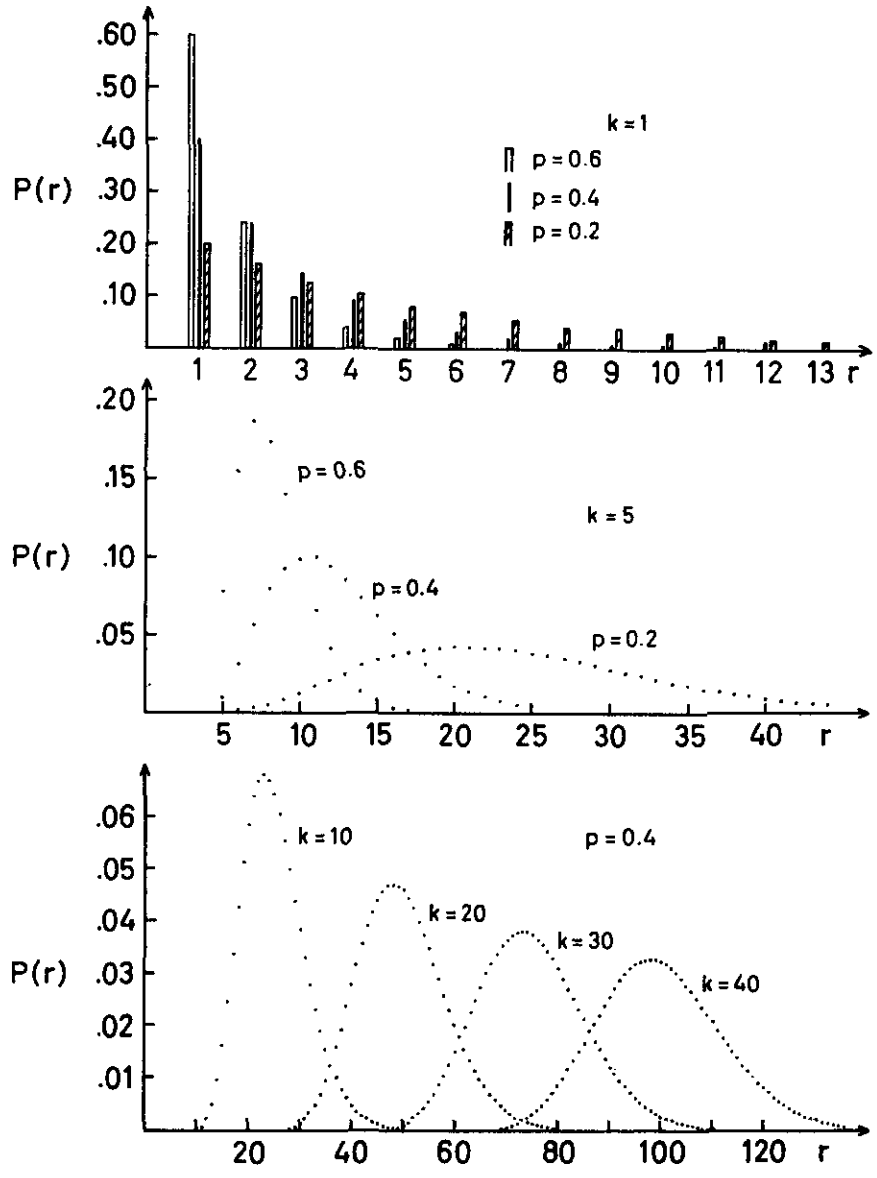

Fig. 4.2. The negative binomial distribution (Exercise 4.5) for indicated values of the parameters k,p. For k=1 one has the geometric distribution (Exercise 4.4).

4.2 THE MULTINOMIAL DISTRIBUTION

4.2.1 Definition and properties

The generalization of the binomial conditions to the case with more than two possible outcomes of an experimental trial leads to the *multinomial distribution* law.

Let the possible outcomes define a set of categories or classes $A_1, A_2, \ldots, A_k$. For each trial the probability of an outcome in the specific class A_i is p_i. Then since every trial must give some outcome, the probabilities must add to unity,

$$\sum_{i=1}^{k} p_i = 1. \tag{4.15}$$

Now we assume that the outcomes of different trials are independent. After n independent trials the probability of having a final result with $r_1, r_2, \ldots, r_k$ outcomes in the different classes is given by

$$M(\underline{r};n,\underline{p}) = \frac{n!}{r_1!r_2!\ldots r_k!} p_1^{r_1} p_2^{r_2} \ldots p_k^{r_k}. \tag{4.16}$$

Eq.(4.16) gives the multinomial distribution of the variables $\underline{r} = \{r_1, r_2, \ldots, r_k\}$ for the parameters n and $\underline{p} = \{p_1, p_2, \ldots p_k\}$. The r_i are not all independent, since

$$\sum_{i=1}^{k} r_i = n. \tag{4.17}$$

Evidently the multinomial distribution eq.(4.16) includes the binomial distribution eq.(4.1) as a special case. In addition it has the following properties:

(i) The expectation value for class A_i is $E(r_i) = np_i$.

(ii) The variance for class A_i is $V(r_i) = np_i(1-p_i)$.

(iii) The covariance for classes A_i, A_j is $\mathrm{cov}(r_i, r_j) = -np_ip_j$.

(iv) When n becomes large the multinomial distribution tends to a multinormal distribution.

To prove properties (i) and (ii), notice that the probability for getting r_i outcomes in the class i in a total of n trials is

$$\binom{n}{r_i} p_i^{r_i} (1-p_i)^{n-r_i}.$$

This is an example of the binomial case: either a trial gives an outcome in the class A_i (the probability for this occurrence being p_i) or it does not (probability $1-p_i$). Therefore the formulae from the binomial distribution eqs.(4.6), (4.7) can be taken over to give the expectation and variance in one particular class.

To prove property (iii), consider the two classes A_i and A_j together. The probability for r_i outcomes in A_i and r_j outcomes in A_j, and with the outcomes of the remaining $n-r_i-r_j$ trials distributed over all other classes, is

$$\frac{n!}{r_i!r_j!(n-r_i-r_j)!} p_i^{r_i} p_j^{r_j} (1-p_i-p_j)^{n-r_i-r_j}.$$

It is easy to show that the expectation value of the product r_ir_j with the p.d.f. of eq.(4.16) is

$$E(r_ir_j) = n(n-1)p_ip_j. \tag{4.18}$$

With this result one gets the covariance from eq.(3.39),

$$\mathrm{cov}(r_ir_j) = E(r_ir_j) - E(r_i)\cdot E(r_j) = n(n-1)p_ip_j - (np_i)(np_j) = -np_ip_j,$$

which was stated above.

Notice that the contents in the two classes are always negatively correlated. In terms of the correlation coefficient from eq.(3.40) one has

$$\rho(r_i,r_j) = \frac{\mathrm{cov}(r_i,r_j)}{\sigma_i\sigma_j} = -\sqrt{\frac{p_ip_j}{(1-p_i)(1-p_j)}}. \tag{4.19}$$

Exercise 4.7: Prove eq.(4.18).

4.2.2 Example: Histogramming events (2)

As an example of the multinomial distribution we consider n events distributed among k bins in a histogram. Then p_i is the probability that an event will fall in bin number i, and r_i is the number of events in this bin. The mean value and variance of the variable r_i are, respectively,

$$E(r_i) = np_i \ , \qquad V(r_i) = np_i(1-p_i) \ .$$

If $p_i \ll 1$, corresponding to a distribution with many bins in general, we have $V(r_i) \simeq np_i = r_i$, and the standard deviation becomes $\sigma(r_i) \simeq \sqrt{r_i}$.

The covariance for the number of events in bins i and j is the negative number

$$\mathrm{cov}(r_i,r_j) = - np_ip_j \ .$$

Thus the correlation between two bins is only negligible if the probabilities for the two bins are small, corresponding to a histogram with many bins.

The multinomial classification considers n, the total number of events, a fixed number (parameter). If instead we regard the number of events r_i in the different bins as independent, random variables of the Poisson types, they will have standard deviations $\sigma(r_i)=\sqrt{r_i}$, and the total number of events $n=\sum_{i=1}^{k} r_i$ will then be a random variable of the Poisson type; see Exercise 4.18 and Sect.4.4.4.

Exercise 4.8: (The generalized hypergeometric distribution (2))

(i) A population consists of N elements which can be classified into k different categories. There are a_1 elements in the first category, a_2 in the second, and so on, $\sum_{i=1}^{k} a_i = N$. Show that, if a random sample of size n is drawn from this population, (that is, if n elements are picked at random among the N elements and without replacement), the probability distribution for obtaining $r_1, r_2, \ldots, r_k$ elements of the different categories is

$$P(\underline{r};N,n,\underline{a}) = \binom{N}{n}^{-1} \prod_{i=1}^{k} \binom{a_i}{r_i} , \qquad r_i = 0,1,\ldots,\min\{a_i,n\},$$

where $\sum_{i=1}^{k} r_i = n$.

(ii) Show that if the population is very large compared to the size of the sample ($N \gg n$), this distribution reduces to the multinomial distribution, of eq.(4.16) with $p_i = a_i/N$, $i=1,2,\ldots,k$.

4.3 THE POISSON DISTRIBUTION

4.3.1 Definition and properties

With binomial conditions it sometimes happens that the rate p of "successes" is very small. In a long series of n trials the total number of successes np may, however, still be considerable. It is therefore appealing to examine mathematically the limiting case of the binomial distribution, when $p \to 0$, $n \to \infty$ in such a way that the product np remains constant and equal to μ, say. From Stirling's formula for the factorial of a large number, $n! \approx \sqrt{2\pi n}\, n^n e^{-n}$, we have under these conditions for the terms of the binomial distribution

$$\frac{n!}{r!(n-r)!} p^r (1-p)^{n-r} \approx \frac{1}{r!} \frac{\sqrt{2\pi n}\, n^n e^{-n}}{\sqrt{2\pi(n-r)}\,(n-r)^{n-r} e^{-(n-r)}} \left(\frac{\mu}{n}\right)^r \left(1-\frac{\mu}{n}\right)^{n-r}$$

$$\approx \frac{1}{r!} \frac{1}{\left(1-\frac{r}{n}\right)^n e^r} \mu^r \left(1-\frac{\mu}{n}\right)^n .$$

Thus, we can write

$$P(r;\mu) = \frac{1}{r!} \mu^r e^{-\mu} , \qquad r=0,1,2,\ldots \qquad (4.20)$$

This is the *Poisson distribution* for the discrete variable r, with the parameter (mean value) μ.

The distribution of eq.(4.20) is evidently correctly normalized. The probability generating function (eq.(3.84)) becomes

$$G(z) \equiv E(z^r) = \sum_{r=0}^{\infty} z^r \frac{1}{r!} \mu^r e^{-\mu} = \sum_{r=0}^{\infty} \frac{1}{r!} (\mu z)^r e^{-\mu} = e^{\mu(z-1)} . \qquad (4.21)$$

From the general expressions eqs.(3.85), (3.86) the expectation and variance of the distribution come out as

$$E(r) = G'(1) = \mu \ ,$$

$$V(r) = G''(1) + G'(1) - \left[G'(1)\right]^2 = \mu \ .$$

Thus the Poisson distribution has the property that the mean equals the variance,

$$E(r) = V(r) = \mu . \qquad (4.22)$$

The Poisson distribution is very asymmetric for small μ and has a tail to the right of the mean. Quantitatively, the asymmetry is expressed by the positive skewness coefficient (eq.(3.20))

$$\gamma_1 = \frac{1}{\sqrt{\mu}} , \tag{4.23}$$

which approaches zero when μ gets large. Asymptotically, when μ goes towards infinity the Poisson distribution becomes identical to the normal distribution. As will be seen from Fig. 4.3 the similarity between these two distributions is rather close already at μ=20.

From eq.(4.20) we observe that

$$P(r;\mu) = P(r-1;\mu) \cdot \frac{\mu}{r} , \tag{4.24}$$

and the probability will therefore increase from r=0,1,2 *etc.* so long as $r < \mu$. The maximum probability is at $r = [\mu]$, and with an equal, adjacent maximum at $\mu-1$ if μ is an integer; see Fig. 4.3.

The Poisson distribution of eq.(4.20) has been tabulated in Appendix Table A3 for values of μ between 0.1 and 20. Appendix Table A4 gives a similar tabulation of the cumulative Poisson distribution

$$F(x;\mu) = \sum_{r=0}^{x} P(r;\mu) . \tag{4.25}$$

The tables of the Poisson distribution involve only one parameter and are easier to work with than the corresponding tables of the two-parameter binomial distribution. Because of the limiting relationship, the tables of the Poisson distribution represent a convenient approximation of the binomial tables when p is small and n sufficiently large (μ=np).

There exists a useful relationship between the cumulative Poisson sum of eq.(4.25) and the cumulative integral of the chi-square distribution, which we shall discuss in Chapter 5,

$$F(x;\mu) = 1 - \int_0^{2\mu} f(u;\nu=2x+2)\,du . \tag{4.26}$$

Here $f(u;\nu)$ is the chi-square p.d.f. with ν degrees of freedom, and the quantity on the right-hand side has been displayed graphically for different ν in Fig.5.2.

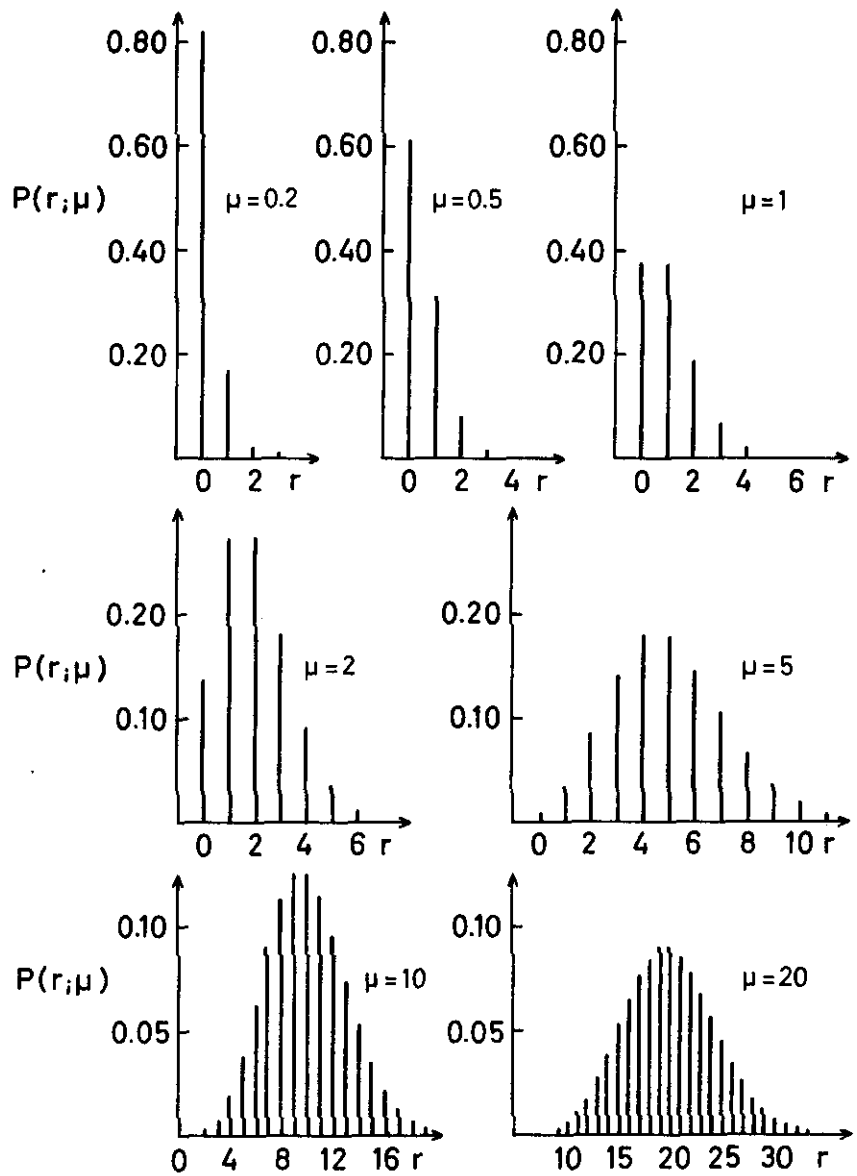

Fig. 4.3. The Poisson distribution for different mean values μ.

From this graph one can therefore read off directly the value of $F(x;\mu)$ on the curve for $v=2x+2$ at the value $u=2\mu$.

Exercise 4.9: Derive the mean value and the variance of the Poisson distribution by carrying out the summations for $E(r)$ and $E(r^2)$ with the probability distribution (4.20).

Exercise 4.10: Show that the Poisson distribution has the probability generating function $G(z) = \exp(\mu(z-1))$.

Exercise 4.11: Show that the asymmetry and kurtosis coefficients of the Poisson distribution are given by $\gamma_1=1/\sqrt{\mu}$ and $\gamma_2=1/\mu$, respectively. Both coefficients tend to zero when μ gets large, indicating an increasing similarity to the normal distribution (Sect.4.8.1).

Exercise 4.12: Show that the Poisson distribution has algebraic moments connected by the relation

$$\mu'_{k+1} = \mu\left(\mu'_k + \frac{d\mu'_k}{d\mu}\right).$$

Exercise 4.13: Show that the Poisson distribution has the characteristic function $\Phi(t) = \exp\left(\mu(e^{it}-1)\right)$.

4.3.2 The Poisson assumptions

Example: Bubbles along a track in a bubble chamber

In the previous section the Poisson distribution was introduced as a limiting case of the binomial distribution. To get more insight into its meaning and applicability as well as the basic assumptions underlying a Poisson law we discuss next in this and the following section two examples of physical processes which are well described by this law. In fact, starting from first principles we shall now derive the Poisson distribution formula.

Let us here think of the distribution of bubbles formed along the paths of charged particles in a bubble chamber. We assume for the moment that the size of the bubbles can be ignored and that the mean number of bubbles per unit length along the track is constant. We denote this average bubble density by g, and consider a small length $\Delta\ell$ of the track. The three basic *Poisson assumptions* may then be formulated as follows:

(i) There is at most one bubble in the interval $[\ell,\ell+\Delta\ell]$.

(ii) The probability for finding one bubble in this interval is proportional to $\Delta\ell$.

(iii) The occurrence of a bubble in the interval $[\ell,\ell+\Delta\ell]$ is independent of the occurrence of bubbles in any other non-overlapping interval.

From assumptions (i) and (ii) the probability that there is one bubble in the interval $[\ell,\ell+\Delta\ell]$ is

$$P_1(\Delta\ell) = g\Delta\ell,$$

while the probability that there is no bubble in the interval is

$$P_0(\Delta\ell) = 1 - P_1(\Delta\ell) = 1 - g\Delta\ell.$$

Assumption (iii) implies that the occurrence of no bubble in $\Delta\ell$ is independent of the presence of no bubbles over the distance ℓ, and hence a factorization of the probability for the occurrence of no bubbles over the length $\ell+\Delta\ell$

$$P_0(\ell+\Delta\ell) = P_0(\ell)P_0(\Delta\ell).$$

Combining the last two expressions we may write

$$\frac{P_0(\ell+\Delta\ell) - P_0(\ell)}{\Delta\ell} = -gP_0(\ell).$$

When $\Delta\ell \to 0$ the left-hand side of this expression is the derivative of $P_0(\ell)$ with respect to ℓ, hence we get

$$\frac{dP_0(\ell)}{d\ell} = -gP_0(\ell). \tag{4.27}$$

The differential equation (4.27) can, because of the assumed constancy of g, easily be solved with the boundary condition $P_0(0) = 1$,

$$P_0(\ell) = e^{-g\ell}. \tag{4.28}$$

Thisformula gives the probability that there are no bubbles over the length ℓ.

Let us next find the probability for observing r bubbles within the length ℓ. Since there can be at most one bubble in the small interval $[\ell,\ell+\Delta\ell]$ we may write

$$P_r(\ell+\Delta\ell) = P_r(\ell)\cdot P_o(\Delta\ell) + P_{r-1}(\ell)\cdot P_1(\Delta\ell),$$

where the first term on the right implies all r bubbles in ℓ, and the second term implies (r-1) bubbles in ℓ and one bubble in $\Delta\ell$. Introducing the probabilities $P_o(\Delta\ell)$ and $P_1(\Delta\ell)$ from assumptions (i) and (ii) we get after rearranging

$$\frac{P_r(\ell+\Delta\ell) - P_r(\ell)}{\Delta\ell} = -gP_r(\ell) + gP_{r-1}(\ell).$$

Again, when $\Delta\ell \to 0$, the left-hand side is a derivative, and leads to the differential equation

$$\frac{dP_r(\ell)}{d\ell} = -gP_r(\ell) + gP_{r-1}(\ell). \qquad (4.29)$$

The solution of this equation is

$$P_r(\ell) = \frac{1}{r!}(g\ell)^r e^{-g\ell}, \qquad (4.30)$$

which gives the distribution of the number of bubbles r in intervals of fixed length ℓ. It is seen that eq.(4.30) gives a Poisson distribution with the parameter $(g\ell)$. Specifically it includes eq.(4.28) as a special case.

It may be appropriate to emphasize that eq.(4.30) describes the frequency distribution for the *discrete* variable r, with ℓ (or, strictly speaking, $g\ell$) as a parameter of the distribution. From the Poisson assumptions one can also turn the problem around and seek the distribution in the *continuous* variable ℓ for specified values of the parameter r. In other words, one can ask for the probability to have a total distance ℓ of the track to find exactly r bubbles, given that the average probability to find a bubble is constant along the track and equal to g per unit length. This problem will be investigated later in Sects.4.6 and 4.7, and as we shall see, it will lead us to a specific class of the gamma distributions, including the well-known exponential distribution law.

Let it also be pointed out that it is essential in the derivation of the formulae above that the average bubble density g per unit length is a *constant*, independent of ℓ. If $g=g(\ell)$ the resulting distribution of bubbles will not be Poisson but some other, generally unknown, distribution; see also Exercise 4.35.

Exercise 4.14: Verify that eq.(4.30) is the solution of eq.(4.29).

Exercise 4.15: Suppose that the average bubble density of minimum ionizing particles of charge $\pm e$ in a bubble chamber is 9 bubbles per cm. What is the probability that a relativistic particle will produce only 1 bubble per cm, equal to the bubble density expected from a quark of charge $\pm 1/3e$? If the minimum ionization were two times larger, *i.e.* 18 bubbles per cm, what would then the probability be for observing a track with 1/9 of this average bubble density?

Exercise 4.16: The number of beam particles per pulse is assumed to be Poisson distributed. If it is known that the average number of particles per pulse is 16, what is the probability that a pulse will have between 12 and 20 particles? (Answer: 0.741.) - See also Exercise 4.40.

Exercise 4.17: In an experimental search for weak neutral currents 9 candidates were observed for the neutrino reactions

$$\nu + p \to \nu + p + \pi^o,$$
$$\nu + p \to \nu + n + \pi^+.$$

The candidates could, however, also be interpreted as background events due to the neutron reactions

$$n + p \to n + p + \pi^o,$$
$$n + p \to n + n + \pi^+.$$

From identified events of the type $n+p\to p+p+\pi^-$ the expected number of background events was estimated to be 4.9. Assuming that the number of background events is Poisson distributed with mean value 4.9, what is the probability to have 9 or more background events? Do you consider that this experiment indicates the presence of weak neutral currents?

4.3.3 Example: Radioactive emissions

A frequently cited example on a Poisson process is that of particle emission from a radioactive source. If the particles are emitted from the source at an average rate of λ_x particles per unit time the number of emissions r_x in fixed time intervals t follows a Poisson law with mean $\lambda_x t$,

$$P(r_x;\lambda_x t) = \frac{1}{r_x!}(\lambda_x t)^{r_x} e^{-\lambda_x t}.$$

Suppose now that the source is placed in surroundings where the background of radioactive emissions is given by an average rate λ_b particles per unit time. Then the number r_b of background emissions in time intervals of length t is Poisson distributed with mean $\lambda_b t$,

$$P(r_b;\lambda_b t) = \frac{1}{r_b!}(\lambda_b t)^{r_b} e^{-\lambda_b t}.$$

Accessible for measurement is the sum of background and source emissions. Let the number of counts in an interval t be denoted by r. The distribution for the variable r can then be expressed as

$$\begin{aligned} P(r;\lambda_x t,\lambda_b t) &= \sum_{r_b=0}^{r} P(r-r_b;\lambda_x t)\cdot P(r_b;\lambda_b t) \\ &= \sum_{r_b=0}^{r}\left[\frac{1}{(r-r_b)!}(\lambda_x t)^{r-r_b} e^{-\lambda_x t}\right]\left[\frac{1}{r_b!}(\lambda_b t)^{r_b} e^{-\lambda_b t}\right] \\ &= \frac{1}{r!}\sum_{r_b=0}^{r}\frac{r!}{(r-r_b)!r_b!}(\lambda_x t)^{r-r_b}(\lambda_b t)^{r_b} e^{-(\lambda_x+\lambda_b)t}, \end{aligned}$$

or,

$$P(r;\lambda_x t,\lambda_b t) = \frac{1}{r!}\big((\lambda_x+\lambda_b)t\big)^r e^{-(\lambda_x+\lambda_b)t} \equiv P\big(r;(\lambda_x+\lambda_b)t\big). \qquad (4.31)$$

Thus the observed emissions are also Poisson distributed, with a rate equal to the sum of the rates of the source and background.

With this example we have arrived at the *addition theorem for Poisson distributed variables*. The generalization to more variables is obvious: The sum of any number of independent Poisson variables is itself a Poisson variable with mean value equal to the sum of the individual means. See also Exercise 4.18 below.

Exercise 4.18: Let $r_1, r_2, \ldots, r_n$ be a set of independent Poisson variables with mean values $\mu_1, \mu_2, \ldots, \mu_n$. Show, using the characteristic function technique of Sect.3.4 and the result of Exercise 4.13, that $r = \sum_{i=1}^{n} r_i$ is also a Poisson variable with mean value $\mu = \sum_{i=1}^{n} \mu_i$.

4.4 RELATIONSHIPS BETWEEN THE POISSON AND OTHER PROBABILITY DISTRIBUTIONS

The Poisson distribution has some interesting connections to other probability distributions which are useful for physical applications. We will in the following indicate how the same mathematical relationships between the Poisson and other distributions can come out when a physical problem is attacked from different viewpoints which may at first appear rather dissimilar. We study in particular the connections between the Poisson and binomial/multinomial distribution laws which are important for many practical problems. In the final section we indicate an extension to the compound Poisson distribution, adequate for the description of chained reactions.

4.4.1 Example: Distribution of counts from an inefficient counter

Suppose that charged particles are counted by some electronic device which has a probability $p < 1$ for registering a particle traversing it; the registration probability is the same for all particles. We assume further that the number of particles n entering and traversing the device in fixed time intervals t is Poisson distributed with mean value ν. We want to find the distribution for the number of registrations r over time intervals of length t.

To have r counts by the inefficient device at least r particles must have traversed it. The probability we seek is therefore obtained by adding all probabilities that will give r registrations. For a given n the probability of getting r counts is given by the binomial law. Thus, when all conditional probabilities are added with n running from r to infinity, we get

$$\begin{aligned} P(r) &= \sum_{n=r}^{\infty} B(r;n,p)\cdot P(n;\nu) \\ &= \sum_{n=r}^{\infty}\left(\frac{n!}{r!(n-r)!}p^r(1-p)^{n-r}\right)\cdot\left(\frac{1}{n!}\nu^n e^{-\nu}\right) \\ &= \frac{1}{r!}(p\nu)^r e^{-\nu}\sum_{n=r}^{\infty}\frac{1}{(n-r)!}\big((1-p)\nu\big)^{n-r} = \frac{1}{r!}(p\nu)^r e^{-\nu}\cdot e^{(1-p)\nu}, \end{aligned}$$

or,

$$P(r) = \frac{1}{r!}(p\nu)^r e^{-p\nu}, \qquad r = 0, 1, \ldots, \qquad (4.32)$$

which is nothing but a Poisson distribution with mean value $p\nu$.

An inefficient counter of the above type has the property that it picks a random sample from the parent population. With a Poisson population the random sample was found to be of the Poisson type. Conversely, the sample will only be Poisson distributed if the population was Poisson.

4.4.2 Example: Subdivision of a counting interval

We assume that a detector registers particles over periods of T seconds and that the number of counts n follows a Poisson law with mean value $\nu=\lambda T$. We want to find the distribution describing the number of counts r in an interval of t seconds, where $t < T$.

From the specification above the counts occur at a rate of λ counts per second. Since the basic process is of the Poisson type the occurrences of events in two non-overlapping time intervals are independent. The probability to have r counts in the interval t and n-r counts in the remaining time T-t is therefore equal to the product of the two Poisson probabilities,

$$P(r) = P(r;\lambda t)\cdot P\big(n-r;\lambda(T-t)\big).$$

This is not the probability distribution we seek, because the normalization is not correct. We need the conditional probability to have r and n-r counts, given a total n counts. The conditional probability is therefore obtained by dividing P(r) by the Poisson probability for n counts in the time T, thus

$$P(r \text{ in } t,\ n-r \text{ in } T-t \,|\, n \text{ in } T) = \frac{P(r;\lambda t)P\big(n-r;\lambda(T-t)\big)}{P(n;\lambda T)}.$$

Substituting the explicit Poisson probabilities on the right-hand side and rearranging terms we get

$$P(r \text{ in } t,\ n-r \text{ in } T-t \,|\, n \text{ in } T) = \frac{n!}{r!(n-r)!}\left(\frac{t}{T}\right)^r\left(1-\frac{t}{T}\right)^{n-r} \equiv B\left(r;n,\frac{t}{T}\right). \quad (4.33)$$

This is a binomial distribution law for the variable r, with parameters n and t/T.

The result found is identical to what we would have obtained with a less sophisticated approach, considering the total number of counts n a fixed number, equal to the number of independent binomial "trials". In this line of

thought a "success" would correspond to having the count occur in the time t, while a "failure" would be that it occurred in the remaining time T-t. With an average counting rate n/T per second, the success rate, or the probability for each count occurring in the time t, is $p=\frac{1}{n}(n/T)t=t/T$. Thus the probability for r counts in t and n-r counts in T-t is given by the expression above.

4.4.3 Relation between binomial and Poisson distributions

Example: Forward-backward classification

The mathematical relationship between the binomial and the Poisson probabilities of the preceding section can be derived from an alternative point of view, assuming two variables related in a binomial distribution, with their sum obeying a Poisson law. The formulation below involving two variables can easily be generalized to several variables, see the subsequent section.

For the moment, let us suppose that we make a classification of particles in two categories, "forward" and "backward", according to their production angle in an overall centre-of-mass system. The total number of particles n is assumed to be a Poisson variable with mean value ν. For any n the number of particles in the forward (f) and backward (b) hemispheres are conditioned through a binomial law, which we write

$$B(f,b;n,p,q) = \frac{n!}{f!b!}p^f q^b.$$

Here p and q are the constants describing the fractions of forward and backward particles, respectively. The joint probability distribution for all three variables f, b and n becomes therefore

$$P(f,b,n) = B(f,b;n,p,q)\cdot P(n;\nu) = \left(\frac{n!}{f!b!}p^f q^b\right)\cdot\left(\frac{1}{n!}\nu^n e^{-\nu}\right).$$

Using the facts that $p+q=1$ and $f+b=n$ this can be written

$$P(f,b,n) = \left(\frac{1}{f!}(\nu p)^f e^{-\nu p}\right)\cdot\left(\frac{1}{b!}(\nu q)^b e^{-\nu q}\right),$$

which is nothing but a product of two Poisson probabilities for the variables f and b with mean values νp and νq, respectively,

$$P(f,b,n) = P(f;\nu p)\cdot P(b;\nu q). \quad (4.34)$$

The explicit n-dependence therefore drops out on the right-hand side.

The last formula could of course have been written down immediately if we had assumed f and b to be two independent Poisson variables.

It will be seen that the present and the preceding section are mathematically equivalent, in the sense that they involve the same Poisson and binomial factors, but in different order.

Exercise 4.19: An experiment is made to measure the Goldhaber asymmetry coefficient

$$\gamma \equiv \frac{f - b}{f + b} \equiv \frac{2f}{n} - 1 .$$

(i) Show that, if f and b are considered two independent Poisson variables, the variance of γ is approximately

$$V(\gamma) \simeq 4fb/(f+b)^3 .$$

(Hint: Use the law of error propagation, eq.(3.77).)

(ii) Show that, if n is considered fixed, with f and b conditioned in a binomial law with constants n,p,q, the variance is given by the exact expression

$$V(\gamma) = 4pq/n ,$$

which coincides with the result from (i), provided p and q are replaced by their estimated values, $p=\hat{p}=f/n$, $q=\hat{q}=b/n$.

4.4.4 Relation between multinomial and Poisson distributions

Example: Histogramming events (3)

Suppose that k variables $r_1, r_2, \ldots, r_k$ are dependently distributed according to the multinomial distribution law, eq.(4.16), in such a way that their sum n is a Poisson variable with mean value ν. The joint distribution is then equal to the product of the multinomial and the Poisson probabilities,

$$P(r_1, r_2, \ldots, r_k, n) = M(\underline{r}; n, \underline{p}) \cdot P(n;\nu)$$
$$= \left(\frac{n!}{r_1! r_2! \ldots r_k!} p_1^{r_1} p_2^{r_2} \cdots p_k^{r_k} \right) \cdot \left(\frac{1}{n!} \nu^n e^{-\nu} \right) .$$

Since $\sum_{i=1}^{k} p_i = 1$, $\sum_{i=1}^{k} r_i = n$, this can be organized to give

$$P(r_1, r_2, \ldots r_k, n) = \left(\frac{1}{r_1!} (\nu p_1)^{r_1} e^{-\nu p_1} \right) \left(\frac{1}{r_2!} (\nu p_2)^{r_2} e^{-\nu p_2} \right) \cdots \left(\frac{1}{r_k!} (\nu p_k)^{r_k} e^{-\nu p_k} \right) . \quad (4.35)$$

Thus the joint distribution is equal to a product of k Poisson distributions, the i-th variable r_i having the mean value (νp_i).

This result can be applied to considerations over the contents of a histogram with k bins. If the total number of entries n is a Poisson variable, it is appropriate to regard the numbers of events in each bin, $r_i, i=1,2,\ldots,k$ as independent Poisson variables. Thus, for each bin, $E(r_i)=V(r_i)=r_i$, as stated in Sect. 4.2.2.

4.4.5 The compound Poisson distribution

Example: Droplet formation along tracks in cloud chamber

Let $r_1, r_2, \ldots, r_n$ be a set of n independent Poisson variables with common mean value μ, and let also n be Poisson distributed with mean value ν. We want to find the distribution of the sum $r=\sum_{i=1}^{n} r_i$.

From the definition of marginal probability the probability P(r) we seek will be the sum of all probabilities that produce exactly r "events", hence

$$P(r) = \sum_{n=0}^{\infty} P(r=\sum_{i=1}^{n} r_i ; \mu) \cdot P(n;\nu) .$$

Here $P(n;\nu)$ is the Poisson distribution for the variable n, and $P(r=\sum_{i=1}^{n} r_i;\mu)$ is the joint probability distribution for the n constituent variables r_i, which according to the addition theorem for Poisson variables (see Sect.4.3.3 and Exercise 4.18) is the Poisson distribution $P(r;n\mu)$. Hence

$$P(r) = \sum_{n=0}^{\infty} \left\{ \left(\frac{1}{r!} (n\mu)^r e^{-n\mu} \right) \left(\frac{1}{n!} \nu^n e^{-\nu} \right) \right\} , \qquad r=0,1,\ldots \quad (4.36)$$

This is the *compound Poisson distribution*.

The compound Poisson distribution has the probability generating function

$$G(z) = \exp\left(\nu e^{\mu(z-1)} - \nu \right) , \quad (4.37)$$

from which all moments can be derived. Specifically, the mean and variance come out as

$$E(r) = \mu\nu , \qquad V(r) = \mu(\mu+1)\nu . \quad (4.38)$$

The compound Poisson distribution is applicable whenever a random process of the Poisson type initiates another. In nature, one can find many examples on chained reactions, where the products from one type of reaction gives rise to a second generation of random events. For instance, it has been suggested (by R.K. Adair and H. Kasha) that the formation of droplets along the tracks of charged particles in a cloud chamber provides an example on chained Poisson processes and should be described by the compound Poisson distribution rather than by the simple Poisson law. The physical argument is that the charged particle on its passage through matter will experience a series of elementary scatterings, the number of which, over fixed lengths, will be Poisson distributed, and further the number of droplets produced in each scattering event is also adequately described by a Poisson law. The number of droplets r over fixed lengths will then be given by eq.(4.36), where μ gives the mean number of droplets per elementary scattering and ν the mean number of scatterings over the fixed length.

Exercise 4.20: In the example in the text, let L denote the fixed length over which the droplets are counted, and put $\nu=\lambda L$ where λ is the average number of scatterings per unit length. Writing the probability generating function in the form $G(z,L)=\exp(\lambda L e^{\mu(z-1)}-\lambda L)$, show that, with $L=\ell_1+\ell_2$,

$$G(z,\ell_1+\ell_2) = G(z,\ell_1)\cdot G(z,\ell_2).$$

This factorization property implies that the number of droplets over each of the two non-overlapping lengths ℓ_1 and ℓ_2 will also be distributed according to the compound Poisson law. The result can obviously be generalized to any number of non-overlapping lengths.

Exercise 4.21: (The general compound Poisson distribution)
(i) The conditions of the previous section can be generalized to situations where the second generation of the branching process is describable by any (*i.e.* not necessarily Poisson) distribution. Specifically, let the number of primaries n be Poisson distributed (as before) with mean value $\nu=\lambda t$. Each primary gives rise to a number of secondaries r_i which are distributed with a common mean value μ. Show that the distribution of the total number of secondaries $r=\sum_{i=1}^{n} r_i$ is given by the *general compound Poisson distribution*

$$P(r) = \sum_{n=0}^{\infty} P\Big(r=\sum_{i=1}^{n} r_i;\mu\Big)\Big(\frac{1}{n!}(\lambda t)^n e^{-\lambda t}\Big),$$

where $P\big(r=\sum_{i=1}^{n} r_i;\mu\big)$ is the joint distribution of the r_i, given n.

(ii) If the constituent variables r_i have the probability generating function g(z), show that the probability generating function of r is given by

$$G(z,t) = \exp\big(\lambda t g(z) - \lambda t\big).$$

(iii) Show that the probability generating function has the factorization property

$$G(z,t_1+t_2) = G(z,t_1)\,G(z,t_2)$$

and interpret the result. Compare Exercise 4.20.

4.5 THE UNIFORM DISTRIBUTION

4.5.1 The uniform p.d.f.

In passing from discrete to continuous random variables the simplest situation one can think of is that there is a single variable x for which the probability density is constant over the region where x is defined. We write

$$f(x) = \frac{1}{b-a}, \qquad a \le x \le b, \tag{4.39}$$

which gives the *uniform* probability density function.

The expectation and variance of x with the uniform p.d.f. become, from eqs.(3.8) and (3.9),

$$E(x) = \int_a^b x f(x)\,dx = \tfrac{1}{2}(a+b), \tag{4.40}$$

$$V(x) = \int_a^b \big(x-E(x)\big)^2 f(x)\,dx = \frac{1}{12}(b-a)^2. \tag{4.41}$$

Since f(x) is symmetric about its mean all odd central moments vanish; the even central moments can be found by trivial integration,

$$\mu_{2k} = \int_a^b \big(x-E(x)\big)^{2k} f(x)\,dx = \frac{1}{(2k+1)}\Big(\frac{b-a}{2}\Big)^{2k}, \qquad k=0,1,\ldots \tag{4.42}$$

The cumulative distribution for the uniform p.d.f. is

$$F(x) = \int_a^x f(x')\,dx' = \frac{x-a}{b-a}, \qquad a \le x \le b. \tag{4.43}$$

Fig. 4.4 illustrates f(x) and F(x) for the uniform distribution.

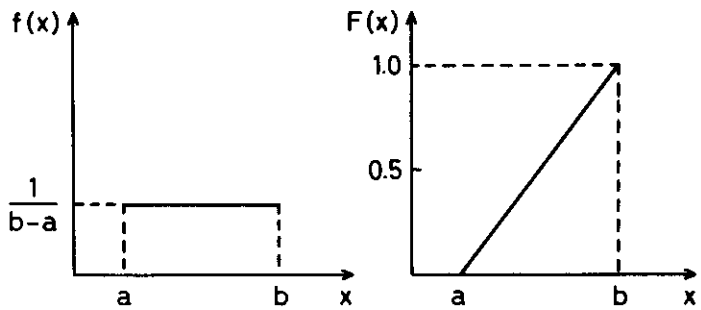

Fig. 4.4. The p.d.f. f(x) and the cumulative distribution F(x) for a uniform distribution between x = a and x = b.

The uniform distribution, although extremely simple, is nevertheless very useful; for example, any distribution of a continuous variable can be transformed into a uniform distribution.

To see this, recall the definition of the cumulative distribution function F(x) for an arbitrary continuous p.d.f. f(x),

$$F(x) = \int_{-\infty}^{x} f(x')dx' .$$

If we make the substitution

$$u = F(x),$$

the new variable u will cover the region $0 \leq u \leq 1$ and be uniformly distributed, because from eq.(3.60) its p.d.f. becomes

$$g(u) = f(x)\left|\frac{dx}{du}\right| = f(x)\left|\left(\frac{dF}{dx}\right)^{-1}\right| = 1 .$$

An example on the usefulness of this transformation is given in Sect.10.4.4.

Exercise 4.22: Show that the skewness and kurtosis coefficients for the uniform distribution are $\gamma_1=0$ and $\gamma_2=-1.2$, respectively.

Exercise 4.23: Show that monoenergetic pions which decay in flight, $\pi \rightarrow \mu + \nu$, produce neutrinos with a uniform energy distribution in the laboratory system. (Hint: The decay is isotropic in the pion rest system.)

Exercise 4.24: Let x be a continuous random variable which is uniformly distributed between 0 and 1. Show that the variable $u=-2\ln x$ has the p.d.f. $g(u)=\frac{1}{2}\exp(-\frac{1}{2}u)$ (the chi-square distribution with 2 degrees of freedom).

4.5.2 Example: Uniform random number generators

As an example of an approximative uniform distribution one can consider the frequency distribution of a sample of numbers obtained from a random number generator.

Truly random numbers can be constructed by rolling dice, dealing cards or be generated by special mechanical machines. Such methods are, however, slow and only of limited practical use.

Large samples of numbers "chosen at random" are frequently needed, for example in Monte Carlo simulations and integrations. Many algorithms have been invented to produce them by computers. These algorithms give prescriptions on how to derive sequences of numbers $x_1, x_2, x_3, \ldots$ which are "evenly" and "randomly" distributed in a given interval. The numbers are obtained by a recurring arithmetic process,

$$x_{i+1} = g(x_i, x_{i-1}, \ldots, x_{i-k}),$$

where g is some generating function and k is usually 1 or 2. Each number x_{i+1} in the sequence will therefore be completely determined by its predecessors and the given starting value(s). Thus the sequence is not random, but it will appear to be so for most practical applications. The sequence will always be periodic, with a cycle of numbers which is repeated endlessly. The length of the periodic cycle is determined by the chosen algorithm, and has an upper limit implied by the computer word length*).

Random numbers generated by recursion formulae are in the technical literature called *pseudo-random* or *quasi-random*.

---

*) A specific algorithm producing sequences of period m is given by the *linear congruential method* as $x_{i+1}=(ax_i+c) \bmod m$, where the constants have been properly adjusted and the user only selects the starting value x_o.

Exercise 4.25: A random number generator produces numbers x_i which are uniformly distributed between 0 and 1. We seek a generator for a new random number y defined over the interval [A,B], which corresponds to the probability density f(y). Defining a function $y_i=y(x_i)$ by

$$x_i = \int_A^{y(x_i)} f(t)dt ,$$

show that y_i has the desired distribution. Specifically, show that the exponential distribution $f(y)=\exp(-y)$ for $0\le y\le\infty$ can be simulated by taking

$$y_i = - \ln(1-x_i) .$$

4.6 THE EXPONENTIAL DISTRIBUTION

4.6.1 Definition and properties

The *exponential* p.d.f. is defined as

$$f(x;\beta) = \frac{1}{\beta} e^{-x/\beta}, \qquad 0 \le x \le \infty, \tag{4.44}$$

for positive values of the scale parameter β. The mean and variance for this distribution are given by, respectively,

$$E(x) = \beta , \tag{4.45}$$

$$V(x) = \beta^2. \tag{4.46}$$

The asymmetry and kurtosis coefficients are constants, and independent of β,

$$\gamma_1 = 2, \qquad \gamma_2 = 6 . \tag{4.47}$$

The exponential distribution describes a variety of physical phenomena. It can easily be derived from the Poisson assumptions for individual "random" events, as we shall see in the next section. Mathematically, the exponential distribution is a special case of the more general gamma distribution discussed in Sect.4.7.

Exercise 4.26: Show that the algebraic moments of the exponential distribution are given by

$$\mu_k = E(k-\beta)^k = k!\beta^k \sum_{r=0}^{k} (-1)^{k-r} \frac{1}{(k-r)!} .$$

4.6.2 Derivation of the exponential p.d.f. from the Poisson assumptions

Let us go back to our earlier example (Sect.4.3.2) with bubbles along the track of a charged particle in a bubble chamber, with the constant g giving the number of (point-like) bubbles per unit length of the track.

We want now to find an expression for the probability that the first bubble on a track occurs at a distance ℓ from the chosen origin. Since "the first bubble in the interval $[\ell,\ell+\Delta\ell]$" is equivalent to having no bubble in $[0,\ell]$ and one bubble in the non-overlapping interval $[\ell,\ell+\Delta\ell]$ the joint probability for the occurrence of these two independent "events" is the product of the probabilities for the individual events. Now we know that the probability for no bubble over the length ℓ is $e^{-g\ell}$ (from eq.(4.28), or the Poisson formula with r=0) and that the probability for one bubble in $\Delta\ell$ is $g\Delta\ell$ (the Poisson assumptions). The probability for the simultaneous events is therefore $(e^{-g\ell})\cdot(g\Delta\ell)$. The probability density, *i.e.* the probability per unit length, for finding the first bubble at the position ℓ becomes

$$f(\ell;g) = ge^{-g\ell}, \qquad 0 \le \ell \le \infty . \tag{4.48}$$

The distribution law (4.48) can also be interpreted as the p.d.f. for the distance ℓ between two consecutive bubbles on the track. Hence the probability for finding intervals of length $\le\ell$ between two adjacent bubbles is given by the cumulative integral

$$F(\ell) = \int_0^{\ell} f(\ell';g)d\ell' = 1 - e^{-g\ell}. \tag{4.49}$$

The probability for intervals $>\ell$ is

$$1 - F(\ell) = e^{-g\ell}. \tag{4.50}$$

The mean size of the intervals between two consecutive bubbles is, of course, $E(\ell)=g^{-1}$.

If the bubbles are of non-negligible size the formula (4.48) still gives the distribution of the distance between adjacent bubble centres. Assuming the bubbles to be of equal size with diameter d the distribution of the gap lengths $x=\ell-d$ between adjacent bubbles can also be expressed by the exponential

law, as

$$f(x;g) = ge^{-gx}, \qquad 0 \le x \le \infty . \tag{4.51}$$

The parameter g will then have the interpretation of the inverse of the mean gap length.

Exercise 4.27: For processes occurring randomly in time the probability density for the events, decays, arrivals of particles into a detector, *etc.*, is expressed as

$$f(t;\tau) = \frac{1}{\tau} e^{-t/\tau}, \qquad 0 \le t \le \infty ,$$

where the parameter τ measures the mean lifetime, or the average time between two consecutive events. Equivalently, in terms of the decay constant $\lambda = 1/\tau$ the p.d.f. is

$$f(t;\lambda) = \lambda e^{-\lambda t}, \qquad 0 \le t \le \infty .$$

Review the considerations of the last section in this context. In particular, find arguments for the common statement that "the exponential distribution has no memory". This property is very useful in practice since it, for example, allows one to measure the lifetimes of unstable particles starting from an arbitrary time t_o after the time (t=0) they were created. See also Exercise 4.28 below.

Exercise 4.28: A neutrino beam is produced by decays of high-energy pions, $\pi \to \mu+\nu$. If the pions have momentum p_π and the available decay region is of length ℓ, what fraction of the pions will produce neutrinos? As a numerical example, take p_π=10.0 GeV/c, ℓ=80 m, m_π=0.1396 GeV/c^2, $c\tau$=7.8 m. (Answer: 13.3%.)

Exercise 4.29: Decays from a radioactive source of decay constant λ are registered by a Geiger counter of resolution time t_o (*i.e.* a detector which fails in recording an event if it occurs separated in time from the previous event by an amount less than t_o). If N' counts are registered over the time t, show that the number of decays N that have actually taken place in the radioactive source is given by

$$N = \frac{1 - e^{-\lambda t}}{e^{-\lambda t_o} - e^{-\lambda t}} N' \equiv aN'$$

where $a > 1$.

Assume that the registered number of counts is a Poisson variable of estimated mean value $\hat{\mu}$=N'. Show that the estimated variance in the true number of decays is $\hat{V}(N) = aN > N$, whereas with a perfect detector, $t_o \ll 1/\lambda$, the estimated variance would have been $\hat{V}(N) = N$.

Exercise 4.30: (The hyperexponential distribution)
A superposition of exponential probability distributions goes under the name of the *hyperexponential distribution*. In terms of a time variable t the p.d.f. is

$$f(t;\underline{p},\underline{\lambda}) = \sum_{i=1}^{k} p_i \lambda_i \exp(-\lambda_i t), \qquad 0 \le t \le \infty ,$$

where p_i denotes the proportion of the i-th process ($\sum_{i=1}^{k} p_i = 1$), for which the decay constant is λ_i. Discuss the properties of this p.d.f.

This distribution describes the effect of exponentials operating in parallel. When exponential distributions of similar type operate in series, the resulting distribution is an Erlangian distribution; see Sect.4.7.2.

4.7 THE GAMMA DISTRIBUTION

4.7.1 Definition and properties

With α and β as two positive constants we define the *gamma distribution* by

$$f(x;\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}, \qquad 0 \le x \le \infty . \tag{4.52}$$

This function is seen to be properly normalized, in virtue of the gamma function,

$$\Gamma(\alpha) \equiv \int_0^\infty y^{\alpha-1} e^{-y} dy \qquad (\alpha > 0) \tag{4.53}$$

which has the property

$$\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1). \tag{4.54}$$

The formula (4.52) produces a variety of shapes for different values of the constant α, as indicated in Fig. 4.5. For $\alpha \le 1$ the distribution is J-shaped, while $\alpha > 1$ gives a unimodal distribution with maximum at $x=(\alpha-1)\beta$. The parameter β is only a scale factor.

When α is an integer, α=k, $\Gamma(k)=(k-1)!$ and the p.d.f. (4.52) is called the *Erlangian distribution*; this distribution law can be derived from first principles and is discussed in the following section. The special case α=1 corresponds to the exponential distribution, which was treated already in Sect.4.6. The special case β=2, $\alpha = \frac{\nu}{2}$ where ν is an integer is equivalent to the chi-square distribution with ν degrees of freedom; this important distribution is discussed quite extensively in Chapter 5.

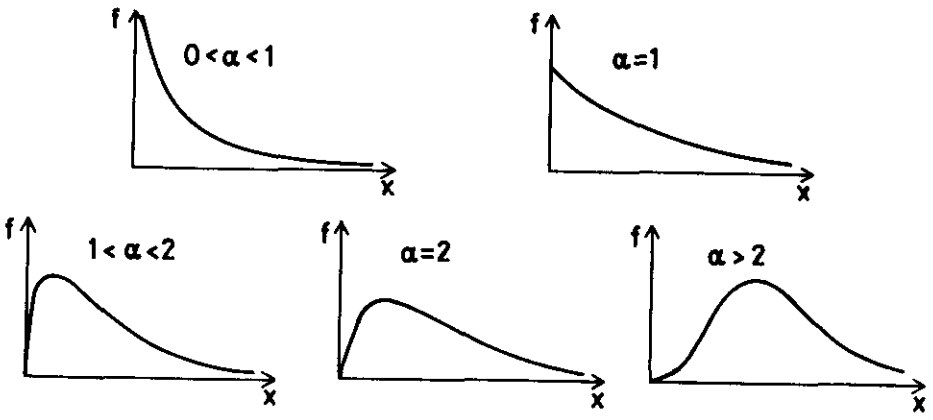

Fig. 4.5. Shapes of the gamma distribution for different parameter values.

Exercise 4.31: Show that the mean and variance of the gamma distribution are given by, respectively,

$$E(x) = \alpha\beta, \qquad V(x) = \alpha\beta^2.$$

Exercise 4.32: Show that the gamma distribution has the characteristic function

$$\Phi(t) = (1 - \beta it)^{-\alpha}.$$

Exercise 4.33: Show that the gamma distribution has asymmetry and kurtosis coefficients given by, respectively,

$$\gamma_1 = 2/\sqrt{\alpha}, \qquad \gamma_2 = 6/\alpha.$$

Exercise 4.34: (The beta distribution)
While the gamma distribution describes variables which are bounded at one side, the *beta distribution*

$$f(x;\mu,\nu) = \frac{\Gamma(\mu+\nu)}{\Gamma(\mu)\Gamma(\nu)} x^{\mu-1}(1-x)^{\nu-1} \qquad 0 \le x \le 1$$

can be used to describe variables which are limited on two sides. The parameters μ and ν are both positive integers; the quantity $B(\mu,\nu) \equiv \Gamma(\mu)\Gamma(\nu)/\Gamma(\mu+\nu)$ is often called the beta function.

(i) Show that the mean and variance of the beta distribution are, respectively,

$$E(x) = \frac{\mu}{\mu+\nu}, \qquad V(x) = \frac{\mu\nu}{(\mu+\nu)^2(\mu+\nu+1)}.$$

(ii) Show that the beta distribution is symmetric about $x=\frac{1}{2}$ when $\mu=\nu$, and reduces to the uniform p.d.f. for $\mu=\nu=1$.

(iii) Sketch the beta distribution for the lowest combinations of the parameters.

4.7.2 Derivation of the gamma p.d.f. from the Poisson assumptions

From a physicist's point of view the importance of the gamma distribution lies mainly in the fact that it, for integer values of the parameter α, can be derived from the Poisson assumptions, and consequently describes random processes of the Poisson type.

For definiteness, let us take the random variable to describe a time interval. If λ is the constant average number of events (decays, accidents, *etc.*) per unit time, the Poisson prediction for the number of events r in the time t is (eq.(4.30))

$$P_r(t;\lambda) = \frac{1}{r!}(\lambda t)^r e^{-\lambda t}.$$

We want to find the distribution law for the time t at which the k-th event occurs. Since the probability for $0,1,\ldots,k-1$ events in the time t is

$$\sum_{r=0}^{k-1} P_r(t;\lambda) = \sum_{r=0}^{k-1} \frac{(\lambda t)^r e^{-\lambda t}}{r!}$$

the probability that there are at least k events in the time t is

$$F_k(t;\lambda) = 1 - \sum_{r=0}^{k-1} \frac{(\lambda t)^r e^{-\lambda t}}{r!}. \tag{4.55}$$

It can be shown by mathematical induction that the sum in eq.(4.55) can be written as an integral,

$$\sum_{r=0}^{k-1} \frac{(\lambda t)^r e^{-\lambda t}}{r!} = \int_{\lambda t}^{\infty} \frac{z^{k-1} e^{-z}}{(k-1)!}\, dz.$$

Hence

$$F_k(t;\lambda) = 1 - \int_{\lambda t}^{\infty} \frac{z^{k-1} e^{-z}}{(k-1)!}\, dz = \int_{0}^{\lambda t} \frac{z^{k-1} e^{-z}}{(k-1)!}\, dz.$$

Replacing z by λz leads to

$$F_k(t;\lambda) = \int_0^t \frac{\lambda^k z^{k-1} e^{-\lambda z}}{(k-1)!}\, dz. \tag{4.56}$$

This is the cumulative integral of a probability density which is seen to be of the form of eq.(4.52), with the parameters $\alpha=k$ (integer), $\beta=\frac{1}{\lambda}$, *i.e.* an Erlangian p.d.f.; we write

$$f(t;k,\lambda) = \frac{\lambda^k}{(k-1)!}\, t^{k-1} e^{-\lambda t}, \qquad 0 \le t \le \infty. \tag{4.57}$$

Since the cumulative distribution of eq.(4.56) gives the probability that there are at least k events in the time t the p.d.f. $f(t;k,\lambda)$ will give the probability density in t for the occurrence of the k-th event, when the events occur independently and at an average rate of λ per unit time. We observe in particular that the p.d.f. describing the occurrence of the first event (a decay, say) is $f(t;1,\lambda)=\lambda e^{-\lambda t}$, as it should be. The formula (4.57) can also be interpreted as giving the distribution for the time elapsed between (k+1) consecutive events, corresponding to k "time gaps". In accordance with this interpretation the expectation value for t is

$$E(t) = \frac{k}{\lambda} = k\tau \tag{4.58}$$

in other words, the mean value of t is k times larger than the average length of the time gap between two consecutive events.

To summarize, the Erlangian p.d.f. describes the (time) distribution of exponentially distributed events occurring in series. (When exponentials operate in parallel the result is the hyperexponential distribution, Exercise 4.30.)

Exercise 4.35: Show that the mean value for the p.d.f. (4.57) is $E(t)=k/\lambda$ (eq. 4.58) and that the variance is $V(t)=k/\lambda^2$.

Exercise 4.36: (Connection between the Erlangian, Poisson, and negative binomial distributions)

In the previous sections it has frequently been emphasized that the physical assumption underlying the Poisson distribution law is that the average event rate is constant. Suppose now that this condition is not fulfilled, and that the average μ is described by an Erlangian formula

$$f(\mu;k,\lambda) = \frac{\lambda^k}{(k-1)!}\, \mu^{k-1} e^{-\lambda\mu} \qquad 0 \le \mu \le \infty$$

Given μ a discrete variable s is supposed to be Poisson distributed with mean value μ. Show that the resulting probability distribution for s is

$$P_k(s;\lambda) = \int_0^\infty d\mu f(\mu;k,\lambda)\cdot P(s;\mu) = \binom{s+k-1}{s}\left(\frac{\lambda}{\lambda+1}\right)^k\left(1-\frac{\lambda}{\lambda+1}\right)^s$$

This is recognized as a negative binomial distribution with parameter $p=\lambda/(\lambda+1)$ (Exercise 4.5) for which the mean value and variance are $E(s)=k/\lambda$ and $V(s)=k/\lambda+k/\lambda^2$, respectively.

4.7.3 Example: On-line processing of batched events

As an application of the previous arguments, let us consider a queueing problem involving the collecting and processing of data in an electronic experiment. The direct measurements from a number of sequential events are stored in an intermediate buffer and subsequently transferred to a central computing unit for processing. If the buffer runs full before the computer is ready with the previous batch of events the transfer of the data is prohibited; the buffer is then simply reset to zero and the data collecting is started anew. How large a fraction of the events will in the long run get lost with this way of operating the system?

The essential quantities in this problem are the event rate, the buffer capacity and the computing speed. We will assume that the events occur independently in time, at an average rate of λ events per second. The buffer has a capacity to store k events, and the computer processes the k events in T seconds. We assume that the transfer time between the buffer and the central memory is negligible and that the time required for resetting the buffer before a new collecting period can also be ignored.

With this formulation we recognize that the probability density function for the time t of a full buffer is given by the Erlangian formula (4.57). The fraction F(T) of rejected events then corresponds to the probability of getting the k-th event before the time T, when the computer is still working with the previous batch of events. Thus,

$$F(T) = \int_0^T f(t;k,\lambda)dt = \int_0^T \frac{\lambda^k t^{k-1}}{(k-1)!}\, e^{-\lambda t} dt.$$

For a numerical example, suppose that the average event rate corresponds to $\lambda=0.5\ s^{-1}$, the buffer capacity k=10, and the computer speed such that one event is processed per second, or T=10 s for one batch of 10 events. With this choice of constants the Erlangian p.d.f. $f(t;k=10,\lambda=\frac{1}{2})$ is identical to a chi-square p.d.f. with $\nu=2k=20$ degrees of freedom. The cumulative chi-square distribution of Fig. 5.2 from Chapter 5 can then be used to read off the value of F(T); we find

$$F(T) = \int_0^{10} f(u;\nu=20)du \simeq 0.03.$$

Thus only about 3% of the events will not be used in this case.

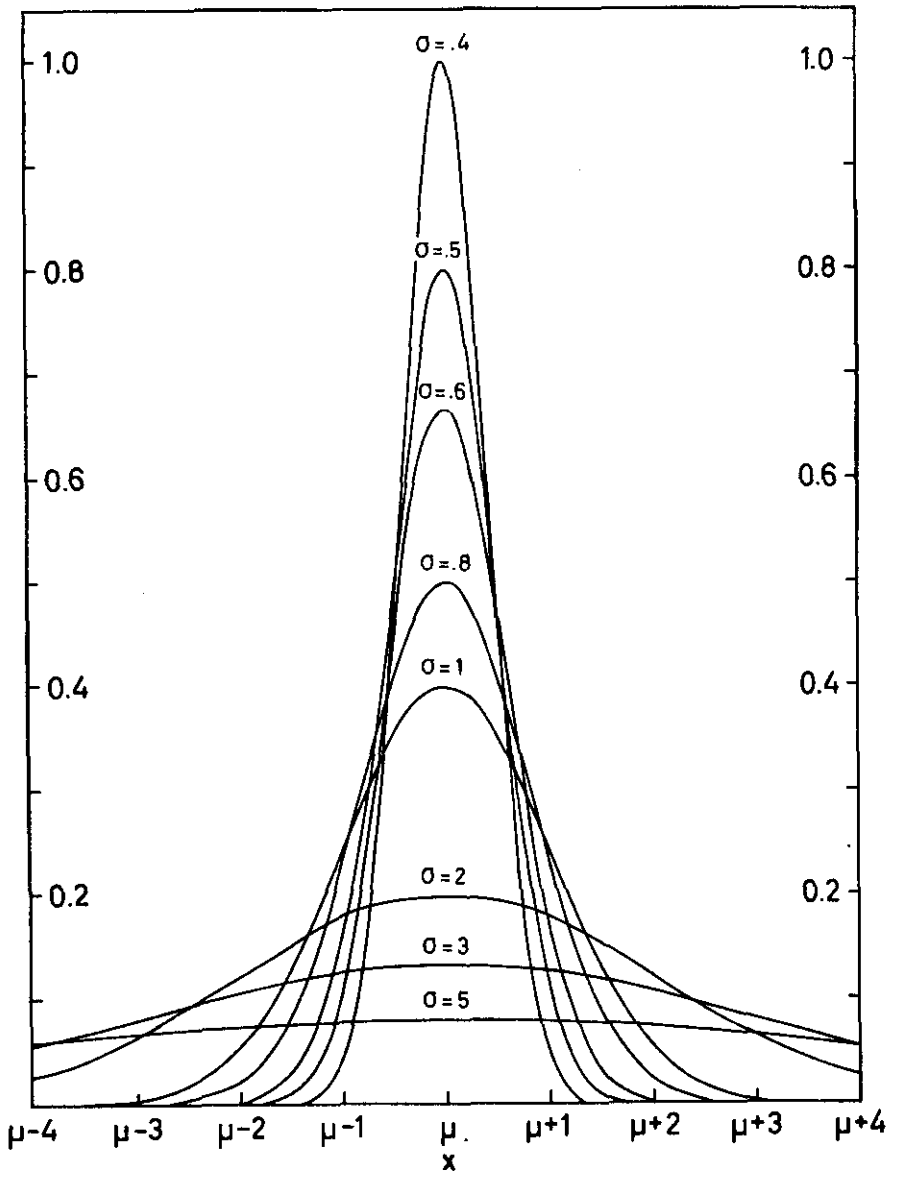

Fig. 4.6. The normal p.d.f. $N(\mu,\sigma^2)$ for different values of the standard deviation σ.

4.8 THE NORMAL, OR GAUSSIAN, DISTRIBUTION

We discuss next the *normal*, or *Gaussian*, probability density function which plays a fundamental role in probability theory and statistics. We consider it first as a function of only one variable; the extension to two and more variables is made in Sects.4.9 and 4.10, respectively.

4.8.1 Definition and properties of $N(\mu,\sigma^2)$

The normal p.d.f. in one dimension has the general form

$$N(\mu,\sigma^2) \equiv f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}(x-\mu)^2/\sigma^2}, \qquad -\infty \le x \le \infty . \tag{4.59}$$

The expectation and the variance of x with this distribution can be found by trivial integration from the definitions eqs.(3.8) and (3.9),

$$E(x) = \int_{-\infty}^{\infty} xf(x)dx = \mu, \tag{4.60}$$

$$V(x) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx = \sigma^2. \tag{4.61}$$

Hence the parameters μ and σ^2 of $N(\mu,\sigma^2)$ have the usual meaning of the mean value and variance (or dispersion) of a distribution.

The normal p.d.f. $N(\mu,\sigma^2)$ is symmetric about $x=\mu$, and hence the median coincides with the mean. The p.d.f. also has its mode (maximum) at $x=\mu$, while the two points of inflection occur at distances $\pm\sigma$ from this value. Figure 4.6 shows different normal distributions with a common mean.

Exercise 4.37: Show that the half-width of $N(\mu,\sigma^2)$ at half-maximum is equal to $\sqrt{2\ln 2}\sigma \approx 1.18\sigma$.

Exercise 4.38: Show that if x is $N(\mu,\sigma^2)$, then ax is $N(a\mu,a^2\sigma^2)$.

4.8.2 The standard normal distribution N(0,1)

The general normal distribution $N(\mu,\sigma^2)$ can be transformed to a convenient form by the *standard transformation*

$$y = \frac{x-\mu}{\sigma}, \tag{4.62}$$

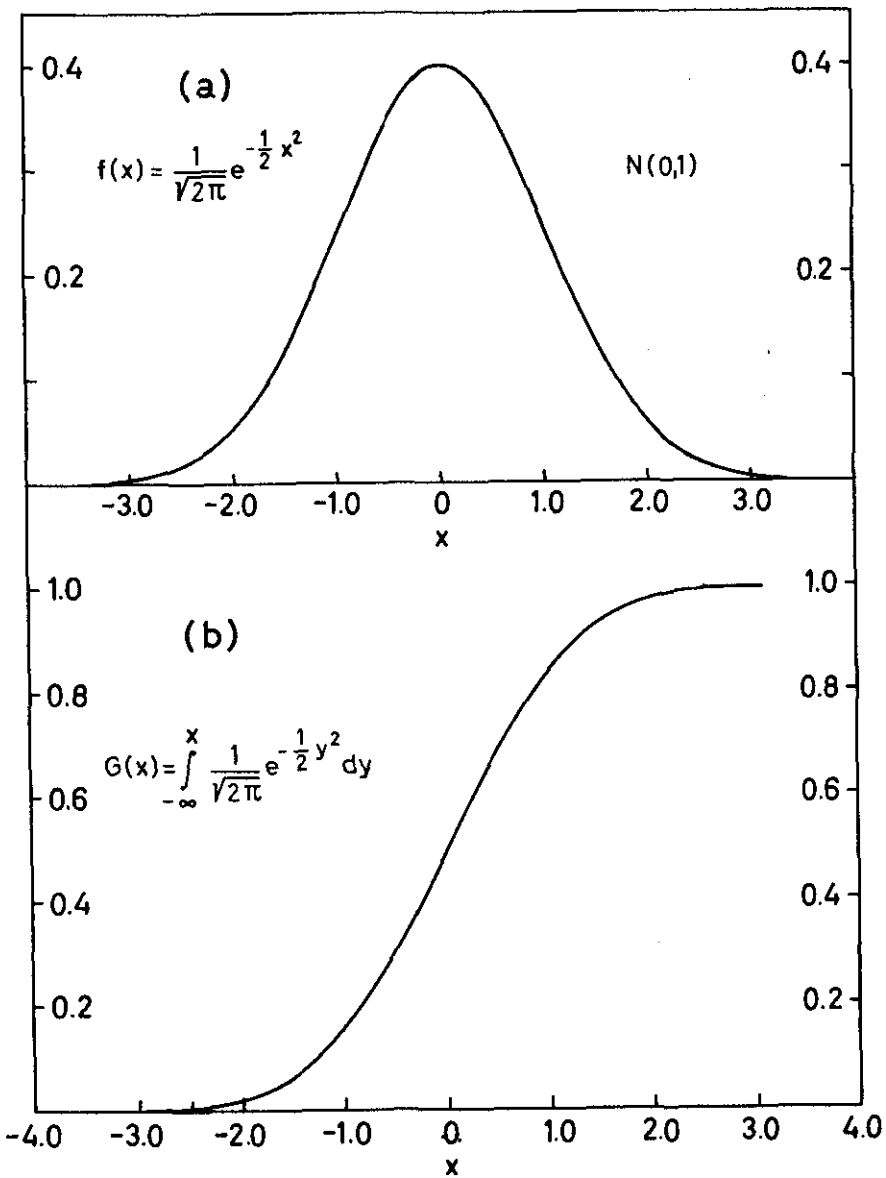

Fig. 4.7. (a) The standard normal p.d.f. N(0,1).
(b) The cumulative standard normal distribution.

which implies a shift of the origin to the mean value and an appropriate change of scale. This gives the *standard normal* p.d.f.

$$N(0,1) \equiv g(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}, \qquad -\infty \leq y \leq \infty, \tag{4.63}$$

which has mean value zero and standard deviation one. This p.d.f. is tabulated in Appendix Table A5.

The *cumulative standard normal distribution* is given by

$$G(y) = \int_{-\infty}^{y} g(y')dy' = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y'^2} dy', \tag{4.64}$$

and is tabulated in Appendix Table A6. This function, which in the literature is often called the standard normal distribution function (compare the footnote in Sect.3.2), has the property

$$G(-y) = 1 - G(y). \tag{4.65}$$

Figure 4.7 shows the standard normal p.d.f. and its cumulative distribution.

Exercise 4.39: Verify that if x is $N(\mu,\sigma^2)$ the variable $y=(x-\mu)/\sigma$ is N(0,1), (eq.(4.63)). Show that $u=y^2$ has the p.d.f. $f(u)= (2\pi)^{-1}u^{-\frac{1}{2}}\exp(-\frac{1}{2}u)$, (the chi-square distribution with one degree of freedom).

4.8.3 Probability contents of $N(\mu,\sigma^2)$

The cumulative standard normal distribution G(y) is used in practice to determine the probability contents of a given interval for a normally distributed value, or *vice versa*, to determine an interval corresponding to a given probability.

Let x be a random variable which is distributed according to the p.d.f. $N(\mu,\sigma^2)$ of eq.(4.59). We want to find the probability that x falls between a lower limit a and an upper limit b. Clearly,

$$P(a \leq x \leq b) = P(x \leq b) - P(x \leq a).$$

On the right-hand-side the inequalities may be expressed in terms of the stan-

dardized variable $(x-\mu)/\sigma$, and we have

$$P(a \le x \le b) = P\left(\frac{x-\mu}{\sigma} \le \frac{b-\mu}{\sigma}\right) - P\left(\frac{x-\mu}{\sigma} \le \frac{a-\mu}{\sigma}\right)$$
$$= \int_{-\infty}^{(b-\mu)/\sigma} g(y')dy' - \int_{-\infty}^{(a-\mu)/\sigma} g(y')dy'.$$

Hence

$$P(a \le x \le b) = G\left(\frac{b-\mu}{\sigma}\right) - G\left(\frac{a-\mu}{\sigma}\right), \tag{4.66}$$

where G is the cumulative standard normal distribution of eq.(4.64). Given μ,σ^2 and the interval [a,b] the actual function values of G can be found from, for example, Fig. 4.7(b) or Appendix Table A6. Using the property of eq.(4.65) we find the probability contents corresponding to the (symmetric) one-, two-, and three standard deviation intervals about the mean μ:

$$\left.\begin{aligned} P(\mu-\sigma \le x \le \mu+\sigma) &= 2G(1) - 1 = 0.6827 \\ P(\mu-2\sigma \le x \le \mu+2\sigma) &= 2G(2) - 1 = 0.9545 \\ P(\mu-3\sigma \le x \le \mu+3\sigma) &= 2G(3) - 1 = 0.9973. \end{aligned}\right\} \tag{4.67}$$

See the illustration of Fig. 4.8.

Suppose next that we have the opposite situation and want to determine the size of an interval which is to correspond to a given probability, say 0.90. Equation (4.66) is now

$$0.90 = G\left(\frac{b-\mu}{\sigma}\right) - G\left(\frac{a-\mu}{\sigma}\right), \tag{4.68}$$

and with given parameters μ,σ there are obviously infinitely many solutions for the unknowns a,b. With the additional requirement that the interval [a,b] should be *symmetric* about μ, however, there is only one unknown, say b. With the use of eq.(4.65) we can then write eq.(4.66) as

$$0.90 = G\left(\frac{b-\mu}{\sigma}\right) - G\left(-\frac{b-\mu}{\sigma}\right) = 2G\left(\frac{b-\mu}{\sigma}\right) - 1,$$

or

$$G\left(\frac{b-\mu}{\sigma}\right) = \tfrac{1}{2}(1+0.90) = 0.95.$$

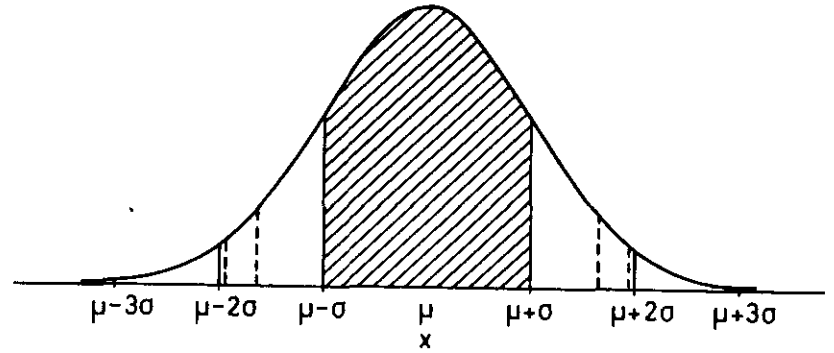

Fig. 4.8. Probability contents of the normal probability distribution $N(\mu,\sigma^2)$. The shaded area corresponds to the symmetric one-standard deviation interval about the mean value, for which the probability is $P(\mu-\sigma \le x \le \mu+\sigma) = 0.6827$; the symmetric regions covering the two- and three standard deviation intervals correspond to probabilities $P(\mu-2\sigma \le x \le \mu+2\sigma) = 0.9545$ and $P(\mu-3\sigma \le x \le \mu+3\sigma) = 0.9973$, respectively, (eq.(4.67)). - The dotted lines indicate the extension of the regions which correspond to probabilities 0.90 and 0.95, (eq.(4.69)).

Interpolation in Appendix Table A6 then gives for the value of the argument of G

$$\frac{b-\mu}{\sigma} = 1.645.$$

Thus, given $N(\mu,\sigma^2)$ the symmetric interval about μ corresponding to the probability 0.90 is

$$[\mu - 1.645\sigma,\ \mu + 1.645\sigma].$$

For any given probability γ between 0 and 1 one can proceed in a similar manner to determine symmetric intervals around μ. For the common choices $\gamma = 0.90,\ 0.95,\ 0.99,\ 0.999$ often referred to, one has

$$\left.\begin{aligned} P(\mu-1.645\sigma \le x \le \mu+1.645\sigma) &= 0.90 \\ P(\mu-1.960\sigma \le x \le \mu+1.960\sigma) &= 0.95 \\ P(\mu-2.576\sigma \le x \le \mu+2.576\sigma) &= 0.99 \\ P(\mu-3.290\sigma \le x \le \mu+3.290\sigma) &= 0.999. \end{aligned}\right\} \tag{4.69}$$

The first two of these intervals are indicated in Fig. 4.8.

For practical purposes it may be useful to observe and remember that, for a normal variable, the symmetric interval covering a probability 0.95 very nearly coincides with the two-standard deviation interval.

Exercise 4.40: If x is normally distributed with mean and variance both equal to 16, what is the probability that $12 \leq x \leq 20$? Compare Exercise 4.16.

4.8.4 Central moments; the characteristic function

The central moments of the normal distribution of eq.(4.59) can be obtained from the general definition by eq.(3.14). Clearly all odd central moments vanish because of the symmetry property of the normal p.d.f. The even moments can be evaluated in a straightforward way by carrying out integrations of the type

$$\mu_{2k} = \int_{-\infty}^{\infty} (x-\mu)^{2k} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}(x-\mu)^2/\sigma^2} dx . \tag{4.70}$$

One finds

$$\mu_{2k} = \frac{(2k)!}{k!}(\tfrac{1}{2}\sigma^2)^k , \qquad k = 0,1,2,\ldots \tag{4.71}$$

Thus all central moments of the normal distribution can be expressed in terms of the variance σ^2. The lowest of the even moments are

$$\mu_2 = \sigma^2 , \qquad \mu_4 = 3\sigma^4 , \qquad \mu_6 = 15\sigma^6 . \tag{4.72}$$

The asymmetry and kurtosis coefficients are both zero,

$$\left.\begin{aligned} \gamma_1 &= \frac{\mu_3}{(\mu_2)^{3/2}} = 0 \\ \gamma_2 &= \frac{\mu_4}{(\mu_2)^2} - 3 = 0 , \end{aligned}\right\} \tag{4.73}$$

since the particular definition of the kurtosis coefficient by eq.(3.21) was made to correspond to $\gamma_2=0$ for the normal p.d.f.

From the general definition of eq.(3.22) one finds the characteristic function for the normal p.d.f. of eq.(4.59),

$$\Phi(t) \equiv E(e^{itx}) = e^{\mu it - \frac{1}{2}\sigma^2 t^2} . \tag{4.74}$$

This form of the characteristic function is used to derive many theoretically important properties of the normal distribution (see, for example, the following *sections). For the purpose of evaluating central moments of the normal p.d.f.* one may consider instead the function (eq.(3.28))

$$\Phi_\mu(t) = e^{-\mu it}\Phi(t) = e^{\frac{1}{2}\sigma^2(it)^2} ,$$

which has the series expansion

$$\Phi_\mu(t) = \sum_{r=0}^{\infty} \frac{(it)^{2r}}{r!}(\tfrac{1}{2}\sigma^2)^r . \tag{4.75}$$

According to eq.(3.27) the central moments are found by taking the derivatives of $\Phi_\mu(t)$ with respect to (it) and putting t=0,

$$\mu_k = \frac{\partial^k}{\partial(it)^k}\left[\sum_{r=0}^{\infty} \frac{(it)^{2r}}{r!}(\tfrac{1}{2}\sigma^2)^r\right]_{t=0} .$$

Since only even powers of t appear in the sum, all odd derivatives will vanish when evaluated for t=0. Thus all odd central moments will be zero and only even moments survive; it will be seen that eq.(4.71) is regained, as expected.

Exercise 4.41: Show that the central moments of the normal distribution satisfy the relation

$$\mu_{2k+2} = \sigma^2\mu_{2k} + \sigma^3 \frac{d\mu_{2k}}{d\sigma} , \qquad k=0,1,\ldots$$

Exercise 4.42: Show that the algebraic moments of the normal p.d.f. are connected through the relation

$$\mu'_{k+2} = 2\mu\mu'_{k+1} + (\sigma^2-\mu^2)\mu'_k + \sigma^3 \frac{d\mu'_k}{d\sigma} , \qquad k=0,1,\ldots$$

4.8.5 Addition theorem for normally distributed variables

It frequently happens that one wants to know the distribution of a variable which is a function of normally distributed variables. In particular, one can have a *linear* function of normal variables. It turns out that a random

variable which is a linear combination of independent, normally distributed random variables is itself normally distributed with mean and variance equal to the sums of, respectively, the means and variances of the constituent variables.

To see this, let x_1 and x_2 be two independent variables with normal distributions $N(\mu_1,\sigma_1^2)$ and $N(\mu_2,\sigma_2^2)$, respectively. Then, with a_1 and a_2 as constants, we know that a_1x_1 and a_2x_2 are independent, normal variables distributed as, respectively, $N(a_1\mu_1,a_1^2\sigma_1^2)$ and $N(a_2\mu_2,a_2^2\sigma_2^2)$ (compare Exercise 4.38). According to eq.(4.74) their characteristic functions are

$$\Phi_{a_1x_1}(t) = e^{(a_1\mu_1)it-\frac{1}{2}(a_1^2\sigma_1^2)t^2}$$

and

$$\Phi_{a_2x_2}(t) = e^{(a_2\mu_2)it-\frac{1}{2}(a_2^2\sigma_2^2)t^2}.$$

We consider the sum

$$y = a_1x_1 + a_2x_2.$$

Because of the independence of the two terms the characteristic function for y is equal to the product of the characteristic functions for a_1x_1 and a_2x_2 (this is the result of Sect.3.5.7). From eq.(3.51), therefore

$$\Phi_y(t) = \Phi_{a_1x_1}(t)\cdot\Phi_{a_2x_2}(t) = e^{(a_1\mu_1+a_2\mu_2)it-\frac{1}{2}(a_1^2\sigma_1^2+a_2^2\sigma_2^2)t}$$

But this is nothing but a new characteristic function of the form of eq.(4.74); hence y is distributed as $N(a_1\mu_1+a_2\mu_2,a_1^2\sigma_1^2+a_2^2\sigma_2^2)$.

The proof can clearly be extended to any number of independent, normal variables. So we may formulate *the addition theorem* for normally distributed variables as follows:

> Let $x_1,x_2,\ldots,x_n$ be independent, normally distributed variables such that x_i is $N(\mu_i,\sigma_i^2)$. Then the linear combination $y=\sum_{i=1}^{n}a_ix_i$ is also a normally distributed variable, with mean $\sum_{i=1}^{n}a_i\mu_i$ and variance $\sum_{i=1}^{n}a_i^2\sigma_i^2$.

An application of this theorem is given in the following section.

4.8.6 Properties of $\bar{x}$ and s^2 for sample from $N(\mu,\sigma^2)$

We assume that n independent random variables are all normally distributed with mean μ and variance σ^2. The variables $x_1,x_2,\ldots,x_n$ are said to be a random sample of size n drawn from the normal population (or universe) $N(\mu,\sigma^2)$. The sample has a mean and a variance given by the general formulae eqs.(3.87) - (3.88),

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i, \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2.$$

From the preceding section we are now able to state the distributional properties of the sample mean $\bar{x}$. With the notation of Sect.4.8.5 $\bar{x}$ is a linear function of the independent, normal variables x_i with all coefficients equal, $a_i=\frac{1}{n}$. Hence, according to the addition theorem for normally distributed variables, $\bar{x}$ will also be a normal variable, with mean $\sum_{i=1}^{n}a_i\mu_i=\mu$ and variance $\sum_{i=1}^{n}a_i^2\sigma_i^2=\sigma^2/n$; thus

$$\bar{x} \text{ is distributed as } N\left(\mu, \frac{\sigma^2}{n}\right).$$

It will be demonstrated in Sect.5.1.6 that the sample variance s^2 for the normal sample is related to the chi-square distribution with n-1 degrees of freedom, $\chi^2(n-1)$. Specifically,

$$\frac{(n-1)s^2}{\sigma^2} \text{ is distributed as } \chi^2(n-1).$$

Furthermore, $\bar{x}$ and s^2 for the normal sample are *independent* variables (see Exercise 5.9). This is a property which is specific for the normal sample. The outstanding position of the normal distribution in this respect is explicit in the following important theorem:

> Given that the random variables $x_1,x_2,\ldots,x_n$ are independent, with *identical normal distributions, then the two variables (statistics)* $\bar{x}=\sum_{i=1}^{n}x_i/n$ and $s^2=\sum_{i=1}^{n}(x_i-\bar{x})^2/(n-1)$ are independent. Conversely, if the mean $\bar{x}$ and variance s^2 of random samples from a population are independent, that population must be normal.

4.8.7 Example: Position and width of resonance peak

To illustrate the implication of the foregoing let us consider as an example the determination of the position and width of a resonance peak in an effective mass spectrum.

For simplicity we assume that the background problem does not exist. From the observation of n events with effective-masses $M_1, M_2, \ldots, M_n$ in the appropriate region of the spectrum the true mass M_o and width Γ_o of the peak can be estimated by, respectively, the arithmetic mean of the sample values

$$\hat{M}_o = \bar{M} = \frac{1}{n} \sum_{i=1}^{n} M_i$$

and the square root of the sample variance, given by

$$\hat{\Gamma}_o^2 = s_M^2 = \frac{1}{n-1} \sum_{i=1}^{n} (M_i - \bar{M})^2 .$$

According to the statement in the preceding section the variables $\bar{M}$ and s_M^2 will be independent if, and only if, the M_i constitute a normal sample. Hence the above estimates $\hat{M}_o$ and $\hat{\Gamma}_o$ will only be independent if the resonance has a strictly normal shape. Any other form of the underlying distribution, for instance a Breit-Wigner shape, implies that the estimates of the peak mass and width will necessarily be dependent. Hence the understanding of the uncertainty in the estimates of these quantities will only be straightforward when the normal peak shape is assumed: the estimates are then uncorrelated and have errors simply given by the square root of the diagonal elements in the covariance matrix. For all other assumptions about the shape of the true resonance peak a more elaborate treatment is required to determine the correlation in the estimates of peak mass and width.

In the discussion above we have tacitly assumed an ideal situation where the measured M_i are precise values, without any experimental uncertainties attached to them. A more realistic experimental situation with a finite inherent error in any measurement can be treated as described in some detail in Sect.6.2.

4.8.8 The Central Limit Theorem

We shall next consider the *Central Limit Theorem*, probably the most important theorem in probability theory, with far-reaching theoretical and practical implications.

We learned already in Sect.3.6 that if $x_1, x_2, \ldots, x_n$ is a set of independent random variables, such that each x_i has a mean value μ_i and a variance σ_i^2, then the sum of these variables, Σx_i, will be distributed with mean and variance equal to, respectively, $\Sigma\mu_i$ and $\Sigma\sigma_i^2$. The distributions of the separate x_i were unspecified except for their means and variances, and so was also the distribution of their sum Σx_i.

The Central Limit Theorem expresses that the distributional properties of the sum Σx_i will be completely known provided that the number of terms in the sum is very large. In the limit when n goes towards infinity, $\sum_{i=1}^{n} x_i$ will be *normally* distributed, with mean value $\sum_{i=1}^{n} \mu_i$ and variance $\sum_{i=1}^{n} \sigma_i^2$.

In a condensed form we may state the Central Limit Theorem as follows:

> Let $x_1, x_2, \ldots, x_n$ be a set of n independent random variables, and each x_i be distributed with mean value μ_i and a finite variance σ_i^2. Then the variable $\left(\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \mu_i\right) \Big/ \left(\sum_{i=1}^{n} \sigma_i^2\right)^{\frac{1}{2}}$ has a limiting distribution which is N(0,1).

The remarkable thing about this powerful theorem is that the assumptions about the x_i are so unrestrictive. In fact, it is sufficient to specify that each x_i should have a finite variance, since then necessarily also the mean will be finite.

We will prove the Central Limit Theorem only for the special case when all n independent x_i have the same mean and variance. The x_i then correspond to a sample from a population with mean μ and variance $\sigma^2 < \infty$. We construct the standardized variable

$$z = \frac{\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \mu_i}{\sqrt{\sum_{i=1}^{n} \sigma_i^2}} = \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma\sqrt{n}} , \qquad (4.76)$$

for which we want to establish the characteristic function $\phi(t)$. Each of the x_i has a characteristic function $\phi(t)$ which contains the arithmetic moments of the x_i by the series expansion of eq.(3.23). Writing out the terms up to second

order moments we have

$$\phi(t) = E(e^{itx_i}) = 1 + (it)\mu + \frac{1}{2!}(it)^2(\sigma^2+\mu^2) + \ldots$$

Since all x_i - and hence the terms $(x_i-\mu)/\sigma\sqrt{n}$ in eq.(4.76) - by assumption are independent, the characteristic function for z is equal to the product of n characteristic functions for each of these terms (Sect.3.5.7). Hence

$$\Phi(t) = \left[E\left(e^{it\frac{x_i-\mu}{\sigma\sqrt{n}}}\right)\right]^n = \left[e^{-\frac{it\mu}{\sigma\sqrt{n}}} E\left(e^{i\frac{t}{\sigma\sqrt{n}}x_i}\right)\right]^n$$

or

$$\Phi(t) = \left[e^{-\frac{it\mu}{\sigma\sqrt{n}}} \phi\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n .$$

Taking the logarithm and expressing $\phi\left(\frac{t}{\sigma\sqrt{n}}\right)$ in terms of the series expansion above we get

$$\ln\phi(t) = n\left\{-\frac{it\mu}{\sigma\sqrt{n}} + \ln\left[1 + \left(\frac{it}{\sigma\sqrt{n}}\right)\mu + \frac{1}{2!}\left(\frac{it}{\sigma\sqrt{n}}\right)^2(\sigma^2+\mu^2) + \ldots\right]\right\},$$

and with the expansion $\ln(1+x)=x-\frac{1}{2}x^2+\ldots$ this can further be written

$$\ln\phi(t) = n\left\{-\frac{it\mu}{\sigma\sqrt{n}} + \left[\left(\frac{it}{\sigma\sqrt{n}}\right)\mu + \frac{1}{2!}\left(\frac{it}{\sigma\sqrt{n}}\right)^2(\sigma^2+\mu^2) - \frac{1}{2}\left(\frac{it}{\sigma\sqrt{n}}\right)^2\mu^2 + \ldots\right]\right\},$$

which simplifies to

$$\ln\phi(t) = -\tfrac{1}{2}t^2 + O(n^{-\frac{1}{2}}).$$

Thus, when $n \to \infty$, $\Phi(t) \to \exp(-\frac{1}{2}t^2)$. But this is precisely the characteristic function for a normal distribution with mean 0 and variance 1 (see eq.(4.74)). In the limit of very large n, therefore, the variable z has the distribution N(0,1), in accordance with the statement made; this completes the proof for the Central Limit Theorem for this particular case.

The Central Limit Theorem is responsible for many other theorems and statements in the theory of probability and statistics. In fact, we have already seen the theorem in operation earlier in this chapter, in the observed tendency towards normality for increasing n of the (discrete) binomial and Poisson distributions.

It is an empirical fact that a variety of phenomena seems to be well described by the normal distribution law. Thus, for example, repeated measurements on the same physical system by a certain apparatus may show a distribution of outcomes which approaches the normal if the number of observations becomes sufficiently large. This is by no means obvious, since even the simplest physical experiment involves many different sources of errors, which can be of instrumental origin as well as have a subjective character. The Central Limit Theorem affords a theoretical explanation of these empirical findings. The total effect registered can be considered as the sum of a "true effect" and a "total error", which is in turn the combined effect of a number of mutually independent elementary errors. When the number of error sources becomes large the total error will be normally distributed, and hence also the total effect observed.

Quite frequently one can envisage that the "true effect" under study is composed of several partial contributions. For example, multiparticle production in high-energy reactions is thought of as being due to different fundamental mechanisms. A dynamical variable describing these reactions will then have contributions from different physical effects. If these effects are independent and many in number, the resultant will be a variable which is normally distributed, irrespective of the spectral forms from the individual elementary effects. In a certain sense the Central Limit Theorem can therefore be said to obscure the fundamental effects in nature.

4.8.9 Example: Gaussian random number generator

An illustrative example with a practical application of the Central Limit Theorem is the construction of a generator for *normally* distributed random numbers.

Suppose that there is available an ordinary random number generator which delivers numbers uniformly distributed over the interval [0,1], as described in Sect.4.5.2, with a mean value $\frac{1}{2}$ and a variance $\frac{1}{12}$. Let x_i be the i-th number in a sequence of n consecutive numbers from this generator. Then, according to the Central Limit Theorem, in the limit of large n, the variable

$$z = \frac{\sum_{i=1}^{n} x_i - \frac{n}{2}}{\sqrt{\frac{n}{12}}}$$

will have a distribution which is exactly normal, with mean zero and variance one. A reasonable approximation to the normal distribution is obtained for n values as small as 10. A practical choice is to take n=12, which gives simply

$$z = \sum_{i=1}^{12} x_i - 6 .$$

This variable, constructed for repeated sequences of size 12, produces a distribution of numbers between -6 and +6, which differs only slightly from N(0,1) and will be sufficiently accurate for most purposes.

An alternative method to obtain normally distributed random numbers is the following: If x_1 and x_2 is a pair of numbers from a generator producing a uniform distribution between 0 and 1, any derived quantity constructed as $z_1 = \sqrt{-2\ell n x_1}\,\cos(2\pi x_2)$ or $z_2 = \sqrt{-2\ell n x_1}\,\sin(2\pi x_2)$ will in the long run be of standard normal form; see Exercise 4.43.

4.9 THE BINORMAL DISTRIBUTION

In generalizing the normal p.d.f. to more than one variable it may be useful to investigate first the case with two variables in some detail. This particular case demonstrates the special features of the normal distribution in a transparent way, and will facilitate the further generalization to many dimensions.

4.9.1 Definition and properties

A joint probability distribution for two random variables x_1 and x_2 is called *binormal* or *two-dimensional normal* if, for $-\infty \le x_1, x_2 \le \infty$,

$$\left.\begin{aligned} f(x_1,x_2) &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\, e^{-\frac{1}{2}Q} , \\ Q &= \frac{1}{1-\rho^2}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)\right] . \end{aligned}\right\} \qquad (4.77)$$

In order that $f(x_1,x_2)$ be non-negative and real, both σ's are required positive,

and $|\rho| < 1$. As one may suspect from the notation the parameters $\mu_2, \mu_2, \sigma_1^2, \sigma_2^2$ will have the meaning of mean values and variances, respectively, while ρ measures the correlation between the two variables.

To see this, it is convenient to work out the characteristic function for x_1 and x_2 with the p.d.f. (4.77) from the general definition of eq. (3.50); the result is

$$\Phi(t_1,t_2) = e^{it_1\mu_1 + it_2\mu_2 + \frac{1}{2}[(it_1)^2\sigma_1^2 + (it_2)^2\sigma_2^2 + (it_1)(it_2)2\rho\sigma_1\sigma_2]} \qquad (4.78)$$

The algebraic moments of the variables x_1 and x_2 can then be obtained from this function, by calculating partial derivatives with respect to (it_1) and (it_2), and evaluating the expressions for $t_1=t_2=0$, as was shown in Sect. 3.5.7. We find, for example,

$$E(x_1) = \left.\frac{\partial\Phi}{\partial(it_1)}\right|_{t_1=t_2=0} = \mu_1 ,$$

$$E(x_1^2) = \left.\frac{\partial^2\Phi}{\partial(it_1)^2}\right|_{t_1=t_2=0} = \mu_1^2 + \sigma_1^2 ,$$

and similarly for the other variable; these are the usual relations between the lowest moments of a random variable (see Sect. 3.3.3). For the product of x_1 and x_2 the expectation is

$$E(x_1x_2) = \left.\frac{\partial^2\Phi}{\partial(it_1)\partial(it_2)}\right|_{t_1=t_2=0} = \mu_1\mu_2 + \rho\sigma_1\sigma_2$$

From eqs. (3.39), (3.40) the correlation coefficient $\rho(x_1,x_2)$ between the two variables therefore becomes

$$\rho(x_1,x_2) = \frac{E(x_1x_2)-E(x_1)E(x_2)}{\sigma_1\sigma_2} = \frac{(\mu_1\mu_2+\rho\sigma_1\sigma_2)-\mu_1\mu_2}{\sigma_1\sigma_2} = \rho$$

Hence the parameter ρ in the p.d.f. (4.77) can be identified as the correlation coefficient between x_1 and x_2 .

If $\rho = 0$ we see from eqs. (4.77) and (4.78) that the p.d.f. and the characteristic function both factorize into two separate parts, one for each variable giving

$$f(x_1,x_2) = \left[\frac{1}{\sqrt{2\pi}\,\sigma_1} e^{-\frac{1}{2}(x_1-\mu_1)^2/\sigma_1^2}\right]\cdot\left[\frac{1}{\sqrt{2\pi}\,\sigma_2} e^{-\frac{1}{2}(x_2-\mu_2)/\sigma_2^2}\right] \tag{4.79}$$

and

$$\phi(t_1,t_1) = \left[e^{it_1\mu_1+\frac{1}{2}(it_1)^2\sigma_1^2}\right]\cdot\left[e^{it_2\mu_2+\frac{1}{2}(it_2)^2\sigma_2^2}\right] \tag{4.80}$$

According to Sects. 3.5.4 and 3.5.7 the variables x_1 and x_2 are therefore also *independent*. - This is a special feature of the normal distribution because, in general, when two variables have zero correlation they need not be independent.

For the general case, when $\rho \neq 0$, the marginal distribution in one variable is found by integrating the binormal p.d.f. (4.77) over the second variable, (this is the definition of a marginal probability density given in Sect. 3.5.5). Thus the marginal distribution in x_1 is obtained by evaluating the integral

$$h_1(x_1) = \int_{-\infty}^{\infty} f(x_1,x_2)\,dx_2 \ .$$

In substituting $f(x_1,x_2)$ from eq. (4.77) the factors can here be organized to give

$$h_1(x_1) = \frac{1}{\sqrt{2\pi}\,\sigma_1} e^{-\frac{1}{2}(x_1-\mu_1)^2/\sigma_1^2}\left\{\frac{1}{\sqrt{2\pi}\,\sigma_2\sqrt{1-\rho^2}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(x_2-C)^2/\sigma_2^2(1-\rho^2)}dx_2\right\},$$

where C in the exponent of the integral is independent of the integration variable x_2, $C \equiv \mu_2 + \rho\sigma_2/\sigma_1(x_1-\mu_1)$. The curly bracket is therefore an integrated normal p.d.f, which gives just 1. Hence the marginal distribution in x_1 is

$$h_1(x_1) = \frac{1}{\sqrt{2\pi}\,\sigma_1} e^{-\frac{1}{2}(x_1-\mu_1)^2/\sigma_1^2} \equiv N(\mu_1,\sigma_1^2) \ . \tag{4.81}$$

Similarly, the marginal distribution in x_2 becomes equal to $N(\mu_2,\sigma_2^2)$. Thus

the projections of the binormal distribution perpendicular to the coordinate axes are always of a normal shape, independent of the correlation between the variables.

Let us next turn to the conditional distributions which, according to Sect.3.5.5, are defined as the ratio between the joint p.d.f. and the marginal distributions (eq.(3.48)). For example, the conditional distribution $f(x_2|x_1)$ for x_2, given x_1, is equal to $f(x_1,x_2)/h_1(x_1)$. From the expressions above it will be seen that the curly bracket corresponds to the integral of $f(x_2|x_1)$ over x_2. Hence the conditional distribution in x_2 for given x_1 is

$$f(x_2|x_1) = N\left(\mu_2+\rho\frac{\sigma_2}{\sigma_1}(x_1-\mu_1),\sigma_2^2(1-\rho^2)\right) \ . \tag{4.82}$$

A corresponding expression is obtained for the other conditional density $f(x_1|x_2)$. Thus any intersection of the two-dimensional surface by a plane perpendicular to a coordinate axis will produce a normal curve. From eq.(4.82) it is interesting to observe that although the mean value of the conditional distribution in x_2 for given x_1 does depend on x_1, the variance does not. Hence the plane intersections for different values of x_1 will be a set of normal curves which all have variance $\sigma_2^2(1-\rho^2)$, but with the mean value $\mu_2+\rho\sigma_2/\sigma_1(x_1-\mu_1)$ growing linearly with x_1.

Other interesting features of the binormal distribution are contained in the exercises below; see in particular 4.46-4.48.

Exercise 4.43: *Let x_1 and x_2 be two independent variables which are uniformly* distributed between 0 and 1. Show that two new variables $z_1 = \sqrt{-2\ln x_1}\cos(2\pi x_2)$, $z_2 = \sqrt{-2\ln x_1}\sin(2\pi x_2)$ will be binormally distributed, with marginal distributions N(0,1) and zero correlation.

Exercise 4.44: Verify that the covariance matrix and its inverse for two variables with the binormal distribution (4.77) are given by, respectively,

$$V = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \qquad V^{-1} = \frac{1}{\sigma_1^2\sigma_2^2(1-\rho^2)}\begin{pmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix}$$

(i) Show that the binormal p.d.f. can be expressed as

$$f(\underline{x}) = (2\pi)^{-1}[|V|]^{-\frac{1}{2}}\exp[-\tfrac{1}{2}(\underline{x}-\underline{\mu})^T V^{-1}(\underline{x}-\underline{\mu})] \tag{4.77b}$$

where $\underline{x} = \{x_1,x_2\}$ and $\underline{\mu} = \{\mu_1,\mu_2\}$.

(ii) Show that the marginal distribution in x_1 is given by a formula of the form of eq.(4.77b), but with V replaced by the "submatrix" obtained by deleting the second row and column from V.

(iii) Show that the conditional distribution in x_2, given x_1, is also given by a formula of the form (4.77b), but with V replaced by V*, where V* is the "submatrix" obtained by deleting the first row and column from V^{-1} and inverting the resultant "submatrix".

Exercise 4.45: Verify that the characteristic function for two variables with a binominal distribution is given by eq.(4.78), which in vector notation reads

$$\phi(\underline{t}) = \exp[(i\underline{t})^T \underline{\mu} + \tfrac{1}{2}(i\underline{t})^T V (i\underline{t})] . \qquad (4.78b)$$

Exercise 4.46: With x_1 and x_2 related in the binormal p.d.f., show that $y = a_1x_1+a_2x_2 \equiv \underline{a}^T\underline{x}$ is $N(a_1\mu_1+a_2\mu_2, a_1^2\sigma_1^2+a_2^2\sigma_2^2+2a_1a_2\rho\sigma_1\sigma_2) = N(\underline{a}^T\underline{\mu}, \underline{a}^T V \underline{a})$. (Hint: Find the characteristic function for y.)

Excercise 4.47: (i) Show that two variables constructed as linear combinations of the variables of a binormal distribution (4.77) will also be binormally distributed. (Hint: Write $y_1=a_1x_1+a_2x_2$, $y_2=b_1x_1+b_2x_2$ and show that the characteristic function $\Phi(t_1,t_2)$ for y_1 and y_2 is of the form of eq. (4.78).)
(ii) Writing $\underline{a} = \{a_1,a_2\}$, $\underline{b} = \{b_1,b_2\}$, show that the new binormal distribution has the marginal distributions $N(\underline{a}^T\underline{\mu}, \underline{a}^T V \underline{a})$ and $N(\underline{b}^T\underline{\mu}, \underline{b}^T V \underline{b})$ for y_1 and y_2, respectively, and that their covariance is $\mathrm{cov}(y_1,y_2)=\underline{a}^T V \underline{b}$. Hence y_1 and y_2 will be independent if, and only if, the transformation makes $\underline{a}^T V \underline{b} = 0$.

Exercise 4.48: Let x_1,x_2 be related in the binormal p.d.f. of eq.(4.77).
(i) Show that a change of variables to y_1,y_2 by the orthogonal transformation

$$y_1 = \frac{1}{\sqrt{2}}\left[\frac{x_1-\mu_1}{\sigma_1} + \frac{x_2-\mu_2}{\sigma_2}\right], \qquad y_2 = \frac{1}{\sqrt{2}}\left[\frac{x_1-\mu_1}{\sigma_1} - \frac{x_2-\mu_2}{\sigma_2}\right],$$

brings the p.d.f. over to the form

$$f(y_1,y_2) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\tfrac{1}{2}\left(\frac{y_1^2}{1+\rho} + \frac{y_2^2}{1-\rho}\right)\right],$$

which shows that y_1 and y_2 will be independent and normal. Here, each of the squared y_i occurs with a coefficient determined by the latent roots (eigenvalues) of the matrix $V(\underline{x})$.
(ii) Show that a subsequent transformation $z_1 = y_1/\sqrt{1+\rho}$, $z_2 = y_2/\sqrt{1-\rho}$, or directly,

$$z_1 = \frac{1}{\sqrt{1+\rho}}\frac{1}{\sqrt{2}}\left[\frac{x_1-\mu_1}{\sigma_1} + \frac{x_2-\mu_2}{\sigma_2}\right], \qquad z_2 = \frac{1}{\sqrt{1-\rho}}\frac{1}{\sqrt{2}}\left[\frac{x_1-\mu_1}{\sigma_1} - \frac{x_2-\mu_2}{\sigma_2}\right]$$

makes $Q = z_1^2 + z_2^2$, and

$$f(z_1,z_2) = \left[\frac{1}{\sqrt{2\pi}}\exp(-\tfrac{1}{2}z_1^2)\right]\cdot\left[\frac{1}{\sqrt{2\pi}}\exp(-\tfrac{1}{2}z_2^2)\right].$$

This shows that z_1 and z_2 will be independent and standard normal. The quantity Q, now in the form of a sum of two squared, independent N(0,1) variables, is seen to be $\chi^2(2)$, in view of the definition of a chi-square variable (see Sect. 5.1.1).

4.9.2 Example: Construction of a binormal random number generator

For Monte Carlo simulations of particle production and subsequent detection in an experimental set-up it is sometimes desirable to have available a method for generating variables which are normally distributed *and* internally correlated. It is known, for example, that a charged particle moving in a uniform magnetic field describes a helix curve with axis along the field direction, which can be parametrized in terms of three quantities $1/\rho$, λ, and ϕ: $1/\rho$ measures the curvature of the helix projection in a plane perpendicular to the field, ϕ gives the azimuthal angle in this plane, and λ the dip angle of the helix relative to the same plane. Further, experience has shown that each of these quantities under measurements can be considered as a normally distributed variable, with a spread around the central (true) value as implied by the accuracy of the measuring system; in addition, $1/\rho$ and ϕ are correlated. To simulate an experiment by the Monte Carlo technique one therefore needs a prescription for obtaining sets of random, normal variables, such that two of them possess a mutual relationship corresponding to a binormal distribution of specified correlation, corresponding to that between the curvature and the azimuthal angle.

We assume that a generator for Gaussian random numbers is available, which upon call produces a "random number" z such that, in the long run, z will be nearly N(0,1). Any independent normal variable of mean value μ and standard deviation σ is then simply constructed as $\mu+\sigma z$.

Suppose that two dependent variables x_1 and x_2 are required to have the correlation coefficient ρ. We then use the Gaussian random number generator to obtain two independent standard normal variables z_1 and z_2, and write

$$\begin{aligned} x_1 &= \frac{1}{\sqrt{2}}\sqrt{1+\rho}\, z_1 + \frac{1}{\sqrt{2}}\sqrt{1-\rho}\, z_2 , \\ x_2 &= \frac{1}{\sqrt{2}}\sqrt{1+\rho}\, z_1 - \frac{1}{\sqrt{2}}\sqrt{1-\rho}\, z_2 . \end{aligned} \qquad (4.83)$$

This transformation represents the solution to our problem, since the joint p.d.f. for x_1 and x_2 becomes

$$f(x_1,x_2) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}(x_1^2 + x_2^2 - 2\rho x_1x_2)\right\}. \qquad (4.84)$$

The variables x_1 and x_2 are therefore binormally distributed with correlation

coefficient ρ, as required, and each variable has marginal distribution N(0,1). Clearly, if non-zero mean values and standard deviations different from 1 are needed, this can be achieved by appropriate location shifts and scalings. Thus the case with parameters μ_1, μ_2, σ_1^2, σ_2^2 is obtained by replacing the left-hand sides in eq.(4.83) by $(x_1-\mu_1)/\sigma_1$ and $(x_2-\mu_2)/\sigma_2$, respectively. The reader will recognize this transformation as the inverse of the general transformation in Exercise 4.48 of the previous section.

4.10 THE MULTINORMAL DISTRIBUTION

4.10.1 Definition and properties

The normal distributions of dimension 1 and 2, as defined by eqs.(4.59) and (4.77), respectively, lead us to search a *multinormal* or *n-dimensional normal distribution*, for which the p.d.f. should be of an exponential type and the exponent have a general quadratic dependence on n variables x_i,

$$\left.\begin{aligned} f(x_1,x_2,\cdots,x_n) &= C\, e^{-\frac{1}{2}Q}\,, \\ Q &= \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}\left(\frac{x_i-\mu_i}{\sigma_i}\right)\left(\frac{x_j-\mu_j}{\sigma_j}\right), \qquad -\infty \le x_i \le \infty\,. \end{aligned}\right\} \tag{4.83}$$

This will be an allowed p.d.f. provided the coefficients c_{ij}, symmetric in the indices, are such that the integral over the n-dimensional real space exists, and C ensures a proper overall normalization.

With an eye to the two-dimensional case, for which vector notation was introduced in Exercise 4.44, it is at once realized that a more compact form of eq.(4.83) is

$$f(\underline{x}) = \frac{1}{(2\pi)^{\frac{1}{2}n}\,|V|^{\frac{1}{2}}}\, e^{-\frac{1}{2}(\underline{x}-\underline{\mu})^T V^{-1}(\underline{x}-\underline{\mu})}\,, \tag{4.84}$$

where $\underline{x} = \{x_1,x_2,\cdots,x_n\}$, $\underline{\mu} = \{\mu_1,\mu_2,\cdots,\mu_n\}$, and V is the symmetric n×n covariance matrix for $\underline{x}$. The denominator of the overall normalization factor includes one $(2\pi)^{\frac{1}{2}}$ for each dimension of $\underline{x}$, and the square root of the determinant of V. Clearly we must have $|V| \neq 0$, and for the integral of $f(\underline{x})$ to exist, V (and V^{-1}) must be positive definite.

The characteristic function for the p.d.f. (4.84) is

$$\Phi(\underline{t}) = e^{(i\underline{t})^T\underline{\mu} + \frac{1}{2}(i\underline{t})^T V(i\underline{t})}\,, \tag{4.85}$$

from which the moments can be calculated in a straightforward manner as described in Sect.3.5.7. For example,

$$E(x_r) = \left.\frac{\partial\Phi}{\partial(it_r)}\right|_{\underline{t}=0} = \mu_r\,,$$

$$E(x_r x_s) = \left.\frac{\partial^2\Phi}{\partial(it_r)\partial(it_s)}\right|_{\underline{t}=0} = \mu_r\mu_s + V_{rs}$$

for any r,s = 1,2,···,n. With $V_{rr} = \sigma_r^2$ and the correlation coefficient between the variables x_r and x_s given by $\rho_{rs} = \rho(x_r,x_s) = V_{rs}/(V_{rr}V_{ss})^{\frac{1}{2}}$ the covariance matrix takes the general form

$$V = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \cdots & \rho_{1n}\sigma_1\sigma_n \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \cdots & \rho_{2n}\sigma_2\sigma_n \\ \vdots & \vdots & & \vdots \\ \rho_{1n}\sigma_1\sigma_n & \rho_{2n}\sigma_2\sigma_n & \cdots & \sigma_n^2 \end{pmatrix}. \tag{4.86}$$

If V is a diagonal matrix, implying that all x_i are uncorrelated, the inverse matrix V^{-1} will also be diagonal. The exponent in eq.(4.84) will then have no terms mixing the different x_i, so that the joint p.d.f. $f(\underline{x})$ will be factorizable into n one-dimensional normal p.d.f.'s for the separate components $x_1,x_2,\cdots,x_n$, showing that they are all independent. - In fact, the variables in a multinormal distribution are independent if, and only if, the covariance matrix is diagonal.

Similar reasoning to that applied for the binormal distribution shows that any projection of the $f(\underline{x})$ from eq.(4.84) to a space of lower dimension gives a new distribution which is of the same form, but with a covariance matrix obtained from the original V by deleting the rows and columns corresponding to the variables projected away. In particular, integrating over all variables except x_i gives the marginal distribution in this variable, which is $N(\mu_i,\sigma_i^2)$. Also, intersecting $f(\underline{x})$ by a plane perpendicular to one of the coordinate axes, say x_i, produces a conditional distribution which is an (n-1)-dimensional normal p.d.f.; the matrix $V = V^*$ in this conditional distribution is obtained by deleting the i-th row and column from the original V^{-1} and inverting the resultant

submatrix.

Other properties of the normal distribution in two dimensions are found to extend to the multi-dimensional case. Thus any linear function of the x_i from a multinormal distribution is itself a (one-dimensional) normally distributed variable (see Exercise 4.49); more generally, a set of variables made up as linear combinations of the multinormal x_i will also be multinormally distributed (Exercise 4.50). This is a vast generalization of the addition theorem for normal variables from Sect.4.8.5.

Exercise 4.49: Let y be a linear function of the x_i with the multinormal distribution (4.84), $y = \sum_{i=1}^{n} a_i x_i = \underline{a}^T \underline{x}$. Show that y is $N(\underline{a}^T \underline{\mu}, \underline{a}^T V \underline{a})$

Exercise 4.50: Let $y_1, y_2, \cdots, y_m$ be a set of linear functions of the x_i with the multinormal p.d.f. (4.84), such that $\underline{y} = S\underline{x}$, where the matrix S is of dimension $m \times n$ $(m \leq n)$. Show that the characteristic function $\Phi(t_1, t_2, \cdots, t_m)$ has a form which implies that the y_i are multinormally distributed with a vector of mean values $S\underline{\mu}$ and covariance matrix SVS^T.

4.10.2 The quadratic form Q

In the multinormal p.d.f. the exponent has been expressed as $-\frac{1}{2}Q$, where Q is the *quadratic form*, or *covariance form*,

$$Q = (\underline{x} - \underline{\mu})^T V^{-1} (\underline{x} - \underline{\mu}) \tag{4.87}$$

in which $\underline{\mu}$ and V are, respectively, the mean value and covariance matrix for the n-component variable $\underline{x}$. For the special cases with $n = 1$ and $n = 2$ it has been indicated earlier that the variable Q has a distribution which is that of a chi-square variable with, respectively, 1 and 2 degrees of freedom (compare Exercises 4.39 and 4.48). It turns out that Q also for a general value of n has a chi-square distribution with n degrees of freedom and thus is a function of one parameter only. This is quite remarkable, in view of the formal structure of Q as a function involving many parameters.

When the covariance matrix is non-singular, as we have assumed, it is always possible to find a linear transformation to a new set of variables y_i from the x_i, which is such that it brings Q to a form with sums of squares in y_i, and with no terms mixing the different y_i. In particular, an orthogonal transformation will produce Q as a sum of squares of n independent y_i, where the coefficients are given by the latent roots of the covariance matrix $V(\underline{x})$; in common parlance, this transformation serves to "diagonalize the covariance matrix".

A further transformation to a new set of scaled variables, say z_i, will express Q as a sum of n squares of independent, standard normal z_i and, in accordance with the definition of Sect.5.1.1, Q is therefore $\chi^2(n)$, a chi-square variable with n degrees of freedom.

If the covariance matrix is singular, $|V| = 0$, its inverse does not exist, and hence eq.(4.84) will loose its usual meaning. There is then at least one linear relation between the x_i, or equivalently, one or more of the x_i is redundant. In this case we shall take eq.(4.84) to mean the multinormal distribution in a space of dimension (n-r), where r is the number of linear redundancies, implying the covariance matrix to be eliminated for the rows and columns corresponding to the redundant x_i's. Similarly, the quadratic form Q can be considered as composed of an equivalent reduced number of terms, making it in this case distributed as $\chi^2(n-r)$ *). This property of quadratic forms of normally distributed variables will be frequently referred to later in this book; since, however, the proof is rather formal and sophisticated its details will not be given here.

Exercise 4.51: (Construction of a multinormal random number generator) Given n independent variables z_i, all $N(0,1)$. Discuss the problem of finding n variables x_i, linearly related to z_i, such that the x_i become multinormally distributed with given covariance matrix.

4.11 THE CAUCHY, OR BREIT-WIGNER, DISTRIBUTION

The *Cauchy distribution* has the form

$$f(x) = \frac{1}{\pi} \frac{1}{1+x^2}, \qquad -\infty \leq x \leq \infty, \tag{4.88}$$

and is an allowed probability density function since the integral of f(x) over all x is equal to one. This distribution is of interest to particle physicists because it produces the Breit-Wigner shape. Unfortunately, however, the function f(x) as defined above is mathematically awkward. Difficulties arise immediately if one tries to evaluate the expectation value of x with the p.d.f. f(x), since the integral

$$\int_{-\infty}^{\infty} x \, \frac{1}{\pi} \frac{1}{1+x^2} \, dx$$

*) The number of independent variables, n-r, is often called the *rank* of the quadratic form.

is not completely convergent. This means that the limiting value

$$\lim_{\substack{L\to\infty \\ L'\to\infty}} \int_{-L}^{L'} x \frac{1}{\pi} \frac{1}{1+x^2}\, dx$$

does not exist, although the principal value, defined with $L' = L$, does exist and is equal to one. Following convention we shall regard the distribution of eq. (4.88) as not possessing a mean. The same convergence situation applies to all other moments x^k. Thus we may say that for the Cauchy distribution no moments are defined, since they all diverge.

One way of getting out of the dilemma is to impose a restriction on the domain of the variable x. The integral of $f(x)$ over a finite interval $[-L,+L]$ is equal to $2/\pi\ (\tan^{-1}L)$. If we therefore redefine our $f(x)$ by this normalization factor and write

$$f'(x) = \frac{1}{2\tan^{-1}L} \cdot \frac{1}{1+x^2}, \qquad -L < x < L, \tag{4.89}$$

this will be an acceptable p.d.f. which is properly normalized and for which all moments exist. From its symmetric form all odd moments of $f'(x)$ vanish identically; in particular, $E(x) = 0$. We also find

$$V(x) = \frac{L}{\tan^{-1}L} - 1. \tag{4.90}$$

This expression for the variance illustrates the state of affairs for the Cauchy p.d.f. (4.88): the tails of this distribution tend so slowly to zero that convergence is prevented. Indeed, when L is permitted to grow indefinitely, the variance can become arbitrarily large, since $V(x)\to\infty$ when $L\to\infty$.

For moderate x values the shape of the Cauchy distribution (4.88) is not very different from the standard normal, as one can see from Fig. 4.9. A quantitative expression of their different tail behaviour is provided by the following table, which gives the fraction of each distribution in both tails beyond the indicated values of $|x|$. For comparison, the table also shows the corresponding fractions for the double exponential distribution (see Exercise 4.55), which in this respect is seen to have an intermediate behaviour.

Distribution	Fraction of distribution in tail					
	$\|x\|$	≥ 1	≥ 2	≥ 3	≥ 4	≥ 6
Standard normal		.3173	.0455	.0027	.00006	-
Double exponential		.3679	.1353	.0498	.0183	.0025
Cauchy		.5000	.2952	.2048	.1560	.1051

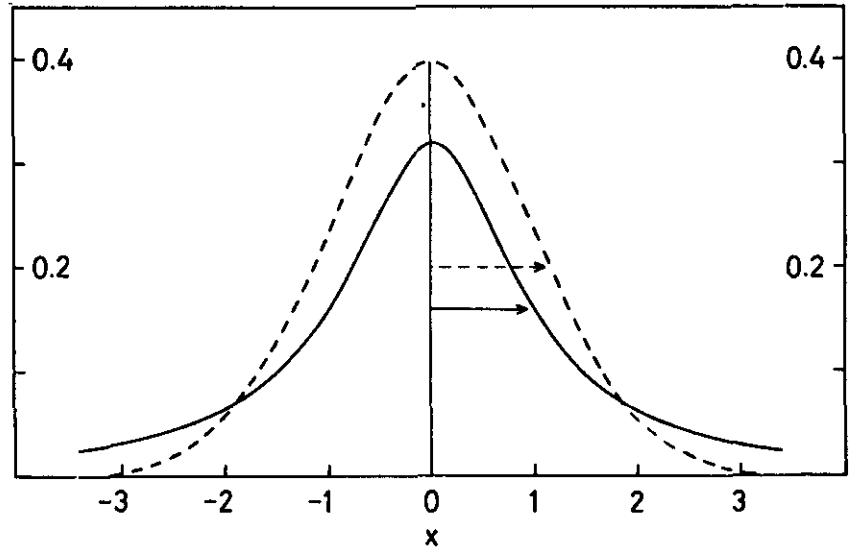

Fig. 4.9. The Cauchy or Breit-Wigner distribution (solid curve) and the standard normal distribution (dashed curve). The half-widths at half-maximum are indicated by arrows of length 1 and $\sqrt{2\ln 2} \approx 1.18$, respectively.

Exercise 4.52: Show that the Breit-Wigner formula for an s-wave resonance of central value M_o and full width Γ at half maximum,

$$f(M;M_o,\Gamma) = \frac{\frac{1}{2}\Gamma}{\pi} \frac{1}{(M-M_o)^2+(\frac{1}{2}\Gamma)^2},$$

corresponds to a Cauchy distribution. Note that with this form, half-maximum occurs for $M = M_o \pm (\frac{1}{2}\Gamma)$, whereas a Gaussian shape $N(M_o, (\frac{1}{2}\Gamma)^2)$ has half-maximum at $M = M_o \pm 1.18(\frac{1}{2}\Gamma)$; compare Exercise 4.37.

Exercise 4.53: Show that the characteristic function for the Cauchy p.d.f. is

$$\Phi(t) = e^{-|t|}.$$

Note that this function has no Taylor expansion around the origin and that therefore the moments of the Cauchy p.d.f. do not exist.

Exercise 4.54: Let $x_1, x_2, \ldots, x_n$ be n independent Cauchy variables, each distributed according to eq.(4.88). Show that $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ has the same distribution.

This result may at first appear somewhat surprising, in view of what has been learned from the Central Limit Theorem. One might perhaps have expected that the arithmetic mean $\bar{x}$ of the n independent variables would become approximately normally distributed for very large n. The appearant discrepancy is due to the fact that the n Cauchy variables do not fulfil the requirement of possessing a *finite* variance, which was essential in deriving the Central Limit Theorem.

Exercise 4.55: (The double exponential distribution)
Discuss the properties of a variable with the p.d.f.

$$f(x) = \frac{1}{2} e^{-|x|}, \qquad -\infty < x < \infty .$$

Note in particular that this distribution has tails which drop off more slowly than the standard normal but faster than the Cauchy distribution.

5. Sampling distributions

The previous chapter has dealt rather extensively with the characteristics and properties of some probability distributions which have been found to describe various physical phenomena quite accurately under certain ideal conditions.

The present chapter will be devoted to a study of the properties of three sampling distributions which are related to the normal. The chi-square, the Student's t, and the F-distributions do not have any direct physical analogues, but can be connected to experimental situations where the normal distribution law is supposed to describe the outcome of measurements. The motivation to study these distributions may perhaps not be clear to the reader at the moment, in which case he should proceed to the following chapters and return to this point when it is found necessary. It may suffice to mention that, although some applications of the sampling distributions are found already in Chapter 7, in connection with simple inference problems involving samples from the normal distribution, a full appreciation of the sampling distributions discussed here will first become evident in the last chapter of the book. Thus a number of examples on hypothesis testing indeed presupposes knowledge on the standard sampling distributions as well as some acquaintance with the related non-central sampling distributions, which are introduced in the exercises of the present chapter.

