Ajustes de modelos teóricos a datos experimentales: el método de cuadrados mínimos

Luciano A. Masullo

Laboratorio 1 (1er Cuatrimestre 2018)
Departamento de Física
Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires

Planteo del problema

- Tenemos un conjunto de mediciones $\{x_i\}_{i=1,\ldots,N}$

Planteo del problema

- Tenemos un conjunto de mediciones $\{x_i\}_{i=1,\dots,N}$
- Tenemos otro conjunto de mediciones $\{y_i\}_{i=1,\dots,N}$

Planteo del problema

- Tenemos un conjunto de mediciones $\{x_i\}_{i=1,\dots,N}$
- Tenemos otro conjunto de mediciones $\{y_i\}_{i=1,\dots,N}$
- Tenemos como **hipótesis** una relación funcional entre las mediciones x e y de la forma $y = f(x)$. Este es nuestro **modelo teórico**.
- Por ejemplo:
  - $y = f(x) = mx + q$ es una relación **lineal** entre x e y
  - $y = f(x) = A\sin(\omega x + \phi) + c$ es una relación **no-lineal** entre x e y

Planteo del problema

- Tenemos un conjunto de mediciones $\{x_i\}_{i=1,\dots,N}$
- Tenemos otro conjunto de mediciones $\{y_i\}_{i=1,\dots,N}$
- Tenemos como **hipótesis** una relación funcional entre las mediciones x e y de la forma $y = f(x)$. Este es nuestro **modelo teórico**.
- Por ejemplo:
  - $y = f(x) = mx + q$ es una relación **lineal** entre x e y
  - $y = f(x) = A\,\mathrm{sin}(\omega x + \phi) + c$ es una relación **no-lineal** entre x e y

Asumiendo un cierto modelo teórico (p. ej. $y = mx + q$) ¿Cómo encuentro los parámetros del modelo (p. ej. m y q) que **mejor explican*** mis datos experimentales?

* Se dice también “que mejor ajustan a” y de ahí viene hablar de “ajuste”

Planteo del problema

- Tenemos un conjunto de mediciones $\{x_i\}_{i=1,\dots,N}$
- Tenemos otro conjunto de mediciones $\{y_i\}_{i=1,\dots,N}$
- Tenemos como **hipótesis** una relación funcional entre las mediciones x e y de la forma $y = f(x)$. Este es nuestro **modelo teórico**.
- Por ejemplo:
  - $y = f(x) = mx + q$ es una relación **lineal** entre x e y ⟶ Tiene solución *analítica*
  - $y = f(x) = A\,\mathrm{sin}(\omega x + \phi) + c$ es una relación **no-lineal** entre x e y ↘ Tiene solución *numérica*

Asumiendo un cierto modelo teórico (p. ej. $y = mx + q$) ¿Cómo encuentro los parámetros del modelo (p. ej. m y q) que **mejor explican*** mis datos experimentales?

* Se dice también “que mejor se ajustan a” y de ahí viene hablar de “ajuste”