5.1 THE CHI-SQUARE DISTRIBUTION

5.1.1 Definition

Let us assume that there is given a set of n mutually independent random variables $x_1, x_2, \ldots, x_n$ which are all normal $N(\mu,\sigma^2)$. We may for instance think of the x_i's as the outcomes of n repeated measurements on the same physical system or n independent observations on the same quantity. Then the x_i's constitute a sample of size n from a population which is normal with mean μ and

variance σ^2. We define the chi-square sum χ^2 by adding the squares of the standardized normal variables $(x_i-\mu)/\sigma$, *viz.*

$$\chi^2 \equiv \sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2. \tag{5.1}$$

The variable χ^2 has a probability density function given by

$$f(\chi^2;n) = \frac{1}{2^{\frac{1}{2}n}\,\Gamma(\frac{1}{2}n)}(\chi^2)^{\frac{1}{2}n-1}e^{-\frac{1}{2}\chi^2}, \qquad 0 \le \chi^2 \le \infty, \tag{5.2}$$

which is called the *chi-square distribution with n degrees of freedom.* In eq. (5.2), Γ is the gamma function, for which $\Gamma(x+1) = x\Gamma(x)$, $\Gamma(\frac{1}{2}) = \sqrt{\pi}$, $\Gamma(1) = 1$.

Notice that the number of degrees of freedom is the number of independent variables making up the χ^2 sum (5.1).

Remark 1. The sample $x_1, x_2, \ldots, x_n$ from $N(\mu,\sigma^2)$ has a sample variance

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2,$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i$ is the sample mean. It was stated in Sect.4.8.6 and will be proved in Sect.5.1.6 that the quantity $(n-1)s^2/\sigma^2$ is distributed as a chi-square variable with n-1 degrees of freedom. This is equivalent to saying that

$$\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sigma}\right)^2 \quad \text{is } \chi^2(n-1), \tag{5.3}$$

in words, this sum of squares has a chi-square distribution with n-1 degrees of freedom. This may at a first glance seem to contradict the definition of the chi-square distribution in this section, which implies that

$$\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2 \quad \text{is } \chi^2(n). \tag{5.4}$$

The disagreement is, however, only apparent, for the following reason: In the expression (5.4), μ is considered a known quantity, given independently of the x_i; in the expression (5.3), on the other hand, $\bar{x}$ is a derived quantity, namely the arithmetic mean of the x_i. When this average has been calculated and adopted as an estimate of the true, but unknown parameter μ, only n-1 of the squares in the sum s^2 are independent quantities. Loosely speaking, one degree of freedom has been lost, since it has been used to estimate the unknown parameter (the central value) of the distribution.

Remark 2. From a mathematical viewpoint the requirement that all the x_i should be similarly distributed, (all $N(\mu,\sigma^2)$), is unnecessarily restrictive. In general, for n independent variables x_i from normal distributions $N(\mu_i,\sigma_i^2)$, the sum of squares $\chi^2 = \sum_{i=1}^{n}y_i^2$ of the standardized quantities $y = (x_i-\mu_i)/\sigma_i$ will have a chi-square distribution with n degrees of freedom.

If the variables y_i in the sum $\chi^2 = \sum_{i=1}^{n}y_i^2$ are not N(0,1), but more generally distributed with unit variances and means different from zero (and not necessarily equal), the variable χ^2 will have a *non-central chi-square distribution* with n degrees of freedom. See Exercise 5.12.

5.1.2 Proof for the chi-square p.d.f.

In our discussion of the standard, or central, chi-square distribution we shall for convencience write u instead of χ^2. Also we shall use the Greek symbol ν for n, to be in accordance with our convention for parameters entering a p.d.f.. Thus the chi-square distribution for u is written

$$f(u;\nu) = \frac{1}{2^{\frac{1}{2}\nu}\,\Gamma(\frac{1}{2}\nu)}u^{\frac{1}{2}\nu-1}e^{-\frac{1}{2}u}, \qquad 0 \le u \le \infty\ ;\ \nu > 0. \tag{5.5}$$

We shall not prove eq.(5.5) for the general case of arbitrary ν. For the most trivial case when $\nu=1$ the reader should have no difficulties in verifying the formula by using the change-of-variable technique outlined in Sect.3.7, recalling only that putting $u = \left(\frac{x-\mu}{\sigma}\right)^2$ implies a two-to-one transformation from x to u; compare Exercise 3.6. Here we shall be satisfied with verifying eq. (5.5) for the case $\nu=2$. The case $\nu=3$ can be treated in an analogous way, (Exercise 5.1). For the general case of an arbitrary ν a proof can be given by mathematical induction (Exercise 5.2), or using a more direct method, (see for instance Kendall and Stuart, Chapter 11, Vol.1).

For $\nu=2$ we have

$$u = \left(\frac{x_1-\mu}{\sigma}\right)^2 + \left(\frac{x_2-\mu}{\sigma}\right)^2,$$

where x_1 has a p.d.f. $f(x_1) = (2\pi\sigma^2)^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\left(\frac{x_1-\mu}{\sigma}\right)^2\right)$ and similarly for x_2. The

joint p.d.f. is equal to the product of the p.d.f.'s of the two independent variables x_1, x_2 (compare Sect.3.5.4). We define two new variables ρ and ϕ by

$$\frac{x_1 - \mu}{\sigma} \equiv \rho \cos\phi , \qquad \frac{x_2 - \mu}{\sigma} \equiv \rho \sin\phi ,$$

where $0 \le \rho \le \infty$, $0 \le \phi \le 2\pi$. The transformation from the set x_1, x_2 to the set ρ, ϕ involves the Jacobian

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial \rho} & \frac{\partial x_1}{\partial \phi} \\ \frac{\partial x_2}{\partial \rho} & \frac{\partial x_2}{\partial \phi} \end{vmatrix} = \begin{vmatrix} \sigma \cos\phi & -\sigma\rho \sin\phi \\ \sigma \sin\phi & \sigma\rho \cos\phi \end{vmatrix} = \sigma^2\rho .$$

Hence the joint p.d.f. in terms of the new variables becomes

$$f(\rho,\phi) = f(x_1)f(x_2)\cdot|J| = \frac{1}{2\pi} e^{-\frac{1}{2}\rho^2}\cdot\rho ,$$

is independent of ϕ. The marginal distribution in ρ obtained by integrating over ϕ (Sect.3.5.5) becomes simply

$$f(\rho) = e^{-\frac{1}{2}\rho^2}\cdot\rho .$$

Since the relation between the variables u and ρ is

$$u = \rho^2$$

the p.d.f. for u is given by

$$f(u) = f(\rho)\cdot\left|\frac{d\rho}{du}\right| = e^{-\frac{1}{2}\rho^2}\rho\cdot\left|\frac{1}{2\rho}\right| = \tfrac{1}{2}e^{-\frac{1}{2}u} ,$$

which is seen to be f(u;2), the chi-square distribution with two degrees of freedom.

5.1.3 Properties of the chi-square distribution

The chi-square distribution of eq.(5.5) is shown in Fig. 5.1 for selected values of the parameter ν.

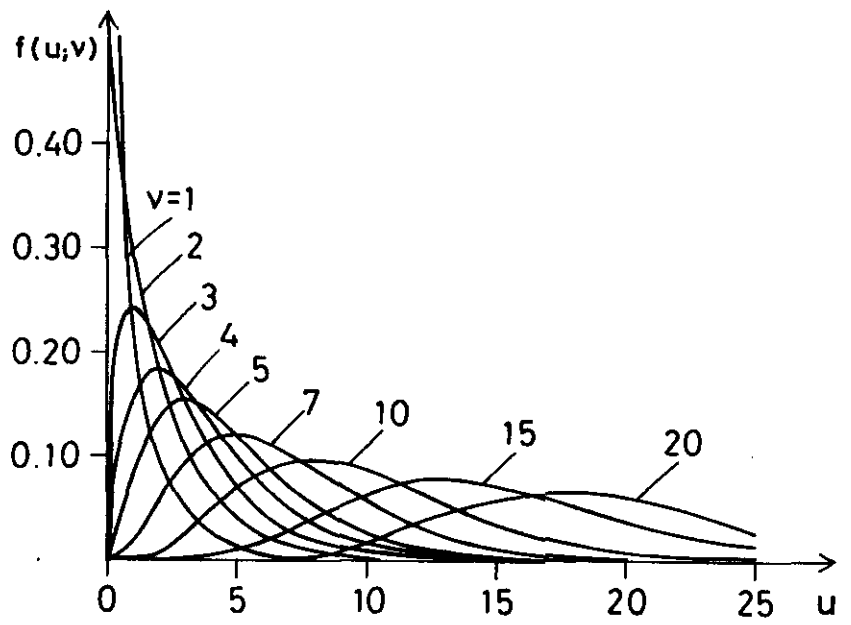

Fig. 5.1. The chi-square distribution for different degrees of freedom ν.

For $\nu \le 2$ the chi-square distribution is monotonically decreasing with increasing u; indeed $\nu=1$ implies an infinite ordinate at u=0. For $\nu > 2$ the distribution has a maximum value (mode) at $u=\nu-2$. It is seen that the chi-square distributions correspond to a special class of the more general gamma distribution (Sect.4.7.1).

The characteristic function for the chi-square distribution is found from the definition eq.(3.22),

$$\Phi(t) = E\left(e^{itu}\right) = \int_0^\infty e^{itu} f(u;\nu)\,du .$$

Inserting the p.d.f. of eq.(5.5) and carrying out the integration leads to

$$\Phi(t) = (1 - 2it)^{-\frac{1}{2}\nu} . \qquad (5.6)$$

When the characteristic function is known one can easily evaluate expectation values by differentiating with respect to (it) and putting t=0, as demonstrated in Sect.3.4. One finds in particular

$$\left.\begin{aligned} E(u) &= \mu_1' = \left.\frac{\partial\Phi}{\partial(it)}\right|_{t=0} = \nu\,, \\ V(u) &= \mu_2 = E(u^2) - \big(E(u)\big)^2 = \left.\frac{\partial^2\Phi}{\partial(it)^2}\right|_{t=0} - \nu^2 = \nu(\nu+2) - \nu^2 = 2\nu\,. \end{aligned}\right\} \tag{5.7}$$

Thus the chi-square distribution has mean value ν and variance 2ν.

In general the algebraic moment of order k_j or the k-th moment about the origin, for the chi-square distribution is given by

$$\mu_k' = E(u^k) = \left.\frac{\partial^k\Phi}{\partial(it)^k}\right|_{t=0} = \nu(\nu+2)\ldots\big(\nu+2(k-1)\big) = 2^k\,\frac{\Gamma(\frac{1}{2}\nu + k)}{\Gamma(\frac{1}{2}\nu)}\,. \tag{5.8}$$

This relation, together with the generally valid formula (3.18), permits the determination of all central moments up to any desired order. In particular, the central moments of orders 3 and 4 are found to be

$$\mu_3 = 8\nu, \qquad \mu_4 = 48 + 12\nu^2. \tag{5.9}$$

Thus the asymmetry and kurtosis coefficients, from their definition by eqs. (3.20) and (3.21), respectively, become

$$\left.\begin{aligned} \gamma_1 &= \frac{\mu_3}{(\mu_2)^{3/2}} = \left(\frac{8}{\nu}\right)^{\frac{1}{2}}, \\ \gamma_2 &= \frac{\mu_4}{(\mu_2)^2} - 3 = \frac{12}{\nu}\,. \end{aligned}\right\} \tag{5.10}$$

These numbers express the tendency seen in Fig. 5.1, that the skewness of the chi-square distribution decreases for increasing ν, while the shape becomes more "bell"-like. Visually the distribution looks "normal" already at $\nu \approx 20$. In the limit $\nu \to \infty$ the coefficients γ_1 and γ_2 are zero, indicating exact symmetry and a peaking equal to that of a normal distribution.

In fact, it can be shown that asymptotically the chi-square distribution does indeed become identical to the normal distribution. To see this it is sufficient to demonstrate that the characteristic functions for the two distributions become equal in the limit of large ν. Guided by the established facts that the mean and variance for the chi-square distribution are given by ν and 2ν, respectively, we form the standardized variable

$$y_1 = \frac{u - \nu}{\sqrt{2\nu}}\,. \tag{5.11}$$

The characteristic function for this variable is

$$\Phi_{y_1}(t) = E\big(e^{ity_1}\big) = E\left(\exp\left(it\,\frac{u-\nu}{\sqrt{2\nu}}\right)\right) = \exp\left(-\frac{\nu it}{\sqrt{2\nu}}\right)\,E\left(\exp\left(\frac{itu}{\sqrt{2\nu}}\right)\right),$$

which can be rewritten as

$$\Phi_{y_1}(t) = \exp\left(-\frac{\nu it}{\sqrt{2\nu}}\right)\,\Phi_u\left(\frac{t}{\sqrt{2\nu}}\right) = \exp\left(-\frac{\nu it}{\sqrt{2\nu}}\right)\,\left(1 - \frac{2it}{\sqrt{2\nu}}\right)^{-\frac{1}{2}\nu}$$

where, in the last step, we have inserted the expression of eq.(5.6) for the characteristic function of the variable u. Taking the logarithm and expanding the last term we have

$$\begin{aligned} \ln\,\Phi_{y_1}(t) &= -\frac{\nu it}{\sqrt{2\nu}} - \tfrac{1}{2}\nu\left(\left(-\frac{2it}{\sqrt{2\nu}}\right) - \frac{1}{2}\left(-\frac{2it}{\sqrt{2\nu}}\right)^2 + \frac{1}{3}\left(-\frac{2it}{\sqrt{2\nu}}\right)^3 - \ldots\right) \\ &= -\frac{1}{2}\,t^2 + \sigma\big(\nu^{-\frac{1}{2}}\big)\,. \end{aligned}$$

When ν goes towards infinity, $\Phi_{y_1}(t) \to e^{-\frac{1}{2}t^2}$; thus in the limit of infinite ν the variable y_1 has the characteristic function of a standard normal variable. Hence the original variable u, for limiting values of ν, will also be normal, namely $N(\nu,2\nu)$.

Mathematically, the approach of $y_1 = (u-\nu)/\sqrt{2\nu}$ to $N(0,1)$ is rather slow. One can show, see Exercises 5.10, 5.11, that the variable

$$y_2 = \sqrt{2u} - \sqrt{2\nu-1} \tag{5.12}$$

represents a better approximation to $N(0,1)$.

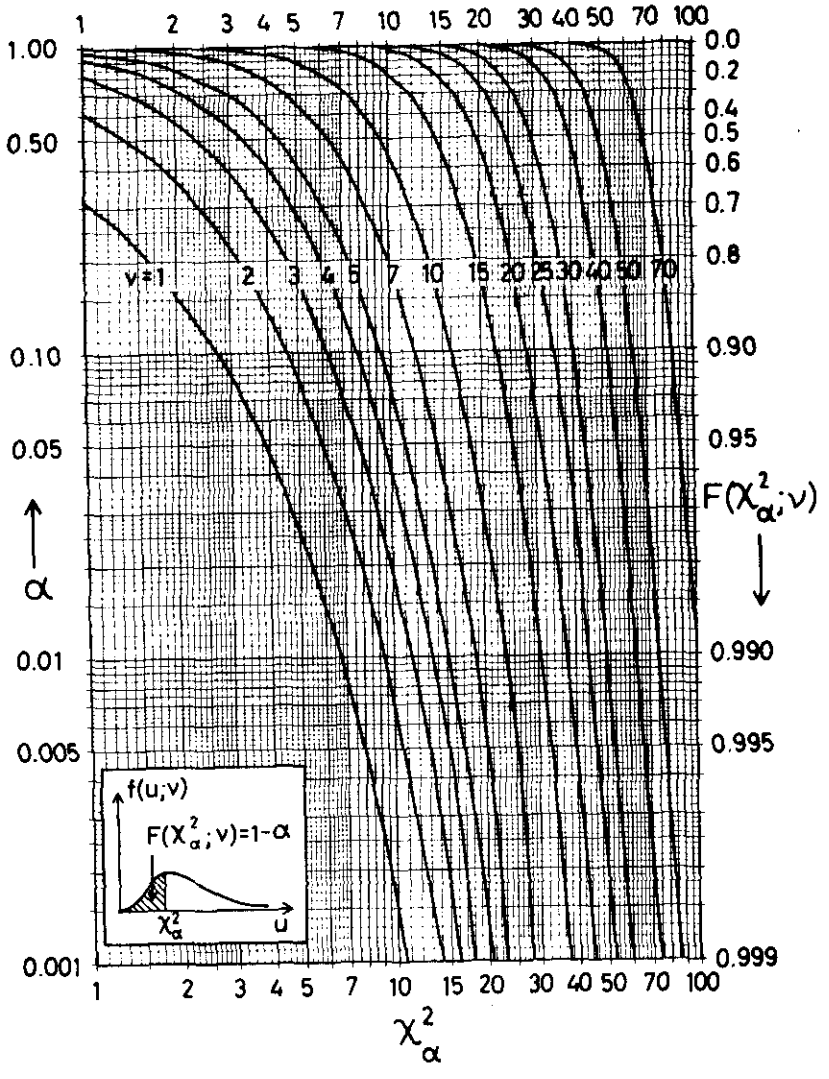

Fig. 5.2. Probability contents of the chi-square distribution.

5.1.4 Probability contents of the chi-square distribution

In practice one is frequently interested in the cumulative chi-square distribution to calculate confidence intervals or for testing hypotheses involving chi-square distributed variables.

Figure 5.2 gives probability contents of the chi-square p.d.f. for different numbers of degrees of freedom. The figure shows a double-logarithmic display of the quantities $F(\chi^2_\alpha;\nu)$ and α *versus* χ^2_α, as implied by the relation

$$F(\chi^2_\alpha;\nu) \equiv \int_0^{\chi^2_\alpha} f(u;\nu)du = 1 - \alpha . \qquad (5.13)$$

Appendix Table A8 gives values of χ^2_α for different ν and specified entries of $F(\chi^2_\alpha;\nu)$.

When the number of degrees of freedom is sufficiently large, $\nu \geq 30$, the probability contents of the chi-square distribution can easily be found using the fact that the variable y_2 of eq.(5.12) (or y_1 of eq.(5.11)) is approximately standard normal. See Exercise 5.8.

It may be worth noting, that the p.d.f. for the variable $F(\chi^2_\alpha;\nu)$ of eq.(5.13) is uniform over the interval [0,1]. (This fact is generally true for any variable defined by the cumulative integral of a p.d.f., see Sect.4.5.1.) We shall see an example of the usefulness of this property in Sect.10.4.4.

5.1.5 Addition theorem for chi-square distributed variables

It has previously been shown that a linear combination of independent, normal variables is itself a normally distributed variable (the addition theorem for normally distributed variables, Sect.4.8.5). A similar theorem holds for a linear combination of independent chi-square variables, and may be stated as follows:

> Let $u_1,u_2,\ldots,u_r$ be a set of independent variables having chi-square distributions with $\nu_1,\nu_2,\ldots,\nu_r$ degrees of freedom, respectively. Then the sum $v=u_1+u_2+\ldots+u_r$ is also a chi-square distributed variable, with $\nu=\nu_1+\nu_2+\ldots+\nu_r$ degrees of freedom.

This theorem is proved simply by noting that the characteristic function for the variable v gets the same form as the characteristic function for an individual u_i. Because of the assumed independence, eq.(3.51) applies, and gives

$$\Phi_v(t) = \Phi_{u_1}(t)\,\Phi_{u_2}(t)\ldots\Phi_{u_r}(t)$$

$$= (1-2it)^{-\frac{1}{2}\nu_1}(1-2it)^{-\frac{1}{2}\nu_2}\ldots(1-2it)^{-\frac{1}{2}\nu_r}$$

$$= (1-2it)^{-\frac{1}{2}(\nu_1+\nu_2+\ldots+\nu_r)}.$$

Hence v is $\chi^2(\nu_1+\nu_2+\ldots+\nu_r)$.

The addition theorem for chi-square variables in fact may appear intuitively correct, because the number of degrees of freedom is nothing but the number of *independent terms* making up the χ^2 sum.

5.1.6 Proof that $(n-1)s^2/\sigma^2$ for sample from $N(\mu,\sigma^2)$ is $\chi^2(n-1)$

Before leaving the chi-square distribution we want to prove the important fact about the variance of a normal sample which was briefly mentioned in Sect.3.10.2, and emphasized in Sect.4.8.6 as well as in the beginning of this chapter. Specifically, if $x_1,x_2,\ldots,x_n$ is a sample from $N(\mu,\sigma^2)$, with mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i$ and variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$, we shall prove that the variable

$$\frac{(n-1)s^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sigma}\right)^2$$

is chi-square distributed with n-1 degrees of freedom.

From the independent variables x_i, which are all $N(\mu,\sigma^2)$, we make a change to a new set of variables y_i by *Helmert's transformation*,

$$y_1 = \frac{x_1-x_2}{\sqrt{2}}$$

$$y_2 = \frac{x_1+x_2-2x_3}{\sqrt{6}}$$

$$\vdots$$

$$y_{n-1} = \frac{x_1+x_2+\ldots+x_{n-1}-(n-1)x_n}{\sqrt{n(n-1)}}$$

$$y_n = \frac{x_1+x_2+\ldots+x_n}{\sqrt{n}} = \sqrt{n}\,\bar{x}.$$

This is an orthonormal transformation, the coefficients a_{ij} in $y_i = \sum_{j=1}^{n}a_{ij}x_j$

satisfying

$$\sum_{i=1}^{n}a_{ij}a_{ik} = \delta_{jk}.$$

The independent variables y_i are all normally distributed, each being $N(0,\sigma^2)$. Now one has

$$(n-1)s^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2 = \sum_{i=1}^{n}x_i^2 - n\bar{x}^2 = \sum_{i=1}^{n}y_i^2 - y_n^2 = \sum_{i=1}^{n-1}y_i^2.$$

Therefore

$$\frac{(n-1)s^2}{\sigma^2} = \sum_{i=1}^{n-1}\left(\frac{y_i}{\sigma}\right)^2,$$

where the right-hand side involves a sum of squares of independent standard normal variables, as required for the definition of a chi-square variable. Since the number of independent terms is n-1, the variable $(n-1)s^2/\sigma^2$ is $\chi^2(n-1)$.

The proof that $\bar{x}$ and s^2 for a normal sample are also independent variables is left as an exercise for the reader (Exercise 5.9).

Exercise 5.1: Verify eq.(5.5) for the case $\nu=3$. (Hint: Use the procedure in the text for $\nu=2$, putting

$$\frac{x_1-\mu}{\sigma} = \rho\cos\theta\cos\phi, \quad \frac{x_2-\mu}{\sigma} = \rho\cos\theta\sin\phi, \quad \frac{x_3-\mu}{\sigma} = \rho\sin\theta.)$$

Exercise 5.2: Prove eq.(5.5) by mathematical induction.

Exercise 5.3: For the chi-square p.d.f with ν degrees of freedom show that $E(u^{\frac{1}{2}\kappa}) = 2^{\frac{1}{2}\kappa}\Gamma\left[\frac{1}{2}(\nu+\kappa)\right]/\Gamma(\frac{1}{2}\nu)$ for all positive and negative integers κ satisfying $\nu+\kappa>0$. Note that this is a more general result than that implied by eq.(5.8), where k is assumed to be a positive integer.

Exercise 5.4: Show that the chi-square distribution has the characteristic function of eq.(5.6).

Exercise 5.5: Verify eqs.(5.8)-(5.10).

Exercise 5.6: Let $x_1,x_2,\ldots,x_r$ be r independent variables which are all uniformly distributed over the interval [0,1]. Show that $u=-2\ln(x_1x_2\ldots x_r)$ is $\chi^2(2r)$. (Hint: See Exercise 4.24.)

Exercise 5.7: From Fig. 5.2, write down values of $F(\chi^2;\nu)$ taking χ^2 equal to the mode of the chi-square distribution for $\nu=3,4,5,10,20,30$ degrees of freedom. What is the limiting value when $\nu\to\infty$?

Exercise 5.8: From Appendix Table A8, verify that $F(\chi^2_\alpha=40.3;\nu=30) = 0.900$. Construcing two approximate normal variables by eqs.(5.11), (5.12) in the text, *viz.* $y_1=(\chi^2_\alpha-\nu)/\sqrt{2\nu}$ and $y_2=\sqrt{2\chi^2_\alpha}-\sqrt{2\nu-1}$, show that the corresponding probabilities obtained from Appendix Table A6, are 0.908 and 0.903, respectively. Repeating the problem for the same ν and a higher χ^2_α, for instance $\chi^2_\alpha=53.7$, show that the y_2 approximation becomes increasingly better than the y_1 approximation.

Exercise 5.9: (Independence of $\bar{x}$ and s^2 for normal sample)

Show that the mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and the variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$ for a sample from $N(\mu,\sigma^2)$ are independent variables. (Hint: Make use of the identity

$$\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2 = \frac{(n-1)s^2}{\sigma^2} + \left(\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right)^2$$

to show that the joint characteristic function of the variables $(n-1)s^2/\sigma^2$ and $\left(\sqrt{n}(\bar{x}-\mu)/\sigma\right)^2$ factorize into their individual characteristic functions. According to Sect.3.5.7 these variables are therefore independent, as will be the variables s^2 and $\bar{x}$.)

Exercise 5.10: (The chi-distribution)

(i) Show that the variable $\chi=\sqrt{u}$ has the probability density function

$$f(\chi;\nu) = \frac{1}{2^{\frac{1}{2}\nu-1}\Gamma(\frac{1}{2}\nu)}\chi^{\nu-1}e^{-\frac{1}{2}\chi^2}, \qquad 0 \le \chi \le \infty.$$

(ii) Show that the algebraic moments for this p.d.f. become in general,

$$\mu'_k = 2^{\frac{1}{2}k}\Gamma\left(\tfrac{1}{2}(\nu+k)\right)/\Gamma(\tfrac{1}{2}\nu),$$

and that in particular the even moments result as

$$\mu'_{2k} = \nu(\nu+2)\ldots\left(\nu+2(k-1)\right).$$

The odd moments can be expressed by the first, μ'_1,

$$\mu'_{2k+1} = (\nu+1)(\nu+3)\ldots(\nu+2k-1)\mu'_1.$$

Use Stirling's expansion

$$\ln\Gamma(x+1) = \tfrac{1}{2}\ln(2\pi) + (x+\tfrac{1}{2})\ln x - x + \frac{1}{12x} - \frac{1}{360x^3} + \ldots$$

to show that

$$\mu'_1 = \sqrt{\nu}\left(1 - \frac{1}{4\nu} + \frac{1}{32\nu^2} + \frac{5}{128\nu^3} + \ldots\right).$$

(iii) With the preceding results and the general relationship between algebraic and central moments, eq.(3.18), show that the lowest central moments for this distribution are

$$\mu_2 = \frac{1}{2} - \frac{1}{8\nu} + \sigma\left(\nu^{-1}\right), \quad \mu_3 = \frac{1}{4\sqrt{\nu}} + \sigma\left(\nu^{-\frac{1}{2}}\right), \quad \mu_4 = \frac{3}{4} - \frac{3}{8\nu} + \sigma\left(\nu^{-1}\right).$$

Here the convergence of $\sigma\left(\nu^{-1}\right)$ towards zero is faster than ν^{-1}, and similarly for $\sigma\left(\nu^{-\frac{1}{2}}\right)$.

Exercise 5.11: Use the results of Exercise 5.10 to show that the variable $\sqrt{2\chi^2}=\sqrt{2u}$ has a distribution with

$$\mu'_1 = \sqrt{2\nu-1} + \sigma\left(\nu^{-3/2}\right), \qquad \mu_2 = 1 - \frac{1}{4\nu} + \sigma\left(\nu^{-1}\right),$$
$$\gamma_1 = \frac{1}{\sqrt{2\nu}} + \sigma\left(\nu^{-\frac{1}{2}}\right), \qquad \gamma_2 = \frac{3}{4\nu^2} + \sigma\left(\nu^{-2}\right).$$

This shows that $\sqrt{2u}$ is distributed about the mean $\sqrt{2\nu-1}$ to order $\nu^{-3/2}$ with a variance 1 to order ν^{-1}. By comparison with eqs.(5.10) it is seen that $\sqrt{2u}$ tends towards normality much faster than u.

Exercise 5.12: (The non-central chi-square distribution)

(i) If $y_1,y_2,\ldots,y_\nu$ is a set of independent variables which are $N(\mu_i,1)$, the variable $u' \equiv \sum_{i=1}^{n} y_i^2$ is a non-central chi-square variable with ν degrees of freedom and non-central parameter $\lambda = \sum_{i=1}^{\nu}\mu_i^2$; for short, u' is $\chi'^2(\nu,\lambda)$. Show that u' has the characteristic function

$$\Phi(t) = (1 - 2it)^{-\frac{1}{2}\nu}\exp\left(\frac{\lambda it}{1 - 2it}\right).$$

Note that if all $\mu_i=0$, $\lambda=0$ and $\Phi(t)$ reduces to the form of eq.(5.6).

(ii) Show that the mean and the variance for u' are given by $(\nu+\lambda)$ and $(2\nu+4\lambda)$, respectively.

(iii) It can be shown (see for instance Kendall and Stuart, Chapter 24, Vol.2) that the p.d.f. for u' is given by

$$f(u';\nu,\lambda) = \frac{1}{2^{\frac{1}{2}\nu}\Gamma(\frac{1}{2})}(u')^{\frac{1}{2}\nu-1}e^{-\frac{1}{2}(u'+\lambda)}\sum_{r=0}^{\infty}\frac{(\lambda u')^r}{(2r)!}\frac{\Gamma(\frac{1}{2}+r)}{\Gamma(\frac{1}{2}\nu+r)}, \qquad 0 \le u' \le \infty.$$

Verify that, when $\lambda=0$, $f(u';\nu,\lambda)$ reduces to the ordinary (central) chi-square distribution with ν degrees of freedom, eq.(5.5).

It can also be shown that the variable

$$u'/\left(\frac{\nu+2\lambda}{\nu+\lambda}\right) = u'/\left(1 + \frac{\lambda}{\nu+\lambda}\right)$$

is approximately distributed like a central chi-square variable according to eq.(5.5), but with a parameter $\nu' = (\nu+\lambda)^2/(\nu+2\lambda)$ where ν' is, in general, fractional. This fact is frequently used to find approximate values of the integral of a non-central chi-square variable, by interpolating in the tables (curves) for the central chi-square distribution.

5.2 THE STUDENT'S T-DISTRIBUTION

5.2.1 Definition

Let x be a standard normal variable N(0,1) and u a chi-square variable with ν degrees of freedom $\chi^2(\nu)$, and assume that x and u are independent. Define a variable t by

$$t \equiv \frac{x}{\sqrt{u/\nu}}, \qquad -\infty \le t \le \infty; \qquad \nu > 0. \tag{5.14}$$

This variable then has a p.d.f. given by

$$f(t;\nu) = \frac{\Gamma\left(\frac{1}{2}(\nu+1)\right)}{\sqrt{\pi\nu}\,\Gamma(\frac{1}{2}\nu)} \cdot \frac{1}{\left(1+\frac{t^2}{\nu}\right)^{\frac{1}{2}(\nu+1)}}, \tag{5.15}$$

which is called the *Student's t-distribution with ν degrees of freedom.*

Remark 1. If x is not an N(0,1) variable, but more generally N(μ,1), and u is $\chi^2(\nu)$ as above, with x and u independent, the variable $t' = x/\sqrt{u/\nu}$ has a *non-central t-distribution*. See Exercise 5.20.

Remark 2. To motivate a study of the Student's t-variable, recall the by now wellknown properties of the mean $\bar{x}$ and variance s^2 of a sample $x_1, x_2, \ldots, x_n$ from $N(\mu,\sigma^2)$,

$$\bar{x} \quad \text{is} \quad N\left(\mu,\frac{\sigma^2}{n}\right), \quad \text{where} \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

$$\frac{(n-1)s^2}{\sigma^2} \quad \text{is} \quad \chi^2(n-1), \quad \text{where} \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2.$$

Moreover, $\bar{x}$ and s^2 are independent variables (Exercise 5.9). Consequently the two independent variables

$$\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \quad \text{and} \quad \frac{(n-1)s^2}{\sigma^2}$$

being, respectively N(0,1) and $\chi^2(n-1)$, satisfy the requirements specified in the beginning of this section, with the trivial difference that the chi-square variable has n-1 degrees of freedom, instead of ν. A variable constructed from these two variables as

$$t = \frac{\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)s^2}{\sigma^2}\Big/(n-1)}} = \frac{\bar{x}-\mu}{s/\sqrt{n}}$$

is therefore a Student's t-variable with n-1 degrees of freedom.

5.2.2 Proof for the Student's t p.d.f.

To prove that the variable t as defined by eq.(5.14) has the p.d.f. of eq.(5.15) one may proceed as follows:

The joint p.d.f. of x and u, because of their independence, is given by the product of the individual p.d.f.'s,

$$f(x,u;\nu) = \left(\frac{1}{\sqrt{2\pi}}\,e^{-\frac{1}{2}x^2}\right)\cdot\left(\frac{1}{2^{\frac{1}{2}\nu}\,\Gamma(\frac{1}{2}\nu)}\,u^{\frac{1}{2}\nu-1}e^{-\frac{1}{2}u}\right).$$

Transforming to a new set of variables,

$$t = \frac{x}{\sqrt{u/\nu}}, \qquad v = u,$$

where $-\infty \le t \le \infty$, $0 \le v \le \infty$, the Jacobian of the transformation is

$$J = \begin{vmatrix} \frac{\partial x}{\partial t} & \frac{\partial x}{\partial v} \\ \frac{\partial u}{\partial t} & \frac{\partial u}{\partial v} \end{vmatrix} = \begin{vmatrix} \sqrt{u/\nu} & 0 \\ 0 & 1 \end{vmatrix} = \sqrt{\frac{v}{\nu}}.$$

In terms of the new variables the joint p.d.f. can be written as

$$f(t,v;\nu) = f(x,u;\nu)|J| = \frac{1}{\sqrt{\pi\nu}\,\Gamma(\frac{1}{2}\nu)\,2^{\frac{1}{2}(\nu+1)}}\,v^{\frac{1}{2}(\nu+1)-1}e^{-\frac{1}{2}v\left(1+\frac{t^2}{\nu}\right)}.$$

Since we are only interested in the variable t we proceed to find the marginal distribution in this variable by integrating over v,

$$f(t;\nu) = \int_0^\infty f(t,v;\nu)\,dv = \frac{1}{\sqrt{\pi\nu}\,\Gamma(\frac{1}{2}\nu)\,2^{\frac{1}{2}(\nu+1)}}\int_0^\infty v^{\frac{1}{2}(\nu+1)-1}e^{-\frac{1}{2}v\left(1+\frac{t^2}{\nu}\right)}\,dv, \tag{5.16}$$

which is seen to lead to eq.(5.15).

5.2.3 Properties of the Student's t-distribution

The Student's t-distribution of eq.(5.15) is shown in Fig. 5.3 for a few values of the parameter ν.

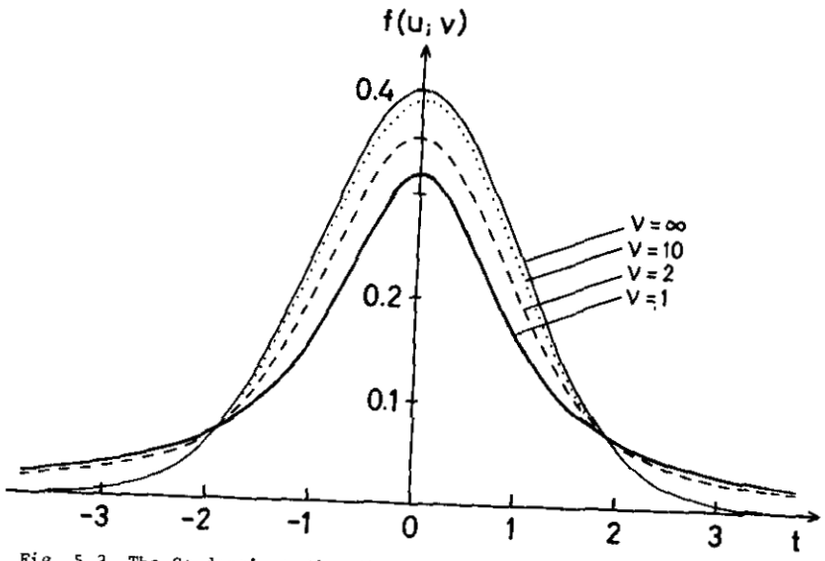

Fig. 5.3. The Student's t-distribution for different degrees of freedom ν. The special cases $\nu=1$ and $\nu=\infty$ correspond to the Cauchy (Breit-Wigner) and the standard normal distributions, respectively.

The properties of this distribution may be enumerated as follows:

(i) $f(t;\nu)$ is unimodular and symmetric about $t=0$.

(ii) For $\nu=1$, $f(t;1) = \frac{1}{\pi}\,\frac{1}{1+t^2}$, that is, the Student's t-distribution with one degree of freedom is identical to the Cauchy distribution.

(iii) When $\nu \to \infty$, $f(t;\nu) \to \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}t^2}$ which means that, asymptotically, the Student's t-distribution becomes equal to the standard normal distribution.

(iv) The mean value and variance of the distribution are given by

$$\left.\begin{aligned} E(t) &= 0 \\ V(t) &= \frac{\nu}{\nu-2}\,, \qquad \nu > 2\,. \end{aligned}\right\} \tag{5.17}$$

The asymmetry and kurtosis coefficients are, respectively,

$$\left.\begin{aligned} \gamma_1 &= 0 \\ \gamma_2 &= \frac{6}{\nu-4}\,, \qquad \nu > 4\,. \end{aligned}\right\} \tag{5.18}$$

These results indicate the similarity between the Student's t-distribution for general ν and the standard normal distribution.

5.2.4 Probability contents of the Student's t-distribution

The cumulative Student's t-distribution is used to evaluate confidence intervals or for testing hypotheses involving t-distributed variables.

Writing

$$F(t_\alpha;\nu) \equiv \int_{-\infty}^{t_\alpha} f(t;\nu)\,dt = 1-\alpha\,, \tag{5.19}$$

where $f(t;\nu)$ is given by eq.(5.15), Appendix Table A7 gives values of t_α corresponding to specified entries for $F(t_\alpha;\nu)$ and α for different degrees of freedom. The table includes the special case $\nu=\infty$, which we have seen is identical to the standard normal distribution. In fact the reader, who already from Sect.4.8.3 is acquainted with the probability contents of N(0,1), will easily establish the connection between Appendix Tables A6 and A7, and thereby be able to make use of the latter. See Exercises 5.17 and 5.18.

Exercise 5.13: Evaluate the integral of eq.(5.16) and thereby complete the proof for the p.d.f. of the Student's t-variable, eq.(5.15).

Exercise 5.14: Verify that the Student's t-distribution specializes to the Cauchy distribution for $\nu=1$ and to the standard normal distribution for $\nu \to \infty$.

Exercise 5.15: Derive the properties listed under (iv) in Sect.5.2.3.

Exercise 5.16: Show that the algebraic moments μ_k' for the t-distribution exist only if $k < \nu$, and that the even moments μ_{2r}' are given by

$$E(t^{2r}) = E\left[\left(\frac{x}{\sqrt{u/\nu}}\right)^{2r}\right] = \nu^r\, E(x^{2r})\cdot E(u^{-r}), \qquad 2r < \nu,$$

because of the independence of the variables x and u; $E(x^{2r})$ and $E(u^{-r})$ are the expectations for N(0,1) and $\chi^2(\nu)$, respectively. Compare Sect.4.8.4 and Exercise 5.3.

Exercise 5.17: Show that, given λ, one can for the Student's t-distribution calculate the limits $\pm b$ in the relation

$$P(-b \le t \le b) = \int_{-b}^{b} f(t;\nu)dt = \gamma$$

by the use of Appendix Table A7. Write down values of b taking γ=0.95 for ν=1,5,10,30,60,∞. Note that for $\nu=\infty$, the limits b=±2.00 correspond to the probability content 0.954 of N(0,1).

Exercise 5.18: Calculate $\sigma = \sqrt{V(t)}$ (eq.(5.17)) and find values of

$$P(-\sigma \le t \le \sigma) = \int_{-\sigma}^{\sigma} f(t;\nu)dt$$

for ν=3,5,10,30,60,∞.

Exercise 5.19: Show that if t^2 is taken as a variable instead of t, the p.d.f. is

$$f(t^2;\nu) = \frac{\Gamma\left(\frac{1}{2}(\nu+1)\right)}{\sqrt{\nu\pi}\;\Gamma\left(\frac{1}{2}\nu\right)} \frac{t^{-1}}{\left(1+\frac{t^2}{\nu}\right)^{\frac{1}{2}(\nu+1)}}.$$

This is the same form as a special case of the F-distribution (with ν_1=1) to be discussed in the next paragraph.

Exercise 5.20: (The non-central t-distribution)
Let x be a normal variable with unit variance but with mean different from zero, *i.e.* x is N(δ,1). If the variable u is $\chi^2(\nu)$ and x and u are independent, the variable $t' \equiv x/\sqrt{u/\nu}$ for $-\infty \le t' \le \infty$ has a p.d.f. given by ($\lambda=\delta^2$)

$$f(t';\nu,\lambda) = e^{-\frac{1}{2}\lambda} \sum_{r=0}^{\infty} \frac{1}{r!}\left(\frac{\lambda}{2}\right)^r \frac{\Gamma\left(\frac{1}{2}(\nu+1)+r\right)}{\Gamma\left(\frac{1}{2}+r\right)\Gamma\left(\frac{1}{2}\nu\right)} \left(\frac{1}{\nu}\right)^{\frac{1}{2}+r} \frac{1}{\left(1+\frac{t'^2}{\nu}\right)^{\frac{1}{2}(\nu+1)+r}},$$

which is called the non-central t-distribution with ν degrees of freedom and non-central parameter δ. Verify that this distribution specializes to the usual (central) t-distribution of eq.(5.15) when δ=0.

When the non-central t-distribution is regarded as a function of t'^2 it is a special case of the non-central F-distribution (for ν_1=1) which is introduced in Exercise 5.29.

5.3 THE F-DISTRIBUTION

5.3.1 Definition

Let u_1 and u_2 be two independent (central) chi-square variables with, respectively, ν_1 and ν_2 degrees of freedom, *i.e.* u_1 is $\chi^2(\nu_1)$, u_2 is $\chi^2(\nu_2)$. Define a variable F by

$$F \equiv \frac{u_1/\nu_1}{u_2/\nu_2} = \frac{\nu_2}{\nu_1}\frac{u_1}{u_2}, \qquad 0 \le F \le \infty; \qquad \nu_1,\nu_2 > 0. \tag{5.20}$$

This variable has the p.d.f.

$$f(F;\nu_1,\nu_2) = \frac{\Gamma\left(\frac{1}{2}(\nu_1+\nu_2)\right)}{\Gamma\left(\frac{1}{2}\nu_1\right)\Gamma\left(\frac{1}{2}\nu_2\right)} \left(\frac{\nu_1}{\nu_2}\right)^{\frac{1}{2}\nu_1} \frac{F^{\frac{1}{2}\nu_1-1}}{\left(1+\frac{\nu_1 F}{\nu_2}\right)^{\frac{1}{2}(\nu_1+\nu_2)}}, \tag{5.21}$$

which is called the *F-distribution with* (ν_1,ν_2) *degrees of freedom.*

Remark 1. If u_1 is not a central chi-square variable as was assumed above, but instead a non-central chi-square variable with ν_1 degrees of freedom, then, with u_2 as $\chi^2(\nu_2)$ and u_1 and u_2 independent, the variable $F' = (u_1/\nu_1)/(u_2/\nu_2)$ has a *non-central F-distribution.* See Exercise 5.29.

Remark 2. In practice one encounters chi-square variables in the form of sample variances for normally distributed variables. For definiteness, let $x_1,x_2,..,x_n$ and $y_1,y_2,\ldots,y_m$ be two independent samples from the same population $N(\mu,\sigma^2)$, for example two series of independent measurements. The sample variances are

$$s_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2, \qquad \text{where} \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i,$$

$$s_2^2 = \frac{1}{m-1}\sum_{i=1}^{m}(y_i-\bar{y})^2, \qquad \text{where} \qquad \bar{y} = \frac{1}{m}\sum_{i=1}^{m}y_i.$$

Then $(n-1)s_1^2/\sigma^2$ and $(m-1)s_2^2/\sigma^2$ are two independent chi-square variables with, respectively, n-1 and m-1 degrees of freedom, satisfying the requirements specified above for an F-variable; taking the ratio according to the definition of eq. (5.20) we find

$$F = \frac{\frac{(n-1)s_1^2}{\sigma^2}/(n-1)}{\frac{(m-1)s_2^2}{\sigma^2}/(m-1)} = \frac{s_1^2}{s_2^2} . \qquad (5.22)$$

This variable then has an F-distribution with (n-1, m-1) degrees of freedom; for obvious reasons F is here called the *variance ratio*.

5.3.2 Proof for the F p.d.f.

To prove that the variable F defined by eq.(5.20) has the p.d.f. of eq.(5.21) one can follow the reasoning that was used to establish the Student's t - p.d.f. in Sect.5.2.2. From the joint p.d.f. of u_1 and u_2, a transformation is made to the new set of variables

$$F = \frac{u_1/\nu_1}{u_2/\nu_2} , \qquad v = u_2 ,$$

where $0 \le F \le \infty$, $0 \le v \le \infty$. Applying the change of variable technique and eliminating the auxiliary variable v by integrating the joint p.d.f. over this variable, the reader should be able to verify eq.(5.21).

5.3.3 Properties of the F-distribution

Figure 5.4 shows a sketch of the F-distribution for a few combinations of the parameters ν_1, ν_2.

The following features characterize the F-distribution:

(i) $f(F;\nu_1,\nu_2)$ is monotonically decreasing if $\nu_1 \le 2$, while it has a maximum for $\nu_1 > 2$, the mode being

$$F_{mode} = \frac{\nu_1-2}{\nu_1}\frac{\nu_2}{\nu_2+2} , \qquad \nu_1 > 2 . \qquad (5.23)$$

(ii) The distribution is positively skew and has a mean value and variance given by

$$\left.\begin{aligned} E(F) &= \frac{\nu_2}{\nu_2-2} , && \nu_2 > 2, \\ V(F) &= \frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)} , && \nu_2 > 4. \end{aligned}\right\} \qquad (5.24)$$

Note that whereas the mode, when it exists, is always <1, the mean value is >1.

(iii) For $\nu_1=1$ the F-distribution specializes to

$$f(F;1,\nu_2) = \frac{\Gamma\left[\frac{1}{2}(\nu_2+1)\right]}{\sqrt{\pi\nu_2}\,\Gamma(\frac{1}{2}\nu_2)} \frac{F^{-\frac{1}{2}}}{\left(1+\frac{F}{\nu_2}\right)^{\frac{1}{2}(\nu_2+1)}} ,$$

which is nothing but a Student's t-distribution with ν_2 degrees of freedom, when the latter is regarded as a function of $t^2=F$; compare Exercise 5.19.

(iv) The limiting properties are such that for ν_1 fixed, $\nu_2 \to \infty$,

$$f(F;\nu_1,\nu_2) \to \frac{1}{2^{\frac{1}{2}\nu_1}\,\Gamma(\frac{1}{2}\nu_1)} (\nu_1 F)^{\frac{1}{2}\nu_1-1} e^{-\frac{1}{2}(\nu_1 F)} \cdot \frac{1}{\nu_1} ,$$

that is, the variable $(\nu_1 F)$ approaches $\chi^2(\nu_1)$; see Exercise (5.25).

For $\nu_1 \to \infty$, $\nu_2 \to \infty$ the F-distribution tends to normal. The approach to normality is, however, rather slow.

(v) The quantity $z = \frac{1}{2}\ln F$ has a distribution which is close to normal, with approximate mean $\frac{1}{2}\left(\frac{1}{\nu_2}-\frac{1}{\nu_1}\right)$ and variance $\frac{1}{2}\left(\frac{1}{\nu_1}+\frac{1}{\nu_2}\right)$. See Exercise 5.28.

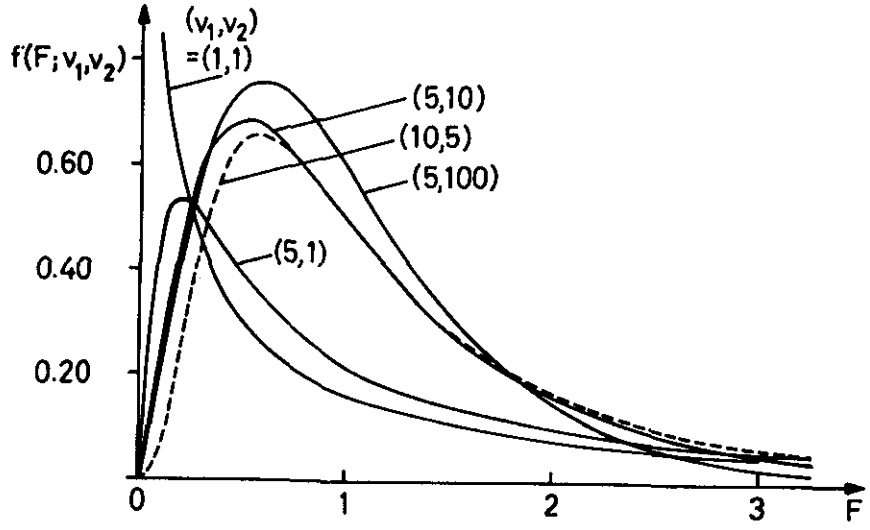

Fig. 5.4. The F-distribution for selected degrees of freedom (ν_1,ν_2).

5.3.4 Probability contents of the F-distribution

The cumulative F-distribution is defined by

$$F(x_\alpha;\nu_1,\nu_2) = \int_0^{x_\alpha} f(F;\nu_1,\nu_2)dF = 1-\alpha\,, \qquad (5.25)$$

where the integrand is the p.d.f. of eq.(5.21). Appendix Table A9 gives values of x_α for specified entries of $F(x_\alpha;\nu_1,\nu_2)$ and different degrees of freedom (ν_1,ν_2).

In practice, the probability contents of the F-distribution is used in connection with hypothesis testing of variances for normal samples; for an example, see Sect.14.3.3.

Exercise 5.21: If the variable F has an F-distribution with (ν_1,ν_2) degrees of freedom, what is the distribution for the variable $1/F$?

Exercise 5.22: In Sect.5.3.2, carry out the details of the proof for the p.d.f. of the F variable.

Exercise 5.23: Verify the formulae for E(F) and V(F) given in the text, eqs. (5.24).

Exercise 5.24: Show that the algebraic moment μ_k' for the F-distribution exists only if $k < \frac{1}{2}\nu_2$, and is then given by

$$\mu_k' = E(F^k) = E\left(\left(\frac{u_1/\nu_1}{u_2/\nu_2}\right)^k\right) = \left(\frac{\nu_2}{\nu_1}\right)^k E(u_1^{\,k})E(u_2^{\,-k}) = \left(\frac{\nu_2}{\nu_1}\right)^k \cdot \frac{\Gamma(\frac{1}{2}\nu_1+k)}{\Gamma(\frac{1}{2}\nu_1)}\;\frac{\Gamma(\frac{1}{2}\nu_2-k)}{\Gamma(\frac{1}{2}\nu_2)}\,,$$

since the expectations $E(u_1^{\,k})$ and $E(u_2^{\,-k})$ are evaluated for $\chi^2(\nu_1)$ and $\chi^2(\nu_2)$, respectively. Compare Exercise 5.3.

Exercise 5.25: Verify the special forms for the F-distribution stated under (iii) and (iv) in Sect.5.3.3.

Exercise 5.26: Calculate $\sigma = \sqrt{V(F)}$ (eq.(5.24)) for the F-distribution with fixed $\nu_1=5$ and $\nu_2=10$, 20, 60, respectively. From Appendix Table A9, find by interpolation $F(3\sigma;\nu_1,\nu_2)$ for the three combinations of (ν_1,ν_2). What is the limiting value when $\nu_2 \to \infty$?

Exercise 5.27: When $\nu_1 \to \infty$, $\nu_2 \to \infty$, what are the limiting values for the mean and the variance of the F-distribution? Compare the corresponding entries of Appendix Table A9.

Exercise 5.28: (The z-distribution)
(i) Put $z \equiv \frac{1}{2}\ln F$. Show that the variable z has the p.d.f.

$$f(z;\nu_1,\nu_2) = \frac{2\Gamma\left(\frac{1}{2}(\nu_1+\nu_2)\right)}{\Gamma(\frac{1}{2}\nu_1)\,\Gamma(\frac{1}{2}\nu_2)}\,\nu_1^{\frac{1}{2}\nu_1}\nu_2^{\frac{1}{2}\nu_2}\,\frac{e^{\nu_1 z}}{(\nu_1 e^{2z}+\nu_2)^{\frac{1}{2}(\nu_1+\nu_2)}}\,,$$

where $-\infty \le z \le \infty$.

(ii) Show that the characteristic function for z is such that

$$\ln\Phi(t) = \frac{it}{2}\left(\frac{1}{\nu_2}-\frac{1}{\nu_1}\right) - \left(\frac{it}{2}\right)^2\left(\frac{1}{\nu_1}+\frac{1}{\nu_2}\right) - \cdots$$

and thereby verify the statement made under (v) in Sect.5.3.3, that z is approximately normally distributed with mean value $\frac{1}{2}\left(\frac{1}{\nu_2}-\frac{1}{\nu_1}\right)$ and variance $\frac{1}{2}\left(\frac{1}{\nu_1}+\frac{1}{\nu_2}\right)$.

Exercise 5.29: (The non-central F-distribution)
(i) Let u_1' be a non-central chi-square variable with ν_1 degrees of freedom and non-central parameter λ as defined in Exercise 5.12, and let u_2 be a (central) chi-square variable with ν_2 degrees of freedom. If u_1' and u_2 are independent the variable $F' \equiv (u_1'/\nu_1)/(u_2/\nu_2)$, where $0 \le F' \le \infty$, has a p.d.f. given by

$$f(F';\nu_1,\nu_2,\lambda) = e^{-\frac{1}{2}\lambda}\sum_{r=0}^{\infty}\frac{1}{r!}\left(\frac{\lambda}{2}\right)^r\frac{\Gamma\left(\frac{1}{2}(\nu_1+\nu_2)+r\right)}{\Gamma(\frac{1}{2}\nu_1+r)\Gamma(\frac{1}{2}\nu_2)}\left(\frac{\nu_1}{\nu_2}\right)^{\frac{1}{2}\nu_1+r}\frac{(F')^{\frac{1}{2}\nu_1-1+r}}{\left(1+\frac{\nu_1 F'}{\nu_2}\right)^{\frac{1}{2}(\nu_1+\nu_2)+r}}\,,$$

which is called the non-central F-distribution with (ν_1,ν_2) degrees of freedom and non-central parameter λ. Show that, for $\lambda=0$, $f(F';\nu_1,\nu_2,\lambda)$ reduces to the ordinary (central) F-distribution of eq.(5.21).

(ii) In view of a statement in Exercise 5.12 the variable $u_1'/\left(\frac{\nu_1+2\lambda}{\nu_1+\lambda}\right)$ will have an approximate central chi-square distribution with parameter $\nu_1' = (\nu_1+\lambda)^2/(\nu_1+2\lambda)$. Hence $\left(u_1'/\left(\frac{\nu_1+2\lambda}{\nu_1+\lambda}\right)\right)/\nu_1' = u_1'/(\nu_1+\lambda)$ is an approximate central chi-square variable divided by its number of degrees of freedom. Show that F' can be written

$$F' = \frac{\nu_1+\lambda}{\nu_1}F\,,$$

where F approximately has a central F-distribution with parameters (ν_1',ν_2), ν_1' being in general fractional.

5.4 LIMITING PROPERTIES - CONNECTION BETWEEN PROBABILITY DISTRIBUTIONS

The connection between the sampling distributions of the present chapter and some of the probability distributions discussed in the previous chapter is illustrated by Fig. 5.5.

Note the central position of the normal distribution as a limiting case of the three sampling distributions (chi-square, F, Student's t) as well *as of the three discrete distributions (multinomial, binomial, Poisson).*

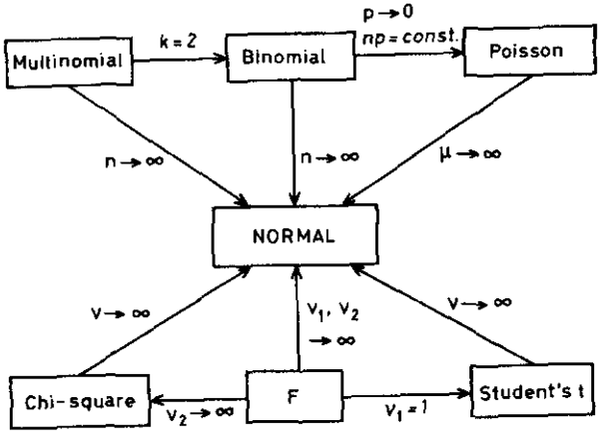

Fig. 5.5. Relations between probability distributions.

Exercise 5.30: Find appropriate positions in Fig. 5.5 of the exponential, the gamma, and the Cauchy distributions.

6. Comparison of experimental data with theory

In the preceding chapters we have investigated features of probability distributions which are frequently used in physics. Experimental findings can, however, not always be directly compared to the ideal mathematical distributions. Quite often a theoretical model will have to be modified in some way before one can make a meaningful comparison between prediction and observation. The reason for this can be that the theoretical p.d.f. will only describe an experiment performed under certain ideal conditions which are not fulfilled in practice; for instance, the lifetime distribution law $f(t;\lambda) = \lambda e^{-\lambda t}$ for $0 \leq t \leq \infty$ assumes that a particle detector of infinite dimensions is available. It may also be necessary to correct for experimental uncertainties and various types of systematic effects.

The purpose of the present chapter, in spite of its rather ambitious headline, is only to outline some of the preparatory work that may be necessary before the experimental data can be used to elicit information on unknown parameters, or agreement checked with other experiments or theoretical models. The parameter estimation problem is discussed quite extensively in Chapters 8-11, whereas a treatment of various tests of goodness-of-fit is postponed to Chapter 14.

6.1 REJECTION OF BAD MEASUREMENTS

It often happens during the accumulation of data that one of the measurements differs substantially from the others. In such a situation one may perhaps suspect that some mistake has occurred, for instance that an erroneous figure has been recorded by the observer, and there should then be no objection to simply discard this single observation.

The situation is frequently not so clear, as when there are several observations which seem to deviate considerably from the majority. It is often a matter of taste to judge which measurements should be kept and which should be

discarded. Several attempts have been made to define objective criteria for the rejection of bad measurements; it appears, however, that such criteria are all more or less arbitrary. Some of the suggested rules will, for instance, allow the rejection of relatively many events in a low-statistics experiment but hardly any events in an experiment with high statistics.

In general, in throwing away suspicious observations one must keep in mind that this rejection should not reduce the reliability of the inferences made on the basis of the remaining sample. In each specific case, therefore, one ought to investigate what might be the reason for the large deviation, such as an erroneous measurement, bias of some type, or just a statistical fluctuation. If this inquiry shows that the inconsistent value is most reasonably ascribed to a "wrong" measurement, it may safely be rejected; otherwise it is probably most correct to keep it.

6.2 EXPERIMENTAL ERRORS ON MEASUREMENTS. THE RESOLUTION FUNCTION

The data with which we have to deal, are subject to observational errors. Due to the uncertainty in the measurements a variable whose true value is x may be observed at some different value x'. Because of the uncertainties inherent in the measuring systems one can even observe events where ideally there should be none. This suggests that before any comparison is made with the data the ideal theoretical distribution for x ought to be adjusted to correspond to some modified or expected distribution for the measurable quantity x'.

A function $r(x';x)$ which describes the distribution of the measurable quantity x' for a given true value x is called the *resolution function* for the variable x. If the true p.d.f. is $f(x;\underline{\theta})$, the p.d.f. for x' is given by

$$f'(x') = \int f(x;\underline{\theta})r(x';x)dx\,. \tag{6.1}$$

In eq.(6.1) the true variable x is integrated over and replaced by the observable variable x'; the original theoretical distribution $f(x;\underline{\theta})$ is "smeared out" by the experimental resolution function into its *resolution transform* $f'(x')$. This "observable p.d.f." will depend on the parameters $\underline{\theta}$ in the original p.d.f. as well as any new constants entering the resolution function.

The resolution function will in general distort the original p.d.f.. However, if the resolution function is strongly varying about the peak value x, so that effectively $r(x';x)$ may be approximated by a delta function $\delta(x'-x)$, eq.(6.1) gives the resolution transform as

$$f'(x') = f(x';\underline{\theta}), \tag{6.2}$$

implying that for an experiment with very good precision the shape of the p.d.f. is unaffected. For the other extreme, if the experimental resolution is very poor, $r(x';x)$ being large also over the regions where the original p.d.f. is small, approximating $f(x';\underline{\theta})$ by a delta function $\delta(x-x_o)$ leads to

$$f'(x') = r(x';x_o). \tag{6.3}$$

Thus the observable distribution is nothing but the resolution function itself.

In many less extreme situations it is possible to give analytical expressions for the resolution function $r(x';x)$ and perform the integration of eq.(6.1) analytically. For instance, it is quite often reasonably assumed that repeated measurements on the true value x would lead to observations x' which are normally distributed about x. Hence the resolution function can be taken as

$$r(x';x) = \frac{1}{\sqrt{2\pi}\,R}\,e^{-\frac{1}{2}\left(\frac{x'-x}{R}\right)^2}, \tag{6.4}$$

where the standard deviation R measures the experimental uncertainty; R is often called the *resolution width*.

In some cases other simple analytical expressions may be used for $r(x';x)$. In other cases the folding-in of the experimental resolution by eq. (6.1) has to be done numerically. We discuss some examples in the following.

6.2.1 Example: Gaussian resolution function and exponential p.d.f.

When the original p.d.f. is exponential,

$$f(x;\lambda) = \lambda e^{-\lambda x}, \qquad 0 \le x \le \infty, \tag{6.5}$$

and the resolution function is of Gaussian form over the range where x is defined, the resolution transform of $f(x;\lambda)$ becomes

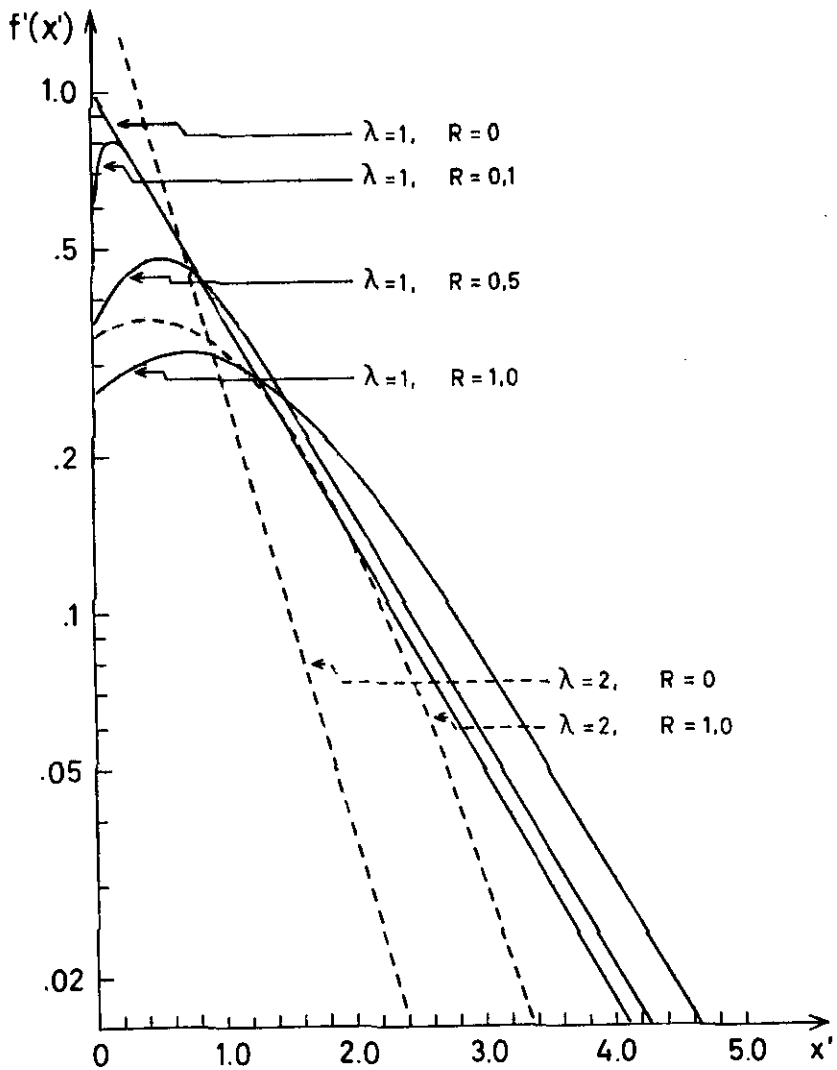

Fig. 6.1. Logarithmic display showing the effect of Gaussian resolution functions on exponential p.d.f.'s for different values of the damping constant λ and the resolution width R. The curves for R = 0 (straight lines) correspond to the unmodified, original p.d.f.'s of eq.(6.5).

$$f'(x') \sim \int_0^\infty \lambda e^{-\lambda x} \cdot e^{-\frac{1}{2}\left(\frac{x'-x}{R}\right)^2} dx .$$

This integral can be evaluated to give the following form,

$$f'(x') \sim e^{\frac{1}{2}(\lambda R)^2} \cdot G\left(\frac{x'}{R} - \lambda R\right) \cdot \lambda e^{-\lambda x'}, \qquad 0 \le x' \le \infty, \qquad (6.6)$$

where G is the cumulative standard normal distribution introduced in Sect.4.8.2 and tabulated in Appendix Table A6.

In practice an ideal behaviour of the form of eq.(6.5) is expected, for example, for particle lifetimes, transverse momenta and four-momentum transfers. The assumption of a normal-shaped resolution function appears reasonable for many experimental set-ups.

Fig. 6.1 illustrates how the original p.d.f. of eq.(6.5) is modified by eq.(6.4) into different observable p.d.f.'s of eq.(6.6) for different numerical values for the constants λ and R.

6.2.2 Example: Gaussian resolution function and Gaussian p.d.f.

Let the original p.d.f. for the variable x be normal with mean value x_o and standard deviation Γ,

$$f(x;x_o,\Gamma) = \frac{1}{\sqrt{2\pi}\,\Gamma} e^{-\frac{1}{2}(x-x_o)^2/\Gamma^2}, \qquad -\infty \le x \le \infty . \qquad (6.7)$$

If the resolution function is also normal of width R over the entire spectrum,

$$r(x';x) = \frac{1}{\sqrt{2\pi}\,R} e^{-\frac{1}{2}(x'-x)^2/R^2},$$

the integration of eq.(6.1) yields for the resolution transform

$$f'(x') = \frac{1}{\sqrt{2\pi(\Gamma^2+R^2)}} e^{-\frac{1}{2}(x'-x_o)^2/(\Gamma^2+R^2)}, \qquad -\infty \le x' \le \infty . \qquad (6.8)$$

Hence, the p.d.f. for the observable x' will also be normal, with the mean value of the true distribution, but with a variance equal to the sum of the variances of the original distribution and the resolution function.

6.2.3 Example: Breit-Wigner resolution function and Breit-Wigner p.d.f.

Suppose that the ideal parameterization is given by a Cauchy, or Breit-Wigner, formula

$$f(x;x_o,\Gamma) = \frac{\Gamma}{\pi}\frac{1}{(x-x_o)^2+\Gamma^2}\,, \qquad -\infty \le x \le \infty. \tag{6.9}$$

For an analytical evaluation of the observable distribution the most convenient form of the resolution function is now another function of the same type, say

$$r(x';x) = \frac{R}{\pi}\frac{1}{(x'-x)^2+R^2}\,.$$

Then eq.(6.1) leads to the resolution transform

$$f'(x') = \frac{\Gamma+R}{\pi}\frac{1}{(x'-x_o)^2+(\Gamma+R)^2}\,, \qquad -\infty \le x' \le \infty. \tag{6.11}$$

In other words, the observable distribution will also be a Breit-Wigner curve, with width equal to the sum of the resonance width and the resolution width.

Exercise 6.1: With a Gaussian resolution function and a Breit-Wigner p.d.f., how would you find the observable p.d.f.?

6.2.4 Example: Width of a resonance

The two preceding examples suggest that for studies of peaks in effective mass spectra it may be convenient to employ the same functional form (Gaussian or Breit-Wigner) for the resolution function and for the p.d.f. describing the physical effect. When this is possible the width of the resulting observable peak Γ_{obs} is simply related to the true resonance width Γ and the resolution width R. Thus, if both the true resonance peak and the resolution function have normal shapes, the width of the resonance is given by

$$\Gamma = \sqrt{\Gamma_{obs}^2 - R^2}\,, \qquad \text{(two Gaussian shapes).} \tag{6.12}$$

For the Breit-Wigner case, the relationship between the widths is linear,

$$\Gamma = \Gamma_{obs} - R, \qquad \text{(two Breit-Wigner shapes).} \tag{6.13}$$

In practice, therefore, the two parameterizations may lead to different estimates of the true resonance width. The difference between the results

will depend on the relative magnitude of the measured width and resolution width, being small when this ratio is large (high resolution). It is recommended to apply both sets of parameterizations; in the case of a poor agreement it may be necessary to study more closely the validity of the assumptions. In any case, rather than using a rough approximation, a superior approach would be to use the experimentally observed shape of the resolution function and perform a numerical integration of eq.(6.1).

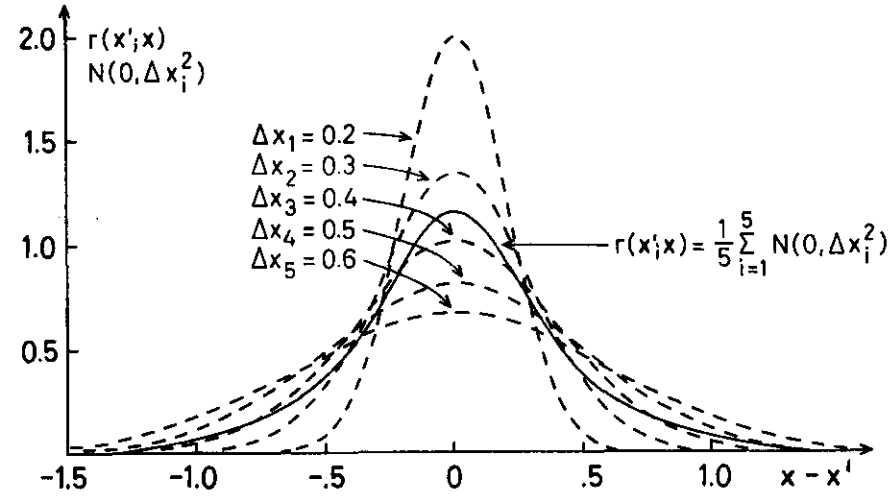

Fig. 6.2. Five measurements of different accuracy are represented by normal p.d.f.'s $N(0,\Delta x_i^2)$ (dashed curves) and correspond to the resolution function r(x';x) shown by the full-drawn curve.

6.2.5 Experimental determination of the resolution function; ideogram

It is often difficult to give a good analytical approximation for the resolution function. This is sometimes the case when the uncertainty in the variable varies from one measurement to another. One may, for instance, think of different bubble chamber events contributing to the same bin of a histogram, where the errors on the individual events can be largely different. In such cases an average resolution function can be determined experimentally by plotting the data in an *ideogram* in the following manner.

Let the measurement x_i have an experimental error Δx_i. One will then assign a normal probability distribution to this measurement, with a standard deviation corresponding to the experimental error. When this is done for all measurements in a certain region of the histogram, and the (normalized) Gaussians are subsequently centered about some common, arbitrarily chosen value, the added contributions give the shape of the experimental resolution function; see illustration by Fig. 6.2.

6.3 SYSTEMATIC EFFECTS. DETECTION EFFICIENCY

In many experiments the detectors used to register the signals do not have the same sensitivity for all types of reactions. The detection efficiency may, for example, depend on the location of the interaction point in the detector, on the emission angle of the particles, on their momenta, *etc*. The experimental bias introduced by imperfect detection ability can for many experiments be very serious, because it leads to loss of information and less reliable conclusions. One therefore tries to design the experiment in such a way that the detection efficiency becomes as high as possible. Since, however, perfect detection can never be achieved in practice due to high costs, time-consumption *etc*., all kinds of possible loss and systematic effects that will distort the data must be checked and estimated.

A probability density function $f(x;\underline{\theta})$ which appears suggestive to describe the phenomenon under study will sometimes be mathematically defined over non-observable values of the physical variable*). This situation can be handled by *truncation* of the p.d.f. in the following way. Let us assume that the observable part of the spectrum of x lies between some definite limits A and B. We then require the p.d.f. to be zero outside these limits and write our new p.d.f. as

$$f'(x;\underline{\theta}) = \frac{f(x;\underline{\theta})}{\int_A^B f(x;\underline{\theta})dx} = \frac{f(x;\underline{\theta})}{F(B) - F(A)}, \qquad A \le x \le B, \qquad (6.14)$$

*) In the remainder of this section we assume that, if necessary, the experimental resolution has already been folded into $f(x;\underline{\theta})$.

where F denotes the cumulative distribution for the ideal p.d.f.. Thus the truncation implies a renormalization of the ideal distribution over the observable region of the variable.

The truncation of a theoretical p.d.f. can formally be regarded as a special case of a more general handling of experimental bias which follows from incomplete detection ability. Detection inefficiency can in principle be handled according to two different approaches which differ in basic philosophy, consisting in, respectively,

(i) modifying the ideal p.d.f. (exact method)

(ii) weighting of the observed events (approximate method)

Essentially, the idea behind the first method is to apply the correction to the theoretical model, leaving the data as they were observed, whereas with the second method one keeps the theoretical model unchanged and adjusts the experimental data.

Let us discuss the treatment according to method (i) in more detail. We shall modify the ideal theoretical description to obtain an observable p.d.f. which in turn can be compared directly with the observations. Although exact, this method may be difficult, if not impossible, to carry out in practice.

Suppose that we make observations on some variable x in order to estimate the parameters of the ideal p.d.f. $f(x;\underline{\theta})$. Because of an imperfect detection apparatus the distribution that can be observed is not $f(x;\underline{\theta})$, but some distorted distribution $f'(x;\underline{\theta})$ which is related to the ideal p.d.f. through the detection efficiency. In general, this efficiency will be dependent on the variable x in which we are interested, as well as on one or more additional variables, y say. The additional variables may also be dependent on x, so that to cover the most general case we write

$$f'(x;\underline{\theta}) = \frac{\int f(x;\underline{\theta})D(x,y)P(y|x)dy}{\iint f(x;\underline{\theta})D(x,y)P(y|x)dydx}. \qquad (6.15)$$

Here $D(x,y)$ is the detection efficiency, and $P(y|x)$ the conditional distribution of y, given x. Since the dependence on y is assumed to be of no interest to our problem this variable is integrated over in the numerator. The integration over y as well as x in the denominator ensures that the distorted p.d.f. $f(x;\theta)$

is correctly normalized over the observable region for x.

The detection efficiency D(x,y), or acceptance, is a property of the detecting system and can in principle be assumed known or measurable. The conditional probability $P(y|x)$, on the other hand, depends on some physical hypothesis that may, or may not be known prior to the experiment. If it is not known *a priori* $P(y|x)$ must be inferred from the data at the expence of the precision in the estimation of the unknown $\underline{\theta}$.

In the general case, with explicit y-dependence for D(x,y) and $P(y|x)$ and with integration limits for y depending on x, the evaluation of the distorted p.d.f. can therefore represent a formidable task. If, however, the detection efficiency should happen to depend on x only, and the limits for the y-integration are independent of x, the observable p.d.f. is easier to find, since eq.(6.15) then reduces to

$$f'(x;\underline{\theta}) = \frac{f(x;\underline{\theta})D(x)}{\int f(x;\underline{\theta})D(x)dx} \quad . \tag{6.16}$$

It is seen that eq.(6.14) for a truncated p.d.f. represents a special case of the last formula.

Method (ii), which is only approximately correct, but perhaps more frequently used in practice, applies correcting weights to the individual observed events, equal to the reciprocal of the detection efficiency. The treatment assumes a subsequent comparison of the distribution of these weighted events with the original p.d.f.. Thus the philosophy is now to adjust the data, rather than the theoretical model. If one event is observed at a particular value x_i of the variable we say that the corrected number of events is w_i, the weight w_i being equal to the inverse of the detection probability for this particular event,

$$w_i = \frac{1}{D(x_i,y_i)} \quad . \tag{6.17}$$

We will see later that the introduction of weighted events gives rise to specific problems which require special attention for the Maximum-Likelihood and the Least-Squares methods for parameter estimation, Sects.9.11 and 10.6, respectively.

In the following sections we shall now give some examples on the application of the procedures outlined above. The three first examples deal with method (i), whereas the fourth applies the weighting method (ii).

6.3.1 Example: Truncation of an exponential distribution

As a first example on the modification of an ideal probability distribution by truncation, let us take the distribution law $f(t;\lambda) = \lambda e^{-\lambda t}$ describing unstable particles with decay constant λ for lifetimes satisfying $0 \leq t \leq \infty$. If the geometry of the apparatus allows the detection of an event only if $t_{min} \leq t \leq t_{max}$ we should, according to eq.(6.14) take for the truncated p.d.f.

$$f'(t;\lambda) = \frac{e^{-\lambda t}}{\int_{t_{min}}^{t_{max}} e^{-\lambda t}dt} = \frac{e^{-\lambda t}}{e^{-\lambda t_{min}} - e^{-\lambda t_{max}}} \quad ,$$

which is properly normalized over the region of detectable flight-times and hence describes the observable events.

6.3.2 Example: Truncation of a Breit-Wigner distribution

A second example where truncation is always used in practice is for the Breit-Wigner parameterization of a resonance $f(M;M_o,\Gamma) = \frac{\Gamma}{\pi}\left((M-M_o)^2+\Gamma^2\right)^{-1}$. When the observations are restricted to a finite mass region $M_A \leq M \leq M_B$, the truncated p.d.f. according to eq.(6.14) is

$$f'(M;M_o,\Gamma) = \frac{\frac{\Gamma}{\pi}\frac{1}{(M-M_o)^2+\Gamma^2}}{\int_{M_A}^{M_B}\frac{\Gamma}{\pi}\frac{dM}{(M-M_o)^2+\Gamma^2}} = \frac{\Gamma}{\tan^{-1}\left(\frac{M_B-M_o}{\Gamma}\right) - \tan^{-1}\left(\frac{M_A-M_o}{\Gamma}\right)}\cdot\frac{1}{(M-M_o)^2+\Gamma^2} \quad .$$

For this p.d.f. the expectation of M is

$$E(M) = \int_{M_A}^{M_B} Mf'(M;M_o,\Gamma)dM = M_o + \frac{\Gamma}{2}\cdot\frac{\ln\left[\left(\left(\frac{M_B-M_o}{\Gamma}\right)^2+1\right)\Big/\left(\left(\frac{M_A-M_o}{\Gamma}\right)^2+1\right)\right]}{\tan^{-1}\left(\frac{M_B-M_o}{\Gamma}\right) - \tan^{-1}\left(\frac{M_A-M_o}{\Gamma}\right)} \quad ,$$

which will be different from M_o, the central value of the resonance peak, unless the interval $[M_A,M_B]$ is symmetric around M_o.

Exercise 6.2: With a symmetric mass interval $[M_o - m, M_o + m]$ around the peak value M_o, show that the variance for the truncated Breit-Wigner p.d.f. is $V(M) = (m/\Gamma)/\arctan(m/\Gamma) - 1$. Note that $\lim_{m\to\infty} V(M) = \infty$; compare Sect.4.11.

6.3.3 Example: Correcting for finite geometry - modifying the p.d.f.

Let us assume that we want to determine some parameters $\underline{\theta}$ related to the spectrum of the momentum p of neutral particles which are observed in a detector through their decays into charged products. For definiteness, let us think of $K^o \to \pi^+\pi^-$ in a bubble chamber. With a limited detection volume the probability to detect and identify a produced K^o will depend on the location of the production point as well as on the momentum and direction of the K^o. Fast particles are more likely to escape the bubble chamber, as will be those whose direction implies a short path length. One will also often expect a lower detection probability for K^o's decaying very close to the production point since such events can be difficult to distinguish from other topologies. To correct for the loss of short-range K^o we require the projected path length to be larger than a certain minimum value. If the K^o proper flight-time follows a distribution law $e^{-t/\tau}$ where τ is the mean K^o lifetime, the detection probability is given by

$$D(p,\lambda,\phi) = e^{-t_{min}/\tau} - e^{-t_{max}/\tau},$$

where t_{min} is the minimal detectable proper flight-time corresponding to the chosen cut on the range, and where t_{max} is the potential flight-time.

Both t_{min} and t_{max} are inversely proportional to the momentum p, the "interesting" variable. They will also involve the "nuisance variables" λ and ϕ giving the direction of the line-of-flight. It is for us necessary to establish a relationship between p and λ,ϕ, which for a given p expresses the distribution of the angles, $P(\lambda,\phi|p)$. Usually this relationship has to be inferred from the same data which we want to use for the estimation of the unknown parameters $\underline{\theta}$. When the dependence $P(\lambda,\phi|p)$ has been established, in a functional form or by a numerical mapping, the observable distribution for the momentum can be found from eq.(6.15), which in this case becomes

$$f'(p,\underline{\theta}) = \frac{1}{N}\int_\Omega f(p,\underline{\theta})\left\{\exp\left(-\frac{t_{min}(p,\lambda,\phi)}{\tau}\right) - \exp\left(-\frac{t_{max}(p,\lambda,\phi)}{\tau}\right)\right\}P(\lambda,\phi|p)\,d\Omega, \quad (6.18)$$

where N is a normalization constant.

6.3.4 Example: Correcting for unobservable events - weighting of the events

Suppose some theory relates the physical constant $\underline{\theta}$ to the production angle for a pion measured relatively to the incoming antiproton direction in the reaction $\bar{p}p \to \pi^+\pi^+\pi^-\pi^-\pi^o$. When observations are made in a bubble chamber the events where one of the charged pions has been emitted along the direction of the magnetic field will often not be successfully analyzed because of an unmeasurable momentum. Such events will therefore largely be missing in the sample. If all events with a charged particle in an "unmeasurable" cone $d\Omega$ around the magnetic field direction are excluded from the sample, the total loss of events can be corrected for using the remaining sample, under the assumption that these events are free from experimental bias.

The reaction $\bar{p}p \to \pi^+\pi^+\pi^-\pi^-\pi^o$ is charge symmetric and has a rotational symmetry around the collision axis. For each observed event we therefore make a rotation of all charged pion directions around this axis and determine the total probability P_π^i that at least one of the tracks will correspond to a laboratory angle within the cone $d\Omega$. The event is then assigned a weight

$$w_i = \frac{1}{1 - P_\pi^i}.$$

When all events from the purified sample are plotted with their individual weights the resulting corrected experimental distribution can be compared with the unmodified theoretical model.

6.4 SUPERIMPOSED PROBABILITY DENSITIES

An experimentally observable quantity often has a possible origin in several physical processes. The overall probability density can then be thought of as a sum of terms where each term gives the contribution from one process;

we write

$$f(x;\underline{\alpha},\underline{\theta}) = \sum_j \alpha_j f_j(x;\underline{\theta}_j),$$

where $f_j(x;\underline{\theta}_j)$ denotes the distribution expected from the j-th contribution alone and α_j the probability for this process. For incoherent processes, $\sum_j \alpha_j = 1$ to keep the overall $f(x;\underline{\alpha},\underline{\theta})$ properly normalized.

In many situations one will be mainly interested in only one or a few of the contributions, and will consider the rest as some type of background effect. Sometimes this background is well understood. Unfortunately, however, the background sources are frequently more or less unknown, so that the mathemathical description of this part of the total amplitude may be rather uncertain. The conclusions that can be drawn about the interesting physical effect will then be correspondingly uncertain. In experimental set-ups one will therefore in general try to keep the background rate, or noise level, as small as possible.

6.4.1 Example: Particle beam with background

A particle beam of mean momentum p_o and standard deviation Δp is of normal shape between the limits p_A and p_B. From outside sources the beam is contaminated by a background of particles uniformly distributed in momentum. The total momentum p.d.f. can then be written

$$f(p;\alpha_1,\alpha_2,p_o,\Delta p) = \alpha_1 e^{-\frac{1}{2}\left(\frac{p-p_o}{\Delta p}\right)^2} + \alpha_2, \qquad p_A \le p \le p_B.$$

Exercise 6.3: In the example above determine the relative magnitude of the beam and the background rates.

6.4.2 Example: Resonance peaks in an effective-mass spectrum

The resonances $\eta(549)$ and $\omega(783)$ are abundantly produced in the reaction $\pi^+ p \to \pi^+ p \pi^+ \pi^- \pi^o$ at intermediate energies and are detected as enhancements in the neutral $\pi^+\pi^-\pi^o$ system. An effective-mass plot for this combination shows nice peaks at about 550 and 780 MeV over a smooth background and may therefore be used to estimate the masses and widths of the η and ω mesons. In terms of the true three-pion effective-mass M the overall p.d.f. must be composed of two

Breit-Wigner parts BW_η and BW_ω describing the resonances, and some function B describing the background. In an obvious notation the ideal p.d.f. is

$$f(M;\underline{\alpha},\underline{\theta}) = \alpha_\eta BW_\eta(M;M_\eta,\Gamma_\eta) + \alpha_\omega BW_\omega(M;M_\omega,\Gamma_\omega) + (1-\alpha_\eta-\alpha_\omega)B(M).$$

If the observations are restricted to a finite mass interval and the experimental resolution taken into account, the overall p.d.f. which should be compared with the observed distribution is

$$f'(M';\underline{\alpha},\underline{\theta}) = \int_{M_A}^{M_B} f(M;\underline{\alpha},\underline{\theta}) r(M';M)\,dM \Big/ \int_{M_A}^{M_B}\int_{M_A}^{M_B} f(M;\underline{\alpha},\underline{\theta}) r(M';M)\,dM dM'.$$

7. Statistical inference from normal samples

The previous chapters of this book have mainly dealt with probability theory. In particular, Chapters 3,4 and 5 were devoted to supply the reader with some general background as well as detailed knowledge on specific probability distributions, which either describe physical phenomena directly, or turn out to be useful for the handling and interpretation of experimental data. In Chapter 6 we saw how it might be necessary to modify the ideal theoretical distributions to prepare the ground for a later comparison between theory and experiment.

We shall now make the transition from probability theory to the domain of statistics. We do this by discussing in this chapter some simple examples on statistical inference about the parameters in the normal probability distribution. The normal p.d.f. is assumed to give an adequate description of some population, or universe, for instance the outcomes of infinitely many measurements on the same physical quantity. The numerical values of the parameters in the p.d.f. may, however, not be completely known. From a restricted number of observations, assumed to be representative for the universe, it is possible to make inferences about the unknown quantities. We will build our examples around the notion of a *confidence interval*, which we shall meet again later in Chapters 9-11 when we come to the different methods which are generally applicable for estimating unknown parameters.

We begin this chapter by giving first in Sect.7.1 a set of formal definitions and some general remarks on the problem of making inferences about an underlying population from a given sample. From Sect.7.2 onwards we shall specifically assume that the sample originates from a normal parent population.

7.1 DEFINITIONS

Let $x_1,x_2,\ldots,x_n$ be a random sample from a population with a probability density function which depends on a parameter θ which is not known but

whose numerical value we want to estimate from the sample. If $t=t(x_1,x_2,\ldots,x_n)$ is a function of the sample variables which does not depend on any unknown parameters, we call t a *statistic*; we assume here that t has some correspondence to θ*). Let f(t) be the probability density function for the variable t. We will assume that, according to a given prescription, it is possible to determine two values t_a and t_b such that the integral of f(t) between the limits t_a and t_b is equal to some fixed number γ, $0 \leq \gamma \leq 1$. If, for the parameter θ, the probability is such that

$$P(t_a \leq \theta \leq t_b) = \gamma, \tag{7.1}$$

the closed interval $[t_a,t_b]$ is called a *100γ % confidence interval for* θ. The number γ is called the *confidence coefficient*, and the numbers t_a and t_b the *confidence limits*. See Fig. 7.1.

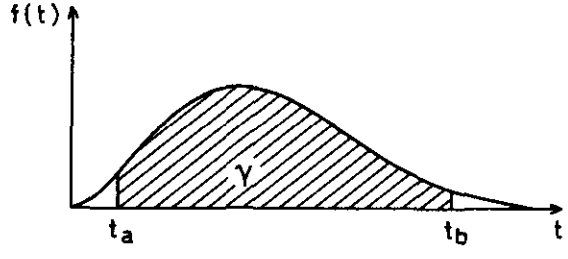

Fig. 7.1. Illustration of the concepts of confidence coefficient and confidence limits.

If we are given one particular sample of size n, say n measurements, the limits t_a and t_b are definite numbers which can be calculated from the measurements. From this particular sample, therefore, a certain interval

*) We shall in Chapter 8 identify t as an estimator for the parameter θ, or, more generally, an estimator for some function of θ.

$[t_a,t_b]$ is deduced, and this interval will either include the true value θ or it will not. A second sample, also of size n, will usually lead to a different interval, which may, or may not include the parameter θ. The relevant statement for one particular sample is therefore, either

$$P(t_a \leq \theta \leq t_b) = 1, \qquad \text{if } \theta \text{ is within } [t_a,t_b],$$

or

$$P(t_a \leq \theta \leq t_b) = 0, \qquad \text{if } \theta \text{ is not within } [t_a,t_b].$$

The meaning of the probability statement eq.(7.1) is the following: There is a probability γ that the *random interval* $[t_a,t_b]$ will cover the true value θ. If a large number of samples of size n are examined, that is, if the experiment is repeated many times under the same conditions, then θ will be included in the calculated intervals $[t_a,t_b]$ in 100γ % of these experiments. In other words, it is expected that, in the long run, the calculated limits t_a and t_b are such that the statement

$$t_a \leq \theta \leq t_b \tag{7.2}$$

will be true in 100γ % of the cases. Thus the confidence coefficient reflects the reliability one may attach to the inequality statement (7.2).

Whenever it is desired to make a statement about an unknown parameter θ in terms of a confidence interval one is confronted with a dilemma. Choosing a wide interval corresponds to having a large probability that the unknown parameter indeed does belong to the interval, but a rather vague statement is then expressed about the parameter itself. On the other hand, giving a narrow interval would imply a more precise specification of the parameter, but then our statement is less likely to be true. It appears that in this situation most people prefer the former alternative and make their assertions taking γ, the confidence coefficient, as a number in the neighbourhood of 1. Common choices are 0.90, 0.95, 0.99.

7.2 CONFIDENCE INTERVALS FOR THE MEAN

In many situations it is reasonable to assume that the result of a measurement of some quantity μ is a random variable x that has a normal distribution about the mean value μ. We will now assume that μ is an unknown quantity, about which we are supposed to say something on the basis of a series of n independent measurements, made under the same conditions. In other words, we are to make an inference about μ from a random sample of size n from the normal population $N(\mu,\sigma^2)$. Here the variance σ^2 gives a measure of the accuracy in the measurements, and we must treat separately the two conceivable cases, when σ^2 is a known number (Sect.7.2.1), and when σ^2 is unknown (Sect.7.2.2).

7.2.1 Case with σ^2 known

Our starting point is a random sample of size n from a normal population $N(\mu,\sigma^2)$, for instance n independent observations $x_1,x_2,\ldots,x_n$ of known error σ. When inferences are to be made about the unknown population mean μ it is suggestive to consider the sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$; this statistic is distributed as $N(\mu,\sigma^2/n)$, and as we have seen repeatedly, (see for instance Sect. 4.8.6), the variable

$$\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

has a distribution which is standard normal N(0,1). We can therefore apply the considerations of Sect.4.8.3 concerning the probability contents of a normal distribution. For instance, a probability content equivalent to ±2 standard deviations is implied by writing

$$P(-2 \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq 2) = \int_{-2}^{2} g(y)dy = 0.954\,, \tag{7.3}$$

where g(y) is the standard normal p.d.f. of eq.(4.63). Eq.(7.3) is a probability statement, expressing that the probability is 0.954 that the random variable $y = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ will have some value between -2 and +2.

We can rewrite the probability statement in the form

$$P\left(\bar{x} - 2\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + 2\frac{\sigma}{\sqrt{n}}\right) = 0.954. \qquad (7.4)$$

This relation apparently considers μ as a variable and $\bar{x} - 2\frac{\sigma}{\sqrt{n}}$, $\bar{x} + 2\frac{\sigma}{\sqrt{n}}$ as two numbers defining an interval. We realize, however, that since $\bar{x}$ is a random variable, the quantities $\bar{x} - 2\frac{\sigma}{\sqrt{n}}$ and $\bar{x} + 2\frac{\sigma}{\sqrt{n}}$ are also random variables; hence it is justified to call the interval $[\bar{x} - 2\frac{\sigma}{\sqrt{n}}, \bar{x} + 2\frac{\sigma}{\sqrt{n}}]$ a *random interval*. We may read eq.(7.4) as a probability statement about μ: Prior to the repeated, independent measurements there is a probability 0.954 that the random interval $[\bar{x} - 2\frac{\sigma}{\sqrt{n}}, \bar{x} + 2\frac{\sigma}{\sqrt{n}}]$ *will include* the unknown, but fixed value μ. Other probability statements can of course be written taking other intervals corresponding to other probabilities. The point is, that all statements of this sort, which can be made before any measurements are actually performed, belong to *probability theory*. Generally we may write

$$P\left(a \le \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \le b\right) = \int_a^b g(y)dy = \gamma, \qquad (7.5)$$

where the limits a and b for a given γ can be found from Appendix Table A6.

As soon as the measured numbers are at hand and we are given a particular set of n observations $x_1, x_2, \ldots, x_n$, we may pass to the domain of *statistics*, and make inferences about the unknown μ on the basis of the observations.

For definiteness, let us establish a 95.4% confidence interval for the mean μ in $N(\mu, 3^2)$, given that four independent measurements, with a known, common error σ=3, have led to the numbers

2.2, 4.3, 1.7, 6.6.

The sample mean is

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 3.7.$$

The confidence limits corresponding to a symmetric 95.4% confidence interval for μ are then given by

$$\bar{x} - 2\frac{\sigma}{\sqrt{n}} = 3.7 - 2\frac{3}{\sqrt{4}} = 0.7$$

and

$$\bar{x} + 2\frac{\sigma}{\sqrt{n}} = 3.7 + 2\frac{3}{\sqrt{4}} = 6.7.$$

Therefore, an inference to be made from the measurements is that the symmetric 95.4% confidence interval for the mean μ is the interval [0.7, 6.7].

It should be clear from the reasoning above that it is essential that σ is a known number. If σ were not known, the confidence limits $(\bar{x} - 2\frac{\sigma}{\sqrt{n}})$ and $(\bar{x} + 2\frac{\sigma}{\sqrt{n}})$ could not have been calculated from the measurements, and hence no inference could have been made about μ based on N(0,1).

In practice the situation is often that the error on the measurements is not known exactly. However, the size of the sample may sometimes be sufficiently large to allow the approximation of σ^2 by the observed sample variance s^2, and the procedure above can be applied to find confidence intervals for μ. If σ^2 is not known, and the sample size is small ($n \lesssim 20$), the procedure of the subsequent section should be used.

Exercise 7.1: For the numerical example given in the text, what is the symmetric 90% confidence interval for μ?

Exercise 7.2: Given 6 independent measurements 10.7, 9.7, 13.3, 10.2, 8.9, 11.6 of known error σ=2. Assuming a normal sample, find symmetric confidence intervals for μ corresponding to (a) γ=0.90, (b) γ=0.95, (c) γ=0.99.

Exercise 7.3: Given that a normal distribution has variance σ^2, what is the sample size needed if the symmetric 95.4% confidence interval for μ shall have a length equal to (a) σ, (b) $\sigma/2$?

Exercise 7.4: Measurements on the momentum of monoenergetic beam tracks on bubble chamber pictures have led to the following sequence of numbers in units of GeV/c: 18.87, 19.55, 19.32, 18.70, 19.41, 19.37, 18.84, 19.40, 18.78, 18.76. We assume that this sample of size 10 originates from a normal distribution.
If the measuring machine has a known accuracy corresponding to an uncertainty of 300 MeV/c in the momentum determination, find a 95% confidence interval for the beam momentum.

7.2.2 Case with σ^2 unknown

We turn to the problem of finding a confidence interval for the mean μ of a normal distribution when we are not so fortunate as to know the variance σ^2.

The tools needed to handle this situation have in fact already been

provided by the remark of Sect.5.2.1. We have seen that if $x_1, x_2, \ldots, x_n$ is a random sample from $N(\mu,\sigma^2)$ two variables can be formed, which have wellknown properties, namely

$$\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \quad \text{is } N(0,1), \qquad \text{where } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

$$\frac{(n-1)\,s^2}{\sigma^2} \quad \text{is } \chi^2(n-1), \qquad \text{where } s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2,$$

and these variables are independent. Therefore the variable

$$t = \frac{\dfrac{\bar{x}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\dfrac{(n-1)s^2}{\sigma^2}\Big/(n-1)}} = \frac{\bar{x}-\mu}{s/\sqrt{n}} \tag{7.6}$$

is a Student's t-variable with (n-1) degrees of freedom. We note that from the construction of t, the unknown parameter σ^2 drops out, and we are left with a variable which has only μ as an unknown constituent. It is also worth comparing with the previous case: With σ^2 known, the variable constructed was $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$, which is distributed as N(0,1); in the present case where σ^2 is assumed unknown, the variable needed is $\frac{\bar{x}-\mu}{s/\sqrt{n}}$, which has a Student's t-distribution with (n-1) degrees of freedom.

For the variable constructed by eq.(7.6) we may write down probability statements analogous to eq.(7.5),

$$P\left(a \le \frac{\bar{x}-\mu}{s/\sqrt{n}} \le b\right) = \int_a^b f(t;n-1)\,dt = \gamma, \tag{7.7}$$

where f(t;n-1) is the Student's t probability density function for (n-1) degrees of freedom, given by eq.(5.15). Since f(t;n-1) has symmetry about t=0 it is customary to choose intervals [a,b] which are symmetric. Values for b in the relation

$$P\left(-b \le \frac{\bar{x}-\mu}{s/\sqrt{n}} \le b\right) = \int_{-b}^{b} f(t;n-1)\,dt = \gamma \tag{7.8}$$

can be deduced from Appendix Table A7 for different number of degrees of freedom and for the usually chosen values of the confidence coefficient γ, (compare Exercise 5.17).

Rewriting the argument in the left-hand side of eq.(7.8) gives a probability statement about the unknown μ,

$$P\left(\bar{x} - b\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + b\frac{s}{\sqrt{n}}\right) = \gamma, \tag{7.9}$$

which may be compared to eq.(7.4), valid in the previous case when σ^2 was known. For a specified value of γ the corresponding value of b will be dependent on the number of degrees of freedom. The size of the random interval $[\bar{x} - b\frac{s}{\sqrt{n}}, \bar{x} + b\frac{s}{\sqrt{n}}]$ for a given γ is large for very small values of (n-1), but approaches the size of the corresponding intervals in N(0,1) when the number of degrees of freedom becomes large. This is so because the Student's t-distribution has N(0,1) as a limiting distribution when $n \to \infty$; (compare Sect.5.2.3).

For illustration, let us return to the numerical example of the previous section, with the measurements 2.2, 4.3, 1.7, 6.6 from $N(\mu,\sigma^2)$, where now μ as well as σ^2 are unknown. We calculate

$$\bar{x} = 3.70,$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 = 5.01.$$

Searching a confidence interval which can be compared to the symmetric 95.4% (or ±2 standard deviation) confidence interval derived for the case when σ^2 was known, we observe that Appendix Table A7 has entries corresponding to probability contents of 0.025 in the tails of the Student's t-distribution. For 3 degrees of freedom, we find b = 3.182, and the confidence limits are given by the numbers

$$\bar{x} - b\frac{s}{\sqrt{n}} = 3.70 - 3.182\,\frac{\sqrt{5.01}}{\sqrt{4}} = 0.14,$$

$$\bar{x} + b\frac{s}{\sqrt{n}} = 3.70 + 3.182\,\frac{\sqrt{5.01}}{\sqrt{4}} = 7.26.$$

Thus the symmetric 95% confidence interval for μ obtained from the four measurements of unknown experimental precision is the interval [0.14, 7.26]. Notice that this interval is larger than the corresponding 95.4% confidence interval [0.7, 6.7] obtained in the previous example when σ^2 was assumed known.

Exercise 7.5: For the numerical example above, what is the symmetric 90% confidence interval for μ? Compare this result with that of Exercise 7.1.

Exercise 7.6: Six independent observations from a population $N(\mu,\sigma^2)$ are given by the numbers 10.7, 9.7, 13.3, 10.2, 8.9, 11.6. With σ^2 unknown, find symmetric confidence intervals for μ corresponding to (a) $\gamma = 0.90$, (b) $\gamma = 0.95$, (c) $\gamma = 0.99$. (Compare Exercise 7.2.)

Exercise 7.7: With the observations of Exercise 7.4, what is the symmetric 95% confidence interval for the beam momentum if the accuracy of the measuring instrument is not known prior to the measurements?

7.3 CONFIDENCE INTERVALS FOR THE VARIANCE

As before we will assume that $x_1,x_2,\ldots,x_n$ is a random sample from $N(\mu,\sigma^2)$, but now we want to discuss how we can find confidence intervals for σ^2 and thereby make inferences about this parameter. Again we must treat separately two cases, first assuming μ to be known (Sect.7.3.1), and next assuming μ unknown (Sect.7.3.2).

7.3.1 Case with μ known

This may correspond to an experimental situation where repeated measurements are performed on a known quantity μ using a measuring device of unknown precision.

A statistic which has correspondence to the variance σ^2 is the sum $\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$. As we have seen before (Sect.5.1.1) the variable

$$\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2$$

will be distributed as $\chi^2(n)$. For the chi-square distribution with n degrees of freedom it is possible to find two numbers, a and b, such that, for $0\le\gamma\le1$,

$$P\left\{a \le \sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2 \le b\right\} = \int_a^b f(u;n)du = \gamma. \qquad (7.10)$$

Here f(u;n) is the chi-square p.d.f. for n degrees of freedom, given by eq. (5.5). The probability statement can be rewritten as

$$P\left\{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{b} \le \sigma^2 \le \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{a}\right\} = \gamma. \qquad (7.11)$$

For a chosen value of γ there is an infinite number of possible choices for the integration limits a and b for the skew chi-square p.d.f.. It is customary to take the limits such that the two tails below a and above b will correspond to equal probabilities $\frac{1}{2}(1-\gamma)$. Calculations of a and b for given γ and given number of degrees of freedom can then be done in the ordinary manner, using Appendix Table A8.

Let us again take an example. Suppose that it is requested to say something about the accuracy of a new measuring instrument, and for this purpose a calibrated length is measured several times. The outcomes from 10 independent measurements are the numbers

1002, 1000, 997, 1001, 1001, 999, 998, 999, 1000, 1003,

and the true number, μ, is 1000.

If we demand a 95% confidence interval for σ^2, Appendix Table A8 shows that for 10 degrees of freedom the integration limits in eq.(7.10) will correspond to equal probabilities (=0.025) in the two tails of the chi-square p.d.f. provided that we take a=3.247 and b=20.483. The measurements give the squared deviations about the known mean μ as $\sum_{i=1}^{10}(x_i-\mu)^2=30$. Hence an inference from the measurements is that the 95% confidence interval for the variance σ^2 is given by

$$\left[\frac{\sum_{i=1}^{10}(x_i-\mu)^2}{20.483}, \quad \frac{\sum_{i=1}^{10}(x_i-\mu)^2}{3.247}\right] = [1.46, \ 9.23]. \qquad (7.12)$$

It may be noted that in this case with an unsymmetric p.d.f. the interval constructed by letting the tails represent equal probabilities does not correspond to the shortest possible confidence interval for a given γ. The required computation to obtain a *minimal* confidence interval, given γ, is so tedious that it is rarely done in practice.

Exercise 7.8: For the example in the text determine confidence intervals for σ^2 corresponding to (a) $\gamma = 0.90$, (b) $\gamma = 0.99$.

Exercise 7.9: Let 7.3, 6.6, 7.0, 5.1, 7.1, 8.5, 5.9, 6.5, 6.2 be 9 independent measurements from an assumed normal population $N(7,\sigma^2)$. On the basis of these

observations, make inferences about the variance σ^2 corresponding to (a) $\gamma = 0.90$, (b) $\gamma = 0.95$, (c) $\gamma = 0.99$.

7.3.2 Case with μ unknown

When we search a confidence interval for the variance σ^2 but do not know the mean value μ of the population $N(\mu,\sigma^2)$ the sample $x_1,x_2,\ldots,x_n$ provides the arithmetic mean $\bar{x}$, and we can make use of the fact (compare Sect. 5.1.6) that the variable

$$\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sigma}\right)^2$$

is distributed as $\chi^2(n-1)$. Thus the reasoning of the preceding section can be applied to the inference problem, and we may still use the chi-square distribution to obtain confidence intervals for σ^2. However, whereas in the previous case when μ was known we used $\chi^2(n)$, the present case with μ unknown requires $\chi^2(n-1)$.

Instead of eq.(7.10) we shall now have

$$P\left\{a \le \sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sigma}\right)^2 \le b\right\} = \int_a^b f(u;n-1)\,du = \gamma\,, \tag{7.13}$$

where $f(u;n-1)$ is the chi-square p.d.f. for $n-1$ degrees of freedom. For a specified confidence coefficient γ the limits a,b can be determined in the usual manner, entering Appendix Table A8 for $n-1$ degrees of freedom. The probability statement for σ^2, analogous to eq.(7.11), reads

$$P\left\{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{b} \le \sigma^2 \le \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{a}\right\} = \gamma. \tag{7.14}$$

For the numerical example of the previous section the 10 measurements give

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 1000\,, \qquad \sum_{i=1}^{n}(x_i-\bar{x})^2 = 30.$$

If we therefore did not know what the true value μ were, the 95% confidence interval for σ^2, obtained for $(10-1) = 9$ degrees of freedom would be (see Appendix Table A8)

$$\left[\frac{\sum_{i=1}^{10}(x_i-\bar{x})^2}{19.023}\,,\quad \frac{\sum_{i=1}^{10}(x_i-\bar{x})^2}{2.700}\right] = [1.58,\ 11.11]. \tag{7.15}$$

We see, by comparison with eq.(7.12) of the preceding section, that the 95% confidence interval for σ^2 becomes wider when the number of degrees of freedom is reduced from 10 to 9. That is, our knowledge about σ^2 has become less precise in the present case because we also had to adopt the sample mean $\bar{x}$ as an estimate for the unknown population mean μ.

Exercise 7.10: For the example in the text deduce confidence intervals for σ^2 corresponding to (a) $\gamma = 0.90$, (b) $\gamma = 0.99$. (Compare Exercise 7.8.)

Exercise 7.11: If 7.3, 6.6, 7.0, 5.1, 7.1, 8.5, 5.9, 6.5, 6.2 are independent observations from $N(\mu,\sigma^2)$, where μ is unknown, find confidence intervals for σ^2 corresponding to (a) $\gamma = 0.90$, (b) $\gamma = 0.95$, (c) $\gamma = 0.99$. (Compare Exercise 7.9.)

Exercise 7.12: If, in Exercise 7.4, the momentum of the beam particles was known to be 19.08 GeV/c, but the measuring device had an unknown accuracy, find a 90% confidence interval for the error. How would you determine a 68% confidence interval for the error?

7.4 CONFIDENCE REGIONS FOR THE MEAN AND VARIANCE

Suppose we are to give a joint confidence region for the mean and the variance in $N(\mu,\sigma^2)$ on the basis of the sample $x_1,x_2,\ldots,x_n$. To do this we use the fact that for normal samples, the variables $\bar{x}$ and s^2 are independent (Sect. 4.8.6). If, for example, a 95% confidence region is desired we can write two probability statements as

$$P\left\{-a \le \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \le a\right\} = \sqrt{0.95}\,, \tag{7.16}$$

$$P\left\{b \le \sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sigma}\right)^2 \le b'\right\} = \sqrt{0.95}\,, \tag{7.17}$$

and determine the limits a and b,b' from $N(0,1)$ and $\chi^2(n-1)$, respectively. For the independent variables $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ and $\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sigma}\right)^2$ a joint probability

statement is obtained by multiplying the probabilities of eqs.(7.16), (7.17), giving

$$P\left(-a \le \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \le a, \quad b \le \sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma}\right)^2 \le b'\right) = 0.95. \tag{7.18}$$

The inequalities in eq.(7.18) determine a region in the parameter space which is indicated by the shaded area in Fig. 7.2. The region is bounded

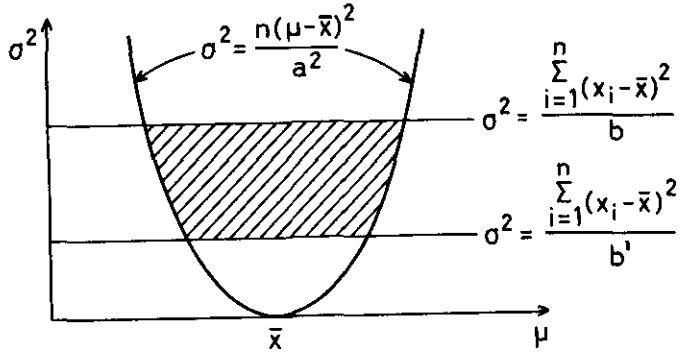

Fig. 7.2. Confidence region (shaded) for the mean μ and variance σ^2 of $N(\mu,\sigma^2)$.

by the two straight lines $\sigma^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2/b'$ and $\sigma^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2/b$, respectively, and the parabola expressing the dependence between σ^2 and μ, $\sigma^2 = n(\mu - \bar{x})^2/a^2$. See further Sect.9.7.5.

8. Estimation of parameters

The general problem of parameter estimation may be sketched as follows:

From a restricted number of observations, assumed to constitute a *random sample*, one wants to gain some knowledge about the underlying population or universe from which the sample emanated. The mathematical form of the parent distribution may be well-defined, but it involves a certain number of parameters, whose numerical values are not known. The measurements should therefore be used to extract the largest possible amount of information about the parameters, specifically, the experimental sample should supply numbers which could be said to represent the numerical values of the parameters.

We have already in Chapter 7 seen examples on *interval estimation* of the parameters in the normal distribution. In the forthcoming chapters we shall study different methods for estimating unknown parameters in general probability distributions. These methods (the Maximum-Likelihood-, Least-Squares- and moments methods) all produce *point estimates*, that is, some definite numbers for the parameters. The methods also give measures for the uncertainties in the estimates and thereby express the confidence that can be attached to the numerical results.

In this chapter we shall take up various general aspects of parameter estimation by discussing in some detail a few of the criteria that should be fulfilled by good and acceptable estimators. Although these criteria will be applied to the specific point estimation methods described in Chapters 9-11 the discussion in the following sections will mainly be of a formal nature. The student who requires just a superficial knowledge of these rather theoretical features may therefore be satisfied with reading only the first two sections of this chapter.

8.1 DEFINITIONS

The term *estimator* denotes in the following a function of the observations, or the method or prescription used to find a value for an unknown parameter. By an *estimate* we mean the numerical value of the parameter obtained with the estimator for a particular set of observations. If the parameter is θ, its estimate is denoted by $\hat{\theta}$. The term *statistic* was introduced in Sect. 7.1 as a function of one or more random variables that does not depend on any unknown parameters. In the general case we will let the statistic $t = t(x_1,x_2,\ldots,x_n)$ be an estimator of the unknown θ or of some function of θ.

If a continuous or discrete population has the probability distribution $f(x;\theta)$, the *likelihood* of the observations $x_1,x_2,\ldots,x_n$ for a specific θ is given by

$$L(x_1,x_2,\ldots,x_n|\theta) = \prod_{i=1}^{n} f(x_i;\theta). \tag{8.1}$$

The product expresses the joint conditional probability for obtaining the measurements $x_1,x_2,\ldots,x_n$, given θ. We shall later also think of $L(x_1,x_2,\ldots,x_n|\theta)$ as a function of θ, calling it a *likelihood function*. We will use the symbol L or $L(\underline{x}|\theta)$ to denote the likelihood (*i.e.* the number expressing the joint probability) as well as the likelihood function (*i.e.* the function of θ). A third interpretation regards L as a joint probability density function of observable quantities x_i, for a given θ. Hopefully, in the following, the precise meaning of the symbol L, or $L(\underline{x}|\theta)$, will be clear from the context in which it appears.

8.2 PROPERTIES OF ESTIMATORS

Observations are random variables. Any function of the observations will also be a random variable, which may take on a variety of values. If a statistic $t = t(x_1,x_2,\ldots,x_n)$ is used as an estimator for the parameter θ it will therefore give rise to a distribution of estimates $\hat{\theta}$. The individual estimates obtained are of less interest than their overall distribution, because this distribution will reflect the quality of the estimator when it is used many times. We will therefore judge the merits of an estimator from the characteristics of the distribution of its estimates.

A good estimator should in the long run produce estimates which do not systematically deviate from the true parameter value, and its accuracy should increase with the number of observations. Frequently there are several estimators which fulfil these requirements and hence can reasonably be thought of for estimating an unknown parameter. If so, one estimator can be said to be superior to the others if its distribution of estimates shows the best "concentration" about the true parameter value. "Concentration" may for this purpose be expressed by giving the variance as a measure of the spread of the distribution about its central value.

In the forthcoming sections we will discuss the following optimum properties that are desired for good estimators: consistency, unbiassedness, minimum variance, efficiency and sufficiency. Only rarely will the conceivable estimators for a parameter possess all the good properties. One may therefore have to choose between them, and in each specific case decide which of the ideal properties that can be abandoned, taking also practical considerations into account.

8.3 CONSISTENCY

It is intuitively clear that a desirable property of an estimator is that its estimates converge towards the true parameter value when the number of observations is increased. Such a property is called *consistency*.

Mathematically, the estimator t can be said to be consistent if it *converges in probability* to θ. If the estimate θ_n is obtained from a sample of size n, then, given any positive ε and η, an N should exist such that

$$P(|\theta_n - \theta| > \varepsilon) < \eta \tag{8.2}$$

for all $n > N$. In words, the estimator is consistent if, given any small quantity ε, we can find a sample size N such that, for all larger samples, the probability that θ_n differs from the true value by more than ε is arbitrarily close to zero.

As an example, we know from the Law of Large Numbers (Sect.3.10.4) that the arithmetic mean of a sample of n measurements from a population with mean μ and finite variance will converge towards μ as n becomes large,

$$t = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \to \mu \qquad \text{when } n \to \infty .$$

The sample mean is therefore a consistent estimator of the population mean.

Exercise 8.1: Show explicitly that the mean $\bar{x}$ of a sample of size n from the normal population $N(\mu,\sigma^2)$ converges in probability to μ.

8.4 UNBIASSEDNESS

Consistency is a property that describes the behaviour of an estimator when the sample size increases to infinity, but says nothing about its behaviour for finite data sets. For example, in the last section we just saw that the sample mean $\bar{x}$ is a consistent estimator of the population mean μ, but so would also be

$$t' = \bar{x}' = \frac{1}{n-a}\sum_{i=1}^{n} x_i ,$$

where a is any fixed number. Why do we prefer one to the other?

Unbiassedness is an estimator property defined for finite sets of observations. This property is possessed by estimators whose estimates are not systematically shifted from the true parameter value but centered around this value for all sample sizes. As a measure of the centre of a distribution one can use the conventional mean value. Mathematically, the property unbiassedness is therefore defined if, for all sample sizes n, the expectation value of the estimator t is equal to the true parameter value θ,

$$E(t) = \theta . \tag{8.3}$$

If

$$E(t) = \theta + b \tag{8.4}$$

and b is different from zero, the estimator is *biassed*. The bias term $b = b(\theta)$ will for all reasonable estimators be of order $\frac{1}{n}$ or smaller compared to θ.

It is rather trivial to see that the sample mean $\bar{x}$ is an unbiassed estimator of the population mean μ whenever the latter exists. It is also seen that the estimator $\bar{x}'$ above will be a biassed estimator of μ for all a different from zero.

Consiatency and unbiassedness are independent estimator qualities, as neither property implies the other. It is generally accepted that consistency is more important than unbiassedness, partly because bias can often be corrected for. A consistent estimator whose asymptotic distribution has a finite mean will always be asymptotically unbiassed.

8.4.1 Example: s^2 as an estimator of σ^2

We will show that the statistic $s^2 = \frac{1}{n-1}\sum_i (x_i-\bar{x})^2$, the sample variance, is an unbiassed estimator of the population variance σ^2. We write

$$s^2 = \frac{1}{n-1}\sum_i (x_i-\bar{x})^2 = \frac{1}{n-1}\sum_i \left((x_i-\mu) - (\bar{x}-\mu)\right)^2$$

$$= \frac{1}{n-1}\sum_i \left((x_i-\mu) - \frac{1}{n}\sum_j (x_j-\mu)\right)^2 ,$$

where μ is the population mean. Remembering the fact that the different x_i are independent we get for the expectation value

$$E(s^2) = E\left\{\frac{1}{n-1}\sum_i (x_i-\mu)^2\right\} - E\left\{\frac{1}{(n-1)n}\sum_j (x_j-\mu)^2\right\}$$

$$= \frac{1}{n-1}\sum_i E(x_i-\mu)^2 - \frac{1}{(n-1)n}\sum_j E(x_j-\mu)^2 = \frac{1}{n-1}\cdot n\sigma^2 - \frac{1}{(n-1)n}\cdot n\sigma^2 .$$

Hence

$$E(s^2) = \sigma^2 ,$$

which should be compared to eq.(8.3). It is seen that the presence of the factor $\frac{1}{n-1}$ instead of the more intuitive $\frac{1}{n}$ in the definition ensures that the sample variance s^2 is an unbiassed estimator of σ^2.

The statistic $S^2 = \frac{1}{n}\sum_i (x_i-\bar{x})^2$ is a biassed estimator of σ^2, because

$$E(S^2) = E\left\{\frac{n-1}{n}s^2\right\} = \left(1 - \frac{1}{n}\right)\sigma^2 \neq \sigma^2 .$$

We see that the bias is $b(\sigma^2) = -\frac{1}{n}\sigma^2$, which decreases with increasing sample size.

8.4.2 Example: Estimator of the third central moment

This example is presented to illustrate how one from the first intuitive guess can construct the correct form of an unbiassed estimator.

Guided by the preceding example we consider the following sum,

$$\sum_i (x_i-\bar{x})^3 = \sum_i \left((x_i-\mu) - (\bar{x}-\mu)\right)^3$$

$$= \sum_i (x_i-\mu)^3 - 3\sum_i (x_i-\mu)^2(\bar{x}-\mu) + 3\sum_i (x_i-\mu)(\bar{x}-\mu)^2 - \sum_i (\bar{x}-\mu)^3.$$

Recalling the definition of the third central moment μ_3 (Sect.3.3.3) and using the independence of the x_i one finds for the expectation of the different parts,

$$E\left(\sum_i (x_i-\mu)^3\right) = \sum_i E(x_i-\mu)^3 = n\mu_3,$$

$$E\left(-3\sum_i (x_i-\mu)^2(\bar{x}-\mu)\right) = -3E\left[\left(\sum_i (x_i-\mu)^2\right)\left(\frac{1}{n}\sum_j (x_j-\mu)\right)\right] = -3\mu_3,$$

$$E\left(3\sum_i (x_i-\mu)(\bar{x}-\mu)^2\right) = 3E\left[\left(\sum_i (x_i-\mu)\right)\left(\frac{1}{n}\sum_j (x_j-\mu)\right)\left(\frac{1}{n}\sum_k (x_k-\mu)\right)\right] = \frac{3}{n}\mu_3,$$

$$E\left(-\sum_i (\bar{x}-\mu)^3\right) = -E\left[\sum_i \left(\frac{1}{n}\sum_j (x_j-\mu)\right)\left(\frac{1}{n}\sum_k (x_k-\mu)\right)\left(\frac{1}{n}\sum_\ell (x_\ell-\mu)\right)\right] = -\frac{1}{n}\mu_3.$$

Collecting terms lead to

$$E\left(\sum_i (x_i-\bar{x})^3\right) = n\mu_3 - 3\mu_3 + \frac{3}{n}\mu_3 - \frac{1}{n}\mu_3 = \frac{(n-1)(n-2)}{n}\mu_3.$$

By comparison with eq.(8.3) it is seen that an unbiassed estimator of μ_3 is

$$t = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} (x_i-\bar{x})^3.$$

Exercise 8.2: In the binomial distribution $B(r;n,p)=\binom{n}{r}p^r(1-p)^{n-r}$, *where* $r=0,1,\ldots,n$, *show that the unbiassed estimators of p and* p^2 *are, respectively,* r/n *and* $r(r-1)/n(n-1)$.

8.5 MINIMUM VARIANCE AND EFFICIENCY

The requirements of consistency and unbiassedness do not uniquely determine how to choose a good estimator. One finds, for instance, that both the sample mean and the sample median are consistent and unbiassed estimators of the location of a normal population with known variance. However, as it can be shown that the variance of the mean is smaller than the variance of the median, the mean is regarded as a better estimator of the central value. It seems natural, therefore, to use the spread in the estimates as a measure for the acceptability of an estimator. For most distributions encountered in practice, the second central moment, or the variance, will be a good measure of the concentration of the estimates; this is especially so for the many cases where this distribution is approximately normal.

Under fairly general conditions there exists a lower bound on the variance of the estimates derived from an estimator. This lower bound is easily established when considering the likelihood function defined by eq.(8.1). We shall assume that the first two derivatives of $L(\underline{x}|\theta)$ with respect to θ exist for all θ, and that the range of x is independent of θ. Given an estimator t of some function of θ, say $\tau(\theta)$, we define its bias $b(\theta)$ by the relation (compare eq.(8.4)),

$$E(t) = \int\ldots\int t(x_1,x_2,\ldots,x_n)L(x_1,x_2,\ldots,x_n|\theta)dx_1dx_2\ldots dx_n = \tau(\theta) + b(\theta). \quad (8.5)$$

With the assumption that the range of x is independent of θ the differentiation of eq.(8.5) with respect to θ gives

$$\int\ldots\int t\,\frac{\partial L}{\partial\theta}\,d\underline{x} = \int\ldots\int t\,\frac{\partial \ln L}{\partial\theta}\,L\,d\underline{x} = \frac{\partial\tau}{\partial\theta} + \frac{\partial b}{\partial\theta}. \quad (8.6)$$

Since L is the joint probability of the observations,

$$\int\ldots\int L\,d\underline{x} = 1, \quad (8.7)$$

we find by differentiating this relation with respect to θ

$$\int\ldots\int \frac{\partial L}{\partial\theta}\,d\underline{x} = \int\ldots\int \frac{\partial \ln L}{\partial\theta}\,L\,d\underline{x} = 0. \quad (8.8)$$

Multiplying eq.(8.8) by $\tau(\theta)$ and subtracting the result from eq.(8.6) leads to

$$\int\cdots\int\left(t-\tau(\theta)\right)\frac{\partial \ln L}{\partial\theta}\,L\,d\underline{x} = \frac{\partial\tau}{\partial\theta}+\frac{\partial b}{\partial\theta}\,. \tag{8.9}$$

By applying the Schwarz inequality to the integral we obtain the formula

$$\int\cdots\int\left(t-\tau(\theta)\right)^2 L\,d\underline{x}\int\cdots\int\left(\frac{\partial \ln L}{\partial\theta}\right)^2 L\,d\underline{x} \geq \left(\frac{\partial\tau}{\partial\theta}+\frac{\partial b}{\partial\theta}\right)^2, \tag{8.10}$$

which can be written as

$$V(t) \geq \left(\frac{\partial\tau}{\partial\theta}+\frac{\partial b}{\partial\theta}\right)^2 / E\left[\left(\frac{\partial \ln L}{\partial\theta}\right)^2\right]. \tag{8.11}$$

This fundamental inequality for the variance of an estimator is often referred to as the *Cramér–Rao inequality*. When L satisfies the aforementioned regularity conditions one can prove the following relation,

$$E\left[\left(\frac{\partial \ln L}{\partial\theta}\right)^2\right] = E\left(-\frac{\partial^2 \ln L}{\partial\theta^2}\right). \tag{8.12}$$

Thus an alternative form of the Cramér-Rao inequality, which is sometimes easier to evaluate, is

$$V(t) \geq \left(\frac{\partial\tau}{\partial\theta}+\frac{\partial b}{\partial\theta}\right)^2 / E\left(-\frac{\partial^2 \ln L}{\partial\theta^2}\right). \tag{8.13}$$

The lower limit of the variance implied by eqs.(8.11), (8.13) is called the *minimum variance bound*, MVB, and an estimator attaining this limit is called an MVB estimator, or more often, and *efficient estimator*.

From the derivation above it is realized that the variance of an estimator will attain the MVB if the Schwarz inequality applied to eq.(8.9) becomes an equality. The necessary and sufficient condition for this is that $\left(t-\tau(\theta)\right)$ is linearly related to $\frac{\partial \ln L}{\partial\theta}$ for all sets of observations; we write

$$\frac{\partial \ln L}{\partial\theta} = A(\theta)\left(t-\tau(\theta)-b(\theta)\right). \tag{8.14}$$

From eq.(8.13) one then gets a simple formula for the MVB,

$$V(t) = \left(\frac{\partial\tau}{\partial\theta}+\frac{\partial b}{\partial\theta}\right)/A(\theta). \tag{8.15}$$

Efficient estimators exist only for the limited class of problems for

which the likelihood function satiafies the condition (8.14). Most estimators will have a variance larger than the MVB. We therefore define the *efficiency* of an estimator as the ratio between the MVB and the actual variance V(t) of the estimator,

$$\text{Efficiency }(t) = \frac{\text{MVB}}{V(t)}\,. \tag{8.16}$$

It is interesting to note that for distributions where the regularity conditions do not hold the minimum attainable variance may be smaller or larger than the MVB.

For the particular case where t is an estimator of the parameter θ itself, we have $\frac{\partial\tau}{\partial\theta}=1$, and the Cramér-Rao inequalities (8.11), (8.13) become, respectively,

$$V(t) \geq \left(1+\frac{\partial b}{\partial\theta}\right)^2 / E\left[\left(\frac{\partial \ln L}{\partial\theta}\right)^2\right], \tag{8.17}$$

$$V(t) \geq \left(1+\frac{\partial b}{\partial\theta}\right)^2 / E\left(-\frac{\partial^2 \ln L}{\partial\theta^2}\right). \tag{8.18}$$

The efficiency condition eq.(8.14) then reads

$$\frac{\partial \ln L}{\partial\theta} = A(\theta)\left(t-\theta-b(\theta)\right), \tag{8.19}$$

simplifying the MVB formula (8.15) to

$$V(t) = \left(1+\frac{\partial b}{\partial\theta}\right)/A(\theta). \tag{8.20}$$

Exercise 8.3: Prove eq.(8.12). (Hint: Differentiate eq.(8.8) with respect to θ.)

8.5.1 Example: Estimator of the mean in the Poisson distribution

The likelihood for n observations $x_1, x_2, \ldots, x_n$ from the Poisson distribution $f(x;\theta) = \frac{1}{x!}\theta^x e^{-\theta}$ is

$$L(x_1, x_2, \ldots, x_n|\theta) = \prod_{i=1}^{n}\frac{1}{x_i!}\theta^{x_i}e^{-\theta}.$$

Hence we have for the derivative of lnL with respect to the unknown θ,

$$\frac{\partial \ln L}{\partial \theta} = \frac{\partial}{\partial \theta} \sum_{i=1}^{n} \ln\left(\frac{1}{x_i!}\,\theta^{x_i} e^{-\theta}\right) = \sum_{i=1}^{n}\left(\frac{1}{\theta}\,x_i - 1\right)$$

or

$$\frac{\partial \ln L}{\partial \theta} = \frac{n}{\theta}(\bar{x} - \theta),$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Thus $\frac{\partial \ln L}{\partial \theta}$ is of the form of eq.(8.19), with $A(\theta) = \frac{n}{\theta}$, $b(\theta) = 0$. An unbiassed and efficient estimator of the parameter θ is therefore $t = \bar{x}$, the sample mean, with variance given by the MVB formula eq.(8.20), $V(t) = \frac{\theta}{n}$.

Exercise 8.4: Explain why the Cauchy p.d.f. $f(x;\theta) = \pi^{-1}\left[1 + (x-\theta)^2\right]^{-1}$ does not have any efficient estimator of θ. Show that the Cramér-Rao lower bound is $2/n$, where n is the sample size.

Exercise 8.5: In the binomial distribution for which $L(r|\theta) = \binom{n}{r} p^r (1-p)^{n-r}$ show that the unbiassed estimator $t = r/n$ of p (Exercise 8.2) is also efficient. What is V(t)?

8.5.2 Example: Estimators of the mean in the normal p.d.f.

We have seen that the sample mean is a consistent and unbiassed estimator of the mean value in any population of finite variance. To examine further the properties of $\bar{x}$ as an estimator of μ in the normal distribution $N(\mu,\sigma^2)$ with fixed σ^2 we write

$$\frac{\partial}{\partial \mu} \ln L = \frac{\partial}{\partial \mu} \ln\left[\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\tfrac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right)\right] = \frac{n}{\sigma^2}(\bar{x} - \mu).$$

Comparing with eq.(8.19) we see that $\bar{x}$ is an efficient estimator of μ, and that the variance from eq.(8.20) is

$$V(\bar{x}) = \frac{\sigma^2}{n}. \tag{8.21}$$

This is in accordance with our previous knowledge that the variable $\bar{x}$ is distributed as $N(\mu,\sigma^2/n)$.

An alternative consistent and unbiassed estimator of μ is the sample median. It can be shown that, when size of the normal sample gets very large, the median becomes distributed according to $N(\mu,\pi\sigma^2/2n)$; hence the variance of this estimator of μ is

$$V(\text{median}) = \frac{\pi\sigma^2}{2n}. \tag{8.22}$$

This variance is larger than the MVB of eq.(8.21). The asymptotic efficiency of the median as an estimator of the mean in the normal distribution is therefore from eq.(8.16),

$$\text{Efficiency (median)} = \frac{\sigma^2}{n} \Big/ \frac{\pi\sigma^2}{2n} = \frac{2}{\pi} \simeq 0.64. \tag{8.23}$$

Exercise 8.6: Show that the MVB of μ in $N(\mu,\sigma^2)$ can also be found from eq. (8.18).

8.5.3 Example: Estimators of σ^2 and σ in the normal p.d.f.

We next consider the normal distribution $N(0,\sigma^2)$. To find an estimator of σ^2 we write

$$\frac{\partial}{\partial \sigma^2} \ln L = \frac{\partial}{\partial \sigma^2} \ln\left[\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\tfrac{1}{2}x_i^2/\sigma^2\right)\right] = \frac{n}{2\sigma^4}\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \sigma^2\right).$$

Again a comparison with eqs.(8.19), (8.20) shows that the statistic $t = \frac{1}{n}\sum_{i=1}^{n} x_i^2$ is an unbiassed and efficient estimator of the variance σ^2 of $N(0,\sigma^2)$, with variance $V(t) = 2\sigma^4/n$.

If, alternatively, we take σ as the parameter to be estimated we find

$$\frac{\partial}{\partial \sigma} \ln L = \frac{n}{\sigma^3}\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \sigma^2\right),$$

which is not of the form of eq.(8.19). Hence there exists no efficient estimator for the standard deviation σ of the normal p.d.f. $N(0,\sigma^2)$. In the framework and formulation of Sect.8.5.1 there exists, however, an efficient estimator $t = \frac{1}{n}\sum_{i=1}^{n} x_i^2$ for a function of σ, $\tau(\sigma) = \sigma^2$; from eq.(8.15) the variance of t is seen to be

$$V(t) = \frac{\partial(\sigma^2)}{\partial \sigma} \Big/ \frac{n}{\sigma^3} = \frac{2\sigma^4}{n}, \tag{8.24}$$

which is in agreement with our previous finding.

Note that the results above hold also if the sample originates from a normal distribution having a known mean value μ different from zero.

Exercise 8.7: Consider the gamma distribution $f(x;\alpha,\beta) = \left[\Gamma(\alpha)\beta^{\alpha}\right]^{-1} x^{\alpha-1} e^{-x/\beta}$. (a) Assuming α to be known, find an efficient estimator and the MVB of β. (b) Assuming β to be known, does any efficient estimator exist for α?

8.6 SUFFICIENCY

8.6.1 One-parameter case

An estimator t is said to be *sufficient* if it exhausts all information in the observations $x_1, x_2, \ldots, x_n$ regarding the parameter θ. For the normal distribution we have used the sample mean $\bar{x} = \frac{1}{n}\sum_i x_i$ as an estimator of the population mean μ. No extra knowledge on μ can be gained from other functions of the observations, such as $\sum_i |x_i|$, $\sum_i x_i^3$ etc.; the statistic $\bar{x}$ is then a sufficient estimator for μ. Actually any function of $\bar{x}$ provides a sufficient statistic for μ in the normal distribution. Therefore, to choose between the different functions one may have to test also the qualities of consistency, unbiassedness, and efficiency.

To be more precise, let us consider the likelihood function when the p.d.f. is $f(x;\theta)$. Suppose that L can be factorized to give

$$L(\underline{x}|\theta) = \prod_{i=1}^{n} f(x_i;\theta) = G(t|\theta)H(\underline{x}), \tag{8.25}$$

where the function*) G involves the statistic t and the parameter θ, and H is independent of θ, being a function of the observations $\underline{x}$ only. Since G only has reference to the data *via* $t = t(x_1, x_2, \ldots, x_n)$, and is a known number for the given sample, t must supply all the available information in the data regarding θ. It follows that whenever the likelihood function can be written in the form of eq.(8.25), t is a sufficient statistic for the parameter θ.

One can show that a necessary condition for a sufficient statistic to exist is that the p.d.f. belongs to the *exponential family*, defined by

$$f(x;\theta) = \exp\big(B(\theta)C(x) + D(\theta) + E(x)\big), \tag{8.26}$$

where B,C,D,E are functions of the indicated arguments.

For the restricted class of p.d.f.'s satisfying eq.(8.26) one sees that the factorization requirement of eq.(8.25) implies that a sufficient statistic must be expressed by the function C(x),

$$t = \sum_{i=1}^{n} C(x_i). \tag{8.27}$$

It is easily shown that efficient estimators are always sufficient. To see this we take logarithms on both sides of eq.(8.25) and differentiate with respect to θ, getting

$$\frac{\partial \ln L}{\partial \theta} = \frac{\partial \ln G}{\partial \theta}, \tag{8.28}$$

We see that the efficiency condition eq.(8.14) is just a special case of the more general eq.(8.28), with

$$\frac{\partial \ln G}{\partial \theta} = A(\theta)\big(t - \tau(\theta) - b(\theta)\big). \tag{8.29}$$

Under the regularity conditions specified for the likelihood function in Sect.8.5 there is among all sufficient estimators for θ only one efficient estimator. If the p.d.f. belongs to the exponential family and the range of x is independent of θ, it can be shown that sufficient statistics always exist for θ. However, there will be just one sufficient statistic which will satisfy eq.(8.29) and thus estimate some function $\tau(\theta)$ with variance equal to the MVB; compare the example of Sect.8.5.3. Furthermore, for large samples, any function of a sufficient statistic will be an MVB estimator.

Exercise 8.8: Verify that the Poisson distribution $P(x;\theta) = \frac{1}{x!}\theta^x e^{-\theta}$ belongs to the exponential family. Compare Sect.8.5.1.

Exercise 8.9: Show that the Cauchy p.d.f. $f(x;\theta) = \pi^{-1}\big(1 + (x-\theta)^2\big)^{-1}$ does not have a sufficient estimator of θ. Compare Exercise 8.4.

8.6.2 Example: Single sufficient statistics for the normal p.d.f.

We have already seen (Sect.8.5.2) that the sample mean $\bar{x}$ is an efficient estimator of the mean μ in the normal distribution. According to the general statement of the previous section $\bar{x}$ is then also a sufficient statistic for μ. We want to show this more directly, and observe that, generally

$$\sum(x_i-\mu)^2 = \sum\big((\bar{x}-\mu) + (x_i-\bar{x})\big)^2 = \sum(\bar{x}-\mu)^2 + 2(\bar{x}-\mu)\sum(x_i-\bar{x}) + \sum(x_i-\bar{x})^2$$

$$= n(\bar{x}-\mu)^2 + \sum(x_i-\bar{x})^2.$$

*) In the non-regular situation where the range of x may depend on θ, one must check that the function G is the conditional p.d.f. for t, given θ.

Therefore, when σ^2 is known, the likelihood function for μ can be written

$$L(\underline{x};\sigma^2|\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\tfrac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right)$$

$$= \left\{\frac{1}{\sqrt{2\pi}\,\sigma/\sqrt{n}} \exp\left(-\tfrac{1}{2}\left(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}\right)^2\right)\right\}\left\{\frac{n^{-\frac{1}{2}}}{(\sqrt{2\pi}\sigma)^{n-1}} \exp\left(-\tfrac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma}\right)^2\right)\right\}.$$

L is now factorized to the form of eq.(8.25), with the first bracket to be identified with $G(\bar{x}|\mu)$, only dependent on the observations through the statistic $\bar{x}$. Thus $\bar{x}$ is a sufficient statistic for μ. We see explicitly that $G(\bar{x}|\mu)$ is the p.d.f. for the variable $\bar{x}$, namely $N(\mu,\sigma^2/n)$. The second bracket involves the separate sample values, and corresponds to $H(\underline{x})$ of eq.(8.25).

It is easy to see that, given σ^2, the normal p.d.f. is of the exponential form of eq.(8.26) for the parameter μ, because

$$N(\mu,\sigma^2) = \exp\left(\frac{\mu}{\sigma^2}x - \frac{\mu^2}{2\sigma^2} - \frac{x^2}{2\sigma^2} - \tfrac{1}{2}\ln(2\pi\sigma^2)\right).$$

Thus we can make the identification with the functions

$$B(\mu) = \frac{\mu}{\sigma^2}, \quad C(x) = x, \quad D(\mu) = -\frac{\mu^2}{2\sigma^2}, \quad E(x) = -\frac{x^2}{2\sigma^2} - \tfrac{1}{2}\ln(2\pi\sigma^2).$$

If, instead, μ was fixed and σ^2 to be estimated we would write the likelihood function as

$$L(\underline{x};\mu|\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\tfrac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right)$$

$$= \left\{\left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\tfrac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2 \cdot \frac{1}{\sigma^2}\right)\right\} \cdot \left\{(2\pi)^{-\frac{1}{2}n}\right\}.$$

This is again of the factorized form of eq.(8.25), and shows that $\Sigma(x_i-\mu)^2$ is now a sufficient statistic for σ^2. One observes that by inclusion of appropriate factors the first bracket can be identified with a chi-square p.d.f. with n degrees of freedom for the variable $\Sigma(x_i-\mu)^2/\sigma^2$.

It is seen that, estimating σ^2, we could take for the functions of the exponential family, eq.(8.26),

$$B(\sigma^2) = -\frac{1}{2\sigma^2}, \quad C(x) = (x - \mu)^2, \quad D(\sigma^2) = -\tfrac{1}{2}\ln(2\pi\sigma^2), \quad E(x) = 0.$$

8.6.3 Extension to several parameters

If the p.d.f. has k parameters $\theta_1,\theta_2,\ldots,\theta_k$, there will exist a set of r jointly sufficient statistics $t_1,t_2,\ldots,t_r$ where r may be smaller than, equal to, or larger than k, provided that the likelihood function can be factorized in the following manner,

$$L(\underline{x}|\theta_1,\theta_2,\ldots,\theta_k) = G(t_1,t_2,\ldots,t_r|\theta_1,\theta_2,\ldots,\theta_k)H(\underline{x}). \tag{8.30}$$

This is seen to be a generalization of eq.(8.25).

To have a set of k jointly sufficient statistics for the k parameters, the p.d.f. must belong to the exponential family. The generalization of eq. (8.26) to the multi-parameter case is

$$f(x;\underline{\theta}) = \exp\left(\sum_{j=1}^{k} B_j(\underline{\theta})C_j(x) + D(\underline{\theta}) + E(x)\right), \tag{8.31}$$

where $\underline{\theta} = \{\theta_1,\theta_2,\ldots,\theta_k\}$. Writing out the likelihood function one finds by comparison with the factorization property (8.30) that the k joint sufficient statistics of the k parameters must be expressed by the functions C of the observations,

$$t_j = \sum_{i=1}^{n} C_j(x_i), \qquad j = 1,2,\ldots,k. \tag{8.32}$$

8.6.4 Example: Jointly sufficient estimators for μ and σ^2 in $N(\mu,\sigma^2)$

To demonstrate that two jointly sufficient estimators exist for μ and σ^2 in the normal distribution it suffices to show that the p.d.f. belongs to the exponential family of eq.(8.31). We write

$$N(\mu,\sigma^2) = \exp\left(\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{\mu^2}{2\sigma^2} - \tfrac{1}{2}\ln(2\pi\sigma^2)\right),$$

which is of the form of eq.(8.31), with

$$B_1(\mu,\sigma^2) = \frac{\mu}{\sigma^2}, \qquad B_2(\mu,\sigma^2) = -\frac{1}{2\sigma^2},$$

$$C_1(x) = x, \qquad C_2(x) = x^2,$$

$$D(\mu,\sigma^2) = -\frac{\mu^2}{2\sigma^2} - \tfrac{1}{2}\ln(2\pi\sigma^2), \qquad E(x) = 0.$$

According to eq.(8.32) two jointly sufficient statistics for μ and σ^2 are given by

$$t_1 = \sum_{i=1}^{n} C_1(x) = \sum_{i=1}^{n} x_i ,$$

$$t_2 = \sum_{i=1}^{n} C_2(x) = \sum_{i=1}^{n} x_i^2 .$$

However, these estimators of μ and σ^2 are neither unbiassed nor consistent. Considering instead (compare Sect.5.1.6)

$$z_1 = \frac{1}{n} t_1 = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x} ,$$

$$z_2 = \frac{1}{n-1}(t_2 - \frac{1}{n} t_1^2) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = s^2 ,$$

it is seen that these variables define a one-to-one mapping of t_1, t_2 onto z_1, z_2. The statistics z_1 and z_2 are therefore also jointly sufficient estimators for the two parameters. Moreover, these are unbiassed estimators of μ and σ^2, because

$$E(z_1) = \mu ,$$

$$E(z_2) = \sigma^2 .$$

The joint likelihood function for μ and σ^2 can be shown to factorize to the form of eq.(8.30), but since

$$G(\bar{x}, s^2 | \mu, \sigma^2) \neq G_1(\bar{x}|\mu) G_2(s^2|\sigma^2)$$

s^2 is not a single sufficient estimator of σ^2 when μ is unknown, nor is $\bar{x}$ a single sufficient estimator of μ when σ^2 is unknown.

9. The Maximum-Likelihood method

The method of parameter estimation known as the Maximum-Likelihood (ML) method is very general and powerful. For estimation problems where a functional dependence can be written down for the observed variables, the ML method is eminently satisfactory for two reasons: it provides estimators with desirable properties, and the estimators are easy to find. The ML theory has a fundamental position in all problems of parameter estimation where the functional form of the p.d.f. is given. We will therefore treat the ML method somewhat at length and discuss its theoretical aspects as well as its practical implications.

In expositions of the ML method aimed for physicists, it is often the *asymptotic properties* of the ML estimators which are emphasized. For large samples the ML estimates are normally distributed. This nice property makes the determination of variances on ML estimates very simple. The following presentation will also emphasize the asymptotic properties, and these should be fairly easy to extract, without reading the chapter in full. However, it is in the framework of *sufficient statistics* that the ML estimators have their most important properties. We have already given a somewhat theoretical discussion of some of the fundamental properties of estimators in Chapter 8, and we will find in this chapter that the ML estimators possess most of these good properties.

The reader who wants only a first working knowledge of the ML method, and who wants mainly to know its asymptotic properties, can select as an initial reading the sections 9.1, 9.2, 9.5.1, 9.5.4 - 9.5.6, 9.6.1, 9.6.3, 9.7.1, 9.9. In particular one should note that the very simple graphical solution of the ML estimation problem, described in Sects.9.6.1 and 9.6.3, can be applied in many practical situations involving one or two unknown parameters.

9.1 THE MAXIMUM-LIKELIHOOD PRINCIPLE

Consider a p.d.f. $f(x|\theta)$ with an unknown parameter θ to be estimated from the set of observations *) $x_1, x_2, \ldots, x_n$. We have already in Chapter 8 encountered the likelihood function (LF)

$$L(\underline{x}|\theta) \equiv \prod_{i=1}^{n} f(x_i|\theta) \tag{9.1}$$

as the joint conditional probability of the observations $x_1, x_2, \ldots, x_n$ at a fixed θ. Since $f(x|\theta)$ is a probability density function properly normalized to one, we see that when the LF is considered a function of the x_i's the integration over the total sample space Ω yields

$$\int_{\Omega} L(\underline{x}|\theta)\,d\underline{x} = 1 \tag{9.2}$$

for all θ.

In the LF one may regard the observed x_i's as constants and the parameter θ as a variable. According to the *Maximum-Likelihood Principle* we should choose as an estimate of the unknown parameter θ that particular $\hat{\theta}$ within the admissible range of θ which renders L as large as possible **). This means that the estimate $\hat{\theta}$ is such that

$$L(\underline{x}|\hat{\theta}) \geq L(\underline{x}|\theta) \tag{9.3}$$

for all conceivable values of θ. If L is twice differentiable with respect to θ, the value $\hat{\theta}$ may be obtained by solving the equation

*) Each observation often corresponds to more than one measured variable; for instance, x can be a set of two angular quantities specifying a spatial direction (an example is given in Sect.9.5.8). In general, x_i will denote the set of measured quantities for event i.

**) This choice of the "best value of a parameter" as the one that maximizes the conditional probability of x for given θ, is not obvious. From Bayes' Theorem, for instance, a more intuitive choice of a "best value of θ" would be the one which maximizes the joint probability of x and θ; compare Sect.2.4.3.

$$\frac{\partial L(\underline{x}|\theta)}{\partial\theta} = \frac{\partial}{\partial\theta}\prod_{i=1}^{n} f(x_i|\theta) = 0 \tag{9.4}$$

with the condition that the second derivative evaluated at $\theta = \hat{\theta}$ is negative,

$$\left.\frac{\partial^2 L(\underline{x}|\theta)}{\partial^2\theta}\right|_{\theta=\hat{\theta}} = \left.\frac{\partial^2}{\partial\theta^2}\prod_{i=1}^{n} f(x_i|\theta)\right|_{\theta=\hat{\theta}} < 0\,. \tag{9.5}$$

Usually L has only one maximum, and $\hat{\theta}$ is unique. If there is more than one maximum, one should look for supplementary information to choose between the solutions.

Since L and the logarithm of L attain their maxima for the same value of θ, the ML solution may be found from the *likelihood equation*

$$\frac{\partial}{\partial\theta}\ln L(\underline{x}|\theta) = \frac{\partial}{\partial\theta}\sum_{i=1}^{n}\ln f(x_i|\theta) = 0\,. \tag{9.6}$$

This sum is often easier to handle than the product in eq.(9.4). We then require

$$\left.\frac{\partial^2}{\partial\theta^2}\ln L(\underline{x}|\theta)\right|_{\theta=\hat{\theta}} = \frac{\partial^2}{\partial\theta^2}\sum_{i=1}^{n}\left.\ln f(x_i|\theta)\right|_{\theta=\hat{\theta}} < 0\,. \tag{9.7}$$

For the general case with several unknown parameters $\underline{\theta} = \{\theta_1, \theta_2, \ldots, \theta_k\}$ we have to solve the set of k likelihood equations

$$\frac{\partial}{\partial\theta_j}\ln L(\underline{x}|\underline{\theta}) = 0\,, \qquad j = 1,2,\ldots,k, \tag{9.8}$$

to find the ML estimates $\underline{\hat{\theta}} = \{\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k\}$. A sufficient condition that the LF is at an absolute maximum is that the quadratic matrix $U(\underline{\hat{\theta}})$ with elements

$$U_{ij}(\underline{\hat{\theta}}) = \left.\frac{\partial^2 \ln L}{\partial\theta_i\,\partial\theta_j}\right|_{\underline{\theta}=\underline{\hat{\theta}}} \tag{9.9}$$

is negative definite.

In many practical problems the solution of eq.(9.6), or more generally eqs.(9.8), has to be found numerically. In fact, the numerical procedure can in many instances be advantageous, since it gives also directly the variances to be associated with the ML estimates; see Sect.9.6.2 for an example.

9.1.1 Example: Estimate of mean lifetime

Let us assume that we observe the production as well as the decay of neutral kaons in an infinite detector. The p.d.f. is then

$$f(t|\tau) = \frac{1}{\tau} e^{-t/\tau}, \qquad 0 \le t \le \infty,$$

where τ is the K^o mean lifetime to be estimated. From measurements of the K^o momentum and the length between the production and the decay points, the proper flight-time t_i for each event is determined. For n observed events the LF is, according to the definition eq.(9.1),

$$L(t|\tau) = \prod_{i=1}^{n} \frac{1}{\tau} e^{-t_i/\tau},$$

and from eq.(9.6) the ML estimate $\hat{\tau}$ of τ is found by solving the equation

$$\frac{\partial \ln L}{\partial \tau} = \frac{\partial}{\partial \tau} \sum_{i=1}^{n} \left(-\ln\tau - \frac{t_i}{\tau}\right) = \sum_{i=1}^{n} \left(-\frac{1}{\tau} + \frac{t_i}{\tau^2}\right) = 0.$$

This gives

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} t_i = \bar{t}.$$

Hence the ML estimate of the parameter τ is equal to the arithmetic mean of the observed flight-times. The solution obtained does correspond to a maximum of the LF, because

$$\left.\frac{\partial^2 \ln L}{\partial \tau^2}\right|_{\tau=\hat{\tau}} = -n/\hat{\tau}^2 < 0,$$

in accordance with the requirement of eq.(9.7).

Exercise 9.1: Instead of having the lifetime τ as a parameter in the example above, one can alternatively use the decay constant $\lambda = 1/\tau$ as the parameter with the p.d.f. $f(t|\lambda) = \lambda e^{-\lambda t}$. Show that the ML estimate of λ is $\hat{\lambda} = \left(\frac{1}{n}\sum_i t_i\right)^{-1} = 1/\hat{\tau}$.

Exercise 9.2: Assume that we observe the production and decay of particles within a *finite* detector. The p.d.f. is then (compare Sect.6.3.1)

$$f(t;T|\tau) = \frac{1}{\tau} e^{-t/\tau} \Big/ \left(1 - e^{-T/\tau}\right), \qquad 0 \le t \le T.$$

Show that the ML estimate for the parameter τ can be expressed as

$$\tau = \frac{1}{n} \sum_{i=1}^{n} \left[t_i + T_i e^{-T_i/\tau} \Big/ \left(1 - e^{-T_i/\tau}\right)\right] = \bar{t} + \frac{1}{n} \sum_{i=1}^{n} T_i e^{-T_i/\tau} \Big/ \left(1 - e^{-T_i/\tau}\right),$$

where T_i is the potential flight-time for the i-th event. Note that τ enters on the right-hand side in the correction term to the arithmetic mean $\bar{t}$. The solution $\hat{\tau}$ can be obtained by an iteration procedure, taking $\bar{t}$ as the starting value.

9.2 ESTIMATION OF PARAMETERS IN THE NORMAL DISTRIBUTION

To illustrate further the ML estimation of unknown parameters we will consider some useful examples where the LF is written in terms of normal probability density functions.

9.2.1 Estimation of μ; measurements with common error

Let $x_1, x_2, \ldots, x_n$ be n independent measurements on the same unknown quantity μ, and assume that the measurement error is σ, common to all observations. We assume the x_i's to constitute a sample of size n drawn from a normal population $N(\mu,\sigma^2)$, where μ is unknown, σ^2 known.

To estimate the parameter μ by the Maximum-Likelihood method we demand the maximum of

$$L(\underline{x};\sigma|\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\tfrac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right)$$

when L is considered a function of μ. In this case eq.(9.6) reads

$$\frac{\partial \ln L}{\partial \mu} = \frac{\partial}{\partial \mu} \sum_{i=1}^{n} \left(-\tfrac{1}{2}\ln(2\pi\sigma^2) - \tfrac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right) = 0,$$

and the solution is

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}. \tag{9.10}$$

Therefore, the ML estimate of the population mean μ is equal to the sample mean $\bar{x}$.

9.2.2 Estimation of μ; measurements with different errors (weighted mean)

We assume now that the measurements $x_1, x_2, \ldots, x_n$ on the unknown quantity μ have different, but still known errors. If each measurement x_i is

normally distributed with measuring error σ_i *), the LF is

$$L(x_1,x_2,\ldots,x_n;\sigma_1,\sigma_2,\ldots,\sigma_n|\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma_i} \exp\left(-\tfrac{1}{2}\left(\frac{x_i-\mu}{\sigma_i}\right)^2\right).$$

The ML estimate for μ becomes in this situation

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i/\sigma_i^2}{\sum_{i=1}^{n} 1/\sigma_i^2}, \tag{9.11}$$

which is called the *weighted mean* of the observations. We note that the measurements are given weight in inverse proportion to the square of their errors, *i.e.* they are weighted in proportion to their precisions $1/\sigma_i^2$. In the case that the measurements have the same error, $\sigma_i = \sigma$, it is seen that eq. (9.11) reduces to eq. (9.10), as it should.

It will be shown in Sect. 9.5.4 that the variance on the estimate $\hat{\mu}$ of eq. (9.11) is given by

$$V(\hat{\mu}) = \left(\sum_{i=1}^{n} 1/\sigma_i^2\right)^{-1}. \tag{9.12}$$

Exercise 9.3: Show that the ML estimate of σ^2 in $N(\mu,\sigma^2)$ for given μ is $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$.

9.2.3 Simultaneous estimation of mean and variance

It is frequently not realistic to consider the errors connected to the measurements as known quantitites. As an example, suppose we have measured the range of monoenergetic particles in some material. The measurements for n events are $x_1,x_2,\ldots,x_n$, and this sample is supposed to originate from a normal population $N(\mu,\sigma^2)$. Here μ is the true, but unknown mean range, and σ the combined straggling and measuring error, which is also not known, but assumed to be common for all measurements.

To estimate both μ and σ^2 by the ML method we write

*) Note that the x_i originate from different parent populations, and hence do not constitute a sample in the usual sense.

$$L(\underline{x}|\mu,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\tfrac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right).$$

From eq. (9.8) we should now solve the set of two simultaneous equations

$$\frac{\partial \ln L}{\partial \mu} = \sum_{i=1}^{n} \frac{x_i-\mu}{\sigma^2} = 0$$

and

$$\frac{\partial \ln L}{\partial(\sigma^2)} = \sum_{i=1}^{n}\left(-\tfrac{1}{2}\,\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}(x_i-\mu)^2\right) = 0,$$

from which the ML estimates are

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x},$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2.$$

We note that the estimate of μ is equal to $\bar{x}$, as we found in Sect. 9.2.1. The ML estimate of σ^2 is biassed, the relation to the unbiassed estimator $s^2 = \frac{1}{n-1}\Sigma(x_i-\bar{x})^2$ of σ^2 being $\hat{\sigma}^2 = \frac{n-1}{n}s^2$; compare Sect. 8.4.1.

The errors on $\hat{\mu}$ and $\hat{\sigma}^2$ will be derived in Sect. 9.5.5.

Exercise 9.4: Show that the estimates $\hat{\mu}$, $\hat{\sigma}^2$ correspond to a maximum of the LF.

9.3 ESTIMATION OF THE LOCATION PARAMETER IN THE CAUCHY P.D.F.

In the examples of the previous sections the ML solution for the parameters turned out to be unique. As an example where eq. (9.6) may have several solutions, consider the estimation of the parameter θ in the Cauchy distribution $f(x|\theta) = 1/\pi\{1 + (x-\theta)^2\}^{-1}$. With the measurements $x_1,x_2,\ldots,x_n$ we have the LF

$$L(\underline{x}|\theta) = \prod_{i=1}^{n} \frac{1}{\pi}\,\frac{1}{1+(x_i-\theta)^2},$$

and eq. (9.6) becomes

$$\frac{\partial \ln L}{\partial \theta} = \sum_{i=1}^{n} \frac{2(x_i-\theta)}{1+(x_i-\theta)^2} = 0.$$

This is an equation of degree (2n-1) in θ, and up to (2n-1) different solutions exist, n of which will correspond to maxima of the LF. Usually the best value, corresponding to the highest maximum of L, is near to the sample median. The median of the sample may therefore be taken as the starting value in an iterative search for the maximum of L.

9.4 PROPERTIES OF MAXIMUM-LIKELIHOOD ESTIMATORS

The use of the ML method presupposes that the p.d.f. is specified except for the unknown parameters. It may often not be possible to find explicit analytical expressions for the estimators, but by maximizing the likelihood function using iteration or interpolation procedures the estimates themselves can be obtained, irrespective of the analytical form of the estimators. Moreover, as will be shown in the following, the ML estimators possess nearly all the theoretically best estimator properties.

9.4.1 Invariance under parameter transformation

In practice it is often rather arbitrary what physical quantity is chosen as the parameter θ to be estimated. For example, in the lifetime determination in Sect. 9.1.1 we could have used the decay constant λ as parameter instead of the mean lifetime τ. Let $\hat{\theta}$ be the ML solution for the parameter θ. If, alternatively, we had chosen to estimate a function of θ, say $\tau(\theta)$, then the ML solution for this function would be that value $\hat{\tau(\theta)}$ for which $\partial L/\partial\tau = 0$. Since $\partial L/\partial\theta = (\partial L/\partial\tau)(\partial\tau/\partial\theta)$ for all θ, and $\partial L/\partial\theta = 0$ for $\theta=\hat{\theta}$, it follows when $\partial\tau/\partial\theta \neq 0$ that also $\partial L/\partial\tau = 0$ for $\theta = \hat{\theta}$; hence we must have

$$\hat{\tau(\theta)} = \tau(\hat{\theta}). \tag{9.13}$$

Eq.(9.13) expresses the invariance of ML estimates under parameter transformations. We see that it makes no difference to the result for the estimate $\hat{\theta}$ which variable, θ or the function $\tau(\theta)$, is used to maximize the LF. Thus the ML method is free for the arbitrariness discussed in connection with Bayes' Postulate, Sect.2.4.5.

An example of the invariance property has already been demonstrated by Sect.9.1.1 and Exercise 9.1.

9.4.2 Consistency

It can be shown that under very general conditions the ML estimators are consistent. This means that the ML estimates will converge towards the true parameter values when the sample size increases. This applies to the single-parameter as well as to the multi-parameter case. However, it may sometimes happen that the LF has two or more suprema due to some specific form of the p.d.f. which essentially makes the parameter indeterminable.

9.4.3 Unbiassedness

In favourable situations the ML estimators turn out to be unbiassed, irrespective of the size of the sample. This means that for any n, the estimates will be distributed with a mean equal to the true parameter value. For example, it is easily verified that the ML estimator $\bar{t} = \frac{1}{n}\Sigma t_i$ of τ in the p.d.f. $f(t|\tau) = \frac{1}{\tau} e^{-t/\tau}$ (Sect.9.1.1) is unbiassed, since

$$E(\bar{t}) = E\left(\frac{1}{n}\sum_{i=1}^{n} t_i\right) = \frac{1}{n}\cdot nE(t) = \int_0^\infty t\,\frac{1}{\tau}\,e^{-t/\tau}dt = \tau,$$

in accordance with the requirement for an unbiassed estimator, eq.(8.3).

The ML estimators are frequently not unbiassed for finite samples. For instance, the ML estimator of σ^2 in $N(\mu,\sigma^2)$ is $\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$ (see Sect. 9.2.3), and this is a biassed estimator of σ^2, because

$$E(\hat{\sigma^2}) = \left(1 - \frac{1}{n}\right)\sigma^2 \neq \sigma^2.$$

The bias is here $-\frac{1}{n}\sigma^2$, which is negligible for large n.

From the invariance property of ML estimators, we know that for a function $\tau(\theta)$ of the parameter, $\hat{\tau(\theta)} = \tau(\hat{\theta})$, but for expectation values in general we have

$$E\Big(\tau(\theta)\Big) \neq \tau\Big(E(\theta)\Big). \tag{9.14}$$

Therefore, although θ may have an unbiassed estimator, $\tau(\theta)$ need not.

In the asymptotic limit of infinite samples all ML estimators are unbiassed.

Exercise 9.5: Show that the ML estimator of the decay constant λ in the p.d.f. $f(t|\lambda) = \lambda e^{-\lambda t}$ is biassed. Is the estimator consistent?

Exercise 9.6: Show that the ML estimates of μ and σ of the normal distribution $N(\mu,\sigma^2)$ are $\hat{\mu} = \bar{x}$ and $\hat{\sigma} = \left(\frac{1}{n}\Sigma(x_i-\hat{\mu})^2\right)^{\frac{1}{2}}$, respectively. It has been shown in Sect.8.3 and Sect.8.4 that $\bar{x}$ is a consistent and unbiassed estimator of μ. Show that the estimator of σ is biassed, but consistent. Compare the estimate $\hat{\sigma}$ with the estimate of σ^2 derived in Sect.9.2.3 and note that this provides another example on the invariance of ML estimators.

9.4.4 Sufficiency

It was stated in Sect.8.6 that if the likelihood function can be factorized as

$$L(\underline{x}|\theta) = G(t|\theta)H(\underline{x}), \tag{8.25}$$

where t is a function of the observations, $t = t(x_1,x_2,\ldots,x_n)$, then t is a sufficient statistic for θ. This entails that t contains all information in the measurements regarding the parameter θ. A condition for a sufficient statistic to exist is that the p.d.f. $f(x|\theta)$ can be written in the exponential form

$$f(x|\theta) = \exp\left(B(\theta)C(x) + D(\theta) + E(x)\right), \tag{8.26}$$

where B,C,D,E are functions of the indicated arguments.

Since estimating θ from $L(\underline{x}|\theta)$ of eq.(8.25) is tantamount to estimation θ from $G(t|\theta)$, it follows that the estimate $\hat{\theta}$ depends on the observations through the sufficient statistic t alone. This also means that if there exists a sufficient estimator for the parameter θ, the likelihood equation will produce it. It can be shown that sufficient estimators will have the *minimum attainable variance*. This means that if a sufficient estimator exists, the ML method produces the estimate with minimum attainable variance. This fact is probably the strongest argument in favour of the ML method.

The likelihood function of our example from Sect.9.1.1 can be written

$$L(t|\tau) = \prod_{i=1}^{n} \frac{1}{\tau} e^{-t_i/\tau} = \left(\frac{1}{\tau}\right)^n e^{-\Sigma t_i/\tau} = \left(\left(\frac{1}{\tau}\right)^n e^{-n\bar{t}/\tau}\right)\cdot(1),$$

which is explicitly of the form (8.25). This demonstrates that $\bar{t} = \frac{1}{n}\Sigma t_i$, or

any one-to-one function of $\bar{t}$, is a sufficient estimator of τ. We have already shown that the ML estimator of τ is just $\bar{t}$.

The present remarks on sufficiency for ML estimators also apply to the multi-parameter case. For the case with k unknown parameters it can be proved that $r \leq k$ sufficient estimators $t_1,t_2,\ldots,t_r$ can simultaneously have their minimum attainable variance.

9.4.5 Efficiency

The variance of an ML estimator can not be arbitrarily small. The necessary and sufficient condition that an *efficient*, or *minimum variance bound* (MVB), estimator t exists for the parameter θ, is that one can write (Sect.8.5)

$$\frac{\partial \ln L}{\partial \theta} = A(\theta)\left(t - \theta - b(\theta)\right), \tag{8.19}$$

where $b(\theta)$ is the bias of the estimator. Since the ML estimate $\hat{\theta}$ for θ is obtained by equating $\partial \ln L/\partial\theta$ to zero, it follows that the likelihood equation produces the efficient estimator whenever there is one.

The MVB for an efficient estimator t of θ is given by the Cramér-Rao limit, (compare eqs.(8.17), (8.18), (8.20)),

$$V(t) = \frac{\left(1 + \frac{\partial b}{\partial\theta}\right)^2}{E\left[\left(\frac{\partial \ln L}{\partial\theta}\right)^2\right]} = \frac{\left(1 + \frac{\partial b}{\partial\theta}\right)^2}{E\left(-\frac{\partial^2 \ln L}{\partial\theta^2}\right)} = \frac{1 + \frac{\partial b}{\partial\theta}}{A(\theta)}. \tag{9.15}$$

For example, to find an efficient estimator for τ in the p.d.f. $f(t|\tau) = \frac{1}{\tau} e^{-t/\tau}$, it must be possible to write $\partial \ln L/\partial\tau = A(\tau)(\hat{\tau}-\tau-b(\tau))$. We find (Sect.9.1.1)

$$\frac{\partial \ln L}{\partial \tau} = \sum_{i=1}^{n}\left(-\frac{1}{\tau} + \frac{t_i}{\tau^2}\right) = \frac{n}{\tau^2}(\bar{t}-\tau),$$

which is of the form of eq.(8.19), with $A(\tau) = n/\tau^2$, $b(\tau) = 0$. Therefore, the ML estimator $\bar{t} = \frac{1}{n}\Sigma t_i$ is an unbiassed and efficient estimator of τ, with variance $V(\bar{t}) = 1/A(\tau) = \tau^2/n$.

As stated in Sect.8.6 there is among all sufficient estimators of θ only one statistic t which will estimate some function $\tau(\theta)$ with variance equal to the MVB. The MVB can be found if one can write

$$\frac{\partial \ln L}{\partial \theta} = A(\theta)\{t - \tau(\theta) - b(\theta)\}. \tag{8.14}$$

Then (compare eqs.(8.11),(8.13),(8.15)),

$$V(t) = \frac{\left(\frac{\partial \tau}{\partial \theta} + \frac{\partial b}{\partial \theta}\right)^2}{E\left[\left(\frac{\partial \ln L}{\partial \theta}\right)^2\right]} = \frac{\left(\frac{\partial \tau}{\partial \theta} + \frac{\partial b}{\partial \theta}\right)^2}{E\left(-\frac{\partial^2 \ln L}{\partial \theta^2}\right)} = \frac{\frac{\partial \tau}{\partial \theta} + \frac{\partial b}{\partial \theta}}{A(\theta)}. \tag{9.16}$$

For example, from eq.(8.14) it is easily seen that there can be no efficient estimator of σ in $N(0,\sigma^2)$. However, there is an unbiassed and efficient estimator t of $\tau(\sigma) = \sigma^2$, namely $t = \frac{1}{n}\sum_{i=1}^{n} x_i^2$, for which $V(t) = 2\sigma^4/n$ (see Sect.8.5.3).

As will be shown in Sect.9.4.7 the ML estimate $\hat{\theta}$ of θ, for large n, is approximately normally distributed about the true value $\theta = \theta_o$ with variance equal to the MVB, provided that certain regularity conditions hold. This implies that ML estimators are asymptotically efficient, and hence also asymptotically sufficient, since whenever efficiency holds sufficiency holds too.

In the multi-parameter case it can be shown that, for large samples, the ML estimates will, under ordinary regularity conditions, tend to a multinormal distribution.

Exercise 9.7: Explain why the p.d.f. $f(t|\lambda) = \lambda e^{-\lambda t}$ has no efficient estimator for λ itself, but only for the function $\tau(\lambda) = 1/\lambda$. What is the ML estimator $\hat{\lambda}$? Show that the MVB of $\hat{\lambda}$ is λ^2/n.

Exercise 9.8: Show that the weighted mean of eq.(9.11) provides an unbiassed and efficient estimator of the mean μ.

9.4.6 Uniqueness

It is easy to see that, if an efficient estimator exists for some function $\tau(\theta)$, then the ML estimate $\hat{\theta}$ is unique. Differentiating the efficiency condition eq.(8.15) with respect to θ and inserting $\theta = \hat{\theta}$ leads to

$$\left.\frac{\partial^2 \ln L}{\partial \theta^2}\right|_{\theta=\hat{\theta}} = -A(\theta)\left(\frac{d\tau}{d\theta} + \frac{\partial b}{\partial \theta}\right)\Bigg|_{\theta=\hat{\theta}} = -\{A(\hat{\theta})\}^2 V(t) < 0,$$

in virtue of eq.(9.16). Thus every solution of the likelihood equation $\partial \ln L/\partial\theta = 0$ corresponds to a maximum of the LF. Since, for a regular function, there must be a minimum between successive maxima, it follows that there cannot

be more than one maximum in the present case, and hence the solution $\hat{\theta}$ is unique. The proof of uniqueness can in fact be extended to all cases where a single sufficient statistic exists. If there is no sufficient statistic, then the ML estimator is not necessarily unique.

For k parameters, when there is a set of k jointly sufficient statistics, the solution of the set of likelihood equations is unique, provided that the usual regularity conditions are satisfied.

9.4.7 Asymptotic normality of ML estimators

Let us consider again the one-parameter case with p.d.f. $f(x|\theta)$ given n observations $x_1, x_2, \ldots, x_n$. We will outline the proof that for large samples the estimate $\hat{\theta}$ is asymptotically normally distributed about the true value $\theta = \theta_o$ with variance equal to the MVB, provided lnL is twice differentiable in θ and the range of x is independent of θ.

A Taylor expansion about the true value $\theta = \theta_o$ of $\partial \ln L/\partial\theta$ at the ML estimate $\theta = \hat{\theta}$ gives

$$0 = \left.\frac{\partial \ln L}{\partial \theta}\right|_{\hat{\theta}} = \left.\frac{\partial \ln L}{\partial \theta}\right|_{\theta_o} + (\hat{\theta} - \theta_o)\left.\frac{\partial^2 \ln L}{\partial \theta^2}\right|_{\theta^*}, \tag{9.17}$$

where θ^* is some value between $\hat{\theta}$ and θ_o.

It has previously been established in Sect.8.5 that, when L is sufficiently regular,

$$\int \cdots \int \frac{\partial \ln L}{\partial \theta} L \, d\underline{x} = E\left(\frac{\partial \ln L}{\partial \theta}\right) = 0. \tag{8.8}$$

Therefore, from eq.(8.12),

$$V\left(\frac{\partial \ln L}{\partial \theta}\right)\Bigg|_{\hat{\theta}} = E\left[\left(\frac{\partial \ln L}{\partial \theta}\right)^2\right] = E\left(-\frac{\partial^2 \ln L}{\partial \theta^2}\right). \tag{9.18}$$

Writing

$$\left.\frac{\partial \ln L}{\partial \theta}\right|_{\theta_o} = \sum_{i=1}^{n} \left.\frac{\partial \ln f(x_i|\theta)}{\partial \theta}\right|_{\theta_o}, \tag{9.19}$$

the right hand side is a sum of n independent terms $\partial \ln f(x_i|\theta)/\partial\theta$, which has a

mean value of zero and a variance given by eq.(9.18). From the Central Limit Theorem the quantity

$$u = \frac{\partial \ln L}{\partial \theta} \Big/ \left[E\left(- \frac{\partial^2 \ln L}{\partial \theta^2} \right) \right]^{\frac{1}{2}} \tag{9.20}$$

at $\theta = \theta_o$ is a standardized variable with asymptotic distribution N(0,1). The quantity

$$v = - \left. \frac{\partial^2 \ln L}{\partial \theta^2} \right|_{\theta^*} = \sum_{i=1}^{n} \left(- \left. \frac{\partial^2 \ln f}{\partial \theta^2} \right) \right|_{\theta^*}, \tag{9.21}$$

in view of the Law of Large Numbers, is such that, when $n \to \infty$,

$$v \to E\left(- \frac{\partial^2 \ln L}{\partial \theta^2} \right) \tag{9.22}$$

since θ^* lies between $\hat{\theta}$ and θ_o, and the ML estimate $\hat{\theta}$ converges towards θ_o (consistency). Introducing the asymptotic value of v in eq.(9.17) we obtain

$$(\hat{\theta} - \theta_o) \left[E\left(- \frac{\partial^2 \ln L}{\partial \theta^2} \right) \right]^{\frac{1}{2}} = \frac{\partial \ln L}{\partial \theta} \Big/ \left[E\left(- \frac{\partial^2 \ln L}{\partial \theta^2} \right) \right]^{\frac{1}{2}}, \tag{9.23}$$

where the expressions should be evaluated at $\theta = \theta_o$. The quantity on the right is recognized as u from eq.(9.20), which is asymptotically normal with zero mean and unit variance. Therefore the estimate $\hat{\theta}$ must be asymptotically normally distributed about the true value θ_o with variance equal to the MVB,

$$V(\hat{\theta}) = 1 / E\left(- \frac{\partial^2 \ln L(\underline{x}|\theta)}{\partial \theta^2} \right). \tag{9.24}$$

For the case with k parameters $\theta_1, \theta_2, \ldots, \theta_k$ one can prove that the estimates $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$ are asymptotically multinormally distributed about the true parameter values and with covariances given by

$$V_{ij}^{-1}(\hat{\underline{\theta}}) = E\left(- \frac{\partial^2 \ln L(\underline{x}|\underline{\theta})}{\partial \theta_i \partial \theta_j} \right), \qquad i,j = 1,2,\ldots,k. \tag{9.25}$$

9.4.8 Example: Asymptotic normality of the ML estimator of the mean lifetime

We have in the last sections indicated that the ML estimator of the mean lifetime τ in the p.d.f. $f(t|\tau) = \frac{1}{\tau} e^{-t/\tau}$ has a number of optimum properties: it is unique, consistent, unbiassed, sufficient, efficient, and also

invariant under transformation to the decay constant $\lambda = \frac{1}{\tau}$. The estimator of τ can in this respect be regarded superior to the estimator of λ, which is biassed and not efficient.

We shall now explicitly prove that the ML estimate $\hat{\tau} = \bar{t}$ for large samples is normally distributed about the true value τ with variance equal to the MVB. The function $f(t|\tau)$ is twice differentiable with respect to τ and the range of variation for t is independent of τ so the conditions for asymptotic normality are fulfilled. Writing x for the variable instead of t, the p.d.f. $f(x|\tau)$ has the characteristic function $\Phi_x(t)$,

$$\Phi_x(t) \equiv E(e^{itx}) = (1 - it\tau)^{-1}.$$

The characteristic function $\Phi_{\bar{x}}(t)$ for the variable $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ with the p.d.f. $L(\underline{x}|\tau) = \prod_{i=1}^{n} f(x_i|\tau)$ is

$$\Phi_{\bar{x}}(t) = E(e^{it\bar{x}}) = E\left(\exp\left(it \frac{1}{n} \sum_{i=1}^{n} x_i \right) \right).$$

Since the observations are independent this may be written

$$\Phi_{\bar{x}}(t) = \prod_{i=1}^{n} E\left(\exp\left(i \frac{t}{n} x_i \right) \right) = \left(\Phi_x\left(\frac{t}{n} \right) \right)^n = \left(1 - \frac{it\tau}{n} \right)^{-n},$$

When n goes towards infinity, $\Phi_{\bar{x}}(t) \to \exp(it\tau)$, which according to Sect.4.8.4 is the characteristic function for a normal variable with mean value τ and variance zero. Hence the absolute limiting distribution for the ML estimate $\hat{\tau} = \bar{x}$ may be interpreted as an infinitely sharp peak at τ.

If we want to say something about the distribution for n large, but not necessarily infinite, we must study the expansion of $\Phi_{\bar{x}}(t)$ in more detail. We have

$$\left(1 - \frac{it\tau}{n} \right)^{-n} = 1 + it\tau + \tfrac{1}{2}(it)^2(\tau^2 + \tau^2/n) + \ldots$$

and may compare with the characteristic function for a normal variable with mean μ and variance σ^2,

$$\exp\left(it\mu + \tfrac{1}{2}(it)^2\sigma^2 \right) = 1 + it\mu + \tfrac{1}{2}(it)^2(\mu^2 + \sigma^2) + \ldots$$

It is seen, therefore, that $\hat{\tau} = \bar{x}$, for large but finite n, will have a normal distribution with mean value τ and variance τ^2/n, which is equal to the MVB. Accordingly, the standardized quantity

$$u = \frac{\hat{\tau} - \tau}{\tau/\sqrt{n}}$$

is N(0,1) for large samples.

9.5 VARIANCE OF MAXIMUM-LIKELIHOOD ESTIMATORS

How to make the best determination of the errors in the Maximum-Likelihood estimates will depend on the properties of the probability density function and on the size of experimental sample. Since the variances for the estimated parameters are so important, we will explain very explicitly under what conditions the different methods for variance determination should be used. Some formulae are valid only for large samples, which means that they are exact only in the limit when n goes towards infinity. However, many of these socalled *large sample formulae* give good approximations also when n is finite but reasonably large. Other formulae are valid for samples of all sizes; these are often referred to as *small sample formulae*.

When the experimental resolution has been folded into the p.d.f. (see Sect.6.2) the errors calculated from the likelihood function will contain the statistical as well as the experimental uncertainties. If the resolution is not included in the p.d.f., the errors estimated from an ideal theoretical distribution obviously only reflect the statistical uncertainty and not the errors in the measurements. When the estimation of the errors is done directly from the observed data as for instance by the graphical method described later in Sect. 9.6, these errors will clearly contain both the experimental and the statistical uncertainties.

9.5.1 General methods for variance estimation

Let us now regard the likelihood function $L(\underline{x}|\underline{\theta}) = \prod_{i=1}^{n} f(x_i|\underline{\theta})$ as the joint p.d.f. of the n variables $x_1, x_2, \ldots, x_n$ for the k parameters $\theta_1, \theta_2, \ldots, \theta_k$. If the estimates can be written explicitly as functions of the x_i's, *i.e.* $\hat{\theta}_i = \hat{\theta}_i(x_1, x_2, \ldots, x_n)$, the covariance term between $\hat{\theta}_i$ and $\hat{\theta}_j$ may be defined as

$$V_{ij}(\underline{\hat{\theta}}) = \int (\hat{\theta}_i - \theta_i)(\hat{\theta}_j - \theta_j) L(\underline{x}|\underline{\theta}) d\underline{x} \,, \tag{9.26}$$

where the integration is over all n x_i's and $\underline{\theta} = \{\theta_1, \theta_2, \ldots, \theta_k\}$ represent the true values of the parameters. The formula above can be used to find the covariance matrix from the given $f(x|\underline{\theta})$ alone without having any data available. An example on the use of eq.(9.26) is given in Sect.9.5.2.

With eq.(9.26) an equivalent formula for the covariance terms can be obtained where the integrations are to be done with $\underline{\hat{\theta}}$ as variables. The joint probability $L(\underline{x}|\underline{\theta})d\underline{x}$ with n variables must be transformed to the joint probability $L'(\underline{\hat{\theta}}|\underline{\theta})d\underline{\hat{\theta}}$ with k variables. Introducing the Jacobian for the chosen transformation and integrating out a number of (n-k) dummy variables the result is

$$V_{ij}(\underline{\hat{\theta}}) = \int (\hat{\theta}_i - \theta_i)(\hat{\theta}_j - \theta_j) L'(\underline{\hat{\theta}}|\underline{\theta}) d\underline{\hat{\theta}} \,, \tag{9.27}$$

where the new likelihood function L' includes the Jacobian. The transformation from the variables $\underline{x}$ to the variables $\underline{\hat{\theta}}$ is often complicated, and it may be easier to find the covariance from the former expression eq.(9.26). It should be noted that the analytical integration of formulae (9.26), (9.27) can only be carried out in rather favourable cases. They will lead to the covariances expressed as functions of the true (constant) parameters.

Let us now consider the likelihood function $L(\underline{x}|\underline{\theta})$ as a function of $\underline{\theta}$ for given $\underline{x}$. Since $L(\underline{x}|\underline{\theta})$ is normalized to one over the sample space, it is generally not normalized over the parameter space. The variances may then be evaluated from the alternative formula *)

$$V_{ij}(\underline{\hat{\theta}}) = \frac{\int (\theta_i - \hat{\theta}_i)(\theta_j - \hat{\theta}_j) L(\underline{x}|\underline{\theta}) d\underline{\theta}}{\int L(\underline{x}|\underline{\theta}) d\underline{\theta}} \,, \tag{9.28}$$

where the integrations are extended over all k parameters. Again, in fortunate situations, it may be possible to carry out parts of the integrations analytically, or remove common factors in the denominator and numerator. In the general case, approximative values of the covariance between any pair of parameters

*) Equations (9.28),(9.29) formally consider the LF as providing a measure of the distribution of the *variables* $\underline{\theta}$; a discussion of the conceptual difficulties on this point is deferred to Sect.9.7.

is found by numerical integrations of the type

$$V_{ij}(\underline{\theta}) = \frac{1}{N} \sum_{\Delta\theta_\ell} \cdots \sum_{\Delta\theta_m} (\theta_i - \hat{\theta}_i)(\theta_j - \hat{\theta}_j) L(\underline{x}|\underline{\theta}) \Delta\theta_\ell \cdots \Delta\theta_m \qquad (9.29a)$$

where the $\Delta\theta$'s are the proper bin widths for the required integrations over the different parameters, and

$$N = \sum_{\Delta\theta_\ell} \cdots \sum_{\Delta\theta_m} L(\underline{x}|\underline{\theta}) \Delta\theta_\ell \cdots \Delta\theta_m \qquad (9.29b)$$

is a common overall normalization factor.

If, in the last formulation, some of the parameters are not integrated over but kept fixed at their estimated values, this will correspond to obtaining *conditional* errors and covariances for the remaining parameters. This procedure, which gives smaller errors than the preceding approach, is computationally simpler and is, in fact, the standard used in some popular optimization programmes (see Chapter 13). Whenever such a procedure is adopted, this should be explicitly stated to avoid misinterpretation of the numerical results.

9.5.2 Example: Variance of the lifetime estimate

Using eq.(9.26) the variance of the ML estimate $\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} t_i$ for the one-parameter p.d.f. $f(t|\tau) = \frac{1}{\tau} e^{-t/\tau}$ is

$$V(\hat{\tau}) = \int_0^\infty \cdots \int_0^\infty (\hat{\tau}-\tau)^2 \prod_{i=1}^{n} \frac{1}{\tau} e^{-t_i/\tau} dt_i .$$

This can be written

$$V(\hat{\tau}) = \int \left(\frac{1}{n} \sum_{k=1}^{n} t_k\right)\left(\frac{1}{n} \sum_{j=1}^{n} t_j\right) \prod_{i=1}^{n} \frac{1}{\tau} e^{-t_i/\tau} dt_i$$
$$- 2\tau \int \left(\frac{1}{n} \sum_{k=1}^{n} t_k\right) \prod_{i=1}^{n} \frac{1}{\tau} e^{-t_i/\tau} dt_i + \tau^2 \int \prod_{i=1}^{n} \frac{1}{\tau} e^{-t_i/\tau} dt_i ,$$

where the integrations are from zero to infinity for all the n variables t_i. A straightforward computation gives

$$V(\hat{\tau}) = \left(\frac{2}{n}\tau^2 + \frac{n-1}{n}\tau^2\right) - (2\tau^2) + (\tau^2) = \tau^2/n .$$

This $V(\hat{\tau})$ is the same as the MVB derived in Sect.9.4.5 for the efficient estimator $\bar{t}$. The alternative formula, eq.(9.28), gives

$$V(\hat{\tau}) = \left[\int_0^\infty (\hat{\tau}-\tau)^2 \left(\prod_{i=1}^{n} \frac{1}{\tau} e^{-t_i/\tau}\right) d\tau\right] \Big/ \left[\int_0^\infty \left(\prod_{i=1}^{n} \frac{1}{\tau} e^{-t_i/\tau}\right) d\tau\right] ,$$

which leads to

$$V(\hat{\tau}) = \frac{\hat{\tau}^2}{n} \cdot \frac{1 + \frac{6}{n}}{\left(1 - \frac{2}{n}\right)\left(1 - \frac{3}{n}\right)} .$$

Thus, only for infinitely large n do the results of the two procedures, eqs. (9.26) and (9.28), coincide.

Exercise 9.9: Consider the p.d.f. $f(t|\lambda) = \lambda e^{-\lambda t}$. Show, using the approximation of eq.(9.28), that the variance of the ML estimate $\hat{\lambda}$ is

$$V(\hat{\lambda}) = \left(\hat{\lambda}^2/n\right)\left(1 + \frac{2}{n}\right)$$

Compare this with the result of Exercise 9.7.

Exercise 9.10: Show, using eq.(9.26), that the variance $V(\hat{\mu})$ of the ML estimate $\hat{\mu} = \bar{x}$ of the mean in the normal distribution $N(\mu,\sigma^2)$ is equal to the MVB, (Sect.8.5.2). What is the result obtained from eq.(9.28)?

Exercise 9.11: From eq.(9.26), find the variance $V(\hat{\sigma^2})$ of the ML estimate $\hat{\sigma^2} = 1/n \sum (x_i-\mu)^2$ of the parameter σ^2 in $N(\mu,\sigma^2)$. (Compare Exercise 9.3)).

Exercise 9.12: Find, using eq.(9.28) the covariance matrix of the simultaneous ML estimates $\hat{\mu} = \bar{x}$, $\hat{\sigma^2} = 1/n \sum (x_i-\bar{x})^2$ of the mean and variance in $N(\mu,\sigma^2)$, (Sect.9.2.3). Hint: The LF can be written

$$L(\underline{x}|\mu,\sigma^2) = \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\tfrac{1}{2}(x_i-\mu)^2/\sigma^2\right)\right) = (2\pi\sigma^2)^{-\frac{1}{2}n} \exp\left(-\tfrac{1}{2}n(\hat{\mu}-\mu)^2/\sigma^2 - \tfrac{1}{2}n\hat{\sigma^2}/\sigma^2\right)$$

9.5.3 Variance of sufficient ML estimators

The formulae given in Sect.9.5.1 are generally valid for all ML estimators, irrespective of the sample size. In specific cases more convenient formulae may be developed, and we turn now to the situation where the ML estimators are sufficient.

When the p.d.f. $f(x|\theta)$ provides a single sufficient statistic, and consequently an efficient estimator, for the parameter θ, we have already seen

in Sect.9.4.5 that the variance of the ML estimator is given by the minimum variance bound,

$$V(\hat{\theta}) = \left(1 + \frac{\partial b}{\partial \theta}\right)^2 \Big/ E\left(-\frac{\partial^2 \ln L}{\partial \theta^2}\right),$$

where $b(\theta)$ is the bias of the estimator. However, it is not necessary to evaluate the expectation value of $-\partial^2 \ln L/\partial\theta^2$ in this case. From the efficiency condition

$$\frac{\partial \ln L}{\partial \theta} = A(\theta)\{t-\theta-b(\theta)\} \tag{8.19}$$

one finds

$$E\left(-\frac{\partial^2 \ln L}{\partial \theta^2}\right) = A(\theta)\left(1 + \frac{\partial b}{\partial \theta}\right) = -\left.\frac{\partial^2 \ln L}{\partial \theta^2}\right|_{\theta=\hat{\theta}}.$$

Thus,

$$V(\hat{\theta}) = \left(1 + \frac{\partial b}{\partial \theta}\right)^2 \Big/ \left(-\frac{\partial^2 \ln L}{\partial \theta^2}\right)_{\theta=\hat{\theta}}, \tag{9.30}$$

or, for an unbiassed estimator simply

$$V(\hat{\theta}) = 1\Big/\left(-\frac{\partial^2 \ln L}{\partial \theta^2}\right)_{\theta=\hat{\theta}}. \tag{9.31}$$

It should be noted that these relations hold for small samples as well as for large samples, and in particular eq.(9.31) is very useful in practice.

In the multi-parameter case the situation is not so simple. If, however, there exists a set of k jointly sufficient statistics $t_1, t_2, \ldots, t_k$ for the k parameters $\theta_1, \theta_2, \ldots, \theta_k$, it can be shown that the inverse of the covariance matrix of the ML estimates *in large samples* is given by

$$V^{-1}_{ij}(\underline{\hat{\theta}}) = \left(-\frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j}\right)_{\underline{\theta}=\underline{\hat{\theta}}}. \tag{9.32}$$

This is a natural generalization of eq.(9.31). In situations where we do not know k jointly sufficient statistics for the k parameters and where the number of observations is not large, the covariance matrix of the ML estimates may still be found by one of the general methods described in Sect.9.5.1.

Exercise 9.13: It was shown in Chapter 8 that $\bar{x}$ is an unbiassed and efficient estimator of μ in the normal distribution. Find the variance of the ML estimate $\hat{\mu} = \bar{x}$ from eq.(9.31).

Exercise 9.14: The distribution of gap lengths x between adjacent bubbles formed along the tracks of charged particles in a bubble chamber is given by

$$f(x|g) = g\, e^{-gx}, \qquad 0 \le x \le \infty.$$

The mean gap length 1/g is often used as a measure of the ionization. Show that the relative statistical uncertainty for n measured gap lengths is $\Delta(1/g)/(1/g) = 1/\sqrt{n}$.

9.5.4 Example: Variance of the weighted mean

In Sect.9.2.2 we found that the ML estimate for the unknown mean μ of the normally distributed observations $x_1, x_2, \ldots, x_n$ with errors $\sigma_1, \sigma_2, \ldots, \sigma_n$ is given by the weighted mean of the x_i,

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i/\sigma_i^2}{\sum_{i=1}^{n} 1/\sigma_i^2}. \tag{9.11}$$

Since the weighted mean is an unbiassed and efficient ML estimator of μ (compare Exercise 9.8) and

$$\ln L = \sum_{i=1}^{n}\left(-\tfrac{1}{2}\ln(2\pi\sigma_i^2) - \tfrac{1}{2}\left(\frac{x_i-\mu}{\sigma_i}\right)^2\right),$$

the variance can be found from eq.(9.31),

$$V(\hat{\mu}) = 1\Big/\left(-\frac{\partial^2 \ln L}{\partial \mu^2}\right)_{\mu=\hat{\mu}} = 1\Big/\sum_{i=1}^{n}\frac{1}{\sigma_i^2}. \tag{9.12}$$

When the errors σ_i are all equal, $\sigma_i = \sigma$, eq.(9.12) leads to the well-known expression for the error on the mean, $\Delta\hat{\mu} = \sigma/\sqrt{n}$.

9.5.5 Example: Errors in the ML estimates of μ and σ^2 in $N(\mu,\sigma^2)$

In Exercise 9.12 we suggested that the approximate relation eq.(9.28) be used to find the covariance matrix of the ML estimates $\hat{\mu}$ and $\hat{\sigma}^2$ of the mean and variance in the normal distribution $N(\mu,\sigma^2)$. Instead of carrying out the integration of the likelihood function over the parameter space a much easier way to obtain the covariances is to use the fact that $\hat{\mu} = \bar{x}$ and $\hat{\sigma}^2 = 1/n\Sigma(x_i-\bar{x})^2$

are two jointly sufficient statistics for the parameters in the normal p.d.f. (compare Sect.8.6.4). It is therefore appropriate to apply eq.(9.32) in the present case, provided n is large. We have now

$$\ln L = \sum_{i=1}^{n}\left(-\tfrac{1}{2}\ln(2\pi) - \tfrac{1}{2}\ln\sigma^2 - \tfrac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right),$$

so that the elements of the inverse of the covariance matrix $V(\hat{\mu},\hat{\sigma^2})$ become from eq.(9.32),

$$V_{11}^{-1} = -\frac{\partial^2 \ln L}{\partial \mu^2} = \frac{n}{\sigma^2},$$

$$V_{22}^{-1} = -\frac{\partial^2 \ln L}{\partial (\sigma^2)^2} = -\frac{n}{2\sigma^4} + \frac{1}{\sigma^6}\sum_{i=1}^{n}(x_i-\mu)^2 = -\frac{n}{2\sigma^4} + \frac{n\sigma^2}{\sigma^6} = \frac{n}{2\sigma^4},$$

$$V_{12}^{-1} = V_{21}^{-1} = -\frac{\partial^2 \ln L}{\partial \mu \partial \sigma^2} = \frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i-\mu) = \frac{n}{\sigma^4}(\hat{\mu}-\mu),$$

where it is understood that the expressions should be evaluated for $\mu = \hat{\mu}$ and $\sigma^2 = \hat{\sigma^2}$. Accordingly, the inverse of the covariance matrix is

$$V^{-1}(\hat{\mu},\hat{\sigma^2}) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix},$$

which can be inverted to give

$$V(\hat{\mu},\hat{\sigma^2}) = \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{pmatrix}. \qquad (9.33)$$

It is interesting to note that this asymptotic covariance matrix for the ML estimates $\hat{\mu}$ and $\hat{\sigma^2}$ is diagonal. This need not surprise us since we know that $\bar{x}(=\hat{\mu})$ and s^2 $(=\frac{n}{n-1}\hat{\sigma^2})$ from normal samples are independent variables (Sect.4.8.6). Now, since $\bar{x}$ is $N(\mu,\sigma^2/n)$, it is clear that the ML estimate $\hat{\mu}$ will, for all n, have a variance σ^2/n, the MVB. From Sect.5.1.6 we know that $(n-1)s^2/\sigma^2$, and hence $n\hat{\sigma^2}/\sigma^2$, is distributed according to $\chi^2(n-1)$, and has a variance $2(n-1)$. Therefore, $V(\hat{\sigma^2}) = (\sigma^2/n)^2 V(n\hat{\sigma^2}/\sigma^2) = (\sigma^2/n)^2 2(n-1)$ $= 2\sigma^4(n-1)/n^2$, which holds strictly for all n. This is close to the result $2\sigma^4/n$ implied by the asymptotic formula (9.33).

9.5.6 Variance of large-sample ML estimators

It was shown in Sect.9.4.7 that the ML estimate, under very general conditions, will be asymptotically normally distributed about the true value and with variance equal to the MVB. The MVB formulae eqs.(9.31) and (9.32) can then be reformulated in terms of the p.d.f. and used, on the planning stage of the experiment, to predict the errors that must be expected.

For the one-parameter p.d.f. $f(x|\theta)$ the negative of the second derivative of lnL with respect to θ can generally be expressed as

$$-\frac{\partial^2 \ln L}{\partial \theta^2} = \sum_{i=1}^{n}\left(-\frac{\partial^2 \ln f(x_i|\theta)}{\partial \theta^2}\right),$$

and the expectation, or average, of this variable is

$$E\left(-\frac{\partial^2 \ln L}{\partial \theta^2}\right) = \int \sum_{i=1}^{n}\left(-\frac{\partial^2 \ln f}{\partial \theta^2}\right) f\,dx = n\int\left(-\frac{\partial^2 \ln f}{\partial \theta^2}\right) f\,dx.$$

For the first part of the integrand on the right-hand side we have

$$\left(-\frac{\partial^2 \ln f}{\partial \theta^2}\right) = -\frac{\partial}{\partial \theta}\left(\frac{1}{f}\frac{\partial f}{\partial \theta}\right) = \frac{1}{f^2}\left(\frac{\partial f}{\partial \theta}\right)^2 - \frac{1}{f}\frac{\partial^2 f}{\partial \theta^2},$$

and, provided the integration limits for x are independent of θ, therefore

$$E\left(-\frac{\partial^2 \ln L}{\partial \theta^2}\right) = n\int \frac{1}{f}\left(\frac{\partial f}{\partial \theta}\right)^2 dx.$$

Substituting this expression in eq.(9.31) we find the following convenient formula for the average of the inverse of the variance,

$$\overline{V^{-1}(\hat{\theta})} = n\int \frac{1}{f}\left(\frac{\partial f}{\partial \theta}\right)^2 dx. \qquad (9.34)$$

In the multi-parameter case the corresponding formula for the covariance terms, expressed by the p.d.f., is

$$\overline{V_{ij}^{-1}(\hat{\underline{\theta}})} = n\int \frac{1}{f}\left(\frac{\partial f}{\partial \theta_i}\right)\left(\frac{\partial f}{\partial \theta_j}\right) dx. \qquad (9.35)$$

The fact that the ML estimate is asymptotically normally distributed about the true parameter value can, in view of the formal symmetry between variable and mean value in the normal p.d.f., be formally expressed as the

parameter being asymptotically normally distributed about the ML estimate, with a spread around this mean value as implied by the MVB. Writing the LF for the one-parameter case as

$$L \propto \exp\left(-\tfrac{1}{2}\,\frac{(\theta-\hat{\theta})^2}{V(\hat{\theta})}\right),$$

we find at once the simple relationship

$$V(\hat{\theta}) = 1 \Big/ \left(-\frac{\partial^2 \ln L}{\partial\theta^2}\right), \tag{9.36}$$

which should be compared to eq.(9.31).

Clearly, a normal-shaped LF corresponds to a parabolic dependence of lnL on θ, and a constant second derivative.

In the multi-parameter case the LF has asymptotically a multinormal shape. The covariance matrix for $\hat{\underline{\theta}}$ is found by inversion of the matrix with elements given by the constants

$$V^{-1}_{ij}(\hat{\underline{\theta}}) = -\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\,. \tag{9.37}$$

9.5.7 Example: Planning of an experiment; polarization (1)

In the preceding sections several methods have been presented on how to find the variances of ML estimates. Some of these methods can only be used after the experimental data have been collected, while others may be applied already on the planning stage, prior to the data taking.

Suppose we want to study the polarization of antiprotons in antiproton-proton elastic scattering from a double scattering experiment measuring the angle ϕ between the normals of the scattering planes. The p.d.f. is

$$f(x|\alpha) = \tfrac{1}{2}(1 + \alpha x) \qquad -1 \le x \le 1 \tag{9.38}$$

where $x = \cos\phi$, $\alpha = P^2$. We ask: How many events will be needed to obtain a prescribed accuracy on the estimate of α?

For large n the variance of $\hat{\alpha}$ can be calculated from the asymptotic formula, eq.(9.34). We first evaluate the integral

$$\int_{-1}^{+1} \frac{1}{f}\left(\frac{\partial f}{\partial\alpha}\right)^2 dx = \int_{-1}^{+1} \tfrac{1}{2}\,\frac{x^2}{1+\alpha x}\,dx = \frac{1}{2\alpha^3}\left(\ln(1+\alpha) - \ln(1-\alpha) - 2\alpha\right).$$

Hence, from eq.(9.34),

$$V(\hat{\alpha}) = \frac{1}{n}\,\frac{2\alpha^3}{\ln(1+\alpha) - \ln(1-\alpha) - 2\alpha}\,. \tag{9.39}$$

When $\alpha \ll 1$ this can be written

$$V(\hat{\alpha}) \ge \frac{1}{n}\left(3 - \frac{9}{5}\,\alpha^2 + \ldots\right). \tag{9.40}$$

If, prior to the experiment, α is assumed to be approximately 0.1 and we wish to determine the parameter with an uncertainty $\Delta\alpha = 0.01$, we find from eq.(9.40) that more than $3\cdot10^4$ events will be needed.

Exercise 9.15: In the preceding example, what can be said about $V(\hat{\alpha})$ if n is not very large?

9.5.8 Example: Planning of an experiment; density matrix elements (1)

We will now look at a specific example to illustrate how the ML method can be applied to problems where the p.d.f. has a physical origin rather than being of a standard mathematical type.

Consider a meson resonance with spin parity $J^P = 1^-$ decaying into two pseudoscalar mesons with $J^P = 0^-$. The angular distribution of the decay particles, described in the Jackson reference system by the polar angle θ and the azimuth angle ϕ, is

$$f(\cos\theta,\phi|\rho_{00},\rho_{1-1},\mathrm{Re}\rho_{10}) = \frac{3}{4\pi}\left(\tfrac{1}{2}(1-\rho_{00}) + \tfrac{1}{2}(3\rho_{00}-1)\cos^2\theta - \rho_{1-1}\sin^2\theta\cos2\phi - \sqrt{2}\,\mathrm{Re}\rho_{10}\sin2\theta\cos\phi\right), \tag{9.41}$$

where f is properly normalized, since

$$\int_0^{2\pi} d\phi \int_{-1}^{+1} d\cos\theta\, f(\cos\theta,\phi|\rho_{00},\rho_{1-1},\mathrm{Re}\rho_{10}) = 1\,.$$

For each event there are two measured quantities, $\cos\theta_i$ and ϕ_i. The density matrix elements to be estimated are ρ_{00}, ρ_{1-1}, and $\mathrm{Re}\rho_{10}$. The three simultaneous likelihood equations are of order n in the parameters, and the ML estimates can therefore not be found analytically. Hence a numerical procedure is necessary to find the estimates $\widehat{\rho_{00}}$, $\widehat{\rho_{1-1}}$, $\widehat{\mathrm{Re}\rho_{10}}$ and their errors. We may

now ask: Can anything be said about the covariance matrix of the parameters before data are available? It can be verified that for the distribution (9.41) there exists no set of jointly sufficient statistics for the three parameters, or for any combination of two of them. Nor is there a single sufficient statistic for any of the parameters alone. For small samples of data, therefore, little can be said about the errors from the theoretical p.d.f. For large samples, however, the asymptotic covariance terms may in principle be calculated from eq.(9.35).

Alternatively to the above description of the density matrix elements in the Jackson reference system, the spin structure can be described in the so-called dynamic reference system where the "observable" part of the density matrix is diagonal. For vector particles the density matrix can be diagonalized by rotating the Jackson reference system a certain angle Θ about the y-axis. The three independent parameters in the dynamic reference system, are called α, β, and θ. The decay distribution is of the form

$$f = \frac{3}{4\pi}\left[\alpha(\underline{e}\cdot\underline{i})^2 + \beta(\underline{e}\cdot\underline{j})^2 + (1-\alpha-\beta)(\underline{e}\cdot\underline{k})^2\right], \qquad (9.42)$$

where $\underline{e}$ is a unit vector along the decay direction, and $\underline{i}$, $\underline{j}$, $\underline{k}$ are unit vectors along the axes of the dynamic reference system. In terms of the polar angles in the Jackson frame and the rotation angle Θ this distribution can be expressed as

$$f(\cos\theta,\phi|\alpha,\beta,\theta) = \frac{3}{4\pi}\left[\alpha(\cos\phi\sin\theta\cos\Theta - \cos\theta\sin\Theta)^2\right.$$
$$\left. + \beta(\sin\phi\sin\theta)^2 + (1-\alpha-\beta)(\cos\theta\cos\Theta + \cos\phi\sin\theta\sin\Theta)^2\right]. \qquad (9.43)$$

The ML properties of this p.d.f. are the same as those of the p.d.f. eq.(9.41). In particular the solutions for the ML estimates $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\theta}$ have to be found numerically by some optimization procedure.

We shall in Sect.11.2.3 discuss how the parameters in eq.(9.41) can alternatively be obtained by the moments method for parameter estimation.

Exercise 9.16: Show that neither of the two marginal distributions obtained by integrating the p.d.f. (9.41) over ϕ and θ, respectively, will provide any sufficient statistic or any set of jointly sufficient statistics for the unknown parameter(s). Show that the same situation applies for the p.d.f. (9.43).

9.6 GRAPHICAL DETERMINATION OF THE MAXIMUM-LIKELIHOOD ESTIMATE AND ITS ERROR

In many practical problems neither the ML estimate nor its variance can be found analytically. The numerical behaviour of $L(\underline{x}|\underline{\theta})$ as a function of $\underline{\theta}$ can, however, be used to determine the estimate $\hat{\underline{\theta}}$ as well as its error $\Delta\hat{\underline{\theta}}$ graphically when the number of parameters is limited to one or two.

9.6.1 The one-parameter case

When the estimation problem involves a single parameter, one can simply plot the likelihood L for the given observations as a function of θ and read off the ML estimate $\hat{\theta}$ from the graph as that particular value of θ for which the curve peaks. Except for rare situations the curve will have a single maximum, and hence a unique solution for the ML estimate $\hat{\theta}$. If more than one maximum show up within the physically admissible range of θ, one will usually take $\hat{\theta}$ as that value of θ which corresponds to the highest maximum.

In the case of a single maximum, or one dominant maximum well separated from other smaller maxima, one deduces the error in the ML estimate $\hat{\theta}$ by looking up the values of θ for which L has fallen by a factor of $e^{-0.5}$ of its maximum value L(max), as indicated by Fig. 9.1. For a strictly normal LF, shown in Fig. 9.1(a), the two values $\theta = \hat{\theta} \pm \Delta\hat{\theta}$ will correspond to $\mu \pm \sigma$ for a variable distributed according to $N(\mu,\sigma^2)$. The true value of the parameter

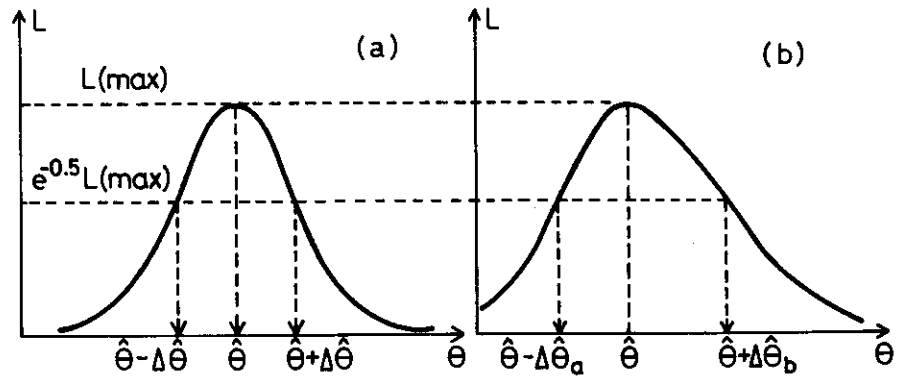

Fig. 9.1. Graphical determination of the ML estimate $\hat{\theta}$ and its error from the one-parametric likelihood function; (a) a symmetric, normal (Gaussian) LF, (b) an unsymmetric LF.

then has a probability 0.683 of being larger than $\hat{\theta} - \Delta\hat{\theta}$ and smaller than $\hat{\theta} + \Delta\hat{\theta}$.

A strictly normal-shaped LF is never attained in real experiments, where the number of observations is finite. In the general case, with an unsymmetric LF, the two values of θ for which $L = L(max)e^{-0.5}$ will give unequal errors on the upper and lower sides of $\hat{\theta}$; see Fig. 9.1(b). The probability that the true value of θ lies between $\hat{\theta} - \Delta\hat{\theta}_a$ and $\hat{\theta} + \Delta\hat{\theta}_b$ will still be approximately equal to 0.68. This will be discussed further in Sect.9.7.1.

9.6.2 Example: Scanning efficiency (3)

We have on two earlier occasions examined the question of how to determine the efficiency in a scanning procedure when two scans have been performed, for example, to locate specified types of events in a batch of bubble chamber films. In Sect.2.3.11 we saw how the efficiencies of the individual scans as well as the overall scanning efficiency, and thereby the total number of events, could be determined from the number of events detected in two independent scans. Error formulae were derived in Sect.4.1.3. The reasoning in these sections rested on the assumption that the numbers of events were large, and also that the efficiencies were not too small.

We will now see how the Maximum-Likelihood method can be used to estimate the scanning efficiencies and the total number of events from the double-scan data. The procedure, introduced by D.A. Evans and W.H. Barkas, extracts more information from the data at hand than does the conventional method described earlier. Contrary to the earlier method the ML approach is also applicable to experiments with low statistics.

We shall assume, as before, that the film has been subjected to two independent scans, and that in each scan all events have the same probability of being detected. The two scans have led to the following results,

N_{12} events were found in scan 1 as well as in scan 2,
N_1 events were found in scan 1 but not in scan 2,
N_2 events were found in scan 2 but not in scan 1.

Then,

$N1 = N_1+N_{12}$ is the number of events found in scan 1,
$N2 = N_2+N_{12}$ is the number of events found in scan 2,
$N12 = N_1+N_2+N_{12}$ is the total number of events found,
N (unknown) is the total number of events in the film,
N-N12 is the number of undetected events in the film.

The scanning process is binomial, since an event is either found by the scanner, or it is not. Therefore, if ε_1 is the probability to observe an event in scan 1, the probability for finding just N1 out of N events in this scan is, according to the binomial distribution law,

$$P_1(N1|N,\varepsilon_1) = \frac{N!}{N1!(N-N1)!}\,\varepsilon_1^{N1}(1-\varepsilon_1)^{N-N1}. \qquad (9.44)$$

For the second scan, the observed N2 events can be divided into two groups, (I) the N_{12} events that have already been found in scan 1, and (II) the N_2 events that have not been recorded in the previous scan. For each of these groups we can apply the binomial distribution law. Thus, for group I the probability to observe N_{12} out of N1 events is

$$P_I(N_{12}|N1,\varepsilon_2) = \frac{N1!}{N_{12}!(N1-N_{12})!}\,\varepsilon_2^{N_{12}}(1-\varepsilon_2)^{N1-N_{12}}. \qquad (9.45)$$

For group II the probability for seeing just N_2 from a total of (N-N1) events that were undetected in scan 1, is

$$P_{II}(N_2|N-N1,\varepsilon_2) = \frac{(N-N1)!}{N_2!(N-N1-N_2)!}\,\varepsilon_2^{N_2}(1-\varepsilon_2)^{N-N1-N_2}. \qquad (9.46)$$

In the expressions for P_1, P_I and P_{II} above the quantities N, ε_1, ε_2 are unknown. The joint probability for seeing N1 events in scan 1, N2 events in scan 2, and N_{12} events in common in the two scans, is the product of the three probabilities,

$$P = P(N1,N2,N_{12}|N,\varepsilon_1,\varepsilon_2) = P_1P_IP_{II}.$$

Organizing factors this can be written

$$P = \frac{N!}{(N-N12)!}\,\varepsilon_1^{N1}\varepsilon_2^{N2}(1-\varepsilon_1)^{N-N1}(1-\varepsilon_2)^{N-N2}\,\frac{1}{N_1!N_2!N_{12}!}, \qquad (9.47)$$

which is symmetric in the indices 1 and 2.

The joint probability P may be interpreted as the likelihood $L(N1,N2,N12|N,\varepsilon_1,\varepsilon_2)$ of the observations $N1,N2,N12 = N1+N2-N_{12}$ for the parameters $N,\varepsilon_1,\varepsilon_2$. One can therefore solve the three simultaneous likelihood equations for these parameters to find their ML estimates $\hat{N},\hat{\varepsilon}_1,\hat{\varepsilon}_2$.

The likelihood equations $\partial \ln L/\partial \varepsilon_1 = 0$ and $\partial \ln L/\partial \varepsilon_2 = 0$ give as expected for the efficiencies of the individual scans

$$\hat{\varepsilon}_1 = \frac{N1}{N}, \qquad \hat{\varepsilon}_2 = \frac{N2}{N}.$$

The likelihood equation $\partial \ln L/\partial N = 0$ can, unfortunately, not be solved analytically. Since, however, we are primarily interested in the number N rather than the individual efficiencies we can integrate the joint probability of eq.(9.47) over the "nuisance" variables $\varepsilon_1,\varepsilon_2$ to obtain a likelihood function involving only the parameter N. The result of this integration is

$$L(N) = \frac{N!(N-N1)!(N-N2)!}{(N-N12)![(N+1)!]^2} \cdot \frac{N1!N2!}{N_1!N_2!N_{12}!} . \qquad (9.48)$$

This form is not particularly suitable for numerical evaluation when the observed numbers are large. However, the expression can be formulated in terms of a recurrence relation,

$$L(N) = \frac{N(N-N1)(N-N2)}{(N-N12)(N+1)^2} \cdot L(N-1) , \qquad (9.49)$$

which is very convenient for numerical calculation. Since the number N of events in the film must be at least as large as the number of events found from the two scans, one can take $L(N-1) = L(N12) = 1$ as a starting value and proceed stepwise to generate as much of the LF as desired.

For a numerical example, suppose that the two independent scans have yielded $N1 = 43$ and $N2 = 48$ events, respectively, with $N_{12} = 25$ events in common. The total number of events recorded by the scans is then $N12 = N1+N2-N_{12} = 66$. Thus we put $L(66) = 1$ and generate new values of L(N) from eq.(9.49) starting with $N = 67$. The result of this computation is shown in Fig. 9.2. From the shape of the LF we conclude that the ML estimate of the unknown number of events is

$$\hat{N} = 81 \pm \frac{8}{6} .$$

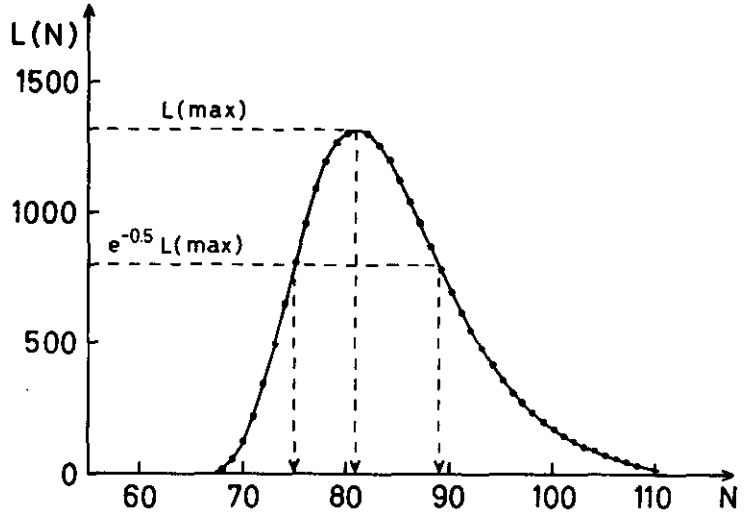

Fig. 9.2. The likelihood function L(N) of eq.(9.49) generated with the numbers $N1=43$, $N2=48$, $N_{12}=25$, starting from $N=67$ ($L(66)=1$).

The estimated efficiencies are

$$\hat{\varepsilon}_1 = \frac{43}{81} = 0.53, \qquad \hat{\varepsilon}_2 = \frac{48}{81} = 0.59.$$

Exercise 9.17: With the observations of the example in the text, what is the total number of events and the scanning efficiencies estimated with the conventional formulae?

9.6.3 The two-parameter case; the covariance ellipse

For the two-parameter case the likelihood function $L(\underline{x}|\theta_1,\theta_2)$ becomes a three-dimensional surface which is less trivial to display. However, the shape of the function can conveniently be visualized by plotting level curves for constant values of $L(\underline{x}|\theta_1,\theta_2)$ in a (θ_1,θ_2) plane, equivalent to drawing intersections between the surface and a set of parallel planes. In the vicinity of a maximum of the function these level curves will be a series of smooth, closed contours around the maximum point $(\hat{\theta}_1,\hat{\theta}_2)$ which can thus be localized to

a required accuracy.

With two parameters the LF will often have more than one maximum. Usually, however, there is little trouble in identifying one particular maximum with the desired ML solution for the parameter estimates. For example, some of the maxima may occur in unphysical regions of the parameter space and can therefore be eliminated right away, or the principal maximum can be overwhelmingly favourable compared to the secondary maxima from the numerical magnitude of their likelihoods. With ill-behaved likelihood functions having two or more maxima of comparable magnitude the ambiguities of the solution may have to be resolved by looking for additional information. An example is given by the case study described in Sect.9.12.

For a regular LF with a single maximum in the parameter region of interest the errors in the ML estimates of the two parameters can be obtained from the specific likelihood contour for which $L = L(\max)e^{-0.5}$. The tangents to this contour parallel to the coordinate axes provide a set of upper and lower errors for the two parameter estimates, as indicated in Fig. 9.3(a). If the LF has the shape of a binormal distribution the errors deduced this way are identical to the standard deviations, as will be demonstrated below.

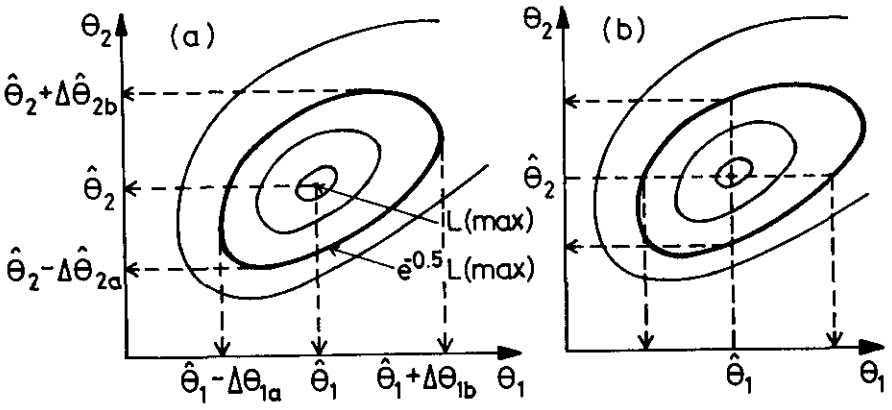

Fig. 9.3. Graphical determination of the ML estimates and their errors from the two-parametric likelihood function; (a) the tangential method, (b) the intersection method (conditional errors).

A second approach determines the "errors" in the parameter estimates from the intersections between the same contour and the two lines $\theta_1 = \hat{\theta}_1$ and $\theta_2 = \hat{\theta}_2$ as indicated in Fig. 9.3(b). With this intersection method the "errors" in either parameter are thus deduced by keeping the other parameter at its estimated value, making these "errors" in general smaller than the errors obtained by the tangential method. Since an asymmetric orientation of the contour $L = L(\max)e^{-0.5}$ relative to the coordinate axes reflects a non-zero correlation between the estimates, these two approaches to error determination will obviously produce identical results in situations with uncorrelated parameters only.

In the *asymptotic limit* of infinitely large samples the LF takes the binormal form

$$L(\theta_1,\theta_2) = L(\max)\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{\theta_1-\hat{\theta}_1}{\sigma_1}\right)^2+\left(\frac{\theta_2-\hat{\theta}_2}{\sigma_2}\right)^2-2\rho\left(\frac{\theta_1-\hat{\theta}_1}{\sigma_1}\right)\left(\frac{\theta_2-\hat{\theta}_2}{\sigma_2}\right)\right]\right\} \quad (9.50)$$

where σ_1^2 and σ_2^2 are the variances for the two ML estimates $\hat{\theta}_1$ and $\hat{\theta}_2$, and ρ their correlation coefficient, as can be verified by calculating the matrix elements $V_{ij}^{-1}(\underline{\hat{\theta}})$ according to eq.(9.37) and inverting the resulting matrix. The contour $L = L(\max)e^{-0.5}$ is now given by a quadratic equation in θ_1 and θ_2,

$$\frac{1}{1-\rho^2}\left[\left(\frac{\theta_1-\hat{\theta}_1}{\sigma_1}\right)^2+\left(\frac{\theta_2-\hat{\theta}_2}{\sigma_2}\right)^2-2\rho\left(\frac{\theta_1-\hat{\theta}_1}{\sigma_1}\right)\left(\frac{\theta_2-\hat{\theta}_2}{\sigma_2}\right)\right] = 1 \quad . \quad (9.51)$$

This is the *covariance ellipse* for the binormal LF. The ellipse is centred at $(\hat{\theta}_1,\hat{\theta}_2)$, and its principal axes make an angle α relative to the coordinate system, where

$$\tan 2\alpha = \frac{2\rho\sigma_1\sigma_2}{\sigma_1^2-\sigma_2^2} \quad . \quad (9.52)$$

Figure 9.4 shows a few examples of covariance ellipses with a common centre and common values of σ_1^2,σ_2^2 but with different ρ. As can be verified from eq.(9.51), the ellipses with such properties are all inscribed in the rectangle defined by the straight lines $\theta_1 = \hat{\theta}_1 \pm \sigma_1$ and $\theta_2 = \hat{\theta}_2 \pm \sigma_2$. In other words, regardless of the value of ρ, the tangents to an ellipse parallel to the coordinate axes will always have distances $\pm\sigma_1, \pm\sigma_2$ from the point $(\hat{\theta}_1,\hat{\theta}_2)$; this serves to justify the tangential method for graphical error determination

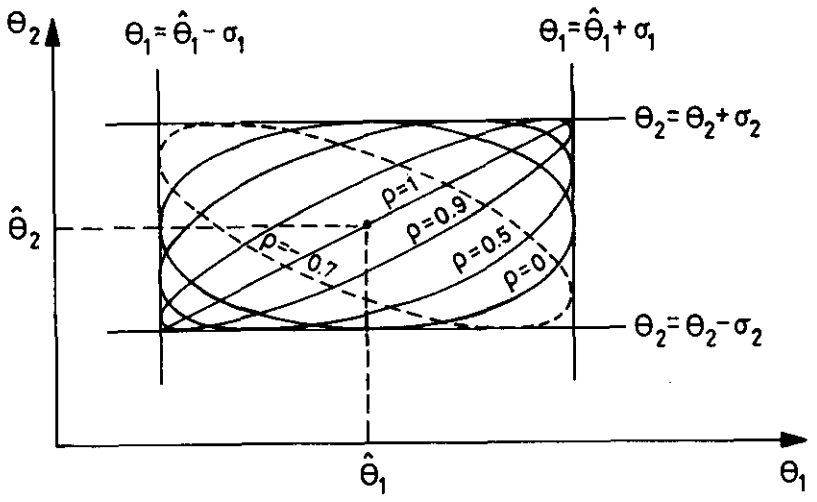

Fig. 9.4. Covariance ellipses for binormal likelihood functions with common maximum $(\hat{\theta}_1,\hat{\theta}_2)$ and common variances $\sigma_1^2=4$, $\sigma_2^2=1$. The ellipses for different values of the correlation coefficient ρ all touch the rectangle defined by $\theta_1=\hat{\theta}_1\pm\sigma_1$, $\theta_2=\hat{\theta}_2\pm\sigma_2$ at four points; for $\rho = \pm 1$ the ellipses degenerate into the diagonals of the rectangle.

(Fig. 9.3(a)). The intersection method for determining the errors, consisting in drawing lines $\theta_1 = \hat{\theta}_1$ and $\theta_2 = \hat{\theta}_2$, is here seen to give intersections with the covariance ellipse at distances $\pm\sigma_1\sqrt{1-\rho^2}$, $\pm\sigma_2\sqrt{1-\rho^2}$ from $(\hat{\theta}_1,\hat{\theta}_2)$. The last observation should make it clear why one must be careful in using this method, since merely quoting the intersection distances as errors will be incomplete, and perhaps even misleading, without a specific statement expressing that these errors are indeed conditional.

It can be shown that determining the errors by the tangents to the contour $L = L(\max)e^{-0.5}$ corresponds to having a probability 0.683 for including the true value of one of the parameters in the interval $[\hat{\theta}_i-\Delta\hat{\theta}_i,\hat{\theta}_i+\Delta\hat{\theta}_i]$ when the second is ignored. The region enclosed by this contour does not represent a joint 68.3% probability for the two parameters, but corresponds to a much lower joint probability, in fact less than 40% in the ideal asymptotic case; this will be discussed further in Sect.9.7.4.

9.7 INTERVAL ESTIMATION FROM THE LIKELIHOOD FUNCTION

The Maximum-Likelihood Principle produces point estimates of the unknown parameters. As it is recognized that there should be a margin of uncertainty associated with an estimate, we have in the previous sections discussed at length how to evaluate its variance. We have seen that the approach to the variance determination is not unique, and that somewhat different results are obtained from the various methods. Therefore, in quoting a result $\hat{\theta} \pm \Delta\hat{\theta}$ for the ML estimate one should indicate how the error was obtained.

Instead of giving the result of an experiment in terms of a point estimate $\hat{\theta}$ and its error $\Delta\hat{\theta}$ one can summarize the outcome of the experiment by performing an interval estimation for the unknown parameter θ. For such a purpose we introduced the concept of a confidence interval in Chapter 7. We associated an estimated interval with a probability content γ, and called it a 100γ % confidence interval for the parameter. The meaning of this was that if the experiment were repeated many times under the same conditions, then, in the long run, the estimated intervals would include the true value of the parameter in 100γ % of these experiments. To arrive at these inferences about the parameter we had to invert probability statements about some function, or statistic, of the observable quantities, whose distributional properties were known. Thus, for instance, inverting a probability statement about the sample mean $\bar{x}$, known to be distributed as $N(\mu,\sigma^2/n)$ with μ unknown, σ^2 known, we obtained a statement expressing that the probability was 0.954 that μ would be larger than $\bar{x} - 2\frac{\sigma}{\sqrt{n}}$ and smaller than $\bar{x} + 2\frac{\sigma}{\sqrt{n}}$; hence we called the random interval $[\bar{x}-2\frac{\sigma}{\sqrt{n}}, \bar{x}+2\frac{\sigma}{\sqrt{n}}]$ a 95.4% confidence interval for μ.

In the asymptotic limit, with *infinite* sample sizes, we can argue in a similar manner to obtain confidence intervals for the unknown parameter θ. We know then that the ML estimate $\hat{\theta}$ has the property of being normally distributed about the true parameter value θ, with a variance given by the minimum variance bound, and this allows us to write down probability statements like

$$P(\theta - 2\sigma < \hat{\theta} < \theta + 2\sigma) = 0.954. \qquad (9.53)$$

The random variable $\hat{\theta}$ here has a probability 0.954 of falling within distance $\pm 2\sigma$ from the true but unknown θ, where σ is implied by the MVB. Rewriting the

inequalities we get the expression

$$P(\hat{\theta} - 2\sigma < \theta < \hat{\theta} + 2\sigma) = 0.954, \qquad (9.54)$$

stating that the probability is 0.954 that the interval $[\hat{\theta}-2\sigma, \hat{\theta}+2\sigma]$ will cover the constant θ. According to the definition of the concept in Chapter 7 the interval $[\hat{\theta}-2\sigma, \hat{\theta}+2\sigma]$ is therefore a 95.4% confidence interval for θ. We note that the inversion of the probability statement (9.53) was particularly simple here, due to the algebraic symmetry between variable and mean value in the normal p.d.f. This asymptotic symmetry persists in the multi-parameter case, and hence permits a similar reasoning with inversion of probability statements to obtain confidence intervals (regions) in the general case with several parameters; see for example the presentation in Chapter 9 in the book by Eadie *et al.*

For *finite* samples we do not know the exact distribution of $\hat{\theta}$. We can therefore not write down statements like eq.(9.53), invert them, and next interpret the results in terms of exact confidence intervals for the unknown constant as we did above.

In the following we shall make use of the likelihood function to perform an interval estimation of θ. We resume and extend our point of view from earlier in this chapter: not only shall we regard the θ value for which the LF is maximal as the "most likely" value of the unknown parameter; other θ values will be considered less likely of being the true value of the parameter, in accordance with the fall-off of the likelihood. Thus, for the set of observations at hand, we shall regard the likelihood function itself as providing a measure of the intensity of our credence in the various conceivable values of the unknown θ. This means that we make an interpretation of the likelihood function as measuring our "degree of belief" in the possible values θ can have, based on our particular observations $x_1, x_2, \ldots, x_n$. Whereas the confidence interval gave a measure for the probability that the true value of the unknown parameter *in the long run* would be included in the estimated interval, an interval estimated from the likelihood function will measure our belief that *the particular set of observations* was generated by a parameter belonging to the estimated interval.

As soon as the likelihood function has been given the intuitive meaning of expressing degree of belief in possible values of θ, it is only natural to associate an interval in θ with a numerical measure of our belief, in proportion to the integral of the LF over this interval. Let us for simplicity assume a normal-shaped LF of mean $\hat{\theta}$ and variance σ^2. Then our belief associated with the specific intervals $[\hat{\theta}-\sigma, \hat{\theta}+\sigma]$, $[\hat{\theta}-2\sigma, \hat{\theta}+2\sigma]$, *etc.* would be proportional to the constants 0.683, 0.954, ···. These numbers therefore provide relative measures of our belief in the specified intervals for θ. We could write, for example,

$$\text{Rel. belief } (\hat{\theta} - 2 < \theta < \hat{\theta} + 2\sigma) = 0.954, \qquad (9.55)$$

which is an expression of the same formal structure as the inverted probability statement eq. (9.54) used to define the 95.4% confidence interval for θ. In the following sections, when we refer to the likelihood function and write

$$P(\theta_a < \theta < \theta_b) = \gamma, \qquad (9.56)$$

we shall take this probability statement to mean "relative belief" in the above sense.

It is customary among physicists to refer to all intervals derived from the likelihood function as confidence intervals. This is in many instances unfortunate, since the meaning is different from what is usually understood by a confidence interval. Intervals obtained by a specific prescription from the likelihood function were originally named *fiducial intervals* by R.A. Fisher. Following a suggestion by D.J. Hudson we shall denote all intervals derived from the likelihood function as *likelihood intervals* to indicate their origin and to distinguish them from confidence intervals which have an entirely different conceptual content.

Finally, let it be mentioned that the use of the likelihood function in statistical inference is by no means a trivial matter. In fact, there has over the years been a great deal of controversy among the specialists, due to their different attitudes to Bayes' Theorem. To indicate how confusion can arise on the subject, let us recall that the likelihood $L(\underline{x}|\theta)$ expresses the

probability to obtain the particular set $\underline{x} = \{x_1, x_2, \ldots, x_n\}$ of observed values, on the condition that the value of the parameter is θ. This probability is connected with the inverse probability $P(\theta|\underline{x})$ for a particular value of θ, given the observations $\underline{x}$, through Bayes' Theorem, which states (eq. (2.26))

$$P(\theta|\underline{x}) \propto L(\underline{x}|\theta)P(\theta),$$

where $P(\theta)$ is the prior probability for θ. In addition to knowing the likelihood $L(\underline{x}|\theta)$ of the observations $\underline{x}$ in our experiment we must have some prior knowledge of the parameter, expressed by $P(\theta)$, in order to obtain the new, posterior probability $P(\theta|\underline{x})$, which in this scheme forms the basis for statistical inference about θ. Thus, we may want to undertake an experiment to get some information on the parameter θ, but unless some previous knowledge already exists on θ, or is simply guessed at, it is not even in principle possible to gain in knowledge by carrying out the experiment. Clearly, the requirement of having previous knowledge in order to learn something new is philosophically disputable. The reader who is interested in underlying philosophies and the relationship between Bayesian and other approaches to interval estimation should consult Kendall and Stuart, Chapters 20 and 21, Vol. 2.

9.7.1 Likelihood intervals; the one-parameter case

Under very general conditions, when the number of observations becomes infinitely large, the likelihood function $L(\underline{x}|\theta) = \prod_{i=n}^{n} f(x_i|\theta)$ gets independent of the sample values $x_1, x_2, \ldots, x_n$ and takes the shape of a normal distribution in θ with mean value $\hat{\theta}$ and variance σ^2 as implied by the MVB. We write

$$L(\underline{x}|\theta) \rightarrow L(\theta) = L(\max)\, e^{-\frac{1}{2}Q}, \tag{9.57}$$

or

$$\ln L(\theta) = \ln L(\max) - \tfrac{1}{2}Q, \tag{9.58}$$

where

$$Q = \left(\frac{\theta - \hat{\theta}}{\sigma}\right)^2. \tag{9.59}$$

With an asymptotic LF of Gaussian form or, equivalently, a parabolic lnL function, intervals for θ can be constructed to correspond to a specified probability content γ, as described for the normal variable in Sect. 4.8.3. In general, two limits θ_a and θ_b can be chosen such as to make

$$P(\theta_a \leq \theta \leq \theta_b) = G\left(\frac{\theta_b - \hat{\theta}}{\sigma}\right) - G\left(\frac{\theta_a - \hat{\theta}}{\sigma}\right) \tag{9.60}$$

where G is the cumulative standard normal distribution. Particularly convenient are the choices which make the intervals symmetric about $\hat{\theta}$, since these are the shortest intervals which have a given probability γ and also produce equal tail probabilities $\frac{1}{2}(1-\gamma)$ in the ends of the distribution *); these can be obtained by simply intersecting the LF or lnL function by straight lines. Thus, the probability statement

$$P(\hat{\theta} - m\sigma \leq \theta \leq \hat{\theta} + m\sigma) = 2\,G(m) - 1 = \gamma \tag{9.61}$$

corresponds to obtaining the symmetric and central m standard deviation likelihood interval $[\hat{\theta}-m\sigma,\ \hat{\theta}+m\sigma]$ of probability γ. The interval is constructed by intersecting the Gaussian $L(\theta)$ by the straight line $L = L(\max)e^{-a}$, or what amounts to the same, by intersecting the parabola lnL by the straight line at distance a below maximum, where $a = \frac{1}{2}Q = \frac{1}{2}m^2$. Specifically,

$$\begin{aligned} &a = 0.5 \text{ gives a } 68.3\% \text{ likelihood interval for } \theta, \\ &a = 2.0 \quad " \quad " \ 95.4\% \quad " \quad " \quad " \ ", \\ &a = 4.5 \quad " \quad " \ 99.7\% \quad " \quad " \quad " \ ". \end{aligned} \tag{9.62}$$

This is illustrated in Fig. 9.5(a).

The above procedure can also be applied when the LF is not of a normal shape (and, equivalently, lnL is not parabolic). Let us assume that $L_\theta(\underline{x}|\theta)$ is a continuous unimodular function of θ and that there exists some transformation $g = g(\theta)$ of the variable θ which transforms the function $L_\theta(\underline{x}|\theta)$ into a Gaussian of unit variance and mean value $\bar{g}$. In terms of the new variable g the LF is

$$L_g(\underline{x}|g) \propto e^{-\frac{1}{2}(g-\bar{g})^2},$$

*) We use the word "symmetric" to characterize the interval limits relative to $\hat{\theta}$, and "central" to indicate that the interval has been chosen to give equal probabilities in the two ends of the distribution.

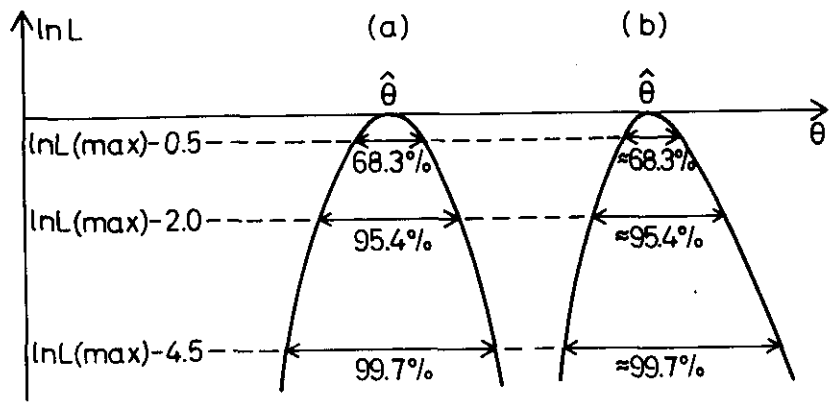

Fig. 9.5. Likelihood intervals for a one-parameter lnL function, obtained from intersection with straight lines lnL = lnL(max)-a; (a) a symmetric, parabolic lnL function, (b) an unsymmetric lnL function.

and in accordance with the invariance property of ML estimates (Sect.9.4.1) the ML solution for g is $\bar{g} = g(\hat{\theta})$.

From $L_g(\underline{x}|g)$ one can find likelihood intervals for $\bar{g}$ in exactly the same way as in the case of a normal-shaped LF. If the likelihood interval for $\bar{g}$ is $[g_a, g_b]$, the corresponding likelihood interval $[\theta_a, \theta_b]$ for the original parameter θ can be found by taking the values θ_a, θ_b as implied by the transformation $g_a = g(\theta_a)$, $g_b = g(\theta_b)$. Thus we should have to make an inverse transformation to obtain θ_a and θ_b explicitly.

The elaborate process of performing a transformation and a subsequent inverse transformation is in fact unnecessary. Since the likelihood itself gives the joint probability for obtaining the observations $x_1, x_2, \ldots, x_n$ and

this probability must be the same whether the parameter is expressed directly as θ or in an implied form $g(\theta)$, we must have, for all θ,

$$L_\theta(\underline{x}|\theta) = L_g(\underline{x}|g).$$

It follows that since, for example, a 95.4% likelihood interval for $\bar{g}$ is found from the intersections of $L_g(\underline{x}|g)$ with the line $L_g = L_g(\max)e^{-2.0}$, the corresponding 95.4% likelihood interval for θ can simply be obtained by finding the intersections of $L_\theta(\underline{x}|\theta)$ with $L_\theta = L_\theta(\max)\, e^{-2.0}$. Similarly, intersecting the non-normal LF by any straight line at a multiple e^{-a} from maximum will directly produce likelihood intervals of probabilities as given for the case with an ideal, normal LF.

Strictly speaking, the last assertion can only be approximately correct. This is so because the assumptions underlying the argumentation above will not be satisfied in the general case. The existence of the transforming function $g(\theta)$ is not granted and will, in fact, only be fulfilled to some approximation, depending on the functional form of the p.d.f. and the sample values, which determine the actual shape of the LF. For practical work this need not disturb us; as long as the graph of the lnL function has a single maximum in the region of interest and does not deviate too much from a parabola we may find its intersections with the straight lines to obtain likelihood intervals of approximate probability contents as given by eq.(9.62); see Fig. 9.5(b).

Because of its simplicity, the intersection procedure is the one most frequently used for interval estimation by physicists. An alternative procedure, equally justified from the intuitive point of view that the likelihood function gives a measure of our belief in the possible values of the unknown parameter, consists in explicitly integrating the LF. For example, if we should choose to have equal probabilities in the two tails of the LF, we could divide the total range for θ, $\theta_L \leq \theta \leq \theta_U$ into a number of cells $\Delta\theta_i$ and determine numerically two values θ_a and θ_b such that

$$\sum_{\theta_i=\theta_L}^{\theta_a} L(\underline{x}|\theta_i)\Delta\theta_i \Big/ \sum_{\theta_i=\theta_L}^{\theta_U} L(\underline{x}|\theta_i)\Delta\theta_i = \tfrac{1}{2}(1-\gamma),$$

and

$$\sum_{\theta_i=\theta_b}^{\theta_U} L(\underline{x}|\theta_i)\Delta\theta_i \Big/ \sum_{\theta_i=\theta_L}^{\theta_U} L(\underline{x}|\theta_i)\Delta\theta_i = \tfrac{1}{2}(1-\gamma) .$$

This would correspond to taking $[\theta_a,\theta_b]$ as a 100γ % central likelihood interval for θ.

Although sufficient for most practical purposes, the methods described above may sometimes not provide an adequate summary of the experiment. This can be the case, for instance, if the LF is very skew, and produces large likelihood intervals which are unsymmetrically positioned relative to the most likely value of the parameter. In such a situation one will usually state the point estimate $\hat{\theta}$ in addition to the interval estimate $[\theta_a,\theta_b]$. In more extreme cases, when the LF is particularly ill-shaped, as is sometimes experienced in low-statistics experiments, it is more informative to present a graph of the whole LF.

Exept for the ideal situation with a strictly normal LF we shall not, in general, be ensured that the intervals obtained directly from the LF (or, equivalently, from lnL) are the shortest intervals for θ corresponding to the given probability γ. If we are interested in obtaining intervals which give as accurate information on the parameter as possible, we should look for some *explicit* transformation into a new variable which is more close to the normal approximation than is the LF, and which will therefore, in the long run, produce tighter intervals, consistent with the minimum variance bound. An example on such a transformation is discussed in the subsequent sections.

9.7.2 Confidence intervals from the Bartlett functions

It was shown in Sect.9.4.7 that the standardized variable

$$S(\theta) \equiv u = \frac{\partial \ln L}{\partial\theta} \Big/ \left[E\left(-\frac{\partial^2 \ln L}{\partial\theta^2}\right)\right]^{\frac{1}{2}} \qquad (9.63)$$

becomes distributed as N(0,1) when n goes towards infinity. One can therefore, as suggested by M.S. Bartlett, use $S(\theta)$ to find the ML estimate $\hat{\theta}$ as well as any confidence interval for θ. This is suggestive, since $S(\theta)$ is usually close to being normally distributed also for finite n.

For small samples an even more applicable function is

$$S_\gamma(\theta) \equiv S(\theta) - \frac{1}{6}\gamma_1\left(S(\theta)^2 - 1\right) , \qquad (9.64)$$

where the last term is a skewness correction to the ordinary Bartlett function of eq.(9.63). The asymmetry coefficient is defined (see Sect.3.3.3) in terms of the second and third central moments of $\partial \ln L/\partial\theta$,

$$\gamma_1 \equiv \frac{\mu_3}{(\mu_2)^{3/2}} , \qquad (3.20)$$

where

$$\mu_2 = V\left(\frac{\partial \ln L}{\partial\theta}\right) = E\left[\left(\frac{\partial \ln L}{\partial\theta}\right)^2\right] = E\left(-\frac{\partial^2 \ln L}{\partial\theta^2}\right) \qquad (9.18)$$

and

$$\mu_3 = E\left[\left(\frac{\partial \ln L}{\partial\theta}\right)^3\right] = 2E\left(\frac{\partial^3 \ln L}{\partial\theta^3}\right) + 3\,\frac{\partial}{\partial\theta}\,E\left(-\frac{\partial^2 \ln L}{\partial\theta^2}\right) . \qquad (9.65)$$

In these expressions the expectations are to be evaluated for the joint p.d.f. $L(\underline{x}|\theta)$.

Exercise 9.18: Derive eq.(9.65).

9.7.3 Example: Confidence intervals for the mean lifetime

To illustrate the use of the Bartlett functions from the previous section let us consider again the ideal lifetime distribution law $f(t|\tau) = \frac{1}{\tau}e^{-t/\tau}$. For n observations,

$$\ln L = \ln\left(\prod_{i=1}^{n}\frac{1}{\tau}e^{-t_i/\tau}\right) = -n\ln\tau - \frac{n\bar{t}}{\tau} ,$$

and the ML estimate becomes $\hat{\tau} = \bar{t} = \frac{1}{n}\Sigma t_i$ as we saw in Sect.9.1.1. The asymmetry coefficient is obtained as $\gamma_1 = 2/\sqrt{n}$ and the Bartlett functions take the forms

$$S(\tau) = \frac{\hat{\tau}-\tau}{\tau/\sqrt{n}} , \qquad (9.66)$$

and

$$S_\gamma(\tau) = \frac{\hat{\tau}-\tau}{\tau/\sqrt{n}} - \frac{1}{3\sqrt{n}}\left[\left(\frac{\hat{\tau}-\tau}{\tau/\sqrt{n}}\right)^2 - 1\right]. \qquad (9.67)$$

The first of these functions appeared already in the example of Sect. 9.4.8, where it was shown explicitly to be a standard normal variable for large n. A probability statement may therefore be written as

$$P\left(-2 \le \frac{\hat{\tau}-\tau}{\tau/\sqrt{n}} \le 2\right) = 0.954, \qquad (9.68)$$

which can be inverted to read

$$P\left(\frac{\hat{\tau}}{1+\frac{2}{\sqrt{n}}} \le \tau \le \frac{\hat{\tau}}{1-\frac{2}{\sqrt{n}}}\right) = 0.954$$

This gives the 95.4% confidence interval for τ,

$$\left[\frac{\hat{\tau}}{1+\frac{2}{\sqrt{n}}}, \frac{\hat{\tau}}{1-\frac{2}{\sqrt{n}}}\right]. \qquad (9.69)$$

To order $\frac{1}{n}$ this interval is symmetric about $\hat{\tau}$.

Taking instead the function $S_\gamma(\tau)$ the analogue to the probability statement (9.68) is

$$P\left\{-2 \le \frac{\hat{\tau}-\tau}{\tau/\sqrt{n}} - \frac{1}{3\sqrt{n}}\left[\left(\frac{\hat{\tau}-\tau}{\tau/\sqrt{n}}\right)^2 - 1\right] \le 2\right\} = 0.954. \qquad (9.70)$$

If we now want the limits of the corresponding confidence interval for τ a second-order equation must be solved for each of these limits. Of the two solutions of each equation we take those which, when $n \to \infty$, coincide with the limits obtained above with the function $S(\tau)$, since $S_\gamma(\tau) \to S(\tau)$ when n becomes very large. The result is that $S_\gamma(\tau)$, for n not too small, provides the 95.4% confidence interval for τ

$$\left[\frac{\hat{\tau}}{\frac{5}{2}-\frac{3}{2}\sqrt{1-\frac{8}{3\sqrt{n}}+\frac{4}{9n}}}, \frac{\hat{\tau}}{\frac{5}{2}-\frac{3}{2}\sqrt{1+\frac{8}{3\sqrt{n}}+\frac{4}{9n}}}\right]. \qquad (9.71)$$

This interval is more symmetric about $\hat{\tau}$ and also shorter than the interval of eq.(9.69).

The intervals of eqs.(9.69), (9.71) are the results of inversion of

probability statements about specified functions of τ. They express the probability that the true value of the parameter will be included between certain limits given by the random variable $\hat{\tau} = \bar{t}$ and the sample size n, and therefore correspond to confidence intervals in the sense of Chapter 7. These intervals will be more symmetric and, in the long run, more accurate than intervals derived directly from the likelihood function.

Exercise 9.19: Using the one-standard deviation limits for the two Bartlett functions $S(\tau)$ and $S_\gamma(\tau)$ in the text, show that the 68.3% confidence intervals obtained for τ are identical.

Exercise 9.20: In a laboratory exercise to measure the mean lifetime τ of Λ hyperons produced in a bubble chamber a student has for 14 event candidates measured the following proper flight times in units of 10^{-10} seconds:

2.4, 1.7, 1.9, 4.5, 0.5, 10.2, 2.7, 0.8, 1.4, 1.6, 1.0, 1.2, 2.8, 0.7.

Assuming the chamber to have infinite dimensions and perfect detection conditions, what is the ML estimate for τ? Plot the lnL function from these assumptions and determine the 68.3% and 95.4% likelihood intervals. Compare these with the corresponding intervals obtained from the Bartlett functions $S(\tau)$ and $S_\gamma(\tau)$.

Exercise 9.21: Consider the p.d.f. $f(t;T|\lambda) = \lambda e^{-\lambda t}/(1-e^{-\lambda T})$ for $0 \le t \le T$, (compare Exercise 9.2). Show that the Bartlett function defined by eq.(9.66) becomes

$$S(\lambda) = \frac{\hat{\lambda}-\lambda}{\frac{\hat{\lambda}}{\sqrt{n}}\left[1 - \frac{1}{n}\sum_i (\lambda T_i)^2 e^{-\lambda T_i}/(1-e^{-\lambda T_i})^2\right]^{\frac{1}{2}}}.$$

Note that when $T_i \to \infty$, $S(\lambda) \to \frac{\hat{\lambda}-\lambda}{\hat{\lambda}/\sqrt{n}}$.

9.7.4 Likelihood regions; the two-parameter case

In a situation involving two parameters we shall regard the likelihood function $L(\underline{x}|\theta_1,\theta_2)$ obtained for a set of observations as containing all information available on the unknown parameters and use it to make inferences about them based on probability statements of the type

$$P(\theta_1^a \le \theta_1 \le \theta_1^b,\ \theta_2^a \le \theta_2 \le \theta_2^b) = \gamma. \qquad (9.72)$$

We will start with the assumption of asymptotic conditions. In the limit of infinitely large samples the LF will be the binormal distribution

$$L(\theta_1,\theta_2) = L(\max)\,\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{\theta_1-\hat\theta_1}{\sigma_1}\right)^2+\left(\frac{\theta_2-\hat\theta_2}{\sigma_2}\right)^2-2\rho\left(\frac{\theta_1-\hat\theta_1}{\sigma_1}\right)\left(\frac{\theta_2-\hat\theta_2}{\sigma_2}\right)\right]\right\} \quad (9.50)$$

equivalent to

$$\ln L(\theta_1,\theta_2) = \ln L(\max) - \tfrac{1}{2}Q \quad (9.73)$$

where

$$Q = \frac{1}{1-\rho^2}\left[\left(\frac{\theta_1-\hat\theta_1}{\sigma_1}\right)^2+\left(\frac{\theta_2-\hat\theta_2}{\sigma_2}\right)^2-2\rho\left(\frac{\theta_1-\hat\theta_1}{\sigma_1}\right)\left(\frac{\theta_2-\hat\theta_2}{\sigma_2}\right)\right] \quad (9.74)$$

Curves for constant likelihood will then be ellipses with centre at the ML estimates $\hat\theta_1,\hat\theta_2$, as shown in Fig. 9.6. Specifically, we recognize $Q = 1$ as giving the covariance ellipse of eq.(9.51).