Método de cuadrados mínimos

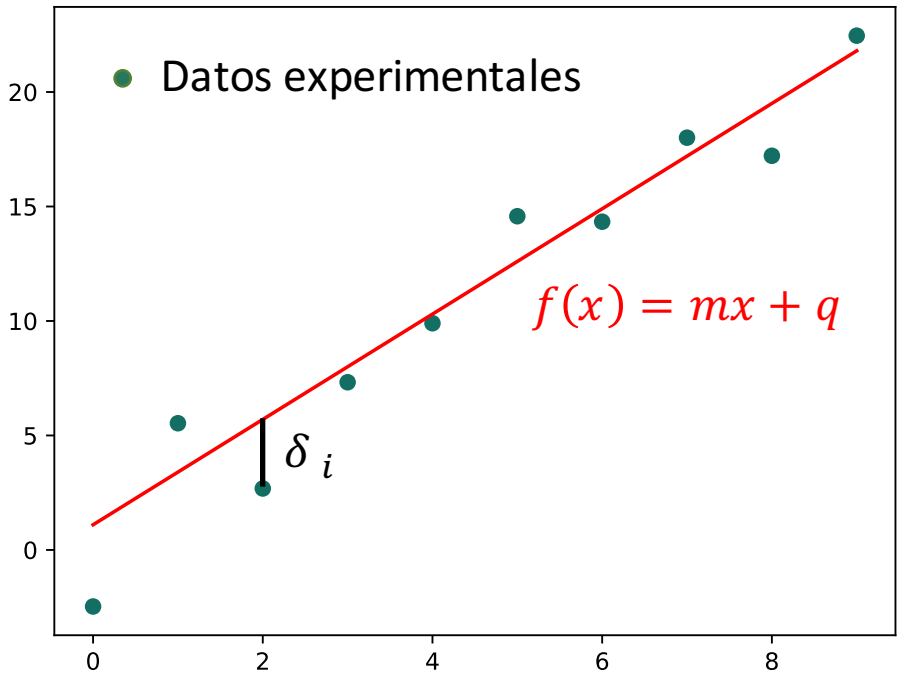
- Criterio: quiero encontrar la función $f(x)$ que minimice la suma de las diferencias al cuadrado entre la predicción $f(x_i)$ y la medición y_i

$$S = \sum_{i=1}^{N} \delta_i^2 = \sum_{i=1}^{N} \left(f\left(x_i\right) - y_i\right)^2$$

- Hipótesis adicionales: la variable aleatoria y tiene distribución gaussiana con valor medio $f(x)$. Además el error relativo en x es mucho menor que el error relativo en y.

- Cuadrados mínimos **caso lineal**: asumo como modelo teórico $y = f(x) = mx + q$, luego

$$S = \sum_{i=1}^{N} \delta_i^2 = \sum_{i=1}^{N} \left|(mx_i + q) - y_i\right|^2$$



Método de cuadrados mínimos

$$S(m,q) = \sum_{i=1}^{N} |(mx_i + q) - y_i|^2 = \sum_{i=1}^{N} y_i^2 + m^2 \sum_{i=1}^{N} x_i^2 + Nq^2 + 2mq \sum_{i=1}^{N} x_i - 2m \sum_{i=1}^{N} x_i y_i - 2q \sum_{i=1}^{N} y_i$$

Método de cuadrados mínimos

$$S(m,q) = \sum_{i=1}^{N} |(mx_i + q) - y_i|^2 = \sum_{i=1}^{N} y_i^2 + m^2 \sum_{i=1}^{N} x_i^2 + Nq^2 + 2mq \sum_{i=1}^{N} x_i - 2m \sum_{i=1}^{N} x_i y_i - 2q \sum_{i=1}^{N} y_i$$

$$\frac{\partial S(m,q)}{\partial m} = 0$$

$$\frac{\partial S(m,q)}{\partial q} = 0$$

Minimización de S en función de m y q

Método de cuadrados mínimos

$$S(m,q) = \sum_{i=1}^{N} |(mx_i + q) - y_i|^2 = \sum_{i=1}^{N} y_i^2 + m^2 \sum_{i=1}^{N} x_i^2 + Nq^2 + 2mq \sum_{i=1}^{N} x_i - 2m \sum_{i=1}^{N} x_i y_i - 2q \sum_{i=1}^{N} y_i$$

$$\left.\begin{aligned} \frac{\partial S(m,q)}{\partial m} &= 0 \\ \frac{\partial S(m,q)}{\partial q} &= 0 \end{aligned}\right] \longrightarrow \begin{aligned} 2m \sum_{i=1}^{N} x_i^2 + 2q \sum_{i=1}^{N} x_i - 2 \sum_{i=1}^{N} x_i y_i &= 0 \\ 2Nq + 2m \sum_{i=1}^{N} x_i - 2 \sum_{i=1}^{N} y_i &= 0 \end{aligned}$$