The quadratic form Q of the two normal variables θ_1 and θ_2 has the remarkable feature that its distributional properties are independent of all five constants $\hat\theta_1,\hat\theta_2,\sigma_1,\sigma_2,\rho$; in fact, Q is distributed as a chi-square variable with 2 degrees of freedom (compare Exercise 4.48). For such a variable we can express a probability by the cumulative integral

$$P(Q \le Q_\gamma) = \int_0^{Q_\gamma} f(Q;\nu=2)\,dQ = \gamma\ . \quad (9.75)$$

where $f(Q;\nu=2)=\frac{1}{2}\exp(-\frac{1}{2}Q)$, according to Sects.5.1.1 and 5.1.2. The integration can therefore be performed explicitly, giving

$$P(Q \le Q_\gamma) = 1 - e^{-\frac{1}{2}Q_\gamma} = \gamma\ . \quad (9.76)$$

Clearly, the condition $Q \le Q_\gamma$ is equivalent to having both variables θ_1,θ_2 at the same time within the region enclosed by the ellipse $Q = Q_\gamma$. Writing

$$P(\theta_1,\theta_2 \text{ within ellipse } Q = Q_\gamma \text{ about } \hat\theta_1,\hat\theta_2) = \gamma$$

we have arrived at a probability statement of the type (9.72). The ellipse $Q = Q_\gamma$ centred at the ML estimates $\hat\theta_1,\hat\theta_2$ has a probability γ of covering both θ_1 and θ_2 simultaneously and therefore represents a 100γ % joint likelihood region for the two parameters.

If the two-parameter LF is of a strictly binormal shape we shall consequently have a simple interpretation of the elliptic likelihood contours

obtained by intersecting lnL by parallel planes lnL = lnL(max)-a: they are the boundaries of joint likelihood regions for the two unknown parameters, and have probability contents as implied by eq.(9.76), where $\frac{1}{2}Q_\gamma = a$. In particular, the covariance ellipse has a probability content $\gamma = 1-e^{-\frac{1}{2}} = 0.393$ and thus represents a 39.3% joint likelihood region for θ_1 and θ_2. The following list of numbers should be compared with its one-parameter analogue, eq.(9.62),

a = 0.5 gives a 39.3% joint likelihood region for θ_1 and θ_2,
a = 2.0 " " 86.5% " " " " " " " " , (9.77)
a = 4.5 " " 98.9% " " " " " " " " .

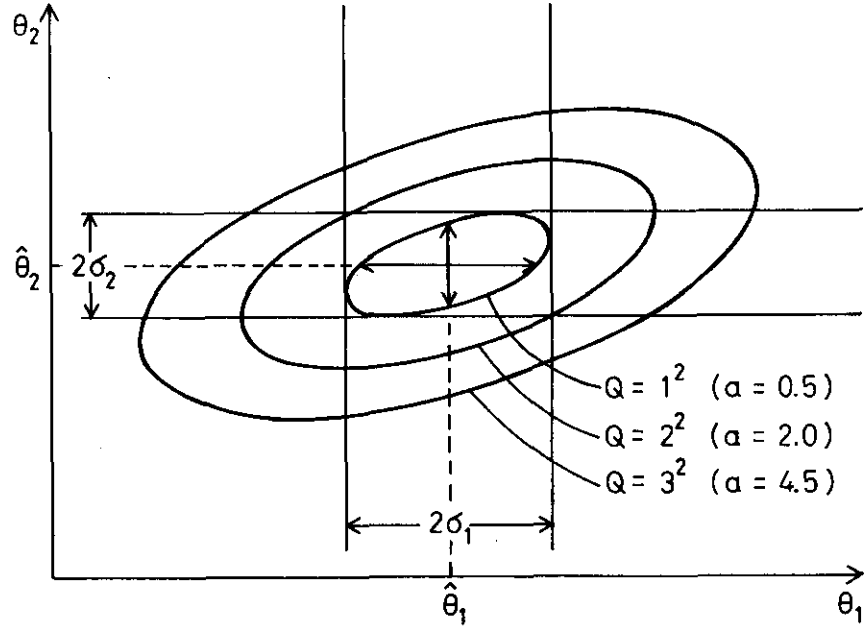

Fig. 9.6. Likelihood regions for a binormal likelihood function. The ellipses are obtained by intersecting the lnL function by planes at distances a below the maximum point $(\hat\theta_1,\hat\theta_2)$ and define joint likelihood regions for θ_1 and θ_2 of probabilities as given by eq.(9.77). The vertical band of width $2\sigma_1$ around $\hat\theta_1$ touching the covariance ellipse $Q = 1$ is a 68.3% likelihood interval for θ_1, irrespective of θ_2; similarly, the horizontal band is a 68.3% likelihood interval for θ_2, ignoring θ_1. Within the covariance ellipse the 68.3% conditional likelihood intervals of length $2\sigma_i\sqrt{1-\rho^2}$ for either parameter are indicated.

From the preceding arguments it is also evident how a joint likelihood region of specified probability γ can be found for two parameters, given a binormal LF. One chooses the elliptic region $Q = 2a$ bounded by the contour $\ln L = \ln L(\max) - a$, where the number a is deduced from eq.(9.76) for the specified γ,

$$a = -\ln(1-\gamma). \tag{9.78}$$

Thus, for example, if we were interested in a 68.3% joint likelihood region for the two parameters, this would correspond to taking $a = 1.15$ and the ellipse $Q = 2.30$.

Although appealing, choosing elliptic areas as joint likelihood regions for two parameters does not represent the only possibility. For instance, the rectangle circumscribing the covariance ellipse gives another convenient joint likelihood region for θ_1 and θ_2. It corresponds to a probability statement of the type of eq.(9.72),

$$P(\hat\theta_1-\sigma_1 \le \theta_1 \le \hat\theta_1+\sigma_1,\ \hat\theta_2-\sigma_2 \le \theta_2 \le \hat\theta_2+\sigma_2) = \gamma(\rho)\ . \tag{9.79}$$

Evidently, the probability content of the rectangle must be larger than that of the covariance ellipse; in fact, it is found to depend on the correlation ρ. With the LF of eq.(9.50) the probability (strictly speaking: our relative belief) of having θ_1 and θ_2 within distances $\pm\sigma_1$ and $\pm\sigma_2$ from the estimated values is given by the integral

$$\left(2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}\cdot L(\max)\right)^{-1}\int_{\hat\theta_1-\sigma_1}^{\hat\theta_1+\sigma_1}\int_{\hat\theta_2-\sigma_2}^{\hat\theta_2+\sigma_2} d\theta_1 d\theta_2 L(\theta_1,\theta_2)\ ,$$

which can be reduced to a function of ρ and determined numerically. The same procedure can obviously be applied to determine the probability $\gamma(\rho,m)$ of any other rectangle specified by $\theta_1 = \hat\theta_1 \pm m\sigma_1$, $\theta_2 = \hat\theta_2 \pm m\sigma_2$. It is left as an exercise to the reader to verify that this probability is given by the formula

$$\gamma(\rho,m) = \frac{1}{\sqrt{2\pi}}\int_{-m}^{+m} dy\ e^{-\frac{1}{2}y^2}\left[\ G\left(\frac{m-\rho y}{\sqrt{1-\rho^2}}\right) - G\left(\frac{-m-\rho y}{\sqrt{1-\rho^2}}\right)\right] \tag{9.80}$$

where G is the cumulative standard normal distribution. As expected from

symmetry reasons, $\gamma(-\rho,m) = \gamma(\rho,m)$. The following table summarizes the numerical values of the probability content for a set of ellipses as well as their circumscribing rectangles for different values of $|\rho|$.

m	Prob. of ellipse, $\gamma=1-e^{-\frac{1}{2}m^2}$	Probability of circumscribing rectangle, $\gamma(\rho,m)$ $\lvert\rho\rvert$ 0.0	0.2	0.4	0.5	0.6	0.7	0.8	0.9	0.95	0.99	1.00
0.5	.1175	.147	.149	.158	.165	.176	.191	.216	.258	.294	.343	.383
1.0	.3935	.466	.471	.486	.498	.514	.534	.561	.596	.622	.655	.683
1.5	.6753	.751	.754	.763	.770	.778	.789	.802	.821	.834	.852	.866
2.0	.8647	.911	.912	.915	.917	.920	.924	.929	.936	.941	.948	.954
2.5	.9561	.975	.975	.976	.977	.977	.978	.979	.982	.983	.986	.988
3.0	.9889	.995	.995	.995	.995	.995	.995	.995	.996	.996	.997	.997

The table values show that $\gamma(\rho,m)$ is a rather sensitive increasing function of $|\rho|$, particularly when $|\rho| \to 1$ and m is small, but becomes almost independent of $|\rho|$ when $m \gtrsim 3$. The numbers imply that, given the size of a circumscribing rectangle, it represents a probability which depends on the magnitude of the correlation between the parameters. On the other hand, *all* ellipses which can be inscribed in the rectangle, corresponding to any value of the correlation, represents *one* common value for the joint probability. These findings may appear somewhat surprising, since one's first impression, say from glancing at Fig. 9.4, is likely to be otherwise and, in particular, that the probability associated with ellipses inscribed in a given rectangle should depend on their size.

It is rather trivial to construct likelihood intervals for each of the parameters considered separately. Writing a probability statement like

$$P(\hat\theta_1-m\sigma_1 \le \theta_1 \le \hat\theta_1+m\sigma_1) = \gamma \tag{9.81}$$

for the first prameter would imply ignoring the second, which means that it can have any value. For this situation, integrating $L(\theta_1,\theta_2)$ with the appropriate normalization over all θ_2 gives the marginal distribution in θ_1, which becomes $N(\hat\theta_1,\sigma_1^2)$ (see the derivation of eq.(4.81) in Sect.4.9.2). From this distribution one finds the likelihood intervals for θ_1 in the usual manner. Specifically, the one-standard deviation (68.3%) likelihood interval $[\hat\theta_1-\sigma_1,\ \hat\theta_1+\sigma_1]$ for θ_1 becomes the infinitely long vertical band touching the covariance ellipse, indicated in Fig. 9.6; similarly, the long horizontal band

is the one-standard deviation likelihood interval $[\hat{\theta}_2-\sigma_2,\ \hat{\theta}_2+\sigma_2]$ for θ_2. In the event of *independent* parameters ($\rho = 0$, a symmetrically positioned ellipse), when the joint probability factorizes into the two marginal probabilities, we therefore find

$$P(\hat{\theta}_1-\sigma_1 \le \theta_1 \le \hat{\theta}_1+\sigma_1, \quad \hat{\theta}_2-\sigma_2 \le \theta_2 \le \hat{\theta}_2+\sigma_2) =$$
$$P(\hat{\theta}_1-\sigma_1 \le \theta_1 \le \hat{\theta}_1+\sigma_1)\cdot P(\hat{\theta}_2-\sigma_2 \le \theta_2 \le \hat{\theta}_2+\sigma_2) = 0.683^2 = 0.466.$$

In a similar manner one can establish *conditional* likelihood intervals for one parameter when the other parameter is kept at a fixed value. For instance, specifying the first parameter to its estimated value $\theta_1 = \hat{\theta}_1$, the second has the conditional distribution $N(\hat{\theta}_2,\sigma_2^2(1-\rho^2))$, (compare eq.(4.82)), showing that $[\hat{\theta}_2-\sigma_2\sqrt{1-\rho^2},\ \hat{\theta}_2+\sigma_2\sqrt{1-\rho^2}]$ in this case is a 68.3% conditional likelihood interval for this parameter.

Our discussion has so far been based on the assumption of an ideal, binormal likelihood function. In real life, with finite samples, the two-parameter LF will not comply with this requirement. As in the one-parameter case we can, however, assume that the departure from asymptotic conditions is not too severe. If the LF has a single maximum and is sufficiently regular in the region of interest we shall again take for granted the existence of some transforming function which brings the LF over to the ideal binormal shape, thereby enabling us to reason about transformed likelihood regions like we did in Sect.9.7.1 for the one-parameter intervals. Consequently, by intersecting the lnL surface by planes lnL = lnL(max)-a we shall take the region enclosed by the contour as an approximate 100γ % joint likelihood region for the two parameters θ_1 and θ_2 where, as before for the ideal binormal case, $\gamma = 1-e^{-a}$.

For an irregular LF, where the existence of a transforming function is an obviously inadequate assumption, it will not be justified to take the numbers $\gamma = 1-e^{-a}$ as approximate measures of the probabilities to be associated with the joint likelihood regions. This applies, in particular, if the LF has more than one maximum and the intersection lnL = lnL(max)-a produces two or more disconnected regions in parameter space. In this situation the experiment is only poorly summarized by specifying the particular regions, and it is more informative to display the LF graphically by a whole set of likelihood contours.

9.7.5 Example: Likelihood region for μ and σ^2 in $N(\mu,\sigma^2)$

We have earlier established that the joint ML estimates of the mean and variance in the normal p.d.f. $N(\mu,\sigma^2)$ are given by $\hat{\mu} = \bar{x} = \frac{1}{n}\Sigma x_i$ and $\hat{\sigma}^2 = \frac{n-1}{n}s^2 = \frac{1}{n}\Sigma(x_i-\bar{x})^2$ (Sect.9.2.3) and that, for large samples, their covariance matrix is diagonal with elements $V_{11} = \sigma^2/n$ and $V_{22} = 2\sigma^4/n$ (Sect. 9.5.5). If in these elements we replace the parameter σ^2 by its estimated value $\hat{\sigma}^2$ we find that the variable Q of eq.(9.74) is given by

$$Q = (\mu-\hat{\mu})^2\frac{n}{\hat{\sigma}^2} + (\sigma^2-\hat{\sigma}^2)^2\frac{n}{2(\hat{\sigma}^2)^2},$$

which describes ellipses centred at $(\hat{\mu},\hat{\sigma}^2)$ in the (μ,σ^2) plane, with half axes proportional to $\hat{\sigma}/\sqrt{n}$ and $\hat{\sigma}^2/\sqrt{n}$. In particular, the ellipse which gives the 95% joint likelihood region for μ and σ^2 corresponds to taking $Q = 2a$ where, from

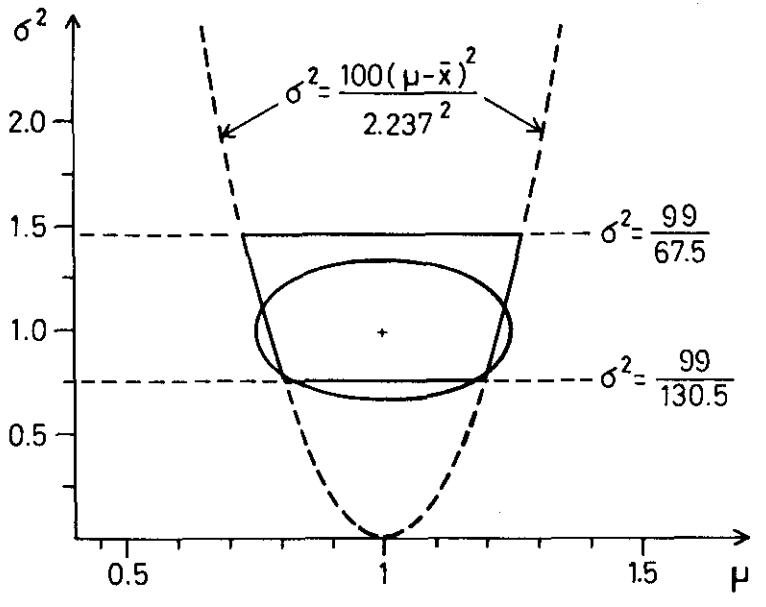

Fig. 9.7. Likelihood region (ellipse) and confidence region (intersected parabola) for μ and σ^2 in $N(\mu,\sigma^2)$.

eq.(9.78), a = − ln(1−0.95) = 2.996; hence its half axes are

$$\sqrt{5.99\cdot\hat{\sigma}^2/n}\ , \qquad \sqrt{5.99\cdot 2(\hat{\sigma}^2)^2/n}\ .$$

It is interesting to compare the elliptic joint likelihood region obtained in this manner with the findings of Sect.7.4, where we used the independence property of the statistics $\bar{x}$ and s^2 to construct a joint confidence region for μ and σ^2, illustrated by the shaded area of Fig. 7.2. A sample of size n = 100 with $\bar{x}$ = 1, s^2 = 1 gives the 95% joint likelihood ellipse of Fig. 9.7. This ellipse is slightly smaller in area than the particular 95% joint confidence region shown in the same figure, bounded by the parabola and the two horizontal lines. These have been determined by demanding equal tail probabilities = $\frac{1}{2}(1-\sqrt{0.95})$ in the ends of the N(0,1) distribution (for the variable $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$) as well as the $\chi^2(n-1)$ distribution (for $\sum_i \left(\frac{x_i-\bar{x}}{\sigma}\right)^2$). This requirement fixes the values of the constants a, b, b' in the probability statement eq.(7.18); we find a = 2.237 in the usual way, and b = 67.5, b' = 130.5 by approximating the chi-square variable to a normal variable of the same mean and variance.

9.7.6 Likelihood regions; the multi-parameter case

To generalize the arguments of Sects.9.7.1 and 9.7.4 to a situation with several parameters we are now looking for the possibility to formulate probability statements of the type

$$P(\theta_1^a \le \theta_1 \le \theta_1^b,\ \cdots,\ \theta_k^a \le \theta_k \le \theta_k^b) = \gamma \tag{9.82}$$

on the basis of a k-dimensional likelihood function $L = L(\underline{x}|\theta_1,\theta_2,\cdots,\theta_k)$. If all parameters are considered to lie between two different limits, the span in parameter space will be a 100γ % joint likelihood region for all the k parameters. If some of them are kept constant, say at their estimated values, the region spanned by the remaining parameters will be a conditional joint likelihood region (if there are at least two parameters left) or a conditional likelihood interval (if only one parameter remains). In any case, our purpose is to find a region for chosen γ, or conversely, to find the value of γ corresponding to a specified region.

Unless we make the simplifying assumption of asymptotic conditions we shall not be able to pursue this subject very far. In doing so it should be noted that with an increasing number of variables (parameters) the approach to *"asymptopia" becomes increasingly slow. Thus larger sample sizes are in* general needed to attain a given accuracy in the approximation to normality when the number of parameters goes from one to two and beyond. In other words, comparing real life to the ideal conditions will generally imply rougher approximations in the multi-parameter case.

A Taylor expansion of lnL about the ML estimate $\underline{\hat{\theta}} = \{\hat{\theta}_1,\hat{\theta}_2,\cdots,\hat{\theta}_k\}$ reads, in general,

$$\ln L = \ln L\Big|_{\underline{\theta}=\underline{\hat{\theta}}} + \sum_{i=1}^{k} \frac{\partial \ln L}{\partial \theta_i}\Big|_{\underline{\theta}=\underline{\hat{\theta}}} (\theta_i-\hat{\theta}_i) + \tfrac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k} \frac{\partial^2 \ln L}{\partial\theta_i\partial\theta_j}\Big|_{\underline{\theta}=\underline{\hat{\theta}}} (\theta_i-\hat{\theta}_i)(\theta_j-\hat{\theta}_j) + \cdots .$$

Here the first derivatives vanish identically. If the sample size is large enough the second derivatives are the negative of the elements of the inverse covariance matrix for the ML estimates (eq.(9.32)). Hence we may write

$$\ln L = \ln L(\max) - \tfrac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k} (\theta_i-\hat{\theta}_i) V_{ij}^{-1}(\underline{\hat{\theta}})(\theta_j-\hat{\theta}_j) + \cdots . \tag{9.83}$$

When higher-order terms are neglected this gives the asymptotic LF as the multinormal distribution in $\underline{\theta}$,

$$L(\underline{\theta}) = L(\max)\ \exp\left[-\tfrac{1}{2}(\underline{\theta}-\underline{\hat{\theta}})^T V^{-1}(\underline{\theta})(\underline{\theta}-\underline{\hat{\theta}})\right] \tag{9.84}$$

where we have written $V^{-1}(\underline{\theta})$ *instead of* $V^{-1}(\underline{\hat{\theta}})$.

Intersecting the hypersurface $L(\underline{\theta})$ by hyperplanes $L = L(\max)e^{-a}$ will now give contours of constant likelihood which define hyperellipsoidal regions in parameter space. Since the quadratic form

$$Q = (\underline{\theta}-\underline{\hat{\theta}})^T V^{-1}(\underline{\theta})(\underline{\theta}-\underline{\hat{\theta}}) \tag{9.85}$$

for multinormal $\underline{\theta}$ is distributed as a chi-square variable with k degrees of freedom (Sect.4.10.2), we can express a probability by integrating the p.d.f. of the $\chi^2(k)$ variable between 0 and some value Q_γ; we write

$$P(Q \le Q_\gamma) = \int_0^{Q_\gamma} f(Q;\nu=k)\,dQ = F_{1-\gamma}(Q=Q_\gamma;\nu=k) = \gamma\ . \tag{9.86}$$

where F is the cumulative integral of the chi-square p.d.f. as defined in Sect.5.1.4. Clearly, $Q \leq Q_\gamma$ is here equivalent to having all parameters $\theta_1, \theta_2, \cdots, \theta_k$ simultaneously within the region enclosed by the hyperellipse $Q = Q_\gamma$. This hyperellipse, centred at $\hat{\underline{\theta}}$ and obtained by the intersecting hyperplane at distance $a = \frac{1}{2}Q_\gamma$ below lnL(max), will therefore be the boundary of a 100γ % joint likelihood region for all k parameters. Corresponding values of γ can be found from graphs or tabulations of the cumulative chi-square distribution, such as Fig. 5.2 or Appendix Table A8. It is important to observe that, for a fixed value of the intersection constant a, the probability associated with the likelihood region drops very quickly when the number of parameters increases. For example, the choice $a = 0.5$ ($Q_\gamma = 1$), which produced $\gamma = 0.683$ for the one-parameter and $\gamma = 0.393$ for the two-parameter case, leads to chi-square probabilities $\simeq 0.20$, $\simeq 0.10$, $\simeq 0.05$ for $k = 3, 4, 5$, as one can see from Fig. 5.2. Likewise, to obtain *a specified probability content* γ, *one will have to take* increasingly large values of a for increasing k. Specifically, to have 68.3% joint likelihood regions, the intersecting hyperplanes must be taken for $a = 1.15$, 1.77, 2.38, and 3.00 when the number of parameters is $k = 2$, 3, 4, and 5, respectively.

To obtain likelihood regions or intervals which are conditional or independent on some of the parameters we must carry out appropriate integrations of the multinormal LF. The background for our remarks here has been presented in Sect.4.10.1. If some of the parameters, say the first ℓ of them, are without interest and can be ignored, the marginal distribution obtained by integrating the LF over these is a multinormal distribution in the $k-\ell$ remaining parameters with the same mean values and covariances as they had in the full LF; this marginal distribution will then provide joint likelihood regions for the $k-\ell$ parameters, independent of the ℓ first, in the manner described above. If, on the other hand, m parameters are kept fixed at their estimated values, the conditional distribution in the remaining variables is also multinormal, of dimension k-m, but with new covariance terms as determined by the covariance matrix V*, which results by deleting the appropriate m rows and columns from the original V^{-1} and inverting the resultant matrix. This distribution would then provide joint likelihood regions for the k-m parameters, conditional on the m

first. More odd situations, in which some parameters are ignored while others are kept fixed, can also be thought of. We leave it to the interested reader to contemplate this matter further and to work out the relevant formulae. The lesson to be learned from our considerations here is that it is extremely important to state explicitly which parameters have been considered jointly estimated and which have been integrated over or kept constant at their estimated values.

9.8 GENERALIZED LIKELIHOOD FUNCTION

With the assumed functional dependence $f(x|\theta)$ between the observable x and the unknown parameter θ we have, given n events (observations) $x_1, x_2, \ldots, x_n$, written the likelihood function as $L(\underline{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$. If it is possible to write the total number ν of expected events as a function of θ, say $\nu = \nu(\theta)$, this "information" may be utilized by constructing a *generalized likelihood function* as

$$\mathcal{L}(n,\underline{x}|\theta) = \frac{\nu^n e^{-\nu}}{n!} L(\underline{x}|\theta) = \frac{\nu^n e^{-\nu}}{n!} \prod_{i=1}^{n} f(x_i|\theta). \qquad (9.87)$$

This expression describes the joint probability for observing just n events, and that these give the results $x_1, x_2, \ldots, x_n$, when the number of observed events is assumed to be Poisson variable with mean value ν.

The advantage of introducing the generalized $\mathcal{L}$ is that the number of observed events n adds an extra constraint in the determination of θ. In problems where the shape of $f(x|\theta)$ is of primary interest one will, however, in general gain fairly little by using $\mathcal{L}$ instead of L. The use of $\mathcal{L}$ is suggestive only in cases where the expected number of events $\nu(\theta)$ can be calculated with considerable accuracy. An example is given in the case study of Sect.9.12.

9.9. APPLICATION OF THE MAXIMUM-LIKELIHOOD METHOD TO CLASSIFIED DATA

When the number of observations is very large the numerical evaluation of the likelihood function may become quite laborious, especially if the p.d.f. $f(x|\theta)$ has a complex form. In such situations one may reduce the amount of computation by grouping the data into subsets or classes and write the

likelihood function as the product of a smaller number of "averaged" p.d.f.'s. A similar grouping of the data is frequently inherent in the experimental set-up itself, as for instance when a counter combination is used to register particles within a certain angular interval. It is clear that the grouping of the data into categories necessarily implies some loss of information; this loss will, however, be modest if the variation of the distribution is small over each interval.

Let the total number of events n be grouped into N classes for different intervals of the variable x. The joint probability to have n_1 events in class 1, n_2 events in class 2, and so on, is given by the multinomial distribution law,

$$L(n_1,n_2,..,n_N|\theta) = n! \prod_{i=1}^{N} \frac{1}{n_i!} p_i^{n_i} , \qquad (9.88)$$

where p_i is the probability for the class i. This probability can be found by integrating the p.d.f. over the width Δx_i of the i-th interval,

$$p_i \approx p_i(\theta) = \int_{\Delta x_i} f(x|\theta)dx .$$

Since L depends on θ only through the p_i, the ML estimate for the parameter is found by seeking the maximum value of the expression

$$\ln L(n_1,n_2,\ldots,n_N|\theta) = \sum_{i=1}^{N} n_i \ln p_i(\theta)$$

in the usual manner.

It is obvious that when the number of classes is large, corresponding to small intervals Δx_i, this method is equivalent to the ordinary procedure for unclassified data. It is also evident that if the p.d.f. varies significantly over each interval, corresponding to few classes, large Δx_i and/or a rapidly varying function, the average p_i will not be a good approximation for the probability of each class. In any case, the grouping of the data into classes implies a loss of information about the parameter and should therefore be avoided if possible.

Exercise 9.22: In the classification above, assume that, for class i, the number of events n_i is a Poisson variable with mean value ν_i. The joint probability, or likelihood, for observing just $n_1,n_2,\ldots,n_N$ events in the N classes is then

$$L(n_1,n_2,\ldots,n_N|\theta) = \prod_{i=1}^{N} \frac{1}{n_i!} \nu_i^{n_i} e^{-\nu_i},$$

where $\nu_i = n \int_{\Delta x_i} f(x|\theta)dx$, $\Sigma\nu_i = \Sigma n_i = n$. Show that, when the number of events in each class is large, $n_i \gg 1$,

$$\ln L \approx \sum_{i=1}^{N} \left(n_i \ln(n_i-1) - \ln(n_i!)\right) - \tfrac{1}{2}X^2,$$

where

$$X^2 = \sum_{i=1}^{N} \left(\frac{n_i-\nu_i}{\sqrt{n_i}}\right)^2 .$$

The maximization of lnL as a function of θ is now the same as minimizing X^2. Hence the ML estimation becomes equivalent to a Least-Squares estimation of the parameter; see Sect.10.5.1.

9.10 COMBINING EXPERIMENTS BY THE MAXIMUM-LIKELIHOOD METHOD

Consider two independent experiments with the purpose of determining the same physical parameter θ from two sets of observations $\underline{x}$ and $\underline{y}$, corresponding to the likelihood functions $L(\underline{x}|\theta)$ and $L(\underline{y}|\theta)$, respectively. The joint probability of all observations is

$$L(\underline{x},\underline{y}|\theta) = \prod_i f_1(x_i|\theta) \prod_j f_2(x_j|\theta) = L(\underline{x}|\theta)L(\underline{y}|\theta), \qquad (9.89)$$

and the ML estimate of θ from the composite experiment is found by maximizing this combined likelihood in the usual manner. Thus the combined estimate $\hat{\theta}$ can be found whenever the likelihood functions of the two experiments are known. In particular for low statistics experiments, when the LF's are far from being of Gaussian shape, this procedure is better than taking some weighted average of the ML estimates from the individual experiments.

In the limiting cases when the LF's are of approximate Gaussian shape, the combination of several experiments becomes very simple, because the combined LF will be nothing but a product of normal p.d.f.'s. Hence the formulae from Sect.9.2.2 for the weighted mean can be applied.

The prescription for combining independent experiments by eq.(9.89) follows as a consequence from Bayes' Theorem. From Sect.2.4 we recall that if

$P(\theta)$ represents the prior knowledge about θ, the posterior knowledge $P(\theta|\underline{x})$ about θ, given the observations $\underline{x}$, is given by (compare eq.(2.26))

$$P(\theta|\underline{x}) = \frac{L(\underline{x}|\theta)P(\theta)}{\int L(\underline{x}|\theta)P(\theta)d\theta} \propto L(\underline{x}|\theta)P(\theta).$$

Before the second set of observations $\underline{y}$ is made $P(\underline{x}|\theta)$ is the new prior knowledge, and the new posterior knowledge becomes

$$P(\theta|\underline{x},\underline{y}) \propto L(\underline{y}|\theta)P(\underline{x}|\theta) = L(\underline{y}|\theta)L(\underline{x}|\theta)P(\theta).$$

For the combined observations we must also have

$$P(\theta|\underline{x},\underline{y}) \propto L(\underline{x},\underline{y}|\theta)P(\theta),$$

and the two expressions are seen to lead to

$$L(\underline{x},\underline{y}|\theta) = L(\underline{x}|\theta)L(\underline{y}|\theta).$$

The present remarks will obviously hold also for the multi-parameter case.

9.11 APPLICATION OF THE MAXIMUM-LIKELIHOOD METHOD TO WEIGHTED EVENTS

It has so far in this chapter been tacitly assumed that the p.d.f. used to construct the likelihood function gives an adequate description of the experimental data. We know, however, that quite frequently will the theoretical, ideal p.d.f. cover unphysical regions for the measurable variable, or the experimental situation be such that the data will be distorted compared to the ideal expectation. We saw in Chapter 6 that it is possible, in favourable cases, to modify the ideal p.d.f. to include corrections for these effects, and construct a modified p.d.f. which will be directly comparable to the raw experimental data. Specifically, we saw that one can take into account

- random observational errors, due to measuring uncertainties, handled by "smearing" the ideal distribution by the experimental resolution function (Sect.6.2),
- systematic effects, due for example to imperfect detection ability (Sect.6.3).

For both types of effects, the new, corrected p.d.f. should then be used in the construction of the LF, and the parameters estimated as before. An example on the effect of folding-in the experimental resolution is given by Exercise 9.23 at the end of this section.

When we have not succeeded in correcting the p.d.f. for experimental biasses, we are in a less fortunate position and have to rely on the approximate method (ii) of Sect.6.3, consisting in applying weight factors to the observed events before comparison is made with the ideal model.

Suppose that we have a biassed sample of observations, due to an uneven detection efficiency. Essentially this means that at the value x_i, where we have observed one event, there should have been w_i events, where w_i is the reciprocal of the detection probability for the observed event. A fairly obvious construction of an approximate likelihood function in this situation is

$$L'(x|\underline{\theta}) = \prod_{i=1}^{n} \left(f(x_i|\underline{\theta})\right)^{w_i}, \tag{9.90}$$

or, equivalently,

$$\ln L'(x|\underline{\theta}) = \sum_{i=1}^{n} w_i \ln f(x_i|\underline{\theta}) = \sum_{i=1}^{n} w_i \ln f_i \tag{9.91}$$

which can be maximized with respect to $\underline{\theta}$ in the usual way.

It can be shown that this approximate method will lead to estimates $\hat{\underline{\theta}}$ which are asymptotically normally distributed about the true parameter values. Since, however, L' is not in a simple way related to the true (unknown) likelihood function of the problem, the usual procedures to determine the errors are no more valid. In particular, taking the inverse of the second derivatives of $-\ln L'$ would underestimate the errors, since it implies the observation of $\Sigma w_i > n$ events.

In the one-parameter case a crude procedure for the determination of the error would be to take the error as deduced from L' by the ordinary methods (for example, the graphical method) and multiply it by a factor equal to the square root of the average weight of the events. This would correspond to having the variance

$$V(\hat{\theta}) = \frac{1}{n} \sum_{i=1}^{n} w_i \Big/ \left(-\frac{\partial^2 \ln L'}{\partial \theta^2}\right), \tag{9.92}$$

which simplifies to the usual large sample formula (9.36) if all weights are equal to one.

For the multi-parameter case with k unknowns it has been shown by F. Solmitz that the variance matrix for $\underline{\hat{\theta}}$ is asymptotically given by

$$V(\underline{\hat{\theta}}) = H^{-1}H'H^{-1}, \tag{9.93}$$

where H^{-1} is the variance matrix to be deduced from L',

$$H_{\ell m} = -\frac{\partial^2 \ln L'}{\partial\theta_\ell \partial\theta_m}, \qquad \ell,m = 1,2,\ldots,k \tag{9.94}$$

and where the matrix H' is defined by

$$H'_{\ell m} = \sum_{i=1}^{n} w_i^2 \left(\frac{\partial \ln f_i}{\partial\theta_\ell}\right)\left(\frac{\partial \ln f_i}{\partial\theta_m}\right), \qquad \ell,m = 1,2,\ldots,k. \tag{9.95}$$

It is seen that eqs.(9.93) - (9.95) reduce to eq.(9.92) if there is only one parameter, and the special case with all $w_i = 1$ reproduce our earlier asymptotic result, eq.(9.37).

There is always loss of information involved in the weighting procedure, but this will be serious only if very large weights occur. As very large weights may arbitrarily increase the variance $V(\underline{\hat{\theta}})$, one will sometimes improve the precision in the parameter estimates by excluding some events from the sample. In bubble chamber experiments, for example, one can usually avoid the unwanted large weights by a suitable choice of fiducial volume.

Exercise 9.23: (The effect of experimental resolution)
An azimuthal angle ϕ is distributed according to the ideal theoretical p.d.f.

$$f(\phi|\alpha) = \frac{1}{2\pi}(1 + \alpha\cos\phi), \qquad 0 \le \phi \le 2\pi,$$

where α is unknown. Measurements on ϕ are associated with Gaussian distributed random errors, *i.e.* the resolution function is of the form $r(\phi';\phi) \sim \exp(-\frac{1}{2}(\phi'-\phi)^2/R^2)$ where R is the resolution width (compare eq.(6.4)).

(i) If $R \ll 2\pi$, show that the resolution transform can be expressed as

$$f'(\phi';R|\alpha) = \frac{1}{2\pi}\left(1 + \alpha\exp(-\tfrac{1}{2}R^2)\cos\phi'\right), \qquad 0 \le \phi' \le 2\pi.$$

(ii) Show that the ML estimates $\hat{\alpha}$ and $\hat{\alpha}'$ for the unknown parameter obtained using the uncorrected p.d.f. $f(\phi|\alpha)$ and the "smeared" p.d.f. $f'(\phi;R|\alpha)$, respectively, are related by

$$\hat{\alpha}' = \hat{\alpha}\exp(-\tfrac{1}{2}R^2).$$

Thus, random observational errors on the data will affect the ML estimate of the parameter. With reasonable good resolution, however, the change is small. For example, a resolution width of 10° will cause only a 1.5% correction.

9.12 A CASE STUDY: AN ILL-BEHAVED LIKELIHOOD FUNCTION

We will now give an example, from a low statistics experiment, on an ill-behaved likelihood function in two parameters. We will show that the introduction of the generalized LF improves the shape somewhat, but only to a very small extent. The real cure of the problem can only be obtained after a close look at the physics involved. In fact, for the present problem two experiments described by slightly different p.d.f.'s should be combined to give a well-behaved LF in the two parameters.

The physics question is the following: Is the decay $K^o \to \pi^+\pi^-\pi^o$ due to the decay of the long-lived K_L^o only, or is it partly due to the decay of the short-lived K_S^o ? In the latter case, CP is violated, and the reaction can give information on the complex CP-violating parameter η, defined by the ratio

$$\frac{\text{Amplitude}\quad (K_S^o \to \pi^+\pi^-\pi^o)}{\text{Amplitude}\quad (K_L^o \to \pi^+\pi^-\pi^o)} \equiv \mathrm{Re}\eta + \mathrm{Im}\eta\,.$$

(It should be noted that the decay $K_L^o \to \pi^+\pi^-\pi^o$ has already been observed and that its decay rate $\Gamma(K_L^o \to \pi^+\pi^-\pi^o)$ is known.)

The random variable (observable) in the problem is the proper flight-time t of the K^o between production and decay, and its p.d.f. is given by

$$f(t|\eta) = C\,N(t|\eta) = C\,\frac{N_o}{2}\,\Gamma(K_L^o \to \pi^+\pi^-\pi^o)\Big[|\eta|^2 e^{-\lambda_S t} + e^{-\lambda_L t} + 2(\mathrm{Re}\eta\cos\delta t - \mathrm{Im}\eta\sin\delta t)e^{-\frac{1}{2}(\lambda_S+\lambda_L)t}\Big] \tag{9.96}$$

for $t_{min} < t < t_{max}$. C is a normalization factor, given by

$$C = 1/\int_{t_{min}}^{t_{max}} N(t|\eta)dt$$

and

N_o = number of K^o's produced at t=0 (known),
λ_S, λ_L = the total decay rates of K_S^o and K_L^o (known),
δ = mass difference between K_L^o and K_S^o (known).

For 180 observed events a contour plot of the likelihood function

$$L(\eta) = \prod_{i=1}^{180} C_i N(t_i|\eta) \qquad (9.97)$$

in the (Reη,Imη) plane is given in Fig. 9.8(a). The LF has two maxima, one pronounced maximum at Reη = − 2.68, Imη = 0.55 and one less pronounced maximum at Reη = 0.15, Imη = −0.06. The difference in lnL between the maxima is 2.18.

We next consider the effect of imposing an absolute normalization given by the known decay rate $\Gamma(K_L^o \to \pi^+\pi^-\pi^o)$. We write the generalized likelihood function as

$$\mathcal{L}(\eta) = \frac{\nu^{180}e^{-\nu}}{180!} L(\eta), \qquad (9.98)$$

where ν is the expected number of events, which can be calculated when the detection efficiency is known.

The contours of the generalized likelihood function are shown in Fig. 9.8(b). One finds still two maxima, they are slightly better separated, but their difference in lnL is now only about 0.24. There is, therefore, no real justification for favouring one solution above the other. Referring to the physics of the problem, one solution suggests the existence of a huge CP-violating effect, while the other solution is consistent with no CP-violation.

One might wonder if the ill-behaved LF and the two solutions are just bad luck in this particular experiment, and due to some large statistical fluctuation in the data. To check this point, many artificial samples consisting of 180 events each, were generated by the Monte Carlo method for Reη = Imη = 0 and the likelihood function constructed. The contours of the LF for a typical artificial sample are shown by the full drawn curves in Fig. 9.8(c). In

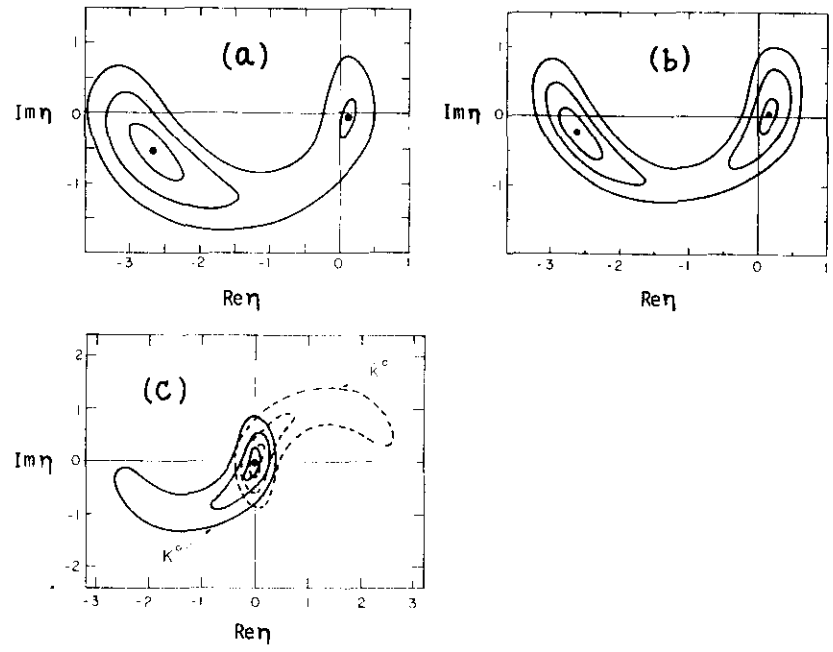

Fig. 9.8. Maximum likelihood points and likelihood contours for the experiment described in the text. (a) The LF of eq.(9.97). (b) The generalized LF of eq.(9.98). (c) Monte Carlo generated LF's assuming decays of initially produced K^o (full curves) and $\bar{K}^o$ (dotted curves).

addition to the maximum at Re$\eta \approx$ Im$\eta \approx 0$ we observe a pronounced tail in the LF favouring large negative values of Reη and Imη. Thus, in a band in the (Reη, Imη) plane the proper flight-time of the K^o is little sensitive to the values of η. The shape of the full likelihood contours of Fig. 9.8(c) indicates that the experiment cannot easily distinguish between a broad range of values of the parameter η. The indeterminacy can, however, be solved by performing a new, but very similar experiment. This rests upon the following observation: If one starts with K-zero with strangeness equal to −1 to study $\bar{K}^o \to \pi^+\pi^-\pi^o$

the last term in the p.d.f. (9.96) changes sign, whereas everything else is unchanged. The contours for a typical LF obtained from 180 artificial $\bar{K}^o$ events are shown by the dotted curves in Fig. 9.8(c). The dotted curves correspond simply to the full drawn curves reflected at the origin. We conclude therefore, that if η is close to zero a good procedure to determine the parameter would be to combine K^o and $\bar{K}^o$ data. This has in fact also been done, and it is found that η is indeed close to zero, and consistent with no CP-violation in $K^o \to \pi^+\pi^-\pi^o$ decay.

10. The Least-Squares method

In this chapter we shall discuss the estimation of parameters by the Least-Squares (LS) method, probably the estimation method most frequently used in practice.

The popularity of the LS method may partly be ascribed to the fact that it has had a long history during which it has been applied to a number of specific problems as well as to problems of more general nature. Besides this importance gained by tradition, the acceptability of the LS method, as for any systematic estimation principle, depends on the properties of the estimators to which it leads. Unlike the Maximum-Likelihood method the Least-Squares method has no general optimum properties to recommend it, even asymptotically. However, for an important class of problems, where the parameter dependence is *linear*, the LS method has the virtue that it, even for small samples, produces unbiassed estimators of minimum variance.

In the following we consider first the simple case with linear parameter dependence, then we proceed to the non-linear case, and further to situations of increasing complexity involving first linear, and later general constraint equations.

10.1 BASIS FOR THE LEAST-SQUARES METHOD

10.1.1 The Least-Squares Principle

Briefly, the basis for the LS estimation method may be stated as follows:

At the observational points $x_1, x_2, \ldots, x_N$ we are given a set of N independent, experimental values $y_1, y_2, \ldots, y_N$. The true values $\eta_1, \eta_2, \ldots, \eta_N$ of the observables are not known, but we assume that some theoretic model exists, which predicts the true value associated with each x_i through some functional dependence,

$$f_i = f_i(\theta_1,\theta_2,\ldots,\theta_L;x_i),$$

where $\theta_1,\theta_2,\ldots,\theta_L$ is a set of parameters, $L \leq N$. According to the *Least-Squares Principle* the best values of the unknown parameters are those which make

$$X^2 \equiv \sum_{i=1}^{N} w_i(y_i-f_i)^2 = \text{minimum}, \tag{10.1}$$

where w_i is the weight ascribed to the i-th observation. The set of parameters $\underline{\hat{\theta}} = \{\hat{\theta}_1,\hat{\theta}_2,\ldots,\hat{\theta}_L\}$ which produces the smallest value for X^2 is called the *Least-Squares estimate of the parameters.*

The weight w_i expresses the accuracy in the measurement y_i. In many situations one assumes that all observations are equally accurate. In such cases the LS solution for the parameters is found by determining the minimum of the unweighted sum of squared deviations, *i.e.* one minimizes the quantity

$$X^2 = \sum_{i=1}^{N} (y_i-f_i)^2. \tag{10.2}$$

This is called *unweighted LS estimation.*

If the errors in the different observations are different but known *the weight of the i-th observation is usually taken equal to its precision,* $w_i = 1/\sigma_i^2$. The quantity to be minimized is then

$$X^2 = \sum_{i=1}^{N} \left(\frac{y_i-f_i}{\sigma_i}\right)^2. \tag{10.3}$$

Quite frequently when the observations are known to be of different accuracy one can not claim a precise knowledge about the errors. Instead one must estimate the error in the individual measurements. Suppose for example, that the measurement y_i gives the number of events in a class i. One could then use the approximation $\sigma_i^2 \approx f_i$, equivalent to considering the true η_i a Poisson variable, with mean value and variance equal to f_i, *i.e.* one puts

$$X^2 = \sum_{i=1}^{N} \frac{(y_i-f_i)^2}{f_i} . \tag{10.4}$$

When f_i is a complicated function one can also for computational convenience see the approximation $\sigma_i^2 \approx y_i$ be used, with

$$X^2 = \sum_{i=1}^{N} \frac{(y_i-f_i)^2}{y_i} . \tag{10.5}$$

This can be called a *simplified LS estimation.*

If the observations are correlated, with errors and covariance terms given by the (symmetric) covariance matrix $V(\underline{y})$, the Least-Squares Principle for finding the best values of the unknown parameters is formulated as

$$X^2 = \sum_{i=1}^{N}\sum_{j=1}^{N} (y_i-f_i)V_{ij}^{-1}(y_j-f_j) = \text{minimum}. \tag{10.6}$$

It has in all formulations above been tacitly assumed that the x_i are precise values, with no errors attached to them. Each x_i may have a preassigned value, or it has been *measured with an error which is negligible* to the error of the corresponding y_i. Alternatively, the "observational point" x_i can stand for a well-defined region from x_i to $x_i+\Delta x_i$, say. Whatever the meaning of x_i the crucial assumption is that it is possible to make a *precise evaluation of* the predicted value f_i corresponding to x_i, or the region from x_i to $x_i+\Delta x_i$. Similarly, the experimental value y_i may be considered as the outcome of a single measurement, or more measurements (an average, say) at the point x_i, eventually in the region from x_i to $x_i+\Delta x_i$. We may well speak of x as an *independent*, and y as a *dependent* variable.

Finally, it may be worthwhile to emphasize that the LS estimation method makes no requirement about the distributional properties of the observables. *In this sense* the LS estimation is *distribution-free.* On the other hand, if the observables are normally distributed, the minimum value X^2_{min} will, under certain conditions, be distributed as a chi-square variable; it will then be possible to give a *quantitative measure* of the overall fit between observations and model based on the properties of the chi-square distribution. This may explain why *chi-square* (or χ^2) *minimization* in common parlance is used (erroneously) as synonymous with Least-Squares *estimation.*

10.1.2 Connection between the LS and the ML estimation methods

Let us now assume that we want to gain information on the true values η_i of the observables on the basis of the observed numbers y_i.

We can then show that the Least-Squares and the Maximum-Likelihood Principles are equivalent under certain conditions. If we assume that the indi-

vidual measurements y_i are *normally* distributed about their true, unknown values η_i with variance σ_i^2, *i.e.* the variable y_i is $N(\eta_i,\sigma_i^2)$, then the likelihood for observing the series $y_1,y_2,\ldots,y_N$ is

$$L = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\,\sigma_i} \exp\left(-\tfrac{1}{2}\left(\frac{y_i-\eta_i}{\sigma_i}\right)^2\right) \propto \exp\left(-\tfrac{1}{2}\sum_{i=1}^{N}\left(\frac{y_i-\eta_i}{\sigma_i}\right)^2\right).$$

According to the Maximum-Likelihood Principle the most probable values of the unknown η_i's are those which make L as large as possible. Evidently, L is at maximum when

$$\sum_{i=1}^{N}\left(\frac{y_i-\eta_i}{\sigma_i}\right)^2 = \text{minimum},$$

which is equivalent to the Least-Squares Principle, provided the weights are identified with the precisions in the individual, independent measurements, $w_i = 1/\sigma_i^2$.

10.2 THE LINEAR LEAST-SQUARES MODEL

The LS estimation of unknown parameters becomes especially attractive if the theoretical model implies a *linear* dependence on the parameters and the weights are *independent* of the parameters. As we shall see in the following the LS method then provides an *exact* solution for the parameters, which can be expressed in a simple closed form. Moreover, the LS estimator in the linear case possesses the theoretical optimum properties of *uniqueness, unbiassedness* and *minimum variance*.

10.2.1 Example: Fitting a straight line (1)

As a simple example on an unweighted LS estimation, suppose that we are given a set of experimental points (x_1,y_1), $(x_2,y_2),\ldots,(x_N,y_N)$ and that we want to find the "best" straight line passing through these points. We try the parameterization

$$f_i = \theta_1 + x_i\theta_2. \tag{10.7}$$

If we assume that all errors can be neglected the LS solution is obtained by determining the minimum of the unweighted sum of squared deviations, eq.(10.2),

$$X^2 = \sum_{i=1}^{N}(y_i-f_i)^2 = \sum_{i=1}^{N}(y_i-\theta_1-x_i\theta_2)^2.$$

When the derivatives of X^2 with respect to θ_1 and θ_2 are put equal to zero we get the two equations

$$\left.\begin{aligned}\frac{\partial X^2}{\partial\theta_1} &= \sum_{i=1}^{N}(-2)(y_i-\theta_1-x_i\theta_2) = 0,\\ \frac{\partial X^2}{\partial\theta_2} &= \sum_{i=1}^{N}(-2x_i)(y_i-\theta_1-x_i\theta_2) = 0.\end{aligned}\right\} \tag{10.8}$$

This set of linear equations can be written in the form

$$\left.\begin{aligned}N\,\theta_1 + \sum_{i=1}^{N}x_i\theta_2 &= \sum_{i=1}^{N}y_i,\\ \sum_{i=1}^{N}x_i\theta_1 + \sum_{i=1}^{N}x_i^2\theta_2 &= \sum_{i=1}^{N}x_iy_i.\end{aligned}\right\} \tag{10.9}$$

The solution for the parameters becomes

$$\left.\begin{aligned}\hat\theta_1 &= \frac{\Sigma x_i^2\Sigma y_i - \Sigma x_iy_i\Sigma x_i}{N\Sigma x_i^2 - (\Sigma x_i)^2},\\ \hat\theta_2 &= \frac{N\Sigma x_iy_i - \Sigma x_i\Sigma y_i}{N\Sigma x_i^2 - (\Sigma x_i)^2},\end{aligned}\right\} \tag{10.10}$$

where the summations are over all observations.

10.2.2 The normal equations

Let us now see how the independent observations $(y_1\pm\sigma_1)$, $(y_2\pm\sigma_2),\ldots,(y_N\pm\sigma_N)$ at the points $x_1,x_2,\ldots,x_N$ can be fitted by the weighted LS method to a linear model with L parameters,

$$f_i = f_i(\theta_1,\theta_2,\ldots,\theta_L;x_i) = \sum_{\ell=1}^{L}a_{i\ell}\theta_\ell, \qquad i = 1,2,\ldots,N, \tag{10.11}$$

where $L \le N$. Here $a_{i\ell}$ is the coefficient appearing with the ℓ-th parameter;

usually $a_{i\ell}$ is a function of the x_i's.

We now seek the parameter values which minimize the quantity of eq. (10.3),

$$X^2 = \sum_{i=1}^{N} \left(\frac{y_i - f_i}{\sigma_i}\right)^2 = \sum_{i=1}^{N} \frac{1}{\sigma_i^2}\left(y_i - \left(\sum_{\ell=1}^{L} a_{i\ell}\theta_\ell\right)\right)^2 . \qquad (10.12)$$

By equating all derivatives $\partial X^2/\partial\theta_k$ to zero we get the L conditions

$$\frac{\partial X^2}{\partial \theta_k} = \sum_{i=1}^{N} (-2) a_{ik} \frac{1}{\sigma_i^2}\left(y_i - \sum_{\ell=1}^{L} a_{i\ell}\theta_\ell\right) = 0 , \qquad k = 1,2,\ldots,L, \qquad (10.13)$$

which can also be written as

$$\sum_{\ell=1}^{L}\left(\sum_{i=1}^{N} \frac{a_{ik}a_{i\ell}}{\sigma_i^2}\right)\theta_\ell = \sum_{i=1}^{N} \frac{a_{ik}y_i}{\sigma_i^2} , \qquad k = 1,2,\ldots,L. \qquad (10.14)$$

We see that eqs.(10.13), (10.14) are the generalizations of eqs.(10.8), (10.9) for the simple unweighted straight-line fit of the previous section.

It may be illuminating to write eqs.(10.14) out:

$$\left.\begin{aligned}
&\sum_{i=1}^{N} \frac{a_{i1}a_{i1}}{\sigma_i^2}\,\theta_1 + \sum_{i=1}^{N} \frac{a_{i1}a_{i2}}{\sigma_i^2}\,\theta_2 + \ldots + \sum_{i=1}^{N} \frac{a_{i1}a_{iL}}{\sigma_i^2}\,\theta_L = \sum_{i=1}^{N} \frac{a_{i1}y_i}{\sigma_i^2} \\
&\sum_{i=1}^{N} \frac{a_{i2}a_{i1}}{\sigma_i^2}\,\theta_1 + \sum_{i=1}^{N} \frac{a_{i2}a_{i2}}{\sigma_i^2}\,\theta_2 + \ldots + \sum_{i=1}^{N} \frac{a_{i2}a_{iL}}{\sigma_i^2}\,\theta_L = \sum_{i=1}^{N} \frac{a_{i2}y_i}{\sigma_i^2} \\
&\quad\vdots \\
&\sum_{i=1}^{N} \frac{a_{iL}a_{i1}}{\sigma_i^2}\,\theta_1 + \sum_{i=1}^{N} \frac{a_{iL}a_{i2}}{\sigma_i^2}\,\theta_2 + \ldots + \sum_{i=1}^{N} \frac{a_{iL}a_{iL}}{\sigma_i^2}\,\theta_L = \sum_{i=1}^{N} \frac{a_{iL}y_i}{\sigma_i^2} .
\end{aligned}\right\} \qquad (10.15)$$

These are *the normal equations* for the L unknown parameters. Since there are L inhomogeneous linear equations for the L unknowns the normal equations provide an *exact* and *unique* solution.

When the number of parameters is small, $L \leq 3$, and the number of observations is not too large the normal equations are easily solved "by hand". If there are more than three parameters, or there are many observations, most people would probably use a computer for the calculations.

Exercise 10.1: In the example of Sect.10.2.1, assume that the observations $y_1, y_2, \ldots, y_N$ have errors $\sigma_1, \sigma_2, \ldots, \sigma_N$. Find the weighted LS solution for the parameters θ_1 and θ_2 of the straight line $f = \theta_1 + x\theta_2$. Numerically, take the four points (x_i, y_i) as (0,2),(2,5),(5,7),(8,10) and assume all y_i's to have a relative error of 10%.

10.2.3 Matrix notation

We shall rephrase the formulae for the linear problem of the last section in terms of matrix notation.

We order the measurements and predictions in two column vectors $\underline{y}$ and $\underline{f}$, both with N elements, and let $\underline{\theta}$ be a column vector with the L parameters, $(L \leq N)$,

$$\underline{y} = \begin{pmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_N \end{pmatrix}, \quad \underline{f} = \begin{pmatrix} f_1 \\ f_2 \\ \cdot \\ \cdot \\ \cdot \\ f_N \end{pmatrix}, \quad \underline{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \cdot \\ \cdot \\ \cdot \\ \theta_L \end{pmatrix}. \qquad (10.16)$$

The errors in $\underline{y}$ are given in a diagonal N by N matrix (diagonal, since the observations were assumed independent),

$$V = V(\underline{y}) = \begin{pmatrix} \sigma_1^2 & 0 & \ldots & 0 \\ 0 & \sigma_2^2 & \ldots & 0 \\ \cdot & \cdot & & \\ \cdot & \cdot & & \\ \cdot & \cdot & & \\ 0 & 0 & \ldots & \sigma_N^2 \end{pmatrix}, \qquad (10.17)$$

and the coefficients are organized in a matrix A with N rows and L columns,

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1L} \\ a_{21} & a_{22} & \ldots & a_{2L} \\ \cdot & \cdot & & \\ \cdot & \cdot & & \\ \cdot & \cdot & & \\ a_{N1} & a_{N2} & \ldots & a_{NL} \end{pmatrix}. \qquad (10.18)$$

The linear dependence of the theoretical predictions on the parameters, eq. (10.11), is then expressed as

$$\underline{f} = A\,\underline{\theta}\,, \tag{10.19}$$

and the quantity to be minimized is

$$X^2 = (\underline{y} - A\underline{\theta})^T V^{-1} (\underline{y} - A\underline{\theta})\,. \tag{10.20}$$

By putting the derivatives of X^2 with respect to $\underline{\theta}$ equal to zero we have the short-hand equivalent of eqs.(10.15),

$$\nabla_{\underline{\theta}} X^2 = -\,2(A^T V^{-1}\underline{y} - A^T V^{-1} A\underline{\theta}) = \underline{0}\,, \tag{10.21}$$

from which

$$(A^T V^{-1} A)\underline{\theta} = A^T V^{-1}\underline{y}\,. \tag{10.22}$$

This compact form should be compared to the normal equations (10.15). Provided the matrix $(A^T V^{-1} A)$ is non-singular, it can be inverted and the solution for $\underline{\theta}$ written in the closed form

$$\underline{\hat{\theta}} = (A^T V^{-1} A)^{-1} A^T V^{-1}\underline{y}\,. \tag{10.23}$$

One observes that to find the solution $\underline{\hat{\theta}}$ of eq.(10.20) it is sufficient that the covariance matrix $V = V(\underline{y})$ is known up to a multiplicative factor. This is also seen directly from eq.(10.23).

The equations written down and solved in matrix notation are in fact more general than indicated, since they hold also when $V(\underline{y})$ is a matrix with non-zero covariance terms. This corresponds to giving up the assumption introduced in the beginning, when it was stated that the measurements $\underline{y}$ should be independent. Relaxing this requirement corresponds of course to the more general formulation of the Least-Squares Principle by eq.(10.6).

We must next find out what can be said about the uncertainties in the linear LS estimate $\underline{\hat{\theta}}$ of the parameters, eq.(10.23). From Sect.3.8 we realize that the errors in the observed quantities $\underline{y}$ are carried over to the derived quantities $\underline{\hat{\theta}}$. Applying the general formula for error propagation, eq.(3.80), to eq.(10.23) we find that the covariance matrix for the LS estimate $\underline{\hat{\theta}}$ is

$$V(\underline{\hat{\theta}}) = \Big((A^T V^{-1} A)^{-1} A^T V^{-1}\Big) V(\underline{y}) \Big((A^T V^{-1} A)^{-1} A^T V^{-1}\Big)^T\,,$$

which simplifies to

$$V(\underline{\hat{\theta}}) = (A^T V^{-1} A)^{-1}\,. \tag{10.24}$$

Clearly, to find $V(\underline{\hat{\theta}})$ the matrix $V = V(\underline{y})$ must be known completely.

Equations (10.23) and (10.24) provide the solution to the linear LS problem for the unknown parameters $\underline{\hat{\theta}}$. It is worth noting that the (symmetric) matrix $(A^T V^{-1} A)^{-1}$ which constitutes the covariance matrix $V(\underline{\hat{\theta}})$, appears as a separate part in the expression for $\underline{\hat{\theta}}$. Thus, no extra calculation is necessary to determine the errors on $\underline{\hat{\theta}}$, as the matrix $(A^T V^{-1} A)^{-1}$ has already been found in obtaining the solution $\underline{\hat{\theta}}$.

Exercise 10.2: Verify eq.(10.24).

10.2.4 Properties of the linear LS estimator

It should be stressed that the LS estimation problem in the general case of a *linear* functional dependence on the parameters, as discussed above, has been given an *exact* solution by the closed form of eq.(10.23). The matrix algebra involves no approximations and the solution is *unique* as long as the matrices are non-singular.

We have found that the Least-Squares Principle has produced $\underline{\hat{\theta}}$ as *linear* estimators, since $\underline{\hat{\theta}}$ come out with a linear dependence on the measurements $\underline{y}$. From the definition of unbiassedness, eq.(8.3), the $\underline{\hat{\theta}}$ are also *unbiassed* estimators, as can be seen by applying the expectation operator to $\underline{\hat{\theta}}$:

$$\begin{aligned} E(\underline{\hat{\theta}}) &= E\Big((A^T V^{-1} A)^{-1} A^T V^{-1}\underline{y}\Big) = (A^T V^{-1} A)^{-1} A^T V^{-1} E(\underline{y}) \\ &= (A^T V^{-1} A)^{-1} A^T V^{-1} A\underline{\theta} = \underline{\theta} \end{aligned} \tag{10.25}$$

when we use the fact that $E(\underline{y}) = \underline{f} = A\underline{\theta}$.

A further optimum property of the linear LS estimators is contained in the *Gauss-Markov theorem:* Among all unbiassed estimators which are linear functions of the observations the LS estimators have the smallest variance.

To prove the Gauss-Markov theorem, consider a vector $\underline{t}$ of estimators, linear in the observations $\underline{y}$.

$$\underline{t} = S\,\underline{y}\,. \tag{10.26}$$

These estimators have expectation $E(\underline{t}) = SE(\underline{y}) = SA\underline{\theta}$ where $\underline{\theta}$ is the original vector of parameters. If the new estimators are to be unbiassed for a set of linear functions of the parameters, say for $C\underline{\theta}$, then also $E(\underline{t}) = C\underline{\theta}$ for all $\underline{\theta}$, hence we must have

$$C = SA\,. \tag{10.27}$$

We want to find out when the covariance matrix $V(\underline{t})$ for the new estimators,

$$V(\underline{t}) = SV(\underline{y})S^T \tag{10.28}$$

has minimal diagonal terms. For this purpose we consider the identity ($V = V(\underline{y})$)

$$\begin{aligned} SV(\underline{y})S^T \equiv & \left(C(A^TV^{-1}A)^{-1}A^TV^{-1}\right)V\left(C(A^TV^{-1}A)^{-1}A^TV^{-1}\right)^T \\ & + \left(S-C(A^TV^{-1}A)^{-1}A^TV^{-1}\right)V\left(S-C(A^TV^{-1}A)^{-1}A^TV^{-1}\right)^T. \end{aligned} \tag{10.29}$$

Here each of the two terms on the right-hand side is of quadratic form UVU^T, which implies non-negative diagonal elements. Only the second term is a function of S, and the sum of the two terms will have strictly minimum diagonal elements when the second term has vanishing elements on the diagonal. This occurs when

$$S = C(A^TV^{-1}A)^{-1}A^TV^{-1}.$$

Therefore, the unbiassed, minimum variance estimators for $C\underline{\theta}$ is

$$\underline{t} = C(A^TV^{-1}A)^{-1}A^TV^{-1}\underline{y} = C\underline{\hat{\theta}}\,, \tag{10.30}$$

where $\underline{\hat{\theta}}$ is the LS estimator for $\underline{\theta}$. Also

$$V(\underline{t}) = C(A^TV^{-1}A)^{-1}C^T. \tag{10.31}$$

10.2.5 Example: Fitting a parabola

We shall work through an example of a linear Least-Squares fit to measurements of different accuracy, which requires a weighted LS estimation.

We want to fit the parabolic parameterization

$$f(\theta_1,\theta_2,\theta_3;x) = \theta_1 + x\theta_2 + x^2\theta_3 \tag{10.32}$$

to the following observations:

i	1	2	3	4
x_i	- 0.6	- 0.2	0.2	0.6
$y_i \pm \sigma_i$	5 ± 2	3 ± 1	5 ± 1	8 ± 2

Thus the problem involves three unknown parameters and four observations.

From eqs.(10.18), (10.19) we can write down the matrix A of dimension 4 x 3 from the observations,

$$A = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \\ 1 & x_4 & x_4^2 \end{pmatrix} = \begin{pmatrix} 1 & -0.6 & (-0.6)^2 \\ 1 & -0.2 & (-0.2)^2 \\ 1 & 0.2 & 0.2^2 \\ 1 & 0.6 & 0.6^2 \end{pmatrix}.$$

The independent measurements define the column vector $\underline{y} = \{5,3,5,8\}$ and the diagonal covariance matrix

$$V = \begin{pmatrix} \sigma_1^2 & 0 & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 \\ 0 & 0 & \sigma_3^2 & 0 \\ 0 & 0 & 0 & \sigma_4^2 \end{pmatrix} = \begin{pmatrix} 2^2 & 0 & 0 & 0 \\ 0 & 1^2 & 0 & 0 \\ 0 & 0 & 1^2 & 0 \\ 0 & 0 & 0 & 2^2 \end{pmatrix}.$$

Because of its diagonal form the inverse of V can easily be written down. We find therefore for the product

$$A^T V^{-1} A = \begin{pmatrix} 1 & 1 & 1 & 1 \\ -0.6 & -0.2 & 0.2 & 0.6 \\ (-0.6)^2 & (-0.2)^2 & 0.2^2 & 0.6^2 \end{pmatrix} \begin{pmatrix} 0.25 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0.25 \end{pmatrix} \begin{pmatrix} 1 & -0.6 & (-0.6)^2 \\ 1 & -0.2 & (-0.2)^2 \\ 1 & 0.2 & 0.2^2 \\ 1 & 0.6 & 0.6^2 \end{pmatrix}$$

$$= \begin{pmatrix} 2.5 & 0 & 0.26 \\ 0 & 0.26 & 0 \\ 0.26 & 0 & 0.068 \end{pmatrix}.$$

The resulting matrix is of dimension 3 x 3, it is non-singular, and can be inverted. The inverse becomes

$$(A^T V^{-1} A)^{-1} = \begin{pmatrix} 0.664 & 0 & -2.54 \\ 0 & 3.85 & 0 \\ -2.54 & 0 & 24.42 \end{pmatrix}.$$

From the formula (10.23) we find the LS solution $\hat{\underline{\theta}}$ by carrying out the multiplication of the matrices,

$$\hat{\underline{\theta}} = (A^T V^{-1} A)^{-1} A^T V^{-1} \underline{y}$$

$$= \begin{pmatrix} 0.664 & 0 & -2.54 \\ 0 & 3.85 & 0 \\ -2.54 & 0 & 24.42 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 \\ -0.6 & -0.2 & 0.2 & 0.6 \\ (-0.6)^2 & (-0.2)^2 & 0.2^2 & 0.6^2 \end{pmatrix} \begin{pmatrix} 0.25 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0.25 \end{pmatrix} \begin{pmatrix} 5 \\ 3 \\ 5 \\ 8 \end{pmatrix}$$

$$= \begin{pmatrix} 3.68 \\ 3.27 \\ 7.81 \end{pmatrix}.$$

Thus the LS solution for the parabola through the given data points is

$$\hat{f} = f(\hat{\underline{\theta}}; x) = 3.68 + 3.27x + 7.81x^2.$$

The accuracy of the estimated parameters can be found from the covariance matrix $V(\hat{\underline{\theta}})$, which according to eq.(10.24) is nothing but the matrix $(A^T V^{-1} A)^{-1}$ above. Hence the estimates of the errors are given by the square roots of the diagonal elements of this matrix,

$$\Delta\hat{\theta}_1 = 0.81, \qquad \Delta\hat{\theta}_2 = 1.96, \qquad \Delta\hat{\theta}_3 = 4.94.$$

The paramaters θ_1 and θ_3 are correlated, with the estimated correlation coefficient

$$\hat{\rho}_{13} = \frac{(A^T V^{-1} A)^{-1}_{13}}{\Delta\theta_1 \Delta\theta_3} = \frac{-2.54}{0.81 \cdot 4.94} = -0.63.$$

Exercise 10.3: Derive the solution $\hat{\underline{\theta}}$ for the problem in the text by solving the normal equations.

Exercise 10.4: Derive the LS solution and its errors for the same problem with all measurement errors equal, $\sigma_i = 2$.

10.2.6 Example: Combining two experiments

To test the $\Delta S/\Delta Q = 1$ rule in weak interactions one can measure the complex quantity

$$x = \frac{\text{Amplitude } (K^0 \to \pi^+ \ell^- \bar{\nu})}{\text{Amplitude } (K^0 \to \pi^- \ell^+ \nu)}$$

which gives a measure of the "violation" of the rule. Let the true values of the observables Rex and Imx be denoted by η_1 and η_2. Experiment A has measured both variables, and obtained the results

$$y_1^A = 0.12, \qquad y_2^A = -0.25$$

with errors given by

$$V(\underline{y}^A) = \begin{pmatrix} 0.01 & -0.01 \\ -0.01 & 0.04 \end{pmatrix}.$$

Experiment B measured η_1 only, with the result

$$y_1^B \pm \sigma_1^B = 0.01 \pm 0.08.$$

We want to find the best combined outcome of the two experiments.

We have two unknowns in this problem, η_1 and η_2, and three measurements, y_1^A, y_2^A and y_1^B. In accordance with our formulation of the linear LS problem we put

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

The measurements are collected in a column vector $\underline{y}$ with 3 elements, and the covariance matrix extended to dimension 3 by 3,

$$\underline{y} = \begin{pmatrix} 0.12 \\ -0.25 \\ 0.01 \end{pmatrix}, \qquad V(\underline{y}) = \left(\begin{array}{cc|c} 0.01 & -0.01 & 0 \\ -0.01 & 0.04 & 0 \\ \hline 0 & 0 & (0.08)^2 \end{array}\right).$$

Carrying out the matrix operations we find

$$\hat{\underline{\eta}} = (A^T V^{-1} A)^{-1} A^T \underline{y} = \begin{pmatrix} 0.052 \\ -0.183 \end{pmatrix}$$

and

$$V(\hat{\underline{\eta}}) = (A^T V^{-1} A)^{-1} = \begin{pmatrix} 0.0039 & -0.0039 \\ -0.0039 & 0.0346 \end{pmatrix}.$$

It is worth noting that due to the correlations the estimate of Imx (= η_2) has changed, although this quantity was only measured in one experiment.

10.2.7 General polynomial fitting

We have in two previous examples considered Least-Squares fits to first-order and second-order polynomials in the variable x (Sects.10.2.1 and 10.2.5, respectively). It is frequently necessary to consider fits to higher-order polynomials of the general form

$$f_i = \sum_{\ell=1}^{L} x_i^{\ell-1} \theta_\ell . \tag{10.33}$$

Since the parameter dependence is still linear, such a problem has an exact solution of the form of eq.(10.23). However, as the power of the polynomial increases the inversion of the matrices involved becomes increasingly intricate. Serious numerical inaccuracies may occur when the degree of the polynomial gets as large as, say, 6 or 7.

10.2.8 Orthogonal polynomials

It is possible to avoid serious rounding-off errors in the matrix operations by rewriting a linear model of the form of eq.(10.33) in terms of *orthogonal polynomials* of x. The effect is to produce matrices of *diagonal* form which can easily be inverted.

Let us consider a case where the measurements are *uncorrelated* and have the *same error* σ. The covariance matrix and its inverse are then multiples of the unit matrix I_N of dimension N x N, $V(\underline{y}) = \sigma^2 I_N$, $V^{-1}(\underline{y}) = \frac{1}{\sigma^2} I_N$. With the linear model

$$f_i = \sum_{\ell=1}^{L} a_{i\ell} \theta_\ell \tag{10.11}$$

the solution of the equivalent unweighted LS problem is given by

$$\hat{\underline{\theta}} = (A^T A)^{-1} A^T \underline{y} ,$$

where A is the matrix of coefficients $a_{i\ell}$.

Suppose now that we have a set of L polynomials $\xi_\ell(x)$, which are orthogonal over the observations,

$$\sum_{i=1}^{N} \xi_k(x_i) \xi_\ell(x_i) = \delta_{k\ell} , \qquad k,\ell = 1,2,\ldots,L. \tag{10.34}$$

We will take as our new model

$$f_i = \sum_{\ell=1}^{L} \xi_\ell(x_i) \omega_\ell , \tag{10.35}$$

where ω_ℓ are the L new parameters. The matrix A and its transposé A^T now have elements

$$(A)_{i\ell} = (A^T)_{\ell i} = a_{i\ell} = \xi_\ell(x_i).$$

The product matrix $A^T A$ is such that

$$(A^T A)_{k\ell} = \sum_{i=1}^{N} (A^T)_{ki} (A)_{i\ell} = \sum_{i=1}^{N} \xi_k(x_i) \xi_\ell(x_i) = \delta_{k\ell}.$$

Hence $A^T A = I_L$, the unit matrix of dimension L x L. The LS solution for the parameter vector $\underline{\omega}$ simplifies to

$$\hat{\underline{\omega}} = A^T \underline{y} \,, \tag{10.36}$$

or, on component form,

$$\hat{\omega}_\ell = (A^T \underline{y})_\ell = \sum_{i=1}^{N} \xi_\ell(x_i) y_i \,, \qquad \ell = 1,2,\ldots,L. \tag{10.37}$$

The covariance matrix for $\hat{\underline{\omega}}$ becomes

$$V(\hat{\underline{\omega}}) = A^T V(\underline{y}) A = \sigma^2 I_L \,, \tag{10.38}$$

i.e. a diagonal matrix.

We see therefore that the LS estimates of the parameters are easily derived in this case, with the errors on the uncorrelated estimates given directly by the (common) error on the measurements.

10.2.9 Example: Fitting a straight line (2)

Let us illustrate the use of orthogonal polynomials by reconsidering the problem of fitting a straight line through the points $(x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N)$, which was first treated in Sect.10.2.1.

The model is now to be written in terms of two parameters ω_1,ω_2 and two orthogonal functions $\xi_1(x_i),\xi_2(x_i)$. From the definition of the mean value $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ we realize that orthogonality is ensured if we take

$$\xi_1(x_i) = 1 \,, \qquad\qquad \xi_2(x_i) = x_i - \bar{x} \,.$$

This choice is, however, not quite in accordance with the formulation of the previous section, since the polynomials are not normalized to one, but give

$$\sum_{i=1}^{N} [\xi_1(x_i)]^2 = N, \qquad\qquad \sum_{i=1}^{N} [\xi_2(x_i)]^2 = \sum_{i=1}^{N} (x_i-\bar{x})^2 .$$

The parameterization for the straight line is

$$f(x) = \omega_1 + (x-\bar{x})\omega_2 \,,$$

and the matrix A, of dimension N by 2,

$$A = \begin{pmatrix} 1 & x_1-\bar{x} \\ 1 & x_2-\bar{x} \\ \vdots & \vdots \\ 1 & x_N-\bar{x} \end{pmatrix} .$$

Because of the relaxation of the normalization condition the product matrix A^TA of dimension 2 x 2 is not a multiple of the unit matrix, but rather

$$A^TA = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ x_1-\bar{x} & x_2-\bar{x} & \ldots & x_N-\bar{x} \end{pmatrix} \begin{pmatrix} 1 & x_1-\bar{x} \\ 1 & x_2-\bar{x} \\ \vdots & \vdots \\ 1 & x_N-\bar{x} \end{pmatrix} = \begin{pmatrix} N & 0 \\ 0 & \Sigma(x_i-\bar{x})^2 \end{pmatrix} .$$

This matrix can trivially be inverted, giving

$$(A^TA)^{-1} = \begin{pmatrix} N^{-1} & 0 \\ 0 & (\Sigma(x_i-\bar{x})^2)^{-1} \end{pmatrix} .$$

The LS solution for the coefficients $\underline{\omega}$ is therefore

$$\hat{\underline{\omega}} = (A^TA)^{-1}A^T\underline{y} = \ldots = \begin{pmatrix} \dfrac{\Sigma y_i}{N} \\ \dfrac{\Sigma x_i y_i - \bar{x}\Sigma y_i}{\Sigma(x_i-\bar{x})^2} \end{pmatrix} ,$$

where the summations go over all measurements.

Exercise 10.5: Show that the result obtained in this section is consistent with the solution found in Sect.10.2.1.

Exercise 10.6: If the (independent) observations y_i have a common error σ, what are the errors on the parameters for the straight line using (a) the model $f_i = \theta_1 + x_i\theta_2$ (Sect.10.2.1), (b) the model $f_i = \omega_1 + (x_i-\bar{x})\omega_2$?

10.3 THE NON-LINEAR LEAST-SQUARES MODEL

We turn now to a more general problem when the predicted values f_i have a *non-linear* dependence on the parameters. It is then not possible to

write down an exact solution as was done for the linear case. Instead one has to perform the minimization by an iterative procedure to find increasingly better approximations for the unknown parameters.

Several iteration methods to locate the minimum values of a general function are outlined in Chapter 13. Here we describe only one method, which goes back to Isaac Newton.

10.3.1 Newton's method

According to the Least-Squares Principle the best estimates of the unknown parameters are the values which minimize the quantity

$$X^2 = (\underline{y} - \underline{f})^T V^{-1}(\underline{y})(\underline{y} - \underline{f}), \qquad (10.6)$$

where $\underline{y}$ is the vector of measurements with covariance matrix $V(\underline{y})$, and $\underline{f}$ the vector of predicted values,

$$\underline{f} = \underline{f}(\underline{\theta}; \underline{x}), \qquad (10.39)$$

which is now non-linear in $\underline{\theta}$.

Suppose that we have found a set of approximate parameter values

$$\underline{\theta}^\nu = \{\theta_1^\nu, \theta_2^\nu, \ldots, \theta_L^\nu\}, \qquad (10.40)$$

which corresponds to the ν-th iteration with the function value X^2_ν, and that a better approximation is needed. We then evaluate the derivatives of X^2 with respect to the parameters at the point $\underline{\theta} = \underline{\theta}^\nu$ (remember that these would vanish if the minimum had been obtained!). In the case of independent measurements, when $V^{-1}_{ii}(\underline{y}) = 1/\sigma_i^2$, we can write

$$g_\ell(\underline{\theta}^\nu) \equiv \frac{\partial X^2}{\partial \theta_\ell} = \sum_{i=1}^{N} \left(-\frac{2}{\sigma_i^2}\right)(y_i - f_i)\frac{\partial f_i}{\partial \theta_\ell}, \qquad \ell=1,2,\ldots,L, \qquad (10.41)$$

where f_i and $\partial f_i/\partial\theta_\ell$ are evaluated for $\underline{\theta}=\underline{\theta}^\nu$. If we can find an increment $\Delta\underline{\theta}^\nu$ to $\underline{\theta}^\nu$ which will make the gradient vector $\underline{g}$ equal to zero,

$$\underline{g}(\underline{\theta}^\nu + \Delta\underline{\theta}^\nu) = \underline{0},$$

the problem has been solved, because then the corrected value

$$\underline{\theta}^{\nu+1} = \underline{\theta}^\nu + \Delta\underline{\theta}^\nu \qquad (10.42)$$

will be the desired solution (provided, of course, that the extremum of X^2 is a minimum).

To find the correction $\Delta\underline{\theta}^\nu$ to the approximate parameter vector $\underline{\theta}^\nu$ we expand the g_ℓ of eqs.(10.41) about $\underline{\theta}^\nu$ and keep terms up to first order. Then we demand

$$g_\ell(\underline{\theta}^\nu) + \frac{\partial g_\ell}{\partial\theta_1}\Delta\theta_1^\nu + \frac{\partial g_\ell}{\partial\theta_2}\Delta\theta_2^\nu + \ldots + \frac{\partial g_\ell}{\partial\theta_L}\Delta\theta_L^\nu = 0, \qquad \ell=1,2,\ldots,L, \qquad (10.43)$$

where the derivatives are evaluated for $\underline{\theta} = \underline{\theta}^\nu$. Equations (10.43) constitute a set of L inhomogeneous, linear equations which can be solved for the unknown $\Delta\theta_\ell^\nu$, $\ell=1,2,\ldots,L$ in the usual way. The coefficient appearing before $\Delta\theta_k^\nu$ in the ℓ-th equation is

$$G_{k\ell} \equiv \frac{\partial g_\ell}{\partial\theta_k} = \frac{\partial^2 X^2}{\partial\theta_k\partial\theta_\ell} = \sum_{i=1}^{N}\left(-\frac{2}{\sigma_i^2}\right)\left[-\frac{\partial f_i}{\partial\theta_k}\frac{\partial f_i}{\partial\theta_\ell} + (y_i - f_i)\frac{\partial^2 f_i}{\partial\theta_k\partial\theta_\ell}\right] \qquad (10.44)$$

and may be found either analytically or numerically, depending on the problem. In eq.(10.44) it is understood that all expressions are evaluated for $\underline{\theta} = \underline{\theta}^\nu$. Obviously $G_{k\ell} = G_{\ell k}$, implying that G is a symmetric matrix.

In vector notation the solution for the corrections $\Delta\underline{\theta}^\nu$ is (compare the formulae of Sects.10.2.2 and 10.2.3)

$$\Delta\underline{\theta}^\nu = -G^{-1}(\underline{\theta}^\nu)\underline{g}(\underline{\theta}^\nu). \qquad (10.45)$$

This relation assumes that G is non-singular, but is valid also when X^2 is constructed for correlated measurements.

The new parameter values $\underline{\theta}^{\nu+1} = \underline{\theta}^\nu + \Delta\underline{\theta}^\nu$ are next used to find the corresponding $X^2_{\nu+1}$, and if this quantity is smaller than X^2_ν a better solution has been obtained. The procedure is then repeated taking $\underline{\theta}^{\nu+1}$ as a new approximate solution, a new correction $\Delta\underline{\theta}^{\nu+1}$ is determined, and so on. The iterations are continued until the improvement in X^2 between two consecutive iterations becomes smaller than some preset number.

If it is found that a new iteration gives $X^2_{\nu+1} > X^2_\nu$ the minimum value has been passed over. One may then redefine the correction to the ν-th step by taking a smaller value, say $\Delta\underline{\theta}^\nu = \frac{1}{2}\Delta\underline{\theta}^\nu$ and repeat the procedure with this step.

It will be seen that if the theoretical value $\underline{f}$ is of second order in the parameters the matrix G will become constant, independent of $\underline{\theta}$, and the exact minimum is found in the first iteration.

10.3.2 Example: Helix parameters in track reconstruction

As an example on an iterative X^2 minimization we will discuss the determination of track parameters for a charged particle moving in a magnetic field, for instance in a bubble chamber or a streamer chamber. We will for simplicity assume that the particle has a constant momentum over the track length considered (neglect ionization loss) and that the magnetic field is uniform. The particle trajectory in space will then be a helix with axis parallel to the direction of the magnetic field. Our problem is to determine the best parameters of this helix on the basis of a series of measured points along the path of the particle.

Let us assume that the N points in space (X_i,Y_i,Z_i), i=1,2,3...,N have been accurately measured with respect to a fixed coordinate system (xyz) with z axis along the direction of the magnetic field. Let us further for the moment assume that the starting point (A,B,C) of the track has been precisely measured. We take this point as the origin of a second coordinate system (x'y'z') with z' axis parallel to the z direction, and with the y' axis along the direction of the tangent to the track projection in the xy plane, see Fig. 10.1. In this relative coordinate system the helix can be parameterized as

$$\begin{aligned} x' &= \rho(\cos\phi - 1) \\ y' &= \rho \sin\phi \\ z' &= \rho\phi \tan\lambda , \end{aligned}$$

where ρ is the radius of curvature of the projected (circular) path in the x'y' plane, and λ the angle between this plane and the tangent to the helix (the "dip" angle). In the fixed coordinate system the helix is expressed as

$$\begin{aligned} x &= A + x'\cos\beta - y'\sin\beta \\ y &= B + x'\sin\beta + y'\cos\beta \\ z &= C + z', \end{aligned}$$

where β is the angle describing the relative orientation of the two coordinate

systems. With ϕ-values corresponding to the measured points, evaluated as indicated later, we have therefore a set of "predicted" points for i=1,2,...,N,

$$\left.\begin{aligned} x_i &= A + \rho(\cos\phi_i - 1)\cos\beta - \rho\sin\phi_i\sin\beta \\ y_i &= B + \rho(\cos\phi_i - 1)\sin\beta + \rho\sin\phi_i\cos\beta \\ z_i &= C + \rho\phi_i\tan\lambda , \end{aligned}\right\} \qquad (10.46)$$

and the LS solution for the unknown parameters is obtained by seeking the minimum of the (unweighted) expression

$$X^2 = \sum_{i=1}^{N} [(X_i-x_i)^2 + (Y_i-y_i)^2 + (Z_i-z_i)^2]. \qquad (10.47)$$

The minimization is most conveniently done in terms of the parameters ρ,β and $\tan\lambda$. With the formulation of Sect.10.3.1 the gradient vector $\underline{g}$ and the matrix G can be calculated from eqs.(10.41), (10.44). One obtains for the components of $\underline{g}$,

$$\begin{aligned} g_\rho &= \frac{\partial X^2}{\partial \rho} = -\frac{2}{\rho}\sum_i [(X_i-x_i)(x_i-A) + (Y_i-y_i)(y_i-B) + (Z_i-z_i)(z_i-C)] \\ g_\beta &= \frac{\partial X^2}{\partial \beta} = -2\sum_i [-(X_i-x_i)(y_i-B) + (Y_i-y_i)(x_i-A)] \\ g_\lambda &= \frac{\partial X^2}{\partial \tan\lambda} = -2\rho\sum_i (Z_i-z_i)\phi_i . \end{aligned}$$

The matrix G has the following non-vanishing elements,

$$\begin{aligned} G_{\rho\rho} &= \frac{\partial^2 X^2}{\partial \rho^2} = \frac{2}{\rho^2}\sum_i [(x_i-A)^2 + (y_i-B)^2 + (z_i-C)^2] \\ G_{\rho\beta} &= G_{\beta\rho} = \frac{2}{\rho}\sum_i [(X_i-x_i)(y_i-B) - (Y_i-y_i)(x_i-A)] \\ G_{\rho\lambda} &= G_{\lambda\rho} = 2\sum_i [(z_i-C) - (Z_i-z_i)]\phi_i \\ G_{\beta\beta} &= \frac{\partial^2 X^2}{\partial \beta^2} = 2\sum_i [(X_i-A)(x_i-A) + (Y_i-B)(y_i-B)] \\ G_{\lambda\lambda} &= \frac{\partial^2 X^2}{\partial (\tan\lambda)^2} = 2\rho^2\sum_i \phi_i^2 , \end{aligned}$$

while $G_{\beta\lambda} = G_{\lambda\beta} = 0$.

The starting values for the iteration procedure can be found as follows:

For the dip angle we take $\tan\lambda^o$ as the value obtained by a LS straight line fit through the N pairs of points (s'_i,z'_i), where z'_i is the measured z-coordinate expressed in the in the (x'y'z') system, and s'_i the distance from the origin of this system to the measured point (x'_i,y'_i,z'_i).

For the radius of curvature ρ^o and the rotation angle β^o we can take the values obtained by a linear LS fit to a circle through the measured projected points (X_i,Y_i,o). Writing the equation for the circle as

$$(x-A)^2 + (y-B)^2 + 2a(x-A) + 2b(y-B) = 0,$$

the fitted values of the parameters a and b give the starting values

$$\rho^o = \sqrt{a^2 + b^2}, \qquad \beta^o = \tan^{-1}\frac{b}{a}.$$

The azimuthal angle ϕ_i can then be expressed for the measured points as

$$\phi_i = \tan^{-1}\frac{y_i - (B-b)}{x_i - (A-a)} - \beta^o, \qquad i=1,2,\ldots,N.$$

When the starting values have been obtained the gradient vector $\underline{g}^o$ and the matrix G^o are evaluated according to the expressions above. The corrections $\Delta\rho^o,\Delta\beta^o,\Delta\tan\lambda^o$ are then evaluated from eq.(10.45) and the starting values replaced by the new approximation

$$\rho^1 = \rho^o + \Delta\rho^o, \qquad \beta^1 = \beta^o + \Delta\beta^o, \qquad \tan\lambda^1 = \tan\lambda^o + \Delta\tan\lambda^o.$$

A new iteration can then be made with this set and the process repeated until a satisfactory stationary value of X^2 is obtained. The iterations are stopped when the sum of the absolute values of the correction terms gets below a preset value, or after a preset maximum number of iterations has been performed.

In practice the assumption of a known origin of the track is not made. Instead one allows the coordinates A,B,C to vary and makes a helix fit with the parameter set A,C,ρ,β,tanλ, or B,C,ρ,β,tanλ. The starting values for A, (or B), and C are then taken as the coordinates of the first measured point, and the starting values for ρ and β must be found by a non-linear LS fit of a circle to the measured projected points.

It should also be mentioned that in practice the measured points are not known with full precision, but are connected with errors, $\Delta X_i,\Delta Y_i,\Delta Z_i$ and correlation terms. The minimization is accordingly carried out for a weighted X^2 function, rather than with the unweighted X^2 of eq.(10.47). The problem then implies more complicated formulae for the gradient vector $\underline{g}$ and the matrix G of second derivatives of X^2. As a result of the complete minimization one will in this situation, in addition to the fitted helix parameters, obtain a set of "improved measurements", or fitted values of the coordinates of the N space points.

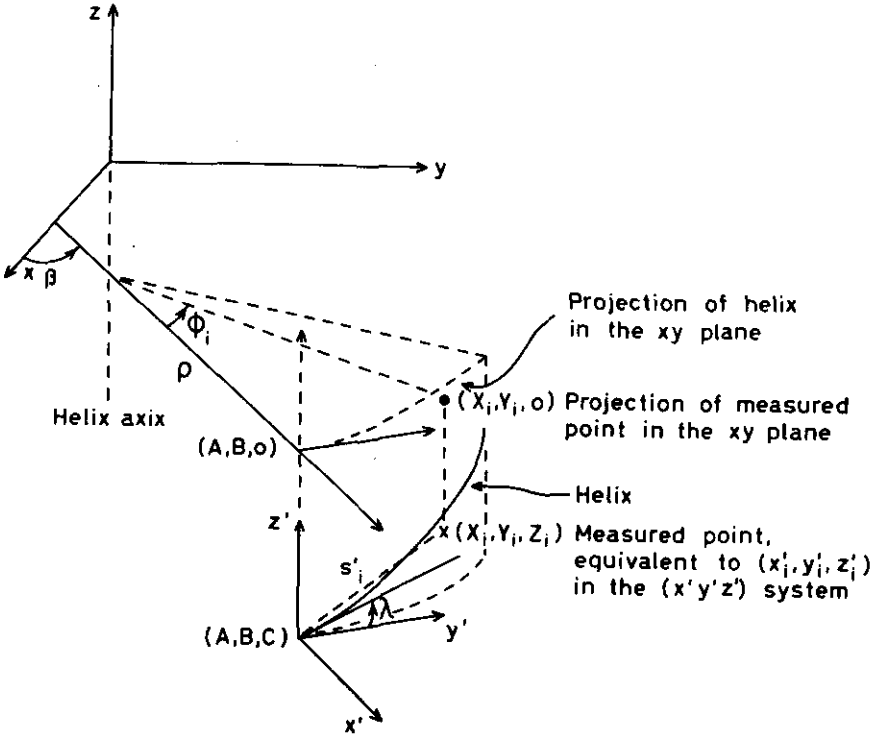

Fig. 10.1. Track reconstruction with helix parameters (see text).

10.4 LEAST-SQUARES FIT

10.4.1 "Improved measurements" (fitted variables) and residuals

In the preceding sections we have regarded the Least-Squares Principle as providing a prescription for how to find the best values of the unknown parameters $\underline{\theta}$, which are supposed to be connected to the true observables $\underline{\eta}$ through some functional dependence.

In many situations the basic unknowns are in fact the observables $\underline{\eta}$ themselves; this was for instance the case in the example of Sect.10.2.6. Characteristic for many situations is that the $\underline{\eta}$ are *directly measureable*. We may take the observations $\underline{y}$ with covariance matrix $V(\underline{y})$ as our initial estimates of the true, but unknown $\underline{\eta}$. According to the Least-Squares Principle we should then adopt as our best estimates of $\underline{\eta}$ those values which minimize the quantity

$$X^2 = \underline{\varepsilon}^T V^{-1} \underline{\varepsilon}\,, \tag{10.48}$$

where $\underline{\varepsilon}$ is the difference between the measured and true values, and may be called the *error due to measurement*,

$$\underline{\varepsilon} = \underline{y} - \underline{\eta}\,. \tag{10.49}$$

When the LS minimization has been performed the final estimates $\hat{\underline{\eta}}$ of the true $\underline{\eta}$ are called the *"improved measurements"*, or *fitted variables*.

The *residuals* $\hat{\underline{\varepsilon}}$ of the LS estimation are defined as the differences between the original measurements and the "improved measurements" from the fit,

$$\hat{\underline{\varepsilon}} = \underline{y} - \hat{\underline{\eta}}\,. \tag{10.50}$$

In other words, the residual is the estimated error due to measurement.

The minimum value obtained for X^2 can be called *the weighted sum of squared residuals* in the fit,

$$X^2_{min} = \hat{\underline{\varepsilon}}^T V^{-1} \hat{\underline{\varepsilon}} = (\underline{y}-\hat{\underline{\eta}})^T V^{-1} (\underline{y}-\hat{\underline{\eta}})\,. \tag{10.51}$$

If the covariance matrix $V(\underline{y})$ is known only up to a constant multiplicative factor σ^2, *i.e.* if

$$V(\underline{y}) = \sigma^2 V_\sigma(\underline{y})\,, \tag{10.52}$$

where $V_\sigma(\underline{y})$ is known, we define in a similar manner a quantity which is usually referred to as *the residual sum of squares*,

$$Q^2_{min} = \hat{\underline{\varepsilon}}^T V_\sigma^{-1} \hat{\underline{\varepsilon}} = (\underline{y}-\hat{\underline{\eta}})^T V_\sigma^{-1} (\underline{y}-\hat{\underline{\eta}})\,. \tag{10.53}$$

The connection between the two quantities X^2_{min} and Q^2_{min} is

$$X^2_{min} = \frac{1}{\sigma^2} Q^2_{min}\,. \tag{10.54}$$

The definitions above make no reference to any specific type of model. In the subsequent sections we shall derive, or merely state, som results which are exact for the linear model, and approximately valid for any general model in which the parameter dependence is not too far from linear.

Exercise 10.7: For the linear parabola fit of Sect.10.2.5, find the "improved measurements" $\hat{\underline{\eta}}$ and the residuals $\hat{\underline{\varepsilon}}$. What is the weighted sum of squared residuals X^2_{min} in this case?

Exercise 10.8: (i) For the linear LS model of Sect.10.2.3, in which $\hat{\underline{\eta}} = A\hat{\underline{\theta}}$, show that the "improved measurements" and their covariance matrix are given by, respectively,

$$\hat{\underline{\eta}} = A(A^T V^{-1} A)^{-1} A^T V^{-1} \underline{y}\,, \qquad V(\hat{\underline{\eta}}) = A(A^T V^{-1} A)^{-1} A^T\,.$$

(ii) Show that the residuals can be expressed as $\hat{\underline{\varepsilon}} = D\underline{\varepsilon}$, where $\underline{\varepsilon}$ is the error due to measurement and

$$D \equiv I_N - A(A^T V^{-1} A)^{-1} A^T V^{-1}\,.$$

(iii) Show that the weighted sum of squared residuals for the linear model reduces to

$$X^2_{min} = \underline{y}^T V^{-1} \hat{\underline{\varepsilon}}\,.$$

Exercise 10.9: With uncorrelated measurements of common error σ and the orthogonal polynomial formulation of Sect.10.2.8, show that the residual sum of squares can be calculated from the simple formula

$$Q^2_{min} = \sum_{i=1}^{N} y_i^2 - \sum_{\ell=1}^{k} \omega_\ell^2 = \underline{y}^T \underline{y} - \hat{\underline{\omega}}^T \hat{\underline{\omega}}\,.$$

10.4.2 Estimating σ^2 in the linear model

We have emphasized that to obtain the solution $\hat{\underline{\theta}}$ (or $\hat{\underline{\eta}}$) of the linear LS problem the covariance matrix $V(\underline{y}) = \sigma^2 V_\sigma(\underline{y})$ of the measurements need only be known up to the multiplicative factor σ^2. However, to find the errors on the estimates the matrix $V(\underline{y})$ must be completely known.

It can be shown that, with $V_\sigma(\underline{y})$ known, the unknown σ^2 can be estimated from the residual sum of squares Q^2_{min}, since

$$s^2 = \frac{Q^2_{min}}{N - L} \tag{10.56}$$

is an unbiassed estimator of σ^2; here, as before, N is the number of observations and L the number of parameters estimated*).

We shall prove that s^2 of eq.(10.56) is an unbiassed estimator of σ^2 for the simplest case where the uncorrelated measurements have a common error. Then

$$V(\underline{y}) = \sigma^2 I_N , \tag{10.57}$$

where I_N is the unit matrix of dimension N × N. The residuals $\hat{\underline{\varepsilon}}$ can now be expressed in terms of the errors due to measurement $\underline{\varepsilon}$. Writing $\underline{y} = A\underline{\theta} + \underline{\varepsilon}$ we have (see Exercise 10.8)

$$\hat{\underline{\varepsilon}} = \underline{y} - \hat{\underline{\eta}} = (A\underline{\theta}+\underline{\varepsilon}) - A(A^TA)^{-1}A^T(A\underline{\theta}+\underline{\varepsilon}) = D\underline{\varepsilon} , \tag{10.58}$$

when we identify

$$D \equiv I_N - A(A^TA)^{-1}A^T. \tag{10.59}$$

The matrix D is *idempotent*, satisfying $D^T = D$, $D^TD = D$. The residual sum of squares therefore becomes

$$Q^2_{min} = \hat{\underline{\varepsilon}}^T\hat{\underline{\varepsilon}} = (D\underline{\varepsilon})^T(D\underline{\varepsilon}) = \underline{\varepsilon}^TD\underline{\varepsilon} = \sum_i D_{ii}\varepsilon_i^2 + \sum_{i\neq j} D_{ij}\varepsilon_i\varepsilon_j . \tag{10.60}$$

If we take the expectation value of this expression the contributions from the last sum will vanish, since we have assumed uncorrelated measurements, for which $E(\varepsilon_i\varepsilon_j) = 0$ when $i\neq j$. Hence

$$E(Q^2_{min}) = E\left(\sum_i D_{ii}\varepsilon_i^2\right) = \sigma^2\sum_i D_{ii} = \sigma^2 trD . \tag{10.61}$$

*) If the L parameters are constrained in K linear algebraic equations the unbiassed estimator of σ^2 is given by $s^2 = Q^2_{min}/(N-L+K)$; see Exercise 10.16.

From the definition of D we find by commuting the matrix $A(A^TA)^{-1}$ by A^T under the trace operation

$$trD = tr\{I_N\} - tr\{A^TA(A^TA)^{-1}\} = tr\{I_N\} - tr\{I_L\} ,$$

where I_L is the identity matrix of dimension L × L. Thus, trD = N-L, and eq. (10.61) leads to

$$E(Q^2_{min}) = \sigma^2(N-L). \tag{10.62}$$

By comparison with the definition of an unbiassed estimator (eq.(8.3)) we see, therefore, that $Q^2_{min}/(N-L)$ will be an unbiassed estimator of σ^2, as was stated above.

Exercise 10.10: Prove that, for any N × N matrix G and a general covariance matrix V,

$$E(\underline{\varepsilon}^TG\underline{\varepsilon}) = tr(VG)$$

in virtue of the definition of the covariances, $E(\varepsilon_i\varepsilon_j) = V_{ij}$.

Exercise 10.11: Generalize the proof in the text and show that $s^2 = Q^2_{min}/(N-L)$ is an unbiassed estimator of σ^2 also if the measurements are correlated and have unequal errors. (Hint: with $V(\underline{y}) = \sigma^2V_\sigma(\underline{y})$ and $V_\sigma(\underline{y})$ known, the residual sum of squares can be expressed as

$$Q^2_{min} = \hat{\underline{\varepsilon}}^TV_\sigma^{-1}\hat{\underline{\varepsilon}} = (D\underline{\varepsilon})^TV_\sigma^{-1}(D\underline{\varepsilon}) = \underline{\varepsilon}^T(D^TV_\sigma^{-1}D)\underline{\varepsilon} ,$$

where

$$D \equiv I_N - A(A^TV_\sigma^{-1}A)^{-1}A^TV_\sigma^{-1}.$$

This matrix has the property that $D^TV_\sigma^{-1}D = V_\sigma^{-1}D = D^TV_\sigma^{-1}$, so that with the result of the preceding exercise, $E(Q^2_{min}) = \sigma^2 trD = \sigma^2(N-L)$.)

10.4.3 The normality assumption; degrees of freedom

For linear models the LS method produces estimators of the unknown parameters $\underline{\theta}$ which are unbiassed and have minimum variance (Sect.10.2.4) and, as we saw in the previous section, it also provides an unbiassed estimator of σ^2. In terms of the error due to measurement $\underline{\varepsilon}$ the only assumption made in deriving these properties is that $E(\underline{\varepsilon}) = 0$, which is required for the proof of unbiassedness. To prove that $Q^2_{min}/(N-L)$ is an unbiassed estimator of σ^2 we also assumed in Sect.10.4.2 that the measurements were uncorrelated, with $E(\varepsilon_i\varepsilon_j) = 0$ for $i\neq j$, but this restriction is not essential (Exercise 10.11). Thus, except

for the first and second moments, no distributional assumptions have been made about the ε_i's; in other words, the optimal properties are *distribution-free*.

Let us now make the further assumption that the *uncorrelated* ε_i's are *normally* distributed with mean value 0 and variance σ_i^2. As uncorrelated, normal variables are independent (Sect.4.10.1) this amounts to assuming that *the N measurements y_i are independent and normally distributed* with variance σ_i^2 about the true values η_i.

If the observables η_i were known the assumptions above would imply that the quantity

$$X^2 = \sum_{i=1}^{N} \left(\frac{\varepsilon_i}{\sigma_i}\right)^2 = \sum_{i=1}^{N} \left(\frac{y_i-\eta_i}{\sigma_i}\right)^2 \qquad (10.63)$$

would be a sum of N independent squared standard normal variables, and would by definition (Sect. 5.1.1) be a chi-square variable with N degrees of freedom.

Since, however, we do not know what the precise values of the true, unknown η_i are we must be satisfied with adopting their *estimated* values $\hat{\eta}_i$ as obtained from the minimization of X^2. Inserted in X^2 this gives the weighted sum of squared residuals,

$$X^2_{min} = \sum_{i=1}^{N} \left(\frac{\hat{\varepsilon}_i}{\sigma_i}\right)^2 = \sum_{i=1}^{N} \left(\frac{y_i-\hat{\eta}_i}{\sigma_i}\right)^2 . \qquad (10.64)$$

Using the normality assumption about the N independent y_i it can be shown, in the case of a linear model with L parameters, that X^2_{min} can be expressed as a sum of (N-L) independent terms, each term being the square of a standardized normally distributed variable. Hence X^2_{min} is χ^2(N-L), a chi-square variable with (N-L) degrees of freedom*).

It has been tacitly assumed so far that the L parameters are independent. If the parameters are internally related in K linear algebraic equations only (L-K) of them are independent, giving (N-(L-K)) independent terms in the X^2_{min} of eq.(10.64). In this situation X^2_{min} is distributed as χ^2(N-L+K). Thus the number of degrees of freedom in the constrained as well as in the unconstrained linear fit is equal to the difference between the number of independent

*) If the ε_i's have mean values different from zero X^2_{min} will have a non-central chi-square distribution; compare Exercise 5.12.

measurements and the number of independent parameters.

If the measurements are not independent it can be shown that the above statements will still be true provided that the measurements are *multinormally* distributed about the true values. When the error due to measurements $\underline{\varepsilon} = \underline{y}-\underline{\eta}$ is multinormally distributed with mean value $\underline{0}$ and constant (non-singular) covariance matrix V, the weighted sum of squared residuals

$$X^2_{min} = \underline{\hat{\varepsilon}}^T V^{-1} \underline{\hat{\varepsilon}} = \sum_{i=1}^{N} \sum_{j=1}^{N} (y_i-\hat{\eta}_i) V^{-1}_{ij} (y_j-\hat{\eta}_j) \qquad (10.65)$$

will be chi-square distributed as before, with (N-L+K) degrees of freedom in the case when there are K constraint equations restricting the L parameters.

It is essentially the optimum properties of *linear* estimators that make X^2_{min} a chi-square variable for normally distributed variables. In a *non-linear* LS estimation the situation is not so simple; the LS estimator is then in general biassed and not of minimum variance, and the exact distribution of X^2_{min} is not known. Asymptotically, for large N, it can be shown, however, that X^2_{min} is approximately chi-square distributed also in this general case.

Finally, let us stress again that for the estimation problem the Least-Squares Principle involved no assumption about the distributional properties of the observations. The commonly used terms "chi-square (or χ^2-) minimization", "χ^2-fitting", *etc.*, the origin of which is evident from the above considerations, are therefore somewhat misleading and should be avoided. As long as the subject is parameter estimation as such one should instead use the appropriate terminology with "Least-Squares...".

As we shall see immediately below the normality assumption about the observations is essential only when it comes to the question of judging how reliable the estimated parameter values are.

Exercise 10.12: For the linear LS model of Sect. 10.2.3, show that, in general,

$$X^2 = X^2_{min} + (\underline{\theta}-\underline{\hat{\theta}})^T A^T V^{-1} A (\underline{\theta}-\underline{\hat{\theta}}).$$

For normally distributed observations each of the three terms is chi-square distributed, as χ^2(N), χ^2(N-L), and χ^2(L), respectively.

10.4.4 Goodness-of-fit

The fact that the weighted sum of squared residuals X^2_{min}, for normally distributed measurements, has a known (*i.e.* chi-square) distribution has an important practical consequence. It implies that the X^2_{min} value obtained in a particular minimization can be used to give a quantitative measure of how close the overall agreement is between the fitted quantities $\hat{\underline{\eta}}$ and the measurements $\underline{y}$. In other words, X^2_{min} will provide a measure of the *goodness-of-fit*.

For definiteness, let us assume that the problem involves ν degrees of freedom; in common parlance we are then dealing with a "ν-*constrained fit*", or a "νC-fit". By comparison with a graph (for example, Fig. 5.2) or a table (for example, Appendix Table A8) of chi-square probabilities we can then deduce the probability corresponding to the contents of the chi-square p.d.f. $f(u;\nu)$ between the values $u = X^2_{min}$ and $u = \infty$. Obviously this *chi-square probability* P_{χ^2} then gives the probability for obtaining, in a new minimization with similar measurements and the same model, a higher value for X^2_{min}. A small value of X^2_{min} corresponds to a large P_{χ^2}, or a "good" fit, while a very large value X^2_{min} implies a small P_{χ^2}, or a "bad" fit.

We have

$$P_{\chi^2} = \int_{X^2_{min}}^{\infty} f(u;\nu)\,du = 1 - F(X^2_{min};\nu), \qquad (10.66)$$

where $F(X^2_{min};\nu)$ is the cumulative chi-square distribution for ν degrees of freedom. Since a cumulative integral is itself a variable which is uniformly distributed between 0 and 1 (compare Sects.4.5.1 and 5.1.4) the chi-square probability P_{χ^2} will also have a *uniform* distribution over the interval $[0,1]$.

If, in a series of similar minimizations, P_{χ^2} turns out to have a non-uniform distribution, this indicates that the assumptions specified in Sect. 10.4.3 are not fulfilled. One may then suspect the measurements or the model, or both, to be unsatisfactory and should examine this further. For example, if P_{χ^2} is strongly peaked at very low probabilities this may reveal a contamination of "wrong" events. Similarly, a skew distribution for P_{χ^2} with an excess on the high (or low) probability side may indicate that the errors in the measurements have systematically been put too high (low).

It is a unique feature of the LS method that assertions can be made about the quality of the fit, and hence of the estimated parameters, from the final numerical value of the optimized quantity. The other estimation methods do not provide this possibility. We shall come back to the question of goodness-of-fit in connection with the subject of hypothesis testing in Chapter 14, where in particular we shall discuss how one can make goodness-of-fit statements when the unknown parameters have been estimated by the ML method (Sect.14.4.3).

10.4.5 Stretch functions, or "pulls"

As discussed in the previous section the weighted sum of squared residuals, X^2_{min}, gives a measure of the similarity between the observations and the fitted values. Specifically, a very large X^2_{min}, or equivalently, a low chi-square probability P_{χ^2} expresses a rather poor overall agreement between data and fitted model, and may lead one to suspect that the model is unsatisfactory.

Quite frequently, however, one finds that an unexpected large value of X^2_{min} not necessarily has to be ascribed to a wrong model or hypothesis, as it can simply be due to a large contribution from one, or a few, of the N points. Thus, rather than immediately abandoning the model from a "bad" X^2_{min} a quick look at the data points is recommended. Sometimes this inspection shows at once that one of the input values was incorrect, and a satisfactory fit can be obtained when the mistake is corrected.

A closer study of the fit can be done by looking at the residuals $\hat{\varepsilon}_i = y_i - \hat{\eta}_i$, which directly measure the deviations between the observations and the fitted values. To allow for different accuracies it is reasonable to judge an $\hat{\varepsilon}_i$ relatively to the uncertainty, or standard deviation $\sigma(\hat{\varepsilon}_i)$, in this quantity. Thus the examination of the fit should be done in terms of the variables

$$z_i \equiv \frac{\varepsilon_i}{\sigma(\varepsilon_i)}, \qquad i=1,2,\ldots,N. \qquad (10.67)$$

z_i is called the *stretch function* or *"pull"* for the i-th observation. Considering uncorrelated observations and a sufficiently linear estimation problem we have

$$\sigma^2(\hat{\varepsilon}_i) = V_{ii}(\underline{y}-\hat{\underline{\eta}}) = V_{ii}(\underline{y}) - 2\mathrm{cov}(\underline{y},\hat{\underline{\eta}})_{ii} + V_{ii}(\hat{\underline{\eta}})$$

$$= V_{ii}(\underline{y}) - V_{ii}(\hat{\underline{\eta}}). \qquad (10.68)$$

Hence, the i-th "pull" can be expressed as

$$z_i = \frac{y_i - \hat{\eta}_i}{\sqrt{\sigma^2(y_i) - \sigma^2(\hat{\eta}_i)}} . \qquad (10.69)$$

Clearly the minus sign in the denominator here has its origin in the fact that the two quantities in the numerator are completely (positively) correlated.

The "pull" z_i is anticipated to have a distribution which is fairly close to N(0,1). If, in a particular fit, one of the z_i's deviates very much from the others in magnitude the corresponding data point should be examined, and perhaps abondoned if it looks suspicious (Sect.6.1). This critique of the data is most likely to be useful when the number of degrees of freedom ν is fairly large. For 1C-fits (ν=1) one sees that all "pulls" are of the same magnitude, and they contain no more information than the value X^2_{min}.

In the long run, if the shape of the observed distribution of a "pull" z_i is shifted relatively to zero this demonstrates a certain bias in the i-th observation. Similarly, if the observed "pull" distribution is substantially broader (narrower) than N(0,1) the error in the i-th observation has probably consistently been taken too small (large).

10.5 APPLICATION OF THE LEAST-SQUARES METHOD TO CLASSIFIED DATA

10.5.1 Construction of X^2

In practice one often groups the measurements (events) according to some classification scheme, for example by plotting a histogram, before the actual estimation of the parameters.

Let the range of the variable*) x be divided into N mutually exclusive classes (bins) and denote by n_i the number of the n observations $x_1, x_2, \ldots, x_n$ belonging to the i-th class. We assume that, with the parameters $\underline{\theta} = \{\theta_1, \theta_2, \ldots, \theta_L\}$ we know the probability $p_i = p_i(\underline{\theta})$ of getting an observation

*) In this section x may denote a one-dimensional or a multi-dimensional variable.

in class i. In the case of a continuous variable x we may have to find p_i by integrating a probability density function over the width Δx_i of the i-th class. The expected number of observations in this class is

$$f_i(\underline{\theta}) = np_i(\underline{\theta}), \qquad (10.70)$$

and the normalization condition $\sum_{i=1}^{N} p_i = 1$ implies

$$\sum_{i=1}^{N} n_i = \sum_{i=1}^{N} f_i(\underline{\theta}) = n. \qquad (10.71)$$

For a given n the numbers of observations n_i are multinomially distributed over the N classes with covariance matrix

$$V(\underline{y}) = \begin{pmatrix} np_1(1-p_1) & -np_1p_2 & \cdots & -np_1p_N \\ -np_1p_2 & np_2(1-p_2) & \cdots & -np_2p_N \\ \cdot & \cdot & & \\ \cdot & \cdot & & \\ \cdot & \cdot & & \\ -np_1p_N & -np_2p_N & \cdots & np_N(1-p_N) \end{pmatrix}. \qquad (10.72)$$

Because of the normalization condition this matrix is singular ($|V| = 0$) and can not be inverted. The Least-Squares Principle as formulated by eq.(10.6) is therefore not applicable to this case. However, if we omit one of the n_i, say n_N, as it is redundant, the remaining (N-1) n_i will correspond to a covariance matrix V^* which is simply V(y) with its N-th row and column deleted. We could then reformulate the Least-Squares Principle for finding the best values of the parameters $\underline{\theta}$ by demanding the minimum of the quantity

$$X^2 = (\underline{y}-\underline{np})^T(V^*)^{-1}(\underline{y}-\underline{np}) = \sum_{i=1}^{N-1}\sum_{j=1}^{N-1}(n_i-np_i)(V^*)^{-1}_{ij}(n_j-np_j). \qquad (10.73)$$

It can be verified by the reader that the inverse of the matrix V^* is

$$(V^*)^{-1} = \frac{1}{n}\begin{pmatrix} p_1^{-1}+p_N^{-1} & p_N^{-1} & \cdots & p_N^{-1} \\ p_N^{-1} & p_2^{-1}+p_N^{-1} & \cdots & p_N^{-1} \\ \cdot & \cdot & & \\ \cdot & \cdot & & \\ \cdot & \cdot & & \\ p_N^{-1} & p_N^{-1} & \cdots & p_{N-1}^{-1}+p_N^{-1} \end{pmatrix}. \qquad (10.74)$$

The double sum in X^2 above can therefore be written as

$$X^2 = \frac{1}{n}\left\{\sum_{i=1}^{N-1}\frac{(n_i-np_i)^2}{p_i} + \frac{1}{p_N}\sum_{i=1}^{N-1}\sum_{j=1}^{N-1}(n_i-np_i)(n_j-np_j)\right\}$$

$$= \frac{1}{n}\left\{\sum_{i=1}^{N-1}\frac{(n_i-np_i)^2}{p_i} + \frac{1}{p_N}\left[\sum_{i=1}^{N-1}(n_i-np_i)\right]^2\right\}$$

$$= \frac{1}{n}\left\{\sum_{i=1}^{N-1}\frac{(n_i-np_i)^2}{p_i} + \frac{1}{p_N}(n_N-np_N)^2\right\},$$

or,

$$X^2 = \sum_{i=1}^{N}\frac{(n_i-np_i)^2}{np_i} = \sum_{i=1}^{N}\left(\frac{n_i-np_i}{\sqrt{np_i}}\right)^2. \qquad (10.75)$$

The last expression restores the symmetry in all N classes and corresponds to the formulation of eq.(10.3) with $f_i=np_i$, $\sigma_i=\sqrt{np_i}$. The expression (10.75) could have been written down at once, from the assumption that the number of events n_i is Poisson distributed with mean and variance equal to np_i. The algebra above thus demonstrates again the mathematical equivalence between two different points of view, the first considering N (dependent) multinomially distributed variables conditioned on their sum, the second considering N independent Poisson variables; compare Sect.4.4.4.

Generally, in the covariance matrix of eq.(10.72), if the number of classes is large such that all p_i's are small, the off-diagonal terms become negligible, and

$$V_{ii}(\underline{y}) = \sigma_i^2 = np_i(1-p_i) \simeq np_i = f_i. \qquad (10.76)$$

Therefore, whenever the LS estimates of the parameters are found by minimizing

$$X^2 = \sum_{i=1}^{N}\frac{(n_i-f_i)^2}{\sigma_i^2} \simeq \sum_{i=1}^{N}\frac{(n_i-f_i)^2}{f_i} \qquad (10.77)$$

the method implies a Poisson approximation for the individual n_i, (mean and variance = f_i). Equating the derivatives of X^2 to zero gives now the following set of equations,

$$\frac{\partial X^2}{\partial\theta_\ell} = -2\sum_{i=1}^{N}\left(\frac{n_i-f_i}{f_i} + \frac{1}{2}\left(\frac{n_i-f_i}{f_i}\right)^2\right)\frac{\partial f_i}{\partial\theta_\ell} = 0, \qquad \ell = 1,2,\ldots,L. \quad (10.78)$$

Even for simple models it may be difficult to solve eqs.(10.78) analytically, and often a numerical minimization of eq.(10.77) is used instead. It can be shown, however, that for large number of events n, the influence from the second term in the parentheses of eqs.(10.78) becomes small. Neglecting this term corresponds to regarding the f_i in the denominator of eq.(10.77) as constants independent of $\underline{\theta}$ and leads to a simpler form of eqs.(10.78),

$$\frac{\partial X^2}{\partial\theta_\ell} = -2\sum_{i=1}^{N}\frac{n_i-f_i}{f_i}\frac{\partial f_i}{\partial\theta_\ell} = 0, \qquad \ell = 1,2,\ldots,L, \quad (10.79)$$

which may be easier to handle.

The variance σ_i^2 is sometimes approximated by n_i instead of by f_i. If, rather than eq.(10.77), we minimize

$$X^2 = \sum_{i=1}^{N}\frac{(n_i-f_i)^2}{\sigma_i^2} \simeq \sum_{i=1}^{N}\frac{(n_i-f_i)^2}{n_i}, \qquad (10.80)$$

the solution for $\underline{\theta}$ is more sensitive to statistical fluctuations in the observed data. It can be shown, however, that for large numbers of events the solutions obtained from the two formulations (*i.e.* eqs.(10.77), (10.80)) coincide. It can further be shown that the formulations correspond to estimators which, for large samples, possess optimum theoretical properties: they are consistent, asymptotically normal, and efficient (*i.e.* give minimum variance).

Asymptotically, in the limit of large numbers, the X^2 of eq.(10.73) will if $E(n_i) = np_i$, be distributed as $\chi^2(N-1)$, as will also the alternative and approximate expressions for X^2 given above. Compared to the situation of Sect.10.4.3 one degree of freedom has been lost because of the normalization condition $\sum_{i=1}^{N} n_i = n$, which leaves only (N-1) of the observations independent.

When the minimization has been performed the minimum value X^2_{min} can therefore be used to give an approximate measure of the goodness-of-fit, by comparison with the chi-square distribution with (N-1-L) degrees of freedom. Here, as before, L is the number of (independent) parameters estimated. From eqs. (10.77) and (10.80) it is clear that X^2_{min} is $\chi^2(N-1-L)$ to the extent that the numbers are large enough to justify the assumption that $(n_i-f_i)/\sqrt{f_i}$ and

$(n_i-f_i)/\sqrt{n_i}$ are approximate standard normal variables*).

10.5.2 Choice of classes

There is to date no generally accepted prescription for how to subdivide the variable range into categories or classes. In many experiments, however, the set-up of the apparatus itself implies a grouping of the data, for example defined by the angular acceptance of an electronic counter. When the grouping of the data is not *a priori* given, essentially two different approaches are used for the subdivision of the range of the variable,

(i) the *equal-width* method, where the range is divided into classes of equal width, and

(ii) the *equal-probability* method, where the range is divided into classes of equal expected probability.

We have already assumed that the number N of classes is large to justify the approximation $\sigma_i^2 \simeq f_i$ in eq.(10.77), or $\sigma_i^2 \simeq n_i$ in eq.(10.80). If X^2_{min} is further to be used to measure the quality of the fit, a few more conditions have to be fulfilled.

Firstly, it is not allowed to choose the limits of the classes in such a way as to make X^2_{min} as small as possible. This follows from the fact that the statistic X^2_{min} will only be an approximate chi-square variable if the class boundaries for x are not random variables. In most practical work, the grouping is made from computational convenience. The second condition, already mentioned, is that the number of expected observations within each class must be "large", which is necessary to approximate $(n_i-f_i)/\sqrt{f_i}$ (or $(n_i-f_i)/\sqrt{n_i}$) to a standard normal variable. Fortunately, for practical purposes the expected numbers need not be very large; in fact it is customary to require a minimum of five entries in each class. It has been verified, using the equal-width subdivision, that one or two classes may be allowed to have expectations even less

*) Estimating parameters by minimizing eq.(10.77) is known in the literature as the *minimum* χ^2 *method*, whereas minimizing eq.(10.80) is sometimes called the *modified minimum* χ^2 *method*. In accordance with our remarks at the end of Sect.10.4.3 we discourage the use of these terms. In particular, we shall refer to the estimation by eq.(10.80) as the *simplified Least-Squares method*.

than five, provided the number of degrees of freedom is sufficiently large, say at least six. Since the probabilities frequently drop off at the ends of the variable range, one may often have to use wider widths at the extremes of the range to get sufficiently large expected numbers in these classes.

10.5.3 Example: Polarization of antiprotons (2)

Let us use the simple polarization example described under the Maximum-Likelihood method in Sect.9.5.7 to illustrate how the different Least-Squares formulations of Sect.10.5.1 lead to minimizations of varying complexity.

We assume that n double-scattering events have been observed and plotted in a histogram as a function of $x \equiv \cos\phi$, where ϕ is the angle between the normals of the two scattering planes. The histogram has N bins, and the i-th bin, extending from x_i to $x_i+\Delta x_i$, contains n_i observed events. The expected frequency for this bin is found by integrating the p.d.f. of eq.(9.38) over the bin width; hence the expected number of events is

$$f_i = n \int_{x_i}^{x_i+\Delta x_i} \tfrac{1}{2}(1+\alpha x)\,dx = n(a_i+b_i\alpha),$$

where

$$a_i \equiv \tfrac{1}{2}\Delta x_i, \quad b_i \equiv \tfrac{1}{2}\Delta x_i(x_i+\tfrac{1}{2}\Delta x_i).$$

With the Poisson approximation, $\sigma_i^2 \simeq f_i$, the exact and the simplified consequence of eq.(10.77) (*i.e.* eqs.(10.78) and (10.79), respectively) lead to equations of degree 2N and N in the parameter α. The parameter estimate $\hat{\alpha}$ and its error must therefore be found by some numerical method.