Minimización de S en función de m y q

Método de cuadrados mínimos

$$S(m,q) = \sum_{i=1}^{N} |(mx_i + q) - y_i|^2 = \sum_{i=1}^{N} y_i^2 + m^2 \sum_{i=1}^{N} x_i^2 + Nq^2 + 2mq \sum_{i=1}^{N} x_i - 2m \sum_{i=1}^{N} x_i y_i - 2q \sum_{i=1}^{N} y_i$$

$$\left.\begin{aligned} \frac{\partial S(m,q)}{\partial m} &= 0 \\ \frac{\partial S(m,q)}{\partial q} &= 0 \end{aligned}\right] \longrightarrow \left[\begin{aligned} 2m \sum_{i=1}^{N} x_i^2 + 2q \sum_{i=1}^{N} x_i - 2 \sum_{i=1}^{N} x_i y_i &= 0 \\ 2Nq + 2m \sum_{i=1}^{N} x_i - 2 \sum_{i=1}^{N} y_i &= 0 \end{aligned}\right] \longrightarrow$$

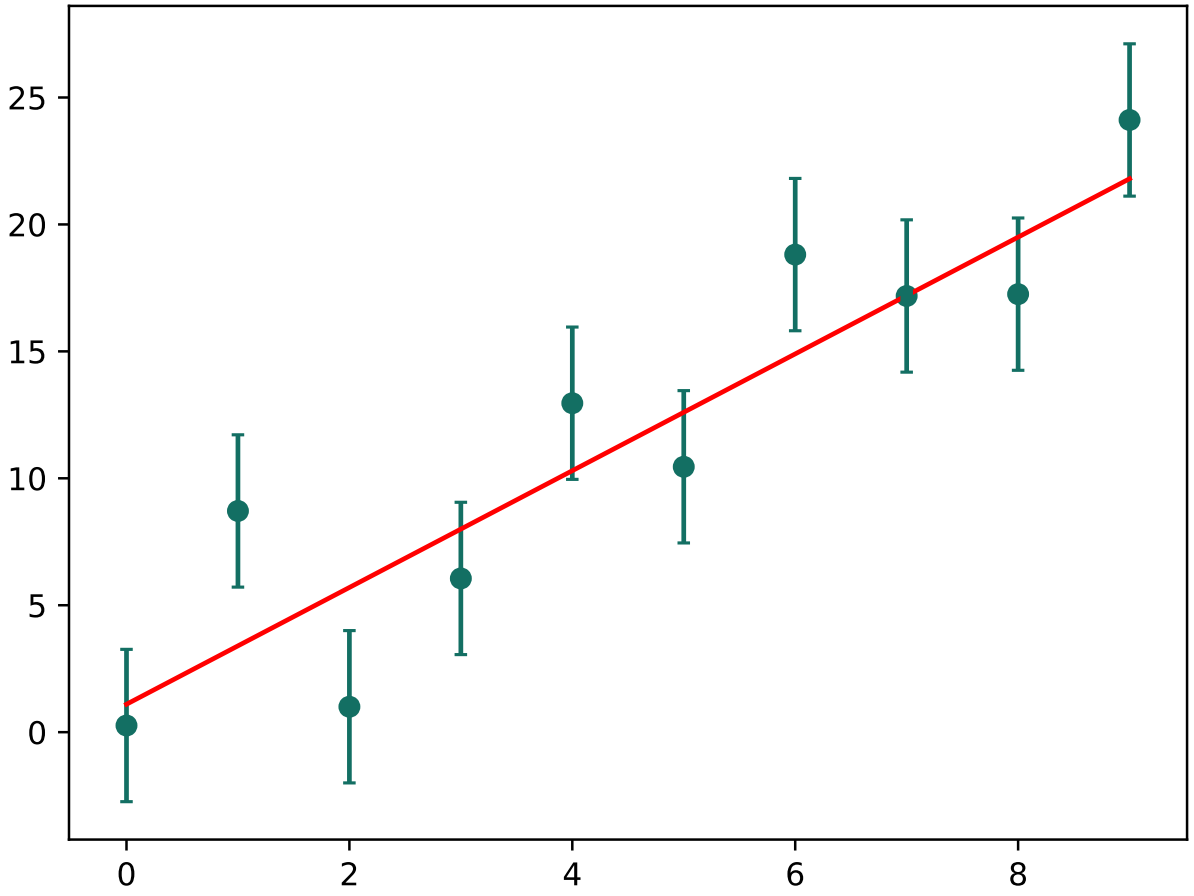
$$m = \frac{N \sum x_i y_i - \sum x_i \sum y_i}{N \sum x_i^2 - (\sum x_i)^2}$$