With the alternative approximation, $\sigma_i^2 \simeq n_i$, however, we have, from eq.(10.80) for the simplified LS method,

$$X^2 = \sum_{i=1}^{N} \frac{[n_i - n(a_i+b_i\alpha)]^2}{n_i}, \qquad (10.81a)$$

which is a function of second order in α. The solution of the equation $dX^2/d\alpha=0$ can therefore be stated explicitly as

$$\hat{\alpha} = \frac{1}{n}\sum_{i=1}^{N}\left(b_i - \frac{na_ib_i}{n_i}\right) \Big/ \sum_{i=1}^{N}\frac{b_i^2}{n_i}. \qquad (10.82)$$

Since X^2 can be expressed as

$$X^2 = X^2_{min} + \left(\sum_{i=1}^{N} \frac{n_i^2 b_i^2}{n_i}\right)(\alpha-\hat{\alpha})^2 \qquad (10.81b)$$

the variance of the LS estimate $\hat{\alpha}$ is exactly given by the inverse of the coefficient to the parabolic term, compare Sect.10.9.2. Hence an analytic formula can also be given for the error $\Delta\hat{\alpha}$,

$$\Delta\hat{\alpha} = \frac{1}{n}\left(\sum_{i=1}^{N} \frac{b_i^2}{n_i}\right)^{-\frac{1}{2}}. \qquad (10.83)$$

The reader may find it a rewarding exercise to rephrase the last example in terms of the linear model formulation of Sect.10.2.3, and verify that the expressions for $\hat{\alpha}$ and $\Delta\hat{\alpha}$ above correspond to the general formulae eqs.(10.23) and (10.24), respectively; compare also Exercise 10.12.

10.5.4 Example: Angular momentum analysis (1)

Suppose that in a preliminary study of the pion-nucleon interaction we want to examine the production and subsequent two-pion decay of a boson B, *viz.*

$$\pi + N \rightarrow B + N \qquad (10.84)$$
$$\hookrightarrow \pi_a + \pi_b$$

If little is known about the interaction a good way of starting an investigation is to study the angular distribution of the decay pions in the rest frame of the system B as a function of the mass of this system.

Let the decay angle $\Omega=(\theta,\phi)$ be defined as the direction of the π_a momentum in the B rest frame relatively to the quantization axis. The angular decay distribution can then generally be expressed in terms of the spherical harmonic functions $Y_J^m(\cos\theta,\phi)$, and the density matrix elements $\rho_{mm'}$, as

$$f(\cos\theta,\phi) = \sum_{m,m'} Y_J^m(\cos\theta,\phi)\rho_{mm'}Y_J^{m'}(\cos\theta,\phi)^*, \qquad (10.85)$$

where J is the spin of B. From the properties of the spherical harmonics

$$\left.\begin{aligned} &\int_{4\pi} Y_j^m(\Omega)Y_{j'}^{m'}(\Omega)^* d\Omega = \delta_{jj'}\delta_{mm'}, \\ &\int_{4\pi} Y_j^m(\Omega)d\Omega = 0, \qquad \ldots \end{aligned}\right\} \qquad (10.86)$$

the distribution can also be written as a linear combination of the Y_j^m,

$$f(\cos\theta,\phi) \equiv W(\Omega) = \sum_{j=0,2,4,\ldots}^{j=2J} \sum_{m=-j}^{m=+j} c_{jm}Y_j^m(\Omega). \qquad (10.87)$$

In eq.(10.87) only spherical harmonics with even values of j occur, implying that in the two-pion decay of a spin J boson the angular distribution will be a polynomial in $\cos\theta$ of degree at most 2J; this is known as the *maximum complexity theorem*.

The expansion coefficients c_{jm} in eq.(10.87) constitute the set of unknown parameters which we want to estimate. Since, for a given J, there are (J+1)(2J+1) terms in the double sum this gives the number of elements in the parameter vector $\underline{c}$. The c_{jm} are not all independent, since they must satisfy the normalization condition

$$\int_{4\pi} W(\Omega)d\Omega = 1. \qquad (10.88)$$

For the chosen mass region of the two-pion system B in (10.84) we classify the events in N angular intervals $\Delta\Omega_i$. The probability for the i-th interval is

$$p_i(\underline{c}) = \int_{\Delta\Omega_i} W(\Omega)d\Omega = \sum_{j,m} c_{jm}\int_{\Delta\Omega_i} Y_j^m(\Omega)d\Omega, \qquad i = 1,2,\ldots,N, \qquad (10.89)$$

where the summation goes over all (J+1)(2J+1) combinations of the indices j,m. With a total number of n events the predicted number for the i-th interval is

$$f_i(\underline{c}) = n\, p_i(\underline{c}), \qquad i = 1,2,\ldots,N. \qquad (10.90)$$

The vector of observed numbers of events in the N intervals is $\underline{y}=\{n_1,n_2,\ldots,n_N\}$, where

$$\sum_{i=1}^{N} n_i = \sum_{i=1}^{N} f_i(\underline{c}) = n. \qquad (10.91)$$

If the number of intervals is sufficiently large to justify the approximation that each n_i is an independent Poisson variable with mean and variance equal to f_i, the unknown parameters would be found by minimizing eq.(10.77), or rather eq.(10.80) if the approximation $\sigma_i^2 \simeq n_i$ is acceptable. Clearly, with

either formulation, a general numerical minimization procedure is called for.

In actual experiments the determination of the coefficients c_{jm} as functions of the mass of the system B can serve different purposes. Firstly, it may be used to determine the spin of the decaying boson from the maximum complexity theorem in the following manner: Starting from the lowest possible spin value one evaluates consecutively the sets of coefficients $\hat{c}_{jm}$ for increasing values of J. From a certain J value on, say from $J=J_{max}$, the goodness-of-fit will not become significantly better with increasing J, and all coefficients $\hat{c}_{jm}$ for $j \geq 2J_{max}$ will be compatible with zero. The value J_{max} then determines a lower limit for the spin of the boson B. Secondly, the behaviour of the coefficients with different j,m may sometimes give evidence for the presence of more than one resonance in a certain mass region. Finally, if the spins of the produced resonances are well-known, the magnitude of the coefficients $\hat{c}_{jm}$ (or, equivalently, the density matrix elements ρ_{mm}) can supply information on the production process, and be used to test production models.

Exercise 10.13: Derive eq.(10.87) from eq.(10.85). (Hint: Use the property $Y_\ell^{m*} = (-1)^m Y_\ell^{-m}$ and the product theorem for spherical harmonics,

$$Y_\ell^m Y_\ell^{m'} = \frac{2\ell+1}{4\pi} \sum_j \langle \ell,m;\ell,m' | j,m+m' \rangle \sqrt{\frac{4\pi}{2j+1}}\, Y_j^{m+m'} \langle \ell,0;\ell,0 | j,0 \rangle$$

with the fact that the Clebsch-Gordan coefficients $\langle \ell,0;\ell,0|j,0\rangle$ vanish for odd j.)

10.6 APPLICATION OF THE LEAST-SQUARES METHOD TO WEIGHTED EVENTS

So far we have in this chapter tacitly assumed that the predictions of the theoretical model are directly comparable to the experimental observation. In practice, however, the observations are frequently known to be biassed. As was discussed in Chapter 6, the observational biases can be divided into random (or statistical) errors and systematic errors. We recall from Sect.6.2 that random observational errors can be taken care of by "smearing" the ideal p.d.f. with the experimental resolution function to obtain a modified p.d.f. ("resolution transform") which can then be directly compared to the observations. Further, from Sect.6.3, systematic observational errors can be handled by two basically different approaches. The first of these is

(i) the "exact method", in which one modifies the ideal p.d.f. (*i.e.* the theoretical model) to give an "observable p.d.f.", which is subsequently compared to the observations. This approach requires that the experimental detection ability is known over the whole range of the observables.

When method (i) is not applicable one has to resort to

(ii) the "approximate method", in which one modifies the raw observations by assigning different weights to the individual observed events. The weight w_i assigned to an event is equal to the inverse of the probability for detecting this event; in other words, if one event was observed one assumes that there would have been w_i events if the detection had been perfect.

Clearly, whenever the theoretical model can be properly modified, by folding-in of the experimental resolution or by the "exact method", there is no further need for changes in the LS estimation procedure as described in the previous sections of this chapter. In fact, the most essential restriction we have put on a theoretical model is that it should give predictions which coincide with the expectation values of the observations. With the "approximate method", however, specific problems arise for the LS parameter estimation, in particular it becomes more difficult to get reliable values of the errors on the estimated parameters.

Let us assume that the observations have been classified into N bins as in Sect.10.5.1, and that it is now not meaningful to compare the predicted number of events f_i directly to the observed number of events n_i in the i-th bin. Using the "approximate method" we would say that with a perfect detection apparatus we should have observed in bin i a number of events equal to

$$n_i' = \sum_{j=1}^{n_i} w_{ij}\,, \tag{10.92}$$

where w_{ij} is the inverse of the detection probability for event j within the i-th bin. It is then suggestive to write down the following two alternative expressions to be minimized in the case of weighted events,

$$X^2 = \sum_{i=1}^{N} \frac{(n_i'-f_i)^2}{f_i} \qquad (10.93)$$

and

$$X^2 = \sum_{i=1}^{N} \frac{(n_i'-f_i)^2}{n_i'} \,. \qquad (10.94)$$

These expressions are the analogues of, respectively, eqs.(10.77) and (10.80). Again, if $E(n_i') = f_i$ and the normality assumption is reasonably fulfilled, X^2_{min} is expected to give an approximate measure of the goodness-of-fit.

If the weights w_{ij} are not too large and not too different, either of the alternative expressions (10.93) and (10.94) works satisfactorily. Experience shows, however, that curious results may sometimes be obtained for the parameter estimates and their errors if some of the events have very large weights. One should therefore be cautious and verify that the weights of the individual events do not deviate too much from the average. If the weights are large mainly in one or a few bins one may improve the reliability of the parameter estimates by simply omitting these bins in the minimization. In general, the inclusion of events with large weights tends to increase the estimated errors on the parameters.

Finally, let it be mentioned that an approach between the "exact" and the "approximate" methods consists in using the expression

$$X^2 = \sum_{i=1}^{N} \frac{(n_i-f_i')^2}{f_i'} \qquad (10.95)$$

with

$$f_i' = f_i \cdot D_i, \qquad D_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{1}{w_{ij}} \,. \qquad (10.96)$$

This would correspond to adopting the philosophy behind the "exact method", where the observations are kept unchanged and the model adjusted to account for the imperfect detection. However, the procedure is no longer exact since the modification of the ideal model rests on an *estimate* of the average detection efficiency D_i for the i-th bin, as obtained from the observed events in the bin. Since an error is inherent in this estimate it is likely that the use of eq.(10.95) implies a reduced reliability of the parameter estimates. Clearly, the method is most reliable for small weights, or nearly equal weights. If all weights are identical eq.(10.95) and eq.(10.93) give the same results.

10.7 LINEAR LEAST-SQUARES ESTIMATION WITH LINEAR CONSTRAINTS

It frequently happens that the observables $\underline{\eta}$ in an LS estimation are related through algebraic constraint equations. While the original measurements $\underline{y}$ are subject to experimental inaccuracies and may not strictly satisfy the constraints the "improved measurements" $\underline{\hat{\eta}}$, the estimates of the true, unknown values, should do so.

In general, two different approaches can be used to solve a minimization problem where the observables are restricted by algebraic constraints. The first is the *elimination method*, in which one makes use of the constraint equations to eliminate a number of the variables and subsequently performs the minimization in terms of the reduced number of unknowns. The second approach is the *method of Lagrangian multipliers*, in which one increases the number of unknowns in the minimization by adding a set of Lagrangian multipliers, one for each constraint equation. Thus both methods reformulate the constrained X^2 minimization to an (ordinary) unconstrained minimization, which can be carried out according to the procedures outlined earlier in this chapter.

In the following sections we will first (in Sect.10.7.1) see how the two methods work on a simple example before we proceed (Sect.10.7.2) to develop the general formulation of the method of Lagrangian multipliers for a linear X^2 minimization with linear constraint equations. As we shall see an *exact* solution is obtained for this problem, with closed formulae expressing a modification of the unconstrained linear LS solution.

10.7.1 Example: Angles in a triangle

To illustrate the two approaches to a linear LS estimation under linear constraints, let us consider the following simple example: The three angles of a triangle have been measured independently, giving the results

$$y_1 = 63^\circ, \qquad y_2 = 34^\circ, \qquad y_3 = 85^\circ.$$

We assume for simplicity that all measurements have an error $\sigma=1^\circ$. We want the LS estimates of the true values η_1,η_2,η_3 of the angles, which must satisfy the requirement that their sum be equal to 180°.

We observe that the measurements y_i fail to satisfy the constraint by the amount $\sum_{i=1}^{3} y_i - 180^\circ = 2^\circ$. To find the "improved measurements" $\hat{\eta}_i$

according to the LS Principle we seek the solution of the constrained minimization problem

$$\left.\begin{aligned} X^2(\eta_1,\eta_2,\eta_3) &= \sum_{i=1}^{3}\left(\frac{y_i-\eta_i}{\sigma_i}\right)^2 = \text{minimum}, \\ \sum_{i=1}^{3}\eta_i &- 180^o = 0. \end{aligned}\right\} \qquad (10.97)$$

Adopting first the eliminiation method we use the constraint equation to eliminate one of the unknowns, say η_3, substitute in X^2 and minimize with respect to the remaining two variables; thus we consider the unconstrained case

$$X^2(\eta_1,\eta_2) = \left(\frac{y_1-\eta_1}{\sigma_1}\right)^2 + \left(\frac{y_2-\eta_2}{\sigma_2}\right)^2 + \left(\frac{y_3-(180^o-\eta_1-\eta_2)}{\sigma_3}\right)^2 = \text{minimum}. \qquad (10.98)$$

This trivial minimization problem has the solution

$$\hat{\eta}_1 = \frac{1}{3}(180^o+2y_1-y_2-y_3) = 62\tfrac{1}{3}^o,$$

$$\hat{\eta}_2 = \frac{1}{3}(180^o-y_1+2y_2-y_3) = 33\tfrac{1}{3}^o,$$

and from the constraint equation we find

$$\hat{\eta}_3 = 180^o - \hat{\eta}_1 - \hat{\eta}_2 = 84\tfrac{1}{3}^o.$$

We find therefore, not unexpectedly, that the "improved measurements" are obtained by subtracting the measured "excess" of 2^o equally from all three observations.

If, instead, we adopt the method of Lagrangian multipliers we will reformulate the problem (10.97) to the equivalent form with the Lagrangian multiplier λ,

$$X^2(\eta_1,\eta_2,\eta_3,\lambda) = \sum_{i=1}^{3}\left(\frac{y_i-\eta_i}{\sigma_i}\right)^2 + 2\lambda\left(\sum_{i=1}^{3}\eta_i - 180^o\right) = \text{minimum}. \qquad (10.99)$$

This is a linear LS problem for the four unknowns $\eta_1,\eta_2,\eta_3,\lambda$. Explicitly, the normal equations become

$$\left.\begin{aligned} 0 &= \frac{\partial X^2}{\partial \eta_1} = -\frac{2}{\sigma^2}(y_1-\eta_1) + 2\lambda, \\ 0 &= \frac{\partial X^2}{\partial \eta_2} = -\frac{2}{\sigma^2}(y_2-\eta_2) + 2\lambda, \\ 0 &= \frac{\partial X^2}{\partial \eta_3} = -\frac{2}{\sigma^2}(y_3-\eta_3) + 2\lambda, \\ 0 &= \frac{\partial X^2}{\partial \lambda} = 2\left(\sum_{i=1}^{3}\eta_i - 180^o\right). \end{aligned}\right\} \qquad (10.100)$$

Multiplying the first three equations by -1 and the last by $\frac{1}{\sigma^2}$, we find by adding all expressions an equation for λ,

$$0 = \frac{2}{\sigma^2}\left(\sum_{i=1}^{3} y_i - 180^o\right) - 6\lambda,$$

which leads to

$$\hat{\lambda} = \frac{1}{3}\frac{1}{\sigma^2}\left(\sum_{i=1}^{3} y_i - 180^o\right).$$

The estimates for the angles, obtained from the three first equations, are

$$\hat{\eta}_j = y_j - \sigma^2\cdot\hat{\lambda} = y_j - \frac{1}{3}\left(\sum_{i=1}^{3} y_i - 180^o\right).$$

These relations, symmetric in all indices from 1 to 3, show in a transparent way how the original measurements are corrected by subtracting equal amounts of the total measured "excess" of 2^o from each of the measurements.

Exercise 10.14: Show that the covariance matrix for the estimates $\hat{\underline{\eta}}$ can be expressed as

$$V(\hat{\underline{\eta}}) = V(\underline{y}) - \frac{1}{3}\sigma^2\begin{Bmatrix}1\\1\\1\end{Bmatrix}(1,1,1) = \sigma^2\begin{bmatrix} 2/3 & -1/3 & -1/3 \\ -1/3 & 2/3 & -1/3 \\ -1/3 & -1/3 & 2/3 \end{bmatrix},$$

showing that the "improved measurements" become correlated but have smaller errors (diagonal terms) than the original measurements. Compare eq. (10.112) of the next section.

10.7.2 Linear LS model with linear constraints; Lagrangian multipliers

We will consider a general linear LS estimation problem with linear constraint equations, in the form

$$\left.\begin{aligned} X^2(\underline{\theta}) &= (\underline{y}-A\underline{\theta})^T V^{-1}(\underline{y}-A\underline{\theta}) = \text{minimum}, \\ B\underline{\theta} - \underline{b} &= \underline{0}. \end{aligned}\right\} \qquad (10.101)$$

Here the L parameters $\underline{\theta}$ are restricted through the K constraint equations expressed by the matrix B of dimension K x L, and the K-component vector $\underline{b}$. The other symbols are the same as used before for the unconstrained linear case.

We introduce a K-component vector $\underline{\lambda} = \{\lambda_1, \lambda_2, \ldots, \lambda_K\}$ of Lagrangian multipliers and reformulate the problem (10.101) to an unconstrained linear minimization for the L + K unknowns $\underline{\theta}$ and $\underline{\lambda}$,

$$X^2(\underline{\theta},\underline{\lambda}) = (\underline{y}-A\underline{\theta})^T V^{-1}(\underline{y}-A\underline{\theta}) + 2\underline{\lambda}^T(B\underline{\theta}-\underline{b}) = \text{minimum}. \qquad (10.102)$$

If we here equate to zero the derivatives of X^2 with respect to θ_ℓ, $\ell=1,2,\ldots,L$ and λ_k, $k=1,2,\ldots,K$ we get the normal equations, which in vector notation can be written as

$$\nabla_\theta X^2 = -2(A^T V^{-1}\underline{y} - A^T V^{-1} A\underline{\theta}) + 2B^T\underline{\lambda} = \underline{0}, \qquad (10.103)$$

$$\nabla_\lambda X^2 = 2(B\underline{\theta}-\underline{b}) = \underline{0}. \qquad (10.104)$$

These are L + K linear equations for the unknowns. Obviously, eqs.(10.104) are the constraint equations regained, whereas eqs.(10.103) are the analogues of eqs. (10.21) for the unconstrained case, now modified by the $\underline{\lambda}$-term due to the constraints. It will be seen that eqs.(10.100) of the previous section represent a special case of the formulae above.

Let us introduce the abbreviations

$$C \equiv A^T V^{-1} A, \qquad \underline{c} \equiv A^T V^{-1}\underline{y}, \qquad (10.105)$$

which bring the simultaneous linear equations to the form

$$\left.\begin{aligned} C\underline{\theta} + B^T\underline{\lambda} &= \underline{c}, \\ B\underline{\theta} &= \underline{b}. \end{aligned}\right\} \qquad (10.106)$$

If the inverse of C exists we can premultiply the first of these equations by BC^{-1} and substitute for $B\underline{\theta}$ from the last equation. This gives an equation involving the unknowns $\underline{\lambda}$ only,

$$\underline{b} + BC^{-1}B^T\underline{\lambda} = BC^{-1}\underline{c}.$$

Writing

$$V_B \equiv BC^{-1}B^T \qquad (10.107)$$

and assuming that this symmetric matrix has an inverse, the solution for the Lagrangian multipliers becomes

$$\hat{\underline{\lambda}} = V_B^{-1}(BC^{-1}\underline{c}-\underline{b}). \qquad (10.108)$$

When the Lagrangian multipliers are substituted back in eq.(10.106) we obtain the solution for the parameters $\underline{\theta}$,

$$\hat{\underline{\theta}} = C^{-1}\underline{c} - C^{-1}B^T V_B^{-1}(BC^{-1}\underline{c}-\underline{b}). \qquad (10.109)$$

Equations (10.108) and (10.109) provide an exact solution, since all matrices and vectors are known quantities. It is interesting to observe that the combination $C^{-1}\underline{c}$, which is nothing but the solution of the unconstrained minimization, enters in $\hat{\underline{\lambda}}$ as well as in $\hat{\underline{\theta}}$. In fact, the parenthesis $(BC^{-1}\underline{c}-\underline{b})$ measures how much the observations $\underline{y}$ violate the constraint equations; compare the example of the previous section. As for $\hat{\underline{\theta}}$ we see that the effect of the constraint equations has been to correct the solution $C^{-1}\underline{c}$ of the unconstrained minimization by an amount proportional to the "violation" term $(BC^{-1}\underline{c}-\underline{b})$.

We note further that the Lagrangian multipliers $\hat{\underline{\lambda}}$ as well as the parameters $\hat{\underline{\theta}}$ have a linear dependence on the observations $\underline{y}$ through the vector $\underline{c}$. It is seen by application of the expectation operator that

$$E(C^{-1}\underline{c}) = C^{-1}E(\underline{c}) = C^{-1}A^T V^{-1}E(\underline{y}) = C^{-1}A^T V^{-1}A\underline{\theta} = \underline{\theta}.$$

Hence the expectation of the term $(BC^{-1}\underline{c}-\underline{b})$ vanishes in virtue of the constraint equations, giving

$$\left.\begin{aligned} E(\hat{\underline{\lambda}}) &= \underline{0}\,, \\ E(\hat{\underline{\theta}}) &= \underline{\theta}\,. \end{aligned}\right\} \qquad (10.110)$$

The last result shows that the parameter estimates are *unbiassed*, as they were in the unconstrained case.

The linear dependence on $\underline{y}$ permits us to establish the covariance matrix for the estimates $\hat{\underline{\theta}}$, by applying the usual law of error propagation. From eq.(10.109), recalling that $\underline{c} = A^T V^{-1} \underline{y}$, we have

$$V(\hat{\underline{\theta}}) = \left[C^{-1}A^TV^{-1} - C^{-1}B^TV_B^{-1}BC^{-1}A^TV^{-1}\right]V\left[C^{-1}A^TV^{-1} - C^{-1}B^TV_B^{-1}BC^{-1}A^TV^{-1}\right]^T.$$

Using the definitions of C and V_B together with the fact that these matrices, and hence also their inverse matrices C^{-1} and V_B^{-1} are symmetric, this expression can be simplified to

$$V(\hat{\underline{\theta}}) = C^{-1}(I_N - B^TV_B^{-1}BC^{-1}),$$

or

$$V(\hat{\underline{\theta}}) = C^{-1} - (BC^{-1})^TV_B^{-1}(BC^{-1}). \qquad (10.111)$$

In eq.(10.111) C^{-1} is the covariance matrix for the unconstrained parameters, and as the diagonal elements of the term $(BC^{-1})^TV_B^{-1}(BC^{-1})$ are always non-negative, we see that the constraint equations will lead to a *reduction of the parameter errors* (*i.e.* the diagonal terms) compared to the unconstrained case. For the off-diagonal terms no similar statement can be made in the general case, as the covariances between different parameters can be smaller or larger than in the unconstrained case.

The "improved measurements" $\hat{\underline{\eta}} = A\hat{\underline{\theta}}$ and their errors are given by the formulae

$$\left.\begin{aligned} \hat{\underline{\eta}} &= A\left[C^{-1}\underline{c} - C^{-1}B^TV_B^{-1}(BC^{-1}\underline{c}-\underline{b})\right], \\ V(\hat{\underline{\eta}}) &= AC^{-1}A^T - A(BC^{-1})^TV_B^{-1}(BC^{-1})A^T, \end{aligned}\right\} \qquad (10.112)$$

which should be compared to the corresponding expressions for the unconstrained problem, given in Exercise 10.8.

Exercise 10.15: Show that the estimates for the Lagrangian multipliers $\hat{\lambda}$ of eq.(10.108) correspond to a covariance matrix $V(\hat{\underline{\lambda}}) = V_B^{-1}$, and that the estimated parameters and Lagrangian multipliers are uncorrelated, $\mathrm{cov}(\hat{\underline{\theta}},\hat{\underline{\lambda}}) = \underline{0}$.

Exercise 10.16: (i) With the notation of Sect.10.7.2, show that the residuals for the linear LS fit with linear constraints can be written as $\underline{\varepsilon} = D\underline{\varepsilon}$ where $\underline{\varepsilon}$ is the vector of errors due to measurement, and

$$D \equiv I_N - AC^{-1}A^TV^{-1} + AC^{-1}B^TV_B^{-1}BC^{-1}A^TV^{-1}.$$

This matrix D differs from the corresponding expression for the unconstrained case (see Exercise 10.8) only through the last term.

(ii) If the measurements are uncorrelated and have equal errors, *i.e.* $V(\underline{y}) = \sigma^2 I_N$, D simplifies to

$$D \equiv I_N - A(A^TA)^{-1}A^T + A(A^TA)^{-1}B^T\left[B(A^TA)^{-1}B^T\right]^{-1}B(A^TA)^{-1}A^T,$$

which represents the extension of eq.(10.59) valid for the unconstrained case. Show that this D is also an idempotent matrix satisfying $D^T=D$, $D^TD=D$. Follow the reasoning of Sect.10.4.2 and verify that $E(Q^2_{min})=\sigma^2(N-L+K)$, showing that $Q^2_{min}/(N-L+K)$ is an unbiassed estimator of σ^2.

(iii) Generalize the proof that $Q^2_{min}/(N-L+K)$ is an unbiassed estimator of σ^2 to the case when the measurements have correlation terms and not necessarily equal errors. (Hint: See Exercise 10.11.)

10.8 GENERAL LEAST-SQUARES ESTIMATION WITH CONSTRAINTS

We have in the preceding sections discussed how the LS method can be used to estimate unknown parameters in various problems of increasing complexity. We will now turn to the most general situation, where the estimation problem involves observable quantities as well as unobservable unknowns, which are connected through a set of general, *i.e.* non-linear, algebraic restrictions.

We shall in the following section develop the formulation of the iterative procedure using the method of Lagrangian multipliers, without making any reference to a special physical problem. The reader may find it useful to have in mind the kinematic analysis of a particle reaction (an example follows in Sect.10.8.2), where the momentum and energy conservation laws constitute a set of restrictions relating the various momenta and angles for the particle combination defining the kinematic hypothesis. Some of the quantities have been measured to a certain accuracy (say, the momenta and angles of curved tracks in a bubble chamber), and some are completely unknown (the variables for an unseen

particle, momenta for short, straight tracks, *etc.*). The purpose of the LS estimation is to investigate the kinematic hypothesis; for a successful minimization the constraint equations will supply estimates for the unmeasured variables as well as "improved measurements" for the measured quantities.

10.8.1 The iteration procedure

We will, as before, let $\underline{\eta}$ be a vector of N observables, for which we have the first approximation values (measurements) $\underline{y}$, with errors contained in the covariance matrix $V(\underline{y})$. In addition we have a set of J unmeasurable variables $\underline{\xi} = \{\xi_1,\xi_2,\ldots,\xi_J\}$. The N measurable and the J unmeasurable variables are related and have to satisfy a set of K constraint equations,

$$f_k(\eta_1,\eta_2,\ldots,\eta_N,\xi_1,\xi_2,\ldots,\xi_J) = 0, \qquad k=1,2,\ldots,K.$$

According to the Least-Squares Principle we should adopt as our best estimates of the unknowns $\underline{\eta}$ and $\underline{\xi}$ those values for which

$$\left.\begin{aligned} &X^2(\underline{\eta}) = (\underline{y}-\underline{\eta})^T V^{-1}(\underline{y})(\underline{y}-\underline{\eta}) = \text{minimum}, \\ &\underline{f}(\underline{\eta},\underline{\xi}) = \underline{0}. \end{aligned}\right\} \qquad (10.113)$$

The general, constrained LS problem of eqs.(10.113) can be solved by eliminating K unknowns from the constraint equations, substituting in X^2 and minimizing this function with respect to the N+J-K remaining variables. The elimination method, however, has the disadvantage that it does not give any prescription on which variables one should eliminate from the constraint equations. If these are non-linear, the actual minimization of the function X^2 may develop quite differently, depending on the elimination made. The Lagrange multiplier method, on the other hand, avoids the preference of any of the unknown variables and treats them all on an equal footing. Accordingly, although this approach implies more variables in the minimization, its feature of symmetry in the variables is considered a greater virtue, and is preferred in practice.

We proceed therefore to solve the problem of eqs.(10.113) by the method of the Lagrangian multipliers. We introduce K additional unknowns $\underline{\lambda} = \{\lambda_1,..,\lambda_K\}$

and rephrase the problem by requiring

$$X^2(\underline{\eta},\underline{\xi},\underline{\lambda}) = (\underline{y}-\underline{\eta})^T V^{-1}(\underline{y})(\underline{y}-\underline{\eta}) + 2\underline{\lambda}^T\underline{f}(\underline{\eta},\underline{\xi}) = \text{minimum}. \qquad (10.114)$$

We now have a total of N+J+K unknowns. When the derivatives of X^2 with respect to all unknowns are put equal to zero we get the following set of equations, written in vector form,

$$\left.\begin{aligned} &\nabla_\eta X^2 = -2V^{-1}(\underline{y}-\underline{\eta}) + 2F_\eta^T\underline{\lambda} = \underline{0}, && \text{(N equations)} \\ &\nabla_\xi X^2 = 2F_\xi^T\underline{\lambda} = \underline{0}, && \text{(J equations)} \\ &\nabla_\lambda X^2 = 2\,\underline{f}(\underline{\eta},\underline{\xi}) = \underline{0}, && \text{(K equations)} \end{aligned}\right\} \qquad (10.115)$$

where the matrices F_η (of dimension KxN) and F_ξ (dimension KxJ) are defined by

$$(F_\eta)_{ki} \equiv \frac{\partial f_k}{\partial \eta_i}, \qquad (F_\xi)_{kj} \equiv \frac{\partial f_k}{\partial \xi_j}. \qquad (10.116)$$

Thus, removing the nuicance factors 2, the equations are

$$V^{-1}(\underline{\eta}-\underline{y}) + F_\eta^T\underline{\lambda} = \underline{0}, \qquad (10.117)$$

$$F_\xi^T\underline{\lambda} = \underline{0}, \qquad (10.118)$$

$$\underline{f}(\underline{\eta},\underline{\xi}) = \underline{0}. \qquad (10.119)$$

The solution of the set of equations (10.117) - (10.119) for the N+J+K unknowns must in the general case*) be found by iterations, producing successively better approximations.

Let us suppose that iteration number ν has been performed and that it is necessary to find a still better solution. For the ν-th iteration the approximative solution is given by the values $\underline{\eta}^\nu,\underline{\xi}^\nu,\underline{\lambda}^\nu$, corresponding to the function value $(X^2)^\nu$. We perform a Taylor expansion of the constraint equations (10.119) in the point $(\underline{\eta}^\nu,\underline{\xi}^\nu)$,

*) The set (10.117)-(10.119) reduces of course to eqs.(10.106) for a linear problem with no unmeasurable variables.

$$f_k^\nu + \sum_{i=1}^{N}\left(\frac{\partial f_k}{\partial \eta_i}\right)^\nu (\eta_i^{\nu+1} - \eta_i^\nu) + \sum_{j=1}^{J}\left(\frac{\partial f_k}{\partial \xi_j}\right)^\nu (\xi_j^{\nu+1} - \xi_j^\nu) + \ldots = 0,\ k=1,2,\ldots,K.$$

When the terms of second and higher orders are neglected this can be written

$$\underline{f}^\nu + F_\eta^\nu(\underline{\eta}^{\nu+1} - \underline{\eta}^\nu) + F_\xi^\nu(\underline{\xi}^{\nu+1} - \underline{\xi}^\nu) = \underline{0}, \tag{10.120}$$

where all superscripts ν indicate that $\underline{f}^\nu, F_\eta^\nu, F_\xi^\nu$ are to be evaluated at the point $(\underline{\eta}^\nu, \underline{\xi}^\nu)$, the ν-th iteration. Equations (10.117) and (10.118) now read

$$V^{-1}(\underline{\eta}^{\nu+1} - \underline{y}) + (F_\eta^T)^\nu \underline{\lambda}^{\nu+1} = \underline{0}, \tag{10.121}$$

$$(F_\xi^T)^\nu \underline{\lambda}^{\nu+1} = \underline{0}. \tag{10.122}$$

These equations, together with the expanded constraint equations (10.120) will make it possible to express all unknowns of the $(\nu+1)$-th iteration by the quantities of the preceding iteration.

If we eliminate $\underline{\eta}^{\nu+1}$ from (10.121) and substitute in (10.120) we get a relation involving only $\underline{\lambda}^{\nu+1}$ and $\underline{\xi}^{\nu+1}$,

$$\underline{f}^\nu + F_\eta^\nu\left[\underline{y} - V(F_\eta^T)^\nu \underline{\lambda}^{\nu+1}) - \underline{\eta}^\nu\right] + F_\xi^\nu(\underline{\xi}^{\nu+1} - \underline{\xi}^\nu) = \underline{0},$$

or,

$$\underline{r} + F_\xi^\nu(\underline{\xi}^{\nu+1} - \underline{\xi}^\nu) = S\underline{\lambda}^{\nu+1}, \tag{10.123}$$

when we introduce the notation

$$\underline{r} \equiv \underline{f}^\nu + F_\eta^\nu(\underline{y} - \underline{\eta}^\nu), \tag{10.124}$$

$$S \equiv F_\eta^\nu V (F_\eta^T)^\nu. \tag{10.125}$$

Clearly, S is a symmetric matrix of dimension K × K. Multiplying eq.(10.123) from the left by S^{-1} gives an expression for $\underline{\lambda}^{\nu+1}$ that can be substituted in eq.(10.122), which in turn leads to an equation with only $\underline{\xi}^{\nu+1}$ unknown,

$$(F_\xi^T)^\nu S^{-1}\left[\underline{r} + F_\xi^\nu(\underline{\xi}^{\nu+1} - \underline{\xi}^\nu)\right] = \underline{0}.$$

Solving this equation for $\underline{\xi}^{\nu+1}$ and substituting back into eq.(10.123) we find $\underline{\lambda}^{\nu+1}$, and finally eq.(10.121) will give $\underline{\eta}^{\nu+1}$. Thus we find in succession

$$\underline{\xi}^{\nu+1} = \underline{\xi}^\nu - (F_\xi^T S^{-1} F_\xi)^{-1} F_\xi^T S^{-1} \underline{r}, \tag{10.126}$$

$$\underline{\lambda}^{\nu+1} = S^{-1}\left[\underline{r} + F_\xi(\underline{\xi}^{\nu+1} - \underline{\xi}^\nu)\right], \tag{10.127}$$

$$\underline{\eta}^{\nu+1} = \underline{y} - VF_\eta^T \underline{\lambda}^{\nu+1}. \tag{10.128}$$

The linearized equations are therefore solved in such a way that the "completely unknown" $\underline{\xi}^{\nu+1}$ are found first, next the Lagrangian multipliers $\underline{\lambda}^{\nu+1}$ and finally the "improved measurements" $\underline{\eta}^{\nu+1}$.

In eqs.(10.126)-(10.128) the matrices F_η, F_ξ, S and the vector $\underline{r}$ are evaluated at the point $(\eta^\nu, \underline{\xi}^\nu)$. We note that in deriving the formulae it has been tacitly assumed that the inverse of the matrices S and $(F_\xi^T S^{-1} F_\xi)$ exists.

With the new values for $\underline{\eta}^{\nu+1}, \underline{\xi}^{\nu+1}$ and $\underline{\lambda}^{\nu+1}$ we calculate the value of the function $(X^2)^{\nu+1}$ for the $(\nu+1)$-th iteration and compare it to the previous value $(X^2)^\nu$. With an improved solution the new point $(\underline{\eta}^{\nu+1}, \underline{\xi}^{\nu+1})$ is used for a new Taylor expansion of the constraint equations and the process is started over again. The iterations should be continued until a satisfactory solution has been found. Usually one would repeat the calculations until the change in X^2 between successive steps becomes small. One may also have to check that the differences $\Delta\underline{\eta}$, $\Delta\underline{\xi}$ converge properly. General convergence criteria can hardly be given, but must be decided for the separate problems. It is generally valid, that in order to optimize the convergence of an iteration procedure one should be careful in giving good starting values $\underline{\eta}^o$, $\underline{\xi}^o$ for the iterations, since these determine how many steps will be necessary to reach the desired minimum.

The distinction between the two types of variables $\underline{\eta}$ and $\underline{\xi}$ lies in fact in the choice of starting values. For the measurable variables $\underline{\eta}$ one should start with $\underline{\eta}^o = \underline{y}$, the measured values; for the unmeasurable variables $\underline{\xi}$ the starting values $\underline{\xi}^o$ should be evaluated from the most convenient constraint equations inserting the measurements $\underline{\eta}^o$ for $\underline{\eta}$.

To summarize, an iteration will consist of the following items:

(i) Evaluate the vector $\underline{r}$ from eq.(10.124) and the matrix S from eq.(10.125).

(ii) Find the new vector $\underline{\xi}^{\nu+1}$ of unmeasurable variables from eq.(10.126).

(iii) Find the new vector $\underline{\lambda}^{\nu+1}$ of Lagrangian multipliers from eq.(10.127).

(iv) Find the new vector $\underline{\eta}^{\nu+1}$ of measurable quantities from eq.(10.128).

(v) Calculate the new value $(X^2)^{\nu+1}$.

(vi) Compare results with previous iteration.
Proceed to (i) if new iteration is required.
Stop, if satisfactory solution has been obtained.

When the final step has been made the covariances of the estimates $\hat{\underline{\eta}}$ and $\hat{\underline{\xi}}$ should be found; see Sect.10.8.3.

Exercise 10.17: Show that the value of X^2 for the $(\nu+1)$-th step is

$$(X^2)^{\nu+1} = (\underline{\lambda}^{\nu+1})^T S \underline{\lambda}^{\nu+1} + 2(\underline{\lambda}^{\nu+1})^T \underline{f}^{\nu+1},$$

where the matrix S is evaluated for the ν-th iteration.

10.8.2 Example: Kinematic analysis of a V^0 event (1)

Let us apply the previous formulation to the kinematic analysis of a V^0 as seen in a bubble chamber. We suppose that the two tracks have been measured and that the track reconstruction has provided for each track a first approximation for the kinematic variables 1/P (inverse momentum), λ (dip angle), and ϕ (azimuth angle) as well as a covariance matrix for these variables. We want to analyze this event for the kinematic hypothesis

$$\Lambda \to p + \pi^- .$$

If the origin of the Λ is unspecified the magnitude of the momentum as well as the direction of the decaying particle are unknowns. The problem therefore involves three completely unknown variables,

$$\underline{\xi} = \{P_\Lambda, \lambda_\Lambda, \phi_\Lambda\},$$

and six measurable unknowns,

$$\underline{\eta} = \{P_p, \lambda_p, \phi_p, P_\pi, \lambda_\pi, \phi_\pi\}.$$

The algebraic constraints are the four equations describing momentum and energy conservation, which are obtained by equating the following expressions to zero,

$$f_1 = -P_\Lambda \cos\lambda_\Lambda \cos\phi_\Lambda + P_p \cos\lambda_p \cos\phi_p + P_\pi \cos\lambda_\pi \cos\phi_\pi ,$$

$$f_2 = -P_\Lambda \cos\lambda_\Lambda \sin\phi_\Lambda + P_p \cos\lambda_p \sin\phi_p + P_\pi \cos\lambda_\pi \sin\phi_\pi ,$$

$$f_3 = -P_\Lambda \sin\lambda_\Lambda + P_p \sin\lambda_p + P_\pi \sin\lambda_\pi ,$$

$$f_4 = -\sqrt{P_\Lambda^2 + m_\Lambda^2} + \sqrt{P_p^2 + m_p^2} + \sqrt{P_\pi^2 + m_\pi^2} .$$

Since the problem involves 4 constraints and 3 unmeasured unknowns we are therefore dealing with a 1C-fit.

From the definitions of eqs.(10.116) we see that the matrices F_η (of dimension 4×6) and F_ξ (dimension 4×3) as obtained from the derivatives of the four constraint functions f_k with respect to the measurable and unmeasurable variables, are, respectively,

$$F_\eta = \begin{bmatrix} \cos\lambda_p\cos\phi_p & -P_p\sin\lambda_p\cos\phi_p & -P_p\cos\lambda_p\sin\phi_p & \cos\lambda_\pi\cos\phi_\pi & -P_\pi\sin\lambda_\pi\cos\phi_\pi & -P_\pi\cos\lambda_\pi\sin\phi_\pi \\ \cos\lambda_p\sin\phi_p & -P_p\sin\lambda_p\sin\phi_p & +P_p\cos\lambda_p\cos\phi_p & \cos\lambda_\pi\sin\phi_\pi & -P_\pi\sin\lambda_\pi\sin\phi_\pi & +P_\pi\cos\lambda_\pi\cos\phi_\pi \\ \sin\lambda_p & P_p\cos\lambda_p & 0 & \sin\lambda_\pi & P_\pi\cos\lambda_\pi & 0 \\ \dfrac{P_p}{\sqrt{P_p^2+m_p^2}} & 0 & 0 & \dfrac{P_\pi}{\sqrt{P_\pi^2+m_\pi^2}} & 0 & 0 \end{bmatrix}$$

and

$$F_\xi = \begin{bmatrix} -\cos\lambda_\Lambda\cos\phi_\Lambda & P_\Lambda\sin\lambda_\Lambda\cos\phi_\Lambda & P_\Lambda\cos\lambda_\Lambda\sin\phi_\Lambda \\ -\cos\lambda_\Lambda\sin\phi_\Lambda & P_\Lambda\sin\lambda_\Lambda\sin\phi_\Lambda & -P_\Lambda\cos\lambda_\Lambda\cos\phi_\Lambda \\ -\sin\lambda_\Lambda & -P_\Lambda\cos\lambda_\Lambda & 0 \\ \dfrac{-P_\Lambda}{\sqrt{P_\Lambda^2+m_\Lambda^2}} & 0 & 0 \end{bmatrix}.$$

To start the iterations we take the measurements as the initial $\underline{\eta}^0$,

$$\underline{\eta}^0 = \{P_p^0, \lambda_p^0, \phi_p^0, P_\pi^0, \lambda_\pi^0, \phi_\pi^0\}.$$

For $\underline{\xi}^0$ we would take, for example, the value $\underline{\xi}^0 = \{P_\Lambda^0, \lambda_\Lambda^0, \phi_\Lambda^0\}$, where the com-

ponents are obtained by demanding the first three constraint functions equal to zero, *i.e.* momentum conservation satisfied. The fourth constraint function will then in general not be strictly zero. Thus we will have an initial value of the vector $\underline{r}$ from eq.(10.124),

$$\underline{r} = \underline{f}^{o} = \{0,\ 0,\ 0,\ f_4^o\}.$$

Inserting the approximations $(\underline{\eta}^o, \underline{\xi}^o)$ we can find F_η^o, F_ξ^o and obtain the 4 × 4 matrix S from eq.(10.125),

$$S = F_\eta^o V (F_\eta^T)^o.$$

Inverting this matrix we can next find the values of $\underline{\xi}^1$, $\underline{\lambda}^1$, $\underline{\eta}^1$ in succession from eqs.(10.126)~(10.127), calculate X^2 with these first estimates, and continue the process.

If we had measured the coordinates of the origin of the Λ in addition to its decay point the line-of-flight of this particle would have been known, equivalent to including λ_Λ and ϕ_Λ among the measurable unknowns $\underline{\eta}$. The only completely unknown variable would then be P_Λ, the magnitude of the momentum, corresponding to a 3C-fit in this case. See also Sect.14.4.5.

10.8.3 Calculation of errors

The errors in the final estimates of the measurable and unmeasurable variables from the general LS fit of Sect.10.8.1 are found by applying the law of error propagation. Let us consider the estimates $\hat{\underline{\eta}} = \underline{\eta}^{\nu+1}$ and $\hat{\underline{\xi}} = \underline{\xi}^{\nu+1}$ as functions of the measurements $\underline{y}$,

$$\left.\begin{aligned} \hat{\underline{\eta}} &= \underline{g}(\underline{y}), \\ \hat{\underline{\xi}} &= \underline{h}(\underline{y}), \end{aligned}\right\} \tag{10.129}$$

where the actual forms of the functions $\underline{g}$ and $\underline{h}$ are found from eqs.(10.126)-(10.128), using eq.(10.124) for $\underline{r}$,

$$\left.\begin{aligned} \underline{g} &= \underline{y} - V F_\eta^T S^{-1}\left[I_K - F_\xi (F_\xi^T S^{-1} F_\xi)^{-1} F_\xi^T S^{-1}\right]\left[\underline{f} + F_\eta(\underline{y}-\underline{\eta})\right], \\ \underline{h} &= \underline{\xi} - (F_\xi^T S^{-1} F_\xi)^{-1} F_\xi^T S^{-1}\left[\underline{f} + F_\eta(\underline{y}-\underline{\eta})\right]. \end{aligned}\right\} \tag{10.130}$$

In these formulae $\underline{\eta}$ and $\underline{\xi}$, as well as the function $\underline{f}$ and all matrices F and S are evaluated in the last, *i.e.* the ν-th iteration. To the approximation of a linear dependence on $\underline{y}$ the covariance matrices for $\hat{\underline{\eta}}$ and $\hat{\underline{\xi}}$ are given by eq.(3.80), the law of propagation of errors, as

$$\left.\begin{aligned} V(\hat{\underline{\eta}}) &= \left(\frac{d\underline{g}}{d\underline{y}}\right) V(\underline{y}) \left(\frac{d\underline{g}}{d\underline{y}}\right)^T, \\ V(\hat{\underline{\xi}}) &= \left(\frac{d\underline{h}}{d\underline{y}}\right) V(\underline{y}) \left(\frac{d\underline{h}}{d\underline{y}}\right)^T, \\ \mathrm{cov}(\hat{\underline{\eta}}, \hat{\underline{\xi}}) &= \left(\frac{d\underline{g}}{d\underline{y}}\right) V(\underline{y}) \left(\frac{d\underline{h}}{d\underline{y}}\right)^T, \end{aligned}\right\} \tag{10.131}$$

where the derivatives of $\underline{g}$ and $\underline{h}$ with respect to $\underline{y}$ denote matrices of dimensions N × N and J × N, respectively. From eqs.(10.130) we can express these quantities as

$$\left.\begin{aligned} \frac{d\underline{g}}{d\underline{y}} &\equiv I_N - V(\underline{y})\left[F_\eta^T S^{-1} F_\eta - F_\eta^T S^{-1} F_\xi (F_\xi^T S^{-1} F_\xi)^{-1} F_\xi^T S^{-1} F_\eta\right], \\ \frac{d\underline{h}}{d\underline{y}} &\equiv -(F_\xi^T S^{-1} F_\xi)^{-1} F_\xi^T S^{-1} F_\eta. \end{aligned}\right\} \tag{10.132}$$

In these expressions we observe that the matrices F_η, F_ξ enter only *via* the three *possible combinations of the type* $F^T S^{-1} F$. One of these, $F_\xi^T S^{-1} F_\xi$, appeared already as a part of the solution for the unmeasurable variables $\hat{\underline{\xi}}$, eq.(10.126). With the abbreviations

$$\left.\begin{aligned} G &\equiv F_\eta^T S^{-1} F_\eta, \qquad H \equiv F_\eta^T S^{-1} F_\xi, \\ U^{-1} &\equiv F_\xi^T S^{-1} F_\xi, \end{aligned}\right\} \tag{10.133}$$

we find after a little algebra that eqs.(10.131) lead to

$$\left.\begin{aligned} V(\hat{\underline{\eta}}) &= V(\underline{y})\left[I_N - (G - HUH^T) V(\underline{y})\right], \\ V(\hat{\underline{\xi}}) &= U, \\ \mathrm{cov}(\hat{\underline{\eta}}, \hat{\underline{\xi}}) &= -V(\underline{y}) H U. \end{aligned}\right\} \tag{10.134}$$

These error formulae imply that the errors on the fitted quantities $\hat{\underline{\eta}}$ in general are smaller than the errors in the observations $\underline{y}$, and that the fitted quantities will be correlated, even if the measurements were independent.

Exercise 10.18: Show that the covariance matrix for the residuals $\hat{\underline{\varepsilon}} = \underline{y} - \hat{\underline{\eta}}$, to the approximation of a linear relationship between $\hat{\underline{\eta}}$ and $\underline{y}$, is

$$V(\hat{\underline{\varepsilon}}) \equiv V(\underline{y}) + V(\hat{\underline{\eta}}) - 2\mathrm{cov}(\underline{y},\hat{\underline{\eta}}) \approx V(\underline{y}) - V(\hat{\underline{\eta}}) = V(\underline{y})(G-HUH^T)V(\underline{y}).$$

10.9 CONFIDENCE INTERVALS AND ERRORS FROM THE X^2 FUNCTION

10.9.1 Basis for the determination of LS confidence intervals

With a theoretical model which is linear in the parameters $\underline{\theta}$ we have the general expression for the X^2 function

$$X^2(\underline{\theta}) = (\underline{y}-A\underline{\theta})^T V^{-1}(\underline{y}-A\underline{\theta}), \tag{10.20}$$

where V^{-1} is the inverse of the covariance matrix $V(\underline{y})$ for the N observed quantities y. When $V(\underline{y})$ is independent of the parameters we found (in Sect.10.2.3) that the minimization of $X^2(\underline{\theta})$ with respect to $\underline{\theta}$ led to the LS estimate

$$\hat{\underline{\theta}} = (A^TV^{-1}A)^{-1}A^TV^{-1}\underline{y}, \tag{10.23}$$

and, from the law of error propagation, the covariances of these estimates were found to be

$$V(\hat{\underline{\theta}}) = (A^TV^{-1}A)^{-1}. \tag{10.24}$$

Simple algebra then leads to the following relation (see Exercise 10.12):

$$X^2(\underline{\theta}) = X^2_{min} + (\underline{\theta}-\hat{\underline{\theta}})^TV^{-1}(\hat{\underline{\theta}})(\underline{\theta}-\hat{\underline{\theta}}). \tag{10.136}$$

Equation (10.136) provides the basis for the determination of confidence intervals (regions) with the LS method. When the observations y are *multinormally* distributed about the true values $\underline{\eta} = A\underline{\theta}$, the estimates $\hat{\underline{\theta}}$, which are linearly related to $\underline{y}$, will also be normally distributed (compare Sect.4.8.5). Under these conditions, each of the three terms in eq.(10.136) will be chi-square distributed. For example, with N independent observations and L unconstrained parameters, $X^2(\underline{\theta})$ is $\chi^2(N)$, X^2_{min} is $\chi^2(N-L)$, and the quadratic (covariance) form $(\underline{\theta}-\hat{\underline{\theta}})^TV^{-1}(\hat{\underline{\theta}})(\underline{\theta}-\hat{\underline{\theta}})$ is $\chi^2(L)$. In the case of constrained parameters the last term will be distributed as $\chi^2(L-K)$, where K is the number of linear constraints. In general, the confidence regions we are seeking will be found by intersecting the

$X^2(\underline{\theta})$ surface by planes

$$X^2(\underline{\theta}) = X^2_{min} + a, \tag{10.137}$$

where the constant a is appropriately chosen. The corresponding probabilities are determined from the chi-square distribution with a number of degrees of freedom equal to the number of independent parameters.

When $X^2(\underline{\theta})$ is not a quadratic function in the parameters, for example if the model is not a linear function of the parameters and/or the covariance matrix is not independent of $\underline{\theta}$, it is still customary to use the intersection approach (the "graphical method") to establish confidence regions for the parameters. However, since the exact distribution of the quantities is then not known, the associated probabilities as deduced from a comparison with the chi-square distribution will only be approximately correct in these cases.

10.9.2 LS errors and confidence intervals, the one-parameter case

We assume that the LS estimation problem involves a single parameter θ. When the X^2 function is Taylor expanded around the extremum (minimum) point $\theta = \hat{\theta}$ where the first derivative vanishes we have in general

$$X^2(\theta) = X^2_{min} + \tfrac{1}{2}\left.\frac{d^2X^2}{d\theta^2}\right|_{\theta=\hat{\theta}}(\theta-\hat{\theta})^2 + \ldots \tag{10.138}$$

Here the second derivative of X^2 must be positive to ensure that the extremum is a minimum value. Clearly, to the extent that the higher-order terms can be neglected the second derivative specifies the "width" of $X^2(\theta)$ around the minimum point $\hat{\theta}$, that is, it determines how "precise" the minimum has been located. Thus the error in the LS estimate $\hat{\theta}$ must in general be related to the second derivative of the function X^2 evaluated for $\theta = \hat{\theta}$.

For a *linear* LS problem fulfilling the conditions specified in Sect. 10.2.3, with a *constant* covariance matrix $V(\underline{y})$ for the observations, the function X^2 is strictly of second order in the parameter θ. The second derivative of X^2 with respect to θ is then a constant and the series expansion (10.138) terminates with the second term, so that we have the parabolic dependence

$$X^2(\theta) = X^2_{min} + \tfrac{1}{2}\frac{d^2X^2}{d\theta^2}(\theta-\hat{\theta})^2. \tag{10.139}$$

Since, from eq.(10.136) we also have

$$X^2(\theta) = X^2_{min} + \frac{1}{V(\hat{\theta})}(\theta-\hat{\theta})^2, \tag{10.140}$$

it is seen from the two last equations by identification of the coefficients of the quadratic terms that

$$V(\hat{\theta}) = 2\left(\frac{d^2X^2}{d\theta^2}\right)^{-1}. \tag{10.141}$$

This expression for the variance of the LS estimate $\hat{\theta}$ can be compared to the similar expression obtained for the large-sample ML estimate, eq.(9.36). It should be emphasized that the formulae (10.139) - (10.141) are *exact* under the conditions specified above.

In the case of a *non-linear* LS problem, or in general, with an X^2 function which is not of strictly parabolic form, one can still expect to find the variance - and hence the error - of the estimate $\hat{\theta}$ from the formula

$$V(\hat{\theta}) = 2\left(\frac{d^2X^2}{d\theta^2}\right)^{-1}_{\theta=\hat{\theta}}. \tag{10.142}$$

which will be correct to the approximation that the higher-order terms of eq. (10.138) are small.

Under the assumption of unbiassed and normally distributed observations we can find (exact or approximate) confidence intervals for θ by seeking the intersections of the (exact or approximate) parabolic function $X^2(\theta)$ by the straight lines

$$X^2(\theta) = X^2_{min} + a. \tag{10.143}$$

Here, in the one-parameter case, the values $a = 1^2$, 2^2, and 3^2 for the intersection distance from minimum lead to confidence intervals of probabilities 68.3, 95.4, and 99.7%, respectively, which correspond to one, two, and three standard deviation intervals when $X^2(\theta)$ is strictly parabolic in θ. Hence there is a close analogy between the interval estimation from the X^2 function considered here and the likelihood function as described for the one-parameter case in Sect.9.7.1.

10.9.3 LS errors and confidence regions, the multi-parameter case

The generalization of the preceding section to the situation with more than one parameter is quite straightforward. The Taylor expansion of the X^2 function around the minimum value $\underline{\theta} = \underline{\hat{\theta}}$ reads, in general,

$$X^2(\underline{\theta}) = X^2_{min} + \tfrac{1}{2}\sum_{ij}\left(\frac{\partial^2X^2}{\partial\theta_i\partial\theta_j}\right)_{\underline{\theta}=\underline{\hat{\theta}}}(\theta_i-\hat{\theta}_i)(\theta_j-\hat{\theta}_j) + \cdots \tag{10.144}$$

By comparison with eq.(10.136), which is valid for the ideal case with a linear parameter dependence and constant covariance matrix for the observations, one finds that the elements of the covariance matrix for the LS estimate $\underline{\theta}$ can be expressed as

$$V^{-1}_{ij}(\underline{\hat{\theta}}) = \tfrac{1}{2}\left(\frac{\partial^2X^2}{\partial\theta_i\partial\theta_j}\right)_{\underline{\theta}=\underline{\hat{\theta}}}. \tag{10.145}$$

This formula is also exact to the extent that X^2 has a quadratic dependence upon $\underline{\theta}$. It can be compared to the similar expression obtained for the covariances of ML estimates, eq.(9.32).

In the simplest situation with a linear model and a parameter independent covariance matrix for the normally distributed observations, the double sum in eq.(10.144) - which is identical to the covariance form of eq.(10.136) - is a chi-square variable for which the number of degrees of freedom is equal to the number of estimated parameters minus the number of linear constraints, if any. Specifically, with only two independent parameters the intersections between the $X^2(\underline{\theta})$ surface and the parallel planes at distance a from the minimum X^2_{min} will be a set of concentric ellipses which define joint confidence regions for the two parameters, whose probability content is determined by $\chi^2(2)$. Hence the elliptic confidence regions obtained by taking $a = 1^2$, 2^2, 3^2 will have associated probabilities 39.3, 86.5, 98.9%, respectively, in complete analogy with the joint likelihood regions for the asymptotic two-parameter case discussed in Sect. 9.7.4. In the general multi-parameter case, the intersection between the $X^2(\underline{\theta})$ hypersurface and the hyperplane at distance a above the minimum will produce a hyperelliptic joint confidence region for all the parameters, for which the confidence coefficient is exactly given by the cumulative integral, up to the value a, of the chi-square p.d.f. with a number of degrees of freedom equal to the

number of independent parameters. Evidently, the associated probabilities for fixed values $a = 1^2, 2^2, \cdots$ will decrease quickly when the number of parameters increases. Conversely, to have a specified probability content for the joint confidence region, larger values of a must be taken for increasing number of parameters.

From the joint confidence region for all parameters considered simultaneously one can also deduce conditional confidence regions (intervals) for subsets of the parameters, by seeking the intersections between this region and lines for fixed values of the remaining parameters, for example their estimated values. The arguments are the same as in Sects.9.7.4 and 9.7.6 for the asymptotic likelihood function. The specific choice $a = 1^2$ will as before supply the errors in the parameter estimates by the hypersurface circumscribing the joint confidence region, as well as the conditional errors obtained by keeping some parameters at their estimated values. In particular, with two independent parameters the situation is completely analogous to that described in detail for the binormal likelihood function, Sects.9.6.3 and 9.7.4.

Irrespective of whether the actual minimization of $X^2(\underline{\theta})$ involves a linear model or not, the errors in the LS estimates are by convention determined from the intersecting hyperplane at one unit above the minimum X^2_{min}, and confidence regions deduced with the choices $a = 1^2, 2^2, \cdots$. In all situations when $X^2(\underline{\theta})$ is not of second order in the parameters and/or the observations are not normally distributed, a comparison with the chi-square distribution will obviously provide only approximate values for the probabilities associated with the different regions. How good the approximation is will in general depend on the magnitude of the higher-order terms in the series expansion of $X^2(\underline{\theta})$ and the validity of the normality assumption for the observations.

11. The method of moments

The parameter estimators constructed by the *method of moments* (MM) are consistent but in general neither as efficient as the Maximum-Likelihood estimators, nor sufficient. However, although the qualities of efficiency and sufficiency are important an estimation method should not be judged from its theoretical optimum properties alone, but also for its applicability to practical problems. In particle physics, the moments method because of its feasibility has been widely used in experiments to determine polarization and density matrix elements. The MM estimates can be easily obtained, since their evaluation only involves calculation of expectation values and averages of specified functions over the experimental sample.

11.1 BASIS FOR THE SIMPLE MOMENTS METHOD

Given a probability density function $f(x|\underline{\theta})$ with unknown parameters $\underline{\theta} = \{\theta_1, \theta_2, \ldots, \theta_k\}$ we want to estimate these parameters from a set of observations $x_1, x_2, \ldots, x_n$. The r-th algebraic moment of the population is defined by (see eq.(3.13) of Sect.3.3.3)

$$\mu'_r(\underline{\theta}) = \int_\Omega x^r f(x|\underline{\theta}) dx , \qquad r = 1,2,\ldots, \tag{11.1}$$

where Ω is the domain of x. A reasonable estimator of $\mu'_r(\underline{\theta})$ is then the arithmethic mean of the r-th power of the observations x_i,

$$m'_r = \frac{1}{n} \sum_{i=1}^{n} x_i^r , \qquad r = 1,2,\ldots \tag{11.2}$$

By equating the different moments of the parent population, which are functions of the unknown $\underline{\theta}$, to the numerical values of the corresponding sample moments we get

$$\left.\begin{aligned} \mu_1'(\underline{\theta}) &= m_1' \\ \mu_2'(\underline{\theta}) &= m_2' \\ \mu_3'(\underline{\theta}) &= m_3' \\ &\textit{etc.} \end{aligned}\right\} \tag{11.3}$$

A set of equations (11.3) can therefore be found and solved to give the MM estimates $\hat{\underline{\theta}} = \{\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k\}$. As a limited number of moments usually will not contain all information about the p.d.f. the MM estimators will in general be less efficient than the Maximum-Likelihood estimators.

The estimator m_r' of eq.(11.2) is an unbiassed estimator of μ_r', since

$$E(m_r') = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i^r\right) = \mu_r' , \tag{11.4}$$

in accordance with the requirement (8.3) for an unbiassed estimator.

To evaluate the variance of m_r' we observe that

$$V(m_r') = E\left((m_r' - \mu_r')^2\right) = E(m_r'^2) - \mu_r'^2 = E\left(\left(\frac{1}{n}\sum_{i=1}^{n} x_i^r\right)\left(\frac{1}{n}\sum_{j=1}^{n} x_j^r\right)\right) - \mu_r'^2$$

$$= \frac{1}{n^2}E\left(\sum_i x_i^{2r} + \sum_i \sum_{j\neq i} x_i^r x_j^r\right) - \mu_r'^2 = \frac{1}{n^2}\left(n\mu_{2r}' + n(n-1)\mu_r'^2\right) - \mu_r'^2,$$

where we use the fact that the x_i are independent. Hence

$$V(m_r') = \frac{1}{n}\left(\mu_{2r}' - \mu_r'^2\right), \tag{11.5}$$

which shows that the variance of the sample moment of a given order is dependent on the population moment of *twice* this order; $V(m_r')$ may therefore, even when n is large, be of considerable magnitude for higher moments if the p.d.f. has substantial tails. This explains why the simple method of taking the moments of the variable x itself (*i.e.* eqs.(11.3)) is rather seldom used in practice.

Exercise 11.1: Show that the covariance between m_r' and m_s' is given by

$$\mathrm{cov}(m_r', m_s') = n^{-1}(\mu_{r+s}' - \mu_r'\mu_s').$$

Exercise 11.2: Show that the MM estimators of the first algebraic moment and the second central moment of any p.d.f. are $\hat{\mu} = \frac{1}{n}\sum_i x_i = \bar{x}$ and $\hat{\sigma^2} = \frac{1}{n}\sum_i (x_i - \bar{x})^2$, respectively. Show that the variance in the MM estimate $\hat{\mu}$ is

$$V(\hat{\mu}) = \frac{1}{n}(\mu_2' - \mu_1'^2) = \sigma^2/n \simeq \hat{\sigma^2}/n .$$

11.2 GENERALIZED MOMENTS METHOD

Instead of using a set of powers of the variable x to estimate the unknown $\underline{\theta}$, one can select a set of independent functions of x and proceed in a similar way to construct estimators for these functions in terms of appropriate averages of the functions evaluated for the sample values $x_1, x_2, \ldots, x_n$ *).

11.2.1 One-parameter case

In the simplest situation, when there is only one unknown parameter, it will suffice to consider a single function g(x). The expectation of g(x), or the first moment of this function, for the p.d.f. $f(x|\theta)$ is defined by (compare eq.(3.6) of Sect.3.3.1)

$$E\big(g(x)\big) \equiv \gamma(\theta) = \int_\Omega g(x)\, f(x|\theta)\,dx . \tag{11.6}$$

An estimator of $\gamma(\theta)$ is now the average of g(x) over the sample,

$$\hat{\gamma}(\theta) = \overline{g(x)} = \frac{1}{n}\sum_{i=1}^{n} g(x_i) . \tag{11.7}$$

The variance of this linear combination is

$$V(\hat{\gamma}) = \left(\frac{1}{n}\right)^2 V\left(\sum_{i=1}^{n} g(x_i)\right) = \frac{1}{n}\, V\big(g(x)\big), \tag{11.8}$$

where for $V\big(g(x)\big)$ we may insert its estimated value obtained from the sample,

$$V\big(g(x)\big) \simeq s_g^2 = \frac{1}{n-1}\sum_{i=1}^{n}\big(g(x_i) - \overline{g(x)}\big)^2 . \tag{11.9}$$

Thus we take

$$V(\hat{\gamma}) \simeq \frac{1}{n(n-1)}\sum_{i=1}^{n}\big(g(x_i) - \overline{g(x)}\big)^2 . \tag{11.10}$$

An alternative form, convenient for numerical computation, is obtained after a little algebra,

$$V(\hat{\gamma}) \simeq \frac{1}{n(n-1)}\left\{\sum_{i=1}^{n} g^2(x_i) - \frac{1}{n}\left(\sum_{i=1}^{n} g(x_i)\right)^2\right\} . \tag{11.11}$$

---

*) We will in the following allow x to have several components; x_i will therefore denote all measured quantities for the i-th event.

To summarize, eq.(11.7) will produce the MM estimate for the function $\gamma(\theta)$, and eq.(11.10) (or alternatively. eq.(11.11)) an estimate of its variance. These estimates must next be "inverted" to give the desired $\hat{\theta}$ and its error.

Exercise 11.3: Verify that the choice $g(x) = x$ reduces the formulae of the last section to the previously established expressions for the MM estimate of μ and its variance (Exercise 11.2).

11.2.2 Multi-parameter case

Let us now assume that the estimation problem involves k unknown parameters, and that we have selected a set of k linearly independent functions $g_1(x)$, $g_2(x),\ldots,g_k(x)$. With the p.d.f. $f(x|\theta)$ the functions have expectation values given by

$$E\big(g_r(x)\big) \equiv \gamma_r(\underline{\theta}) = \int_{\Omega} g_r(x)\ f(x|\underline{\theta})dx\,, \qquad r=1,2,\ldots,k. \tag{11.12}$$

These expectation values are functions of the unknown $\underline{\theta}$, and for their estimators we may take

$$\hat{\gamma}_r(\underline{\theta}) = \overline{g_r(x)} = \frac{1}{n}\sum_{i=1}^{n} g_r(x_i), \qquad r=1,2,\ldots,k. \tag{11.13}$$

These expressions are of course similar to the formulae written down for the one-parameter case. For the covariance terms between the $\overline{g_r(x)}$ we generalize eqs. (11.10)-(11.11) to give the elements of the covariance matrix for $\hat{\underline{\gamma}}=\{\hat{\gamma}_1,\ldots,\hat{\gamma}_k\}$,

$$V_{rs}(\hat{\underline{\gamma}}) \simeq \frac{1}{n}\frac{1}{(n-1)}\sum_{i=1}^{n}\Big(g_r(x_i) - \overline{g_r(x)}\Big)\Big(g_s(x_i) - \overline{g_s(x)}\Big)$$

$$= \frac{1}{n(n-1)}\Big\{\sum_{i=1}^{n} g_r(x_i)g_s(x_i) - \frac{1}{n}\Big(\sum_{i=1}^{n} g_r(x_i)\Big)\Big(\sum_{i=1}^{n} g_s(x_i)\Big)\Big\}. \tag{11.14}$$

In practice one will employ functions $g_r(x)$ for which the expectations have simple relationships to the parameters $\underline{\theta}$, and which are such that the variances on the estimates $\hat{\underline{\theta}}$ will not become too large.

11.2.3 Example: Density matrix elements (2)

We shall see how the generalized moments method can be used to find the density matrix elements in the example which was introduced in Sect.9.5.8 in

connection with the Maximum-Likelihood approach to the estimation problem. We assume that a sample of n resonance events has been obtained, each event corresponding to two measured quantities $\cos\theta_i$, ϕ_i. The theoretical distribution for the decay of a vector meson into two pseudoscalar mesons is

$$f(\cos\theta,\phi|\rho_{00},\rho_{1-1},\mathrm{Re}\rho_{10}) = \frac{3}{4\pi}\Big(\tfrac{1}{2}(1-\rho_{00}) + \tfrac{1}{2}(3\rho_{00}-1)\cos^2\theta$$

$$- \rho_{1-1}\sin^2\theta\cos2\phi - \sqrt{2}\ \mathrm{Re}\rho_{10}\sin2\theta\cos\phi\Big)\,, \tag{9.41}$$

where $-1 \le \cos\theta \le +1$, $0 \le \phi \le 2\pi$, and $\rho_{00},\rho_{1-1},\mathrm{Re}\rho_{10}$ are the three unknown parameters.

Guided by the form of the p.d.f. we now define three function g_1,g_2,g_3 of the angular variables as

$$\left.\begin{aligned} g_1(\cos\theta,\phi) &= \cos^2\theta\,,\\ g_2(\cos\theta,\phi) &= \sin^2\theta\cos2\phi,\\ g_3(\cos\theta,\phi) &= \sin2\theta\cos\phi\,. \end{aligned}\right\} \tag{11.15}$$

Calculating the expectation values of these functions for the p.d.f. of eq.(9.41) and equating the expectations to the corresponding sample means we get

$$\left.\begin{aligned} E\big(g_1(\cos\theta,\phi)\big) &= \frac{1}{5}(1+2\rho_{00}) = \frac{1}{n}\textstyle\sum\cos^2\theta_i &&= \bar{g}_1\,,\\ E\big(g_2(\cos\theta,\phi)\big) &= -\frac{4}{5}\rho_{1-1} = \frac{1}{n}\textstyle\sum\sin^2\theta_i\cos2\phi_i &&= \bar{g}_2\,,\\ E\big(g_3(\cos\theta,\phi)\big) &= -\frac{4}{5}\sqrt{2}\mathrm{Re}\rho_{10} = \frac{1}{n}\textstyle\sum\sin2\theta_i\cos\phi_i &&= \bar{g}_3\,, \end{aligned}\right\} \tag{11.16}$$

where the summations are extended over all observed events.

The set of linear equations can easily be inverted to give the MM estimates of the density matrix elements. To find the errors in these estimates let us for convenience write the linear relationship as

$$\hat{\underline{\rho}} \equiv \underline{c} + S\bar{\underline{g}}\,, \tag{11.17}$$

where $\hat{\underline{\rho}} = \{\rho_{00},\rho_{1-1},\mathrm{Re}\rho_{10}\}$, $\bar{\underline{g}} = \{\bar{g}_1,\bar{g}_2,\bar{g}_3\}$, $\underline{c} = \{-\frac{1}{5},0,0\}$, and

$$S = \begin{pmatrix} 5/2 & 0 & 0 \\ 0 & -5/4 & 0 \\ 0 & 0 & -5/4\sqrt{2} \end{pmatrix}. \qquad (11.18)$$

When the variance matrix $V(\hat{\underline{\gamma}}) = V(\hat{\underline{g}})$ has been calculated from the measurements using eq.(11.14), the law of propagation of errors, eq.(3.80), applies and gives the variance matrix $V(\hat{\underline{\rho}})$ of the density matrix elements as

$$V(\hat{\underline{\rho}}) = SV(\hat{\underline{g}})S^T, \qquad (11.19)$$

or, explicitly

$$V(\hat{\underline{\rho}}) = \frac{25}{32} \begin{pmatrix} 8V_{11} & -4V_{12} & -2\sqrt{2}V_{13} \\ -4V_{12} & 2V_{22} & \sqrt{2}V_{23} \\ -2\sqrt{2}V_{13} & \sqrt{2}V_{23} & V_{33} \end{pmatrix},$$

where V_{rs} is short for $V_{rs}(\hat{\underline{g}})$ from eq.(11.14).

The previous considerations assume that our sample consists of n events that all represent true decays of the specific type $1^- \to 0^- + 0^-$. Most often it is not experimentally possible to obtain a pure sample of resonance events, and it is necessary to perform some kind of background subtraction. One popular way of doing this is to determine the density matrix elements ρ_a using all the n_a events in the resonance region of the effective mass plot, and then to calculate the same quantities ρ_b using the events in two adjacent mass regions. The number n_b of background events within the resonance region can be estimated from the shape of the effective mass spectrum. Under the assumption that this background is well described by the events in the neighbouring regions we can estimate the density matrix elements ρ of the resonance by the approximation

$$\rho = \frac{n_a\rho_a - n_b\rho_b}{n_a - n_b}. \qquad (11.20)$$

The uncertainty in this quantity is estimated as

$$\Delta\rho = \frac{1}{|n_a - n_b|}\sqrt{(n_a\Delta\rho_a)^2 + (n_b\Delta\rho_b)^2}, \qquad (11.21)$$

where $\Delta\rho_a$, $\Delta\rho_b$ are the errors connected to the estimates of ρ_a,ρ_b.

Exercise 11.4: The two-body decay $\frac{3}{2}^+ \to \frac{1}{2}^+ + 0^-$ can in the Jackson reference system be described by the p.d.f.

$$f = \frac{3}{4\pi}\left(\frac{1}{6}(1+4\rho_{33}) + \frac{1}{2}(1-4\rho_{33})\cos^2\theta - \frac{2}{\sqrt{3}}\mathrm{Re}\rho_{31}\sin 2\theta\cos\phi - \frac{2}{\sqrt{3}}\mathrm{Re}\rho_{3-1}\sin^2\theta\cos 2\phi\right),$$

where θ is the polar angle, $0 \le \theta \le \pi$, and ϕ the azimuth angle, $0 \le \phi \le 2\pi$. Show that the three density matrix elements ρ_{33},$\mathrm{Re}\rho_{31}$,$\mathrm{Re}\rho_{3-1}$ and their errors can be obtained by the moments method using the trial function of eq.(11.15).

Exercise 11.5: The decay $2^+ \to 0^- + 0^-$ has the distribution

$$\begin{aligned} f = \frac{15}{16\pi}\Big\{ & 3\rho_{00}(\cos^2\theta - \tfrac{1}{3})^2 + 4\rho_{11}\sin^2\theta\cos^2\theta + \rho_{22}\sin^4\theta \\ & - 2\cos\phi\sin 2\theta\left(\mathrm{Re}\rho_{21}\sin^2\theta + \sqrt{6}\mathrm{Re}\rho_{10}(\cos^2\theta - \tfrac{1}{3})\right) \\ & - 2\cos 2\phi\sin^2\theta\left(2\rho_{1-1}\cos^2\theta - \sqrt{6}\mathrm{Re}\rho_{20}(\cos^2\theta - \tfrac{1}{3})\right) \\ & + 2\mathrm{Re}\rho_{2-1}\cos 3\phi\sin^2\theta\sin 2\theta + \rho_{2-2}\cos 4\phi\sin^4\theta\Big\}. \end{aligned}$$

The nine density matrix elements are not all independent since the normalization condition requires $\rho_{00} + 2\rho_{11} + 2\rho_{22} = 1$. Discuss how a set of trial functions can be chosen for this p.d.f., which will lead to MM estimates for the parameters.

Exercise 11.6: Show that the formulae for the decays $1^- \to 0^- + 0^-$ (eq.(9.41)) and $2^+ \to 0^- + 0^-$ (Exercise 11.5) follow from the general formula eq.(10.85). (Hint: The density matrix ρ is Hermitean.)

11.3 MOMENTS METHOD WITH ORTHONORMAL FUNCTIONS

The method outlined in Sect.11.2.2 becomes especially simple if the p.d.f. can be expressed as

$$f(x|\underline{\theta}) = 1 + \sum_{r=1}^{k} \theta_r\xi_r(x), \qquad 0 \le x \le 1, \qquad (11.22)$$

where the $\xi_r(x)$ constitute a set of k *orthonormal functions*, satisfying

$$\int_0^1 \xi_r(x)\xi_s(x)dx = \delta_{rs}, \qquad r,s=1,2,\ldots,k, \qquad (11.23)$$

and

$$\int_0^1 \xi_r(x)dx = 0, \qquad r=1,2,\ldots,k. \qquad (11.24)$$

Then eq.(11.12) yields the expectation of $\xi_r(x)$ simply as

$$E\left(\xi_r(x)\right) = \theta_r . \qquad (11.25)$$

Therefore, an unbiassed estimator of θ_r is provided by the average value of the function $\xi_r(x)$ over the sample,

$$\hat{\theta}_r = \overline{\xi_r(x)} = \frac{1}{n}\sum_{i=1}^{n}\xi_r(x_i), \qquad r=1,2,\ldots,k. \qquad (11.26)$$

Since the functions $\xi_r(x)$ are orthogonal the covariance terms between different $\hat{\theta}_r, \hat{\theta}_s$ vanish. The error in $\hat{\theta}_r$ can be found from the approximate formula for the variance, eq.(11.14), which becomes

$$V_{rr}(\hat{\underline{\theta}}) \equiv s_r^2 = \frac{1}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n}\xi_r^2(x_i) - \hat{\theta}_r^2\right) \simeq \frac{1}{n-1}\left(1-\hat{\theta}_r^2\right). \qquad (11.27)$$

11.3.1 Example: Polarization of antiprotons (3)

We reconsider the polarization example described under the Maximum-Likelihood method in Sect.9.5.7 and under the Least-Squares method in Sect.10.5.3. The distribution of the angle ϕ between the normals of the two scattering planes is given by

$$f(\cos\phi|\alpha) = \tfrac{1}{2}(1 + \alpha\cos\phi), \qquad -1 \le \cos\phi \le 1, \qquad (9.38)$$

where the unknown is $\alpha=P^2$, the square of the polarization.

The p.d.f. is not of the form of eq.(11.22). However, in terms of a new variable

$$y = \tfrac{1}{2}(\cos\phi + 1), \qquad 0 \le y \le 1, \qquad (11.28)$$

we can write the p.d.f. in the required form

$$f'(y|\alpha') = 1 + \alpha'\xi(y)$$

if the new parameter is taken to be $\alpha' = \alpha/\sqrt{3}$ and

$$\xi(y) = \sqrt{3}(2y-1).$$

The function $\xi(y)$ satisfies the two conditions of eqs.(11.23), (11.24). Hence the results of the previous section can be applied directly. An unbiassed estimator of α' is therefore given by (eq.(11.26))

$$\hat{\alpha'} = \overline{\xi(y)} = \frac{1}{n}\sum_{i=1}^{n}\sqrt{3}(2y_i-1),$$

for which the variance is asymptotically (eq.(11.27))

$$V(\hat{\alpha'}) = \frac{1}{n-1}\left(1-\hat{\alpha'}^2\right).$$

In terms of the original variable $\cos\phi$ the estimate of the parameter α and its variance become

$$\hat{\alpha} = \sqrt{3}\,\hat{\alpha'} = \frac{3}{n}\sum_{i=1}^{n}\cos\phi_i , \qquad (11.29)$$

$$V(\hat{\alpha}) = 3V(\hat{\alpha'}) = \frac{1}{n-1}\left(3-\hat{\alpha}^2\right). \qquad (11.30)$$

It is interesting to compare this last result with the corresponding result for the variance obtained with the Maximum-Likelihood method. For large n the variance of the ML estimate of α takes the smallest possible value, given by eq.(9.39). Hence the asymptotic efficiency of the moments estimator is

$$\text{Efficiency}\ (\hat{\alpha}_{MM}) = \frac{V_{ML}(\hat{\alpha})}{V_{MM}(\hat{\alpha})} = \frac{\frac{1}{n}\,\frac{2\hat{\alpha}^3}{\ln(1+\hat{\alpha}) - \ln(1-\hat{\alpha}) - 2\hat{\alpha}}}{\frac{1}{n-1}\left(3-\hat{\alpha}^2\right)}, \qquad (11.31)$$

which for large n and $\hat{\alpha}<<1$ gives

$$\text{Efficiency}\ (\hat{\alpha}_{MM}) \simeq 1 - \frac{4}{15}\hat{\alpha}^2. \qquad (11.32)$$

Thus, for small polarizations P, the MM estimator is nearly fully efficient. Numerically, the asymptotic efficiency is 0.99997 for P = 0.10, and 0.998 for P = 0.30.

See further Chapter 12.

Exercise 11.7: Show that a more general expression than eq.(11.30) is

$$V(\hat{\alpha}) = \frac{1}{n-1}\left(\frac{9}{n}\sum_{i=1}^{n}\cos^2\phi - \hat{\alpha}^2\right).$$

This variance formula is valid for all sample sizes.

11.3.2 Example: Angular momentum analysis (2)

We go back to the example of Sect.10.5.4 and write the formula (10.87) for the angular distribution of a two-pion decay of a spin J boson in the form

$$W(\Omega) = \frac{1}{4\pi} + \sum_{j=2,4,\ldots}^{2J} c_{jm} Y_j^m(\Omega) , \tag{11.33}$$

where the $Y_j^m(\Omega)$ are the well-known spherical harmonic functions. All conditions specified in Sect.11.3 for the use of orthogonal functions are satisfied, except for the trivial difference in variable region. We have now

$$\left.\begin{aligned} &\int_{4\pi} W(\Omega)\,d\Omega = 1 , \\ &\int_{4\pi} Y_j^m(\Omega) Y_{j'}^{m'}(\Omega)^* d\Omega = \delta_{jj'}\delta_{mm'} , \\ &\int_{4\pi} Y_j^m(\Omega)\,d\Omega = 0 . \end{aligned}\right\} \tag{11.34}$$

Hence the results of Sect.11.3 apply, giving the unbiassed estimates of the expansion coefficients c_{jm} as the average values of the Y_j^m over the experimental sample, (eq.(11.26))

$$\hat{c}_{jm} = \overline{Y_j^m(\Omega)} = \frac{1}{n} \sum_{i=1}^{n} Y_j^m(\cos\theta_i, \phi_i) , \tag{11.35}$$

where n is the number of events. The estimates are uncorrelated and have variances given by eq.(11.27),

$$V(\hat{c}_{jm}) \simeq \frac{1}{n-1} (1-\hat{c}_{jm}^2) . \tag{11.36}$$

It will be seen that estimating by the moments method with orthogonal functions, as described here, is computationally much simpler than the Least-Squares estimation of Sect.10.5.4.

11.3.3 Confidence intervals for MM estimates

Working with orthonormal functions has the further advantage that confidence intervals can be obtained for the parameters *independently*. For the r-th parameter we have from the Central Limit Theorem, when n becomes large,

$$\frac{\sum_{i=1}^{n} g_r(x_i) - n\theta_r}{\sqrt{n\sigma_r^2}} = \frac{\hat{\theta}_r - \theta_r}{\sigma_r/\sqrt{n}} \to N(0,1) , \tag{11.37}$$

where $\hat{\theta}_r$ is the MM estimate of θ_r from eq.(11.26), and where σ_r^2 is the (unknown) variance in the distribution of $g_r(x)$ around θ_r. For large n, $\sigma_r/\sqrt{n}$ may be approximated by $s_r/\sqrt{n}$, where s_r^2 is the sample variance of $g_r(x)$ from eq.(11.27); hence

$$\frac{\hat{\theta}_r - \theta_r}{s_r/\sqrt{n}} \to N(0,1) \qquad \text{when } n \to \infty . \tag{11.38}$$

The standard normal distribution can therefore, when n is not too small, be used to derive approximate confidence intervals for each θ_r from the observed $\hat{\theta}_r$, s_r.

11.4 COMBINING MM ESTIMATES FROM DIFFERENT EXPERIMENTS

Suppose that two experiments have estimated the parameters $\underline{\theta}$ by the moments method using the same sets of orthogonal functions. The experiments have been based on n_1 and n_2 events, respectively, and have given the estimates $\underline{\theta}^{(1)}$ and $\underline{\theta}^{(2)}$. The result of the combined experiment for the r-th component in $\underline{\theta}$ is

$$\hat{\theta}_r = \frac{1}{n_1+n_2} \sum_{i=1}^{n_1+n_2} g_r(x_i) = \frac{n_1}{n_1+n_2} \hat{\theta}_r^{(1)} + \frac{n_2}{n_1+n_2} \hat{\theta}_r^{(2)} ,$$

which is nothing but a simple weighted combination of the MM estimates from the individual experiments. The variance of this combined estimate of θ_r is

$$V(\hat{\theta}_r) \simeq \frac{1}{n_1+n_2-1} (1-\hat{\theta}_r^2) .$$

If the two experiments have used the more general moments method of Sect.11.2.2 the MM estimates of each experiment will be associated with a non-diagonal covariance matrix. The combined estimate and its error must then be found by a more elaborate procedure, for example by a Least-Squares approach as described for the example of Sect.10.2.6.

12. A simple case study with application of different parameter estimation methods

In practical problems to determine numerical values of physical constants on the basis of sets of experimental data the physicist may have the opportunity to choose between different estimation methods. His actual choice of estimators will usually depend on statistical as well as other criteria. General statistical properties that should be possessed by good estimators were considered in Chapter 8. These optimum properties are:

- *consistency*, which means that the estimator gives rise to estimates that converge towards the true parameter value when the number of observations is increased;
- *unbiassedness*, which means that, regardless of the sample size, the estimator produces estimates that are not systematically shifted from the true parameter value;
- *efficiency*, which means that the distribution of the estimates has *minimum variance* about the central value (equal to the true value of the parameter for unbiassed estimators);
- *sufficiency*, which means that the estimator exhausts all information in the observations regarding the unknown parameter.

In choosing between possible estimators the physicist will also take other factors into consideration. Preferentially,

- the establishing of necessary formulae for computation should be as simple as possible;
- the computer programming should not be too complicated; for example, relevant software should be available for matrix inversion and function optimization;
- the method should make economic use of computer time.

Some of the ideal theoretical properties and the practical demands will frequently be conflicting. In practice one will therefore have to give

priority to certain factors and sacrifice others which are assumed less important.

In this chapter we will consider a particularly simple example with simulated polarization experiments where the estimation involves a single parameter . The purpose is, firstly, to recall how the parameter estimate and its error can be obtained by the different estimation methods described earlier and, secondly, to demonstrate that these estimators produce mutually consistent results.

12.1 SIMULATION OF POLARIZATION EXPERIMENTS

We will take for a case study the simple one-parameter example of Sects.9.5.7, 10.5.3, and 11.3.1, describing the polarization of antiprotons in $\bar{p}p$ elastic scattering. The theoretical distribution for the angle ϕ between the scattering normals is given by

$$f(x|\alpha) = \tfrac{1}{2}(1 + \alpha x) \tag{12.1}$$

where $x = \cos\phi$ is restricted to the interval $[-1,+1]$. For this class of underlying distributions we have generated artificial event samples corresponding to specified values of the parameter α by the "hit-and-miss" Monte Carlo method using a computer equipped with a uniform random number generator. The number generator delivers numbers r which are uniformly distributed between 0 and 1. An event candidate can be constructed from two consecutive numbers r_1 and r_2 from the generator by defining the angle ϕ_i through the relation

$$\cos\phi_i = 2r_1 - 1, \tag{12.2}$$

and this candidate is accepted and included in the artificial event sample if, for the specified α,

$$f(\cos\phi_i|\alpha) = \tfrac{1}{2}(1 + \alpha\cos\phi_i) > r_2. \tag{12.3}$$

We have carried out simulations using two different values of the parameter α, α=0.09 (corresponding to a polarization $P=\sqrt{\alpha}=0.3$) and α=0.25 (P=0.5). For each α, four independent, artificial samples were generated with n=10, 100, 1000, and 10000, giving altogether 8 independent, simulated experi-

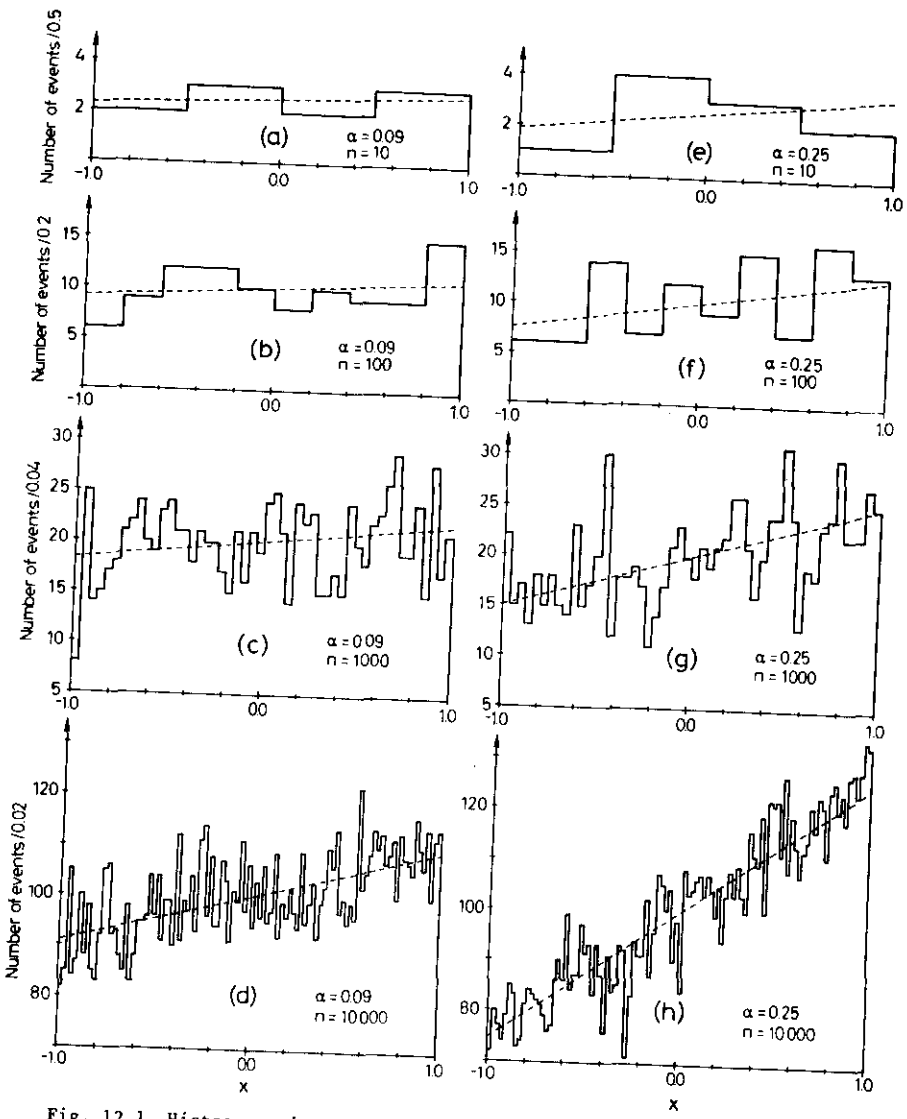

Fig. 12.1. Histograms in x=cosϕ for 8 Monte Carlo simulated polarization experiments generated with the indicated values of the parameter α. The dashed straight lines correspond to the theoretical distribution $f(x|\alpha)$ of eq.(12.1) normalized to the generated number of events n.

ments. Figure 12.1 shows histograms in x=cosϕ for the "events" from these 8 "experiments", together with curves showing the underlying "theoretical" distributions $f(x|\alpha)$ with the α-values used in the simulations, normalized to the number of events in the "experiments". It is seen that all histograms match well with the "theoretical" distributions, and it appears justified to consider the generated event samples as fairly typical for physical events originating from the corresponding distributions. We will therefore, in the following, regard the generated samples as if they consisted of real, observed events in 8 different experiments, and use them to estimate the unknown parameter α by the usual estimation methods described in Chapters 9-11.

12.2 APPLICATION OF DIFFERENT ESTIMATION METHODS

We proceed now to see how the data of the simulated experiments can be used to obtain estimates of the unknown parameter and its error. In the following sections we briefly recapitulate how $\hat{\alpha}$ and $\Delta\hat{\alpha}$, or $V(\hat{\alpha})$, can be found by the different estimators using explicit formulae and graphical methods. The numerical results obtained by the different methods for the 8 experiments are all summarized in Table 12.1, and will be discussed further in Sect.12.3 below.

12.2.1 The method of moments

The present example was treated in Sect.11.3.1 using the moments method with orthogonal functions. The MM estimate of α is simply given by (eq.(11.29))

$$\hat{\alpha} = \frac{3}{n}\sum_{i=1}^{n} x_i, \tag{12.4}$$

and for large samples the error in $\hat{\alpha}$ can be found from the asymptotic expression for the variance, (eq.(11.30)),

$$V(\hat{\alpha}) = \frac{1}{n-1}\,(3-\hat{\alpha}^2). \tag{12.5}$$

For small samples the error can be obtained from the more general variance formula (Exercise 11.7)

$$V(\hat{\alpha}) = \frac{1}{n-1}\left(\frac{9}{n}\sum_{i=1}^{n} x_i^2 - \hat{\alpha}^2\right) \tag{12.6}$$

The last formula has been used to obtain the errors for the generated samples with n not exceeding 100, although one may expect the numerical result for $V(\hat{\alpha})$ to be rather sensitive to the particular sample values x_i, given n.

12.2.2 The Maximum-Likelihood method

The likelihood function is given by

$$L(x_1,x_2,\ldots,x_n|\alpha) = \prod_{i=1}^{n} \tfrac{1}{2}(1+\alpha x_i)$$

from which

$$\ln L = -n\ln 2 + \sum_{i=1}^{n} \ln(1+\alpha x_i). \qquad (12.7)$$

In Fig. 12.2 lnL is shown as a function of α for the 8 experiments. For each experiment the ML estimate $\hat{\alpha}$ corresponds to the peak of the lnL function, and the error in α is determined by intersecting the function by a straight line at a distance 0.5 below its maximum value. As can be seen from Fig. 12.2 the lnL function, even for the smallest samples, has an almost symmetric and parabolic shape. The error $\Delta\hat{\alpha}$ can therefore for each experiment be taken as the average of the distances $\Delta\hat{\alpha}_L$ and $\Delta\hat{\alpha}_U$ defined by the lower and upper intersection points; see Fig. 12.2(a).

In addition to the estimated errors obtained by the graphical method Table 12.1 also gives the errors deduced from the large-sample formula (9.39) derived for the present p.d.f. in Sect.9.5.7,

$$V(\hat{\alpha}) = \frac{1}{n}\,\frac{2\alpha^3}{\ln(1+\alpha) - \ln(1-\alpha) - 2\alpha}\,. \qquad (12.8)$$

12.2.3 The Maximum-Likelihood method for classified data

When the events are classified in N bins with n_i observed events in the i-th bin the ML solution for the unknown parameter α is found by maximizing the expression (compare Sect.9.9)

$$\ln L(n_1,n_2,\ldots,n_N|\alpha) = \sum_{i=1}^{N} n_i \ln p_i(\alpha),$$

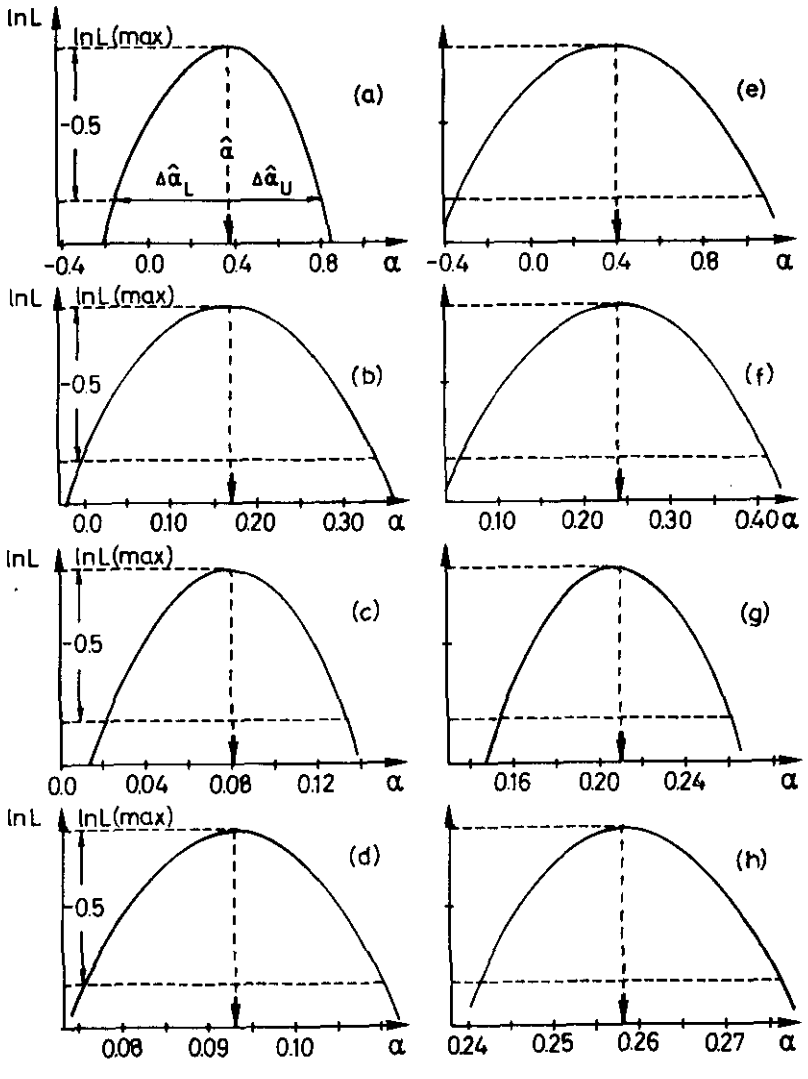

Fig. 12.2. lnL of eq.(12.7) as a function of the parameter α for the 8 simulated experiments.

where $p_i(\alpha)$ is the probability for the i-th bin, given α. For the bin extending from x_i to $x_i + \Delta x_i$ we have

$$p_i(\alpha) = \int_{x_i}^{x_i+\Delta x_i} \tfrac{1}{2}(1 + \alpha x)dx = a_i + \alpha b_i, \qquad (12.9a)$$

where

$$a_i \equiv \tfrac{1}{2}\Delta x_i, \quad b_i \equiv \tfrac{1}{2}\Delta x_i(x_i + \tfrac{1}{2}\Delta x_i). \qquad (12.9b)$$

Hence we can write

$$\ln L(n_1,n_2,\ldots,n_N|\alpha) = \sum_{i=1}^{N} n_i \ln\left\{1 + \alpha(x_i + \tfrac{1}{2}\Delta x_i)\right\} + \text{const} \qquad (12.10)$$

where the argument for the logarithm is taken at the central value of the i-th bin.

Although the ML method with classified data was introduced to save computation, and only is of practical interest when n is large, we have applied it here for demonstration purposes for n as small as 100 (10 classes). The results obtained are given in Table 12.1, together with the errors in the estimates as deduced by the graphical method using the points where the lnL function is 0.5 below its maximum value.

12.2.4 The Least-Squares method

With n events distributed in N bins the usual form for the function to be minimized is

$$X^2 = \sum_{i=1}^{N} \frac{\left(n_i - np_i(\alpha)\right)^2}{np_i(\alpha)}, \qquad (12.11)$$

where $p_i(\alpha)$, the probability for the i-th bin, is given by eqs.(12.9a,b).

The quantity X^2 is shown graphically in Fig. 12.3 as a function of α for the 8 experiments, with their minimum values indicated, corresponding to the LS estimates $\hat{\alpha}$. Also indicated in each graph is the straight line at distance 1.0 above the function minimum, which determines the error in each estimate.

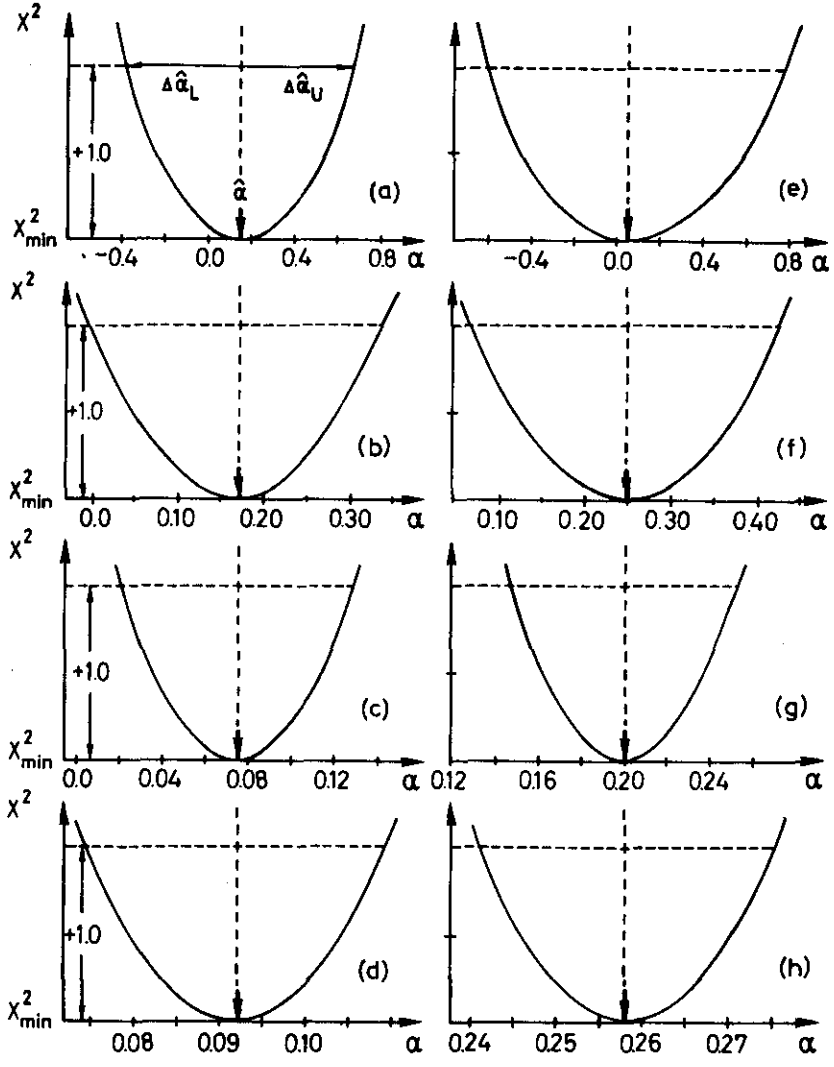

Fig. 12.3. X^2 of eq.(12.11) as a function of the parameter α for the 8 simulated experiments.

12.2.5 The simplified Least-Squares method

As we saw in Sect.10.5.3, the simplified LS method with

$$X^2 = \sum_{i=1}^{N} \frac{\left(n_i - np_i(\alpha)\right)^2}{n_i} \tag{12.12}$$

and $np_i(\alpha)$ linear in the parameter α, implies that X^2 is of second order in α. Hence the analytical solutions can be written down for the LS estimate and its error [eqs.(10.82), (10.83)],

$$\hat{\alpha} = \frac{1}{n} \sum_{i=1}^{N} \left(b_i - \frac{n a_i b_i}{n_i} \right) \Big/ \sum_{i=1}^{N} \frac{b_i^2}{n_i} , \tag{12.13}$$

$$\Delta\hat{\alpha} = \frac{1}{n} \left(\sum_{i=1}^{N} \frac{b_i^2}{n_i} \right)^{-\frac{1}{2}} . \tag{12.14}$$

Table 12.1 gives the numerical results obtained for the estimated parameter and its error with the simplified LS method as well as the ordinary LS method applied to all 8 experiments, including the two experiments with sample size n=10, which do not fulfil the usual requirements on the number of bins and their contents as discussed in Sects.10.5.1, 10.5.2; for the latter the numbers are given in parenthesis.

12.3 DISCUSSION

12.3.1 The estimated parameters and their errors

From the numerical values of Table 12.1 the following conclusions may be drawn:

- for each generated sample (experiment) the estimated values of the parameter by the five different procedures are generally in good agreement, except when the sample size is very small (n=10);
- within each sample the parameter errors estimated by the different procedures are roughly equal;
- the estimated errors are inversely proportional to the square root of the sample size.

For the larger samples the first result is not surprising, since all

Table 12.1 Summary of results from estimations

Generated sample (experiment)	Estimation method	Number of bins N	Estimated parameter value $\hat{\alpha}$	Estimated parameter error $\Delta\hat{\alpha}$		
				Analytical	Graphical	MVB
(a)	Moments	–	0.42	0.62	–	
	ML	–	0.38	0.52	0.47	
α=0.09	ML, classified data	–	–	–	–	0.54
	LS, ordinary	(4)	(0.15)	–	(0.53)	
n=10	LS, simplified	(4)	(0.16)	(0.55)	–	
(b)	Moments	–	0.172	0.175	–	
	ML	–	0.170	0.171	0.173	
α=0.09	ML, classified data	10	0.177	–	0.174	0.173
	LS, ordinary	10	0.175	–	0.172	
n=100	LS, simplified	10	0.188	0.165	–	
(c)	Moments	–	0.078	0.055	–	
	ML	–	0.080	0.054	0.056	
α=0.09	ML, classified data	50	0.080	–	0.056	0.054
	LS, ordinary	50	0.075	–	0.054	
n=1000	LS, simplified	50	0.109	0.052	–	
(d)	Moments	–	0.093	0.0173	–	
	ML	–	0.093	0.0173	0.0173	
α=0.09	ML, classified data	100	0.093	–	0.0172	0.0173
	LS, ordinary	100	0.092	–	0.0175	
n=10000	LS, simplified	100	0.095	0.0172	–	
(e)	Moments	–	0.21	0.44	–	
	ML	–	0.40	0.52	0.71	
α=0.25	ML, classified data	–	–	–	–	0.54
	LS, ordinary	(4)	(0.05)	–	(0.69)	
n=10	LS, simplified	(4)	(0.40)	(0.43)	–	
(f)	Moments	–	0.215	0.164	–	
	ML	–	0.240	0.170	0.178	
α=0.25	ML, classified data	10	0.251	–	0.180	0.170
	LS, ordinary	10	0.250	–	0.180	
n=100	LS, simplified	10	0.224	0.154	–	
(g)	Moments	–	0.211	0.055	–	
	ML	–	0.210	0.054	0.054	
α=0.25	ML, classified data	50	0.207	–	0.054	0.054
	LS, ordinary	50	0.200	–	0.057	
n=1000	LS, simplified	50	0.215	0.054	–	
(h)	Moments	–	0.262	0.0171	–	
	ML	–	0.258	0.0170	0.0172	
α=0.25	ML, classified data	100	0.259	–	0.0168	0.0170
	LS, ordinary	100	0.258	–	0.0170	
n=10000	LS, simplified	100	0.260	0.0169	–	

five estimators are consistent and asymptotically unbiassed. For the very small samples with n=10, even the methods which utilize all information in the (unbinned) data, *i.e.* the moments and the ML method, give numerically different values for the estimated parameters; however, considering the magnitude of the estimated errors, these results are not incompatible.

The dependence of the estimated errors upon the sample size is as expected. For the p.d.f. of the present example, eq.(12.1), the minimum variance bound, MVB, can be evaluated from the fundamental Cramér-Rao inequality (8.11); one finds

$$V(\hat{\alpha}) = \frac{1}{n} \frac{2\alpha^3}{\ln(1+\alpha) - \ln(1-\alpha) - 2\alpha} ,$$

which is nothing but the ML large-sample variance of eq.(12.8). The corresponding MVB error $\Delta\hat{\alpha}$ obtained for each generated sample is also given in Table 12.1. It is seen that the errors estimated by the different procedures are close to the MVB error for all sample sizes. This means that all five estimation procedures have a high efficiency also for small n. (However, no estimation procedure for α can be fully efficient for all n, since the p.d.f. of eq.(12.1) does not belong to the exponential family and therefore does not have any sufficient estimator for α; Sect.8.6.1.)

Some of the estimated errors in Table 12.1 are somewhat smaller than the corresponding MVB error. This need not disturb us, since the MVB is to be understood as the lower limit of the *expected* value of the estimated variance, and thus represents no absolute minimum for this quantity. Hence if many new samples were generated, with similar number of events, these could give smaller or larger estimated errors than the actual table value for the given methods, but in such a way that their average value, for any method, would always be at least as large as the MVB error.

12.3.2 Goodness-of-fit

As was emphasized in Chapter 10 the Least-Squares method has an advantage over other parameter estimation procedures in that it can provide a direct measure of the goodness-of-fit between a fitted model and the experimental data, since the minimum value obtained for the optimized function, under certain

conditions, has welldefined distribution properties. Specifically, if the number of events is not too small, X^2_{min} is a chi-square variable with a number of degrees of freedom equal to the number of independent terms in the X^2 sum minus the number of independent parameters estimated; the corresponding chi-square probability P_{χ^2} is then the probability for obtaining a higher value of X^2_{min} than that observed, and can be found, for example, from the graph of Fig. 5.2.

Table 12.2 gives the minimum value X^2_{min} and the deduced chi-square probability P_{χ^2} for the ordinary and the simplified LS fits to the generated samples of size $n \geq 100$. In each fit the number of degrees of freedom is (N-1) - 1 = N-2, where N is the number of bins used.

Table 12.2.

Generated sample (experiment)	Number of bins N	Ordinary LS X^2_{min}	Ordinary LS P_{χ^2}	Simplified LS X^2_{min}	Simplified LS P_{χ^2}
(b) α=0.09 n=100	10	4.6	0.80	4.5	0.82
(c) α=0.09 n=1000	50	38.5	0.82	47.3	0.56
(d) α=0.09 n=10000	100	49.0	>0.99	47.9	>0.99
(f) α=0.25 n=100	10	11.7	0.18	11.8	0.17
(g) α=0.25 n=1000	50	39.5	0.80	41.7	0.72
(h) α=0.25 n=10000	100	39.0	>0.99	39.5	>0.99

The numbers of Table 12.2 show that the chi-square probabilities from the ordinary and the simplified LS methods are similar. The high probabilities indicate that the LS fits are "good". In particular, the exceedingly high probabilities obtained for the large sample experiments in this case are very likely just a reflection of a well-behaved random number generator, which has produced artificial event samples which are extremely close to the ideal continuous distributions of eq.(12.1).

For small samples the distribution of the X^2_{min} statistic is not known, and the LS estimation is not associable with a chi-square probability expressing the goodness-of-fit. Similarly, regardless of the sample size, the moments and the ML estimation methods provide no direct measures for the goodness-of-fit. One may of course calculate a corresponding X^2 value from eq.(12.11) or (12.12) using the fitted parameter values by these methods and a reasonable binning of the data, but the X^2 statistic constructed this way is generally not simply chi-

square distributed, and one will therefore in general not be able to assign a chi-square probability for the goodness-of-fit. Only if there is a very large number of observations, corresponding to a substantial number of events in the separate bins, can the X^2 statistic as obtained by inserting the ML estimates for the parameters be regarded as approximately chi-square distributed (see Sect.14.4.3), and a chi-square probability for the goodness-of-fit be deduced from, for example, a standard graph of the cumulative chi-square distribution.

In principle, the numerical value obtained for lnL(max) could also be used to supply information on the goodness-of-fit if the distributional properties of the statistic lnL(max) were known. This is generally not the case. One can, however, construct an *approximate* probability distribution of lnL(max) corresponding to the specific $\hat{\alpha}$ and n by using the Monte Carlo technique to generate a large number of event samples, all of size n, and determine for these (independent) samples the frequency distribution of the values obtained for lnL(max). Since lnL(max) depends on the parameter α, only simulated experiments producing fitted parameter values very close to the specific $\hat{\alpha}$ should be used in deriving this frequency distribution. The probability to obtain a smaller value than the actually observed lnL(max) for the specific $\hat{\alpha}$ can then be estimated as the integrated value from $-\infty$ up to lnL(max) of the derived frequency distribution, thus providing the desired measure of goodness-of-fit.

Figure 12.4 shows the frequency distribution for lnL(max) and the corresponding cumulative distribution F obtained this way on the basis of 100 independent simulated experiments with n=10 which all gave estimated parameter values in the interval [0.37,0.41]. We take Fig. 12.4(a) to represent an approximate distribution of lnL(max) for the two small sample experiments (a) and (e) from Table 12.1, for which the ML method gave the estimated parameter value $\hat{\alpha}$ equal to 0.39 and 0.40, respectively. Since the actual values obtained for lnL(max) were -6.68 for experiment (a) and -6.79 for experiment (e), we find from Fig. 12.4(b) that the corresponding estimated ML probabilities become 0.46 and 0.04 for these experiments. Proceeding in a similar manner to obtain the approximate lnL(max) distributions for the two experiments (b) and (f) in Table 12.1, both with n=100, we estimate their ML probabilities to be 0.55 and 0.12, respectively; these numbers compare reasonably with the chi-square probabilities as given in Table 12.2, being ≈0.80 and ≈0.18 for these experiments.

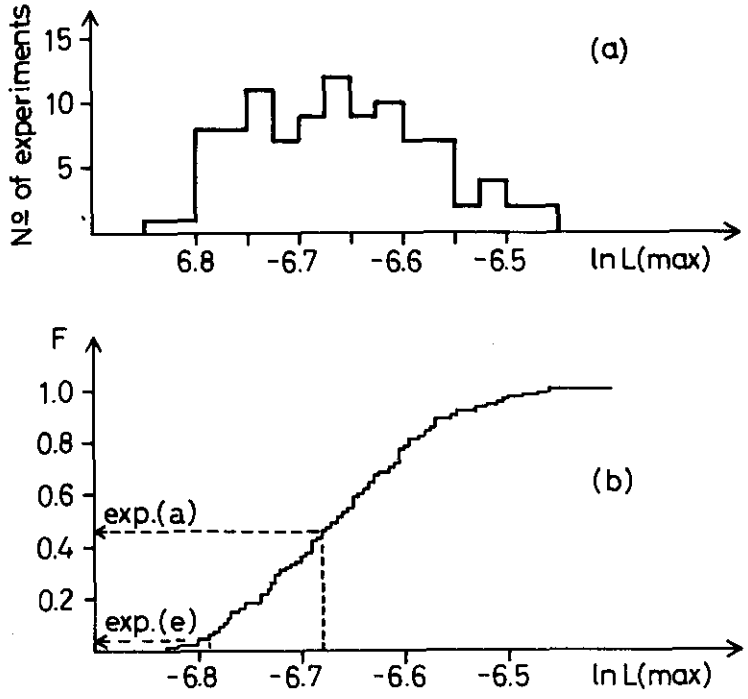

Fig. 12.4. (a) Distribution of lnL(max) obtained for 100 simulated experiments with n=10 and $\hat{\alpha} \in [0.37,0.41]$. (b) The cumulative integral of the distribution in (a).

13. Minimization procedures

There exists a variety of problems which require the optimization of some function with respect to a set of parameters. The Maximum-Likelihood and the Least-Squares methods have in common that the optimum values of the unknown parameters are determined by seeking the extremum of a function for which an explicit dependence on the parameters has been written down. Since maximizing the likelihood function is equivalent to minimizing its negative value we may formulate the ML as well as the LS methods for parameter estimation as problems of minimizing a function with respect to its variables.

The numerical minimization procedures to be outlined in this chapter are of rather general applicability, the only assumption being, for some methods, an approximate quadratic behaviour of the function in the immediate vicinity of its minimum. In particular, when minimizing the negative logarithm of the likelihood function or the sum of squared deviations by the Least-Squares method, the function is often to a good approximation quadratic in the variables near its minimum.

Effective use of the techniques described in the following presupposes that the procedures have been programmed for high-speed computers. Although much of the basic theory dates back to Newton's time or even earlier, it is now hardly conceivable to apply these techniques for hand calculation. One may conveniently think of the procedure as a programmed subroutine or algorithm which is called with assigned values of the variables (parameters) and which returns the function value and sometimes its derivatives.

The renewed interest in the minimization problem during the last years has resulted in new minimization procedures as well as improvements and extensions to old ones. We shall emphasize the principles behind selected methods and describe how they work. The reader who needs more detailed information and further theoretical justification should consult more specialized literature.

Several very good and flexible minimization programmes*) of rather

*) An example is MINUIT of the CERN Program Library.

general nature are presently available as more or less standard equipment of large computer installations. Due to the finite work length of digital computers a user may sometimes have to worry about numerical inaccuracies arising from rounding-off errors, underflow and overflow. Difficulties of this kind require special attention and can be handled by various elaborate methods. A discussion of these more technical points is, however, outside the scope of the present text.

13.1 GENERAL REMARKS

In this chapter the real function to be minimized is denoted by $F(\underline{x})$, with the variable $\underline{x} = \{x_1, x_2, \ldots, x_n\}$ corresponding to the n parameters.

A minimum point $\underline{x}^o$ of $F(\underline{x})$ is defined as a point where $F(\underline{x}) > F(\underline{x}^o)$ for all points $\underline{x}$ near $\underline{x}^o$. From elementary calculus we know that a minimum point must be of one of the following types:

(i) a *stationary point*, where all the derivatives $\partial F/\partial x_i$ are zero,

(ii) a *cusp*, where some of the derivatives are zero and others do not exist,

(iii) an *edge point*, that is, a point lying on the boundary of the allowed variable region.

We will here consider functions for which an explicit analytic expression is not specified or is very complicated. The usual way of finding the extrema of a function by equating all its first derivatives to zero is then not directly applicable. A reasonable approach in this situation is to perform a mapping or search over the variable space to locate the minima of $F(\underline{x})$. Such a search can be done in many ways, defining different minimization procedures.

The function to be minimized often has more than one minimum. Since it seems to be rather difficult to define minimization procedures which will surely produce the absolute minimum of a function, or the *global minimum*, we will at first anticipate the procedures to lead to the nearest *local minimum*.

To be able to propose "intelligent" minimization methods let us first study the function $F(\underline{x})$ near some arbitrary point $\underline{c}$. Intuitively we expect all

the derivatives of a physically meaningful $F(\underline{x})$ to exist in the region of interest. We may therefore perform a Taylor series expansion of $F(\underline{x})$ around the point $\underline{c}$ and write

$$F(\underline{x}) = F(\underline{c}) + \underline{g}^T(\underline{x}-\underline{c}) + \tfrac{1}{2}(\underline{x}-\underline{c})^T G(\underline{x}-\underline{c}) + \ldots \qquad (13.1)$$

where $\underline{g}^T$ is the transposed gradient vector with elements $g_i=\partial F/\partial x_i$, and the matrix G has elements $G_{ij}=\partial^2 F/\partial x_i \partial x_j$ with the indices i,j running from 1 to n. The derivatives are to be evaluated at $\underline{x} = \underline{c}$.

In eq.(13.1) $F(\underline{c})$ is a constant and therefore supplies no information about the location of a minimum. In the second term the vector $\underline{g}$ is expected to vary considerably over parameter space, being close to zero in the neighbourhood of a stationary minimum. The components of the product $\underline{g}^T(\underline{x}-\underline{c})$ will tell us in which direction $F(\underline{x})$ changes most rapidly, but not how far we should go to possibly reach the minimum. Information about the required step size can be gained from the third term of eq.(13.1). The matrix G, derived from the second derivatives of $F(\underline{x})$, will usually have a modest variation over parameter space, being constant for a function $F(\underline{x})$ of strictly quadratic form. Clearly, if $F(\underline{x})$ is to possess any minimum value at all there must be some restrictions on the symmetric n x n matrix G. At a stationary minimum G is positive-definite.

For a specified problem the choice of minimization method should depend on the information available on the function $F(\underline{x})$. In general, the more information about $F(\underline{x})$ actually used in the minimization, the more efficient we expect the method to be. One can conveniently consider the following situations in levels of increasing knowledge about $F(\underline{x})$,

(i) only $F(\underline{x})$ is known,

(ii) $F(\underline{x})$ as well as its first derivatives are known,

(iii) $F(\underline{x})$, its first and second derivatives are known and reasonably continuous.

We shall in the following consider minimization methods which assume different knowledge about $F(\underline{x})$. If the derivatives are not known analytically they may have to be obtained numerically.

The minimization procedures we describe can conveniently be divided into two main classes, the *step methods* and the *gradient methods*. The step methods do not use any information about the derivatives of $F(\underline{x})$ when the step length and step direction have been chosen, whereas the gradient methods do.

Common to many methods is that they need the repetition of a certain procedure to make the function converge towards a minimum. One must therefore formulate convergence criteria which will ensure that the process is brought to an end as soon as the criteria are fulfilled. The repetitions can be stopped, for example, if the change in the function value for two consecutive iterations is smaller than a preassigned number.

When a minimum has been obtained for $F(\underline{x})$ it remains to find the errors on the parameters. For the last of the minimization procedures described here, the Davidon variance algorithm, the covariance matrix is obtained by the algorithm itself. For the other minimization methods special variance algorithms must be applied, based upon reasonable assumptions about $F(\underline{x})$ and the ideas of Chapters 9 and 10, whenever these are applicable.

In Sect.13.6 we will discuss the situation which arises when the function $F(\underline{x})$ is constrained through limited allowed regions for the parameters.

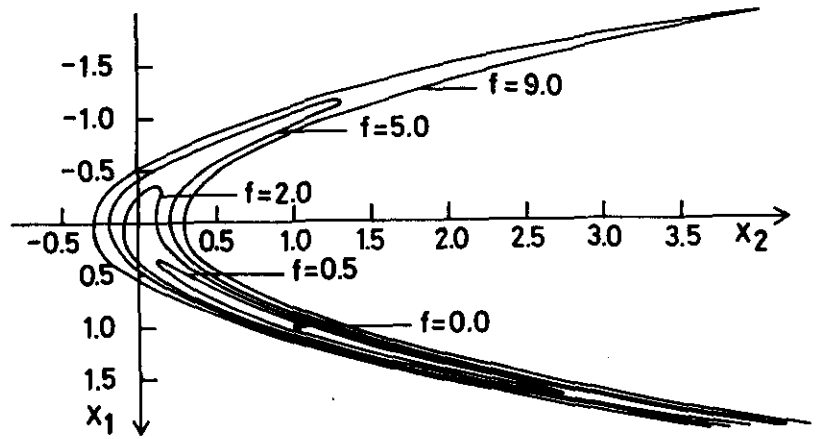

Fig. 13.1. "Rosenbrock's curved valley", $F(x_1,x_2)=100(x_2-x_1^2)^2+(1-x_1)^2$, (Exercise 13.1).

Exercise 13.1: An often used function for testing the efficiency of minimization procedures is "Rosenbrock's curved valley"

$$F(x_1,x_2) = 100(x_2-x_1^2)^2 + (1-x_1)^2.$$

This function defines a narrow parabolic valley as indicated by the contour diagram of Fig. 13.1. Show that

(i) the minimum of the function is given by $F(1,1) = 0$,

(ii) the components of the gradient vector $\underline{g}$ at an arbitrary point (x_1,x_2) are given by

$$g_1 = 400x_1^3 - 400x_1x_2 + 2x_1 - 2, \qquad g_2 = -200x_1^2 + 200x_2,$$

(iii) the elements of the matrix G are given by

$$G_{11} = 1200x_1^2 - 400x_2 + 2, \quad G_{12} = G_{21} = -400x_1, \quad G_{22} = 200.$$

13.2 STEP METHODS

The step methods presented below are more or less empirical and do not have any real theoretical basis. Nevertheless, for many minimization problems simple step methods perform equally well as the better grounded gradient methods.

13.2.1 Grid search and random search

The most elementary minimization procedure consists in mapping the function values in a grid over the entire parameter space and keeping the point with the lowest function value as the best point.

In a one-dimensional grid search one just calculates the function values $F(x)$ at points equally spaced Δx apart. One of these points must then lie within $\frac{1}{2}\Delta x$ from the true minimum, but since the minimum need not be closest to the point with the smallest F-value it is for reasonably smooth functions assumed that this simple grid search in one variable will only localize the minimum to within a distance Δx.

In two variables a grid search locates the minimum within a rectangle $\Delta x_1\Delta x_2$, in three variables within a volume $\Delta x_1\Delta x_2\Delta x_3$, *etc*. With more variables the grid search obviously requires a rapidly increasing number of function evaluations. For example, to localize a minimum to within 1% of the range of one variable by this technique, 100 function evaluations are necessary, while with five variables the number of evaluations required is 10^{10}. Clearly, therefore, a simple-minded grid search should only be used for a rather crude mapping over the parameter space when more than two or three parameters are involved. The best point found

in the crude scan can then be used as a starting point for a more refined minimization method. Indeed, whenever complicated functions are involved an introductory grid mapping is recommended to find good starting values for a subsequent iteration method.

The mapping procedure runs into obvious difficulties if the parameter range is very large (infinite). In this case one can in practice start the mapping with a reasonable interval for x within the allowed region, and later shift the range if the smallest value of the function turns out to be at the boundary of the chosen interval.

The simple grid search can be characterized as a blind or unintelligent method, since it does not take account of what could be learned about the function along the way. Assuming a reasonably smooth function the method certainly involves many redundant function evaluations in regions of the parameter space where the function values are not small. By performing the grid search in more stages, however, the method can be made more efficient.

A multi-stage grid search can be done in the following way: in the first stage a crude grid mapping is made all over the parameter space, confining the minimum to a restricted volume element. In the second stage a new grid search is performed within this volume, limiting the minimum to an even smaller region, and so on.

With many parameters, instead of performing a systematic grid search all over parameter space, good results are often obtained by a Monte Carlo search, choosing points $\underline{x}$ randomly in that region of parameter space where one expects the minimum to be. The Monte Carlo mapping is usually made with a Gaussian random number generator constructing a set of trial points $\underline{x}$ around some first-guess value $\underline{x}_0$ of specified width. The point giving the lowest function value is then retained and can be used as a start point for a more advance minimization technique.

Exercise 13.2: A function of four variables defined within a finite space is to be minimized by a grid search. The minimum should be localized to within one per mille of the range in each variable. Show that in a simple grid search 10^{12} function evaluations are necessary, and that 3.10^4 evaluations are necessary in a three-stage grid search.

13.2.2 Minimum along a line; the success-failure method

In the one-parameter case and in variation methods varying only one parameter at a time the problem is to search for a minimum in one direction. The following prescription for finding the minimum along a line, the *success-failure method*, has been given by H.H. Rosenbrock.

Let $F(x)$ be a function of the variable (parameter) x and suppose that a first guess of the minimum is $x = x_0$. For a given step of length s we evaluate $F(x+s)$ and proceed according to the new function value obtained:

(i) If $F(x+s) \leq F(x)$ call the trial a success, replace x by $x+s$ and s by αs, where the expansion factor α satisfies $\alpha > 1$.

(ii) If $F(x+s) > F(x)$ call the trial a failure, replace s by βs, where the contraction factor β satisfies $-1 < \beta < 0$.

The procedure is repeated until the difference between the function values of the two last successes is less than a specified amount. Good empirical values for the expansion and contraction factors are $\alpha = 3.0$ and $\beta = -0.4$.

It should be noted that when a success is immediately followed by a failure the middle one of the last three points corresponds to a smaller function value than the outer two, and hence a local minimum must lie somewhere between the outer points. This suggests the formulation of an alternative convergence criterion: If the distance Δx between the outer points is smaller than a preassigned value the procedure should be stopped; the best point then corresponds to the last success and the true minimum will be less than a distance Δx from this point.

In this situation when a success is immediately followed by a failure one can, in fact, do even better. Since experience indicates that functions often have an approximate quadratic behaviour near a minimum we predict F^o to coincide with the minimum of the parabola passing through the last three points x_1, x_2 (success), x_3 (failure). Denoting the corresponding function values by F_1, F_2, F_3 the interpolated minimum is at x^o, given by

$$x^o = \tfrac{1}{2} \frac{(F_1-F_2)(x_2^2-x_3^2) + (F_3-F_2)(x_1^2-x_2^2)}{(F_1-F_2)(x_2-x_3) + (F_3-F_2)(x_1-x_2)} . \qquad (13.2)$$

The success-failure method combined with a quadratic interpolation appears in practice to be very efficient and is therefore well suited to find the minimum of a function of only one variable.

Exercise 13.3: Prove eq.(13.2).

Exercise 13.4: Show that an alternative form for the interpolated value x^o of eq.(13.2) is

$$x^o = \tfrac{1}{2} \frac{(F_1-F_2)\alpha(2x_1+2s+\alpha s) + (F_3-F_2)(2x_1+s)}{(F_1-F_2)\alpha + (F_3-F_2)} .$$

Exercise 13.5: The function $F(x) = x^2-x-1$ with the true minimum $F^o=-1.25$ for $x^o = 0.5$ is to be minimized by the success-failure method using the start point $x_0 = 1$, an initial step size $s = 1$, and the constants $\alpha = 3.0$, $\beta = -0.4$. Show that after four function evaluations (three steps) the minimum is bracketed between $x = -0.6$ and $x = +2$ ($\Delta x = 2.6$), and verify that the interpolated minimum occurs at the correct point $x^o = 0.5$. Ignoring the interpolation, show that after seven function evaluations (six steps) the minimum is bracketed between $x = -0.168$ and $x = 1.08$ ($\Delta x = 1.248$).

Exercise 13.6: The results of the first few steps in minimizing some function $F(x)$ according to the success-failure method are given below ($\alpha = 2.0$, $\beta = -0.3$):

Step i	0	1	2	3
x	-1.5	-1	0	2
F(x)	4.75	2	1	17
s	0.5	1.0	2.0	-0.6

Show that the best estimate of the minimum is $x^o = -1/3$.

13.2.3 The coordinate variation method

An intuitively simple method for finding a minimum of a function of several variables, $F(\underline{x}) = F(x_1, x_2, \ldots, x_n)$ is to seek a minimum along each coordinate axis in turn. This is the idea behind the *one-by-one variation method* or the *single-parameter-*, or the *coordinate variation method*, which works as follows:

Starting at a point P_0 one first finds a minimum P_1 parallel to the x_1 axis, for example by the success-failure method. Next, from this minimum one finds a minimum P_2 along the x_2 axis, and so on for all coordinate axes. Of course, when the minimum along x_n has been obtained after one full cycle or stage, the function is no longer at a minimum in the other coordinates. One therefore

starts searching along x_1 again, and completes a new stage. The process is continued until some specified convergence criterion is fulfilled. A reasonable stopping criterion could be that the difference between the best function values in two subsequent full cycles should be less than some prescribed amount.

The one-by-one variation method is illustrated in Figs. 13.2(a) and (b) for two two-dimensional problems involving, respectively, weakly and strongly correlated variables. Curves have been drawn for constant values of $F(x_1,x_2)$. Starting at P_0 the successive minima along the lines parallel to the x_1,x_2 axes are found at $P_1,P_2,\ldots$

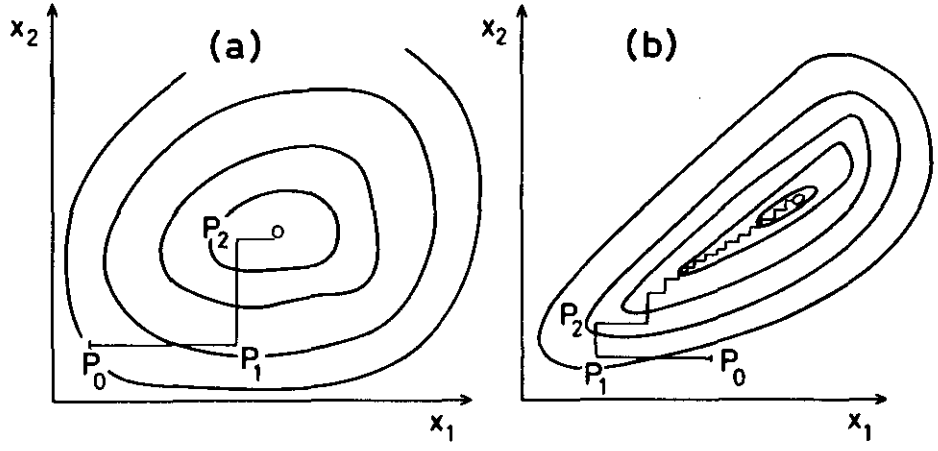

Fig. 13.2. The one-by-one variation method for finding the minimum of a function of two variables; (a) weakly correlated variables, (b) strongly correlated variables.

Although the one-by-one variation method usually does converge, it may require a large number of steps before the convergence is reached. In cases with strongly correlated variables the method is unacceptably slow, since the approach towards the minimum goes by an inefficient zig zag curve crossing the sides of the "valley" to which the minimum belongs; see Fig. 13.2(b).

13.2.4 The Rosenbrock method

Rosenbrock's algorithm is an extension and improvement of the one-by-one variation method and has been found very useful in practice.

The principle of this method can be illustrated for the case of two variables by referring to Fig. 13.3. Starting at the point P_0 the minima P_1 and P_2 are first found by searching along the coordinate axes x_1 and x_2, respectively. The idea is now to search next along a "best" line, defined by the direction from P_0 to P_2. This line corresponds to the direction of the overall improvement in the first stage and is accordingly expected to be a good direction for further search. The coordinate system is therefore rotated so that one axis, x_1 say,

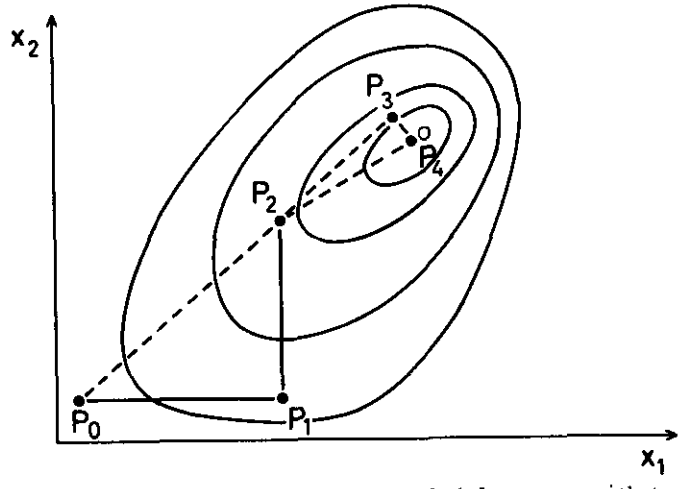

Fig. 13.3. Illustration of the Rosenbrock method for a case with two variables.

points along the "best" line. When a minimum P_3 has subsequently been found along the "best" line a new search is made in a perpendicular direction, giving a minimum P_4. The process is repeated along a new "best" direction defined by the line joining P_2 and P_4, and so on until the convergence criteria are fulfilled.

In n dimensions the Rosenbrock recipe is as follows: Let $\underline{\xi}_i$, $i=1,2,\ldots,n$ denote the set of n orthogonal directions. We find the minimum of $F(\underline{x})$ along each of these directions in turn, starting from the point P_0 and after completing the cycle reaching the point P_1. Let s_i be the size of the step taken to reach the minimum along $\underline{\xi}_i$. The n steps of the first stage can then be characterized

by the n linearly independent vectors

$$\underline{a}_i = \sum_{j=i}^{n} s_j \underline{\xi}_j \,, \qquad i = 1,2,\ldots,n. \qquad (13.3)$$

It is seen that $\underline{a}_1$ is the sum of all steps in the stage, *i.e.* the vector connecting P_0 and P_1, $\underline{a}_2$ is the sum of all steps except the first, *etc.* Obviously the result of all n steps could have been obtained with one single step in the direction $\underline{\eta}_1 = \underline{a}_1/|\underline{a}_1|$, the "best" direction. According to the Gram-Schmidt orthogonalization method we can construct a whole set of n orthogonal directions by defining the other vectors $\underline{\eta}_2,\ldots,\underline{\eta}_n$ as

$$\underline{\eta}_i = \underline{b}_i/|\underline{b}_i| \,, \qquad i = 2,3,\ldots,n, \qquad (13.4)$$

where

$$\underline{b}_i = \underline{a}_i - \sum_{j=1}^{i-1} (\underline{a}_i \underline{\eta}_j)\underline{\eta}_j \,, \qquad i = 2,3,\ldots,n. \qquad (13.5)$$

With this set of orthogonal directions the procedure is repeated for the point P_1, giving P_2, and so on.

The Rosenbrock method usually works well if the number of variables is not too large, but when the number increases its efficiency goes down.

Exercise 13.7: Verify that two different unit vectors $\underline{\eta}_i, \underline{\eta}_j$ defined by eqs.(13.4), (13.5) are orthogonal.

13.2.5 The simplex method

A frequently used step method for minimizing a function of many variables is the *simplex method* invented by J.A. Nelder and N. Mead. The method is based on the evaluation of the function value $F(x_1,x_2,\ldots x_n)$ at n+1 points forming a general simplex*), followed by the replacement of the vertex with the highest function value by a new and better point, if possible. The new point is obtained by a specific algorithm, and leads to a new simplex better adapted to the function.

Let us assume that the points $P_1,P_2,\ldots,P_{n+1}$ defining the current simplex

*) A *simplex* is defined as the simplest n-dimensional geometrical figure specified by giving its n+1 vertices. It is a triangle for n=2, a tetrahedron for n=3, *etc.*

are given. For short we write $F(P_i)$ for the function value at P_i.

We start the procedure by evaluating all $F(P_i)$ and determine the "highest" point P_h and the "lowest" point P_l in the simplex, where

$$\left.\begin{aligned} P_h &\text{ is such that } F(P_h) = \max\{F(P_1),\ldots,F(P_{n+1})\}, \\ P_l &\text{ is such that } F(P_l) = \min\{F(P_1),\ldots,F(P_{n+1})\}. \end{aligned}\right\} \qquad (13.6)$$

Next we define the centroid ("centre-of-mass") $\bar{P}$ of all points in the simplex except P_h,

$$\bar{P} = \frac{1}{n}\left(\sum_{i=1}^{n+1} P_i - P_h\right). \qquad (13.7)$$

The recipe is now to replace P_h by a new point with lower functional value lying on the line through P_l and $\bar{P}$. Three operations can be used, *reflecion, contraction* and *expansion*.

The first trial is done by reflecting P_h about $\bar{P}$, defining a new point P* by the relation

$$P^* = (1+\alpha)\bar{P} - \alpha P_h \,, \qquad (13.8)$$

where the reflection coefficient α is a positive constant. Three situations are possible:

(i) If $F(P^*) < F(P_l)$ the reflection has produced a new minimum. To see if we can do even better we make an expansion along the line and go beyond P* to a new point P**, defined by

$$P^{**} = \gamma P^* + (1-\gamma)\bar{P} \,, \qquad (13.9)$$

where the expansion coefficient γ is greater than unity. If $F(P^{**}) < F(P_l)$ we replace P_h by P** and restart the process. If $F(P^{**}) \geq F(P_l)$ the expansion has failed and we replace P_h by P* before restarting.

(ii) If $F(P_l) \leq F(P^*) < F(P_h)$ the reflection has given a better point than the previous highest value. P_h is replaced by P* and the process is restarted.

(iii) If $F(P^*) \geq F(P_h)$ the reflection has failed and P^* is unacceptable. With a contraction along the line we try a new point P^{**} between P_h and $\bar{P}$, such that

$$P^{**} = \beta P_h + (1-\beta)\bar{P} , \qquad (13.10)$$

where the contraction coefficient β lies between 0 and 1. If $F(P^{**}) < F(P_h)$ P_h is replaced by P^{**} and the process restarted. If $F(P^{**}) > \min\{F(P_h), F(P^*)\}$ (the contracted point is worse than the better of P_h and P^*) the whole attempt has failed. The simplex is then contracted towards the lowest point by replacing all P_i's by $\frac{1}{2}(P_i+P_1)$ before restarting the procedure.

How the simplex method works is illustrated for the two-dimensional case by Fig. 13.4, where contour lines have been drawn for constant values of $F(x_1,x_2)$. A simplex (triangle in this case) with vertices P_1,P_2,P_3 is shown together with the calculated points $\bar{P}$, P^*, P^{**} in one iteration. The point $P_h=P_1$

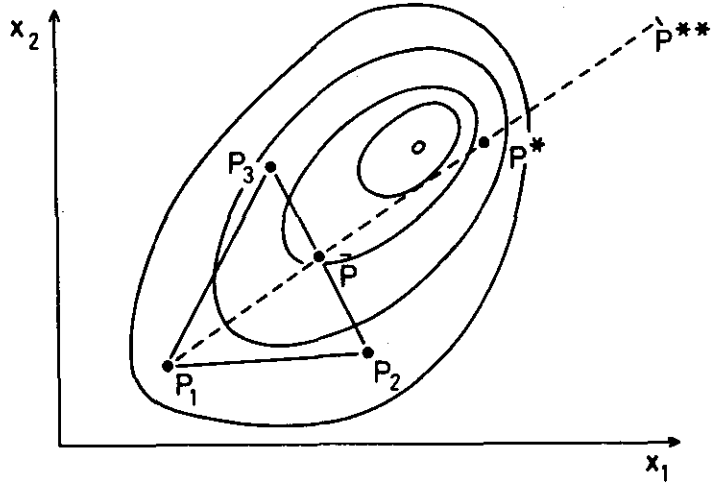

Fig. 13.4. Illustration of the simplex method for a case with two variables.

will in this iteration be replaced by P^* and the next simplex will have the vertices $P_1=P^*,P_2,P_3$.

The procedure with reflections, expansions and contractions runs into difficulties if the new point, P^* or P^{**}, falls too close to $\bar{P}$, since the simplex then collapses into a surface of smaller dimension from which it can never recover. Recommended values for the coefficients α,β and γ are, respectively, 1, 0.5 and 2. Initial values of the simplex may be chosen at random or evaluated from a given point according to a prescribed algorithm. For this method a convenient convergence test is to calculate the difference $F(P_h) - F(P_1)$ for each iteration and to stop when this difference falls below a preset value. Finally, after convergence, $\bar{P}$ and $F(\bar{P})$ are calculated, and the minimum point is taken as P_1 or $\bar{P}$, whichever produces the lowest F-value.

The simplex method does not directly produce the covariance matrix at the estimated minimum. It is, however, a rather effective method: it requires few function evaluations, one or two per iteration, each search is made in an "intelligent" direction pointing from the highest function value to the average of the better values of the simplex; furthermore, as the method is designed to take as large steps as possible it is rather insensitive to shallow local minima or fine structures in the function, implying a generally good adaption to the landscape and a quick contraction to the overall minimum.

Exercise 13.8: The function $F(x_1,x_2) = x_1^2 + x_2^2$ is given together with an initial simplex defined by $P_1= (1,1)$, $P_2= (1,-2)$, $P_3= (-1,0)$. Show, using the simplex minimization procedure with coefficients $\alpha = 1$, $\beta = 0.5$, $\gamma = 2$, that $F(P_h) - F(P_1) < 1$ after three iterations, and that the estimated minimum at this stage is at (3/8, 5/16).

13.3 GRADIENT METHODS

The *gradient methods* utilize the derivatives of $F(\underline{x})$ in the minimization process to predict new trial points relatively far away from the last point. They have a better theoretical foundation than the simple step methods, they are also more complicated, but do not always produce better results.

In this section we will present two gradient methods which both use the first derivatives of $F(\underline{x})$ and which have been widely used in practice, the classical *steepest descent method* invented by Cauchy, and *Davidon's variance algorithm*. A third gradient method, *Newton's method*, which uses also the second derivatives

of $F(\underline{x})$ to calculate the step sizes, was described already in Sect.10.3.1.

13.3.1 Numerical calculation of derivatives

In many cases the analytical expressions for the derivatives of the function are very complicated or can not be found at all. General programmes minimizing by the gradient methods should therefore be supplied with algorithms for determining the derivatives of a function from finite differences.

The gradient $\underline{g}(\underline{x})$ of $F(\underline{x})$ can be estimated numerically at each point $\underline{x}$ of parameter space by performing repeated calculations of $F(\underline{x})$ varying one parameter at a time. For the i-th component of $\underline{g}$, with a positive increment Δx_i, we may take

$$g_i = \frac{\partial F}{\partial x_i} = \frac{F(x_1,\ldots,x_i+\Delta x_i,\ldots,x_n) - F(x_1,\ldots,x_i,\ldots,x_n)}{\Delta x_i} , \qquad (13.11)$$

or, alternatively,

$$g_i = \frac{\partial F}{\partial x_i} = \frac{F(x_1,\ldots,x_i,\ldots,x_n) - F(x_1,\ldots,x_i-\Delta x_i,\ldots,x_n)}{\Delta x_i} . \qquad (13.12)$$

Either alternative requires n+1 function evaluations to give $\underline{g}$. In choosing the size of the increments one should keep in mind the finite precision of the computer and avoid too small Δx_i. A third and better alternative for estimating g_i is to take the average of the expressions above,

$$g_i = \frac{\partial F}{\partial x_i} = \frac{F(x_1,\ldots,x_i+\Delta x_i,\ldots,x_n) - F(x_1,\ldots,x_i-\Delta x_i,\ldots,x_n)}{2\Delta x_i} , \qquad (13.13)$$

which, however, implies 2n evaluations of the function to obtain $\underline{g}$. For functions having a nearly quadratic dependence on the parameters the last formula gives approximately correct derivatives, independent of the size of the increment.

The numerical evaluation of the second derivatives $G_{ij} = \partial^2 F/\partial x_i \partial x_j$ is even more time-consuming. However, the diagonal elements of the matrix G come out as by-products when estimating the gradient by the symmetric method, *i.e.* eq.(13.13), since one may write

$$G_{ii} = \frac{\partial^2 F}{\partial x_i^2} = \frac{F(x_1,\ldots,x_i+\Delta x_i,\ldots,x_n) - F(x_1,\ldots,x_i-\Delta x_i,\ldots,x_n) - 2F(\underline{x})}{(\Delta x_i)^2} . \qquad (13.14)$$

For the off-diagonal elements of G we get, using symmetrical steps,

$$G_{ij} = \frac{\partial^2 F}{\partial x_i \partial x_j} = \big(F(x_i+\Delta x_i, x_j+\Delta x_j) + F(x_i-\Delta x_i, x_j-\Delta x_j) - F(x_i+\Delta x_i, x_j-\Delta x_j) - F(x_i-\Delta x_i, x_j+\Delta x_j)\big)/4\Delta x_i \Delta x_j , \qquad (13.15)$$

where we have simplified notation and only specified the dependence on the x_i, x_j variables. Equation (13.15) shows that four new function evaluations are required for each off-diagonal element. Since there are $n(n-1)/2$ independent off-diagonal elements in a symmetric n x n matrix we see that $2n(n-1)$ extra function evaluations have to be done to estimate G.

If the second derivatives can be assumed approximately constant over small regions near $\underline{x}$, symmetrical steps will not be necessary to find G. The off-diagonal elements may then be calculated from the formula

$$G_{ij} = \frac{\partial^2 F}{\partial x_i \partial x_j} = \big(F(x_i+\Delta x_i, x_j+\Delta x_j) + F(x_i, x_j) - F(x_i+\Delta x_i, x_j) - F(x_i, x_j+\Delta x_j)\big)/\Delta x_i \Delta x_j , \qquad (13.16)$$

rather than from eq.(13.15). With eq.(13.16) only one new function evaluation is needed per off-diagonal element of G in addition to those required for the first derivatives.

Exercise 13.9: Show that the error in the i-th component of the gradient vector $\underline{g}$ as calculated by the unsymmetrical expressions eqs.(13.11), (13.12) to the lowest order in the Taylor expansion is $\frac{1}{2}\Delta x_i(\partial^2 F/\partial x_i^2)$.

Exercise 13.10: Consider the simple function $F(x) = x^2$ with slope 2 in the point $x = 1$. Calculate numerically the gradient at $x = 1$ by the three expressions given in the text, eqs.(13.11),(13.12),(13.13) in turn, using (i) $\Delta x = 1$, (ii) $\Delta x = 0.1$.

Exercise 13.11: Verify the formula eq.(13.14) for the diagonal elements of the matrix G by applying eq.(13.13) twice.

Exercise 13.12: Verify the expressions given for the off-diagonal elements of G, eqs.(13.15),(13.16).

Exercise 13.13: Consider the quadratic function $F(x_1,x_2) = x_1^2+x_1x_2+2x_2^2$ for which the matrix of second derivatives is constant and equal to

$$G = \begin{pmatrix} 2 & 1 \\ 1 & 4 \end{pmatrix} .$$

Show that all numerical expressions for the second derivatives calculated for the point (1,1) with step sizes $\Delta x_1 = \Delta x_2 = 1$ in this case produce exact results.

Exercise 13.14: Given the function $F(x_1,x_2) = x_1^3+x_1x_2+x_1x_2^2+x_2^2$. Calculate the first and second derivatives in the point (1,1) analytically and numerically, using (i) $\Delta x_1 = \Delta x_2 = 1$, (ii) $\Delta x_1 = \Delta x_2 = 0.1$.

13.3.2 Method of steepest descent

The *steepest descent method* can be thought of as a natural improvement of the one-by-one variation method in situations where the derivatives of F(x) are known.

From the starting point P_0 we here seek a minimum of $F(\underline{x})$ along the direction in parameter space where the function decreases most rapidly. When the minimum P_1 in this direction has been found the process is repeated searching a second and better minimum in a direction orthogonal to the first, giving P_2, and so on until a satisfactory convergence has been obtained.

The direction $\underline{\xi}$ of the steepest descent in the point $P_0=\underline{x}_0$ has components

$$\xi_i = \frac{\partial F}{\partial x_i} \Big/ \left(\sum_{j=1}^{n} \left(\frac{\partial F}{\partial x_j} \right)^2 \right)^{\frac{1}{2}} , \qquad i = 1,2,\ldots,n \qquad (13.17)$$

where the derivatives are evaluated at P_0. If the search along the direction $\underline{\xi}$ has lead to the minimum P_1 the new direction $\underline{\eta}$ of steepest descent in the point P_1 is perpendicular to the previous direction $\underline{\xi}$. To see this, introduce a variable s on a line along the direction $\underline{\xi}$ through P_0. All points on this line then satisfy $\underline{x} = \underline{x}_0+s\underline{\xi}$. The minimum point P_1 on the line is defined by the requirement $\partial F/\partial x=0$. With the derivatives evaluated at P_1 we have

$$0 = \frac{\partial F}{\partial s} = \sum_{i=1}^{n} \left(\frac{\partial F}{\partial x_i} \right) \left(\frac{\partial x_i}{\partial s} \right) = \sum_{i=1}^{n} \left(\frac{\partial F}{\partial x_i} \right) \xi_i \propto \underline{\eta} \cdot \underline{\xi} ,$$

which implies that two consecutive directions of search will always be orthogonal.

The steepest descent method is illustrated for a case with two variables in Fig. 13.5. In this situation the method is equivalent to the one-by-one variation method except for a rotation of the coordinate axes; in the general case with more than two dimensions the two methods are not equivalent.

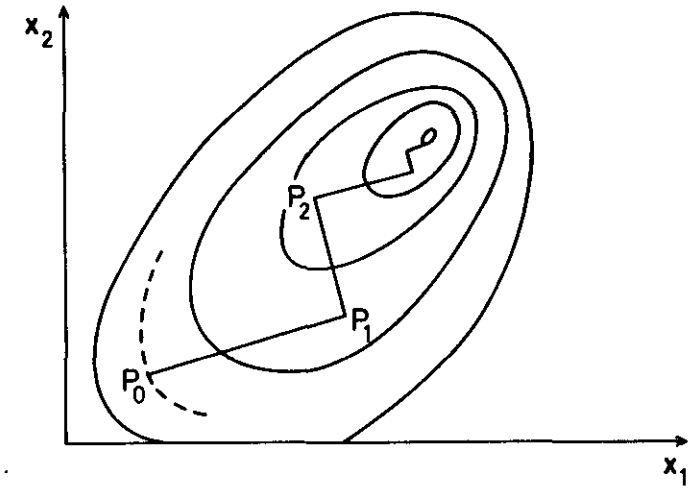

Fig. 13.5. Illustration of the method of steepest descent for a case with two variables.

The method of steepest descent ensures a better convergence than the simple one-by-one variation method. Still it may be rather slow when the variables have a complicated interdependency in F and the choice of starting point P_0 has not been fortunate.

Exercise 13.15: For the "Rosenbrock curved valley" of Exercise 13.1, show how different choices of the starting point P_0 for the steepest descent method will lead to minimizations of different efficiency.

13.3.3 The Davidon variance algorithm

The essential feature of *Davidon's variance algorithm* is that the function is made to approach its minimum by letting the covariance matrix $V(\underline{x}) = G^{-1}$ undergo successive approximations. This means that a simultaneous convergence is obtained towards the function minimum and the true covariance matrix. In this

process the variances are never calculated directly, as must be done with other methods, and thereby one saves the evaluation of second derivatives and the inversion of the resulting matrix. Instead successive approximations are made for $V(\underline{x})$ using only the function $F(\underline{x})$ and its gradient $\underline{g}$.

In the following we merely give the prescriptions of the method without any comments on its theoretical foundation. For complete proofs the reader is referred to the original paper by W.C. Davidon.

Before starting the iterations the method requires the knowledge of an approximate minimum point x^0 with corresponding values F^0, $\underline{g}^0$ known and a first estimate V^0 of the covariance matrix. For the latter it will suffice to have the diagonal elements only, and if these are completely unknown one may simply start with V^0 as the unit matrix; in general, the better the estimate of V^0 the faster the convergence. Initially given is also a set of step constants α, β, which from computational experience may be reasonably taken as $\alpha = 10^{-3}$, $\beta = 10$, and a convergence constant, $\varepsilon = 0.1$, say.

Suppose that an iteration has yielded the values $\underline{x}$, F, $\underline{g}$ and V and that a further iteration is needed to obtain a new set, $\underline{x}^*$, F^*, $\underline{g}^*$, V^*. The algorithm then consists of the following items:

(i) Define $\underline{x}^* = \underline{x} - V\,\underline{g}$, and compute $F^* = F(\underline{x}^*)$ and $\underline{g}^* = \underline{g}(\underline{x}^*)$.

(ii) Define a residual vector $\underline{r} = V\,\underline{g}^*$ and a number $\rho = \underline{g}^* \cdot \underline{r}$ ($\underline{r}$ will vanish if the exact minimum is found, because then $\underline{g}^* = \underline{0}$; ρ is a measure of the perpendicular distance to the minimum).
If $\rho < \varepsilon$, stop. The final estimate for the minimum is then $\underline{x}^0 = \underline{x}^*$.
If $\rho \geq \varepsilon$, proceed to (iii).

(iii) Define $\gamma = -\,\underline{g} \cdot \underline{r}/\rho$.
If $-\frac{\alpha}{1+\alpha} \leq \gamma < \frac{\alpha}{1-\alpha}$, define $\lambda = \alpha$.
If $-\frac{\beta}{\beta+1} \leq \gamma < -\frac{\alpha}{1+\alpha}$, define $\lambda = -\frac{\gamma}{\gamma+1}$.
If $-\frac{\beta}{\beta-1} \leq \gamma < -\frac{\beta}{\beta+1}$, define $\lambda = \beta$,
If none of these three, define $\lambda = \frac{\gamma}{\gamma+1}$.
Define $V^*_{ij} = V_{ij} + (\lambda-1) r_i r_j/\rho$.

(iv) If $F^* < F$, define $\underline{x} = \underline{x}^*$, $F = F^*$, $\underline{g} = \underline{g}^*$, $V = V^*$.
If $F^* \geq F$, define $V = V^*$.
Proceed to (i) for a new iteration.

It should be stressed that this minimization algorithm gives an exact covariance matrix V if $F(\underline{x})$ is a quadratic function.

Experience shows that the method leads to a fast convergence for various types of problems. In fact it can be proven that when $F(\underline{x})$ is of quadratic form in the n parameters the true minimum of the function and the exact covariance matrix are always found within n iterations; see Exercise 13.16.

Exercise 13.16: The function $F(x) = x^2$ is to be minimized by the Davidon variance algorithm from the starting value (approximate minimum) $x^0 = 2$, $V^0 = 1$ for the constants $\alpha = 10^{-3}$, $\beta = 10$, $\varepsilon = 0.1$. Perform the calculations of the steps (i) - (iv) in the text and verify that the method finds the true function minimum and exact variance after only one full stage. Make a sketch of the function and mark the quantities calculated.

13.4 MULTIPLE MINIMA

One of the most basic problems in the field of minimization is that of deciding whether the correct minimum has been found. It has in the previous sections been more or less implicitly assumed that the different iterative operations will eventually lead to just one local minimum of the function, and that this minimum is identical to the desired point. It is, however, not at all evident that the procedure will end up with the global (lowest) minimum of the function, nor is there any guarantee that the minimum found is the nearest to the starting point for the minimization.

In minimizing a function with several minima the task usually belongs to one of the following three categories in descending order of difficulty:

(i) All minima are of interest and should be found. An obvious although rather primitive way to look for all minima is to make a complete mapping of $F(\underline{x})$ over the entire parameter space, but this may imply a large number of function evaluations and be prohibited for economical reasons. Another approach is to perform the minimum search from several starting values. This may, however, lead to arbitrary results unless one has some prior knowledge as to the location of the possible minima. For the general case there seems to be no exhaustive and

practical guide on how to find all minima of a complicated function of many variables.

(ii) Only the global minimum is of interest. Many authors have described procedures to search for the global minimum of a function. It seems, however, as if no safe and simple method has been invented yet. For details on this advanced subject the reader is referred to more specialized literature, for example I.M. Gelfand and M.L. Tsetlin, and A.A. Goldstein and J.F. Price.

(iii) Only one minimum corresponds to a physical solution and is of interest. In fact, in many physics problems approximate values of some of the parameters may often be known in advance. An appealing approach is then first to fix these parameters at their assumed values and minimize the function with respect to the remaining parameters. The parameter set corresponding to the minimum for this simplified problem is next used as a starting point for a complete minimization involving all the parameters.

13.5 EVALUATION OF ERRORS

Implicit in the previous considerations is that our primary interest lies in the particular set of parameter values which minimizes the function $F(\underline{x})$ rather than the function minimum itself. It remains now to discuss how one can evaluate the uncertainties (errors) associated with these parameter values.

With the Davidon variance algorithm the error determination becomes especially simple, since the covariance matrix is obtained directly in the minimization process as part of the algorithm itself. The errors, derived from the diagonal elements of this matrix, are even exact for functions of quadratic form.

For the other minimization procedures the parameter errors can be found for specially constructed F's, using the ideas of Chapters 9 and 10. For instance, for a Least-Squares minimization with one parameter a one-standard deviation confidence interval is derived from the points where the sum of squares is 1.0 above its minimum value F^o. Similarly, in a Maximum-Likelihood estimation of a single parameter the error corresponding to a one-standard deviation confidence interval is determined from the points where the negative log likelihood is 0.5 above minimum. Thus for both methods the computational problem involved in the determination of the errors is to find the parameter values that correspond to the function values $F = F^o + a$, where the constant a is chosen so as to give the desired confidence interval. With the ML method it is important to remember that the confidence intervals obtained in this way have the meaning of likelihood intervals, as discussed for the one-, two- and multi-parameter cases in Sects.9.7.1, 9.7.4 and 9.7.6, respectively. The numerical problem itself is trivial when the function, as is often the case, is approximately quadratic around its minimum. If one suspects that the minimum region is not sufficiently quadratic the determination of the parameter values that make $F = F^o + a$ is not simple.

A special technique for the handling of ill-behaved functions, used for example in the MINUIT minimization programme, can be sketched as follows:

Suppose that a crude estimate of the covariance matrix exists and that the minimum value F^o of the non-parabolic function $F(\underline{x})$ has been located at $\underline{x}^o$. We want to find the error $\sigma(x_i)$ in the parameter value x_i^o corresponding to an increase in F by the amount a from the minimum when only the parameter x_i is varied and all the others are kept fixed at their values at the minimum, $x_1^o,\ldots,x_{i-1}^o,x_{i+1}^o,\ldots,x_n^o$.

Let us refer to Fig. 13.6, where the minimum point of the function is called A. If the crudely estimated covariance matrix implies an error σ_0 in x_i^o, the parabola $F(x_i) = F^o + a\sigma_0^{-2}(x_i-x_i^o)^2$ intersects the straight line $F = F^o + a$ at the point B. The function value for $x_i = B$ corresponds to a lower point B'. A second parabola with minimum at A, passing through B' intersects the straight line at the point C, with a corresponding function value C', above the line. A third parabola through the points A, B', C' gives an intersection with the straight line at D and a new point D' on the curve, which also is below the line. A new parabola through B', C', D' gives an intersection with the line at E, and this time the function value E' coincides with E within some given tolerance. Hence the required point has been found and $\sigma(x_i)$ can be determined.

This method can handle rather pathological functions, but will then be time-consuming.

13.6 MINIMIZATION WITH CONSTRAINTS

The parameters of the function to be minimized are frequently restricted to a limited region in parameter space through constraint equations or inequalities due to physics requirements. In our discussion of minimization so far the

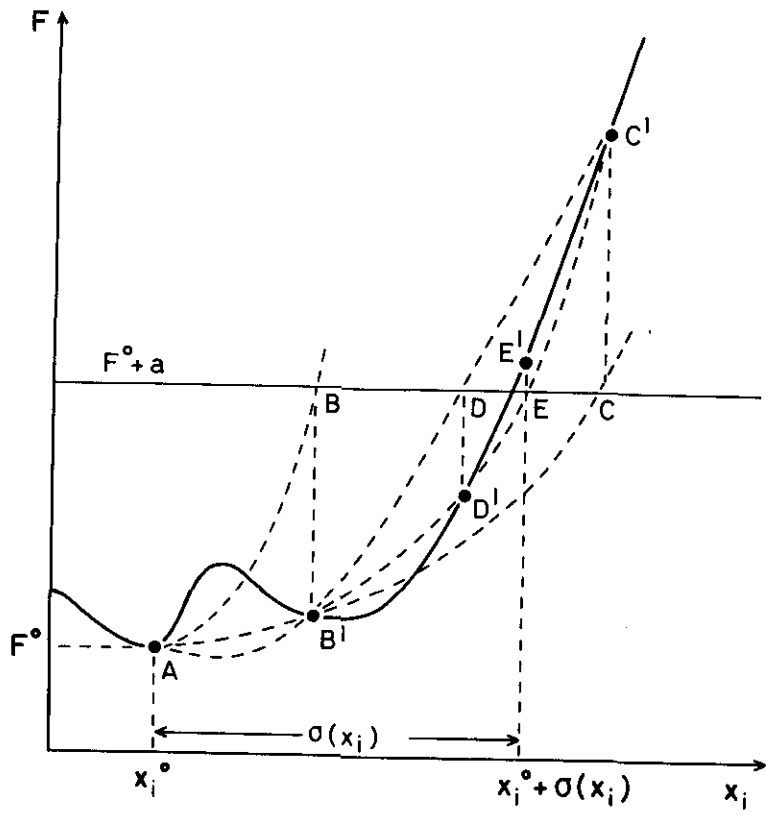

Fig. 13.6. Error determination for an ill-behaved function (see text).

variables have been assumed to be without restrictions. We will now see how different types of constrained problems may be handled.

The general problem in this section is to minimize the function $F(\underline{x})$ subject to one or more of the following conditions:

$$\text{(i)} \quad \begin{cases} f_1(\underline{x}) = 0 \\ f_2(\underline{x}) = 0 \\ \vdots \end{cases} \qquad \text{constraint equations}$$

$$\text{(ii)} \quad \begin{cases} a_1 \le x_1 \le b_1 \\ a_2 \le x_2 \le b_2 \\ \vdots \end{cases} \qquad \text{simple constant limits}$$

$$\text{(iii)} \quad \begin{cases} l_1(\underline{x}) \le x_1 \le h_1(\underline{x}) \\ l_2(\underline{x}) \le x_2 \le h_2(\underline{x}) \\ \vdots \end{cases} \qquad \text{simple variable limits}$$

$$\text{(iv)} \quad \begin{cases} u_1(\underline{x}) \le v_1(\underline{x}) \le w_1(\underline{x}) \\ u_2(\underline{x}) \le v_2(\underline{x}) \le w_2(\underline{x}) \\ \vdots \end{cases} \qquad \text{implicit variable limits}$$

In (i) - (iv), a_i, b_i, $i=1,2,\ldots,n$ are constants while all functions depend on $\underline{x}$. The general conditions of type (iv) actually include all conditions of the types (ii) and (iii), but for practical and illustrative purposes they may be separated as above.

In the special field of *linear programming*, where the function $F(\underline{x})$ and its constraints are linear in the parameters, the ideas discussed in the following are not of practical use. When $F(\underline{x})$ is linear it is the constraints that make it possible for the function to take a minimum value at a finite point at the boundary of the allowed region. Thus, in the field of linear programming the constraints are essential to obtain a minimum, whereas in our problems the constraints are more or less regarded as nuisance.

Constraints can be taken care of by applying special procedures for

constrained minimization. We will, however, only consider techniques for modifying the function $F(\underline{x})$ in such a way that ordinary, unconstrained, minimization procedures may be applied. Well-known techniques are

- elimination of variables using the constraint equations,
- introduction of Lagrangian multipliers,
- change of variables to eliminate constraints,
- introduction of penalty functions.

In particular, when $F(\underline{x})$ is to be minimized under k constraint equations $f_i(\underline{x}) = 0$, $i=1,2,\ldots,k$, corresponding to case (i) above, one may use the constraint equations to eliminate variables in $F(\underline{x})$. An example on the elimination approach was given in Sect.10.7.1. Alternatively one can, as discussed in Chapter 10, introduce Lagrangian multipliers $\underline{\lambda} = \{\lambda_1,\lambda_2,\ldots,\lambda_k\}$ and construct a modified function $T(\underline{x},\underline{\lambda})$, where

$$T(\underline{x},\underline{\lambda}) = F(\underline{x}) + \sum_{i=1}^{k} \lambda_i f_i(\underline{x}) \qquad (13.18)$$

which is then minimized with respect to $\underline{x}$ and $\underline{\lambda}$. The method with the Lagrangian multipliers obviously suffers from the drawback of an increased number of parameters in the minimization, but has otherwise virtues of its own.

In the following sections we will see how the two techniques involving change of variables and penalty functions can be applied to the situations (ii) - (iv) above.

13.6.1 Elimination of constraints by change of variables

An important practical case of constrained minimization occurs when some or all variables are bounded by constant limits, $a_i \le x_i \le b_i$. In a minimization by a grid search such inequality constraints are easily taken care of by limiting the search region. With the other minimization procedures one may eliminate the explicit inequality constraints by changing to a new set of unrestricted variables. To find the minima one simply expresses $F(\underline{x})$ by the new variables $\underline{y}$ through a straightforward substitution and minimizes with respect to $\underline{y}$. The minima obtained in $\underline{y}$-space are then transformed back to $\underline{x}$-space. The errors in $\underline{x}^0$ must be found from an ordinary computation of error propagation.

If, for instance, x_i has to be non-negative, $0 \le x_i \le \infty$, either of the two following variable changes can be applied,

$$x_i = y_i^2 \qquad \text{or} \qquad x_i = e^{y_i}. \qquad (13.19)$$

For problems which require $0 \le x_i \le 1$ transformations like

$$x_i = \sin^2 y_i \qquad \text{or} \qquad x_i = \frac{e^{y_i}}{e^{y_i} - e^{-y_i}} \qquad (13.20)$$

will remove the constraint. With general constant limits $a_i \le x_i \le b_i$ an often used transformation is

$$x_i = a_i + (b_i - a_i)\sin^2 y_i. \qquad (13.21)$$

The variable changes above will not introduce new minima in $\underline{x}$. The transformations involving $\sin^2 y_i$ actually produce for each minimum in $\underline{x}$-space many equally-spaced minima in $\underline{y}$-space. This should, however, not cause difficulties provided that the minimization procedure used does not involve so long steps that intermediate hills are crossed.

Exercise 13.17: A function $F(x_1,x_2,x_3)$ is to be minimized under the constraints

$$0 \le x_1 \le \infty,$$
$$0 \le x_2 \le 1.$$

With the substitutions

$$x_1 = e^{y_1}, \qquad x_2 = \sin^2 y_2, \qquad x_3 = y_3,$$

verify that the minimum can be obtained by an unconstrained minimization in $\underline{y}$-space followed by the transformation of the minimum point back to $\underline{x}$-space. If the covariance of the minimum point in $\underline{y}$-space is $V(\underline{y})$, show that the covariance of the minimum in $\underline{x}$-space is $V(\underline{x}) = SV(\underline{y})S^T$, where

$$S = \begin{pmatrix} e^{y_1} & 0 & 0 \\ 0 & \sin 2y_2 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Exercise 13.18: The minimum of a function $F(x_1,x_2)$ subject to the constraint

$$0 \le x_1 \le x_2 \le \infty$$

can be obtained by an unconstrained minimization with direct substitution of

$$x_1 = y_1^2 , \qquad x_2 = y_1^2 + y_2^2 .$$

Consider the explicit function $F(x_1,x_2) = -x_1^2 - x_1x_2 + x_2^2$. Show analytically that the minimum is at $x_1 = x_2 = 0$.

13.6.2 Penalty functions; Carroll's response surface technique

When one has simple variable limits, $l_i(\underline{x}) \le x_i \le h_i(\underline{x})$, or implicit variable limits, $u_i(\underline{x}) \le v_i(\underline{x}) \le w_i(\underline{x})$, it may be difficult to find transformations which will eliminate the constraints. However, it is possible to modify the function $F(\underline{x})$ by adding a so-called *penalty function* to it so that the function value becomes large at the boundary of the allowed region. In this way one "fools" the minimization procedure to search only inside the allowed region provided that a permissible starting point was chosen. The penalty function should have an effect just on the boundary, to ensure that no new minima are introduced and that no minima are lost.

To be specific, consider the minimization of $F(\underline{x})$ with the constraint $p(\underline{x}) \ge 0$. We can then define a modified function $T(\underline{x})$ such that

$$\left.\begin{aligned} T(\underline{x}) &= F(\underline{x}) && \text{if} \quad p(\underline{x}) \ge 0, \\ T(\underline{x}) &= F(\underline{x}) + cp^2(\underline{x}) && \text{if} \quad p(\underline{x}) < 0, \end{aligned}\right\} \tag{13.22}$$

and minimize $T(\underline{x})$ as before. The constant c should be chosen large compared to $F(\underline{x})$, and $F(\underline{x})$ should be continuous at the boundary. This method can easily be extended to more constraints.

One sees from the example above a useful feature of introducing the "penalty" $cp^2(\underline{x})$ in $T(\underline{x})$. Without the penalty function the constraints $p(\underline{x}) \ge 0$ could be used as a test criterion during the minimization of $F(\underline{x})$, but then a point for which $p(\underline{x}) < 0$ would supply no information about the direction for the continued search. When the "penalty" is incorporated in $T(\underline{x})$ the knowledge gained on the gradient will help speed up the return to the allowed region of parameter space.

Another popular way of introducing penalty functions is by *Carroll's response surface technique*. Suppose the constraints are redefined to the form $p_i(\underline{x}) \ge 0$, for i=1,2,...,m. The modified function $T(\underline{x},r)$ to be minimized is then defined by

$$T(\underline{x},r) = F(\underline{x}) + r\sum_{i=1}^{m} c_i/p_i(\underline{x}) , \tag{13.23}$$

where the sum represents the "penalty". The c_i's are positive weights of the individual "penalties", while $r(\ge 0)$ weights the sum relatively to $F(\underline{x})$. The minimization is then handled as a succession of unconstrained minimizations for different values of r, starting with a (finite) r>0 and reducing it stepwize to zero.

Exercise 13.19: The function F(x) is to be minimized within the region $a \le x \le b$ using Carroll's response surface technique. Show that the function T(x,r) of eq.(13.23) with equal weights c_i can be written

$$T(x,r) = F(x) + r\left(\frac{1}{x-a} + \frac{1}{b-x}\right) .$$

13.6.3 Example: Determination of resonance production

An effective mass spectrum of the $\pi^+\pi^-$ system in the reaction $\pi^-p \to \pi^-p\pi^+\pi^-$ shows enhancements corresponding to the production of $\rho(765)$ and $f(1260)$. The experimental spectrum is to be compared with the theoretical distribution

$$f(M,\underline{\alpha}) = \alpha_\rho BW_\rho(M) + \alpha_f BW_f(M) + \alpha_b B(M),$$

where $BW_\rho(M)$, $BW_f(M)$ are Breit-Wigner functions describing the resonances and B(M) a background term, all of which are normalized and known. The production fractions $\alpha_\rho,\alpha_f,\alpha_b$ are the unknown parameters, which must satisfy the following constraints:

$$(1) \quad \begin{cases} 0 \le \alpha_\rho \le 1 \\ 0 \le \alpha_f \le 1 \\ 0 \le \alpha_b \le 1 \\ \alpha_\rho + \alpha_f + \alpha_b = 1. \end{cases}$$

The last (normalization) condition can be used to eliminate α_b from $f(M,\underline{\alpha})$, leaving two unknown parameters α_ρ and α_f. If now the fractions α_ρ,α_f are to be estimated, for example by the Least-Squares method, this corresponds to minimizing a function $F(\alpha_\rho,\alpha_f)$ of summed squared deviations, subject to the constraints

$$(2)\quad \begin{cases} 0 \le \alpha_\rho \le 1 \\ 0 \le \alpha_f \le 1 \\ 0 \le 1 - \alpha_\rho - \alpha_f \le 1. \end{cases}$$

To make use of an unconstrained minimization procedure with Carroll's response surface technique we must rewrite the constraints (2) to the form $p_i(\underline{x}) \ge 0$, $i=1,2,\ldots,m$. The equivalent set of constraints is

$$(3)\quad \begin{cases} \alpha_\rho \ge 0 \\ 1 - \alpha_\rho \ge 0 \\ \alpha_f \ge 0 \\ 1 - \alpha_f \ge 0 \\ 1 - \alpha_\rho - \alpha_f \ge 0. \end{cases}$$

One constraint, $1 - \alpha_\rho - \alpha_f \ge 1$, in (2) is implicitly given by the others, leaving five constraints in (3). Thus the modified function of eq.(13.23) with equal weights c_i can be written

$$T(\underline{\alpha},r) = F(\alpha_\rho,\alpha_f) + r\left(\frac{1}{\alpha_\rho} + \frac{1}{1-\alpha_\rho} + \frac{1}{\alpha_f} + \frac{1}{1-\alpha_f} + \frac{1}{1-\alpha_\rho-\alpha_f}\right).$$

From this example it is realized that, depending on which variable is eliminated from the normalization condition, somewhat different estimates may be found for the production fractions. The differences will, however, be reflected by the errors connected to the estimates.

13.7 CONCLUDING REMARKS

The merit of a minimization method is often judged on its ability to handle problems of large complexity. The sophisticated methods designed to search in "intelligent" directions with variable step sizes may in this sense be considered superior to the more simple-minded step methods. However, which of the many minimization procedures is the best for practical use, depends to a large extent on the actual function to be minimized. It is therefore difficult to give specific recommendations. With simple functions of only a few parameters the simple methods are often found to produce satisfactory results. With complicated functions involving many parameters the more refined techniques are necessary, and

the choice of procedure should be made also considering where in the region of parameter space the minimization is to be started. Quite often a method which works well in the distant regions is less efficient when approaching the minimum. This suggests that the different methods be applied in sequence, depending on the convergence reached. Indeed, the general purpose programmes available incorporate several minimization techniques and allow the user to change from one method to another in one run.

14. Hypothesis testing

14.1 INTRODUCTORY REMARKS

In the preceding chapters main emphasis has been on the area of statistical inference which deals with parameter estimation, that is, point estimation and interval estimation of the parameters of a distribution. This chapter will be devoted to another domain of statistical inference, namely that of testing statistical hypotheses. This subject is closely related to parameter point estimation and to the determination of confidence intervals. While the estimation problem generally amounts to seeking estimates of unknown parameters, our concern under hypothesis testing will be to decide whether, from a statistical point of view, some given mathematical model with pre-assigned or estimated values of the parameters is acceptable in light of the observations.

Suppose for instance that we have measured the proper decay time of a sample of Ξ^0 hyperons and from these measurements estimated the Ξ^0 mean lifetime. We may then ask the question: Does this estimate agree with the prediction of the $\Delta I=\frac{1}{2}$ rule that the Ξ^0 lives twice as long as the Ξ^-? In different words, do the observations throw doubt upon or even disprove the validity of the $\Delta I=\frac{1}{2}$ rule?

Instead of giving answers "yes" or "no" to questions of the above type it is customary to rephrase the problem in terms of *a test of a statistical hypothesis* which follows as a consequence of the assumed physical law. In the situation above let τ_o be the mean lifetime of the Ξ^0 as implied by the $\Delta I=\frac{1}{2}$ rule and the known Ξ^- lifetime, and let τ be the value indicated by an experiment. We may then explicitly want to test if τ is equal to τ_o within the experimental errors; we write

$$H_o: \ \tau = \tau_o$$

and call H_o *the null hypothesis*. The other possibility, τ is different from τ_o, can be called *the alternative hypothesis* to H_o, and is written

$$H_1: \ \tau \neq \tau_o.$$

When the experiment has been performed the actually observed lifetime τ_{obs} should enable us to express a probability or "likelihood" that the null hypothesis H_o is correct. Whether the null hypothesis should be rejected as false will in general depend on the alternative hypothesis to which it is compared.

It will be our concern in this chapter to discuss how various hypotheses can be put to test. If the hypothesis under test involves only specified values of certain parameters, as H_o in the case considered above, the test is called a *parametric test*. The formulation of such tests follows closely the lines of thought for the construction of confidence intervals for unknown parameters as discussed in Chapter 7.

Non-parametric tests deal with questions like: Is an experimental distribution of a specific shape? Or, are two given experimental distributions of the same form? In the former case the null hypothesis can be that the observations come from, say, a normal population while the alternative hypothesis can be that they originate from any other parent distribution; for this kind of problem one can formulate *goodness-of-fit tests* for H_o. In the comparison of experimental distributions the null hypothesis need not make any specific assumption about the shapes of the parental distributions, except that these are equal. If the test formulation is made independent of the form of the underlying distributions, and therefore is valid for all distributional forms, it is called *a distribution-free test*.

Our discussion in the following will be limited to a rather elementary survey of the general principles involved, which should, however, be sufficient to tackle many practical problems met in the everyday life of an experimental physicist.

It should be stressed from the outset that an experimentalist will often, on the basis of his own observations only, not be able to prove or disprove a fundamental theoretical idea or hypothesis which motivated his experiment. This is particularly true in particle physics, which commonly calls for higher statistics than can be obtained in a single experiment. In essence the individual experiment determines the probability to obtain the observed result assuming the hypothesis to be true, and quite often the experimentalist will have to leave the problem at this stage. Further measurements or other types of

experiments may have to be awaited before a final statement or decision can be made. For instance, repeated experiments may be necessary to get a sufficiently low overall probability to justify the rejection of some theoretical hypothesis. The decision is not taken until it has been demonstrated with overwhelming plausibility that the theoretical idea was wrong.

This is a somewhat different situation compared to the practice in industry, insurance companies *etc.*, where quite often decisions have to be taken even though the risk of making the wrong decision can be rather high. High-risk decisions also occur in particle physics, but then on a more elementary level. For example, for an individual particle reaction a selection is made between the kinematically permitted hypotheses on the basis of the probabilities of the various fits. The physicist, however, will usually have the possibility to correct his final sample of accepted events for the possible "background" of incorrectly assigned kinematical hypotheses. Therefore, although the decisions about the individual events are taken with a rather large risk of being wrong, one tries in the end to correct for the wrong decisions made.

14.2 OUTLINE OF GENERAL METHODS

To introduce the concepts involved in hypothesis testing suppose that we have two hypotheses completely specified by two different values of a parameter θ entering a p.d.f. $f(x|\theta)$ *). The null hypothesis H_o assumes $\theta = \theta_o$, and the alternative hypothesis H_1 assumes $\theta = \theta_1$. When a hypothesis is completely specified, as both H_o and H_1 in this case, it is called a *simple* hypothesis. A hypothesis is called *composite* if any parameter is not specified completely. For instance, an alternative H_1: $\theta > \theta_1$ is a composite hypothesis.

We shall now define criteria for when to accept the null hypothesis H_o (and reject H_1), and when to accept the alternative hypothesis H_1 (rejecting H_o), on the basis of a given observation x_{obs}.

Assuming the null hypothesis H_o to be true we can find a region R in the sample space W for the observation x such that the probability that x belongs to R is equal to any preassigned numerical value. The region R is called the

*) In this section x can be a directly observable quantity or more generally some function of the observables, corresponding to the notion of a *statistic* (compare Chapters 7 and 8).

region of rejection or the *critical region* for H_o, while (W-R) is called the *acceptance region* for H_o. The two regions are separated by the *critical value* x_c; see the illustration in Fig. 14.1. The implication is that if the observed value x_{obs} falls in R (*i.e.* x_{obs} exceeds x_c in Fig. 14.1) we shall reject H_o, otherwise we shall accept it.

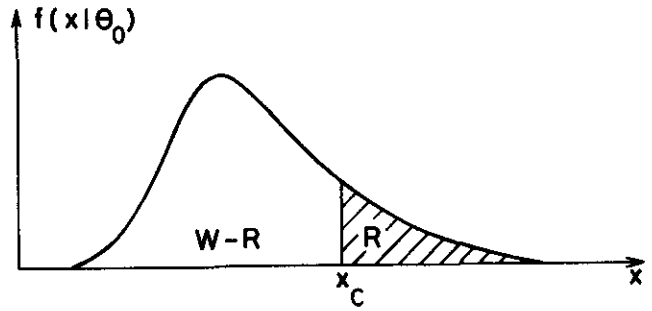

Fig. 14.1. Illustration of critical region R (rejection region) and acceptance region W-R for the test statistic x.

The preassigned probability α that the observation x will belong to the region R is called the *significance*, or the *size of the test*, and determines the *significance level* at 100α %.

From this definition there is obviously a probability α that the observed value x_{obs} will fall into R also when H_o is true. Therefore, in 100α % of all decisions H_o will be rejected when it should, in fact, have been accepted. The mistake we do by rejecting H_o when it is true is called a *Type I error*, or an *error of the first kind*. Since we want to commit such an error as rarely as possible a low numerical value should be taken for α.

There is, however, another possible mistake which can occur, namely that we accept H_o as true when it is, in fact, false. This is called a *Type II error*, or an *error of the second kind*; the probability of its occurrence β depends on the alternative hypothesis H_1. With reference to the illustrations in Fig. 14.2 we can in an obvious notation state the above definitions as follows,

$$\text{Prob(Type I error)} = \alpha = \int_R f(x|\theta_o)dx = \int_{x_c}^{\infty} f(x|\theta_o)dx\,, \tag{14.1}$$

$$\text{Prob(Type II error)} = \beta = \int_{W-R} f(x|\theta_1)dx = \int_{-\infty}^{x_c} f(x|\theta_1)dx\,. \tag{14.2}$$

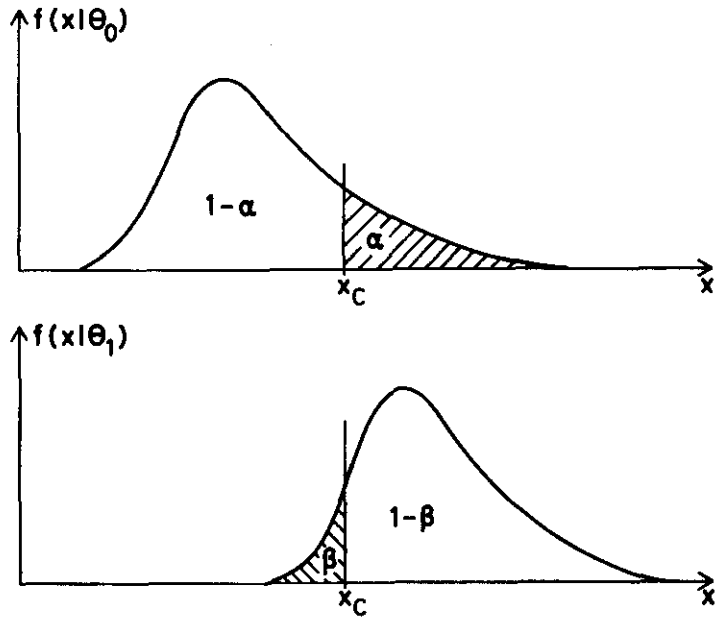

Fig. 14.2. Illustration of Type I error α and Type II error β.

The power of a test is defined as the probability of rejecting a hypothesis when it is false. We have for the power of the test of the null hypothesis H_o against the alternative H_1:

$$\text{Power} = 1-\beta = \int_R f(x|\theta_1)dx = \int_{x_c}^{\infty} f(x|\theta_1)dx\,. \tag{14.3}$$

From the previous considerations it seems reasonable to choose the critical region R such that, for a specified significance level, the power gets as high as possible. This vague statement about the choice of *the best critical region* can be given a more quantitative form in terms of the Neyman-Pearson test and the likelihood-ratio test to be discussed later. Before we proceed to discuss these explicit tests we will illustrate with a specific example from particle physics some of the concepts introduced above.

14.2.1 Example: Separation of one-π^o and multi-π^o events

A wellknown situation involving decisions between different hypotheses occurs in the identification of particle reactions based on kinematical fitting of measured events.

Let us consider antiproton-proton annihilations into four-pronged events in a hydrogen bubble chamber. A large fraction of the events do not satisfy the kinematical constraints imposed by the energy and momentum conservation laws if one assumes that only four charged pions were created in the annihilation process. In these events probably one or more neutral particles were produced, and the missing-mass as deduced from the measured visible particles (tracks) is consistent with the production of at least one π^o. From the measured missing-mass values one wants to separate these events into two groups, one sample consisting of events where there is a single π^o (the "1C-channel"), the second sample consisting of events in which two or more π^o's were produced (the "no-fit channel").

In the language of hypothesis testing the problem is phrased as follows: The individual events are to be put to test for the null hypothesis

$$H_o\colon\ \bar{p}p \to \pi^+\pi^+\pi^-\pi^-\pi^o$$

against the alternative hypothesis

$$H_1\colon\ \bar{p}p \to \pi^+\pi^+\pi^-\pi^-M\,,$$

(here M denotes "more than one neutral pion"), and the missing-mass squared m^2 is to be used as a test statistic. The critical value m_c^2 is most reasonably taken somewhere beyond the squared π^o mass, corresponding to a one-sided test

for H_o. If the missing-mass squared for an event comes out smaller than m_c^2 the hypothesis H_o is accepted; we then call the event a one-π^o event. If $m^2 > m_c^2$ the hypothesis H_o is rejected and the alternative H_1 accepted, corresponding to a multi-π^o event.

In choosing the critical value m_c^2 we are faced with a dilemma. Suppose, for example, that the observed missing-mass spectrum for the total event sample has the shape of Fig. 14.3(a). If the critical value is set at a low mass, not much larger than $m_{\pi^o}^2$, we shall be ensured that relatively few true multi-π^o events will wrongly be taken for one-π^o events, but at the same time many true one-π^o events will then be lost from the final one-π^o sample. On the other hand, if the critical value is taken at a high mass, the loss of true one-π^o events will be small, but the contamination of multi-π^o events in the one-π^o sample will with this choice be considerable. Clearly, the decision of a critical value requires a two-way consideration of the conflicting demands of a smallest possible loss of true one-π^o events and a smallest possible contamination of multi-π^o events in the final one-π^o sample.

This example illustrates the general situation when it is desired to minimize the probability α for rejecting H_o when it is true, *i.e.* commit a Type I error (corresponding to losing true one-π^o events) and simultaneously minimize the probability β for accepting H_o when it really is false, *i.e.* commit a Type II error (contamination of multi-π^o events). To keep the significance α at a low value and at the same time have an optimum of the power $(1-\beta)$ of the test a compromise is called for.

With reference to Fig. 14.3(b) suppose that we know the distribution $f(m^2|H_o)$ in the variable m^2 for true one-π^o events and the corresponding distribution $f(m^2|H_1)$ for the true multi-π^o events. For any chosen significance α we can then, by numerical integration of $f(m^2|H_o)$ according to eq.(14.1), determine the critical value m_c^2 and subsequently, by integration of $f(m^2|H_1)$, determine β from eq.(14.2). The result of this computation for the power considered as a function of the significance is that $(1-\beta)$ from the value zero increases very rapidly with increasing values of α up to about 0.1. For any significance greater than, say, 0.15 the power is practically equal to 1. Thus, depending on whether it is the purity or the size of the one-π^o sample which is of greatest

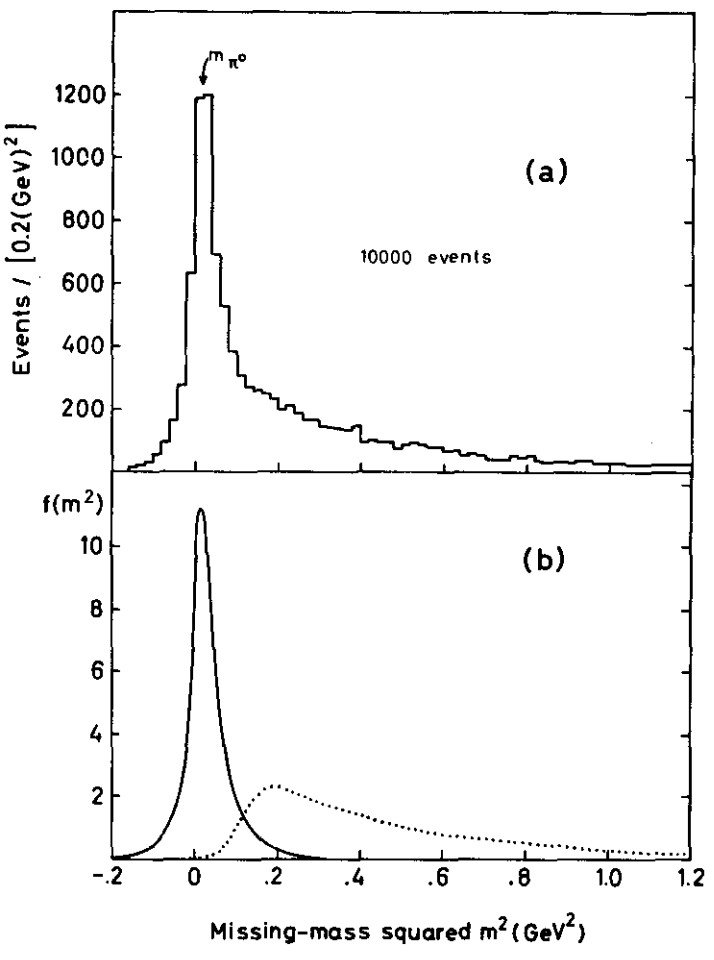

Fig. 14.3. (a) Distribution of the missing-mass squared m^2 in $\bar{p}p$ annihilations into four charged pions plus at least one neutral pion; data from an experiment with 1.2 GeV/c antiprotons incident in a hydrogen bubble chamber. (b) Probability density functions in m^2 for one-π^o events (full curve) and multi-π^o events (dotted curve).

importance one would in this case probably fix the significance level somewhere between 2 and 10%, corresponding to taking the critical value m_c^2 in the region 0.2 to 0.1 GeV^2.

For completeness it should be mentioned that in the present example a better test statistic exists for the test of the hypothesis H_o against the alternative H_1. When the uncertainties on the measured angles and momenta are known a Least-Squares minimization can be performed as described in Sects.10.8, 10.8.1 with contraints from energy and momentum conservation assuming an unmeasured π^o. The X^2 function then contains more information than the missing-mass alone, and as X^2_{min} will be an approximate chi-square variable with one degree of freedom, it can be used as a test statistic for a goodness-of-fit test; see also Sect.14.4.5.

14.2.2 The Neyman-Pearson test for simple hypotheses

The Neyman-Pearson lemma states that, for a fixed significance level, the best critical region R should include those values of x for which $f(x|\theta_1)$ is as large as possible relative to $f(x|\theta_o)$.

It can easily be verified that this choice of R maximizes the power of the test H_o: $\theta = \theta_o$ against H_1: $\theta = \theta_1$. Consider one measurement x and define the significance α of the test by the integral of eq.(14.1). The region R should obviously include all points where $f(x|\theta_o) = 0$ and $f(x|\theta_1) > 0$, since these points do not contribute to α. For $f(x|\theta_o) > 0$ the power of the test of H_o against H_1 may be written (from eq.(14.3))

$$1-\beta = \int_R f(x|\theta_1)dx = \int_R \left(\frac{f(x|\theta_1)}{f(x|\theta_o)}\right) f(x|\theta_o)dx = \left.\frac{f(x|\theta_1)}{f(x|\theta_o)}\right|_{x=\xi} \int_R f(x|\theta_o)dx, \quad (14.4)$$

where ξ is a point within R. Since the last integral is nothing but the constant α, the power of H_o will be maximal if the region R is chosen such that the ratio $f(x|\theta_1)/f(x|\theta_o)$ is as large as possible. The best critical region therefore consists of points satisfying the inequality

$$\frac{f(x|\theta_1)}{f(x|\theta_o)} > k, \quad (14.5)$$

where k is determined by the preassigned significance α.

For the case with a series of measurements $x_1, x_2, \ldots, x_n$ the p.d.f. $f(x|\theta)$ is replaced by the joint distribution function, the likelihood function

$$L(\underline{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta).$$

The criterion (14.5) now becomes

$$\frac{L(\underline{x}|\theta_1)}{L(\underline{x}|\theta_o)} > k \quad (14.6)$$

with the condition

$$\int_R L(\underline{x}|\theta_o)d\underline{x} = \alpha \quad (14.7)$$

which replaces eq.(14.1).

The critical region constructed in accordance with the rules eqs. (14.6), (14.7) will provide the maximum power of a simple null hypothesis against a simple alternative hypothesis for the given significance level. Equivalently, this region minimizes the probability of Type II errors.

For the case with many measurements the critical region R in $\underline{x}$-space may be difficult to find since eq.(14.7) constitutes an n-dimensional integral. In practice one will seek the critical region for some test statistic, expressed as a function of the x_i. For instance, it will be convenient to seek the critical region for the sample mean $\bar{x}$ when testing on a population mean μ, or the sample variance s^2 when the test involves the population variance σ^2. In any case, to determine the critical region one will have to integrate over the probability density function for the test statistic actually used.

It should be noted that the Neyman-Pearson test is applicable only in testing *simple* hypotheses. For composite hypotheses one can only rarely find a test which is more powerful than any other test. The latter type of problem will be discussed in Sect.14.2.4 in connection with the likelihood-ratio test.

14.2.3 Example: Neyman-Pearson test on the Ξ^o mean lifetime

For the Ξ^o lifetime problem introduced in Sect.14.1, consider the two following simple hypotheses about the mean value τ, both with the functional

form $f(t|\tau) = \frac{1}{\tau}\exp(-t/\tau)$, (to get simple relations the lifetime is here expressed in units of the theoretical prediction from the $\Delta I=\frac{1}{2}$ rule),

$$H_o\colon\ \tau = 1,$$

$$H_1\colon\ \tau = 2.$$

The condition given by eq.(14.6) now reads, assuming n observations,

$$\frac{L(\underline{t}|\tau=2)}{L(\underline{t}|\tau=1)} = \frac{\prod_{i=1}^{n} \frac{1}{2}\exp(-t_i/2)}{\prod_{i=1}^{n}\exp(-t_i)} = (\tfrac{1}{2})^n \exp(\tfrac{1}{2}\sum_{i=1}^{n} t_i) > k.$$

In terms of $\bar{t} = \frac{1}{n}\sum_{i=1}^{n} t_i$, this can be written,

$$\bar{t} > 2\,(\frac{1}{n}\ln k + \ln 2) \equiv T_n. \qquad (14.8)$$

The best critical region in $\bar{t}$ space is therefore the set of values satisfying the inequality (14.8), where T_n is a constant which can be calculated for any given significance α. We see that the sample mean $\bar{t}$ is a natural test statistic for this test about the population mean τ. However, to find R (or T_n) it is necessary to know the p.d.f. $f_n(\bar{t})$ for $\bar{t}$. This p.d.f. is particularly simple for the two limiting cases, n=1 and n very large. Fixing the significance level to 5% we will now study these extreme cases.

(i) n=1. The p.d.f. for the test statistic $\bar{t}$ is in this case the same as the p.d.f. for t,

$$f_1(\bar{t}) = \frac{1}{\tau}\exp(-\bar{t}/\tau).$$

The significance α determines the lower limit T_1 of the critical region for $\bar{t}$ through the definition, eq.(14.1),

$$0.05 = \alpha = \int_{T_1}^{\infty} e^{-\bar{t}}d\bar{t}\,.$$

Hence the lower limit is

$$T_1 = -\ln\alpha \simeq 3.00.$$

The power of the test $H_o\colon \tau = 1$ against $H_1\colon \tau = 2$ becomes

$$1-\beta = \int_{T_1}^{\infty} \tfrac{1}{2}\, e^{-\bar{t}/2}d\bar{t} = \sqrt{\alpha} \simeq 0.22.$$

With just one observation of the Ξ^o lifetime we will therefore reject the null hypothesis H_o if the observed value $\bar{t}_{obs}$ is larger than $T_1 \simeq 3.00$. The probability that we accept H_o when the alternative hypothesis H_1 is true (*i.e.* we commit a Type II error) is very large, namely $\beta \simeq 0.78$.

(ii) n large. In this case the p.d.f. for the test statistic $\bar{t}$ can be approximated to a normal p.d.f. with mean τ and variance τ^2/n (this was shown in Sect. 9.4.8),

$$f_n(\bar{t}) = N(\tau,\tau^2/n) = \frac{1}{\sqrt{2\pi}\,\tau/\sqrt{n}}\exp\left(-\tfrac{1}{2}\,\frac{(\bar{t}-\tau)^2}{\tau^2/n}\right).$$

The lower limit T_n of the critical region for $\bar{t}$ is now implied by the integral

$$0.05 = \alpha = \int_{T_n}^{\infty} N(1,\tfrac{1}{n})d\bar{t} = 1 - \int_{-\infty}^{T_n} N(1,\tfrac{1}{n})d\bar{t} = 1 - G\left(\frac{T_n-1}{1/\sqrt{n}}\right),$$

where G is the cumulative standard normal distribution function. From Appendix Table A6 we find the value of the argument of G to be 1.645; hence

$$T_n = 1 + \frac{1.645}{\sqrt{n}}\,.$$

The power of the test $H_o\colon \tau = 1$ against $H_1\colon \tau = 2$ now becomes

$$1-\beta = \int_{T_n}^{\infty} N(2,\tfrac{4}{n})d\bar{t} = 1 - \int_{-\infty}^{T_n} N(2,\tfrac{4}{n})d\bar{t} = 1 - G\left(\frac{T_n-2}{2/\sqrt{n}}\right).$$

Thus, both the critical value T_n and the power $(1-\beta)$ depend on the number of observations n. Numerically, with n=100 we find T_{100}=1.16 and $(1-\beta)$=0.99999. The fact that the limiting value of the power is unity expresses that the test H_o against H_1 is *consistent*.

From the points (i) and (ii) we see, therefore, that for the fixed significance $\alpha=0.05$, the power of the test H_o against H_1 increases very rapidly with the number of observations. If instead we had fixed the significance at $\alpha=0.01$ we would have found $(1-\beta)=0.10$ for one observation, and $(1-\beta)=0.9994$ for $n=100$. Thus the power of the test H_o against H_1 is weakened when the significance level is lowered.

Exercise 14.1: For the Ξ^o lifetime example, consider the two simple hypotheses

H_o: $\tau = 1$, ($\Delta I = \frac{1}{2}$ rule),

H_1: $\tau = 1/4$, ($\Delta I = 3/2$ rule).

For the significance $\alpha=0.05$, show that the power of the test H_o against H_1 with only one observation is 0.19.

14.2.4 The likelihood-ratio test for composite hypotheses

The Neyman-Pearson test is only applicable when the null hypothesis H_o and the alternative H_1 are both simple. In situations where either H_o or H_1, or both, are composite hypotheses one may apply the *likelihood-ratio test*.

We will consider a p.d.f. $f(x|\underline{\theta})$ with parameters $\underline{\theta}=\{\theta_1,\theta_2,\ldots,\theta_k\}$ which belong to the parameter space Ω. We assume that H_o puts some conditions on at least one of the parameters. If H_o is true the parameters are therefore restricted to lie in a subspace ω of the total space Ω. On the basis of a sample of size n from $f(x|\underline{\theta})$ we want to test the hypothesis

$$H_o: \underline{\theta} \text{ belongs to } \omega. \tag{14.9}$$

Given the observations $x_1,x_2,\ldots,x_n$ the likelihood function is

$$L = \prod_{i=1}^{n} f(x_i|\underline{\theta}).$$

The maximum of this function over the total space Ω is denoted by $L(\hat{\Omega})$. Within the subspace ω, specified by the restrictions from H_o, the maximum of the likelihood function is $L(\hat{\omega})$. We then define the *likelihood-ratio* by the quotient

$$\lambda \equiv \frac{L(\hat{\omega})}{L(\hat{\Omega})}. \tag{14.10}$$

This ratio is obviously non-negative, and since the maximum of L within the subspace ω cannot exceed the maximum value over the entire space Ω, λ must be a quantity between 0 and 1.

The likelihood-ratio λ is a function of the observations. If λ turns out to be in the neighbourhood of 1 the null hypothesis H_o is such that it renders $L(\hat{\omega})$ close to the maximum $L(\hat{\Omega})$, and hence H_o will have a large probability of being true. On the other hand, a small value of λ will indicate that H_o is unlikely. The variable λ is therefore intuitively a reasonable test statistic for H_o, and the likelihood-ratio test principle states that the critical region for λ is given by

$$0 < \lambda < \lambda_\alpha. \tag{14.11}$$

Here λ_α must be adjusted to correspond to the chosen significance α, where

$$\alpha = \int_0^{\lambda_\alpha} g(\lambda|H_o)d\lambda \tag{14.12}$$

and $g(\lambda|H_o)$ is the p.d.f. for λ under the assumption that H_o is true.

If the p.d.f. $g(\lambda|H_o)$ is not known it is still possible - as will become clear from an example in the next section - to use the likelihood-ratio test provided that one knows the distribution of some function of λ, which has a monotonic behaviour in λ. Let $y = y(\lambda)$ be a monotonic function of λ and, assuming H_o true, let the p.d.f. for y be $h(y|H_o)$. Then the relation

$$\alpha = \int_0^{\lambda_\alpha} g(\lambda|H_o)d\lambda = \int_{y(0)}^{y(\lambda_\alpha)} h(y|H_o)dy \tag{14.13}$$

gives the critical region in terms of the variable y,

$$y(0) < y < y(\lambda_\alpha), \tag{14.14}$$

and this again can be inverted to give λ_α.

It frequently occurs that it is not possible to put the likelihood-ratio λ in a unique and exact correspondence to a statistic whose distribution is known exactly. Under such circumstances it will be more complicated to

construct a valid test for H_o, and approximative procedures must be tried. Fortunately a satisfactory solution often exists when one is dealing with large samples. For example, if the null hypothesis fixes the values of r of the parameters it can be shown that, for H_o true, the statistic $-2 \ln \lambda$ tends asymptotically to a chi-square variable with r of degrees of freedom. The statistic $-2 \ln \lambda$ can therefore be used to test H_o, taking the critical region at the right-hand tail of the appropriate chi-square distribution.

It should be stressed that the asymptotic behaviour of the likelihood-ratio depends essentially on the same conditions that give the optimum properties of the Maximum-Likelihood estimators.

14.2.5 Example: Likelihood-ratio test on the mean of a normal p.d.f.

To get acquainted with the terminology and idea behind the likelihood-ratio test let us consider a common practical problem. Suppose we want to test whether the mean θ_1 of a normal p.d.f.

$$N(\theta_1,\theta_2) = \frac{1}{\sqrt{2\pi\theta_2}} \exp\left(-\tfrac{1}{2} \frac{(x-\theta_1)^2}{\theta_2}\right)$$

has a specified value μ_o, given a sample of n observations $x_1, x_2, \ldots, x_n$. The null hypothesis to be tested is then

$$H_o: \theta_1 = \mu_o \tag{14.15}$$

and the alternative is

$$H_1: \theta_1 \neq \mu_o. \tag{14.16}$$

For both hypotheses the second parameter θ_2 is some positive number. The total parameter space Ω is here the set of all points in the positive half plane spanned by θ_1 and θ_2, where $-\infty < \theta_1 < \infty$ and $\theta_2 > 0$. Under the null hypothesis H_o the parameter θ_1 is fixed to take the value μ_o; thus the subspace ω is the line defined by $\theta_1 = \mu_o$ for positive values of θ_2.

The likelihood function is

$$L(\underline{x}|\theta_1,\theta_2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\theta_2}} \exp\left(-\tfrac{1}{2} \frac{(x_i-\theta_1)^2}{\theta_2}\right). \tag{14.17}$$

The ML estimators of θ_1 and θ_2 over the entire parameter space Ω are as found in Sect.9.2.3,

$$\hat{\theta}_1 = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}, \tag{14.18}$$

and

$$\hat{\theta}_2 = \frac{1}{n} \sum_{i=1}^{n} (x_i-\bar{x})^2. \tag{14.19}$$

The absolute maximum of the likelihood function is found by inserting $\hat{\theta}_1$ and $\hat{\theta}_2$ in eq.(14.17), giving

$$L(\hat{\Omega}) = \left(\frac{n}{2\pi\sum(x_i-\bar{x})^2}\right)^{n/2} e^{-n/2}. \tag{14.20}$$

Within the subspace ω the ML estimator of θ_2 is found by inserting $\theta_1 = \mu_o$ in eq.(14.17) and maximizing with respect to θ_2; one finds

$$\hat{\theta}_2 = \frac{1}{n} \sum_{i=1}^{n} (x_i-\mu_o)^2. \tag{14.21}$$

We note that the ML estimators of θ_2 are different in the two spaces ω and Ω. For H_o true, the maximum of the likelihood function is

$$L(\hat{\omega}) = \left(\frac{n}{2\pi\sum(x_i-\mu_o)^2}\right)^{n/2} e^{-n/2}. \tag{14.22}$$

From the definition, eq.(14.10), the likelihood-ratio is, therefore,

$$\lambda = \left(\frac{\sum(x_i-\bar{x})^2}{\sum(x_i-\mu_o)^2}\right)^{n/2}. \tag{14.23}$$

Since

$$\sum_{i=1}^{n} (x_i-\mu_o)^2 = \sum_{i=1}^{n} (x_i-\bar{x})^2 + n(\bar{x}-\mu_o)^2,$$

we may write

$$\lambda = \left(\frac{1}{1 + \frac{t^2}{n-1}}\right)^{n/2}, \tag{14.24}$$

where the variable t is defined by

$$t = \frac{\sqrt{n}(\bar{x}-\mu_o)}{\sqrt{\sum(x_i-\bar{x})^2/(n-1)}} = \frac{\frac{\bar{x}-\mu_o}{\sigma/\sqrt{n}}}{\sqrt{\left(\frac{(n-1)s^2}{\sigma^2}\right)/(n-1)}} . \qquad (14.25)$$

With reference to Sect.5.2.1, remark 2, we see that this is a Student's t-variable with n-1 degrees of freedom. The variable has the p.d.f. $f(t;\nu)$ of eq. (5.15) with $\nu = n-1$.

The p.d.f. $g(\lambda)$ for λ may now be derived from the t-distribution, and thus provide the critical region for λ with a given significance α; compare eq. (14.13). However, this variable transformation is not necessary, because λ is a monotonic function of t^2. Since H_o false corresponds to λ small and t^2 very large, we see that a critical region defined by $0 < \lambda < \lambda_\alpha$ is equivalent to $t^2 > t_\alpha^2$. Thus the critical region for t corresponding to the significance α consists of the two intervals

$$-\infty < t < -t_{\alpha/2}, \qquad t_{\alpha/2} < t < +\infty . \qquad (14.26)$$

Here $t_{\alpha/2}$ is defined by

$$\alpha/2 = \int_{-\infty}^{-t_{\alpha/2}} f(t;n-1)dt = \int_{t_{\alpha/2}}^{\infty} f(t;n-1)dt . \qquad (14.27)$$

With the critical region consisting of two separate parts we are here dealing with a <u>two-sided test</u>. If the observed value $t = t_{obs}$ satisfies one of the inequalities (14.26) the hypothesis H_o is rejected. If the observed value satisfies

$$-t_{\alpha/2} < t < t_{\alpha/2}$$

the hypothesis H_o is accepted.

Consider, for instance, an experiment with n=20 and a chosen significance α=0.05. For this case Appendix Table A7 gives $t_{.025}$=2.093. If therefore the observed $|t|$-value computed from eq.(14.25) is larger than 2.093, the null hypothesis should be rejected. The corresponding critical value for λ is found by inserting $t_{.025}$=2.093 into eq.(14.24), giving $\lambda_{.05}$=0.125.

Let us compare this exact calculation of the critical region for λ to

the approximation introduced at the end of the preceding section. According to the statement made $-2 \ln \lambda$ is approximately chi-square distributed with a number of degrees of freedom equal to the number of parameters which were fixed by the hypothesis H_o. In the present case H_o fixes one parameter, and eqs.(14.24), (14.25) give

$$-2 \ln \lambda = n \ln\left(1 + \frac{n(\bar{x}-\mu_o)^2}{\sum(x_i-\bar{x})^2}\right) .$$

When a Taylor expansion is performed and the sample variance replaced by σ^2 for n large we get the asymptotic expression

$$-2 \ln \lambda \simeq \left(\frac{\bar{x}-\mu_o}{\sigma/\sqrt{n}}\right)^2 .$$

For a chi-square variable with one degree of freedom we find from Appendix Table A8 that a <u>one-sided test</u> for the significance α=0.05 corresponds to the critical value $-2 \ln \lambda_{.05}$=3.841, from which $\lambda_{.05}$=0.147. Comparing this to the exact value 0.125 obtained above, we see that even for a sample size as small as 20 the asymtotic $-2 \ln \lambda$ approximation is reasonable.

To compute the power of the test we have to consider the alternative hypothesis H_1: $\theta_1 \neq \mu_o$. The variable t as defined by eq.(14.25) may still be used as a test statistic, but if H_1 is true, the statistic has a non-central Student's t-distribution (compare Sect.5.2.1, remark 1 and Exercise 5.20) with non-central parameter

$$\delta = \frac{\theta_1-\mu_o}{\sigma/\sqrt{n}} . \qquad (14.28)$$

It is seen that δ gives a measure of how much the alternative deviates from the null hypothesis.

In accordance with the definition of the power of a test *the power function* is defined as

$$1-\beta(\delta) = \int_0^{\lambda_\alpha} g(\lambda|H_1,\delta)d\lambda = \int_{-\infty}^{-t_{\alpha/2}} f(t;n-1,\delta)dt + \int_{t_{\alpha/2}}^{\infty} f(t;n-1,\delta)dt ,$$

where $g(\lambda|H_1,\delta)$ is the distribution of λ for a given δ when H_1 is true, and $f(t;n-1,\delta)$ is the non-central t-distribution with n-1 degrees of freedom. The

cumulative distribution of $f(t;n-1,\delta)$ has been tabulated for various parameter combinations in, for instance, "Tables of the non-central t-distribution" by G.J. Resnikoff and G.J. Lieberman. In Fig 14.4 the power function is shown for

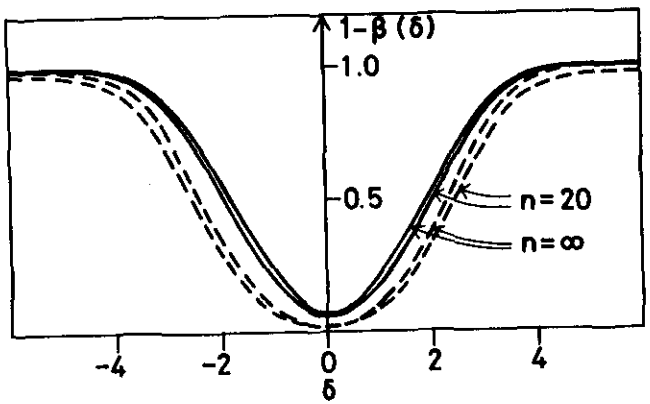

Fig. 14.4. Power function for Student's t test on the mean value; curves are given for sample sizes n=20 and n=∞, and for significances α=0.05 (full curves) and α=0.02 (dotted curves).

different values of α and n. The following features emerge from the figure:

(i) The power function is symmetric about $\delta=0$, as already suggested by the proposition we are checking.

(ii) The power is at its minimum for $\delta=0$ when it equals the significance, and increases with increasing magnitude of $|\delta|$.

(iii) The power increases with increasing significance.

(iv) For a given δ the power increases with the number of observations.

These results would obviously also be found if, instead of testing on t, we had considered an upper-tail test with the statistic t^2, which has a non-central F-distribution if the alternative hypothesis is true.

Exercise 14.2: For the Ξ^0 lifetime example introduced in Sect.14.2.3, consider the test of the simple null hypothesis H_0: $\tau = 1$ against the composite alternative hypothesis H_1: $\tau \neq 1$. Show that with the sample $t_1,t_2,\ldots,t_n$ the likelihood-ratio becomes

$$\lambda = \bar{t}^n \exp(-n(\bar{t}-1)),$$

where $\bar{t} = \Sigma t_i/n$. For n reasonably large, when $\bar{t}$ is $N(1,1/n)$, show that $\alpha=0.05$ corresponds to a critical region $0 < \lambda < \lambda_{.05}$ where

$$\lambda_{.05} = \left(1 + \frac{1.645}{\sqrt{n}}\right)^n \exp(-1.645\sqrt{n}),$$

and that $\lambda_{.05}$ approaches unity when $n \to \infty$.

14.3 PARAMETRIC TESTS FOR NORMAL VARIABLES

We shall now concentrate on various test problems which involve the parameters of the normal distribution. The first of these is the following: Given a set of observations, assumed to represent a random sample from a normal population, we want to test whether these observations are in agreement with some particular value of the mean or variance of the normal distribution. Simple tests of this type are treated in Sect.14.3.1. In the subsequent sections we will be concerned with constructing tests for the comparison of parameters of two normal distributions. In Sect.14.3.7 we discuss how the mean values in more than two normal distributions can be compared.

The main point in all the examples considered below is to construct an appropriate test statistic, that is, some function involving the sample variables but no unspecified parameters, and which has a known distribution function. This distribution will enable us to determine a critical region for the test statistic in question, the size of the region being determined by the chosen significance level of the test.

14.3.1 Tests of mean and variance in $N(\mu,\sigma^2)$

Let $x_1,x_2,\ldots,x_n$ be a random sample from the normal population $N(\mu,\sigma^2)$. We want first to see how the mean value of the distribution can be put to test and write the null hypothesis as

$$H_0: \mu = \mu_0, \tag{14.30}$$

where μ_0 is some specific number. The alternative is the composite hypothesis

$$H_1: \mu \neq \mu_0. \tag{14.31}$$

We then have a problem which involves a reasoning quite analogous to that used in establishing confidence intervals for the mean of a normal distribution, which

was carried out in some detail in Sect.7.2. Depending on whether σ^2 is a known quantity or not, the appropriate test statistic to test $H_o: \mu = \mu_o$ is (see Sect. 7.2.1 and 7.2.2 for the arguments)

$$\left.\begin{aligned} &\frac{\bar{x}-\mu_o}{\sigma/\sqrt{n}}, \quad \text{which is } N(0,1) \quad (\sigma^2 \text{ known}) \\ &\frac{\bar{x}-\mu_o}{s/\sqrt{n}}, \quad \text{which is } t(n-1) \quad (\sigma^2 \text{unknown}) \end{aligned}\right\} \tag{14.32}$$

Here $N(0,1)$, is the usual abbreviation for a standard normal variable, while $t(n-1)$ is a Student's t-variable with n-1 degrees of freedom; $\bar{x}$ and s^2 are the wellknown sample characteristics,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2.$$

In testing the mean we are likely to reject H_o (and accept H_1) if the observations are such that the sample mean $\bar{x}$ is either much too small or much too large compared to the hypothetical value μ_o, therefore a two-sided test is applied. Fixing the significance α we will use the appropriate distribution (*i.e.* the standard normal or the Student's t, depending on whether σ^2 is known or not) to determine the two critical values of the test statistic, which correspond to a probability $\frac{1}{2}\alpha$ in either tail of the distribution. If it is found that the observed value of the test statistic falls in any of the two tails (*i.e.* it belongs to the critical region) then we shall reject the hypothesis $H_o: \mu = \mu_o$ at the significance level 100α %; otherwise the null hypothesis is accepted.

A similar procedure is applicable if we want to test the variance of the normal distribution. The null hypothesis is

$$H_o: \sigma^2 = \sigma_o^2, \tag{14.33}$$

where σ_o^2 is specified, and the alternative (composite) hypothesis can be

$$H_1: \sigma^2 \neq \sigma_o^2. \tag{14.34}$$

The technique developed earlier (see Sects.7.3.1 and 7.3.2) for establishing a confidence interval for the variance of a normal distribution can now be used. We recall that a variable constructed from the sample variance s^2 as $(n-1)s^2/\sigma_o^2$ will have a chi-square distribution with n-1 degrees of freedom. If we should happen to know what the true mean μ of the population is we may replace $\bar{x}$ by μ in the evaluation of s^2, and the chi-square variable $(n-1)s^2/\sigma_o^2$ will then have n degrees of freedom. Thus the appropriate test statistic to be used for testing $H_o: \sigma^2 = \sigma_o^2$ is

$$\left.\begin{aligned} &(n-1)s^2/\sigma_o^2 = \sum_{i=1}^{n}(x_i-\mu)^2/\sigma_o^2, \quad \text{which is } \chi^2(n) \quad (\mu \text{ known}), \\ &(n-1)s^2/\sigma_o^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2/\sigma_o^2, \quad \text{which is } \chi^2(n-1) \quad (\mu \text{ unknown}). \end{aligned}\right\} \tag{14.35}$$

In testing the variance we are likely to reject H_o (and accept H_1) if s^2 from the measurements is either much too small or much too large compared to the value σ_o^2 specified by H_o. Therefore, again, a two-sided test is appropriate. Since the chi-square distribution is unsymmetric the choice of the two critical regions is not unambiguous, but common practice is to take equal probabilities ($= \frac{1}{2}\alpha$) at both ends of the distribution.

In general, two-sided tests for the null hypothesis H_o will always be appropriate whenever the alternative hypothesis are of the forms of eqs.(14.31), (14.34). If the alternatives are formulated as, for example,

$$H_1: \mu > \mu_o,$$

or

$$H_1: \sigma^2 > \sigma_o^2,$$

one-sided tests must be applied to test H_o.

14.3.2 Comparison of means in two normal distributions

We shall assume that $x_1, x_2, \ldots, x_n$ is a sample of size n from $N(\mu_1, \sigma_1^2)$ and $y_1, y_2, \ldots, y_m$ a sample of size m from $N(\mu_2, \sigma_2^2)$. We want to test the hypothesis that the two normal distributions have the same mean,

$$H_o: \mu_1 = \mu_2, \tag{14.36}$$

against the alternative that the means are different

$$H_1: \mu_1 \neq \mu_2. \tag{14.37}$$

The formalism covers situations frequently met in practice when it is desired to check whether two series of numbers really are measurements on the same physical quantity. Although simply stated, the null hypothesis is not always easily tested unless some simplifying assumptions can be made about the variances σ_1^2 and σ_2^2. We shall consider the following situations:

(i) σ_1^2 and σ_2^2 are known,
(ii) σ_1^2 and σ_2^2 are unknown, but equal,
(iii) σ_1^2 and σ_2^2 are unknown and different.

(i) σ_1^2 and σ_2^2 are known

We know that the arithmetic means $\bar{x}$ and $\bar{y}$ will be normally distributed as $N(\mu_1,\sigma_1^2/n)$ and $N(\mu_2,\sigma_2^2/m)$, respectively. According to the addition theorem for normal variables, (Sect.4.8.5), the difference $\bar{x}-\bar{y}$ will also be normal, with mean $(\mu_1-\mu_2)$ and variance $(\sigma_1^2/n+\sigma_2^2/m)$. Thus the variable

$$\frac{(\bar{x}-\bar{y}) - (\mu_1-\mu_2)}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}}$$

will be N(0,1). To test the null hypothesis of equal means, when $\mu_1-\mu_2=0$, we consequently use the test statistic

$$\frac{\bar{x} - \bar{y}}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}} \qquad (14.38)$$

which is N(0,1) under H_o. The problem is therefore reduced to a wellknown one, being exactly analogous to the testing of the mean of a normal distribution when its variance is known, which was discussed in the preceding section.

(ii) σ_1^2 and σ_2^2 are unknown, but equal

If the variances of the populations are not known it is still possible to carry out the desired test, provided that it can be assumed that the two variances are of equal size. To construct a test statistic for this case we recall that with the two samples $x_1,x_2,\ldots,x_n$ from $N(\mu_1,\sigma_1^2)$ and $y_1,y_2,\ldots,y_m$ from $N(\mu_2,\sigma_2^2)$ we can immediately write down a set of four variables which are all

independent and have wellknown distributions:

$\bar{x}$, which is $N(\mu_1,\sigma_1^2/n)$, $\qquad (n-1)s_1^2/\sigma_1^2$, which is $\chi^2(n-1)$,

$\bar{y}$, which is $N(\mu_2,\sigma_2^2/m)$, $\qquad (m-1)s_2^2/\sigma_2^2$, which is $\chi^2(m-1)$.

These variables are the entities from which we must try to build up a new variable that may serve as a test statistic. The two normal variables can be combined in the usual way such that

$$\frac{(\bar{x}-\bar{y}) - (\mu_1-\mu_2)}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}} \qquad \text{is } N(0,1).$$

Adding the two chi-square variables gives a new chi-square variable with a number of degrees of freedom equal to $(n-1)+(m-1)=n+m-2$ (the addition theorem for chi-square variables, Sect.5.1.5), thus

$$(n-1)s_1^2/\sigma_1^2 + (m-1)s_2^2/\sigma_2^2 \qquad \text{is } \chi^2(n+m-2).$$

The standard normal and the chi-square variables constructed this way are furthermore independent, and according to Sect.5.2.1 they can be used to form a Student's t-variable by the prescription of eq.(5.14). Hence

$$\frac{\dfrac{(\bar{x}-\bar{y}) - (\mu_1-\mu_2)}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}}}{\sqrt{\left[(n-1)s_1^2/\sigma_1^2 + (m-1)s_2^2/\sigma_2^2\right]/(n+m-2)}} \qquad (14.39)$$

will be a Student's t-variable with (n+m-2) degrees of freedom.

Note that the complicated expression (14.39) is no test statistic yet, since it includes the parameters σ_1^2 and σ_2^2, which were assumed to be unknown. However, <u>if $\sigma_1^2 = \sigma_2^2$</u>, the expression can be simplified further since the population variances drop out, giving

$$\frac{(\bar{x}-\bar{y}) - (\mu_1-\mu_2)}{S\sqrt{\frac{1}{n} + \frac{1}{m}}}, \qquad (14.40)$$

where S^2 is the *pooled estimate* of the population variance σ^2 $(=\sigma_1^2=\sigma_2^2)$,

$$S^2 \equiv \frac{1}{n+m-2}\left((n-1)s_1^2 + (m-1)s_2^2\right) = \frac{1}{n+m-2}\left(\sum_{i=1}^{n}(x_i-\bar{x})^2 + \sum_{i=1}^{m}(y_i-\bar{y})^2\right). \quad (14.41)$$

Therefore, to test the hypothesis of equal means in the case of unknown, but equal variances, we shall take as our test statistic

$$\frac{\bar{x}-\bar{y}}{S\sqrt{\frac{1}{n}+\frac{1}{m}}}, \quad (14.42)$$

which is t(n+m-2). Thus the problem is analogous to the one discussed previously in Sect.14.3.1, when we tested the mean of a normal distribution with an unknown variance.

It is seen from this rather lengthy derivation that unless the two unknown population variances σ_1^2 and σ_2^2 are equal, we shall in general not be able to obtain a variable for which we know the explicit distribution. The general case $\sigma_1^2 \neq \sigma_2^2$ can in fact not be treated exactly, and approximate methods must be applied.

(iii) σ_1^2 and σ_2^2 are unknown, and different

Given two samples from assumed normal distributions, one does not usually have any exact knowledge about the population variances σ_1^2 and σ_2^2, nor does one have any particular reason to believe that they are of equal size. Thus the conditions specified in (i) or (ii) above are not fulfilled in general.

When the samples are reasonably large, it is fair to say that the sample variances are good estimates of the population variances. Replacing σ_1^2, σ_2^2 by their estimates s_1^2, s_2^2 in eq.(14.36) gives

$$\frac{(\bar{x}-\bar{y}) - (\mu_1-\mu_2)}{\sqrt{s_1^2/n + s_2^2/m}}. \quad (14.43)$$

To test the hypothesis H_o: $\mu_1=\mu_2$ one will therefore in this situation use the test statistic

$$\frac{\bar{x}-\bar{y}}{\sqrt{s_1^2/n + s_2^2/m}}, \quad (14.44)$$

which is approximately standard normal if H_o is true.

The approach described above can be used for the practical purpose of checking consistency between two sets of observations. If the first set $x_1,x_2,\ldots,x_n$ implies an experimental result $(\bar{x} \pm \Delta\bar{x})$ and the second set $y_1,y_2,\ldots,y_m$ similarly $(\bar{y} \pm \Delta\bar{y})$ a test for the hypothesis that the two experiments have measured the same physical quantity (*i.e.* $\mu_1=\mu_2$) can be based on the test statistic

$$z = \frac{\bar{x}-\bar{y}}{\sqrt{(\Delta\bar{x})^2 + (\Delta\bar{y})^2}}, \quad (14.45)$$

which is assumed to be N(0,1). In fact, this is the prescription used by most physicists to test if two experimental results are compatible. Thus the assumption of normally distributed observations is implicit whenever the above prescription is used, although this is seldom explicitly stated in practice.

Exercise 14.3: Experimental values (as of April 1976) for the Ξ^0 and Ξ^- lifetimes are, respectively,

$$\tau_o = (2.96 \pm 0.12) \times 10^{-10} \text{ sec},$$

$$\tau_- = (1.652 \pm 0.023) \times 10^{-10} \text{ sec}.$$

Show that the null hypothesis H_o: $\tau_o = 2\tau_-$ (prediction of the $\Delta=\frac{1}{2}$ rule) can be accepted at a significance level of 0.5% against the alternative hypothesis H_1: $\tau_o \neq 2\tau_-$.

14.3.3 Comparison of variances in two normal distributions

For the two normal samples, $x_1,x_2,\ldots,x_n$ from $N(\mu_1,\sigma_1^2)$ and $y_1,y_2,\ldots,y_m$ from $N(\mu_2,\sigma_2^2)$ we can also be interested in testing the hypothesis of equal variances,

$$H_o: \sigma_1^2 = \sigma_2^2, \quad (14.46)$$

against the alternative

$$H_1: \sigma_1^2 > \sigma_2^2. \quad (14.47)$$

For the moment we assume that both μ_1 and μ_2 are unknown. Since we know that the population variances are estimated by the sample variances it is reasonable to try to establish a test statistic using the by now wellknown properties that

$$\frac{(n-1)s_1^2}{\sigma_1^2} = \frac{1}{\sigma_1^2}\sum_{i=1}^{n}(x_i-\bar{x})^2 \qquad \text{is } \chi^2(n-1),$$

$$\frac{(m-1)s_2^2}{\sigma_2^2} = \frac{1}{\sigma_2^2}\sum_{i=1}^{m}(y_i-\bar{y})^2 \qquad \text{is } \chi^2(m-1).$$

With two independent chi-square variables it is natural to construct an F-variable by dividing each of them by its number of degrees of freedom and then take their ratio,

$$\frac{\dfrac{(n-1)s_1^2}{\sigma_1^2}/(n-1)}{\dfrac{(m-1)s_2^2}{\sigma_2^2}/(m-1)} = \frac{s_1^2}{s_2^2}\cdot\frac{\sigma_2^2}{\sigma_1^2}, \tag{14.48}$$

which, according to the definition in Sect.5.3.1, gives an F-variable with $(n-1, m-1)$ degrees of freedom. For the test of H_o against H_1 we shall therefore use the simple test statistic

$$\frac{s_1^2}{s_2^2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \Big/ \frac{1}{m-1}\sum_{i=1}^{m}(y_i-\bar{y})^2 \tag{14.49}$$

and take the critical region at the upper tail of the appropriate F-distribution. Conversely, if the alternative were

$$H_1: \sigma_1^2 < \sigma_2^2, \tag{14.50}$$

we would take the critical region in the lower part of the distribution.

If the population means μ_1 and μ_2 were known one could reason in a similar manner; it is readily seen that the appropriate test statistic in this case is obtained by replacing in eq.(14.49) the sample averages $\bar{x}$ and $\bar{y}$ by the population means μ_1 and μ_2, giving

$$\frac{s_1^2}{s_2^2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\mu_1)^2 \Big/ \frac{1}{m-1}\sum_{i=1}^{m}(y_i-\mu_2)^2. \tag{14.51}$$

This statistic will be distributed as an F-variable with (n,m) degrees of freedom.

14.3.4 Summary table

Table 14.1 summarizes relevant details about different tests involving the mean and variance of normal distributions under various specified conditions as discussed in the previous sections and exemplified in the following.

Table 14.1 Tests for the mean and variance of normal distributions

H_o	Conditions	Test statistic	Test distribution
$\mu=\mu_o$	σ^2 known	$\dfrac{\bar{x}-\mu_o}{\sigma/\sqrt{n}}$	N(0,1)
	σ^2 unknown	$\dfrac{\bar{x}-\mu_o}{s/\sqrt{n}}$	t(n-1)
$\sigma^2=\sigma_o^2$	μ known	$(n-1)s^2/\sigma_o^2 = \sum_{i=1}^{n}(x_i-\mu)^2/\sigma_o^2$	$\chi^2(n)$
	μ unknown	$(n-1)s^2/\sigma_o^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2/\sigma_o^2$	$\chi^2(n-1)$
$\mu_1-\mu_2=0$	$\sigma_1^2=\sigma_2^2=\sigma^2$ known	$\dfrac{\bar{x}-\bar{y}}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}}$	N(0,1)
	σ_1^2 and σ_2^2 known	$\dfrac{\bar{x}-\bar{y}}{\sqrt{\sigma_1^2/n+\sigma_2^2/m}}$	
	$\sigma_1^2=\sigma_2^2=\sigma^2$ unknown	$\dfrac{\bar{x}-\bar{y}}{S\sqrt{\frac{1}{n}+\frac{1}{m}}}$, $S^2 \equiv \frac{1}{n+m-2}\left((n-1)s_1^2+(m-1)s_2^2\right)$	t(n+m-2)
	$\sigma_1^2 \neq \sigma_2^2$ unknown	$\dfrac{\bar{x}-\bar{y}}{\sqrt{s_1^2/n+s_2^2/m}}$	not exactly known, $\approx$ N(0,1)
$\frac{\sigma_1^2}{\sigma_2^2}=1$	μ_1 and μ_2 known	$\dfrac{s_1^2}{s_2^2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\mu_1)^2 \Big/ \frac{1}{m-1}\sum_{i=1}^{m}(y_i-\mu_2)^2$	F(n,m)
	μ_1 and μ_2 unknown	$\dfrac{s_1^2}{s_2^2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \Big/ \frac{1}{m-1}\sum_{i=1}^{m}(y_i-\bar{y})^2$	F(n-1,m-1)

14.3.5 Example: Comparison of results from two different measuring machines

As a first example on tests involving parameters of normal distributions, suppose that we want to compare measurements on beam tracks in a bubble chamber from two different measuring devices with unspecified precisions.

For definiteness we assume that a monoenergetic beam of momentum P_o=24.90 GeV/c is incident in the bubble chamber, and that the multiple Coulomb scattering on the tracks can be considered negligible. Then the error on the measured track momentum will be due solely to the inaccuracy of the measuring device. From the geometrical reconstruction method applied to the measured tracks one expects that it is the inverse of the momentum - rather than the momentum itself - which is an approximate normal variable. Suppose that 20 tracks have been measured with the two machines A and B, giving the mean values and standard deviations for the inverse momentum 1/P as displayed in the following table.

Machine	$\langle\frac{1}{P}\rangle = \frac{1}{20}\sum_{i=1}^{20}\frac{1}{P_i}$ in units $10^{-3}\cdot(\mathrm{GeV/c})^{-1}$	$s = \left(\frac{1}{20-1}\sum_{i=1}^{20}\left(\frac{1}{P_i} - \langle\frac{1}{P}\rangle\right)^2\right)^{\frac{1}{2}}$ in units $10^{-3}\cdot(\mathrm{GeV/c})^{-1}$
A	40.12	0.46
B	40.32	0.25

Let us first check the supposition that each machine really measures the incident momentum, which corresponds to $1/P_o = 40.16\cdot10^{-3}(\mathrm{GeV/c})^{-1}$. For each machine this implies a test on the mean value of a normal distribution, with the condition that the variance is unknown. With the formulation of the present chapter we write

$$H_o: \frac{1}{\bar{P}} = \frac{1}{P_o},$$

$$H_1: \frac{1}{\bar{P}} \neq \frac{1}{P_o}.$$

According to Sect.14.3.1 the appropriate test statistic is

$$t = \frac{\langle\frac{1}{P}\rangle - \frac{1}{P_o}}{s/\sqrt{20}},$$

which has a Student's t-distribution with 19 degrees of freedom. If we choose a significance of α=0.05 we find from Appendix Table A7 that the critical region for the two-sided test with equal probabilities $\frac{1}{2}\alpha$ in the two tails is given by

$$|t| > t_{.025} = 2.09.$$

From the observed numbers in the table above we find that the observations with machines A and B correspond to

$$t^A_{obs} = -0.40, \qquad t^B_{obs} = +1.60.$$

Hence both series of observations produce values of t in the acceptance region, and there is no reason to suspect the measurements for either machine.

Secondly, we want to test the hypothesis that the two measuring machines have the same precision. This amounts to testing whether the variances σ_A^2 and σ_B^2 are equal, since

$$H_o: \frac{1}{\sigma_A^2} = \frac{1}{\sigma_B^2}.$$

The alternative hypothesis may state that A is less precise than B,

$$H_1: \frac{1}{\sigma_A^2} < \frac{1}{\sigma_B^2}.$$

In this case we know the mean values of the two (normal) distributions, $\mu_A=\mu_B=1/P_o$ and, in accordance with Sect.14.3.3, we form the test statistic

$$F = \frac{s_A^2}{s_B^2},$$

which has an F-distribution with (20,20) degrees of freedom. Fixing the significance level at 5% the critical region corresponding to a one-sided test is given by, according to Appendix Table A8,

$$F > F_{.95} = 2.16.$$

From the observed numbers we have the actual number

$$F_{obs} = \frac{0.46^2}{0.25^2} = 3.39,$$

which falls within the critical region. Hence we must reject the hypothesis of equal precisions at the chosen significance level. Indeed, we find that there is a probability of less than 0.5% that a larger ratio F would be observed, if H_o were true. Hence we are likely to accept the alternative hypothesis in this case, and conclude that machine A is less precise than machine B.

14.3.6 Example: Significance of signal above background

It often happens that an experimental set-up to study a particular effect implies the registration of various background effects which tend to obscure the "real" effect under study. Whenever the background becomes appreciable compared to the total signal it will be increasingly difficult to decide whether an apparent effect is real or just represents a statistical fluctuation of the background. We are therefore interested in finding criteria for assigning "reality" to various phenomena and do so, for example, by attaching a significance level to the observed enhancement in an effective-mass distribution or the dip seen in a four-momentum-transfer spectrum, *etc.*

To be specific, let us consider the effective-mass histogram in Fig. 14.5, where an accumulation of events is observed centered at a mass value $M=M_o$

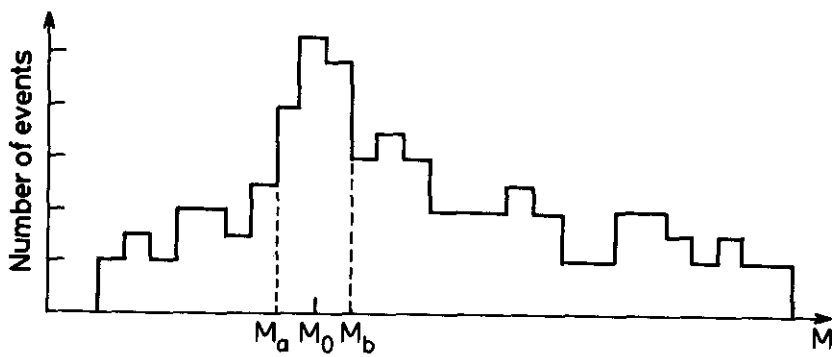

Fig. 14.5. Experimental effective-mass plot with enhancement.

and extending over a few bins. We ask the following questions:

(1) What is the probability that the observed effect <u>at this particular mass value</u> ($M=M_o$) is a statistical fluctuation of the background?

(2) What is the probability to observe such a fluctuation <u>at any position</u> in the spectrum?

To answer question (1), let the total number of events in the interval $[M_a,M_b]$ ("the bump region") be denoted by N and the number of background events in the same region be denoted by B. If we assume that all events in the bump region are due to the background only, we will formulate a null hypothesis as

$$H_o: N = B. \tag{14.52}$$

It sometimes happens that the background distribution is known from some theory or model. In this fortunate situation B and its variance V(B) can be calculated directly. More often the background is not known, but is approximated by a polynomial of appropriate order and fitted to the observed spectrum on both sides of the peak or bump. Another procedure, which may be equally well justified, is to estimate the number B by first forgetting about the bump region and drawing smooth curves through the data points on each side of the bump, and then finally interpolate over the bump region. The error in such an eye-estimate of $\hat{B}$ may be taken as half the difference between the largest and the smallest "reasonable" values of the estimated number $\hat{B}$ of background events.

The total number of events N in the bump region may be regarded as a Poisson variable. Under the assumption that the enhancement is a fluctuation, equivalent to assuming H_o true, the variance of the total number is then

$$V(N) = N = B. \tag{14.53}$$

If B has been estimated by ignoring the enhancement region, as for instance from a polynomial fit or by the hand-drawn curves as described above, it is correct to regard N and B as *uncorrelated* variables. Then the variance on their difference becomes

$$V(N-B) = V(N) + V(B) = B + V(B).$$

If the number of events is not too small, it is further justified to approximate the Poisson variable N to a *normal* variable. To test the hypothesis H_o it is therefore suggestive to adopt the test statistic

$$d = \frac{N - B}{\left(V(N-B)\right)^{\frac{1}{2}}} \simeq \frac{N - \hat{B}}{\left(\hat{B} + V(\hat{B})\right)^{\frac{1}{2}}}, \qquad (14.54)$$

which will be an approximate standard normal variable if H_o is true. The test problem is therefore analogous to that of testing the mean of a normal distribution, which was considered in Sect.14.3.1.

We are now in a position to state an answer to our question (1): The probability of observing a *positive* fluctuation at least as large as the bump observed around the mass value $M=M_o$ is simply given by

$$P(d; M=M_o) = 1 - G(d) = G(-d), \qquad (14.55)$$

where G has the usual meaning of the cumulative standard normal distribution.

It is seen that d measures the excess of events above the background in units of standard deviations. Common practice is to express the significance of an enhancement by quoting the number of standard deviations as defined by eq.(14.54). Evidently, small values of d are compatible with the assumption that the background is responsible for the bump, while very large values of d may be regarded as support for an alternative interpretation, for instance the presence *of a resonance.*

It should be stressed that it is very important in determining the significance of an enhancement to take into account the uncertainty in the background estimate. It is seen that the term $V(\hat{B})$ in the denominator of eq.(14.54) tends to lower the number of standard deviations, that is, it reduces the significance of the peak. Since, for given total number of events N and estimated background $\hat{B}$, a broad accumulation covering many bins will usually have a larger $V(\hat{B})$ than has a narrow peak, it will in general be more difficult to detect a broad resonance than a narrow one.

We proceed next to answer the question (2) raised earlier. We assume that we have already determined the probability $P(d; M=M_o)$ that the d standard deviation effect at the particular mass value $M=M_o$ is a statistical fluctuation

and that we want to find the probability P(d) that a bump of d standard deviations may appear at any place in the histogram. If the bump extends over k bins and the total number of bins in the histogram is n, the central value of the bump may be located in (n-k+1) different bins over the plot. Therefore the probability to observe a fluctuation of at least d standard deviations anywhere in the histogram is

$$P(d) = 1 - \left(1 - P(d; M=M_o)\right)^{n-k+1}.$$

If d is large this may be written as

$$P(d) \simeq (n-k+1)\ P(d; M=M_o). \qquad (14.56)$$

Thus the probability to observe a highly significant peak of d standard deviations anywhere in an effective-mass plot increases with the number of bins in the plot and decreases with the width of the peak. For reactions with many particles in the final state where many mass combinations can be formed, the chance for observing large effects somewhere in any of the effective-mass distributions may certainly become quite appreciable.

To exemplify the last point, let us estimate the number of "many-σ fluctuations" expected per year in effective-mass plots from bubble chamber experiments performed all over the world. The following numbers apply to the year 1970:

total number of events measured	$= 2 \cdot 10^6$,
average number of mass combinations per event	$= 15$,
number of combinations in each histogram	$= 3000$,
number of bins per histogram	$= 40$.

This gives an estimated number of bins per year equal to

$$\frac{2\cdot 10^6 \times 15 \times 40}{3000} = 4\cdot 10^5 \text{ bins/yr}.$$

Since the probability of a positive fluctuation of minimum 4σ is $3.2\cdot 10^{-5}$ in any of these bins we must expect total of ≃ 13 occurrences per year of "effects" of at least 4 standard deviations in magnitude. This example illustrates why it has

become customary to require five or more standard deviations to claim any enhancement in an effective-mass distribution as evidence for a new resonance.

Exercise 14.4: In an experiment to search for and study the rate of elastic antineutrino-electron scattering, *viz.*

$$(1)\quad \bar{\nu}_e + e^- \rightarrow \bar{\nu}_e + e^-$$

Reines, Gurr and Sobel have utilized low-energy antineutrinos from a nuclear reactor. Evidence for the reaction would mainly come from a difference in the counting rates observed when the reactor was ON and when it was OFF. For six different electron energy intervals one registered the counts over periods of many days. The observed counting rates with the reactor ON and OFF are given in the table below (columns II and III), together with the estimated errors due to instrumental instabilities in the runs (column IV):

I Electron energy (MeV)	II Counts/day (over 64.6 days) with reactor ON	III Counts/day (over 60.7 days) with reactor off	IV Stability error in any run	V Counts/day associated with reactor	VI Standard deviations
1.5 - 2.0	30.6 ± .69	26.9 ± .67	± .60	3.7 ± 1.28	2.89
2.0 - 2.5	10.5 ±	9.1 ±	± .38		
2.5 - 3.0	4.0 ±	3.2 ±	± .17		
3.0 - 3.5	1.5 ±	0.6 ±	± .08		
3.5 - 4.0	0.54 ±	0.35 ±	± .03		
4.0 - 4.5	0.40 ±	0.21 ±	± .01		

(i) Assume the number of counts per day to be Poisson variables and fill into columns II and III the errors (standard deviations) on the average numbers of events per day. (The errors for the first energy bin are given for check.)
(ii) Find the difference in the observed counting rates with the reactor ON and with the reactor OFF and determine the error on these differences (column V). (Hint: Assume the instrumental instability to be independent of the number of counts.)
(iii) In accordance with common practice among physicists, express the size of the observed reactor associated rates in units of standard deviations (column VI).
(iv) At a chosen significance level of 0.1%, do the reactor associated rates of the separate energy bins support the hypothesis of a real (positive, non-zero) effect in any bin? (Hint: Assume the reactor associated rate to be a normal variable.)
(v) If the data are lumped into one group covering all energy bins, what is the combined evidence for a real signal of reaction (1)?

14.3.7 Comparison of means in N normal distributions; scale factor

We saw in Sect.14.3.2 how the mean values of two normal distributions could be tested for equality by construction of an appropriate test statistic, which was exactly or approximately N(0,1), or a Student's t-variable, depending on whether the variances of the two normal distributions were known or not. These methods can therefore be used to test whether two experimental results are mutually consistent provided, of course, that the underlying assumption of normally distributed observations is reasonably satisfied.

It is frequently desired to check the internal consistency between more than two experimental results. Suppose, for example, that there are N experiments, each reporting an observed value x_i with error Δx_i. We want to find out whether these N measurements*) are compatible within the errors. This we formulate as a test problem where the null hypothesis under test assumes that the x_i originate from normal populations which all have the same mean value,

$$H_o: \mu_1 = \mu_2 = \ldots = \mu_N. \qquad (14.57)$$

The alternative hypothesis is any other possibility, where some of the means are not equal to the others, corresponding to bias in some of the experiments.

If the null hypothesis had specified the value μ of the common population mean, and also the standard deviations σ_i of the individual normal distributions, then the quantity $\sum_{i=1}^{N}(x_i-\mu)^2/\sigma_i^2$ would by definition be a chi-square variable with N degrees of freedom and an appropriate test statistic. However, since H_o does not specify the common population mean it must be estimated from the data. An estimate for μ is the weighted mean value for all N measurements,

$$\hat{\mu} = \bar{x} = \sum_{i=1}^{N} w_i x_i \Big/ \sum_{i=1}^{N} w_i , \qquad (14.58)$$

where the weight of the i-th observation is taken equal to inverse of the square of its error, $w_i = 1/\Delta x_i^2$. An estimate for the error in μ is

$$\hat{\sigma} = \Delta\bar{x} = \left(\sum_{i=1}^{N} w_i \right)^{-\frac{1}{2}} . \qquad (14.59)$$

*) In practice, each reported value x_i will often be an average value from a series of observations, and the Δx_i the corresponding error in this average; compare Sect.14.3.2.

(For H_o true and normally distributed observations these are the Maximum-Likelihood estimates of the population parameters, given by eqs.(9.11),(9.12), when the σ_i are approximated by the observed errors Δx_i.)

If H_o is true we expect that the weighted sum of the squared deviations from the weighted mean value

$$X^2 = \sum_{i=1}^{N} w_i (x_i - \bar{x})^2 \qquad (14.60)$$

will be an approximate chi-square variable with N-1 degrees of freedom; hence X^2 is a suggestive test statistic for the hypothesis of equal population means. If the population means are not equal (*i.e.* if H_o is not true) we are likely to obtain values of X^2 which are, on the average, too large. Consequently we shall use a one-sided test for equality of the means, with critical region at the upper tail of the $\chi^2(N-1)$ distribution.

Whenever the observed value X^2_{obs} as calculated from eq.(14.60) comes out with a magnitude comparable to the mean value N-1 for the $\chi^2(N-1)$ distribution this is taken as support for the assumption of consistent measurements. One will then adopt the weighted mean value $\bar{x}$ and the error $\Delta\bar{x}$ in this quantity from eqs.(14.58),(14.59) as the best estimates for the true value μ and its error σ.