$$q = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{N \sum x_i^2 - (\sum x_i)^2}$$

Minimización de S en función de m y q

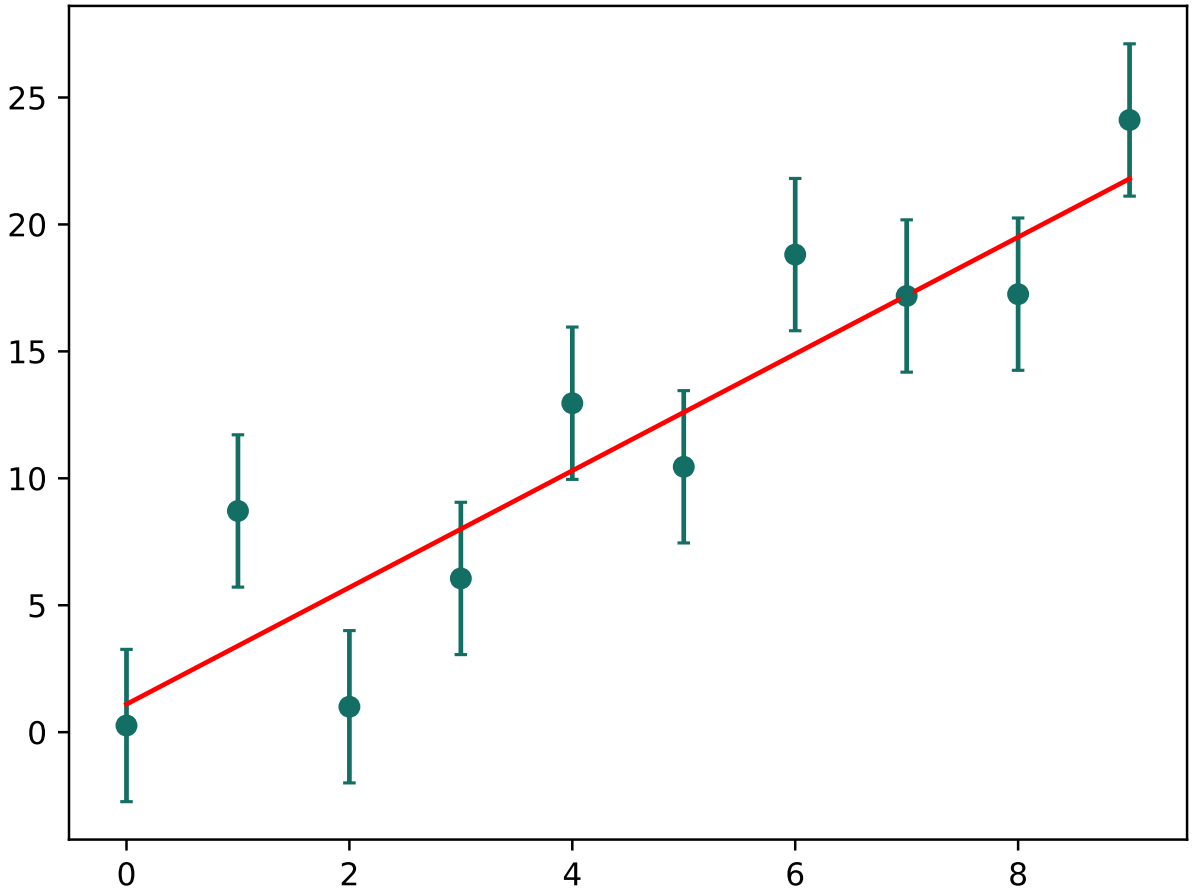
Encontramos los parámetros m y q que minimizan la suma cuadrática de las diferencias entre el modelo y los datos