If the χ^2 test on the mean values fails at the chosen significance level, *i.e.* X^2_{obs} exceeds the critical value $\chi^2_\alpha(N-1)$, the reason can sometimes be that one, or a few, of the measurements deviates strongly from all others. Such deviations are conveniently detected by plotting the data in an *ideogram* as described in Sect.6.2.5. One assigns to each measurement a normalized Gaussian curve of mean x_i and width Δx_i and sums all the curves. If the resulting envelope has a single fairly smooth peak centered at $\bar{x}$ the measurements are probably reasonably consistent, but if some secondary peak shows up clearly separated from the peak at $\bar{x}$ this indicates that the measurements responsible for the secondary peak are incompatible with the rest. If no natural explanation can be found for the odd behaviour it is recommended to reject the deviating measurements and repeat the calculations with the reduced number of observations.

When X^2_{obs} is large compared to the expectation value, but not so large that the hypothesis of equal mean values must be abandoned, it has in particle

physics become customary to retain the averaging of the weighted observations but to introduce a *scale factor* to increase the errors. Since one usually does not know which, if any, of the measurements or experiments may be wrong one makes the rather arbitrary assumption that all experiments have underestimated their errors, and in the same proportion, so that all errors should be *adjusted* upwards by some common factor S. The Particle Data Group accordingly define S by

$$S = \left(\frac{X^2_{obs}}{N-1}\right)^{\frac{1}{2}} . \qquad (14.61)$$

The above procedure is used for the calculation of S if all experiments have errors of about the same size. The new X^2_{obs} obtained with the adjusted errors will then of course be just N-1. If the experiments have widely varying errors the Particle Data Group recommend a slight modification which consists in disregarding the less precise experiments in the evaluation of S, thereby obtaining a scale factor which is sensitive to the most precise measurements.

Note that the scaling procedure for errors in no way affects the weighted average value $\bar{x}$ from eq.(14.58), but only amounts to increasing the error in this quantity by the factor S compared to the unscaled error from eq.(14.59).

It should also be stressed that it is of great importance that the individual error estimates Δx_i have been checked before they are used to calculate a weighted mean from eq.(14.58) and a numerical value for the test statistic of eq.(14.60). For instance, it is recommended to make sure whenever possible that these errors are not smaller than those implied by the minimum variance bound.

Exercise 14.5: Five bubble chamber experiments have estimated the mass of the Ω^- hyperon as follows: (1673.0 ± 8.0)MeV, (1673.3 ± 1.0)MeV, (1671.8 ± 0.8)MeV, (1674.2 ± 1.6)MeV, (1671.9 ± 1.2)MeV. Are these results consistent at a significance level of 5%? An old measurement using nuclear emulsions estimated the mass of a negative particle at (1620 ± 25)MeV. Is this result consistent with the more recent bubble chamber data for the Ω^-?

Exercise 14.6: Six different experiments have measured the complex CP violation parameter η_{+-0} in the decay $K^o \to \pi^+\pi^-\pi^o$. The values reported for the real and imaginary parts of this parameter are the following:

Experiment	Re η_{+-o}	Im η_{+-o}
1	0.01 ± 0.21	0.33 ± 0.46
2	0.47 ± 0.20	- 0.10 ± 0.34
3	0.08 ± 0.44	0.32 ± 0.40
4	- 0.30 ± 0.28	0.08 ± 0.58
5	0.05 ± 0.30	- 0.15 ± 0.45
6	2.75 ± 0.62	0.50 ± 0.62

Considering the real and the imaginary parts of the parameter separately, are the results from the different experiments consistent at the 5% level? What are the scale factors for the two series of observations using all measurements? What would the scale factors be if the last experiment was disregarded?

Exercise 14.7: The mean lifetime τ of the π^o meson has been measured by 11 experiments, six of which used nuclear emulsions as detectors and five used counter technique. The following data have been reported:

Nuclear emulsion technique		Counter technique
τ (in units of 10^{-16}sec)	Number of events	τ (in units of 10^{-16}sec)
1.9 ± 0.5	76	1.05 ± 0.18
2.3 ± 1.1	45	0.730 ± 0.105
2.8 ± 0.9	88	0.6 ± 0.2
1.7 ± 0.5	75	0.56 ± 0.06
1.6 ± 0.6	67	0.9 ± 0.068
1.0 ± 0.5	232	

(i) Verify that the errors given by the emulsion experiments are larger than the theoretically smallest possible errors. (Hint: The MVB for τ in the p.d.f. $f(t|\tau) = 1/\tau \exp(-t/\tau)$ is τ^2/n.)
(ii) Show that the emulsion experiments are internally consistent.
(iii) Are the counter experiments internally consistent?
(iv) Do the two techniques lead to consistent results?

14.4 GOODNESS-OF-FIT TESTS

Up to this point we have in this chapter considered various types of parametric tests, in which the task has been to decide whether, in the light of a set of observations, the numerical values of certain parameters describing a probability distribution are compatible with some specified values defining the parametric hypothesis.

We will now proceed to discuss more general problems involving non-parametric hypotheses, with which we shall be concerned for the remainder of the chapter. We start by considering *tests of goodness-of-fit*. For definiteness,

let $x_1, x_2, \ldots, x_n$ be sample values for a random variable x whose true probability function f(x), continuous or discrete, is not known and let $f_o(x)$ be some particular specified distribution. To test the simple hypothesis

$$H_o: f(x) = f_o(x) \tag{14.62}$$

on the basis of the sample values is then a typical goodness-of-fit problem.

In testing goodness-of-fit we shall, as before, need a test statistic whose distribution, assuming H_o true, defines a critical region and an acceptance region with probabilities α and $1-\alpha$, respectively. The situation can be different from that of the previous sections in that we may now not formulate an alternative hypothesis H_1, since H_1 can be the ensemble of all conceivable hypotheses different from H_o. Thus H_1 is often left unspecified, and the power of the test not taken into account.

The goodness-of-fit test most commonly used is *Pearson's χ^2 test*, which will be discussed extensively in the following sections. This test is strictly speaking exact for large samples only, and otherwise approximate. A second test for goodness-of-fit, the *likelihood-ratio test*, is valid for all samples sizes, but since its test statistic has a distribution which is not known in general, this test is less applicable to practical problems; moreover, since the likelihood-ratio test for large samples can be shown to be equivalent to Pearson's χ^2 test, it will not be considered in this book. Instead we shall discuss the *Kolmogorov-Smirnov test*, which is particularly useful in the case of small samples when the conditions for Pearson's χ^2 test are not satisfied.

14.4.1 Pearson's χ^2 test

In this and subsequent sections, the presentation resembles that of Sects. 10.4 and 10.5 regarding concepts and notation. However, while in these sections the topic was parameter estimation (by the Least-Squares approach), the estimation problem as such will be of subordinate importance to us here. In fact, to emphasize that our present concern is the goodness-of-fit between observed data and some theoretical model we shall for the moment assume that the data have *not* been used to infer the numerical values of parameters in the model. In other words, the distribution defining H_o is presently assumed to be specified independently of the observations which constitute the basis for the test of fit. The necessary modifications for a situation where the data have been

used to specify parameters in H_o will be considered in Sects. 14.4.3 and 14.4.4 below.

We assume now that n observations on the variable x belong to N mutually exclusive classes, such as successive intervals in a histogram, non-overlapping regions in two-dimensional plot, *etc.* The number of events $n_1, n_2, \ldots, n_N$ in the different classes will then be multinomially distributed, with probabilities p_i for the individual classes as determined by the underlying distribution. The simple hypothesis we wish to test specifies the class probabilities according to some precription,

$$H_o: p_1 = p_{o1},\ p_2 = p_{o2},\ \ldots,\ p_N = p_{oN}, \tag{14.63}$$

where

$$\sum_{i=1}^{N} p_{oi} = 1 . \tag{14.64}$$

Equivalently, the hypothesis specifies the numbers predicted for the different classes, given that the total number in all classes is n. To test whether the set of predicted numbers np_{oi} is compatible with the set of observed numbers n_i we take as our test statistic the quantity

$$X^2 = \sum_{i=1}^{N} \frac{(n_i - np_{oi})^2}{np_{oi}} , \tag{14.65}$$

or, in an equivalent form which is easier to compute,

$$X^2 = \frac{1}{n} \sum_{i=1}^{N} \frac{n_i^2}{p_{oi}} - n . \tag{14.66}$$

When H_o is true this statistic is approximately chi-square distributed with N-1 degrees of freedom.

To qualify this statement we observe that each term in eq. (14.65) is the square of a quantity $(n_i - np_{oi})/\sqrt{np_{oi}}$, where the numerator measures the difference between the observed and hypothetical value for class i; if H_o is true this difference has expectation value zero. Also, the number of observations in any class i can be considered a Poisson variable, with mean value equal to its variance, and for H_o true this mean is equal to np_{oi}. If, further, np_{oi} is large enough to approximate a Poisson variable to a normal variable, then $(n_i - np_{oi})/\sqrt{np_{oi}}$ is approximately N(0,1). The statistic X^2 constructed as a sum of squares of such variables will accordingly be an approximate chi-square variable with a number of degrees of freedom equal to the number of independent terms in the sum, which is here N-1 due to the normalization condition.

If H_o is true, and the experiment is repeated many times under the same conditions with n observations, the actual values obtained for X^2_{obs} will therefore be distributed nearly like $\chi^2(N-1)$; in particular, the average value for X^2_{obs} will be $\simeq N-1$ and the variance $\simeq 2(N-1)$. If, on the other hand, H_o is not true, the expectation for each n_i is not np_{oi}, and the sum of squared terms $(n_i - np_{oi})/\sqrt{np_{oi}}$ will tend to become on the average larger than if H_o were true. Hence it seems reasonable to reject H_o if X^2_{obs} becomes too large and to adopt a one-sided test for H_o taking the critical region at the upper tail of the appropriate chi-square distribution. - See also Exercise 14.8 below.

It has been the practice at times to reject H_o for very small as well as very large values of X^2_{obs}, *i.e.* to use a two-sided test, the argument being that a very small X^2_{obs} might indicate some bias in the data towards the hypothetical values. Although such a bias may certainly cause an improbable, small value for X^2_{obs}, it appears on the other hand even more unlikely to obtain a low X^2_{obs} value using a wrong hypothesis; the two-sided test thus seems less justified and has been abandoned.

Since the Pearson χ^2 test is insensitive to the signs of the differences $(n_i - np_{oi})$ it is strongly recommended, also in the cases when there is no reason to suspect the hypothesis from the X^2_{obs} value, to examine the hypothetical and observed numbers and look for systematic trends over the variable range. Quite frequently a hypothesis which corresponds to an acceptable chi-square probability can be ruled out from the pattern of signs in the deviations between data and model; see the example in Sect. 14.6.7.

Exercise 14.8: (Expectation values of the X^2 statistic)
(i) Show that, for any underlying multinomial distribution with class probabilities $p_1, p_2, \ldots, p_N$ the expectation value of Pearson's test statistic eq. (14.66) is

$$E(X^2) = \sum_{i=1}^{N} \frac{p_i(1-p_i)}{p_{oi}} + n\left(\sum_{i=1}^{N} \frac{p_i^2}{p_{oi}} - 1 \right) .$$

In particular, if H_o is true (*i.e.* if $p_i = p_{oi}$ for all i), one has, *for any sample*

size n,

$$E(X^2|H_o) = N-1 .$$

This represents a generalization of the already established asymptotic result, when X^2 is $\chi^2(n-1)$ and therefore has mean value N-1.
(ii) Show, by minimizing $E(X^2)$ with respect to the p_i under the constraint $\Sigma p_i = 1$, that as $n\to\infty$, $E(X^2)$ will be minimal if $p_i = p_{oi}$. (Hint: Use the Lagrangian multiplier method.) Hence, for any hypothesis H_1 specifying a set of class probabilities differing from H_o one will have, asymptotically,

$$E(X^2|H_1) > N-1 ,$$

suggesting that the critical region for Pearson's χ^2 test should be at the upper tail of the chi-square distribution.
(iii) If all class probabilities are equal under H_o, *i.e.* if all $p_{oi} = \frac{1}{N}$, show that, *for any sample size*,

$$E(X^2|H_1) = (N-1) + (n-1)(N\Sigma p_i^2 - 1),$$

which is always larger than N-1 for hypotheses differing from H_o.

14.4.2 Choice of classes for Pearson's χ^2 test

The problem of how to divide the variable range into classes or bins, and thereby fix the number of classes, was also considered in connection with the Least-Squares method of parameter estimation, Sect.10.5.2.

The asymptotic chi-square behaviour of the X^2 statistic for the Pearson χ^2 test of goodness-of-fit is, strictly speaking, only proved to be correct if the class division is made without any reference to the observations. This is so because the formulation above assumed the observations to be randomly (multinomially) distributed over a set of predefined classes, each corresponding to a specified probability. In practice, however, the choice of class boundaries is often made after the data have been obtained and the general pattern of the observations has emerged and can be taken into account. This practice is justified by the fact that, for infinite n, the distribution of X^2 will be $\chi^2(N-1)$ for any partition with N classes, provided H_o is true.

Pearson's χ^2 test relies on the approximation of a multinomial to a multinormal distribution, since it assumes an approximate standard normal behaviour of all terms $(n_i - np_{oi})/\sqrt{np_{oi}}$, and this requires sufficiently large numbers of expected events within each of the N classes. Clearly, the grouping of individual observations into a class and representing them by some common average value of the variable necessarily implies a certain loss of information which is in itself unwanted. In applying this test to compare model and data the physicist may therefore experience the choice between Scylla and Charybdis: he should

use relatively few classes to comply with the normality requirement, and many classes to reduce the loss of information from the data.

To solve this dilemma it is usually recommended to group the observations as little as possible and in such a way that the number of events within all classes will be in accordance with the normality requirement, which is customarily specified as a minimum of 5 expected events in each class. If the number of degrees of freedom is not too small, say at least 6, one or two classes may be allowed to have even less than 5 expected events.

For observations on a *discrete* variable the class division will usually represent no special problem, since the class boundaries will more or less suggest themselves and possible necessary groupings of events, particularly in the tails of the distribution, can be made to correspond to the minimum requirements mentioned above.

For a *continuous* variable there is no natural suggestion for the class determination from the hypothetical distribution itself, and essentially two different approaches can be thought of for the subdivision of the variable range (Sect.10.5.2): Either the range is divided into classes of *equal width*, or it is divided to correspond to classes of *equal probability*. The equal-width method is arithmetically simpler than the equal-probability method which may require a somewhat higher level of computational sophistication and therefore, apparently, enjoys less popularity among physicists. The equal-probability method can, however, be shown to be theoretically advantageous.

Assuming equal-probability partition and sufficiently large samples, it is possible to establish a relation for the optimum number of classes which maximizes an approximate power function for Pearson's χ^2 test (see Kendall and Stuart, Chapter 30, Vol.2). The optimum number of classes is found to increase in proportion to $n^{2/5}$ for fixed power and significance or, *equivalently*, *the optimum* expected event number in each class increases as $n^{3/5}$. Specifically, maximizing at a power level $1-\beta = \frac{1}{2}$ for $n = 200$ one finds that the optimum class number N is 31 (27) for a significance level of 5% (1%), corresponding to between 6 and 8 expected events per class, just barely in accordance with the normality requirement. For higher power levels the optimum number of classes is lowered, corresponding to more expected events per class.

Whether the equal-probability method generally improves the power of Pearson's χ^2 test compared to the equal-width method is, however, not at all clear. Indeed, one may suspect that the degree of "fit" will be most critical at the extremes of the variable range, and under such circumstances the equal-probability method may well result in a loss of sensitivity in these regions.

14.4.3 Degrees of freedom in Pearson's χ^2 test

Tests for goodness-of-fit are frequently used in conjunction with estimation problems. Typically, some theory or model is available which includes a certain number of parameters that are either known to an unsatisfactory precision, or not known at all. This model may be said to constitute a composite hypothesis, since it is not completely specified, but spans a certain region of parameter space. If then the data are used to obtain estimates for the unknowns the effect is to reduce the allowed region of parameter space to a single point. This corresponds to making the model a simple hypothesis, which is subsequently put to test for goodness-of-fit.

For a Least-Squares estimation we know from Sects.10.4.3 and 10.4.4 that the comparison between data and fitted model is made using the chi-square distribution with a number of degrees of freedom equal to the number of independent observations minus the number of independent parameters estimated. This procedure is exact only in the limit of infinitely many observations and with a linear parameter dependence; otherwise it is an approximation. Thus, if there are L parameters in H_o which are estimated by the LS method and N classes subject to an overall normalization condition, Pearson's χ^2 test for goodness-of-fit consists in comparing the fitted (minimum) value X^2_{min} to the chi-square distribution with (N-1-L) degrees of freedom.

Whenever unknown parameters are to be inferred from the data it is generally recommended to use the Maximum-Likelihood method for this estimation and to use the X^2 statistic of eqs.(14.65) or (14.66) with the estimated predicted frequencies $p_{oi} = \hat{p}_{oi}$ for testing the goodness-of-fit. The comparison will then be made to the chi-square distribution with a number of degrees of freedom which is smaller than N-1. Two possibilities can be thought of:

(i) The data were grouped in N classes and the L unknown parameters estimated by the multinomial Maximum-Likelihood method as described in Sect.9.9. - The X^2 statistic can then be shown to have asymptotically a $\chi^2(N-1-L)$ distribution, just as if the parameters were estimated by the Least-Squares method.

(ii) The L unknown parameters were estimated by the ordinary Maximum-Likelihood method from the original ungrouped observations. - In this situation it is a little more problematic to perform the goodness-of-fit test, because the X^2 statistic does no longer have a simple chi-square distribution. Assuming again N terms in the X^2 sum it can be shown that X^2 will have a distribution which is bounded by two chi-square distributions with (N-1) and (N-1-L) degrees of freedom, respectively. When N is a large number the difference between the two distributions may be ignored, and the critical value for X^2 at a given significance can be found from the corresponding $\chi^2(N-1)$ distribution. For N small, however, it may be necessary to check that the calculated X^2 exceeds the critical value for both distributions $\chi^2(N-1)$ and $\chi^2(N-1-L)$ before rejecting the fit.

14.4.4 General χ^2 tests for goodness-of-fit

In our considerations so far we have assumed that the test statistic X^2 has been expressed in terms of N class probabilities, which are not all independent but must add to unity. This is equivalent to requiring equally many predicted and observed events when summed over all classes, and implies a reduction in the number of degrees of freedom of one unit in the comparison for goodness-of-fit. Quite frequently, however, the model which we wish to test gives definite predictions for the absolute numbers of expected events f_i in the separate classes. Rather than eqs.(14.63),(14.64) the hypothesis is

$$H_o: \quad f_1 = f_{o1}, \quad f_2 = f_{o2}, \ \ldots, \quad f_N = f_{oN}. \tag{14.67}$$

and the natural test statistic is then

$$X^2 = \sum_{i=1}^{N} \frac{(n_i - f_{oi})^2}{f_{oi}} \tag{14.68}$$

which, for H_o true, is assumed to be approximately $\chi^2(N)$ when all N classes have sufficiently many events. Again, if H_o involves parameters whose numerical values are inferred from the data by the usual estimation methods, the number of degrees of freedom will have to be reduced as described in the previous section.

For more complicated problems, where there are constraint equations relating measurable and unmeasurable quantities (parameters), and also correla-

tion terms between the measurements, a χ^2 test for goodness-of-fit can be based on a test statistic of the type

$$X^2 = (\underline{y}-\underline{f}_o)^T V^{-1}(\underline{y})(\underline{y}-\underline{f}_o) \qquad (14.69)$$

where $\underline{f}_o$ is the vector of fitted quantities defining H_o and $\underline{y}$ the observations with covariance matrix $V(\underline{y})$. In particular, if a Least-Squares estimation has been used, the number of degrees of freedom is (K-J), where K is the number of constraints and J the number of unmeasurables in the problem.

For a chosen significance level 100α % of the χ^2 test the lower limit χ^2_α of the critical region is generally given by the relation

$$\alpha = \int_{\chi^2_\alpha}^{\infty} f(u;\nu)du = 1 - F(u=\chi^2_\alpha;\nu) \qquad (14.70)$$

where $f(u;\nu)$ is the chi-square p.d.f., $F(u;\nu)$ the cumulative integral of the same p.d.f., and ν the appropriate number of degrees of freedom. If X^2_{obs} as calculated with the observations exceeds the critical value χ^2_α implied by eq.(14.70) (and obtained, for example, from a tabulation with fixed percentage points such as Appendix Table A8) the hypothesis H_o will have to be rejected at the chosen significance level; if X^2_{obs} comes out smaller than χ^2_α there is no reason to reject H_o on the basis of the χ^2 test.

Common practice among physicists is to convert the actually observed value X^2_{obs} for the fit to an equivalent chi-square probability P_{χ^2} as implied by the relation

$$P_{\chi^2} = \int_{X^2_{obs}}^{\infty} f(u;\nu)du = 1 - F(u=X^2_{obs};\nu)\,. \qquad (14.71)$$

This probability is most conveniently obtained from curves of the cumulative chi-square distribution, such as Fig. 5.2, but can also be found by interpolation in the standard tables with fixed percentage points.

It may be appropriate to stress that although a very bad fit (with a high X^2_{obs} value and low P_{χ^2}) can be a sufficient reason for rejecting a hypothesis, a good fit is in itself inconclusive as long as other hypotheses have not been tried. In fact, instead of using the phrase "we accept the hypothesis" it will perhaps be more appropriate to express the matter as "we fail to reject the hypothesis". A χ^2 test for goodness-of-fit may, however, provide additional support for a hypothesis which is already plausible for other reasons. Thus, when a physicist says he accepts a certain hypothesis he probably has been convinced from physical arguments rather than purely statistical considerations.

14.4.5 Example: Kinematic analysis of a V^o event (2)

To illustrate the use of the χ^2 test for goodness-of-fit we shall go back to the example of Sect.10.8.2 and see how this test can be used to decide the identity of neutral strange particles observed as V^o events in a bubble chamber. Depending on whether the positive decay particle is a proton or a pion there are two hypotheses for each V^o:

H_o : V^o is $\Lambda \to p + \pi^-$,

H_1 : V^o is $K^o_s \to \pi^+ + \pi^-$.

We assume now that the flight direction of the V^o has been determined from the measured production and decay points, and that the momenta and angles of both decay particles have been measured. The measurable variables $\underline{\eta}$ then constitute an 8-component vector, involving two angles for the V^o, and momentum and two angles for each of the two decay particles. The only unmeasurable variable $\underline{\xi}$ in the problem is (the magnitude of) the V^o momentum. Since there are four constraint equations for energy and momentum conservation this corresponds to a 3C-fit for the constrained Least-Squares estimation of all kinematic variables in a specified hypothesis.

If the correct hypothesis has been used for the fit the variable X^2 of eq.(14.69) with the fitted quantities $\underline{f}_o = \hat{\underline{\eta}}$ obtained in the final iteration of the minimization procedure, will be approximately $\chi^2(3)$. Fixing, for example, the significance level at 1% the critical value, as read off from Appendix Table A8, is $\chi^2_{.01} = 11.345$. Hence, at the 1% level, we shall reject a hypothesis if the observed value X^2_{obs} exceeds $\chi^2_{.01}$, and otherwise accept it.

Specifically, let us consider an event for which relevant numbers are given in the following table. The angles for the V^o have been obtained from the production and decay points with measured (x,y,z) coordinates (-44.4±.14,-1.8±.17, -16.2±.26) and (-28.8±.15,-5.3±.16,-16.0±.26) in cm, respectively. All measured quantities have uncorrelated errors.

		Momentum (MeV/c)	Dip angle (radians)	Azimuth angle (radians)
Measured quantities	p π^+	1535±72 1479±60	0.022±0.006 0.019±0.006	6.107±0.007 6.111±0.006
	π^-	378±18	-0.097±0.016	5.768±0.014
Fitted quantities, hypothesis H_0	p π^-	1564±72 354±11	0.022±0.006 -0.091±0.016	6.106±0.007 5.781±0.012
Fitted quantities, hypothesis H_1	π^+ π^-	1831±51 381±18	0.024±0.006 -0.118±0.016	6.124±0.006 5.719±0.011

In view of the preassigned 1% significance level and corresponding critical value of the test statistic, we shall from the observed values $X^2_{obs}(H_0)=3.6$ and $X^2_{obs}(H_1) = 26.7$ accept H_0 and reject H_1. We thus conclude that this particular V^0 is a Λ.

With the present example the numbers assure, with overwhelming plausibility, that the correct identity has been established for the V^0. In other cases the situation may be not so simple. For example, if both hypotheses give $X^2_{obs} < \chi^2_\alpha$ and thus are acceptable at the chosen significance α, the V^0 is kinematically ambiguous. If this ambiguity can not be resolved by ionization or other criteria, the practice among physicists is *either* to accept only the hypothesis with lowest X^2_{obs} (highest chi-square probability P_{χ^2}), rejecting the other, *or* to accept both hypotheses, including the V^0 in the two samples of Λ and K^0_s events with weights in accordance with the P_{χ^2} values for the two fits. - If both hypotheses give $X^2_{obs} > \chi^2_{.01}$ the event is rejected, and both final samples corrected for a 1% loss of true events.

14.4.6 The Kolmogorov-Smirnov test

Pearson's χ^2 test is undoubtedly the most popular non-parametric test used by physicists. However, other goodness-of-fit tests exist which avoid the binning of individual observations and may be more sensitive to the data. The most important of these tests is probably the *Kolmogorov-Smirnov test*, which in particular for small samples is superior to the χ^2 test, and has many nice properties when applied to problems in which no parameters are estimated.

Given n independent observations on the variable x we form an *ordered sample* by arranging the observations in ascending order of magnitude, $x_1, x_2, .., x_n$.

The cumulative distribution for this sample of size n is now defined by

$$S_n(x) = \begin{cases} 0 & x < x_1 \\ \frac{i}{n} & x_i \le x < x_{i+1} \\ 1 & x \ge x_n . \end{cases} \qquad (14.72)$$

Thus $S_n(x)$ is an increasing step function with a step of height $\frac{1}{n}$ at each of the observational points $x_1, x_2, \ldots, x_n$.

The Kolmogorov-Smirnov test involves a comparison between the observed cumulative distribution function $S_n(x)$ for the sample and the cumulative distribution function $F_o(x)$ which would occur under some theoretical model. We state the null hypothesis as

$$H_o: \quad S_n(x) = F_o(x) . \qquad (14.73)$$

For H_o true one expects that the difference between $S_n(x)$ and $F_o(x)$ at any point should be reasonably small. The Kolmogorov-Smirnov test looks at the difference $S_n(x) - F_o(x)$ at all observed points and takes as a test statistic*) the maximum of the absolute value of this quantity, thus

$$D_n = \max \left| S_n(x) - F_o(x) \right| . \qquad (14.74)$$

It can be shown that provided no parameter in $F_o(x)$ has been determined from the data, and assuming H_o true, The variable D_n has a distribution which is independent of $F_o(x)$, *i.e.* D_n is *distribution-free*. This holds irrespective of the sample size.

To be able to use D_n as a test statistic for testing H_o one must know its distribution function. It was shown by Kolmogorov that D_n has a cumulative distribution which for large n is given by

$$\lim_{n\to\infty} P\left(D_n \le \frac{z}{\sqrt{n}}\right) = 1 - 2\sum_{r=1}^{\infty} (-1)^{r-1} e^{-2r^2z^2} \qquad (14.75)$$

This relation is approximately valid already at $n \approx 80$. For finite n the D_n distributions can be found from recurrence relations.

Appendix Table A10 gives the exact critical values d_α of the test statistic D_n for $n \le 100$ as well as approximate values for the limiting case of n

---

*) Alternative formulations of related tests use the statistics

$$D_n^+ = \max\left(S_n(x) - F_o(x)\right) \quad \text{or} \quad D_n^- = \max\left(F_o(x) - S_n(x)\right)$$

large (last row), for different values of the significance α. It turns out that the approximate values obtained with the limiting entries are always larger than the exact ones. For instance, taking $\alpha = 0.05$ in a case with $n = 80$ the exact critical value of D_n is $d_{.05} = 0.1496$, while the approximate value becomes $1.358/\sqrt{80} = 0.1518$. If the null hypothesis H_o is tested at a significance level of 5% it should therefore be rejected if the largest observed deviation between $S_{80}(x)$ and $F_o(x)$ exceeds 0.15.

It will be seen from Appendix Table A10 that, if the sample size is small, rather large differences must be found between the cumulative distributions in order to detect significant deviations between data and hypothesis; hence unless this difference is considerable one shall not be able to falsify H_o. Indeed, the numbers of this table can be taken as an illustration of the general difficulty in constructing effective tests for small data samples.

It is worth noting that since D_n for H_o true has a distribution which is universal and independent of the theoretical $F_o(x)$, and furthermore is known for all n, one may use D_n to construct *confidence bands* for any continuous distribution function F(x). Whatever the true F(x) is we may write a probability statement about D_n as

$$P\left(D_n = \max |S_n(x) - F_o(x)| > d_\alpha\right) = \alpha , \qquad (14.76)$$

where as before d_α is the critical value of D_n corresponding to the significance α. The statement can be inverted to give a confidence statement about F(x),

$$P\left(S_n(x) - d_\alpha < F(x) < S_n(x) + d_\alpha \quad \text{all } x\right) = 1 - \alpha . \qquad (14.77)$$

This means that, at any point x, the cumulative distribution function F(x) will have a probability $(1-\alpha)$ of being larger than $(S_n(x) - d_\alpha)$ but smaller than $(S_n(x) + d_\alpha)$. Therefore, if one constructs a band of width $\pm d_\alpha$ around the empirical cumulative distribution $S_n(x)$ the probability is $(1-\alpha)$ that the true F(x) will lie entirely within this band. This provides an extremely simple and direct method for estimating a cumulative distribution function at a given confidence level. Obviously this inversion of the goodness-of-fit test into a confidence statement about F(x) rests upon the simple way D_n was defined to give a measure of the deviation between $S_n(x)$ and $F_o(x)$.

The technique described here can be used, for instance, to plan what

size an experimental sample should have in order to provide F(x) to a required precision. Numerically, let us demand an accuracy of better than 0.20 anywhere on F(x) at a confidence level of 90%. Then, from Appendix Table A10 we see from the entries for $\alpha = 0.10$ that $n \gtrsim 35$ will be necessary to have $D_n \leq 0.20$. With a required accuracy of better than 0.05 at the same confidence level we find from the asymptotic entry that the condition on n is $1.22/\sqrt{n} \leq 0.05$, which implies $n \gtrsim 600$.

It should be stressed that the considerations above apply only to situations where no unknown parameters are involved. If some of the parameters entering $F_o(x)$ have been estimated using the data the statistic D_n is no longer independent of $F_o(x)$, and the critical values d_α can not be obtained using the universal tables. However, in some fortunate situations the Kolmogorov-Smirnov test can still be used even in the presence of unknown nuisance parameters, provided appropriate tables over percentage points are available. For example, for the important class of problems where the estimation involves the mean value of an exponential distribution such tables can be found in a recent article by J. Durbin.

14.4.7 Example: Goodness-of-fit in a small sample

To illustrate the use of the Kolmogorov-Smirnov test of goodness-of-fit we consider a typical low-statistics experiment. Suppose that for 30 events one has measured the proper flight-time of neutral kaons decaying into the semileptonic final state $\pi^+e^-\nu$. With the kaons produced in an initially pure strangeness +1 state one can predict the p.d.f. $f_o(t)$ for the flight-time t under the assumption that only the $\bar{K}^o$ component with strangeness -1 contributes to the $\pi^+e^-\nu$ final state. This assumption defines the null hypothesis H_o which we want to test.

Figure 14.6(a) shows the step function $S_{30}(t)$ obtained for the sample of 30 flight-times and the predicted cumulative distribution function $F_o(t)$. From the largest deviation betweeen the experimental and theoretical curves we determine the actual value of the Kolmogorov test statistic of eq.(14.74) as

$$D_{30} = \max |S_{30}(t) - F_o(t)| = 0.17 .$$

From Appendix Table A10 we see that at the commonly chosen significance levels up to 10% we shall not be able to reject H_o on the basis of the 30 observations with

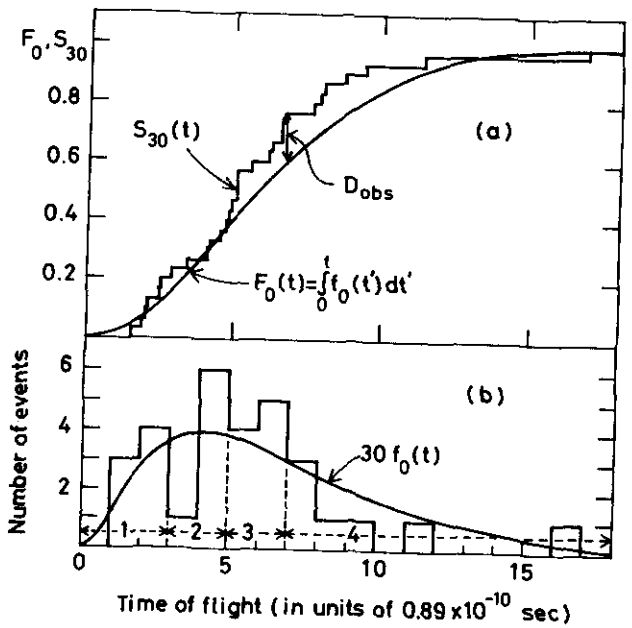

Fig. 14.6. Comparison of predicted and experimental distribution of flight times; (a) cumulative distribution (Kolmogorov-Smirnov test), (b) differential distribution (Pearson's χ^2 test).

this test. Indeed, for H_o true, we find by extrapolation of the table entries for n = 30 that there is a probability of about 0.25 that a larger maximum deviation than 0.17 would be found between the observed cumulative distribution function and the predicted $F_o(t)$.

For comparison, let us use the same observations to test H_o by the χ^2 method. For the flight-times of Fig. 14.6(a) a grouping of the data in the 4 intervals 0-3, 3-5, 5-7, and 7-18 (in units of the K^o_s mean lifetime) fulfil the recommendation given earlier for Pearson's χ^2 test with nearly equal probabilities in all bins and at least 5 entries in any bin, see Fig. 14.6(b). Assuming the X^2 statistic to be sufficiently well approximated to a chi-square variable under these circumstances one finds a chi-square probability of about 0.40 (X^2_{obs}=3.0 and 3 degrees of freedom).

14.5 TESTS OF INDEPENDENCE

Frequently, when data are available in differential form specifying several properties or attributes, it is desired to test whether these properties are independent of each other. The motivation for carrying out a test of independence can sometimes be a profound theoretical conjecture, for example, a scaling prediction for the shape of a spectrum of a kinematical variable. Before a claim is made on scaling behaviour it is then necessary to establish that the spectrum in question remains unchanged when, say, an incident energy is increased. More often the motivation is less subtle. The experimenter may simply want to find out whether observed events are uniformly distributed along a band in a Dalitz plot, whether transverse and longitudinal momenta are uncorrelated, *etc.*

An assumption of independence in the variables x,y,... can be stated as a null hypothesis where the joint probability distribution factorizes into separate probability distributions for the individual variables, *i.e.*

$$H_o: f(x,y,\cdots) = f_1(x)\, f_2(y)\cdots . \qquad (14.78)$$

A test problem of this kind can be approached along different lines of thought, some of which consisting in rephrasing the problem to make it analogous to those discussed in Sects.14.6 below.

It turns out that the χ^2 test is also adequate for providing answers to test problems of the above type, since a test statistic can be constructed in analogy with the Pearson statistic of eq.(14.65). To demonstrate this we shall be satisfied with considering a problem in two dimensions only, which will be sufficient for most practical purposes. The extension to higher dimensions may become somewhat awkward regarding notation but is conceptually simple and should be borne in mind by the experimenter working with high-statistics data samples.

14.5.1 Two-way classification; contingency tables

Suppose that observations can be classified according to two different attributes or properties A and B, and that there are I categories for the first attribute, $A_1, A_2, \ldots, A_I$, and J categories for the second, $B_1, B_2, \ldots, B_J$. Let the number of observations with attributes A_i og B_j be denoted by n_{ij} and let the total number of observations be n. With the notation

$$n_{i\cdot} \equiv \sum_{j=1}^{J} n_{ij}\,, \qquad n_{\cdot j} \equiv \sum_{i=1}^{I} n_{ij}\,, \tag{14.79}$$

the normalization condition is

$$\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} = \sum_{i=1}^{I} n_{i\cdot} = \sum_{j=1}^{J} n_{\cdot j} = n\ . \tag{14.80}$$

The numbers can conveniently be written in a *contingency table*, as in Fig. 14.7. A similar table*) can be written for the cell probabilities $P(A_i \cap B_j) \equiv p_{ij}$. If these probabilities were specified by some theory or model as definite numbers $p_{ij} = p^{o}_{ij}$, corresponding to a simple hypothesis, the agreement between the

	B_1	B_2	B_3	. . . .	B_J	$n_{i\cdot}$
A_1	n_{11}	n_{12}	n_{13}		n_{1J}	$n_{1\cdot}$
A_2	n_{21}	n_{22}	n_{23}		n_{2J}	$n_{2\cdot}$
A_3	n_{31}	n_{32}	n_{33}		n_{3J}	$n_{3\cdot}$
	⋮					⋮
A_I	n_{I1}	n_{I2}	n_{I3}		n_{IJ}	$n_{I\cdot}$
$n_{\cdot j}$	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$	. . . .	$n_{\cdot J}$	n

Fig. 14.7. Contingency table for two-way classification.

observed and predicted distributions could be checked by taking the sum

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - np^{o}_{ij})^2}{np^{o}_{ij}} \tag{14.81}$$

as a test statistic for an ordinary Pearson χ^2 test of goodness-of-fit with (IJ-1) degrees of freedom.

We assume now that our interest is not in the goodness-of-fit as such, but rather in the problem of deciding, on the basis of the data available,

*) A contingency table for probabilities was given already in Sect.2.3.12.

whether the A and B classifications can be said to be independent. The independence would imply that the conditional probability for attribute B_j, given A_i, is the same whatever A_i, and *vice versa*. Equivalently, the probability for having simultaneously properties A_i and B_j is equal to the product of probabilities for the separate occurrences. Thus we can state our (composite) null hypothesis as

$$H_o\colon\ P(A_i \cap B_j) = P(A_i)\cdot P(B_j) \qquad \text{all } i,j$$

(see Sect.2.3.6 for the general definition of independence). With a somewhat simplified notation we may write

$$H_o\colon\ p_{ij} = p_{i\cdot}\, p_{\cdot j} \qquad \text{all } i,j \tag{14.82}$$

where the marginal probabilities for attributes A_i and B_j are given by

$$p_{i\cdot} \equiv P(A_i) = \sum_{j=1}^{J} p_{ij}\,, \qquad p_{\cdot j} \equiv P(B_j) = \sum_{j=1}^{J} p_{ij}\,. \tag{14.83}$$

These marginal probabilities must satisfy the condition for exhaustive sets

$$\sum_{i=1}^{I} p_{i\cdot} = \sum_{j=1}^{J} p_{\cdot j} = 1\,, \tag{14.84}$$

which implies that only I-1 of the row probabilities and J-1 of the column probabilities are independent. These probabilities may be estimated from the observations as

$$\hat{p}_{i\cdot} = n_{i\cdot}/n\,, \qquad \hat{p}_{\cdot j} = n_{\cdot j}/n\ . \tag{14.85}$$

The individual cell probabilities may therefore, if H_o is true, be estimated by the products of the appropriate estimated row and column probabilities, $\hat{p}_{ij} = \hat{p}_{i\cdot}\,\hat{p}_{\cdot j} = n_{i\cdot}n_{\cdot j}/n^2$. A suggestive test statistic for testing independence in a two-way classification is consequently

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - n_{i\cdot}n_{\cdot j}/n)^2}{n_{i\cdot}n_{\cdot j}/n}\,, \tag{14.86}$$

or, in a form more convenient for computation,

$$X^2 = n\left\{\sum_{i=1}^{I}\sum_{j=1}^{J} \frac{n_{ij}^2}{n_{i\cdot}n_{\cdot j}} - 1\right\}. \tag{14.87}$$

Under H_o, X^2 will be approximately chi-square distributed, provided the numbers of events in the different cells are sufficiently large. The number of

degrees of freedom is equal to the number of independent observations minus the number of independently estimated unknowns, that is,

$$(IJ-1) - [(I-1) + (J-1)] = (I-1)(J-1) . \qquad (14.88)$$

Exercise 14.9: Show that the $\hat{p}_{i\cdot}$ (and similarly the $\hat{p}_{\cdot j}$) of eq.(14.85) are the Maximum-Likelihood estimates of the row (column) probabilities. (Hint: Write down the likelihood for obtaining the observations $n_{1\cdot}, n_{2\cdot}, \ldots, n_{I\cdot}$, and use the Lagrangian multiplier method to find the I probabilities $p_{1\cdot}, p_{2\cdot}, \ldots, p_{I\cdot}$ which maximize the log likelihood function under the constraint $\Sigma p_{i\cdot} = 0$.)

Exercise 14.10: In an I x I contingency table, show that the hypothesis of complete symmetry, H_o: $p_{ij} = p_{ji}$, $i,j = 1,2,\ldots,I$, can be tested by the statistic

$$\chi^2 = \sum_{i<j} \sum \frac{(n_{ij} - n_{ji})^2}{n_{ij} + n_{ji}}$$

which, under H_o, is asymptotically distributed as $\chi^2\{\frac{1}{2}I(I-1)\}$.

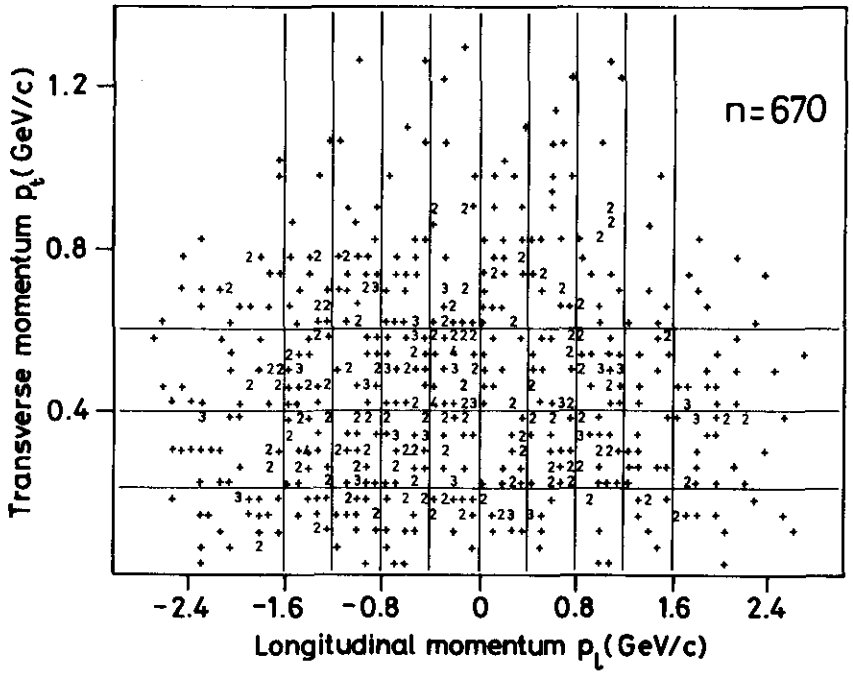

Fig. 14.8. Scatter diagram of centre-of-mass momentum components of Λ hyperons.

14.5.2 Example: Independence of momentum components

As an example on a test of independence in a two-way classification, consider the data of Fig. 14.8 which shows in the form of a scatter diagram the distribution of centre-of-mass momentum components for a sample of 670 Λ hyperons produced in pp collisions at 19 GeV/c.

We want to test, on the basis of this information whether the longitudinal and transverse momentum components can be regarded as independent variables in the underlying distribution. The following contingency table summarizes the data according to a chosen subdivision with 4 categories for the transverse momentum (classification A) and 10 categories for the longitudinal momentum classification B):

	B_1	B_2	B_3	B_4	B_5	B_6	B_7	B_8	B_9	B_{10}	$n_{i\cdot}$
A_1	20	8	13	12	13	11	7	5	7	10	106
A_2	20	23	22	26	15	14	21	20	9	21	191
A_3	21	22	15	24	35	10	19	22	8	15	191
A_4	20	18	25	19	25	25	19	18	6	7	182
$n_{\cdot j}$	81	71	75	81	88	60	66	65	30	53	670

By carrying out the summation according to eq.(14.87) with the numbers in this table one finds that the data correspond to $\chi^2_{obs} = 39.8$. The number of degrees of freedom, from eq.(14.88), is equal to $(4-1)(10-1) = 27$. Hence the chi-square probability for independence in the two momentum components is deduced to be about 5% .

14.6 TESTS OF CONSISTENCY AND RANDOMNESS

When a set of observations is used to estimate the parameters entering a p.d.f., it is generally tacitly assumed that the observations represent a sample which has been drawn at random from the population or universe. Thus the notion of a *random sample* presupposes that the observations acquired are typical and representative for the underlying distribution. If the assumption about randomness is not fulfilled the conclusions concerning the properties of the

population may be wrong or misleading. It is therefore of importance to have available some standard procedures for testing whether sets of observations may be regarded as random and free of systematic effects.

Particle physicists frequently find themselves in situations which call for investigation of randomness and consistency. When observations have been obtained through a series of measurements extending over time or space it may be necessary to check that the experimental conditions have remained the same throughout the experiment. Similarly, the data may have been acquired in two or more runs with complicated experimental set-ups, or collected by different laboratories participating in a collaboration experiment. In such situations, before any inferences are made from the combined data, it is important that consistency checks are performed to ensure that systematic differences do not exist between the separate samples. Likewise, before different experimental estimates of some parameter are combined to obtain a "best average" or "pooled estimate", it must be checked that the individual estimates do not depend on particular assumptions which are different for the different estimates. For example, if the values of mass and width of a resonance have been estimated by different groups it will be unjustified to deduce pooled estimates of the resonance parameters if the values reported by the separate groups have been derived from the raw observations using dissimilar assumptions about the resonance shape.

We saw in Sect.14.3 how tests of consistency can be formulated for observations which are *normally* distributed. Indeed, given the task of testing compatibility between experimental results, most physicists would without hesitance apply the procedures of Sects.14.3.2 and 14.3.7. Only seldom is the normality of the observations explicitly demonstrated to justify the use of these classical tests. It appears that normality is often taken for granted, although it is, admittedly, a very specific assumption about the nature of the universe from which the observations originate, and certainly not an indisputable fact.

Fortunately, theoretical studies have shown that certain test procedures are relatively insensitive to the specific form of the underlying distribution. These procedures possess a property aptly called *robustness*. This applies, for example, to the tests on population means referenced above, and may justify their use in cases where no dramatic deviation from normal behaviour

is expected. On the other hand, tests on population variances are very sensitive to departures from normality, restricting the usefulness of the procedures of Sect.14.3.3 to situations where the observations are manifestly close to normal.

In the following we shall for the investigation of randomness and consistency formulate various tests which avoid making specific assumptions about the form of the populations. These *distribution-free* tests are therefore generally valid, regardless of the underlying population; the distribution of their test statistic is determined by the number of equivalent permutations of elementary, equiprobable events and can, at least in principle, for finite samples be derived from purely combinatorial arguments. In abandoning the common normal theory methods for the more general distribution-free procedures one may have to pay a certain price, that of loosing "efficiency", or relative power. However, although it is generally true that distribution-free approaches are less efficient than tailored tests based on normality assumptions, theoretical investigations have shown that for populations which are not normal, the distribution-free tests may even be superior.

To test consistency between two or more experimental samples we shall from now on make no further assumption about the underlying distributions except that they are all equal. For observations of the continuous type these *tests of homogeneity* imply the null hypothesis

$$H_o: f_1(x) = f_2(x) = \cdots \qquad (14.89)$$

where f(x) is left unspecified. We will describe three different tests which can be used when the comparison involves only two samples. Of these, the *run test* is particularly simple, since it requires little more than just counting, while the *Kolmogorov-Smirnov* and the *Wilcoxon rank sum two-sample tests* are somewhat more sophisticated and may require some computational effort. The recommendation for practical work is, if inconsistency is suspected, to apply the run test first to see whether this simple test is capable of rejecting H_o; if the run test is inconclusive, the other tests should be applied in turn. These tests exploit more fully the information in the data and will be more powerful in detecting possible inconsistencies. - The run test has other useful applications; for example it can be used to give a rough check as to whether a set of observations is free from systematic trends. It can also be used to supplement Pearson's χ^2 test for

goodness-of-fit, of which it is independent under some conditions. This is an interesting feature because, in general, different tests on the same data are not independent, and hence the combining of outcomes of different tests is not trivial.

With more than two samples the hypothesis (14.89) can be tested by the *Kruskal-Wallis rank test*. When the underlying distributions are of the discrete type, the analogous multi-sample hypothesis can be examined by applying the well-known χ^2 test; this is described in Sect.14.6.12.

14.6.1 Sign test

A simple way of recording data is to note only whether each observation is smaller than, or larger than, some specified value. Although this rough method may imply the loss of a considerable amount of information in the observations, it is possible to construct useful tests for this kind of data. These *sign tests* are based on the binomial distribution law, which describes experiments with only two possible outcomes for individual events.

Let the variable x have a distribution of *median* value μ. We want to test the simple null hypothesis

$$H_o: \quad \mu = \mu_o \tag{14.90}$$

against the composite alternative

$$H_1: \quad \mu \neq \mu_o . \tag{14.91}$$

Suppose that of n observations on x, r observations are smaller than μ_o, while n - r are larger than this value. Clearly, if H_o is true, the distribution of the variables r and n - r will correspond to the binomial law with equal probabilities for the two outcomes $x < \mu_o$ and $x > \mu_o$ for the individual observations. Thus, H_o true implies an expected value of r equal to $\frac{1}{2}n$, and very small values of r as well as very large values (near n) are unlikely. To test H_o at the significance α we may therefore use the number r as a test statistic and take the rejection region at the two tails of the binomial distribution $B(r;n,p=\frac{1}{2})$. That is, we shall reject the assumption of a population median equal to μ_o if, among n observations, the number of times r when x is smaller than μ_o is such that

$$r < r_{\alpha/2} \quad \text{or} \quad r > r_{1-\alpha/2} . \tag{14.92}$$

The critical values $r_{\alpha/2}$ and $r_{1-\alpha/2}$ can be determined from tabulations of the cumulative binomial distribution. Since the statistic r is discrete one can, for a probability α in the lower tail, define the critical value r_α as that integer value which satisfies the inequality

$$\sum_{r=0}^{r_\alpha} B(r;n,\tfrac{1}{2}) \leq \alpha < \sum_{r=0}^{r_\alpha+1} B(r;n,\tfrac{1}{2}), \tag{14.93}$$

or, with the notation of Appendix Table A2,

$$F(x=r_\alpha;n,p=\tfrac{1}{2}) \leq \alpha < F(x=r_\alpha+1;n,p=\tfrac{1}{2}).$$

The sign test can be used to test whether a variable when recorded as a function of time remains "constant" and equal to a fixed value, or tends to change with time. Suppose, for instance, that in a bubble chamber exposure one may want to check that the (average) number of beam particles per pulse remains the same during the whole run, and equal to the optimum number requested by the designers of the experiment. If the beam intensity should vary significantly during the exposure, either by falling below the requested value (with the consequence of loss of useful events) or by rising well above it (too dirty pictures), one would adjust the experimental conditions and monitor the beam in such a way that the intensity is brought back to the optimum value.

Numerically, let us take a sample size 20 and choose a significance level of 20%. By looking up the entries for n = 20, p = 0.50 in Appendix Table A2 we see that r = 6 corresponds to a probability which is smaller than $\alpha/2 = 0.10$ (since $F(6;20,0.50) = 0.0577$), while r = 7 gives a probability higher than this value ($F(7;20,0.50) = 0.1316$). Hence the lower critical value is $r_{\alpha/2} = r_{.10} = 6$. Similarly, the upper critical value is $r_{1-\alpha/2} = r_{.90} = 14$, the two values being symmetrically positioned relatively to the expectation value for r, which is here $\frac{1}{2}\cdot 20 = 10$. - Therefore, if during the exposure we made a count on 20 randomly selected pictures and found that the number of tracks was smaller than the optimum number μ_o in more than 6 but in less than 14 pictures, then we would be satisfied with the state of affairs, and believe in the assumption of a constant beam intensity. If, on the other hand, we found that the number of pictures having less than μ_o tracks was *either* ≤ 6, *or* ≥ 14, we would reject H_o on the basis of this test. If repeated counts on a new sample of 20 pictures gave similar results, or

if a closer examination of the number tracks on each picture suggested a shift towards, say, a lower intensity, we would presumably adjust the conditions to bring the intensity up to the optimal before continuing the exposure.

The sign test assumes, for H_o true, each of the n observations $x_1, x_2, \ldots, x_n$ to have a constant probability for being smaller than the specified value μ_o. In this respect the order of the individual observations is immaterial, and the sequence is random under H_o. Instead of classifying the observations relative to a fixed, outside value μ_o, one can design a simple test of randomness by classifying the individual measurements with reference to one of the sample values, for example, relative to the sample median, see Sect.14.6.4.

14.6.2 Run test for comparison of two samples

To introduce the run test, suppose that we have two series of observations which have been arranged in terms of increasing (or decreasing) magnitude, giving the two ordered samples $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_m$. Without loss of generality we assume for simplicity in the following that $n \leq m$, since the case with $n > m$ can be considered by merely interchanging the x and y notation. We then combine the two ordered samples and arrange all the (n+m) observations in increasing order of magnitude, the result being for instance a series of the form

$$x_1 \quad x_2 \quad y_1 \quad x_3 \quad y_2 \quad x_4 \quad x_5 \quad x_6 \quad y_3 \ \ldots$$

The assumption of a common parent population implies that the x's and y's in this series should be well mixed, and if such a pattern is obtained it will be taken as support for H_o, while a pattern with, say, a preponderance of x's to the left of the chain and y's to the right will indicate systematic differences between the two sets of data.

A *run* is defined as a sequence of symbols of the same kind. Thus the chain above starts with a run of two x's, then follows a run of one y, and so on, altogether six runs are shown. The null hypothesis suggests that the number of runs r will be fairly large for the particular values of n and m. If the number of runs comes out very small compared to the maximum possible one might suspect that the probability for a definite outcome of one measurement has not remained the same in the x and y series. If, on the other hand, the number of runs is very large it is possible that the two series have not been obtained independently. Thus r very small or r very large could indicate that the assumption of consistency does not hold.

To use the number of runs r as a test statistic for the hypothesis H_o we have to find the probability distribution of r assuming H_o to be true. A total of (n+m) quantities can be arranged in (n+m)! different ways. Since, however, the mutual order within the x's as well as within the y's, has already been fixed, we can only have $(n+m)!/(n!m!) = \binom{n+m}{n}$ different permutations and, provided H_o is true, each of these permutations will have the same probability of occuring. To find the probability for a particular number of runs, say r, one must count all permutations giving rise to just r runs. This is a combinatorial problem which can be solved in a straightforward manner. The results of the computation is that the probability distribution for r is given by $(2 \leq r \leq n + m)$

$$\left.\begin{aligned} p(r=2k) &= 2\,\frac{\binom{n-1}{k-1}\binom{m-1}{k-1}}{\binom{n+m}{n}}, && \text{r even,} \\ p(r=2k-1) &= \frac{\binom{n-1}{k-2}\binom{m-1}{k-1}+\binom{n-1}{k-1}\binom{m-1}{k-2}}{\binom{n+m}{n}} && \text{r odd.} \end{aligned}\right\} \qquad (14.94)$$

This distribution has mean and variance given by

$$E(r) = \frac{2nm}{n+m} + 1, \qquad V(r) = \frac{2nm(2nm-n-m)}{(n+m)^2(n+m-1)}. \qquad (14.95)$$

To test H_o it is customary to take the critical region only at the lower tail of the run distribution. The argument for adopting a left-tail test is that it appears rather improbable to obtain a very large number of runs if H_o is not true. A shift in location, or a difference in dispersion between the two parent populations would most likely lead to less randomness and greater order in the combined series of x's and y's. Since all reasonable ("smooth") alternatives to H_o therefore in all likelihood would give rise to predominantly small values of r, the critical region should be restricted to the lower tail of the distribution.

Critical values of the run statistic are given in Appendix Table A11 for sample sizes up to 15 and for 4 different significance levels. Since r is a dicrete variable the critical value r_α corresponding to the significance α is

taken as that integer which satisfies the inequality

$$\sum_{r=2}^{r_\alpha} p(r) \leq \alpha < \sum_{r=2}^{r_\alpha+1} p(r) . \qquad (14.96)$$

The run test is simple and easy to apply when tabulations are at hand. For example, with two samples of size $n=6$, $m=8$ one finds by looking up the appropriate entries in Appendix Table AII that the assumption of consistency between the two samples will have to be rejected at the significance level 5% (or 1%) if the number of runs observed is ≤ 4 (or ≤ 3). However, it is clear that of all the information contained in the data, only little is actually used by this test. This means that the run test may fail to reject a hypothesis of consistency, while other, more elaborate tests which explore the data more thoroughly, may produce evidence for rejection.

For large samples, which in practice often is taken to mean m and n larger than 10, the probability distribution of eq.(14.94) is very close to normal. The appropriate run test statistic in this situation is

$$z = \frac{r - E(r)}{\sqrt{V(r)}} \qquad (14.97)$$

which is approximately $N(0,1)$.

The run test in the form described above is one of the least powerful distribution-free tests. It is in fact only meaningful when applied to samples of comparable size. If one sample is very much larger than the other then, in the combined ordered series, the observations from the smallest sample will almost certainly be separated from each other by observations from the larger sample; hence the number of runs will tend to become maximum, regardless of whether the assumption of identical parent populations is true or not.

Other tests based on runs have been devised, some of which exploit more fully the information in the data; one, for example, takes as a test statistic the length of the longest run. For a description of this and other run tests the reader should consult more specialized literature.

14.6.3 Example: Consistency between two effective-mass samples

To study the production of π^0 in antineutrino induced reactions in a heavy liquid bubble chamber two laboratories have measured electron-positron pairs, or γ's, pointing towards the reaction point and calculated the effective-mass $M_{\gamma\gamma}$ of pairs of γ's. One wants to test if the results from the two laboratories are consistent, given the following ordered samples of effective-masses (numbers in MeV):

Lab. X ($n=28$): 81, 82, 87, 93, 102, 104, 108, 112, 116, 122, 125, 131, 131, 133, 134, 139, 139, 142, 144, 146, 152, 156, 182, 202, 206, 216, 226, 270.

Lab. Y ($m=32$): 8, 12, 14, 16, 22, 26, 26, 50, 64, 68, 76, 79, 83, 88, 96, 97, 98, 99, 103, 105, 107, 113, 114, 115, 126, 128, 130, 132, 138, 150, 169, 171.

With these numbers the combined ordered series is

$y_1\ y_2\ y_3\ y_4\ y_5\ y_6\ y_7\ y_8\ y_9\ y_{10}\ y_{11}\ y_{12}\ x_1\ x_2\ y_{13}$

$x_3\ y_{14}\ x_4\ y_{15}\ y_{16}\ y_{17}\ y_{18}\ x_5\ y_{19}\ x_6\ y_{20}\ y_{21}\ x_7\ x_8\ y_{22}$

$y_{23}\ y_{24}\ x_9\ x_{10}\ x_{11}\ y_{25}\ y_{26}\ y_{27}\ x_{12}\ x_{13}\ y_{28}\ x_{14}\ x_{15}\ y_{29}\ x_{16}$

$x_{17}\ x_{18}\ x_{19}\ x_{20}\ y_{30}\ x_{21}\ x_{22}\ y_{31}\ y_{32}\ x_{23}\ x_{24}\ x_{25}\ x_{26}\ x_{27}\ x_{28}$,

which means that the observed number of runs is equal to 24. From eq.(14.95) the expected value and variance for the variable r with sample sizes $n=28$ and $m=32$ are, respectively,

$$E(r) = \frac{2\cdot 28\cdot 32}{28+32} + 1 \simeq 30.87 ,$$

$$V(r) = \frac{2\cdot 28\cdot 32(2\cdot 28\cdot 32-28-32)}{(28+32)^2(28+32-1)} \simeq 14.62 .$$

Using the large sample approximation eq.(14.97) for the test statistic we find that the actual value from the observations is

$$z_{obs} = \frac{24 - 30.87}{\sqrt{14.62}} = -1.80 .$$

Since the probability for a standard normal variable to be smaller than -1.80 is

$G(-1.80) = 1 - G(1.80) \simeq 0.036$, the hypothesis of consistency between the two samples will therefore have to be rejected from the run test at the 5% level.

By inspection of the original sample values it is seen that the striking difference between the two series of measurements is the large number of small $M_{\gamma\gamma}$ values observed by laboratory Y which are missing for laboratory X. As physicists we may try to explain the discrepancy between the data sets as being due to an experimental bias: The events with small values of $M_{\gamma\gamma}$ could be "wrong" events, in which, for example, one γ from π^0 decay was erroneously combined with a bremsstrahlung pair from the same γ. Therefore, given the data sets above, it is suggestive to ask laboratory Y to look more closely at their low-mass events on the scan table.

If one disregards the seven events with $M_{\gamma\gamma} < 40$ MeV from laboratory Y, one finds that the data for the samples of size 28 and 25 correspond to a run probability of about 0.17. Hence with the reduced number of events the run test finds no incompatibility between the two samples even at the 10% level.

14.6.4 Run test for checking randomness within one sample

The run test procedure for testing whether two samples have the same parent population can readily be adopted to test the assumption that a series of observations obtained sequentially, for instance by measurements performed at different times, can be considered free from systematic trends. The hypothesis of randomness is that all observations measure the same quantity, which implies that the order of the observations is immaterial.

Let the elements in a time-ordered series of observations be classified relatively to some value of the sample, such that an observation above this value is labelled by A and an observation below it by B. Observations coinciding with the chosen value can be ignored. The hypothesis H_o implies that at every position in the sequence the probability to have an A is the same, *i.e.* the probability for an A remains constant along the sequence, and likewise for B. The resulting series of A's and B's is then a pattern of symbols with properties analogous to the series of x's and y's of Sect. 14.6.2. We may therefore use the formulae for the run statistic given earlier. These become particularly simple if we choose to classify the observations relatively to the sample median, since by definition the number of A's and B's will then be equal. By putting $n = m$ in eq. (14.95) the

mean and variance for the number of runs become, respectively,

$$E(r) = n+1, \qquad V(r) = \frac{n(n-1)}{2n-1}. \tag{14.98}$$

For large samples the run statistic is then approximately $N(n, \frac{1}{2}n)$.

14.6.5 Example: Time variation of beam momentum

As an application of the run test on one series of observations, let us consider measurements on the beam momentum in a bubble chamber experiment. Suppose that measurements on 30 rolls of film ordered according to the time of exposure gave the following values for the average momentum of the incident tracks (numbers in GeV/c):

18.90	18.88	18.94	18.91	18.96	19.05	19.06	19.08	19.03	19.10
19.07	19.12	19.13	19.10	19.15	19.20	19.17	19.14	19.14	19.10
19.11	19.08	19.08	19.07	19.03	18.98	19.00	18.97	18.94	18.95

Do these numbers support the hypothesis H_o of a constant beam momentum during the exposure?

Already from an inspection of the *control chart* for the measurements above, shown in Fig. 14.9, one would be inclined to reject the assumption of constancy in the beam momentum. Indeed, the evidence from this chart is that the momentum has first increased, then decreased during the exposure.

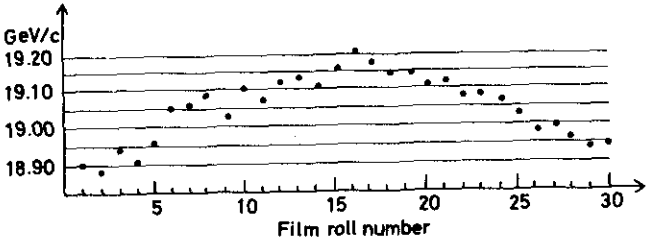

Fig. 14.9. Control chart for beam measurements.

We seak a numerical measure for our disbelief in H_o. The median value for the observations is 19.07 GeV/c, and the classification relative to this value gives the following series after the two observations coinciding with the

median are ignored:

B B B B B B B A B A A A A A A A A A A A A A B B B B B B

This series has a considerable degree of order with only 5 runs among the 28 symbols. From Appendix Table A11 we see that, for $n = m = 14$, the critical values r_α for the significances $\alpha = 0.05$, 0.025, 0.01, and 0.005 are, respectively, 10, 9, 8, and 7. Hence the probability to have as little as 5 runs must be considerably smaller than 1 ‰.

From eq.(14.98) the expectation value and variance for the number of runs r are, respectively,

$$E(r) = 14+1 = 15, \qquad V(r) = \frac{14 \cdot 13}{2 \cdot 14 - 1} \simeq 6.74 .$$

If we adopt the large sample approximation in this case the actual value of the approximate N(0,1) test statistic of eq.(14.97) becomes

$$z_{obs} = \frac{5 - 15}{\sqrt{6.74}} = -3.85 .$$

Hence the probability to have 5 or less runs among the 28 symbols with this approximation and the accuracy of Appendix Table A6 is $G(-3.85) = 1 - 0.99994 = 6 \cdot 10^{-5}$.

Exercise 14.11: For the example above, show that, for H_o true, the exact probability to have 5 or less runs is $5.97 \cdot 10^{-5}$.

14.6.6 Run test as a supplement to Pearson's χ^2 test

The Pearson χ^2 test for goodness-of-fit is based on a test statistic which is a sum of terms, each involving the square of the deviation between observed and predicted value in the different classes. From the construction of the test statistic this test therefore has the defect that knowledge regarding the signs of the deviations in the individual classes gets lost, and so does the order in which these deviations occur. The χ^2 test is, in other words, insensitive to the pattern of signs in the deviations. This pattern evidently contains some useful information about the correspondence between experiment and prediction, and should be possible to explore by a run test, which appears to suggest itself as an attractive supplement to the χ^2 test. In fact, since it utilizes

a limited amount of information in the observations, the run test is particularly meaningful only when used in conjunction with a χ^2 test on the same data.

For definiteness, consider the three situations sketched in Fig. 14.10. In (a) the prediction from the hypothesis under test roughly follows the observations over the variable range, resulting in a series of deviations between observed and hypothetical values which alternate in sign and hence give a fairly large number of runs. If the hypothetical distribution differs substantially from the observed in location, as in (b), it is clear that there will be a sequence of positive signs followed by a sequence of negative signs. Similarly,

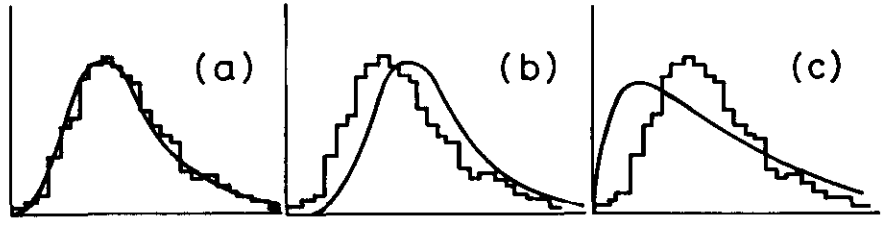

Fig. 14.10. Observed and hypothetical distributions, (a) comparable in shape and location, (b) differing in location, (c) differing in shape.

if the distributions differ mainly in shape, as in (c), the signs will occur in sequence of negative, positive, negative. Thus in both situations (b) and (c) the signs tend to being equal over large parts of the variable range, in contrast to the more random pattern expected if the two distributions had greater similarity. Since, therefore, alternatives to the true hypothesis most likely will lead to few runs, the critical region for the run test must be at the lower tail of the run distribution.

Let us assume that the observations are grouped in classes for the χ^2 test, and that the total number of expected events under H_o is the same as the total number of observed events. Suppose further that there are n classes where the deviation between observed and predicted value is positive, and in classes where it is negative. We count the number of runs in the sequence of signs and can find the run probability in the usual manner for the given n,m.

The implicit assumption in this procedure is that the order in which

the n + m signs occur is immaterial if H_o is true. In any class the probability to have a positive deviation is the same as the probability for a negative deviation. This will be the case if the conditions for the χ^2 test are satisfied, since then the number in any class is normally distributed. It has been verified, however, from an extensive series of random sampling experiments (F.N.David) that the procedure is valid even when the probability of obtaining a positive deviation in a class is four times that of obtaining a negative.

It can be shown that, for a simple hypothesis H_o, the run test is *asymptotically independent* of the χ^2 test, whereas if parameters in H_o are estimated from the data (except the overall normalization) the two tests are not independent and hence it is less meaningful to apply both. When the two tests can be regarded as independent they can be combined into a single test. Let P_1 be the probability of a value of X^2 larger than that observed and P_2 the probability of a value of r not larger than that observed. If the sample size is sufficiently large to approximate both probabilities to continuous variables, uniformly distributed between 0 and 1, then the variable

$$u = -2(\ell n P_1 + \ell n P_2) \tag{14.99}$$

will be $\chi^2(4)$, (see Exercise 5.6) and an appropriate test statistic. For this combined test the critical region is taken at the upper tail of the chi-square distribution.

14.6.7 Example: Comparison of experimental histogram and theoretical distribution

To illustrate the use of the run test as a supplement to the Pearson χ^2 test for goodness-of-fit we shall refer to Fig. 14.11. The histogram shows the observed distribution of the squared four-momentum transfer t from the target proton to a negative pion in the annihilation

$$\bar{p} + p \to \pi^+ + \pi^+ + \pi^- + \pi^-$$

at 1.2 GeV/c antiproton momentum. The smooth curve gives the distribution expected from a multi-Regge model of a particular form and specifies the simple null hypothesis H_o to be tested on the basis of the data. The area under the curve has been normalized to the number of events (n = 990) in the histogram.

From inspection of the diagram one gets the impression that the observed spectrum is somewhat more concentrated around t=0 than the theoretical

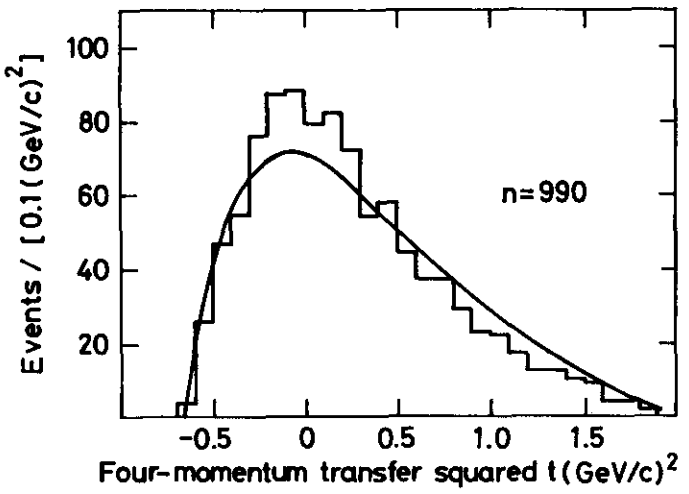

Fig. 14.11. Experimental and predicted distribution of four-momentum transfer.

distribution. To test the null hypothesis quantitatively one could use the class subdivision as implicit by the binning of the histogram. However, in view of the requirement of not too few events in each class, it is reasonable to group together several bins at the upper part of the spectrum. If this is done above 1.6 $(GeV/c)^2$ there will be a total of 24 classes, and only one class, the first bin, has an expected number of events below the recommended minimum of 5. The expected number $np_{oi} = f_{oi}$ within each class can be obtained by numerical integration of the curve; the result of the computation is

$$X^2_{obs} = \sum_{i=1}^{24} \frac{(n_i - f_{oi})^2}{f_{oi}} = \sum_{i=1}^{24} \frac{n_i^2}{f_{oi}} - n = 30.6 .$$

Because of the normalization constraint on the theoretical model the χ^2 test has $\nu = 24-1 = 23$ degrees of freedom. The observed value X^2_{obs} then corresponds to a chi-square probability $P_{\chi^2} = P_1 = 0.14$. Hence, from the χ^2 test there is little reason to suspect the hypothesis.

It will be seen from Fig. 14.11 that there are 17 bins where the histogram lies above the theoretical curve and 7 bins where the opposite occurs, with

an observed number of runs equal to 5. The probability to have no more than 5 runs among the 7+17 symbols is, from eq.(14.94), only $P_2 = 0.0034$. Hence from the run test we would certainly reject the proposition H_o.

For the combined test the actual value of the approximate $\chi^2(4)$ statistic u of eq.(14.99) is

$$u_{obs} = -2(\ln 0.14 + \ln 0.0034) = 15.3$$

which, from Appendix Table A8, corresponds to a combined probability < 0.005.

14.6.8 Kolmogorov-Smirnov test for comparison of two samples

The consistency between two experimental distributions of a continuous variable can also be checked by applying the *Kolmogorov-Smirnov two-sample test*. This is a distribution-free test which involves the comparison of the two cumulative sample distributions, analogous to the Kolmogorov-Smirnov goodness-of-fit test described earlier for the comparison one sample and a specified distribution defining H_o.

Let $S_m(x)$ and $S_n(x)$ be the cumulative distributions for the two (ordered) samples of sixe m and n, respectively (compare eq.(14.72) of Sect.14.4.6). The Kolmogorov-Smirnov two-sample test statistic D_{mn} is the maximum deviation between these two step functions over the entire variable range,

$$D_{mn} = \max|S_m(x) - S_n(x)|. \qquad (14.100)$$

Critical values of this statistic can be found, for instance, from Biometrika Tables for Statisticians, Vol.II, for samples sizes up to 25. In the limiting case, it can be shown that, for identical parent populations, *i.e.* for H_o true,

$$\lim_{m,n\to\infty} P\left(D_{mn} \le z\sqrt{\frac{1}{m}+\frac{1}{n}}\right) = 1 - 2\sum_{r=1}^{\infty}(-1)^{r-1} e^{-2r^2z^2}. \qquad (14.101)$$

This formula is analogous to eq.(14.75) for the one-sample case, and the probability distribution for the two-sample statistic D_{mn} is therefore related to the distribution for the one-sample statistic D_n. For m and n not too small, critical values D_α of D_{mn} can consequently be obtained from the corresponding critical values d_α of D_n, which have been tabulated in Appendix Table A10. The relationship is

$$D_\alpha = d_\alpha\sqrt{1+\frac{n}{m}} \qquad (14.102)$$

Hence, for a given significance level, the critical values in the two-sample test are always larger than the critical values for the one-sample goodness-of-fit test, in accordance with common sense expectation. However, when one sample is very much larger than the other, corresponding to very small steps in its cumulative distribution function, the critical values for the two-sample test are only slightly larger than the table values for the smallest sample size. In the extreme case, with $m \gg n$, the two-sample comparison is evidently identical to the goodness-of-fit test for one sample.

For an application of the Kolmogorov-Smirnov two-sample test, let us go back to the example of Sect.14.6.3. The cumulative distributions for the two effective-mass samples have been plotted in Fig. 14.12; the largest vertical

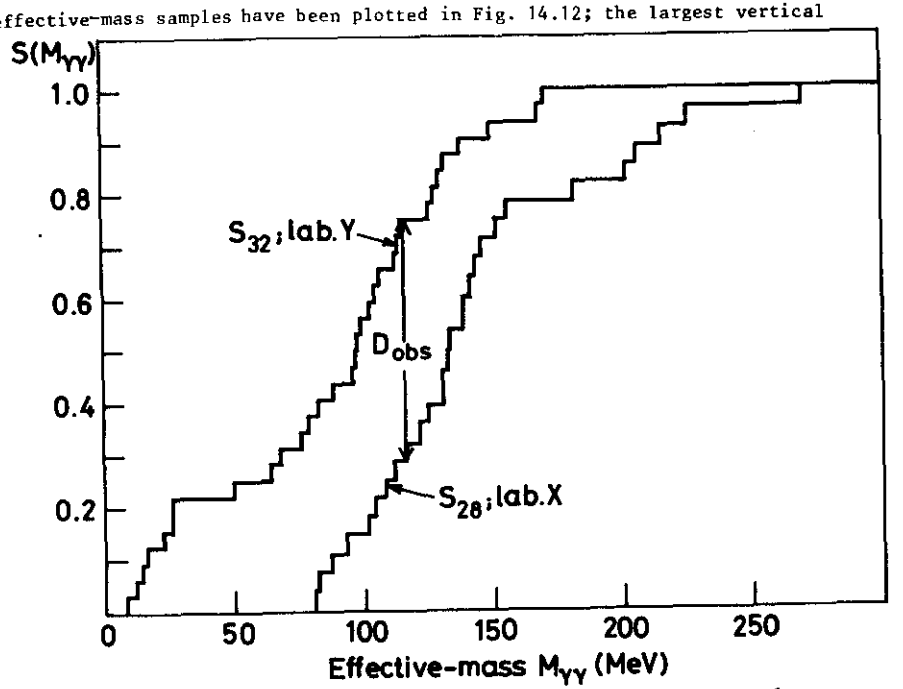

Fig. 14.12. Cumulative distributions of two effective-mass samples.

separation between the step functions is

$$D_{obs} = \max|S_{32}(M_{YY}) - S_{28}(M_{YY})| = \frac{24}{32} - \frac{8}{28} = 0.46 .$$

From Appendix Table A10 the critical value for a one-sample Kolmogorov-Smirnov test at a 5% significance level with $n = 28$ is $d_{.05} = 0.2499$. Hence, from eq. (14.102), the critical value for the two-sample test is

$$D_{.05} = 0.2499 \cdot \sqrt{1 + \frac{28}{32}} = 0.34 .$$

Since the observed largest deviation exceeds the critical value, $D_{obs} > D_{.05}$, we must reject the hypothesis of consistency between the two samples at the 5% level on the basis of the Kolmogorov-Smirnov test. This is of course not at all surprising, since even the run test would reject our assumption at the 5% significance level, as we saw in Sect.14.6.3.

Exercise 14.12: Repeat the Kolmogorov-Smirnov test for the example in text using the reduced data samples as explained in Sect.14.6.3. Show that this test, in contrast to the run test, rejects the hypothesis of consistency at the 5% level, also for the revised samples.

14.6.9 Wilcoxon's rank sum test for comparison of two samples

We have so far given several prescriptions for the comparison of two samples, and will now introduce the *Wilcoxon two-sample test*, or *Wilcoxon's rank sum test* for the same problem. We assume again that we have two ordered samples $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_m$ $(n \leq m)$; we want to test the hypothesis H_o that the two populations from which these samples originate are identical.

As described for the run test in Sect.14.6.2 we arrange the $(n+m)$ observations in increasing order of magnitude. In this combined ordered sample each observation is assigned a *rank*, equal to the order in which the observation occurs in the series. If some observations happen to be identical ("ties") they are all assigned the average value of the ranks these observations would have if they were distinguishable. The Wilcoxon test statistic W is now constructed as the sum of the n ranks for the observations from the x sample. If H_o is true we expect the x and the y observations to be well mixed in the combined series, and hence the value of W should be not "too small" and not "too large". Conversely, if the value for W comes out either "very small" or "very large" this would mean

that the x observations occur predominantly in one end of the combined ordered series, indicating that the hypothesis H_o is less likely to be correct. Hence it seems resonable to adopt a two-sided test for H_o.

Assuming H_o to be true the smallest possible value W_{min} for the rank sum W of the x sample is obtained when all x's are to the left of the combined ordered series and all y's to the right. Then W is nothing but the sum of the n first integer numbers or $W_{min} = n(n+1)/2$. Similarly, the largest possible value W_{max} will occur when all the x's are to the right and all y's to the left; hence W_{max} is the sum of all integers from $(m+1)$ to $(m+n)$, or $W_{max} = n(n+1)/2 + mn$. The distribution of probabilities p(W) for W between W_{min} and W_{max} can be obtained in a similar manner as indicated for the run statistic. The distribution is symmetric, and has mean and variance equal to

$$E(W) \equiv \bar{W} = \frac{n}{2}(n+m+1), \qquad V(W) = \frac{nm}{12}(n+m+1). \tag{14.103}$$

Critical values for the statistic W are reproduced in Appendix Table A12 for sample sizes up to 25 and for 6 different significances α. The table assumes one-sided tests, the critical value W_α being defined as that integer value for which

$$\sum_{W_{min}}^{W_\alpha} p(W) \leq \alpha < \sum_{W_{min}}^{W_\alpha+1} p(W) . \tag{14.104}$$

To obtain the critical values corresponding to a two-sided test at a significance level 100α % one reads off the lower critical value W_ℓ as the table entry in the appropriate column for $\alpha/2$. The upper critical value W_u can then be obtained from the symmetry property of the distribution, which implies

$$W_u = \bar{W} + (\bar{W} - W_\ell) = 2\bar{W} - W_\ell . \tag{14.105}$$

The value of $2\bar{W}$ is also given in the table for each n,m combination.

The distribution for W can be shown to tend to normal for n and m large. For large samples one can therefore use Wilcoxon's test with the approximate N(0,1) statistic

$$z = \frac{W - \bar{W} \pm \frac{1}{2}}{\sqrt{V(W)}} \tag{14.106}$$

where a "continuity correction" of $-\frac{1}{2}$ or $+\frac{1}{2}$ is added to the numerator depending on whether an upper or lower tail probability is being calculated. The normal approximation is good for most practical purposes with n and m both larger than 10. Even if n is smaller than 10 the approximation is fair, provided that m is not too much larger than n (fairly symmetric probability distribution) and the significance α nor too small (below 0.01, say).

Exercise 14.13: (Wilcoxon's rank sum test for randomness within one sample)
Discuss how the Wilcoxon rank sum test can be used to test whether a series of measurements is free from systematic trends.

Exercise 14.14: (Wilcoxon's rank sum test for independence)
A distribution of two variables f(x,y) is such that the variable y can take on only two values. Show how Wilcoxon's two-sample test can be used to test whether x and y are independent variables.

14.6.10 Example: Consistency test for two sets of measurements of the π^o lifetime

The mean lifetime of the π^o meson has been measured by several experiments which utilize essentially two different techniques. One set of experiments has used nuclear emulsions as detecting device, the other has used counter detectors (see Exercise 14.7). Ignoring for the moment the different accuracies of the experiments the results can be summarized by the following numbers giving the measured mean lifetime in units of 10^{-16} seconds:

Counter technique:	0.56,	0.6,	0.73,	0.9,	1.05	
Nuclear emulsions:	1.0 ,	1.6,	1.7,	1.9,	2.3,	2.8

The question is, do the different techniques provide consistent results at a significance level of 5%?

We want here to apply a two-sided Wilcoxon rank sum test for the hypothesis of equal population means on the basis of the two samples of sizes n = 5, m = 6. From Appendix Table A12 the lower critical limit for the W statistic with these numbers is $W_\ell = W_{.025} = 18$. The upper critical limit becomes $W_u = 2\bar{W} - W_\ell = 60 - 18 = 42$, which is also given in the table. Hence we shall reject the hypothesis of a common population mean at the 5% level if we find a rank sum for the smallest sample which is ≤ 18 or ≥ 42.

The observations above correspond to the following combined ordered sample,

<u>0.56</u> <u>0.6</u> <u>0.73</u> <u>0.9</u> 1.0 <u>1.05</u> 1.6 1.7 1.9 2.3 2.8

where the contributions from the counter experiments have been underlined. Thus the actual value of the rank sum becomes

$$W_{obs} = 1 + 2 + 3 + 4 + 6 = 16$$

which is below the critical limit for the test with the chosen significance level. Accordingly, from the Wilcoxon rank sum test the two sets of measurements of the π^o lifetime are not consistent at the 5% level.

The ordered sample of the two measurement series corresponds to 4 runs. It will be seen from Appendix Table A11 that the critical number of runs for n=5, m=6 and α=0.05 is $r_{.05} = 3$. Hence, from the run test we would have no reason to claim that the results of the two sets of measurements are not consistent at a significance 0.05. In fact, the probability to have 4 or less runs is 0.0644. This illustrates that the Wilcoxon test is more capable than the simple run test in rejecting a hypothesis.

14.6.11 Kruskal-Wallis rank test for comparison of several samples

When more than two data samples are to be checked for consistency, the two-sample tests described in the previous sections may be applied for a pairwise comparison of any two samples. With J samples this would imply a total of $\frac{1}{2}J(J-1)$ two-sample comparisons to be carried out. Also, one may compare each of the J samples with "the average" of all samples by the same two-sample procedures. Obviously, the number of such comparisons will soon become large, making the process a lengthy one if the number of samples is not relatively small. By random fluctuations, even if all samples do originate from the same parent population, one is liable to find at least two samples that appear to be inconsistent, and hence the risk of rejecting a true hypothesis may become considerable. Moreover, this approach will not give a measure of the overall agreement between all samples.

An efficient method for the simultaneous comparison of any number of samples is the *Kruskal-Wallis rank test*. Suppose that the complete set of N observations from J samples is arranged according to magnitude, such that each observations is assigned a rank between 1 and N. For each sample one finds the rank sum W_j as well as the mean rank $\bar{W}_j = W_j/n_j$, where n_j is the number of observations in the j-th sample. If the assumption of J identical parent populations

is correct, *i.e.* if H_o is true, all samples are expected to have the same mean rank

$$E(\bar{W}_j) = \frac{1}{N}\sum_{i=1}^{N} i = \frac{1}{2}(N+1) \qquad (14.107)$$

with the same unbiassed variance

$$V(\bar{W}_j) = \frac{1}{N-1}\sum_{i=1}^{N}\left(i-E(\bar{W}_j)\right)^2 = \frac{1}{12}N(N+1)\ . \qquad (14.109)$$

A suggestive test statistic with weighted contributions from the different samples is

$$H = \sum_{j=1}^{J} n_j\left(\bar{W}_j - \frac{1}{2}(N+1)\right)^2 \Big/ \left(\frac{1}{12}N(N+1)\right), \qquad (14.109)$$

which can be rewritten in the following form, convenient for computation,

$$H = \frac{12}{N(N+1)}\sum_{j=1}^{J}\frac{W_j^2}{n_j} - 3(N+1)\ . \qquad (14.110)$$

Since H will be zero if the $\bar{W}_j$ come out equal, and large of the $\bar{W}_j$ are substantially different, the hypothesis of a common parent population should be rejected if the observed value H_{obs} exceeds the critical value H_α corresponding to the chosen significance α. In order to determine these critical limits one must know the probability dristribution for H assuming the null hypothesis to be true. In principle, these (discrete) probabilities can be obtained from purely combinatorial arguments (assuming no "tied ranks") for any set of samples $n_1, n_2, \ldots, n_J$, starting with small numbers. Unfortunately, unless the numbers are tiny the amount of required work soon becomes formidable, and tabulating according to many arguments also becomes impracticable. Accordingly, Kruskal and Wallis give tables of critical values corresponding to significances between 1 and 10% for 3 samples of size not exceeding 5.

In practice one makes use of the fact that for H_o true, the statistic H for sufficiently large n_j has a chi-square distribution with J-1 degrees of freedom. The χ^2-approximation is generally accepted when *either* J = 3 and all sample sizes are above 5, *or* $J \geq 4$ and all sample sizes above 4.

Exercise 14.15: Justify the statement that H of eq.(14.109) is asymptotically distributed as $\chi^2(J-1)$ if H_o is true.

Exercise 14.16: Show that eq.(14.110) follows from eq.(14.109).

Exercise 14.17: Show that a test with the statistic $N/(N-1)\cdot H$ for J = 2 is equivalent to the Wilcoxon two-sample test.

14.6.12 The χ^2 test for comparison of histograms

In the preceeding discussion of rank test procedures for the comparison of experimental samples it has been tacitly assumed that the underlying distributions are continuous, in order that the rank assignments be meaningful. The foregoing procedures are therefore not applicable for testing the consistency of samples which are known to originate from discrete populations, neither will they apply for the comparison of samples consisting of "pooled" observations, where individual measurements have been grouped in categories prior to comparison.

To be specific, let us think of the common situation when a set of histograms is to be checked for consistency. The observations have been classified in I bins, equally chosen for all J histograms, and correspond to J independent multinomial distributions, the j-th of which having the following set if bin probabilities,

$$p_{1j}, p_{2j}, \ldots, p_{Ij}\ ; \qquad \sum_{i=1}^{I} p_{ij} = 1\ . \qquad (14.111)$$

The observed number of events in the different bins and the total number of events in the j-th histogram are given as

$$n_{1j}, n_{2j}, \ldots, n_{Ij}\ ; \qquad \sum_{i=1}^{I} n_{ij} = n_{\cdot j}\ . \qquad (14.112)$$

For all histograms the overall number if observations is n,

$$\sum_{j=1}^{J} n_{\cdot j} = \sum_{j=1}^{J}\sum_{i=1}^{I} n_{ij} = n\ . \qquad (14.113)$$

The hypothesis we wish to test is that all parent distributions are identical, corresponding to, for each bin number i, a common probability for all J histograms,

$$H_o:\ p_{i1} = p_{i2} = \cdots = p_{iJ} \qquad i = 1,2,\ldots,I \qquad (14.114)$$

Let us denote the common, unknown, probabilities by $p_{i\cdot}$, $i = 1,2,\cdots,I$. The Maximum-Likelihood estimates for these bin probabilities are the average frequencies observed for each of the bins

$$\hat{p}_{i\cdot} = \sum_{j=1}^{J} n_{ij}/n , \qquad i = 1,2,\cdots,I, \tag{14.115}$$

which are seen to satisfy the requirement $\sum_{i=1}^{I} \hat{p}_{i\cdot} = 1$ in virtue of the constraint on the n_{ij}, eq.(14.113); thus only I-1 of the estimated bin probabilities are independent. The test statistic for the comparison of all histograms simultaneously is constructed as the sum over all histograms and bins of altogether J·I terms, each term being a squared deviation between an observed and estimated number, divided by the estimated number,

$$X^2 = \sum_{j=1}^{J} \sum_{i=1}^{I} \frac{(n_{ij} - \hat{p}_{i\cdot} n_{\cdot j})^2}{\hat{p}_{i\cdot} n_{\cdot j}} . \tag{14.116}$$

If H_o is true and the expected event numbers in all histogram bins fulfil the usual normality requirement, this statistic will be approximately chi-square distributed. The number of degrees of freedom is (I-1)(J-1), corresponding to the present number of independent observations (IJ-J) minus the number of independently estimated parameters, I-1 .

The assumption of consistency between all J histograms is accepted at the chosen significance level 100α % if the calculated value X^2_{obs} comes out smaller than the critical value χ^2_α for the appropriate number of degrees of freedom, and rejected if the opposite occurs. Quite frequently when $X^2_{obs} > \chi^2_\alpha$ the overall inconsistency can be traced to a single histogram, say the j-th, having an exceptionally large contribution to X^2_{obs}. A repeated calculation with the j-th histogram excluded may then show the remaining histograms to be mutually compatible and suggest a critical examination of the data for the odd histogram.

Exercise 14.18: Show that the problem of testing consistency between histograms is equivalent to testing independence in a two-way classification.

Exercise 14.19: Four laboratories participating in a collaboration experiment have scanned their bubble chamber films for five different event topologies and have obtained the following number of events:

Laboratory	Event topology 1	2	3	4	5
A	202	150	107	51	18
B	152	131	70	48	17
C	189	161	108	42	25
D	105	78	52	32	12

Check that the event samples obtained by the different laboratories are fully compatible.

APPENDIX
Statistical Tables

Table A1. The binomial distribution

The table gives values of $B(r;n,p) = \binom{n}{r} p^r (1-p)^{n-r}$ for specified values of n,p and r, where $0 \leq r \leq n$.

The table only has entries for $p \leq 0.50$, but it can be used to find values of $B(r;n,p)$ for $p > 0.50$ by means of the relation

$$B(r;n,p) = B(n-r;n,1-p).$$

n	r	p .01	.02	.03	.05	.10	.15	.20	.25	.30	.40	.50
1	0	.9900	.9800	.9700	.9500	.9000	.8500	.8000	.7500	.7000	.6000	.5000
	1	.0100	.0200	.0300	.0500	.1000	.1500	.2000	.2500	.3000	.4000	.5000
2	0	.9801	.9604	.9409	.9025	.8100	.7225	.6400	.5625	.4900	.3600	.2500
	1	.0198	.0392	.0582	.0950	.1800	.2550	.3200	.3750	.4200	.4800	.5000
	2	.0001	.0004	.0009	.0025	.0100	.0225	.0400	.0625	.0900	.1600	.2500
3	0	.9703	.9412	.9127	.8574	.7290	.6141	.5120	.4219	.3430	.2160	.1250
	1	.0294	.0576	.0847	.1354	.2430	.3251	.3840	.4219	.4410	.4320	.3750
	2	.0003	.0012	.0026	.0071	.0270	.0574	.0960	.1406	.1890	.2880	.3750
	3	.0000	.0000	.0000	.0001	.0010	.0034	.0080	.0156	.0270	.0640	.1250
4	0	.9606	.9224	.8853	.8145	.6561	.5220	.4096	.3164	.2401	.1296	.0625
	1	.0388	.0753	.1095	.1715	.2916	.3685	.4096	.4219	.4116	.3456	.2500
	2	.0006	.0023	.0051	.0135	.0486	.0975	.1536	.2109	.2646	.3456	.3750
	3	.0000	.0000	.0001	.0005	.0036	.0115	.0256	.0469	.0756	.1536	.2500
	4	.0000	.0000	.0000	.0000	.0001	.0005	.0016	.0039	.0081	.0256	.0625
5	0	.9510	.9039	.8587	.7738	.5905	.4437	.3277	.2373	.1681	.0778	.0313
	1	.0480	.0922	.1328	.2036	.3281	.3915	.4096	.3955	.3601	.2592	.1563
	2	.0010	.0038	.0082	.0214	.0729	.1382	.2048	.2637	.3087	.3456	.3125
	3	.0000	.0001	.0003	.0011	.0081	.0244	.0512	.0879	.1323	.2304	.3125
	4	.0000	.0000	.0000	.0000	.0005	.0022	.0064	.0146	.0284	.0768	.1563
	5	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0010	.0024	.0102	.0313
6	0	.9415	.8858	.8330	.7351	.5314	.3771	.2621	.1780	.1176	.0467	.0156
	1	.0571	.1085	.1546	.2321	.3543	.3993	.3932	.3560	.3025	.1866	.0937
	2	.0014	.0055	.0120	.0305	.0984	.1762	.2458	.2966	.3241	.3110	.2344
	3	.0000	.0002	.0005	.0021	.0146	.0415	.0819	.1318	.1852	.2765	.3125
	4	.0000	.0000	.0000	.0001	.0012	.0055	.0154	.0330	.0595	.1382	.2344
	5	.0000	.0000	.0000	.0000	.0001	.0004	.0015	.0044	.0102	.0369	.0937
	6	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0007	.0041	.0156
7	0	.9321	.8681	.8080	.6983	.4783	.3206	.2097	.1335	.0824	.0280	.0078
	1	.0659	.1240	.1749	.2573	.3720	.3960	.3670	.3115	.2471	.1306	.0547
	2	.0020	.0076	.0162	.0406	.1240	.2097	.2753	.3115	.3177	.2613	.1641
	3	.0000	.0003	.0008	.0036	.0230	.0617	.1147	.1730	.2269	.2903	.2734
	4	.0000	.0000	.0000	.0002	.0026	.0109	.0287	.0577	.0972	.1935	.2734
	5	.0000	.0000	.0000	.0000	.0002	.0012	.0043	.0115	.0250	.0774	.1641
	6	.0000	.0000	.0000	.0000	.0000	.0001	.0004	.0013	.0036	.0172	.0547
	7	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0016	.0078
8	0	.9227	.8508	.7837	.6634	.4305	.2725	.1678	.1001	.0576	.0168	.0039
	1	.0746	.1389	.1939	.2793	.3826	.3847	.3355	.2670	.1977	.0896	.0313
	2	.0026	.0099	.0210	.0515	.1488	.2376	.2936	.3115	.2965	.2090	.1094
	3	.0001	.0004	.0013	.0054	.0331	.0839	.1468	.2076	.2541	.2787	.2188
	4	.0000	.0000	.0001	.0004	.0046	.0185	.0459	.0865	.1361	.2322	.2734
	5	.0000	.0000	.0000	.0000	.0004	.0026	.0092	.0231	.0467	.1239	.2188
	6	.0000	.0000	.0000	.0000	.0000	.0002	.0011	.0038	.0100	.0413	.1094
	7	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0004	.0012	.0079	.0313
	8	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0007	.0039
9	0	.9135	.8337	.7602	.6302	.3874	.2316	.1342	.0751	.0404	.0101	.0020
	1	.0830	.1531	.2116	.2985	.3874	.3679	.3020	.2253	.1556	.0605	.0176
	2	.0034	.0125	.0262	.0629	.1722	.2597	.3020	.3003	.2668	.1612	.0703
	3	.0001	.0006	.0019	.0077	.0446	.1069	.1762	.2336	.2668	.2508	.1641
	4	.0000	.0000	.0001	.0006	.0074	.0283	.0661	.1168	.1715	.2508	.2461
	5	.0000	.0000	.0000	.0000	.0008	.0050	.0165	.0389	.0735	.1672	.2461
	6	.0000	.0000	.0000	.0000	.0001	.0006	.0028	.0087	.0210	.0743	.1641
	7	.0000	.0000	.0000	.0000	.0000	.0000	.0003	.0012	.0039	.0212	.0703
	8	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0004	.0035	.0176
	9	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0003	.0020

Table A1. The binomial distribution (continued)

n	r	p = .01	.02	.03	.05	.10	.15	.20	.25	.30	.40	.50
10	0	.9044	.8171	.7374	.5987	.3487	.1969	.1074	.0563	.0282	.0060	.0010
	1	.0914	.1667	.2281	.3151	.3874	.3474	.2684	.1877	.1211	.0403	.0098
	2	.0042	.0153	.0317	.0746	.1937	.2759	.3020	.2816	.2335	.1209	.0439
	3	.0001	.0008	.0026	.0105	.0574	.1298	.2013	.2503	.2668	.2150	.1172
	4	.0000	.0000	.0001	.0010	.0112	.0401	.0881	.1460	.2001	.2508	.2051
	5	.0000	.0000	.0000	.0001	.0015	.0085	.0264	.0584	.1029	.2007	.2461
	6	.0000	.0000	.0000	.0000	.0001	.0012	.0055	.0162	.0368	.1115	.2051
	7	.0000	.0000	.0000	.0000	.0000	.0001	.0008	.0031	.0090	.0425	.1172
	8	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0004	.0014	.0106	.0439
	9	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0016	.0098
	10	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0010
11	0	.8953	.8007	.7153	.5688	.3138	.1673	.0859	.0422	.0198	.0036	.0005
	1	.0995	.1798	.2433	.3293	.3835	.3248	.2362	.1549	.0932	.0266	.0054
	2	.0050	.0183	.0376	.0867	.2131	.2866	.2953	.2581	.1998	.0887	.0269
	3	.0002	.0011	.0035	.0137	.0710	.1517	.2215	.2581	.2568	.1774	.0806
	4	.0000	.0000	.0002	.0014	.0158	.0536	.1107	.1721	.2201	.2365	.1611
	5	.0000	.0000	.0000	.0001	.0025	.0132	.0388	.0803	.1321	.2207	.2256
	6	.0000	.0000	.0000	.0000	.0003	.0023	.0097	.0268	.0566	.1471	.2256
	7	.0000	.0000	.0000	.0000	.0000	.0003	.0017	.0064	.0173	.0701	.1611
	8	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0011	.0037	.0234	.0806
	9	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0052	.0269
	10	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0007	.0054
	11	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0005
12	0	.8864	.7847	.6938	.5404	.2824	.1422	.0687	.0317	.0138	.0022	.0002
	1	.1074	.1922	.2575	.3413	.3766	.3012	.2062	.1267	.0712	.0174	.0029
	2	.0060	.0216	.0438	.0988	.2301	.2924	.2835	.2323	.1678	.0639	.0161
	3	.0002	.0015	.0045	.0173	.0852	.1720	.2362	.2581	.2397	.1419	.0537
	4	.0000	.0001	.0003	.0021	.0213	.0683	.1329	.1936	.2311	.2128	.1208
	5	.0000	.0000	.0000	.0002	.0038	.0193	.0532	.1032	.1585	.2270	.1934
	6	.0000	.0000	.0000	.0000	.0005	.0040	.0155	.0401	.0792	.1766	.2256
	7	.0000	.0000	.0000	.0000	.0000	.0006	.0033	.0115	.0291	.1009	.1934
	8	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0024	.0078	.0420	.1208
	9	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0004	.0015	.0125	.0537
	10	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0025	.0161
	11	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0003	.0029
	12	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002
13	0	.8775	.7690	.6730	.5133	.2542	.1209	.0550	.0238	.0097	.0013	.0001
	1	.1152	.2040	.2706	.3512	.3672	.2774	.1787	.1029	.0540	.0113	.0016
	2	.0070	.0250	.0502	.1109	.2448	.2937	.2680	.2059	.1388	.0453	.0095
	3	.0003	.0019	.0057	.0214	.0997	.1900	.2457	.2517	.2181	.1107	.0349
	4	.0000	.0001	.0004	.0028	.0277	.0838	.1535	.2097	.2337	.1845	.0873
	5	.0000	.0000	.0000	.0003	.0055	.0266	.0691	.1258	.1803	.2214	.1571
	6	.0000	.0000	.0000	.0000	.0008	.0063	.0230	.0559	.1030	.1968	.2095
	7	.0000	.0000	.0000	.0000	.0001	.0011	.0058	.0186	.0442	.1312	.2095
	8	.0000	.0000	.0000	.0000	.0000	.0001	.0011	.0047	.0142	.0656	.1571
	9	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0009	.0034	.0243	.0873
	10	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0006	.0065	.0349
	11	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0012	.0095
	12	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0016
	13	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001
14	0	.8687	.7536	.6528	.4877	.2288	.1028	.0440	.0178	.0068	.0008	.0001
	1	.1229	.2153	.2827	.3593	.3559	.2539	.1539	.0832	.0407	.0073	.0009
	2	.0081	.0286	.0568	.1229	.2570	.2912	.2501	.1802	.1134	.0317	.0056
	3	.0003	.0023	.0070	.0259	.1142	.2056	.2501	.2402	.1943	.0845	.0222
	4	.0000	.0001	.0006	.0037	.0349	.0998	.1720	.2202	.2290	.1549	.0611
	5	.0000	.0000	.0000	.0004	.0078	.0352	.0860	.1468	.1963	.2066	.1222
	6	.0000	.0000	.0000	.0000	.0013	.0093	.0322	.0734	.1262	.2066	.1833
	7	.0000	.0000	.0000	.0000	.0002	.0019	.0092	.0280	.0618	.1574	.2095
	8	.0000	.0000	.0000	.0000	.0000	.0003	.0026	.0082	.0232	.0919	.1833
	9	.0000	.0000	.0000	.0000	.0000	.0000	.0005	.0018	.0066	.0408	.1222
	10	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0003	.0014	.0136	.0611
	11	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0033	.0222
	12	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0005	.0056
	13	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0009
	14	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001
15	0	.8601	.7386	.6333	.4633	.2059	.0874	.0352	.0134	.0047	.0005	.0000
	1	.1303	.2261	.2938	.3658	.3432	.2312	.1319	.0668	.0305	.0047	.0005
	2	.0092	.0323	.0636	.1348	.2669	.2856	.2309	.1559	.0916	.0219	.0032
	3	.0004	.0029	.0085	.0307	.1285	.2184	.2501	.2252	.1700	.0634	.0139
	4	.0000	.0002	.0008	.0049	.0428	.1156	.1876	.2252	.2186	.1268	.0417
	5	.0000	.0000	.0001	.0006	.0105	.0449	.1032	.1651	.2061	.1859	.0916
	6	.0000	.0000	.0000	.0000	.0019	.0132	.0430	.0917	.1472	.2066	.1527
	7	.0000	.0000	.0000	.0000	.0003	.0030	.0138	.0393	.0811	.1771	.1964
	8	.0000	.0000	.0000	.0000	.0000	.0005	.0035	.0131	.0348	.1181	.1964
	9	.0000	.0000	.0000	.0000	.0000	.0001	.0007	.0034	.0116	.0612	.1527
	10	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0007	.0030	.0245	.0916
	11	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0006	.0074	.0417
	12	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0016	.0139
	13	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0003	.0032
	14	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0005
	15	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000

Table A1. The binomial distribution (continued)

n	r	p = .01	.02	.03	.05	.10	.15	.20	.25	.30	.40	.50
16	0	.8515	.7238	.6143	.4401	.1853	.0743	.0281	.0100	.0033	.0003	.0000
	1	.1376	.2363	.3040	.3706	.3294	.2097	.1126	.0535	.0228	.0030	.0002
	2	.0104	.0362	.0705	.1463	.2745	.2775	.2111	.1336	.0732	.0150	.0018
	3	.0005	.0034	.0102	.0359	.1423	.2285	.2463	.2079	.1465	.0468	.0085
	4	.0000	.0002	.0010	.0061	.0514	.1311	.2001	.2252	.2040	.1014	.0278
	5	.0000	.0000	.0001	.0008	.0137	.0555	.1201	.1802	.2099	.1623	.0667
	6	.0000	.0000	.0000	.0001	.0028	.0180	.0550	.1101	.1649	.1983	.1222
	7	.0000	.0000	.0000	.0000	.0004	.0045	.0197	.0524	.1010	.1889	.1746
	8	.0000	.0000	.0000	.0000	.0001	.0009	.0055	.0197	.0487	.1417	.1964
	9	.0000	.0000	.0000	.0000	.0000	.0001	.0012	.0058	.0185	.0840	.1746
	10	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0014	.0056	.0392	.1222
	11	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0013	.0142	.0667
	12	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0040	.0278
	13	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0008	.0085
	14	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0018
	15	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002
	16	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
17	0	.8429	.7093	.5958	.4181	.1668	.0631	.0225	.0075	.0023	.0002	.0000
	1	.1447	.2461	.3133	.3741	.3150	.1893	.0957	.0426	.0169	.0019	.0001
	2	.0117	.0402	.0775	.1575	.2800	.2673	.1914	.1136	.0581	.0102	.0010
	3	.0006	.0041	.0120	.0415	.1556	.2359	.2393	.1893	.1245	.0341	.0052
	4	.0000	.0003	.0013	.0076	.0605	.1457	.2093	.2209	.1868	.0796	.0182
	5	.0000	.0000	.0001	.0010	.0175	.0668	.1361	.1914	.2081	.1379	.0472
	6	.0000	.0000	.0000	.0001	.0039	.0236	.0680	.1276	.1784	.1839	.0944
	7	.0000	.0000	.0000	.0000	.0007	.0065	.0267	.0668	.1201	.1927	.1484
	8	.0000	.0000	.0000	.0000	.0001	.0014	.0084	.0279	.0644	.1606	.1855
	9	.0000	.0000	.0000	.0000	.0000	.0003	.0021	.0093	.0276	.1070	.1855
	10	.0000	.0000	.0000	.0000	.0000	.0000	.0004	.0025	.0095	.0571	.1484
	11	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0026	.0242	.0944
	12	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0006	.0081	.0472
	13	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0021	.0182
	14	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0004	.0052
	15	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0010
	16	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001
	17	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
18	0	.8345	.6951	.5780	.3972	.1501	.0536	.0180	.0056	.0016	.0001	.0000
	1	.1517	.2554	.3217	.3763	.3002	.1704	.0811	.0338	.0126	.0012	.0001
	2	.0130	.0443	.0846	.1683	.2835	.2556	.1723	.0958	.0458	.0069	.0006
	3	.0007	.0048	.0148	.0473	.1680	.2406	.2297	.1704	.1046	.0246	.0031
	4	.0000	.0004	.0016	.0093	.0700	.1592	.2153	.2130	.1681	.0614	.0117
	5	.0000	.0000	.0001	.0014	.0218	.0787	.1507	.1988	.2017	.1146	.0327
	6	.0000	.0000	.0000	.0002	.0052	.0301	.0816	.1436	.1873	.1655	.0708
	7	.0000	.0000	.0000	.0000	.0010	.0091	.0350	.0820	.1376	.1892	.1214
	8	.0000	.0000	.0000	.0000	.0002	.0022	.0120	.0376	.0811	.1734	.1669
	9	.0000	.0000	.0000	.0000	.0000	.0004	.0033	.0139	.0386	.1284	.1855
	10	.0000	.0000	.0000	.0000	.0000	.0001	.0008	.0042	.0149	.0771	.1669
	11	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0010	.0046	.0374	.1214
	12	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0012	.0145	.0708
	13	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0045	.0327
	14	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0011	.0117
	15	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0031
	16	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0006
	17	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001
	18	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
19	0	.8262	.6812	.5606	.3774	.1351	.0456	.0144	.0042	.0011	.0001	.0000
	1	.1586	.2642	.3294	.3774	.2852	.1529	.0685	.0268	.0093	.0008	.0000
	2	.0144	.0485	.0917	.1787	.2852	.2428	.1540	.0803	.0358	.0046	.0003
	3	.0008	.0056	.0161	.0533	.1796	.2428	.2182	.1517	.0869	.0175	.0018
	4	.0000	.0005	.0020	.0112	.0798	.1714	.2182	.2023	.1491	.0467	.0074
	5	.0000	.0000	.0002	.0018	.0266	.0907	.1636	.2023	.1916	.0933	.0222
	6	.0000	.0000	.0000	.0002	.0069	.0374	.0955	.1574	.1916	.1451	.0518
	7	.0000	.0000	.0000	.0000	.0014	.0122	.0443	.0974	.1525	.1797	.0961
	8	.0000	.0000	.0000	.0000	.0002	.0032	.0166	.0487	.0981	.1797	.1442
	9	.0000	.0000	.0000	.0000	.0000	.0007	.0051	.0198	.0514	.1464	.1762
	10	.0000	.0000	.0000	.0000	.0000	.0001	.0013	.0066	.0220	.0976	.1762
	11	.0000	.0000	.0000	.0000	.0000	.0000	.0003	.0018	.0077	.0532	.1442
	12	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0004	.0022	.0237	.0961
	13	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0085	.0518
	14	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0024	.0222
	15	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0005	.0074
	16	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0018
	17	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0003
	18	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	19	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000

Table A1. The binomial distribution (continued)

n	r	p .01	.02	.03	.05	.10	.15	.20	.25	.30	.40	.50
20	0	.8179	.6676	.5438	.3585	.1216	.0388	.0115	.0032	.0008	.0000	.0000
	1	.1652	.2725	.3364	.3774	.2702	.1368	.0576	.0211	.0068	.0005	.0000
	2	.0159	.0528	.0988	.1887	.2852	.2293	.1369	.0669	.0278	.0031	.0002
	3	.0010	.0065	.0183	.0596	.1901	.2428	.2054	.1339	.0716	.0123	.0011
	4	.0000	.0006	.0024	.0133	.0898	.1821	.2182	.1897	.1304	.0350	.0046
	5	.0000	.0000	.0002	.0022	.0319	.1028	.1746	.2023	.1789	.0746	.0148
	6	.0000	.0000	.0000	.0003	.0089	.0454	.1091	.1686	.1916	.1244	.0370
	7	.0000	.0000	.0000	.0000	.0020	.0160	.0545	.1124	.1643	.1659	.0739
	8	.0000	.0000	.0000	.0000	.0004	.0046	.0222	.0609	.1144	.1797	.1201
	9	.0000	.0000	.0000	.0000	.0001	.0011	.0074	.0271	.0654	.1597	.1602
	10	.0000	.0000	.0000	.0000	.0000	.0002	.0020	.0099	.0308	.1171	.1762
	11	.0000	.0000	.0000	.0000	.0000	.0000	.0005	.0030	.0120	.0710	.1602
	12	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0008	.0039	.0355	.1201
	13	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0010	.0146	.0739
	14	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0049	.0370
	15	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0013	.0148
	16	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0003	.0046
	17	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0011
	18	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002
	19	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	20	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
25	0	.7778	.6035	.4670	.2774	.0718	.0172	.0038	.0008	.0001	.0000	.0000
	1	.1964	.3079	.3611	.3650	.1994	.0759	.0236	.0063	.0014	.0000	.0000
	2	.0238	.0754	.1340	.2305	.2659	.1607	.0708	.0251	.0074	.0004	.0000
	3	.0018	.0118	.0318	.0930	.2265	.2174	.1358	.0641	.0243	.0019	.0001
	4	.0001	.0013	.0054	.0269	.1384	.2110	.1867	.1175	.0572	.0071	.0004
	5	.0000	.0001	.0007	.0060	.0646	.1564	.1960	.1645	.1030	.0199	.0016
	6	.0000	.0000	.0001	.0010	.0239	.0920	.1633	.1828	.1472	.0442	.0053
	7	.0000	.0000	.0000	.0001	.0072	.0441	.1108	.1654	.1712	.0800	.0143
	8	.0000	.0000	.0000	.0000	.0018	.0175	.0623	.1241	.1651	.1200	.0322
	9	.0000	.0000	.0000	.0000	.0004	.0058	.0294	.0781	.1336	.1511	.0609
	10	.0000	.0000	.0000	.0000	.0001	.0016	.0118	.0417	.0916	.1612	.0974
	11	.0000	.0000	.0000	.0000	.0000	.0004	.0040	.0189	.0536	.1465	.1328
	12	.0000	.0000	.0000	.0000	.0000	.0001	.0012	.0074	.0268	.1140	.1550
	13	.0000	.0000	.0000	.0000	.0000	.0000	.0003	.0025	.0115	.0760	.1550
	14	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0007	.0042	.0434	.1328
	15	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0013	.0212	.0974
	16	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0004	.0088	.0609
	17	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0031	.0322
	18	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0009	.0143
	19	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0053
	20	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0016
	21	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0004
	22	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001
	23	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	24	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	25	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
30	0	.7397	.5455	.4010	.2146	.0424	.0076	.0012	.0002	.0000	.0000	.0000
	1	.2242	.3340	.3721	.3389	.1413	.0404	.0093	.0018	.0003	.0000	.0000
	2	.0328	.0988	.1669	.2586	.2277	.1034	.0337	.0086	.0018	.0000	.0000
	3	.0031	.0188	.0482	.1270	.2361	.1703	.0785	.0269	.0072	.0003	.0000
	4	.0002	.0026	.0101	.0451	.1771	.2028	.1325	.0604	.0208	.0012	.0000
	5	.0000	.0003	.0016	.0124	.1023	.1861	.1723	.1047	.0464	.0041	.0001
	6	.0000	.0000	.0002	.0027	.0474	.1368	.1795	.1455	.0829	.0115	.0006
	7	.0000	.0000	.0000	.0005	.0180	.0828	.1538	.1662	.1219	.0263	.0019
	8	.0000	.0000	.0000	.0001	.0058	.0420	.1106	.1593	.1501	.0505	.0055
	9	.0000	.0000	.0000	.0000	.0016	.0181	.0676	.1298	.1573	.0823	.0133
	10	.0000	.0000	.0000	.0000	.0004	.0067	.0355	.0909	.1416	.1152	.0280
	11	.0000	.0000	.0000	.0000	.0001	.0022	.0161	.0551	.1103	.1396	.0509
	12	.0000	.0000	.0000	.0000	.0000	.0006	.0064	.0291	.0749	.1474	.0806
	13	.0000	.0000	.0000	.0000	.0000	.0001	.0022	.0134	.0444	.1360	.1115
	14	.0000	.0000	.0000	.0000	.0000	.0000	.0007	.0054	.0231	.1101	.1354
	15	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0019	.0106	.0783	.1445
	16	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0006	.0042	.0489	.1354
	17	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0015	.0264	.1115
	18	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0005	.0129	.0806
	19	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0054	.0509
	20	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0020	.0280
	21	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0006	.0133
	22	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0055
	23	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0019
	24	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0006
	25	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001
	26	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	27	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	28	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	29	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	30	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000

Table A2. The cumulative binomial distribution

The table gives values of $F(x;n,p) = \sum_{r=0}^{x} \binom{n}{r} p^r (1-p)^{n-r}$ for specified values of n,p and x, where $0 \leq x \leq n$.

The table only has entries for $p \leq 0.50$, but it can be used to find values of $F(x;n,p)$ for $p > 0.50$ by means of the relation

$$F(x;n,p) = 1 - F(n-x-1;n,1-p), \qquad (0 \leq x \leq n-1).$$

n	x	p .01	.02	.03	.05	.10	.15	.20	.25	.30	.40	.50
1	0	.9900	.9800	.9700	.9500	.9000	.8500	.8000	.7500	.7000	.6000	.5000
	1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
2	0	.9801	.9604	.9409	.9025	.8100	.7225	.6400	.5625	.4900	.3600	.2500
	1	.9999	.9996	.9991	.9975	.9900	.9775	.9600	.9375	.9100	.8400	.7500
	2	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
3	0	.9703	.9412	.9127	.8574	.7290	.6141	.5120	.4219	.3430	.2160	.1250
	1	.9997	.9988	.9974	.9927	.9720	.9392	.8960	.8438	.7840	.6480	.5000
	2	1.0000	1.0000	1.0000	.9999	.9990	.9966	.9920	.9844	.9730	.9360	.8750
	3	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
4	0	.9606	.9224	.8853	.8145	.6561	.5220	.4096	.3164	.2401	.1296	.0625
	1	.9994	.9977	.9948	.9860	.9477	.8905	.8192	.7383	.6517	.4752	.3125
	2	1.0000	1.0000	.9999	.9995	.9963	.9880	.9728	.9492	.9163	.8208	.6875
	3	1.0000	1.0000	1.0000	1.0000	.9999	.9995	.9984	.9961	.9919	.9744	.9375
	4	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
5	0	.9510	.9039	.8587	.7738	.5905	.4437	.3277	.2373	.1681	.0778	.0313
	1	.9990	.9962	.9915	.9774	.9185	.8352	.7373	.6328	.5282	.3370	.1875
	2	1.0000	.9999	.9997	.9988	.9914	.9734	.9421	.8965	.8369	.6826	.5000
	3	1.0000	1.0000	1.0000	1.0000	.9995	.9978	.9933	.9844	.9692	.9130	.8125
	4	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9997	.9990	.9976	.9898	.9687
	5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
6	0	.9415	.8858	.8330	.7351	.5314	.3771	.2621	.1780	.1176	.0467	.0156
	1	.9985	.9943	.9875	.9672	.8857	.7765	.6554	.5339	.4202	.2333	.1094
	2	1.0000	.9998	.9995	.9978	.9841	.9527	.9011	.8306	.7443	.5443	.3438
	3	1.0000	1.0000	1.0000	.9999	.9987	.9941	.9830	.9624	.9295	.8208	.6562
	4	1.0000	1.0000	1.0000	1.0000	.9999	.9996	.9984	.9954	.9891	.9590	.8906
	5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9998	.9993	.9959	.9844
	6	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
7	0	.9321	.8681	.8080	.6983	.4783	.3206	.2097	.1335	.0824	.0280	.0078
	1	.9980	.9921	.9829	.9556	.8503	.7166	.5767	.4449	.3294	.1586	.0625
	2	1.0000	.9997	.9991	.9962	.9743	.9262	.8520	.7564	.6471	.4199	.2266
	3	1.0000	1.0000	1.0000	.9998	.9973	.9879	.9667	.9294	.8740	.7102	.5000
	4	1.0000	1.0000	1.0000	1.0000	.9998	.9988	.9953	.9871	.9712	.9037	.7734
	5	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9996	.9987	.9962	.9812	.9375
	6	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9998	.9984	.9922
	7	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
8	0	.9227	.8508	.7837	.6634	.4305	.2725	.1678	.1001	.0576	.0168	.0039
	1	.9973	.9897	.9777	.9428	.8131	.6572	.5033	.3671	.2553	.1064	.0352
	2	.9999	.9996	.9987	.9942	.9619	.8948	.7969	.6785	.5518	.3154	.1445
	3	1.0000	1.0000	.9999	.9996	.9950	.9786	.9437	.8862	.8059	.5941	.3633
	4	1.0000	1.0000	1.0000	1.0000	.9996	.9971	.9896	.9727	.9420	.8263	.6367
	5	1.0000	1.0000	1.0000	1.0000	1.0000	.9998	.9988	.9958	.9887	.9502	.8555
	6	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9996	.9987	.9915	.9648
	7	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9993	.9961
	8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
9	0	.9135	.8337	.7602	.6302	.3874	.2316	.1342	.0751	.0404	.0101	.0020
	1	.9966	.9869	.9718	.9288	.7748	.5995	.4362	.3003	.1960	.0705	.0195
	2	.9999	.9994	.9980	.9916	.9470	.8591	.7382	.6007	.4628	.2318	.0898
	3	1.0000	1.0000	.9999	.9994	.9917	.9661	.9144	.8343	.7297	.4826	.2539
	4	1.0000	1.0000	1.0000	1.0000	.9991	.9944	.9804	.9511	.9012	.7334	.5000
	5	1.0000	1.0000	1.0000	1.0000	.9999	.9994	.9969	.9900	.9747	.9006	.7461
	6	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9997	.9987	.9957	.9750	.9102
	7	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9996	.9962	.9805
	8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9997	.9980
	9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table A2. The cumulative binomial distribution (continued)

n	x	p .01	.02	.03	.05	.10	.15	.20	.25	.30	.40	.50
10	0	.9044	.8171	.7374	.5987	.3487	.1969	.1074	.0563	.0282	.0060	.0010
	1	.9957	.9838	.9655	.9139	.7361	.5443	.3758	.2440	.1493	.0464	.0107
	2	.9999	.9991	.9972	.9885	.9298	.8202	.6778	.5256	.3828	.1673	.0547
	3	1.0000	1.0000	.9999	.9990	.9872	.9500	.8791	.7759	.6496	.3823	.1719
	4	1.0000	1.0000	1.0000	.9999	.9984	.9901	.9672	.9219	.8497	.6331	.3770
	5	1.0000	1.0000	1.0000	1.0000	.9999	.9986	.9936	.9803	.9527	.8338	.6230
	6	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9991	.9965	.9894	.9452	.8281
	7	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9996	.9984	.9877	.9453
	8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9983	.9893
	9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9990
	10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
11	0	.8953	.8007	.7153	.5688	.3138	.1673	.0859	.0422	.0198	.0036	.0005
	1	.9948	.9805	.9587	.8981	.6974	.4922	.3221	.1971	.1130	.0302	.0059
	2	.9998	.9988	.9963	.9848	.9104	.7788	.6174	.4552	.3127	.1189	.0327
	3	1.0000	1.0000	.9998	.9984	.9815	.9306	.8389	.7133	.5696	.2963	.1133
	4	1.0000	1.0000	1.0000	.9999	.9972	.9841	.9496	.8854	.7897	.5328	.2744
	5	1.0000	1.0000	1.0000	1.0000	.9997	.9973	.9883	.9657	.9218	.7535	.5000
	6	1.0000	1.0000	1.0000	1.0000	1.0000	.9997	.9980	.9924	.9784	.9006	.7256
	7	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9998	.9988	.9957	.9707	.8867
	8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9994	.9941	.9673
	9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9993	.9941
	10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9995
	11	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
12	0	.8864	.7847	.6938	.5404	.2824	.1422	.0687	.0317	.0138	.0022	.0002
	1	.9938	.9769	.9514	.8816	.6590	.4435	.2749	.1584	.0850	.0196	.0032
	2	.9998	.9985	.9952	.9804	.8891	.7358	.5583	.3907	.2528	.0834	.0193
	3	1.0000	.9999	.9997	.9978	.9744	.9078	.7946	.6488	.4925	.2253	.0730
	4	1.0000	1.0000	1.0000	.9998	.9957	.9761	.9274	.8424	.7237	.4382	.1938
	5	1.0000	1.0000	1.0000	1.0000	.9995	.9954	.9806	.9456	.8822	.6652	.3872
	6	1.0000	1.0000	1.0000	1.0000	.9999	.9993	.9961	.9857	.9614	.8418	.6128
	7	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9994	.9972	.9905	.9427	.8062
	8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9996	.9983	.9847	.9270
	9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9998	.9972	.9807
	10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9997	.9968
	11	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9998
	12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
13	0	.8775	.7690	.6730	.5133	.2542	.1209	.0550	.0238	.0097	.0013	.0001
	1	.9928	.9730	.9436	.8646	.6213	.3983	.2336	.1267	.0637	.0126	.0017
	2	.9997	.9980	.9938	.9755	.8661	.6920	.5017	.3326	.2025	.0579	.0112
	3	1.0000	.9999	.9995	.9969	.9658	.8820	.7473	.5843	.4206	.1686	.0461
	4	1.0000	1.0000	1.0000	.9997	.9935	.9658	.9009	.7940	.6543	.3530	.1334
	5	1.0000	1.0000	1.0000	1.0000	.9991	.9925	.9700	.9198	.8346	.5744	.2905
	6	1.0000	1.0000	1.0000	1.0000	.9999	.9987	.9930	.9757	.9376	.7712	.5000
	7	1.0000	1.0000	1.0000	1.0000	1.0000	.9998	.9988	.9944	.9818	.9023	.7095
	8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9998	.9990	.9960	.9679	.8666
	9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9993	.9922	.9539
	10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9987	.9887
	11	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9983
	12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
14	0	.8687	.7536	.6528	.4877	.2288	.1028	.0440	.0178	.0068	.0008	.0001
	1	.9916	.9690	.9355	.8470	.5846	.3567	.1979	.1010	.0475	.0081	.0009
	2	.9997	.9975	.9923	.9699	.8416	.6479	.4481	.2811	.1608	.0398	.0065
	3	1.0000	.9999	.9994	.9958	.9559	.8535	.6982	.5213	.3552	.1243	.0287
	4	1.0000	1.0000	1.0000	.9996	.9908	.9533	.8702	.7415	.5842	.2793	.0898
	5	1.0000	1.0000	1.0000	1.0000	.9985	.9885	.9561	.8883	.7805	.4859	.2120
	6	1.0000	1.0000	1.0000	1.0000	.9998	.9978	.9884	.9617	.9067	.6925	.3953
	7	1.0000	1.0000	1.0000	1.0000	1.0000	.9997	.9976	.9897	.9685	.8499	.6047
	8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9996	.9978	.9917	.9417	.7880
	9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9997	.9983	.9825	.9102
	10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9998	.9961	.9713
	11	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9994	.9935
	12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9991
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
15	0	.8601	.7386	.6333	.4633	.2059	.0874	.0352	.0134	.0047	.0005	.0000
	1	.9904	.9647	.9270	.8290	.5490	.3186	.1671	.0802	.0353	.0052	.0005
	2	.9996	.9970	.9906	.9638	.8159	.6042	.3980	.2361	.1268	.0271	.0037
	3	1.0000	.9998	.9992	.9945	.9444	.8227	.6482	.4613	.2969	.0905	.0176
	4	1.0000	1.0000	.9999	.9994	.9873	.9383	.8358	.6865	.5155	.2173	.0592
	5	1.0000	1.0000	1.0000	.9999	.9978	.9832	.9389	.8516	.7216	.4032	.1509
	6	1.0000	1.0000	1.0000	1.0000	.9997	.9964	.9819	.9434	.8689	.6098	.3036
	7	1.0000	1.0000	1.0000	1.0000	1.0000	.9994	.9958	.9827	.9500	.7869	.5000
	8	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9992	.9958	.9848	.9050	.6964
	9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9992	.9963	.9662	.8491
	10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9993	.9907	.9408
	11	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9981	.9824
	12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9997	.9963
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9995
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table A2. The cumulative binomial distribution (continued)

n	x	p .01	.02	.03	.05	.10	.15	.20	.25	.30	.40	.50
16	0	.8515	.7238	.6143	.4401	.1853	.0743	.0281	.0100	.0033	.0003	.0000
	1	.9891	.9601	.9182	.8108	.5147	.2839	.1407	.0635	.0261	.0033	.0003
	2	.9985	.9963	.9887	.9571	.7892	.5614	.3518	.1971	.0994	.0183	.0021
	3	1.0000	.9998	.9989	.9930	.9316	.7899	.5981	.4050	.2459	.0651	.0106
	4	1.0000	1.0000	.9999	.9991	.9830	.9209	.7982	.6302	.4499	.1666	.0384
	5	1.0000	1.0000	1.0000	.9999	.9967	.9765	.9183	.8103	.6598	.3288	.1051
	6	1.0000	1.0000	1.0000	1.0000	.9995	.9944	.9733	.9204	.8247	.5272	.2272
	7	1.0000	1.0000	1.0000	1.0000	.9999	.9989	.9930	.9729	.9256	.7161	.4018
	8	1.0000	1.0000	1.0000	1.0000	1.0000	.9998	.9985	.9925	.9743	.8577	.5982
	9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9998	.9984	.9929	.9417	.7728
	10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9997	.9984	.9809	.8949
	11	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9997	.9951	.9616
	12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9991	.9894
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9979
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9997
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
17	0	.8429	.7093	.5958	.4181	.1668	.0631	.0225	.0075	.0023	.0002	.0000
	1	.9877	.9554	.9091	.7922	.4818	.2525	.1182	.0501	.0193	.0021	.0001
	2	.9994	.9956	.9866	.9497	.7618	.5198	.3096	.1637	.0774	.0123	.0012
	3	1.0000	.9997	.9986	.9912	.9174	.7556	.5489	.3530	.2019	.0464	.0064
	4	1.0000	1.0000	.9999	.9988	.9779	.9013	.7582	.5739	.3887	.1260	.0245
	5	1.0000	1.0000	1.0000	.9999	.9953	.9681	.8943	.7653	.5968	.2639	.0717
	6	1.0000	1.0000	1.0000	1.0000	.9992	.9917	.9623	.8929	.7752	.4478	.1662
	7	1.0000	1.0000	1.0000	1.0000	.9999	.9983	.9891	.9598	.8954	.6405	.3145
	8	1.0000	1.0000	1.0000	1.0000	1.0000	.9997	.9974	.9876	.9597	.8011	.5000
	9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9995	.9969	.9873	.9081	.6855
	10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9994	.9968	.9652	.8338
	11	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9993	.9894	.9283
	12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9975	.9755
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9995	.9936
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9988
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
18	0	.8345	.6951	.5780	.3972	.1501	.0536	.0180	.0056	.0016	.0001	.0000
	1	.9862	.9505	.8997	.7735	.4503	.2241	.0991	.0395	.0142	.0013	.0001
	2	.9993	.9948	.9843	.9419	.7338	.4797	.2713	.1353	.0600	.0082	.0007
	3	1.0000	.9996	.9982	.9891	.9018	.7202	.5010	.3057	.1646	.0328	.0038
	4	1.0000	1.0000	.9998	.9985	.9718	.8794	.7164	.5187	.3327	.0942	.0154
	5	1.0000	1.0000	1.0000	.9998	.9936	.9581	.8671	.7175	.5344	.2088	.0481
	6	1.0000	1.0000	1.0000	1.0000	.9988	.9882	.9487	.8610	.7217	.3743	.1189
	7	1.0000	1.0000	1.0000	1.0000	.9998	.9973	.9837	.9431	.8593	.5634	.2403
	8	1.0000	1.0000	1.0000	1.0000	1.0000	.9995	.9957	.9807	.9404	.7368	.4073
	9	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9991	.9946	.9790	.8653	.5927
	10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9998	.9988	.9939	.9424	.7597
	11	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9998	.9986	.9797	.8811
	12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9997	.9942	.9519
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9987	.9846
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9998	.9962
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9993
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999
	17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	18	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
19	0	.8262	.6812	.5606	.3774	.1351	.0456	.0144	.0042	.0011	.0001	.0000
	1	.9847	.9454	.8900	.7547	.4203	.1985	.0829	.0310	.0104	.0008	.0000
	2	.9991	.9939	.9817	.9335	.7054	.4413	.2369	.1113	.0462	.0055	.0004
	3	1.0000	.9995	.9974	.9868	.8850	.6841	.4551	.2631	.1332	.0230	.0022
	4	1.0000	1.0000	.9994	.9980	.9648	.8556	.6733	.4854	.2822	.0696	.0096
	5	1.0000	1.0000	1.0000	.9998	.9914	.9463	.8369	.6678	.4739	.1629	.0318
	6	1.0000	1.0000	1.0000	1.0000	.9983	.9837	.9324	.8251	.6655	.3081	.0835
	7	1.0000	1.0000	1.0000	1.0000	.9997	.9959	.9767	.9225	.8180	.4878	.1796
	8	1.0000	1.0000	1.0000	1.0000	1.0000	.9992	.9933	.9713	.9161	.6675	.3238
	9	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9984	.9911	.9674	.8139	.5000
	10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9997	.9977	.9895	.9115	.6762
	11	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9995	.9972	.9648	.8204
	12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9994	.9884	.9165
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9969	.9682
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9994	.9904
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9978
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9996
	17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	18	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	19	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table A2. The cumulative binomial distribution (continued)

n	x	p .01	.02	.03	.05	.10	.15	.20	.25	.30	.40	.50
20	0	.8179	.6676	.5438	.3585	.1216	.0388	.0115	.0032	.0008	.0000	.0000
	1	.9831	.9401	.8802	.7358	.3917	.1756	.0692	.0243	.0076	.0005	.0000
	2	.9990	.9929	.9790	.9245	.6769	.4049	.2061	.0913	.0355	.0036	.0002
	3	1.0000	.9994	.9973	.9841	.8670	.6477	.4114	.2252	.1071	.0160	.0013
	4	1.0000	1.0000	.9997	.9974	.9568	.8298	.6296	.4148	.2375	.0510	.0059
	5	1.0000	1.0000	1.0000	.9997	.9887	.9327	.8042	.6172	.4164	.1256	.0207
	6	1.0000	1.0000	1.0000	1.0000	.9976	.9781	.9133	.7858	.6080	.2500	.0577
	7	1.0000	1.0000	1.0000	1.0000	.9996	.9941	.9679	.8982	.7723	.4159	.1316
	8	1.0000	1.0000	1.0000	1.0000	.9999	.9987	.9900	.9591	.8867	.5956	.2517
	9	1.0000	1.0000	1.0000	1.0000	1.0000	.9998	.9974	.9861	.9520	.7553	.4119
	10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9994	.9961	.9829	.8725	.5881
	11	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9991	.9949	.9435	.7483
	12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9998	.9987	.9790	.8684
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9997	.9935	.9423
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9984	.9793
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9997	.9941
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9987
	17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9998
	18	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	19	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	20	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
25	0	.7778	.6035	.4670	.2774	.0718	.0172	.0038	.0008	.0001	.0000	.0000
	1	.9742	.9114	.8280	.6424	.2712	.0931	.0274	.0070	.0016	.0001	.0000
	2	.9980	.9868	.9620	.8729	.5371	.2537	.0982	.0321	.0090	.0004	.0000
	3	.9999	.9986	.9938	.9659	.7636	.4711	.2340	.0962	.0332	.0024	.0001
	4	1.0000	.9999	.9992	.9928	.9020	.6821	.4207	.2137	.0905	.0095	.0005
	5	1.0000	1.0000	.9999	.9988	.9666	.8385	.6167	.3783	.1935	.0294	.0020
	6	1.0000	1.0000	1.0000	.9998	.9905	.9305	.7800	.5611	.3407	.0736	.0073
	7	1.0000	1.0000	1.0000	1.0000	.9977	.9745	.8909	.7265	.5118	.1536	.0216
	8	1.0000	1.0000	1.0000	1.0000	.9995	.9920	.9532	.8506	.6769	.2735	.0539
	9	1.0000	1.0000	1.0000	1.0000	.9999	.9979	.9827	.9287	.8106	.4246	.1148
	10	1.0000	1.0000	1.0000	1.0000	1.0000	.9995	.9944	.9703	.9022	.5858	.2122
	11	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9985	.9893	.9558	.7323	.3450
	12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9996	.9966	.9825	.8462	.5000
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9991	.9940	.9222	.6550
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9998	.9982	.9656	.7878
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9995	.9868	.8852
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9957	.9461
	17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9988	.9784
	18	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9997	.9927
	19	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9980
	20	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9995
	21	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999
	22	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	23	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	24	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	25	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
30	0	.7397	.5455	.4010	.2146	.0424	.0076	.0012	.0002	.0000	.0000	.0000
	1	.9639	.8795	.7731	.5535	.1837	.0480	.0105	.0020	.0003	.0000	.0000
	2	.9967	.9783	.9399	.8122	.4114	.1514	.0442	.0106	.0021	.0000	.0000
	3	.9998	.9971	.9881	.9392	.6474	.3217	.1227	.0374	.0093	.0003	.0000
	4	1.0000	.9997	.9982	.9844	.8245	.5245	.2552	.0979	.0302	.0015	.0000
	5	1.0000	1.0000	.9998	.9967	.9268	.7106	.4275	.2026	.0766	.0057	.0002
	6	1.0000	1.0000	1.0000	.9994	.9742	.8474	.6070	.3481	.1595	.0172	.0007
	7	1.0000	1.0000	1.0000	.9999	.9922	.9302	.7608	.5143	.2814	.0435	.0026
	8	1.0000	1.0000	1.0000	1.0000	.9980	.9722	.8713	.6736	.4315	.0940	.0081
	9	1.0000	1.0000	1.0000	1.0000	.9995	.9903	.9389	.8034	.5888	.1763	.0214
	10	1.0000	1.0000	1.0000	1.0000	.9999	.9971	.9744	.8943	.7304	.2915	.0494
	11	1.0000	1.0000	1.0000	1.0000	1.0000	.9992	.9905	.9493	.8407	.4311	.1002
	12	1.0000	1.0000	1.0000	1.0000	1.0000	.9998	.9969	.9784	.9155	.5785	.1808
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9991	.9918	.9599	.7145	.2923
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9998	.9973	.9831	.8246	.4278
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9992	.9936	.9029	.5722
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9998	.9979	.9519	.7077
	17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9994	.9788	.8192
	18	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9998	.9917	.8998
	19	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9971	.9506
	20	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9991	.9786
	21	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9998	.9919
	22	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9974
	23	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9993
	24	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9998
	25	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	26	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	27	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	28	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	29	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	30	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table A3. The Poisson distribution

The table gives values of $P(r;\mu) = \frac{1}{r!}\,\mu^{r}e^{-\mu}$ for the specified values of μ and r.

r \ μ	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
0	.9048	.8187	.7408	.6703	.6065	.5488	.4966	.4493	.4066	.3679
1	.0905	.1637	.2222	.2681	.3033	.3293	.3476	.3595	.3659	.3679
2	.0045	.0164	.0333	.0536	.0758	.0988	.1217	.1438	.1647	.1839
3	.0002	.0011	.0033	.0072	.0126	.0198	.0284	.0383	.0494	.0613
4	.0000	.0001	.0003	.0007	.0016	.0030	.0050	.0077	.0111	.0153
5	.0000	.0000	.0000	.0001	.0002	.0004	.0007	.0012	.0020	.0031
6	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0003	.0005
7	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001

r \ μ	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
0	.3329	.3012	.2725	.2466	.2231	.2019	.1827	.1653	.1496	.1353
1	.3662	.3614	.3543	.3452	.3347	.3230	.3106	.2975	.2842	.2707
2	.2014	.2169	.2303	.2417	.2510	.2584	.2640	.2678	.2700	.2707
3	.0738	.0867	.0998	.1128	.1255	.1378	.1496	.1607	.1710	.1804
4	.0203	.0260	.0324	.0395	.0471	.0551	.0636	.0723	.0812	.0902
5	.0045	.0062	.0084	.0111	.0141	.0176	.0216	.0260	.0309	.0361
6	.0008	.0012	.0018	.0026	.0035	.0047	.0061	.0078	.0098	.0120
7	.0001	.0002	.0003	.0005	.0008	.0011	.0015	.0020	.0027	.0034
8	.0000	.0000	.0001	.0001	.0001	.0002	.0003	.0005	.0006	.0009
9	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0002

r \ μ	2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8	2.9	3.0
0	.1225	.1108	.1003	.0907	.0821	.0743	.0672	.0608	.0550	.0498
1	.2572	.2438	.2306	.2177	.2052	.1931	.1815	.1703	.1596	.1494
2	.2700	.2681	.2652	.2613	.2565	.2510	.2450	.2384	.2314	.2240
3	.1890	.1966	.2033	.2090	.2138	.2176	.2205	.2225	.2237	.2240
4	.0992	.1082	.1169	.1254	.1336	.1414	.1488	.1557	.1622	.1680
5	.0417	.0476	.0538	.0602	.0668	.0735	.0804	.0872	.0940	.1008
6	.0146	.0174	.0206	.0241	.0278	.0319	.0362	.0407	.0455	.0504
7	.0044	.0055	.0068	.0083	.0099	.0118	.0139	.0163	.0188	.0216
8	.0011	.0015	.0019	.0025	.0031	.0038	.0047	.0057	.0068	.0081
9	.0003	.0004	.0005	.0007	.0009	.0011	.0014	.0018	.0022	.0027
10	.0001	.0001	.0001	.0002	.0002	.0003	.0004	.0005	.0006	.0008
11	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0002	.0002
12	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001

r \ μ	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9	4.0
0	.0450	.0408	.0369	.0334	.0302	.0273	.0247	.0224	.0202	.0183
1	.1397	.1304	.1217	.1135	.1057	.0984	.0915	.0850	.0789	.0733
2	.2165	.2087	.2008	.1929	.1850	.1771	.1692	.1615	.1539	.1465
3	.2237	.2226	.2209	.2186	.2158	.2125	.2087	.2046	.2001	.1954
4	.1733	.1781	.1823	.1858	.1888	.1912	.1931	.1944	.1951	.1954
5	.1075	.1140	.1203	.1264	.1322	.1377	.1429	.1477	.1522	.1563
6	.0555	.0608	.0662	.0716	.0771	.0826	.0881	.0936	.0989	.1042
7	.0246	.0278	.0312	.0348	.0385	.0425	.0466	.0508	.0551	.0595
8	.0095	.0111	.0129	.0148	.0169	.0191	.0215	.0241	.0269	.0298
9	.0033	.0040	.0047	.0056	.0066	.0076	.0089	.0102	.0116	.0132
10	.0010	.0013	.0016	.0019	.0023	.0028	.0033	.0039	.0045	.0053
11	.0003	.0004	.0005	.0006	.0007	.0009	.0011	.0013	.0016	.0019
12	.0001	.0001	.0001	.0002	.0002	.0003	.0003	.0004	.0005	.0006
13	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0001	.0002	.0002
14	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001

Table A3. The Poisson distribution (continued)

r \ μ	4.1	4.2	4.3	4.4	4.5	4.6	4.7	4.8	4.9	5.0
0	.0166	.0150	.0136	.0123	.0111	.0101	.0091	.0082	.0074	.0067
1	.0679	.0630	.0583	.0540	.0500	.0462	.0427	.0395	.0365	.0337
2	.1393	.1323	.1254	.1188	.1125	.1063	.1005	.0948	.0894	.0842
3	.1904	.1852	.1798	.1743	.1687	.1631	.1574	.1517	.1460	.1404
4	.1951	.1944	.1933	.1917	.1898	.1875	.1849	.1820	.1789	.1755
5	.1600	.1633	.1662	.1687	.1708	.1725	.1738	.1747	.1753	.1755
6	.1093	.1143	.1191	.1237	.1281	.1323	.1362	.1398	.1432	.1462
7	.0640	.0686	.0732	.0778	.0824	.0869	.0914	.0959	.1002	.1044
8	.0328	.0360	.0393	.0428	.0463	.0500	.0537	.0575	.0614	.0653
9	.0150	.0168	.0188	.0209	.0232	.0255	.0281	.0307	.0334	.0363
10	.0061	.0071	.0081	.0092	.0104	.0118	.0132	.0147	.0164	.0181
11	.0023	.0027	.0032	.0037	.0043	.0049	.0056	.0064	.0073	.0082
12	.0008	.0009	.0011	.0013	.0016	.0019	.0022	.0026	.0030	.0034
13	.0002	.0003	.0004	.0005	.0006	.0007	.0008	.0009	.0011	.0013
14	.0001	.0001	.0001	.0001	.0002	.0002	.0003	.0003	.0004	.0005
15	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0001	.0001	.0002
16	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000

r \ μ	5.1	5.2	5.3	5.4	5.5	5.6	5.7	5.8	5.9	6.0
0	.0061	.0055	.0050	.0045	.0041	.0037	.0033	.0030	.0027	.0025
1	.0311	.0287	.0265	.0244	.0225	.0207	.0191	.0176	.0162	.0149
2	.0793	.0746	.0701	.0659	.0618	.0580	.0544	.0509	.0477	.0446
3	.1348	.1293	.1239	.1185	.1133	.1082	.1033	.0985	.0938	.0892
4	.1719	.1681	.1641	.1600	.1558	.1515	.1472	.1428	.1383	.1339
5	.1753	.1748	.1740	.1728	.1714	.1697	.1678	.1656	.1632	.1606
6	.1490	.1515	.1537	.1555	.1571	.1584	.1594	.1601	.1605	.1606
7	.1086	.1125	.1163	.1200	.1234	.1267	.1298	.1326	.1353	.1377
8	.0692	.0731	.0771	.0810	.0849	.0887	.0925	.0962	.0998	.1033
9	.0392	.0423	.0454	.0486	.0519	.0552	.0586	.0620	.0656	.0688
10	.0200	.0220	.0241	.0262	.0285	.0309	.0334	.0359	.0386	.0413
11	.0093	.0104	.0116	.0129	.0143	.0157	.0173	.0190	.0207	.0225
12	.0039	.0045	.0051	.0058	.0065	.0073	.0082	.0092	.0102	.0113
13	.0015	.0018	.0021	.0024	.0028	.0032	.0036	.0041	.0046	.0052
14	.0006	.0007	.0008	.0009	.0011	.0013	.0015	.0017	.0019	.0022
15	.0002	.0002	.0003	.0003	.0004	.0005	.0006	.0007	.0008	.0009
16	.0001	.0001	.0001	.0001	.0001	.0002	.0002	.0002	.0003	.0003
17	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0001	.0001
18	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000

r \ μ	6.1	6.2	6.3	6.4	6.5	6.6	6.7	6.8	6.9	7.0
0	.0022	.0020	.0018	.0017	.0015	.0014	.0012	.0011	.0010	.0009
1	.0137	.0126	.0116	.0106	.0098	.0090	.0082	.0076	.0070	.0064
2	.0417	.0390	.0364	.0340	.0318	.0296	.0276	.0258	.0240	.0223
3	.0848	.0806	.0765	.0726	.0688	.0652	.0617	.0584	.0552	.0521
4	.1294	.1249	.1205	.1162	.1118	.1076	.1034	.0992	.0952	.0912
5	.1579	.1549	.1519	.1487	.1454	.1420	.1385	.1349	.1314	.1277
6	.1605	.1601	.1595	.1586	.1575	.1562	.1546	.1529	.1511	.1490
7	.1399	.1418	.1435	.1450	.1462	.1472	.1480	.1486	.1489	.1490
8	.1066	.1099	.1130	.1160	.1188	.1215	.1240	.1263	.1284	.1304
9	.0723	.0757	.0791	.0825	.0858	.0891	.0923	.0954	.0985	.1014
10	.0441	.0469	.0498	.0528	.0558	.0588	.0618	.0649	.0679	.0710
11	.0244	.0265	.0285	.0307	.0330	.0353	.0377	.0401	.0426	.0452
12	.0124	.0137	.0150	.0164	.0179	.0194	.0210	.0227	.0245	.0263
13	.0058	.0065	.0073	.0081	.0089	.0099	.0108	.0119	.0130	.0142
14	.0025	.0029	.0033	.0037	.0041	.0046	.0052	.0058	.0064	.0071
15	.0010	.0012	.0014	.0016	.0018	.0020	.0023	.0026	.0029	.0033
16	.0004	.0005	.0005	.0006	.0007	.0008	.0010	.0011	.0013	.0014
17	.0001	.0002	.0002	.0002	.0003	.0003	.0004	.0004	.0005	.0006
18	.0000	.0001	.0001	.0001	.0001	.0001	.0001	.0002	.0002	.0002
19	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0001

Table A3. The Poisson distribution (continued)

r \ μ	7.1	7.2	7.3	7.4	7.5	7.6	7.7	7.8	7.9	8.0
0	.0008	.0007	.0007	.0006	.0006	.0005	.0005	.0004	.0004	.0003
1	.0059	.0054	.0049	.0045	.0041	.0038	.0035	.0032	.0029	.0027
2	.0208	.0194	.0180	.0167	.0156	.0145	.0134	.0125	.0116	.0107
3	.0492	.0464	.0438	.0413	.0389	.0366	.0345	.0324	.0305	.0286
4	.0874	.0836	.0799	.0764	.0729	.0696	.0663	.0632	.0602	.0573
5	.1241	.1204	.1167	.1130	.1094	.1057	.1021	.0986	.0951	.0916
6	.1468	.1445	.1420	.1394	.1367	.1339	.1311	.1282	.1252	.1221
7	.1489	.1486	.1481	.1474	.1465	.1454	.1442	.1428	.1413	.1396
8	.1321	.1337	.1351	.1363	.1373	.1381	.1388	.1392	.1395	.1396
9	.1042	.1070	.1096	.1121	.1144	.1167	.1187	.1207	.1224	.1241
10	.0740	.0770	.0800	.0829	.0858	.0887	.0914	.0941	.0967	.0993
11	.0478	.0504	.0531	.0558	.0585	.0613	.0640	.0667	.0695	.0722
12	.0283	.0303	.0323	.0344	.0366	.0388	.0411	.0434	.0457	.0481
13	.0154	.0168	.0181	.0196	.0211	.0227	.0243	.0260	.0278	.0296
14	.0078	.0086	.0095	.0104	.0113	.0123	.0134	.0145	.0157	.0169
15	.0037	.0041	.0046	.0051	.0057	.0062	.0069	.0075	.0083	.0090
16	.0016	.0019	.0021	.0024	.0026	.0030	.0033	.0037	.0041	.0045
17	.0007	.0008	.0009	.0010	.0012	.0013	.0015	.0017	.0019	.0021
18	.0003	.0003	.0004	.0004	.0005	.0006	.0006	.0007	.0008	.0009
19	.0001	.0001	.0001	.0002	.0002	.0002	.0003	.0003	.0003	.0004
20	.0000	.0000	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0002
21	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001

r \ μ	8.1	8.2	8.3	8.4	8.5	8.6	8.7	8.8	8.9	9.0
0	.0003	.0003	.0002	.0002	.0002	.0002	.0002	.0002	.0001	.0001
1	.0025	.0023	.0021	.0019	.0017	.0016	.0014	.0013	.0012	.0011
2	.0100	.0092	.0086	.0079	.0074	.0068	.0063	.0058	.0054	.0050
3	.0269	.0252	.0237	.0222	.0208	.0195	.0183	.0171	.0160	.0150
4	.0544	.0517	.0491	.0466	.0443	.0420	.0398	.0377	.0357	.0337
5	.0882	.0849	.0816	.0784	.0752	.0722	.0692	.0663	.0635	.0607
6	.1191	.1160	.1128	.1097	.1066	.1034	.1003	.0972	.0941	.0911
7	.1378	.1358	.1338	.1317	.1294	.1271	.1247	.1222	.1197	.1171
8	.1395	.1392	.1388	.1382	.1375	.1366	.1356	.1344	.1332	.1318
9	.1256	.1269	.1280	.1290	.1299	.1306	.1311	.1315	.1317	.1318
10	.1017	.1040	.1063	.1084	.1104	.1123	.1140	.1157	.1172	.1186
11	.0749	.0776	.0802	.0828	.0853	.0878	.0902	.0925	.0948	.0970
12	.0505	.0530	.0555	.0579	.0604	.0629	.0654	.0679	.0703	.0728
13	.0315	.0334	.0354	.0374	.0395	.0416	.0438	.0459	.0481	.0504
14	.0182	.0196	.0210	.0225	.0240	.0256	.0272	.0289	.0306	.0324
15	.0098	.0107	.0116	.0126	.0136	.0147	.0158	.0169	.0182	.0194
16	.0050	.0055	.0060	.0066	.0072	.0079	.0086	.0093	.0101	.0109
17	.0024	.0026	.0029	.0033	.0036	.0040	.0044	.0048	.0053	.0058
18	.0011	.0012	.0014	.0015	.0017	.0019	.0021	.0024	.0026	.0029
19	.0005	.0005	.0006	.0007	.0008	.0009	.0010	.0011	.0012	.0014
20	.0002	.0002	.0002	.0003	.0003	.0004	.0004	.0005	.0005	.0006
21	.0001	.0001	.0001	.0001	.0001	.0002	.0002	.0002	.0002	.0003
22	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0001	.0001	.0001
23	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000

r \ μ	9.1	9.2	9.3	9.4	9.5	9.6	9.7	9.8	9.9	10.0
0	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0000
1	.0010	.0009	.0009	.0008	.0007	.0007	.0006	.0005	.0005	.0005
2	.0046	.0043	.0040	.0037	.0034	.0031	.0029	.0027	.0025	.0023
3	.0140	.0131	.0123	.0115	.0107	.0100	.0093	.0087	.0081	.0076
4	.0319	.0302	.0285	.0269	.0254	.0240	.0226	.0213	.0201	.0189
5	.0581	.0555	.0530	.0506	.0483	.0460	.0439	.0418	.0398	.0378
6	.0881	.0851	.0822	.0793	.0764	.0736	.0709	.0682	.0656	.0631
7	.1145	.1118	.1091	.1064	.1037	.1010	.0982	.0955	.0928	.0901
8	.1302	.1286	.1269	.1251	.1232	.1212	.1191	.1170	.1148	.1126
9	.1317	.1315	.1311	.1306	.1300	.1293	.1284	.1274	.1263	.1251
10	.1198	.1210	.1219	.1228	.1235	.1241	.1245	.1249	.1250	.1251
11	.0991	.1012	.1031	.1049	.1067	.1083	.1098	.1112	.1125	.1137
12	.0752	.0776	.0799	.0822	.0844	.0866	.0888	.0908	.0928	.0948
13	.0526	.0549	.0572	.0594	.0617	.0640	.0662	.0685	.0707	.0729
14	.0342	.0361	.0380	.0399	.0419	.0439	.0459	.0479	.0500	.0521
15	.0208	.0221	.0235	.0250	.0265	.0281	.0297	.0313	.0330	.0347
16	.0118	.0127	.0137	.0147	.0157	.0168	.0180	.0192	.0204	.0217
17	.0063	.0069	.0075	.0081	.0088	.0095	.0103	.0111	.0119	.0128
18	.0032	.0035	.0039	.0042	.0046	.0051	.0055	.0060	.0065	.0071
19	.0015	.0017	.0019	.0021	.0023	.0026	.0028	.0031	.0034	.0037
20	.0007	.0008	.0009	.0010	.0011	.0012	.0014	.0015	.0017	.0019
21	.0003	.0003	.0004	.0004	.0005	.0006	.0006	.0007	.0008	.0009
22	.0001	.0001	.0002	.0002	.0002	.0002	.0003	.0003	.0004	.0004
23	.0000	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0002	.0002
24	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001

Table A3. The Poisson distribution (continued)

r \ μ	11.0	12.0	13.0	14.0	15.0	16.0	17.0	18.0	19.0	20.0
0	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
1	.0002	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
2	.0010	.0004	.0002	.0001	.0000	.0000	.0000	.0000	.0000	.0000
3	.0037	.0018	.0008	.0004	.0002	.0001	.0000	.0000	.0000	.0000
4	.0102	.0053	.0027	.0013	.0006	.0003	.0001	.0001	.0000	.0000
5	.0224	.0127	.0070	.0037	.0019	.0010	.0005	.0002	.0001	.0001
6	.0411	.0255	.0152	.0087	.0048	.0026	.0014	.0007	.0004	.0002
7	.0646	.0437	.0281	.0174	.0104	.0060	.0034	.0019	.0010	.0005
8	.0888	.0655	.0457	.0304	.0194	.0120	.0072	.0042	.0024	.0013
9	.1085	.0874	.0661	.0473	.0324	.0213	.0135	.0083	.0050	.0029
10	.1194	.1048	.0859	.0663	.0486	.0341	.0230	.0150	.0095	.0058
11	.1194	.1144	.1015	.0844	.0663	.0496	.0355	.0245	.0164	.0106
12	.1094	.1144	.1099	.0984	.0829	.0661	.0504	.0368	.0259	.0176
13	.0926	.1056	.1099	.1060	.0956	.0814	.0658	.0509	.0378	.0271
14	.0728	.0905	.1021	.1060	.1024	.0930	.0800	.0655	.0514	.0387
15	.0534	.0724	.0885	.0989	.1024	.0992	.0906	.0786	.0650	.0516
16	.0367	.0543	.0719	.0866	.0960	.0992	.0963	.0884	.0772	.0646
17	.0237	.0383	.0550	.0713	.0847	.0934	.0963	.0936	.0863	.0760
18	.0145	.0255	.0397	.0554	.0706	.0830	.0909	.0936	.0911	.0844
19	.0084	.0161	.0272	.0409	.0557	.0699	.0814	.0887	.0911	.0888
20	.0046	.0097	.0177	.0286	.0418	.0559	.0692	.0798	.0866	.0888
21	.0024	.0055	.0109	.0191	.0299	.0426	.0560	.0684	.0783	.0846
22	.0012	.0030	.0065	.0121	.0204	.0310	.0433	.0560	.0676	.0769
23	.0006	.0016	.0037	.0074	.0133	.0216	.0320	.0438	.0559	.0669
24	.0003	.0008	.0020	.0043	.0083	.0144	.0226	.0328	.0442	.0557
25	.0001	.0004	.0010	.0024	.0050	.0092	.0154	.0237	.0336	.0446
26	.0000	.0002	.0005	.0013	.0029	.0057	.0101	.0164	.0246	.0343
27	.0000	.0001	.0002	.0007	.0016	.0034	.0063	.0109	.0173	.0254
28	.0000	.0000	.0001	.0003	.0009	.0019	.0038	.0070	.0117	.0181
29	.0000	.0000	.0001	.0002	.0004	.0011	.0023	.0044	.0077	.0125
30	.0000	.0000	.0000	.0001	.0002	.0006	.0013	.0026	.0049	.0083
31	.0000	.0000	.0000	.0000	.0001	.0003	.0007	.0015	.0030	.0054
32	.0000	.0000	.0000	.0000	.0001	.0001	.0004	.0009	.0018	.0034
33	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0005	.0010	.0020
34	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0006	.0012
35	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0007
36	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0004
37	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002
38	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001
39	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001
40	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000

Table A4. The cumulative Poisson distribution

The table gives values of $F(x;\mu) = \sum_{r=0}^{x} \frac{1}{r!} \mu^r e^{-\mu}$ for the specified values of μ and x.

x \ μ	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
0	.9048	.8187	.7408	.6703	.6065	.5488	.4966	.4493	.4066	.3679
1	.9953	.9825	.9631	.9384	.9098	.8781	.8442	.8088	.7725	.7358
2	.9998	.9989	.9964	.9921	.9856	.9769	.9659	.9526	.9371	.9197
3	1.0000	.9999	.9997	.9992	.9982	.9966	.9942	.9909	.9865	.9810
4	1.0000	1.0000	1.0000	.9999	.9998	.9996	.9992	.9986	.9977	.9963
5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9998	.9997	.9994
6	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999
7	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

x \ μ	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
0	.3329	.3012	.2725	.2466	.2231	.2019	.1827	.1653	.1496	.1353
1	.6990	.6626	.6268	.5918	.5578	.5249	.4932	.4628	.4337	.4060
2	.9004	.8795	.8571	.8335	.8088	.7834	.7572	.7306	.7037	.6767
3	.9743	.9662	.9569	.9463	.9344	.9212	.9068	.8913	.8747	.8571
4	.9946	.9923	.9893	.9857	.9814	.9763	.9704	.9636	.9559	.9473
5	.9990	.9985	.9978	.9968	.9955	.9940	.9920	.9896	.9868	.9834
6	.9999	.9997	.9996	.9994	.9991	.9987	.9981	.9974	.9966	.9955
7	1.0000	1.0000	.9999	.9999	.9998	.9997	.9996	.9994	.9992	.9989
8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9999	.9998	.9998
9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

x \ μ	2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8	2.9	3.0
0	.1225	.1108	.1003	.0907	.0821	.0743	.0672	.0608	.0550	.0498
1	.3796	.3546	.3309	.3084	.2873	.2674	.2487	.2311	.2146	.1991
2	.6496	.6227	.5960	.5697	.5438	.5184	.4936	.4695	.4460	.4232
3	.8386	.8194	.7993	.7787	.7576	.7360	.7141	.6919	.6696	.6472
4	.9379	.9275	.9162	.9041	.8912	.8774	.8629	.8477	.8318	.8153
5	.9796	.9751	.9700	.9643	.9580	.9510	.9433	.9349	.9258	.9161
6	.9941	.9925	.9906	.9884	.9858	.9828	.9794	.9756	.9713	.9665
7	.9985	.9980	.9974	.9967	.9958	.9947	.9934	.9919	.9901	.9881
8	.9997	.9995	.9994	.9991	.9989	.9985	.9981	.9976	.9969	.9962
9	.9999	.9999	.9999	.9998	.9997	.9996	.9995	.9993	.9991	.9989
10	1.0000	1.0000	1.0000	1.0000	.9999	.9999	.9999	.9998	.9998	.9997
11	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9999
12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

x \ μ	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9	4.0
0	.0450	.0408	.0369	.0334	.0302	.0273	.0247	.0224	.0202	.0183
1	.1847	.1712	.1586	.1468	.1359	.1257	.1162	.1074	.0992	.0916
2	.4012	.3799	.3594	.3397	.3208	.3027	.2854	.2689	.2531	.2381
3	.6248	.6025	.5803	.5584	.5366	.5152	.4942	.4735	.4532	.4335
4	.7982	.7806	.7626	.7442	.7254	.7064	.6872	.6678	.6484	.6288
5	.9057	.8946	.8829	.8705	.8576	.8441	.8301	.8156	.8006	.7851
6	.9612	.9554	.9490	.9421	.9347	.9267	.9182	.9091	.8995	.8893
7	.9858	.9832	.9802	.9769	.9733	.9692	.9648	.9599	.9546	.9489
8	.9953	.9943	.9931	.9917	.9901	.9883	.9863	.9840	.9815	.9786
9	.9986	.9982	.9978	.9973	.9967	.9960	.9952	.9942	.9931	.9919
10	.9996	.9995	.9994	.9992	.9990	.9987	.9984	.9981	.9977	.9972
11	.9999	.9999	.9998	.9998	.9997	.9996	.9995	.9994	.9993	.9991
12	1.0000	1.0000	1.0000	.9999	.9999	.9999	.9999	.9998	.9998	.9997
13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9999
14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table A4. The cumulative Poisson distribution (continued)

x \ μ	4.1	4.2	4.3	4.4	4.5	4.6	4.7	4.8	4.9	5.0
0	.0166	.0150	.0136	.0123	.0111	.0101	.0091	.0082	.0074	.0067
1	.0845	.0780	.0719	.0663	.0611	.0563	.0518	.0477	.0439	.0404
2	.2238	.2102	.1974	.1851	.1736	.1626	.1523	.1425	.1333	.1247
3	.4142	.3954	.3772	.3594	.3423	.3257	.3097	.2942	.2793	.2650
4	.6093	.5898	.5704	.5512	.5321	.5132	.4946	.4763	.4582	.4405
5	.7693	.7531	.7367	.7199	.7029	.6858	.6684	.6510	.6335	.6160
6	.8786	.8675	.8558	.8436	.8311	.8180	.8046	.7908	.7767	.7622
7	.9427	.9361	.9290	.9214	.9134	.9049	.8960	.8867	.8769	.8666
8	.9755	.9721	.9683	.9642	.9597	.9549	.9497	.9442	.9382	.9319
9	.9905	.9889	.9871	.9851	.9829	.9805	.9778	.9749	.9717	.9682
10	.9966	.9959	.9952	.9943	.9933	.9922	.9910	.9896	.9880	.9863
11	.9989	.9986	.9983	.9980	.9976	.9971	.9966	.9960	.9953	.9945
12	.9997	.9996	.9995	.9993	.9992	.9990	.9988	.9986	.9983	.9980
13	.9999	.9999	.9998	.9998	.9997	.9997	.9996	.9995	.9994	.9993
14	1.0000	1.0000	1.0000	.9999	.9999	.9999	.9999	.9999	.9998	.9998
15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9999
16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

x \ μ	5.1	5.2	5.3	5.4	5.5	5.6	5.7	5.8	5.9	6.0
0	.0061	.0055	.0050	.0045	.0041	.0037	.0033	.0030	.0027	.0025
1	.0372	.0342	.0314	.0289	.0266	.0244	.0224	.0206	.0189	.0174
2	.1165	.1088	.1016	.0948	.0884	.0824	.0768	.0715	.0666	.0620
3	.2513	.2381	.2254	.2133	.2017	.1906	.1800	.1700	.1604	.1512
4	.4231	.4061	.3895	.3733	.3575	.3422	.3272	.3127	.2987	.2851
5	.5984	.5809	.5635	.5461	.5289	.5119	.4950	.4783	.4619	.4457
6	.7474	.7324	.7171	.7017	.6860	.6703	.6544	.6384	.6224	.6063
7	.8560	.8449	.8335	.8217	.8095	.7970	.7841	.7710	.7576	.7440
8	.9252	.9181	.9106	.9027	.8944	.8857	.8766	.8672	.8574	.8472
9	.9644	.9603	.9559	.9512	.9462	.9409	.9352	.9292	.9228	.9161
10	.9844	.9823	.9800	.9775	.9747	.9718	.9686	.9651	.9614	.9574
11	.9937	.9927	.9916	.9904	.9890	.9875	.9859	.9841	.9821	.9799
12	.9976	.9972	.9967	.9962	.9955	.9949	.9941	.9932	.9922	.9912
13	.9992	.9990	.9988	.9986	.9983	.9980	.9977	.9973	.9969	.9964
14	.9997	.9997	.9996	.9995	.9994	.9993	.9991	.9990	.9988	.9986
15	.9999	.9999	.9999	.9998	.9998	.9998	.9997	.9996	.9996	.9995
16	1.0000	1.0000	1.0000	.9999	.9999	.9999	.9999	.9999	.9999	.9998
17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999
18	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

x \ μ	6.1	6.2	6.3	6.4	6.5	6.6	6.7	6.8	6.9	7.0
0	.0022	.0020	.0018	.0017	.0015	.0014	.0012	.0011	.0010	.0009
1	.0159	.0146	.0134	.0123	.0113	.0103	.0095	.0087	.0080	.0073
2	.0577	.0536	.0498	.0463	.0430	.0400	.0371	.0344	.0320	.0296
3	.1425	.1342	.1264	.1189	.1118	.1052	.0988	.0928	.0871	.0818
4	.2719	.2592	.2469	.2351	.2237	.2127	.2022	.1920	.1823	.1730
5	.4298	.4141	.3988	.3837	.3690	.3547	.3406	.3270	.3137	.3007
6	.5902	.5742	.5582	.5423	.5265	.5108	.4953	.4799	.4647	.4497
7	.7301	.7160	.7017	.6873	.6728	.6581	.6433	.6285	.6136	.5987
8	.8367	.8259	.8148	.8033	.7916	.7796	.7673	.7548	.7420	.7291
9	.9090	.9016	.8939	.8858	.8774	.8686	.8596	.8502	.8405	.8305
10	.9531	.9486	.9437	.9386	.9332	.9274	.9214	.9151	.9084	.9015
11	.9776	.9750	.9723	.9693	.9661	.9627	.9591	.9552	.9510	.9467
12	.9900	.9887	.9873	.9857	.9840	.9821	.9801	.9779	.9755	.9730
13	.9958	.9952	.9945	.9937	.9929	.9920	.9909	.9898	.9885	.9872
14	.9984	.9981	.9978	.9974	.9970	.9966	.9961	.9956	.9950	.9943
15	.9994	.9993	.9992	.9990	.9988	.9986	.9984	.9982	.9979	.9976
16	.9998	.9997	.9997	.9996	.9996	.9995	.9994	.9993	.9992	.9990
17	.9999	.9999	.9999	.9999	.9998	.9998	.9998	.9997	.9997	.9996
18	1.0000	1.0000	1.0000	1.0000	.9999	.9999	.9999	.9999	.9999	.9999
19	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table A4. The cumulative Poisson distribution (continued)

x \ μ	7.1	7.2	7.3	7.4	7.5	7.6	7.7	7.8	7.9	8.0
0	.0008	.0007	.0007	.0006	.0006	.0005	.0005	.0004	.0004	.0003
1	.0067	.0061	.0056	.0051	.0047	.0043	.0039	.0036	.0033	.0030
2	.0275	.0255	.0236	.0219	.0203	.0188	.0174	.0161	.0149	.0138
3	.0767	.0719	.0674	.0632	.0591	.0554	.0518	.0485	.0453	.0424
4	.1641	.1555	.1473	.1395	.1321	.1249	.1181	.1117	.1055	.0996
5	.2881	.2759	.2640	.2526	.2414	.2307	.2203	.2103	.2006	.1912
6	.4349	.4204	.4060	.3920	.3782	.3646	.3514	.3384	.3257	.3134
7	.5838	.5689	.5541	.5393	.5246	.5100	.4956	.4812	.4670	.4530
8	.7160	.7027	.6892	.6757	.6620	.6482	.6343	.6204	.6065	.5925
9	.8202	.8096	.7988	.7877	.7764	.7649	.7531	.7411	.7290	.7166
10	.8942	.8867	.8788	.8707	.8622	.8535	.8445	.8352	.8257	.8159
11	.9420	.9371	.9319	.9265	.9208	.9148	.9085	.9020	.8952	.8881
12	.9703	.9673	.9642	.9609	.9573	.9536	.9496	.9454	.9409	.9362
13	.9857	.9841	.9824	.9805	.9784	.9762	.9739	.9714	.9687	.9658
14	.9935	.9927	.9918	.9908	.9897	.9886	.9873	.9859	.9844	.9827
15	.9972	.9969	.9964	.9959	.9954	.9948	.9941	.9934	.9926	.9918
16	.9989	.9987	.9985	.9983	.9980	.9978	.9974	.9971	.9967	.9963
17	.9996	.9995	.9994	.9993	.9992	.9991	.9989	.9988	.9986	.9984
18	.9998	.9998	.9998	.9997	.9997	.9996	.9996	.9995	.9994	.9993
19	.9999	.9999	.9999	.9999	.9999	.9999	.9998	.9998	.9998	.9997
20	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9999	.9999	.9999
21	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

x \ μ	8.1	8.2	8.3	8.4	8.5	8.6	8.7	8.8	8.9	9.0
0	.0003	.0003	.0002	.0002	.0002	.0002	.0002	.0002	.0001	.0001
1	.0028	.0025	.0023	.0021	.0019	.0018	.0016	.0015	.0014	.0012
2	.0127	.0118	.0109	.0100	.0093	.0086	.0079	.0073	.0068	.0062
3	.0396	.0370	.0346	.0323	.0301	.0281	.0262	.0244	.0228	.0212
4	.0940	.0887	.0837	.0789	.0744	.0701	.0660	.0621	.0584	.0550
5	.1822	.1736	.1653	.1573	.1496	.1422	.1352	.1284	.1219	.1157
6	.3013	.2896	.2781	.2670	.2562	.2457	.2355	.2256	.2160	.2068
7	.4391	.4254	.4119	.3987	.3856	.3728	.3602	.3478	.3357	.3239
8	.5786	.5647	.5507	.5369	.5231	.5094	.4958	.4823	.4689	.4557
9	.7041	.6915	.6788	.6659	.6530	.6400	.6269	.6137	.6006	.5874
10	.8058	.7955	.7850	.7743	.7634	.7522	.7409	.7294	.7178	.7060
11	.8807	.8731	.8652	.8571	.8487	.8400	.8311	.8220	.8126	.8030
12	.9313	.9261	.9207	.9150	.9091	.9029	.8965	.8898	.8829	.8758
13	.9628	.9595	.9561	.9524	.9486	.9445	.9403	.9358	.9311	.9261
14	.9810	.9791	.9771	.9749	.9726	.9701	.9675	.9647	.9617	.9585
15	.9908	.9898	.9887	.9875	.9862	.9848	.9832	.9816	.9798	.9780
16	.9958	.9953	.9947	.9941	.9934	.9926	.9918	.9909	.9899	.9889
17	.9982	.9979	.9977	.9973	.9970	.9966	.9962	.9957	.9952	.9947
18	.9992	.9991	.9990	.9989	.9987	.9985	.9983	.9981	.9978	.9976
19	.9997	.9997	.9996	.9995	.9995	.9994	.9993	.9992	.9991	.9989
20	.9999	.9999	.9998	.9998	.9998	.9998	.9997	.9997	.9996	.9996
21	1.0000	1.0000	.9999	.9999	.9999	.9999	.9999	.9999	.9998	.9998
22	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9999
23	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

x \ μ	9.1	9.2	9.3	9.4	9.5	9.6	9.7	9.8	9.9	10.0
0	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0000
1	.0011	.0010	.0009	.0009	.0008	.0007	.0007	.0006	.0005	.0005
2	.0058	.0053	.0049	.0045	.0042	.0038	.0035	.0033	.0030	.0028
3	.0198	.0184	.0172	.0160	.0149	.0138	.0129	.0120	.0111	.0103
4	.0517	.0486	.0456	.0429	.0403	.0378	.0355	.0333	.0312	.0293
5	.1098	.1041	.0986	.0935	.0885	.0838	.0793	.0750	.0710	.0671
6	.1978	.1892	.1808	.1727	.1649	.1574	.1502	.1433	.1366	.1301
7	.3123	.3010	.2900	.2792	.2687	.2584	.2485	.2388	.2294	.2202
8	.4426	.4296	.4168	.4042	.3918	.3796	.3676	.3558	.3442	.3328
9	.5742	.5611	.5479	.5349	.5218	.5089	.4960	.4832	.4705	.4579
10	.6941	.6820	.6699	.6576	.6453	.6329	.6205	.6080	.5955	.5830
11	.7932	.7832	.7730	.7626	.7520	.7412	.7303	.7193	.7081	.6968
12	.8684	.8607	.8529	.8448	.8364	.8279	.8191	.8101	.8009	.7916
13	.9210	.9156	.9100	.9042	.8981	.8919	.8853	.8786	.8716	.8645
14	.9552	.9517	.9480	.9441	.9400	.9357	.9312	.9265	.9216	.9165
15	.9760	.9738	.9715	.9691	.9665	.9638	.9609	.9579	.9546	.9513
16	.9878	.9865	.9852	.9838	.9823	.9806	.9789	.9770	.9751	.9730
17	.9941	.9934	.9927	.9919	.9911	.9902	.9892	.9881	.9870	.9857
18	.9973	.9969	.9966	.9962	.9957	.9952	.9947	.9941	.9935	.9928
19	.9988	.9986	.9985	.9983	.9980	.9978	.9975	.9972	.9969	.9965
20	.9995	.9994	.9993	.9992	.9991	.9990	.9989	.9987	.9986	.9984
21	.9998	.9998	.9997	.9997	.9996	.9996	.9995	.9995	.9994	.9993
22	.9999	.9999	.9999	.9999	.9999	.9998	.9998	.9998	.9997	.9997
23	1.0000	1.0000	1.0000	1.0000	.9999	.9999	.9999	.9999	.9999	.9999
24	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table A4. The cumulative Poisson distribution (continued)

x \ μ	11.0	12.0	13.0	14.0	15.0	16.0	17.0	18.0	19.0	20.0
0	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
1	.0002	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
2	.0012	.0005	.0002	.0001	.0000	.0000	.0000	.0000	.0000	.0000
3	.0049	.0023	.0011	.0005	.0002	.0001	.0000	.0000	.0000	.0000
4	.0151	.0076	.0037	.0018	.0009	.0004	.0002	.0001	.0000	.0000
5	.0375	.0203	.0107	.0055	.0028	.0014	.0007	.0003	.0002	.0001
6	.0786	.0458	.0259	.0142	.0076	.0040	.0021	.0010	.0005	.0003
7	.1432	.0895	.0540	.0316	.0180	.0100	.0054	.0029	.0015	.0008
8	.2320	.1550	.0998	.0621	.0374	.0220	.0126	.0071	.0039	.0021
9	.3405	.2424	.1658	.1094	.0699	.0433	.0261	.0154	.0089	.0050
10	.4599	.3472	.2517	.1757	.1185	.0774	.0491	.0304	.0183	.0108
11	.5793	.4616	.3532	.2600	.1848	.1270	.0847	.0549	.0347	.0214
12	.6887	.5760	.4631	.3585	.2676	.1931	.1350	.0917	.0606	.0390
13	.7813	.6815	.5730	.4644	.3632	.2745	.2009	.1426	.0984	.0661
14	.8540	.7720	.6751	.5704	.4657	.3675	.2808	.2081	.1497	.1049
15	.9074	.8444	.7636	.6694	.5681	.4667	.3715	.2867	.2148	.1565
16	.9441	.8987	.8355	.7559	.6641	.5660	.4677	.3751	.2920	.2211
17	.9678	.9370	.8905	.8272	.7489	.6593	.5640	.4686	.3784	.2970
18	.9823	.9626	.9302	.8826	.8195	.7423	.6550	.5622	.4695	.3814
19	.9907	.9787	.9573	.9235	.8752	.8122	.7363	.6509	.5606	.4703
20	.9953	.9884	.9750	.9521	.9170	.8682	.8055	.7307	.6472	.5591
21	.9977	.9939	.9859	.9712	.9469	.9108	.8615	.7991	.7255	.6437
22	.9990	.9970	.9924	.9833	.9673	.9418	.9047	.8551	.7931	.7206
23	.9995	.9985	.9960	.9907	.9805	.9633	.9367	.8989	.8490	.7875
24	.9998	.9993	.9980	.9950	.9888	.9777	.9594	.9317	.8933	.8432
25	.9999	.9997	.9990	.9974	.9938	.9869	.9748	.9554	.9269	.8878
26	1.0000	.9999	.9995	.9987	.9967	.9925	.9848	.9718	.9514	.9221
27	1.0000	.9999	.9998	.9994	.9983	.9959	.9912	.9827	.9687	.9475
28	1.0000	1.0000	.9999	.9997	.9991	.9978	.9950	.9897	.9805	.9657
29	1.0000	1.0000	1.0000	.9999	.9996	.9989	.9973	.9941	.9882	.9782
30	1.0000	1.0000	1.0000	.9999	.9998	.9994	.9986	.9967	.9930	.9865
31	1.0000	1.0000	1.0000	1.0000	.9999	.9997	.9993	.9982	.9960	.9919
32	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9996	.9990	.9978	.9953
33	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9998	.9995	.9988	.9973
34	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9998	.9994	.9985
35	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9997	.9992
36	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9998	.9996
37	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9998
38	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999
39	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	.9999
40	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table A5. The standard normal probability density function

The table gives values of $g(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$ for $0 \le x \le 4.99$.

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.39894	.39892	.39886	.39876	.39862	.39844	.39822	.39797	.39767	.39733
.1	.39695	.39654	.39608	.39559	.39505	.39448	.39387	.39322	.39253	.39181
.2	.39104	.39024	.38940	.38853	.38762	.38667	.38568	.38466	.38361	.38251
.3	.38139	.38023	.37903	.37780	.37654	.37524	.37391	.37255	.37115	.36973
.4	.36827	.36678	.36526	.36371	.36213	.36053	.35889	.35723	.35553	.35381
.5	.35207	.35029	.34849	.34667	.34482	.34294	.34105	.33912	.33718	.33521
.6	.33322	.33121	.32918	.32713	.32506	.32297	.32086	.31874	.31659	.31443
.7	.31225	.31006	.30785	.30563	.30339	.30114	.29887	.29659	.29431	.29200
.8	.28969	.28737	.28504	.28269	.28034	.27798	.27562	.27324	.27086	.26848
.9	.26609	.26369	.26129	.25888	.25647	.25406	.25164	.24923	.24681	.24439
1.0	.24197	.23955	.23713	.23471	.23230	.22988	.22747	.22506	.22265	.22025
1.1	.21785	.21546	.21307	.21069	.20831	.20594	.20357	.20121	.19886	.19652
1.2	.19419	.19186	.18954	.18724	.18494	.18265	.18037	.17810	.17585	.17360
1.3	.17137	.16915	.16694	.16474	.16256	.16038	.15822	.15608	.15395	.15183
1.4	.14973	.14764	.14556	.14350	.14146	.13943	.13742	.13542	.13344	.13147
1.5	.12952	.12758	.12566	.12376	.12188	.12001	.11816	.11632	.11450	.11270
1.6	.11092	.10915	.10741	.10567	.10396	.10226	.10059	.09893	.09728	.09566
1.7	.09405	.09246	.09089	.08933	.08780	.08628	.08478	.08329	.08183	.08038
1.8	.07895	.07754	.07614	.07477	.07341	.07206	.07074	.06943	.06814	.06687
1.9	.06562	.06438	.06316	.06195	.06077	.05959	.05844	.05730	.05618	.05508
2.0	.05399	.05292	.05186	.05082	.04980	.04879	.04780	.04682	.04586	.04491
2.1	.04398	.04307	.04217	.04128	.04041	.03955	.03871	.03788	.03706	.03626
2.2	.03547	.03470	.03394	.03319	.03246	.03174	.03103	.03034	.02965	.02898
2.3	.02833	.02768	.02705	.02643	.02582	.02522	.02463	.02406	.02349	.02294
2.4	.02239	.02186	.02134	.02083	.02033	.01984	.01936	.01888	.01842	.01797
2.5	.01753	.01709	.01667	.01625	.01585	.01545	.01506	.01468	.01431	.01394
2.6	.01358	.01323	.01289	.01256	.01223	.01191	.01160	.01130	.01100	.01071
2.7	.01042	.01014	.00987	.00961	.00935	.00909	.00885	.00861	.00837	.00814
2.8	.00792	.00770	.00748	.00727	.00707	.00687	.00668	.00649	.00631	.00613
2.9	.00595	.00578	.00562	.00545	.00530	.00514	.00499	.00485	.00470	.00457
3.0	.00443	.00430	.00417	.00405	.00393	.00381	.00370	.00358	.00348	.00337
3.1	.00327	.00317	.00307	.00298	.00288	.00279	.00271	.00262	.00254	.00246
3.2	.00238	.00231	.00224	.00216	.00210	.00203	.00196	.00190	.00184	.00178
3.3	.00172	.00167	.00161	.00156	.00151	.00146	.00141	.00136	.00132	.00127
3.4	.00123	.00119	.00115	.00111	.00107	.00104	.00100	.00097	.00094	.00090
3.5	.00087	.00084	.00081	.00079	.00076	.00073	.00071	.00068	.00066	.00063
3.6	.00061	.00059	.00057	.00055	.00053	.00051	.00049	.00047	.00046	.00044
3.7	.00042	.00041	.00039	.00038	.00037	.00035	.00034	.00033	.00031	.00030
3.8	.00029	.00028	.00027	.00026	.00025	.00024	.00023	.00022	.00021	.00021
3.9	.00020	.00019	.00018	.00018	.00017	.00016	.00016	.00015	.00014	.00014
4.0	.00013	.00013	.00012	.00012	.00011	.00011	.00011	.00010	.00010	.00009
4.1	.00009	.00009	.00008	.00008	.00008	.00007	.00007	.00007	.00006	.00006
4.2	.00006	.00006	.00005	.00005	.00005	.00005	.00005	.00004	.00004	.00004
4.3	.00004	.00004	.00004	.00003	.00003	.00003	.00003	.00003	.00003	.00003
4.4	.00002	.00002	.00002	.00002	.00002	.00002	.00002	.00002	.00002	.00002
4.5	.00002	.00002	.00001	.00001	.00001	.00001	.00001	.00001	.00001	.00001
4.6	.00001	.00001	.00001	.00001	.00001	.00001	.00001	.00001	.00001	.00001
4.7	.00001	.00001	.00001	.00001	.00001	.00001	.00000	.00000	.00000	.00000
4.8	.00000	.00000	.00000	.00000	.00000	.00000	.00000	.00000	.00000	.00000
4.9	.00000	.00000	.00000	.00000	.00000	.00000	.00000	.00000	.00000	.00000

Table A6. The cumulative standard normal distribution

The table gives values of $G(y) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx$ for $0 \le y \le 4.99$. $G(-y) = 1 - G(y)$.

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56749	.57142	.57535
.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
2.0	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
2.3	.98928	.98956	.98983	.99010	.99036	.99061	.99086	.99111	.99134	.99158
2.4	.99180	.99202	.99224	.99245	.99266	.99286	.99305	.99324	.99343	.99361
2.5	.99379	.99396	.99413	.99430	.99446	.99461	.99477	.99492	.99506	.99520
2.6	.99534	.99547	.99560	.99573	.99585	.99598	.99609	.99621	.99632	.99643
2.7	.99653	.99664	.99674	.99683	.99693	.99702	.99711	.99720	.99728	.99736
2.8	.99744	.99752	.99760	.99767	.99774	.99781	.99788	.99795	.99801	.99807
2.9	.99813	.99819	.99825	.99831	.99836	.99841	.99846	.99851	.99856	.99861
3.0	.99865	.99869	.99874	.99878	.99882	.99886	.99889	.99893	.99896	.99900
3.1	.99903	.99906	.99910	.99913	.99916	.99918	.99921	.99924	.99926	.99929
3.2	.99931	.99934	.99936	.99938	.99940	.99942	.99944	.99946	.99948	.99950
3.3	.99952	.99953	.99955	.99957	.99958	.99960	.99961	.99962	.99964	.99965
3.4	.99966	.99968	.99969	.99970	.99971	.99972	.99973	.99974	.99975	.99976
3.5	.99977	.99978	.99978	.99979	.99980	.99981	.99981	.99982	.99983	.99983
3.6	.99984	.99985	.99985	.99986	.99986	.99987	.99987	.99988	.99988	.99989
3.7	.99989	.99990	.99990	.99990	.99991	.99991	.99992	.99992	.99992	.99992
3.8	.99993	.99993	.99993	.99994	.99994	.99994	.99994	.99995	.99995	.99995
3.9	.99995	.99995	.99996	.99996	.99996	.99996	.99996	.99996	.99997	.99997
4.0	.99997	.99997	.99997	.99997	.99997	.99997	.99998	.99998	.99998	.99998
4.1	.99998	.99998	.99998	.99998	.99998	.99998	.99998	.99998	.99999	.99999
4.2	.99999	.99999	.99999	.99999	.99999	.99999	.99999	.99999	.99999	.99999
4.3	.99999	.99999	.99999	.99999	.99999	.99999	.99999	.99999	.99999	.99999
4.4	.99999	.99999	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000
4.5	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000
4.6	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000
4.7	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000
4.8	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000
4.9	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000

Table A7. Percentage points of the Student's t-distribution

The table gives values of t_α for different degrees of freedom ν such as to produce specified values $F(t_\alpha;\nu)$, where

$$F(t_\alpha;\nu) = \int_{-\infty}^{t_\alpha} \frac{\Gamma\left(\frac{1}{2}(\nu+1)\right)}{\sqrt{\pi\nu}\,\Gamma\left(\frac{1}{2}\nu\right)} \cdot \frac{1}{\left(1+\frac{t^2}{\nu}\right)^{\frac{1}{2}(\nu+1)}} dt = 1-\alpha .$$

$F(-t;\nu) = 1 - F(t;\nu)$. $\nu = \infty$ corresponds to the standard normal distribution.

ν \ F	.60	.70	.80	.90	.95	.975	.990	.995	.999	.9995
1	.325	.727	1.376	3.078	6.314	12.706	31.821	63.657	318.31	663.62
2	.289	.617	1.061	1.886	2.920	4.303	6.965	9.925	22.327	31.598
3	.277	.584	.978	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	.271	.569	.941	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	.267	.559	.920	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	.265	.553	.906	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	.263	.549	.896	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	.262	.546	.889	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	.261	.543	.883	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	.260	.542	.879	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	.260	.540	.876	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	.259	.539	.873	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	.259	.538	.870	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	.258	.537	.868	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	.258	.536	.866	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	.258	.535	.865	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	.257	.534	.863	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	.257	.534	.862	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	.257	.533	.861	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	.257	.533	.860	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	.257	.532	.859	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	.256	.532	.858	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	.256	.532	.858	1.319	1.714	2.069	2.500	2.807	3.485	3.767
24	.256	.531	.857	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	.256	.531	.856	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	.256	.531	.856	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	.256	.531	.855	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	.256	.530	.855	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	.256	.530	.854	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	.256	.530	.854	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	.255	.529	.851	1.303	1.684	2.021	2.423	2.704	3.307	3.551
50	.255	.528	.849	1.299	1.676	2.009	2.403	2.678	3.261	3.496
60	.254	.527	.848	1.296	1.671	2.000	2.390	2.660	3.232	3.460
70	.254	.527	.847	1.294	1.667	1.994	2.381	2.648	3.211	3.435
80	.254	.527	.846	1.292	1.664	1.990	2.374	2.639	3.195	3.416
90	.254	.526	.846	1.291	1.662	1.987	2.369	2.632	3.183	3.402
100	.254	.526	.845	1.290	1.660	1.984	2.364	2.626	3.174	3.391
110	.254	.526	.845	1.289	1.659	1.982	2.361	2.621	3.166	3.381
120	.254	.526	.845	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	.253	.524	.842	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Table A8. Percentage points of the chi-square distribution

The table gives values of χ^2_α for different degrees of freedom ν such as to produce specified values $F(\chi^2_\alpha;\nu)$, where

$$F(\chi^2_\alpha;\nu) = \int_0^{\chi^2_\alpha} \frac{1}{2^{\frac{1}{2}\nu}\,\Gamma(\frac{1}{2}\nu)}\, u^{\frac{1}{2}\nu-1}\, e^{-\frac{1}{2}u}\, du = 1 - \alpha\ .$$

ν \ F	.005	.010	.025	.050	.100	.200	.250	.500	.750	.800	.900	.950	.975	.990	.995
1	.00004	.00016	.00098	.00393	.0158	.064	.102	.455	1.323	1.642	2.706	3.841	5.024	6.635	7.879
2	.0100	.0201	.0506	.103	.211	.446	.575	1.386	2.773	3.219	4.605	5.991	7.378	9.210	10.597
3	.0717	.115	.216	.352	.584	1.005	1.213	2.366	4.108	4.642	6.251	7.815	9.348	11.345	12.838
4	.207	.297	.484	.711	1.064	1.649	1.923	3.357	5.385	5.989	7.779	9.488	11.143	13.277	14.860
5	.412	.554	.831	1.145	1.610	2.343	2.675	4.351	6.626	7.289	9.236	11.071	12.833	15.086	16.750
6	.676	.872	1.237	1.635	2.204	3.070	3.455	5.348	7.841	8.558	10.645	12.592	14.449	16.812	18.548
7	.989	1.239	1.690	2.167	2.833	3.822	4.255	6.346	9.037	9.803	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	4.594	5.071	7.344	10.219	11.030	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	5.380	5.899	8.343	11.389	12.242	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	6.179	6.737	9.342	12.549	13.442	15.987	18.307	20.483	23.209	25.188
11	2.60	3.05	3.82	4.57	5.58	6.99	7.58	10.34	13.70	14.63	17.28	19.68	21.92	24.73	26.76
12	3.07	3.57	4.40	5.23	6.30	7.81	8.44	11.34	14.85	15.81	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	8.63	9.30	12.34	15.98	16.98	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	9.47	10.17	13.34	17.12	18.15	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	10.31	11.04	14.34	18.25	19.31	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	11.15	11.91	15.34	19.37	20.47	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.08	12.00	12.79	16.34	20.49	21.61	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	12.86	13.68	17.34	21.60	22.76	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	13.72	14.56	18.34	22.72	23.90	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	14.58	15.45	19.34	23.83	25.04	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	15.44	16.34	20.34	24.93	26.17	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	16.31	17.24	21.34	26.04	27.30	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	17.19	18.14	22.34	27.14	28.43	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	18.06	19.04	23.34	28.24	29.55	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	18.94	19.94	24.34	29.34	30.68	34.38	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	19.82	20.84	25.34	30.43	31.79	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	20.70	21.75	26.34	31.53	32.91	36.74	40.11	43.19	46.96	49.65
28	12.46	13.56	15.31	16.93	18.94	21.59	22.66	27.34	32.62	34.03	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	22.48	23.57	28.34	33.71	35.14	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	23.36	24.48	29.34	34.80	36.25	40.26	43.77	46.98	50.89	53.67
40	20.70	22.16	24.43	26.51	29.05	32.34	33.66	39.34	45.62	47.27	51.81	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	34.76	37.69	41.45	42.94	49.34	56.33	58.16	63.17	67.50	71.42	76.16	79.49
60	35.53	37.48	40.48	43.19	46.46	50.64	52.29	59.33	66.98	68.97	74.40	79.08	83.30	88.38	91.95
70	43.27	45.44	48.76	51.74	55.33	59.90	61.70	69.33	77.58	79.72	85.53	90.53	95.03	100.43	104.22
80	51.17	53.54	57.15	60.39	64.28	69.21	71.14	79.33	88.13	90.41	96.58	101.88	106.63	112.33	116.32
90	59.19	61.75	65.64	69.12	73.29	78.56	80.62	89.33	98.65	101.06	107.57	113.15	118.14	124.12	128.30
100	67.32	70.06	74.22	77.93	82.36	87.94	90.13	99.33	109.14	111.67	118.50	124.34	129.56	135.81	140.17

Table A9. Percentage points of the F-distribution

The table gives values of x_α for different degrees of freedom (ν_1,ν_2) such as to produce specified values $F(x_\alpha;\nu_1,\nu_2)$, where

$$F(x_\alpha;\nu_1,\nu_2) = \int_0^{x_\alpha} \frac{\Gamma\left(\frac{1}{2}(\nu_1,\nu_2)\right)}{\Gamma(\frac{1}{2}\nu_1)\Gamma(\frac{1}{2}\nu_2)} \left(\frac{\nu_1}{\nu_2}\right)^{\frac{1}{2}\nu_1} \frac{F^{\frac{1}{2}\nu_1-1}}{\left(1+\frac{\nu_1 F}{\nu_2}\right)^{\frac{1}{2}(\nu_1+\nu_2)}}\, dF = 1 - \alpha\ .$$

The table only has "upper tail" entries corresponding to values F = 0.90, 0.95, 0.975, 0.99, 0.995, 0.999, but it can be used to obtain "lower tail" percentage points corresponding to the complementary F = 0.10, 0.05, 0.025, 0.01, 0.005, 0.001, by means of the relation

$$F(x_{1-\alpha};\nu_2,\nu_1) = \frac{1}{F(x_\alpha;\nu_1,\nu_2)}\ .$$

ν_2 \ ν_1	F	1	2	3	4	5	6	7	8	9	10	12	15	20	30	40	60	120	∞
1	.90	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19	60.71	61.22	61.74	62.26	62.53	62.79	63.06	63.33
	.95	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	250.1	251.1	252.2	253.3	254.3
	.975	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.7	963.3	968.6	976.7	984.9	993.1	1001	1006	1010	1014	1018
	.99	4052	4999	5403	5625	5764	5859	5928	5982	6022	6056	6106	6157	6209	6261	6287	6313	6339	6366
	.995	16211	20000	21615	22500	23056	23437	23715	23925	24091	24224	24426	24630	24836	25044	25148	25253	25359	25465
	.999	4053*	5000*	5404*	5625*	5764*	5859*	5929*	5981*	6023*	6056*	6107*	6158*	6209*	6261*	6287*	6313*	6340*	6366*
2	.90	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.41	9.42	9.44	9.46	9.47	9.47	9.48	9.49
	.95	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.46	19.47	19.48	19.49	19.50
	.975	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.41	39.43	39.45	39.46	39.47	39.48	39.49	39.50
	.99	98.50	99.00	99.17	99.25	99.30	99.32	99.36	99.37	99.39	99.40	99.42	99.43	99.45	99.47	99.47	99.48	99.49	99.50
	.995	198.5	199.0	199.2	199.2	199.3	199.3	199.4	199.4	199.4	199.4	199.4	199.4	199.4	199.5	199.5	199.5	199.5	199.5
	.999	998.5	999.0	999.2	999.2	999.3	999.3	999.4	999.4	999.4	999.4	999.4	999.4	999.4	999.5	999.5	999.5	999.5	999.5
3	.90	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.20	5.18	5.17	5.16	5.15	5.14	5.13
	.95	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.62	8.59	8.57	8.55	8.53
	.975	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.34	14.25	14.17	14.08	14.04	13.99	13.95	13.90
	.99	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.50	26.41	26.32	26.22	26.13
	.995	55.55	49.80	47.47	46.19	45.39	44.84	44.43	44.13	43.88	43.69	43.39	43.08	42.78	42.47	42.31	42.15	41.99	41.83
	.999	167.0	148.5	141.1	137.1	134.6	132.8	131.6	130.6	129.9	129.2	128.3	127.4	126.4	125.4	125.0	124.5	124.0	123.5

* Multiply these numbers by 100.

Table A9. Percentage points of the F-distribution (continued)

ν_2	F	ν_1 = 1	2	3	4	5	6	7	8	9	10	12	15	20	30	40	60	120	∞
4	.90	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.90	3.87	3.84	3.82	3.80	3.79	3.78	3.76
	.95	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.75	5.72	5.69	5.66	5.63
	.975	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	9.84	8.75	8.66	8.56	8.46	8.41	8.36	8.31	8.26
	.99	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.84	13.75	13.65	13.56	13.46
	.995	31.33	26.28	24.26	23.15	22.46	21.97	21.62	21.35	21.14	20.97	20.70	20.44	20.17	19.89	19.75	19.61	19.47	19.32
	.999	74.14	61.25	56.18	53.44	51.71	50.53	49.66	49.00	48.47	48.05	47.41	46.76	46.10	45.43	45.09	44.75	44.40	44.05
5	.90	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.27	3.24	3.21	3.17	3.16	3.14	3.12	3.10
	.95	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.50	4.46	4.43	4.40	4.36
	.975	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52	6.43	6.33	6.23	6.18	6.12	6.07	6.02
	.99	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.38	9.29	9.20	9.11	9.02
	.995	22.78	18.31	16.53	15.56	14.94	14.51	14.20	13.96	13.77	13.62	13.38	13.15	12.90	12.64	12.53	12.40	12.27	12.14
	.999	47.18	37.12	33.20	31.09	29.75	28.84	28.16	27.64	27.24	26.92	26.42	25.91	25.39	24.87	24.60	24.33	24.06	23.79
6	.90	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.90	2.87	2.84	2.80	2.78	2.76	2.74	2.72
	.95	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.81	3.77	3.74	3.70	3.67
	.975	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37	5.27	5.17	5.07	5.01	4.96	4.90	4.85
	.99	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.23	7.14	7.06	6.97	6.88
	.995	18.63	14.54	12.92	12.03	11.46	11.07	10.79	10.57	10.39	10.25	10.03	9.81	9.59	9.36	9.24	9.12	9.00	8.88
	.999	35.51	27.00	23.70	21.92	20.81	20.03	19.46	19.03	18.69	18.41	17.99	17.56	17.12	16.67	16.44	16.21	15.99	15.75
7	.90	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.67	2.63	2.59	2.56	2.54	2.51	2.49	2.47
	.95	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.38	3.34	3.30	3.27	3.23
	.975	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.67	4.57	4.47	4.36	4.31	4.25	4.20	4.14
	.99	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	5.99	5.91	5.82	5.74	5.65
	.995	16.24	12.40	10.88	10.05	9.52	9.16	8.89	8.68	8.51	8.38	8.18	7.97	7.75	7.53	7.42	7.31	7.19	7.08
	.999	29.25	21.69	18.77	17.19	16.21	15.52	15.02	14.63	14.33	14.08	13.71	13.32	12.93	12.53	12.33	12.12	11.91	11.70
8	.90	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.50	2.46	2.42	2.38	2.36	2.34	2.32	2.29
	.95	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.08	3.04	3.01	2.97	2.93
	.975	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.20	4.10	4.00	3.89	3.84	3.78	3.73	3.67
	.99	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.20	5.12	5.03	4.95	4.86
	.995	14.69	11.04	9.60	8.81	8.30	7.95	7.69	7.50	7.34	7.21	7.01	6.81	6.61	6.40	6.29	6.18	6.06	5.95
	.999	25.42	18.49	15.83	14.39	13.49	12.86	12.40	12.04	11.77	11.54	11.19	10.84	10.48	10.11	9.92	9.73	9.53	9.33
9	.90	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.38	2.34	2.30	2.25	2.23	2.21	2.18	2.16
	.95	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.86	2.83	2.79	2.75	2.71
	.975	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.87	3.77	3.67	3.56	3.51	3.45	3.39	3.33
	.99	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.65	4.57	4.48	4.40	4.31
	.995	13.61	10.11	8.72	7.96	7.47	7.13	6.88	6.69	6.54	6.42	6.23	6.03	5.83	5.62	5.52	5.41	5.30	5.19
	.999	22.86	16.39	13.90	12.56	11.71	11.13	10.70	10.37	10.11	9.89	9.57	9.24	8.90	8.55	8.37	8.19	8.00	7.81
10	.90	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.28	2.24	2.20	2.16	2.13	2.11	2.08	2.06
	.95	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.70	2.66	2.62	2.58	2.54
	.975	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.62	3.52	3.42	3.31	3.26	3.20	3.14	3.08
	.99	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.25	4.17	4.08	4.00	3.91
	.995	12.83	9.43	8.08	7.34	6.87	6.54	6.30	6.12	5.97	5.85	5.66	5.47	5.27	5.07	4.97	4.86	4.75	4.64
	.999	21.04	14.91	12.55	11.28	10.48	9.92	9.52	9.20	8.96	8.75	8.45	8.13	7.80	7.47	7.30	7.12	6.94	6.76
11	.90	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.21	2.17	2.12	2.08	2.05	2.03	2.00	1.97
	.95	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.57	2.53	2.49	2.45	2.40
	.975	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.43	3.33	3.23	3.12	3.06	3.00	2.94	2.88
	.99	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	3.94	3.86	3.78	3.69	3.60
	.995	12.23	8.91	7.60	6.88	6.42	6.10	5.86	5.68	5.54	5.42	5.24	5.05	4.86	4.65	4.55	4.44	4.34	4.23
	.999	19.69	13.81	11.56	10.35	9.58	9.05	8.66	8.35	8.12	7.92	7.63	7.32	7.01	6.68	6.52	6.35	6.17	6.00

Table A9. Percentage points of the F-distribution (continued)

ν_2	F	ν_1 = 1	2	3	4	5	6	7	8	9	10	12	15	20	30	40	60	120	∞
12	.90	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.15	2.10	2.06	2.01	1.99	1.96	1.93	1.90
	.95	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.47	2.43	2.38	2.34	2.30
	.975	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.28	3.18	3.07	2.96	2.91	2.85	2.79	2.72
	.99	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.70	3.62	3.54	3.45	3.36
	.995	11.75	8.51	7.23	6.52	6.07	5.76	5.52	5.35	5.20	5.09	4.91	4.72	4.53	4.33	4.23	4.12	4.01	3.90
	.999	18.64	12.97	10.80	9.63	8.89	8.38	8.00	7.71	7.48	7.29	7.00	6.71	6.40	6.09	5.93	5.76	5.59	5.42
13	.90	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.10	2.05	2.01	1.96	1.93	1.90	1.88	1.85
	.95	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.38	2.34	2.30	2.25	2.21
	.975	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.15	3.05	2.95	2.84	2.78	2.72	2.66	2.60
	.99	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.51	3.43	3.34	3.25	3.17
	.995	11.37	8.19	6.93	6.23	5.79	5.48	5.25	5.08	4.94	4.82	4.64	4.46	4.27	4.07	3.97	3.87	3.76	3.65
	.999	17.81	12.31	10.21	9.07	8.35	7.86	7.49	7.21	6.98	6.80	6.52	6.23	5.93	5.63	5.47	5.30	5.14	4.97
14	.90	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.05	2.01	1.96	1.91	1.89	1.86	1.83	1.80
	.95	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.31	2.27	2.22	2.18	2.13
	.975	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	3.05	2.95	2.84	2.73	2.67	2.61	2.55	2.49
	.99	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.35	3.27	3.18	3.09	3.00
	.995	11.06	7.92	6.68	6.00	5.56	5.26	5.03	4.86	4.72	4.60	4.43	4.25	4.06	3.86	3.76	3.66	3.55	3.44
	.999	17.14	11.78	9.73	8.62	7.92	7.43	7.08	6.80	6.58	6.40	6.13	5.85	5.56	5.25	5.10	4.94	4.77	4.60
15	.90	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	2.02	1.97	1.92	1.87	1.85	1.82	1.79	1.76
	.95	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.25	2.20	2.16	2.11	2.07
	.975	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.96	2.86	2.76	2.64	2.59	2.52	2.46	2.40
	.99	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.21	3.13	3.05	2.96	2.87
	.995	10.80	7.70	6.48	5.80	5.37	5.07	4.85	4.67	4.54	4.42	4.25	4.07	3.88	3.69	3.58	3.48	3.37	3.26
	.999	16.59	11.34	9.34	8.25	7.57	7.09	6.74	6.47	6.26	6.08	5.81	5.54	5.25	4.95	4.80	4.64	4.47	4.31
16	.90	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	1.99	1.94	1.89	1.84	1.81	1.78	1.75	1.72
	.95	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.19	2.15	2.11	2.06	2.01
	.975	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.89	2.79	2.68	2.57	2.51	2.45	2.38	2.32
	.99	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.10	3.02	2.93	2.84	2.75
	.995	10.58	7.51	6.30	5.64	5.21	4.91	4.69	4.52	4.38	4.27	4.10	3.92	3.73	3.54	3.44	3.33	3.22	3.11
	.999	16.12	10.97	9.00	7.94	7.27	6.81	6.46	6.19	5.98	5.81	5.55	5.27	4.99	4.70	4.54	4.39	4.23	4.06
17	.90	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.96	1.91	1.86	1.81	1.78	1.75	1.72	1.69
	.95	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.15	2.10	2.06	2.01	1.96
	.975	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.82	2.72	2.62	2.50	2.44	2.38	2.32	2.25
	.99	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.00	2.92	2.83	2.75	2.65
	.995	10.38	7.35	6.16	5.50	5.07	4.78	4.56	4.39	4.25	4.14	3.97	3.79	3.61	3.41	3.31	3.21	3.10	2.98
	.999	15.72	10.66	8.73	7.68	7.02	7.56	6.22	5.96	5.75	5.58	5.32	5.05	4.78	4.48	4.33	4.18	4.02	3.85
18	.90	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.93	1.89	1.84	1.78	1.75	1.72	1.69	1.66
	.95	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.11	2.06	2.02	1.97	1.92
	.975	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.77	2.67	2.56	2.44	2.38	2.32	2.26	2.19
	.99	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	2.92	2.84	2.75	2.66	2.57
	.995	10.22	7.21	6.03	5.37	4.96	4.66	4.44	4.28	4.14	4.03	3.86	3.68	3.50	3.30	3.20	3.10	2.99	2.87
	.999	15.38	10.39	8.49	7.46	6.81	6.35	6.02	5.76	5.56	5.39	5.13	4.87	4.59	4.30	4.15	4.00	3.84	3.67
19	.90	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.91	1.86	1.81	1.76	1.73	1.70	1.67	1.63
	.95	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.07	2.03	1.98	1.93	1.88
	.975	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.72	2.62	2.51	2.39	2.33	2.27	2.20	2.13
	.99	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.84	2.76	2.67	2.58	2.49
	.995	10.07	7.09	5.92	5.27	4.85	4.56	4.34	4.18	4.04	3.93	3.76	3.59	3.40	3.21	3.11	3.00	2.89	2.78
	.999	15.08	10.16	8.28	7.26	6.62	6.18	5.85	5.59	5.39	5.22	4.97	4.70	4.43	4.14	3.99	3.84	3.68	3.51

Table A9. Percentage points of the F-distribution (continued)

ν_2	F	ν_1 = 1	2	3	4	5	6	7	8	9	10	12	15	20	30	40	60	120	∞
20	.90	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.89	1.84	1.79	1.74	1.71	1.68	1.64	1.61
	.95	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.04	1.99	1.95	1.90	1.84
	.975	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.68	2.57	2.46	2.35	2.29	2.22	2.16	2.09
	.99	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.78	2.69	2.61	2.52	2.42
	.995	9.94	6.99	5.82	5.17	4.76	4.47	4.26	4.09	3.96	3.85	3.68	3.50	3.32	3.12	3.02	2.92	2.81	2.69
	.999	14.82	9.95	8.10	7.10	6.46	6.02	5.69	5.44	5.24	5.08	4.82	4.56	4.29	4.00	3.86	3.70	3.54	3.38
21	.90	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92	1.87	1.83	1.78	1.72	1.69	1.66	1.62	1.59
	.95	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.01	1.96	1.92	1.87	1.81
	.975	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.64	2.53	2.42	2.31	2.25	2.18	2.11	2.04
	.99	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.72	2.64	2.55	2.46	2.36
	.995	9.83	6.89	5.73	5.09	4.68	4.39	4.18	4.01	3.88	3.77	3.60	3.43	3.24	3.05	2.95	2.84	2.73	2.61
	.999	14.59	9.77	7.94	6.95	6.32	5.88	5.56	5.31	5.11	4.95	4.70	4.44	4.17	3.88	3.74	3.58	3.42	3.26
22	.90	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.86	1.81	1.76	1.70	1.67	1.64	1.60	1.57
	.95	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	1.98	1.94	1.89	1.84	1.78
	.975	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.60	2.50	2.39	2.27	2.21	2.14	2.08	2.00
	.99	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.67	2.58	2.50	2.40	2.31
	.995	9.73	6.81	5.65	5.02	4.61	4.32	4.11	3.94	3.81	3.70	3.54	3.36	3.18	2.98	2.88	2.77	2.66	2.55
	.999	14.38	9.61	7.80	6.81	6.19	5.76	5.44	5.19	4.99	4.83	4.58	4.33	4.06	3.78	3.63	3.48	3.32	3.15
23	.90	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89	1.84	1.80	1.74	1.69	1.66	1.62	1.59	1.55
	.95	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	1.96	1.91	1.86	1.81	1.76
	.975	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	2.57	2.47	2.36	2.24	2.18	2.11	2.04	1.97
	.99	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.62	2.54	2.45	2.35	2.26
	.995	9.63	6.73	5.58	4.95	4.54	4.26	4.05	3.88	3.75	3.64	3.47	3.30	3.12	2.92	2.82	2.71	2.60	2.48
	.999	14.19	9.47	7.67	6.69	6.08	5.65	5.33	5.09	4.89	4.73	4.48	4.23	3.96	3.68	3.53	3.38	3.22	3.05
24	.90	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.83	1.78	1.73	1.67	1.64	1.61	1.57	1.53
	.95	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.94	1.89	1.84	1.79	1.73
	.975	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.54	2.44	2.33	2.21	2.15	2.08	2.01	1.94
	.99	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.58	2.49	2.40	2.31	2.21
	.995	9.55	6.66	5.52	4.89	4.49	4.20	3.99	3.83	3.69	3.59	3.42	3.25	3.06	2.87	2.77	2.66	2.55	2.43
	.999	14.03	9.34	7.55	6.59	5.98	5.55	5.23	4.99	4.80	4.64	4.39	4.14	3.87	3.59	3.45	3.29	3.14	2.97
25	.90	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87	1.82	1.77	1.72	1.66	1.63	1.59	1.56	1.52
	.95	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.92	1.87	1.82	1.77	1.71
	.975	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.51	2.41	2.30	2.18	2.12	2.05	1.98	1.91
	.99	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.54	2.45	2.36	2.27	2.17
	.995	9.48	6.60	5.46	4.84	4.43	4.15	3.94	3.78	3.64	3.54	3.37	3.20	3.01	2.82	2.72	2.61	2.50	2.38
	.999	13.88	9.22	7.45	6.49	5.88	5.46	5.15	4.91	4.71	4.56	4.31	4.06	3.79	3.52	3.37	3.22	3.06	2.89
26	.90	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86	1.81	1.76	1.71	1.65	1.61	1.58	1.54	1.50
	.95	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.90	1.85	1.80	1.75	1.69
	.975	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	2.49	2.39	2.28	2.16	2.09	2.03	1.95	1.88
	.99	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.81	2.66	2.50	2.42	2.33	2.23	2.13
	.995	9.41	6.54	5.41	4.79	4.38	4.10	3.89	3.73	3.60	3.49	3.33	3.15	2.97	2.77	2.67	2.56	2.45	2.33
	.999	13.74	9.12	7.36	6.41	5.80	5.38	5.07	4.83	4.64	4.48	4.24	3.99	3.72	3.44	3.30	3.15	2.99	2.82
27	.90	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85	1.80	1.75	1.70	1.64	1.60	1.57	1.53	1.49
	.95	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.88	1.84	1.79	1.73	1.67
	.975	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57	2.47	2.36	2.25	2.13	2.07	2.00	1.93	1.85
	.99	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.47	2.38	2.29	2.20	2.10
	.995	9.34	6.49	5.36	4.74	4.34	4.06	3.85	3.69	3.56	3.45	3.28	3.11	2.93	2.73	2.63	2.52	2.41	2.29
	.999	13.61	9.02	7.27	6.33	5.73	5.31	5.00	4.76	4.57	4.41	4.17	3.92	3.66	3.38	3.23	3.08	2.92	2.75

Table A9. Percentage points of the F-distribution (continued)

ν_2	F	ν_1 = 1	2	3	4	5	6	7	8	9	10	12	15	20	30	40	60	120	∞
28	.90	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84	1.79	1.74	1.69	1.63	1.59	1.56	1.52	1.48
	.95	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.87	1.82	1.77	1.71	1.65
	.975	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	2.45	2.34	2.23	2.11	2.05	1.98	1.91	1.83
	.99	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.44	2.35	2.26	2.17	2.06
	.995	9.28	6.44	5.32	4.70	4.30	4.02	3.81	3.65	3.52	3.41	3.25	3.07	2.89	2.69	2.59	2.48	2.37	2.25
	.999	13.50	8.93	7.19	6.25	5.66	5.24	4.93	4.69	4.50	4.35	4.11	3.86	3.60	3.32	3.18	3.02	2.86	2.69
29	.90	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83	1.78	1.73	1.68	1.62	1.58	1.55	1.51	1.47
	.95	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.85	1.81	1.75	1.70	1.64
	.975	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53	2.43	2.32	2.21	2.09	2.03	1.96	1.89	1.81
	.99	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.73	2.57	2.41	2.33	2.23	2.14	2.03
	.995	9.23	6.40	5.28	4.66	4.24	3.98	3.77	3.61	3.48	3.38	3.21	3.04	2.86	2.66	2.56	2.45	2.33	2.21
	.999	13.39	8.85	7.12	6.19	5.59	5.18	4.87	4.64	4.45	4.29	4.05	3.80	3.54	3.27	3.12	2.97	2.81	2.64
30	.90	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.77	1.72	1.67	1.61	1.57	1.54	1.50	1.46
	.95	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.84	1.79	1.74	1.68	1.62
	.975	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.41	2.31	2.20	2.07	2.01	1.94	1.87	1.79
	.99	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.39	2.30	2.21	2.11	2.01
	.995	9.18	6.35	5.24	4.62	4.23	3.95	3.74	3.58	3.45	3.34	3.18	3.01	2.82	2.63	2.52	2.42	2.30	2.18
	.999	13.29	8.77	7.05	6.12	5.53	5.12	4.82	4.58	4.39	4.24	4.00	3.75	3.49	3.22	3.07	2.92	2.76	2.59
40	.90	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.71	1.66	1.61	1.54	1.51	1.47	1.42	1.38
	.95	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.74	1.69	1.64	1.58	1.51
	.975	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.29	2.18	2.07	1.94	1.88	1.80	1.72	1.64
	.99	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.20	2.11	2.02	1.92	1.80
	.995	8.83	6.07	4.98	4.37	3.99	3.71	3.51	3.35	3.22	3.12	2.95	2.78	2.60	2.40	2.30	2.18	2.06	1.93
	.999	12.61	8.25	6.60	5.70	5.13	4.73	4.44	4.21	4.02	3.87	3.64	3.40	3.15	2.87	2.73	2.57	2.41	2.23
60	.90	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.66	1.60	1.54	1.48	1.44	1.40	1.35	1.29
	.95	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.65	1.59	1.53	1.47	1.39
	.975	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.17	2.06	1.94	1.82	1.74	1.67	1.58	1.48
	.99	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.03	1.94	1.84	1.73	1.60
	.995	8.49	5.79	4.73	4.14	3.76	3.49	3.29	3.13	3.01	2.90	2.74	2.57	2.39	2.19	2.08	1.96	1.83	1.69
	.999	11.97	7.76	6.17	5.31	4.76	4.37	4.09	3.87	3.69	3.54	3.31	3.08	2.83	2.55	2.41	2.25	2.08	1.89
120	.90	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65	1.60	1.55	1.48	1.41	1.37	1.32	1.26	1.19
	.95	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.55	1.50	1.43	1.35	1.25
	.975	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.05	1.94	1.82	1.69	1.61	1.53	1.43	1.31
	.99	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.86	1.76	1.66	1.53	1.38
	.995	8.18	5.54	4.50	3.92	3.55	3.28	3.09	2.93	2.81	2.71	2.54	2.37	2.19	1.98	1.87	1.75	1.61	1.43
	.999	11.38	7.32	5.79	4.95	4.42	4.04	3.77	3.55	3.38	3.24	3.02	2.78	2.53	2.26	2.11	1.95	1.76	1.54
∞	.90	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60	1.55	1.49	1.42	1.34	1.30	1.24	1.17	1.00
	.95	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.46	1.39	1.32	1.22	1.00
	.975	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.94	1.83	1.71	1.57	1.48	1.39	1.27	1.00
	.99	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.70	1.59	1.47	1.32	1.00
	.995	7.88	5.30	4.28	3.72	3.35	3.09	2.90	2.74	2.62	2.52	2.36	2.19	2.00	1.79	1.67	1.53	1.36	1.00
	.999	10.83	6.91	5.42	4.62	4.10	3.74	3.47	3.27	3.10	2.96	2.74	2.51	2.27	1.99	1.84	1.66	1.45	1.00

Table A10. Percentage points of the Kolmogorov-Smirnov statistic*

The table gives values d_α for specified values of α such that

$$P(D_n \leq d_\alpha) \simeq 1 - \alpha,$$

where the Kolmogorov-Smirnov test statistic D_n for the sample of size n is the largest deviation between the observed cumulative distribution and the theoretical cumulative distribution.

n \ α	.20	.10	.05	.02	.01
1	.9000	.9500	.9750	.9900	.9950
2	.6838	.7764	.8419	.9000	.9293
3	.5648	.6360	.7076	.7846	.8290
4	.4927	.5652	.6239	.6889	.7342
5	.4470	.5095	.5633	.6272	.6685
6	.4104	.4680	.5193	.5774	.6166
7	.3815	.4361	.4834	.5384	.5758
8	.3583	.4096	.4543	.5065	.5418
9	.3391	.3875	.4300	.4796	.5133
10	.3226	.3687	.4093	.4566	.4889
11	.3083	.3524	.3912	.4367	.4677
12	.2958	.3382	.3754	.4192	.4491
13	.2847	.3255	.3614	.4036	.4325
14	.2748	.3142	.3489	.3897	.4176
15	.2659	.3040	.3376	.3771	.4042
16	.2578	.2947	.3273	.3657	.3920
17	.2504	.2863	.3180	.3553	.3809
18	.2436	.2785	.3094	.3457	.3706
19	.2374	.2714	.3014	.3369	.3612
20	.2316	.2647	.2941	.3287	.3524
21	.2262	.2586	.2872	.3210	.3443
22	.2212	.2528	.2809	.3139	.3367
23	.2165	.2475	.2749	.3073	.3295
24	.2121	.2424	.2693	.3010	.3229
25	.2079	.2377	.2640	.2952	.3166
26	.2040	.2332	.2591	.2896	.3106
27	.2003	.2290	.2544	.2844	.3050
28	.1968	.2250	.2499	.2794	.2997
29	.1935	.2212	.2457	.2747	.2947
30	.1903	.2176	.2417	.2702	.2899
35	.1766	.2019	.2243	.2507	.2690
40	.1655	.1891	.2101	.2349	.2521
45	.1562	.1786	.1984	.2218	.2380
50	.1484	.1696	.1884	.2107	.2260
55	.1416	.1619	.1798	.2011	.2157
60	.1357	.1551	.1723	.1927	.2067
65	.1305	.1491	.1657	.1853	.1988
70	.1259	.1438	.1598	.1786	.1917
75	.1217	.1390	.1544	.1727	.1853
80	.1179	.1347	.1496	.1673	.1795
85	.1144	.1307	.1452	.1624	.1742
90	.1113	.1271	.1412	.1579	.1694
95	.1083	.1238	.1375	.1537	.1649
100	.1056	.1207	.1340	.1499	.1608
≥ 100	$\frac{1.07}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.52}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$

* Abridged and adapted from Table 15.1 in Donald B. Owen: *Handbook of Statistical Tables*, 1962, Addison-Wesley Publishing Company, Inc., Reading, Mass., by permission of the publisher.

Table A11. Critical values of the run statistic

The table gives critical values r_α of the run statistic r with probability distribution p(r) defined in Sect.14.6.2 for sample sizes n,m up to 15 ($n \leq m$). The table assumes a one-sided test of significance α with critical region in the lower tail of the probability distribution,

$$\sum_{r=2}^{r_\alpha} p(r) \leq \alpha < \sum_{r=2}^{r_\alpha+1} p(r).$$

$\alpha = 0.005$

m \ n	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2	-													
3	-	-												
4	-	-	-											
5	-	-	-	-										
6	-	-	-	2	2									
7	-	-	-	2	2	3								
8	-	-	2	2	3	3	3							
9	-	-	2	2	3	3	3	4						
10	-	-	2	3	3	3	4	4	5					
11	-	-	2	3	3	4	4	5	5	5				
12	-	2	2	3	3	4	4	5	5	6	6			
13	-	2	2	3	3	4	5	5	5	6	6	7		
14	-	2	2	3	4	4	5	5	6	6	7	7	7	
15	-	2	3	3	4	4	5	6	6	7	7	7	8	8

$\alpha = 0.010$

m \ n	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2	-													
3	-	-												
4	-	-	-											
5	-	-	-	2										
6	-	-	2	2	2									
7	-	-	2	2	3	3								
8	-	-	2	2	3	3	4							
9	-	2	2	3	3	4	4	4						
10	-	2	2	3	3	4	4	5	5					
11	-	2	2	3	4	4	5	5	5	6				
12	-	2	3	3	4	4	5	5	6	6	7			
13	-	2	3	3	4	5	5	6	6	6	7	7		
14	-	2	3	3	4	5	5	6	6	7	7	8	8	
15	-	2	3	4	4	5	5	6	7	7	8	8	8	9

$\alpha = 0.025$

m \ n	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2	-													
3	-	-												
4	-	-	-											
5	-	-	2	2										
6	-	2	2	3	3									
7	-	2	2	3	3	3								
8	-	2	3	3	3	4	4							
9	-	2	3	3	4	4	5	5						
10	-	2	3	3	4	5	5	5	6					
11	-	2	3	4	4	5	5	6	6	7				
12	2	2	3	4	4	5	6	6	7	7	7			
13	2	2	3	4	5	5	6	6	7	7	8	8		
14	2	2	3	4	5	5	6	7	7	8	8	9	9	
15	2	3	3	4	5	6	6	7	7	8	8	9	9	10

$\alpha = 0.05$

m \ n	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2	-													
3	-	-												
4	-	-	2											
5	-	2	2	3										
6	-	2	3	3	3									
7	-	2	3	3	4	4								
8	2	2	3	3	4	4	5							
9	2	2	3	4	4	5	5	6						
10	2	3	3	4	5	5	6	6	6					
11	2	3	3	4	5	5	6	6	7	7				
12	2	3	4	4	5	6	6	7	7	8	8			
13	2	3	4	4	5	6	6	7	8	8	9	9		
14	2	3	4	5	5	6	7	7	8	8	9	9	10	
15	2	3	4	5	6	6	7	8	8	9	9	10	10	11

Table A12. Critical values of the Wilcoxon rank sum statistic*

The table gives critical values of the Wilcoxon two-sample test statistic W defined in Sect.14.6.9 for sample sizes n,m up to 25 ($n \le m$). The table assumes a one-sided test of significance α as quoted in boldface at the head of the columns, and critical region in the lower tail of the probability distribution. For a two-sided test of significance α, the lower critical value W_ℓ is found as the table entry for $\frac{1}{2}\alpha$, and the upper critical value is deduced as $W_u = 2\bar{W} - W_\ell$, where $2\bar{W}$ is also given in the table.

n = 1

m	0·001	0·005	0·010	0·025	0·05	0·10	$2\bar{W}$
2							4
3							5
4							6
5							7
6							8
7							9
8						—	10
9						1	11
10						1	12
11						1	13
12						1	14
13						1	15
14						1	16
15						1	17
16						1	18
17						1	19
18					—	1	20
19					1	2	21
20					1	2	22
21					1	2	23
22					1	2	24
23					1	2	25
24					1	2	26
25	—	—	—	—	1	2	27

n = 2

m	0·001	0·005	0·010	0·025	0·05	0·10	$2\bar{W}$
2						—	10
3						3	12
4					—	3	14
5					3	4	16
6					3	4	18
7				—	3	4	20
8				3	4	5	22
9				3	4	5	24
10				3	4	6	26
11				3	4	6	28
12			—	4	5	7	30
13			3	4	5	7	32
14			3	4	6	8	34
15			3	4	6	8	36
16			3	4	6	8	38
17			3	5	6	9	40
18		—	3	5	7	9	42
19		3	4	5	7	10	44
20		3	4	5	7	10	46
21		3	4	6	8	11	48
22		3	4	6	8	11	50
23		3	4	6	8	12	52
24		3	4	6	9	12	54
25	—	3	4	6	9	12	56

n = 3

m	0·001	0·005	0·010	0·025	0·05	0·10	$2\bar{W}$
3					6	7	21
4				—	6	7	24
5				6	7	8	27
6			—	7	8	9	30
7			6	7	8	10	33
8		—	6	8	9	11	36
9		6	7	8	10	11	39
10		6	7	9	10	12	42
11		6	7	9	11	13	45
12		7	8	10	11	14	48
13		7	8	10	12	15	51
14		7	8	11	13	16	54
15		8	9	11	13	16	57
16	—	8	9	12	14	17	60
17	6	8	10	12	15	18	63
18	6	8	10	13	15	19	66
19	6	9	10	13	16	20	69
20	6	9	11	14	17	21	72
21	7	9	11	14	17	21	75
22	7	10	12	15	18	22	78
23	7	10	12	15	19	23	81
24	7	10	12	16	19	24	84
25	7	11	13	16	20	25	87

n = 4

m	0·001	0·005	0·010	0·025	0·05	0·10	$2\bar{W}$
4				—	10	11	36
5				10	11	12	40
6			10	11	12	13	44
7			10	11	13	14	48
8			11	12	14	15	52
9		—	11	13	14	16	56
10		10	12	13	15	17	60
11		10	12	14	16	18	64
12		10	13	15	17	19	68
13		11	13	15	18	20	72
14		11	14	16	19	21	76
15		11	15	17	20	22	80
16		12	15	17	21	24	84
17		12	16	18	21	25	88
18		13	16	19	22	26	92
19		13	17	19	23	27	96
20		13	18	20	24	28	100
21		14	18	21	25	29	104
22		14	19	21	26	30	108
23		14	19	22	27	31	112
24		15	20	23	27	32	116
25		15	20	23	28	33	120

* Reproduced with changes in notation from Table 1 in L.R. Verdooren: "Extended tables of critical values for Wilcoxon's test statistic", *Biometrika*, 50 (1963) 177–186, by permission of the publisher.

Table A12. Critical values of the Wilcoxon rank sum statistic (continued)

n = 5

m	0·001	0·005	0·010	0·025	0·05	0·10	$2\bar{W}$
5		15	16	17	19	20	55
6		16	17	18	20	22	60
7	—	16	18	20	21	23	65
8	15	17	19	21	23	25	70
9	16	18	20	22	24	27	75
10	16	19	21	23	26	28	80
11	17	20	22	24	27	30	85
12	17	21	23	26	28	32	90
13	18	22	24	27	30	33	95
14	18	22	25	28	31	35	100
15	19	23	26	29	33	37	105
16	20	24	27	30	34	38	110
17	20	25	28	32	35	40	115
18	21	26	29	33	37	42	120
19	22	27	30	34	38	43	125
20	22	28	31	35	40	45	130
21	23	29	32	37	41	47	135
22	23	29	33	38	43	48	140
23	24	30	34	39	44	50	145
24	25	31	35	40	45	51	150
25	25	32	36	42	47	53	156

n = 6

m	0·001	0·005	0·010	0·025	0·05	0·10	$2\bar{W}$
6		—	23	24	26	28	30	78
7		21	24	25	27	29	32	84
8		22	25	27	29	31	34	90
9		23	26	28	31	33	36	96
10		24	27	29	32	35	38	102
11		25	28	30	34	37	40	108
12		25	30	32	35	38	42	114
13		26	31	33	37	40	44	120
14		27	32	34	38	42	46	126
15		28	33	36	40	44	48	132
16		29	34	37	42	46	50	138
17		30	36	39	43	47	52	144
18		31	37	40	45	49	55	150
19		32	38	41	46	51	57	156
20		33	39	43	48	53	59	162
21		33	40	44	50	55	61	168
22		34	42	45	51	57	63	174
23		35	43	47	53	58	65	180
24		36	44	48	54	60	67	186
25		37	45	50	56	62	69	192

n = 7

m	0·001	0·005	0·010	0·025	0·05	0·10	$2\bar{W}$
7	29	32	34	36	39	41	105
8	30	34	35	38	41	44	112
9	31	35	37	40	43	46	119
10	33	37	39	42	45	49	126
11	34	38	40	44	47	51	133
12	35	40	42	46	49	54	140
13	36	41	44	48	52	56	147
14	37	43	45	50	54	59	154
15	38	44	47	52	56	61	161
16	39	46	49	54	58	64	168
17	41	47	51	56	61	66	175
18	42	49	52	58	63	69	182
19	43	50	54	60	65	71	189
20	44	52	56	62	67	74	196
21	46	53	58	64	69	76	203
22	47	55	59	66	72	79	210
23	48	57	61	68	74	81	217
24	49	58	63	70	76	84	224
25	50	60	64	72	78	86	231

n = 8

m	0·001	0·005	0·010	0·025	0·05	0·10	$2\bar{W}$
8	40	43	45	49	51	55	136
9	41	45	47	51	54	58	144
10	42	47	49	53	56	60	152
11	44	49	51	55	59	63	160
12	45	51	53	58	62	66	168
13	47	53	56	60	64	69	176
14	48	54	58	62	67	72	184
15	50	56	60	65	69	75	192
16	51	58	62	67	72	78	200
17	53	60	64	70	75	81	208
18	54	62	66	72	77	84	216
19	56	64	68	74	80	87	224
20	57	66	70	77	83	90	232
21	59	68	72	79	85	92	240
22	60	70	74	81	88	95	248
23	62	71	76	84	90	98	256
24	64	73	78	86	93	101	264
25	65	75	81	89	96	104	272

n = 9

m	0·001	0·005	0·010	0·025	0·05	0·10	$2\bar{W}$
9	52	56	59	62	66	70	171
10	53	58	61	65	69	73	180
11	55	61	63	68	72	76	189
12	57	63	66	71	75	80	198
13	59	65	68	73	78	83	207
14	60	67	71	76	81	86	216
15	62	69	73	79	84	90	225
16	64	72	76	82	87	93	234
17	66	74	78	84	90	97	243
18	68	76	81	87	93	100	252
19	70	78	83	90	96	103	261
20	71	81	85	93	99	107	270
21	73	83	88	95	102	110	279
22	75	85	90	98	105	113	288
23	77	88	93	101	108	117	297
24	79	90	95	104	111	120	306
25	81	92	98	107	114	123	315

n = 10

m	0·001	0·005	0·010	0·025	0·05	0·10	$2\bar{W}$
10	65	71	74	78	82	87	210
11	67	73	77	81	86	91	220
12	69	76	79	84	89	94	230
13	72	79	82	88	92	98	240
14	74	81	85	91	96	102	250
15	76	84	88	94	99	106	260
16	78	86	91	97	103	109	270
17	80	89	93	100	106	113	280
18	82	92	96	103	110	117	290
19	84	94	99	107	113	121	300
20	87	97	102	110	117	125	310
21	89	99	105	113	120	128	320
22	91	102	108	116	123	132	330
23	93	105	110	119	127	136	340
24	95	107	113	122	130	140	350
25	98	110	116	126	134	144	360

Table A12. Critical values of the Wilcoxon rank sum statistic (continued)

	n = 11							n = 12							
m	0.001	0.005	0.010	0.025	0.05	0.10	$2W$	0.001	0.005	0.010	0.025	0.05	0.10	$2W$	m
11	81	87	91	96	100	106	253								
12	83	90	94	99	104	110	264	98	105	109	115	120	127	300	12
13	86	93	97	103	108	114	275	101	109	113	119	125	131	312	13
14	88	96	100	106	112	118	286	103	112	116	123	129	136	324	14
15	90	99	103	110	116	123	297	106	115	120	127	133	141	336	15
16	93	102	107	113	120	127	308	109	119	124	131	138	145	348	16
17	95	105	110	117	123	131	319	112	122	127	135	142	150	360	17
18	98	108	113	121	127	135	330	115	125	131	139	146	155	372	18
19	100	111	116	124	131	139	341	118	129	134	143	150	159	384	19
20	103	114	119	128	135	144	352	120	132	138	147	155	164	396	20
21	106	117	123	131	139	148	363	123	136	142	151	159	169	408	21
22	108	120	126	135	143	152	374	126	139	145	155	163	173	420	22
23	111	123	129	139	147	156	385	129	142	149	159	168	178	432	23
24	113	126	132	142	151	161	396	132	146	153	163	172	183	444	24
25	116	129	136	146	155	165	407	135	149	156	167	176	187	456	25

	n = 13							n = 14							
m	0.001	0.005	0.010	0.025	0.05	0.10	$2W$	0.001	0.005	0.010	0.025	0.05	0.10	$2W$	m
13	117	125	130	136	142	149	351								
14	120	129	134	141	147	154	364	137	147	152	160	166	174	406	14
15	123	133	138	145	152	159	377	141	151	156	164	171	179	420	15
16	126	136	142	150	156	165	390	144	155	161	169	176	185	434	16
17	129	140	146	154	161	170	403	148	159	165	174	182	190	448	17
18	133	144	150	158	166	175	416	151	163	170	179	187	196	462	18
19	136	148	154	163	171	180	429	155	168	174	183	192	202	476	19
20	139	151	158	167	175	185	442	159	172	178	188	197	207	490	20
21	142	155	162	171	180	190	455	162	176	183	193	202	213	504	21
22	145	159	166	176	185	195	468	166	180	187	198	207	218	518	22
23	149	163	170	180	189	200	481	169	184	192	203	212	224	532	23
24	152	166	174	185	194	205	494	173	188	196	207	218	229	546	24
25	155	170	178	189	199	211	507	177	192	200	212	223	235	560	25

	n = 15							n = 16							
m	0.001	0.005	0.010	0.025	0.05	0.10	$2W$	0.001	0.005	0.010	0.025	0.05	0.10	$2W$	m
15	160	171	176	184	192	200	465								
16	163	175	181	190	197	206	480	184	196	202	211	219	229	528	16
17	167	180	186	195	203	212	495	188	201	207	217	225	235	544	17
18	171	184	190	200	208	218	510	192	206	212	222	231	242	560	18
19	175	189	195	205	214	224	525	196	210	218	228	237	248	576	19
20	179	193	200	210	220	230	540	201	215	223	234	243	255	592	20
21	183	198	205	216	225	236	555	205	220	228	239	249	261	608	21
22	187	202	210	221	231	242	570	209	225	233	245	255	267	624	22
23	191	207	214	226	236	248	585	214	230	238	251	261	274	640	23
24	195	211	219	231	242	254	600	218	235	244	256	267	280	656	24
25	199	216	224	237	248	260	615	222	240	249	262	273	287	672	25

	n = 17							n = 18							
m	0.001	0.005	0.010	0.025	0.05	0.10	$2W$	0.001	0.005	0.010	0.025	0.05	0.10	$2W$	m
17	210	223	230	240	249	259	595								
18	214	228	235	246	255	266	612	237	252	259	270	280	291	666	18
19	219	234	241	252	262	273	629	242	258	265	277	287	299	684	19
20	223	239	246	258	268	280	646	247	263	271	283	294	306	702	20
21	228	244	252	264	274	287	663	252	269	277	290	301	313	720	21
22	233	249	258	270	281	294	680	257	275	283	296	307	321	738	22
23	238	255	263	276	287	300	697	262	280	289	303	314	328	756	23
24	242	260	269	282	294	307	714	267	286	295	309	321	335	774	24
25	247	266	275	288	300	314	731	273	292	301	316	328	343	792	25

Table A12. Critical values of the Wilcoxon rank sum statistic (continued)

	n = 19							n = 20							
m	0.001	0.005	0.010	0.025	0.05	0.10	$2W$	0.001	0.005	0.010	0.025	0.05	0.10	$2W$	m
19	267	283	291	303	313	325	741								
20	272	289	297	309	320	333	760	298	315	324	337	348	361	820	20
21	277	295	303	316	328	341	779	304	322	331	344	356	370	840	21
22	283	301	310	323	335	349	798	309	328	337	351	364	378	860	22
23	288	307	316	330	342	357	817	315	335	344	359	371	386	880	23
24	294	313	323	337	350	364	836	321	341	351	366	379	394	900	24
25	299	319	329	344	357	372	855	327	348	358	373	387	403	920	25

	n = 21							n = 22							
m	0.001	0.005	0.010	0.025	0.05	0.10	$2W$	0.001	0.005	0.010	0.025	0.05	0.10	$2W$	m
21	331	349	359	373	385	399	903								
22	337	356	366	381	393	408	924	365	386	396	411	424	439	990	22
23	343	363	373	388	401	417	945	372	393	403	419	432	448	1012	23
24	349	370	381	396	410	426	966	379	400	411	427	441	457	1034	24
25	356	377	388	404	418	434	987	386	408	419	435	450	467	1056	25

	n = 23							n = 24							
m	0.001	0.005	0.010	0.025	0.05	0.10	$2W$	0.001	0.005	0.010	0.025	0.05	0.10	$2W$	m
23	402	424	434	451	465	481	1081								
24	409	431	443	459	474	491	1104	440	464	475	492	507	525	1176	24
25	416	439	451	468	483	500	1127	448	472	484	501	517	535	1200	25

	n = 25						
m	0.001	0.005	0.010	0.025	0.05	0.10	$2W$
25	480	505	517	536	552	570	1275

Bibliography

General books on probability and statistics

BRADLEY, J.V., *Distribution-Free Statistical Tests*,
Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1968.

BREIMAN, L., *Statistics With a View Toward Applications*,
Houghton Mifflin Company, Boston, 1973.

CRAMÉR, H., *Mathematical Methods of Statistics*,
Princeton University Press, Princeton, 1966.

FISHER, R.A., *Statistical Methods for Research Workers*, Thirteenth Edition,
Oliver and Boyd, Edinburgh - London, 1958.

HOGG, R.V. and CRAIG, A.T.,
Introduction to Mathematical Statistics, Second Edition,
The Macmillan Company, New York,
Collier - Macmillan Limited, London, 1967.

KENDALL, M.G. and STUART, A.,
The Advanced Theory of Statistics, In Three Volumes,
Vol. 1 Distribution Theory Second Edition 1963;
Vol. 2 Inference and Relationship Second Edition 1967;
Vol. 3 Design and Analysis, and Time-Series 1966;
Charles Griffin & Company Limited, London.

LEHMAN, E.L. *Nonparametrics : Statistical Methods Based on Ranks*,
Holden-Day, Inc., San Francisco, 1975.

LINDLEY, D.V. *Introduction to Probability and Statistics*, In Two Volumes,
Part 1 Probability, Part 2 Inference,
Cambridge University Press, London, 1965.

MOOD, A.M., GRAYBILL, F.A., and BOES, D.C.,
Introduction to the Theory of Statistics, Third Edition,
McGraw-Hill Book Company, Inc., New York, 1974.

OSTLE, B., *Statistics in Research Basic Concepts and Techniques for Research Workers*,
The Iowa State College Press, Ames, Iowa, 1956.

SVERDRUP, E. *Lov og tilfeldighet Den praktiske statistikks metode og teknikk*,
I to bind (in Norwegian);
Bind I En elementær innføring, 2. utgave 1973;
Bind II En matematisk videreføring, 1964;
Universitetsforlaget, Oslo.
Laws and Chance Variations Basic Concepts of Statistical Inference, In Two Volumes, (English translation);
Vol. I Elementary Introduction;
Vol. II More Advanced Treatment;
North-Holland Publishing Company, Amsterdam, 1967.

WALPOLE, R.E., *Introduction to Statistics*, Second Edition,
Macmillan Publishing Co., Inc., New York,
Collier Macmillan Publishers, London, 1974.

Books on statistics applied to physics

BRANDT, S., *Statistical and Computational Methods in Data Analysis*, Second Revised Edition, North-Holland Publishing Company, Amsterdam, Elsevier North-Holland, Inc., New York, 1976.

COOPER, B.E., *Statistics for Experimentalists*, Pergamon Press, Oxford-London-Edinburgh-New York-Toronto-Sydney-Paris-Braunschweig, 1969.

EADIE, W.T., DRIJARD, D., JAMES, F.E., ROOS, M., and SADOULET, B., *Statistical Methods in Experimental Physics*, North-Holland Publishing Company, Amsterdam-London, 1971.

JANOSSY, L., *Theory and Practice for the Evaluation of Measurements*, Oxford at the Clarendon Press, Oxford University Press, 1965.

JOHNSON, N.J. and LEONE, F.C., *Statistics and Experimental Design in Engineering and the Physical Sciences*, Volume I, John Wiley & Sons, Inc., New York - London - Sydney, 1964.

MARTIN, B.R., *Statistics for Physicists*, Academic Press, London-New York, 1971.

PARRATT, L.G., *Probability and Experimental Errors in Science*, John Wiley & Sons, Inc., New York-London, 1961.

WINE, R.L., *Statistics for Scientists and Engineers*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1964.

Statistical tables

FISHER, R.A. and YATES, F., *Statistical Tables for Biological, Agricultural and Medical Research*, Oliver and Boyd, Edinburgh, 1938.

MILLER, L.H., Table of percentage points of Kolmogorov statistics, *J. Amer. Statist. Ass.* 51, 111 (1956).

OWEN, D.B., *Handbook of Statistical Tables*, Addison-Wesley Publishing Company, Inc., Reading, Massachusetts, 1962.

RESNIKOFF, G.J. and LIEBERMAN, G.J., *Tables of the non-central t-distribution*, Stanford University Press, 1957.

VERDOREN, L.R., Extended tables of critical values for Wilcoxon's test statistic, *Biometrika* 50, 177 (1963).

PEARSON, E.S. and HARTLEY, H.O. (editors), *Biometrika Tables for Statisticians*, Volumes I and II, Cambridge, 1970 and 1972.

Articles, reports, and lecture notes on statistical estimation in particle physics

ANNIS, M., CHESTON, W., and PRIMAKOFF, H., On statistical estimation in physics, *Rev. Mod. Phys.* 25, 818 (1953).

HUDSON, D.J. *Lectures on Elementary Statistics and Probability*, CERN 63-29 (1963), and *Statistic Lectures II : Maximum Likelihood and Least Squares Theory*, CERN 64-18 (1964).

JAMES, F., *Function Minimization*, CERN 72-21 (1972).

JAUNEAU, L., and MORELLET, D., Notions of statistics and applications, in *Methods in Submuclear Physics*, Volume IV Part 3, Data Handling, edited by M. Nikolič; Gordon and Breach Science Publishers, New York-London-Paris, 1970.

KNOP, R.E. Errors in estimation of total events, *Rev. Sci. Instrum.* 41, 1518 (1970).

OREAR, J., Notes on statistics for physicists, UCRL-8417 (1958).

ROSENFELD, A.H. and HUMPHREY, W.E., Analysis of bubble chamber data, *Ann. Rev. Nucl. Sci.* 13, 103 (1963).

SHEPPEY, G.C., Minimization and curve fitting, in *Programming Techniques*, CERN 68-5 (1968).

SOLMITZ, F., Analysis of experiments in particle physics, *Ann. Rev. Nucl. Sci.* 14, 375 (1964).

Articles referred to in the text

ADAIR, R.K. and KASHA, H., Analysis of some results of quark searches, *Phys. Rev. Letters* 23, 1355 (1969).

BARTLETT, M.S., On the statistical estimation of mean life-times, *Phil. Mag.* 44, 249 (1953), and Estimation of mean lifetimes from multiplate cloud chamber tracks, *Phil. Mag.* 44, 1407 (1953).

DAVID, F.N., A χ^2 "smooth" test for goodness-of-fit, *Biometrika* 34, 299 (1947).

DAVIDON, W.C., Variance algorithm for minimization, *Comp. J.* 10, 406 (1968).

DURBIN, J., Kolmogorov-Smirnov tests when parameters are estimated with applications to tests of exponentiality and tests on spacings, *Biometrika* 62, 5 (1975).

EVANS, D.A. and BARKAS, W.H., Exact treatment of search statistics, *Nucl. Inst. Meth.* 56, 289 (1967).

GELFAND, I.M. and TSETLIN, M.L., The principles of nonlocal search in automatic optimization systems, *Soviet Phys. Dokl.* 6, 192 (1961).

GOLDSTEIN, A.A. and PRICE, J.I., On descent from local minima, *Math. Comput.* 25, 569 (1971).

KRUSKAL, W.H. and WALLIS, W.A., The use of ranks in one-criterion variance analysis, *J. Amer. Statist. Ass.* 47, 583 (1952).

NELDER, J.A. and MEAD, R., A simplex method for function minimization, *Comp. J.* 7, 308 (1965).

PARTICLE DATA GROUP, Review of particle properties, *Rev. Mod. Phys.* 48, No. 2, Part II, April 1976.

REINES, F., GURR, H.S., and SOBEL, H.W., Detection of $\bar{\nu}_e$-e scattering, *Phys. Rev. Letters* 37, 315 (1976).

ROSENBROCK, H., An automatic method for finding the greatest or least value of a function, *Comp. J.* 3, 175 (1960).

WILCOXON, F., Individual comparisons by ranking methods, *Biometrics Bulletin* 1, 80 (1945).

Articles on Monte Carlo methods and multi-dimensional data analysis in particle physics

JAMES, F., *Monte Carlo Phase Space*, CERN 68-15 (1968).

Monte Carlo for particle physicists, in *Methods in Subnuclear Physics*, Volume IV Part 3, Data Handling, edited by M. Nikolić; Gordon and Breach Science Publishers, New York-London-Paris, 1970.

KITTEL, W., VAN HOVE, L., and WOJCIK, W., A Monte Carlo generation method with importance sampling for high energy collisions of hadrons, *Comput. Phys. Comm.* 1, 425 (1970).

BENZÉCRI, J.P., Statistical Methods and Their Adaption in Lorentz Invariant Framework, in *Proceedings of the 3rd Topical Meeting on Multi-Dimensional Analysis of High Energy Data, Nijmegen 8-11 March 1978*, edited by W. Kittel (1978).

FRIEDMAN, J.H., Data analysis technique for high energy particle physics, in *Proceedings of the 1974 CERN School of Computing*, CERN 74-23 (1974).

Multi-dimensional associative searching, in *Proceedings of the 1976 CERN School of Computing*, CERN 76-24 (1976).

INDEX