Método de cuadrados mínimos



De todas las posibles funciones lineales, la función $f(x) = mx + q$ es la que minimiza S

Método de cuadrados mínimos



Algunas consideraciones importantes:

- ¿Es mejor un ajuste que esté contenido dentro de las barras de error de *todos* los puntos que uno que no?
- ¿Qué relación hay entre la cantidad de puntos dentro de cuyas barras de error pasa el ajuste y el error de las mediciones?
- ¿Cuántos puntos espero de cada “lado” de la recta? ¿Por qué?

Parámetros de un MRUV por cuadrados mínimos

- Posición en función del tiempo
- Velocidad en función del tiempo
- Aceleración en función del tiempo

Parámetros de un MRUV por cuadrados mínimos

- Posición en función del tiempo

$$x(t) = x_0 + v_0(t - t_0) + \frac{1}{2}a(t - t_0)^2 \qquad \text{Si } v_0 = 0 \longrightarrow x((t - t_0)^2) = x_0 + \frac{1}{2}a(t - t_0)^2$$

lineal

- Velocidad en función del tiempo

$$v(t) = v_0 + a(t - t_0)$$

- Aceleración en función del tiempo

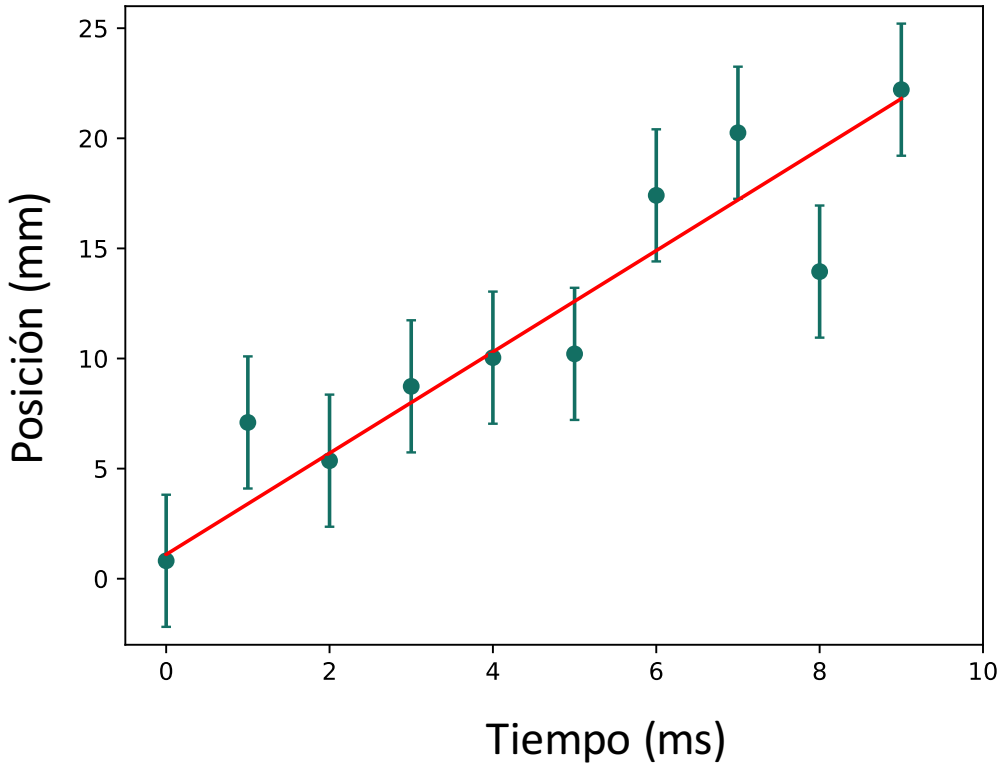
$$a(t) = a$$

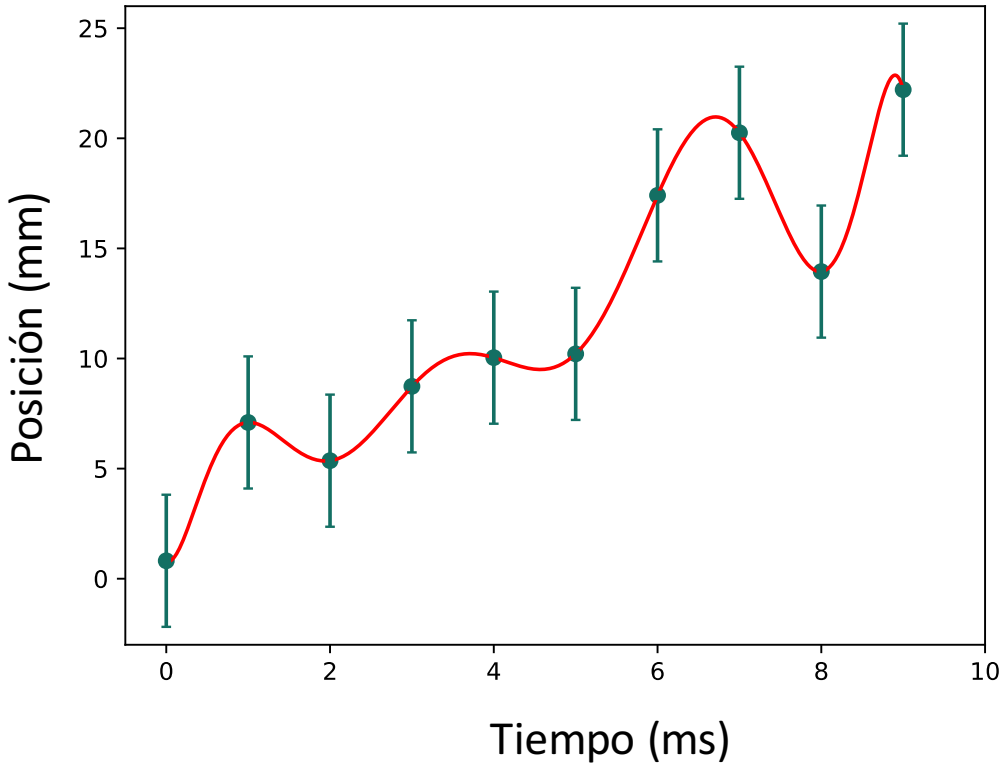
Limitaciones del método y criterios

Siempre es esencial **evaluar críticamente** el resultado de un ajuste:

- ¿Es siempre el mejor modelo el que mejor se ajusta a los datos?
- ¿Es el mejor modelo el que menos hipótesis hace sobre el problema?
- ¿Es razonable plantear un modelo que se ajusta muy bien pero que no puedo predecir con razonamientos y argumentos físicos?

Limitaciones del método y criterios (ejemplo #1)





¿Qué ajuste elegirían? ¿Por qué?

Limitaciones del método y criterios (ejemplo #2)

Tabla de cuatro conjuntos de datos

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Estos datos ajustan exactamente igual de bien a una misma función lineal $y = 3x + 0{,}5$

Adaptado de F.J. Anscombe, "Graphs in Statistical Analysis," American Statistician, 27 (1973), 17-21

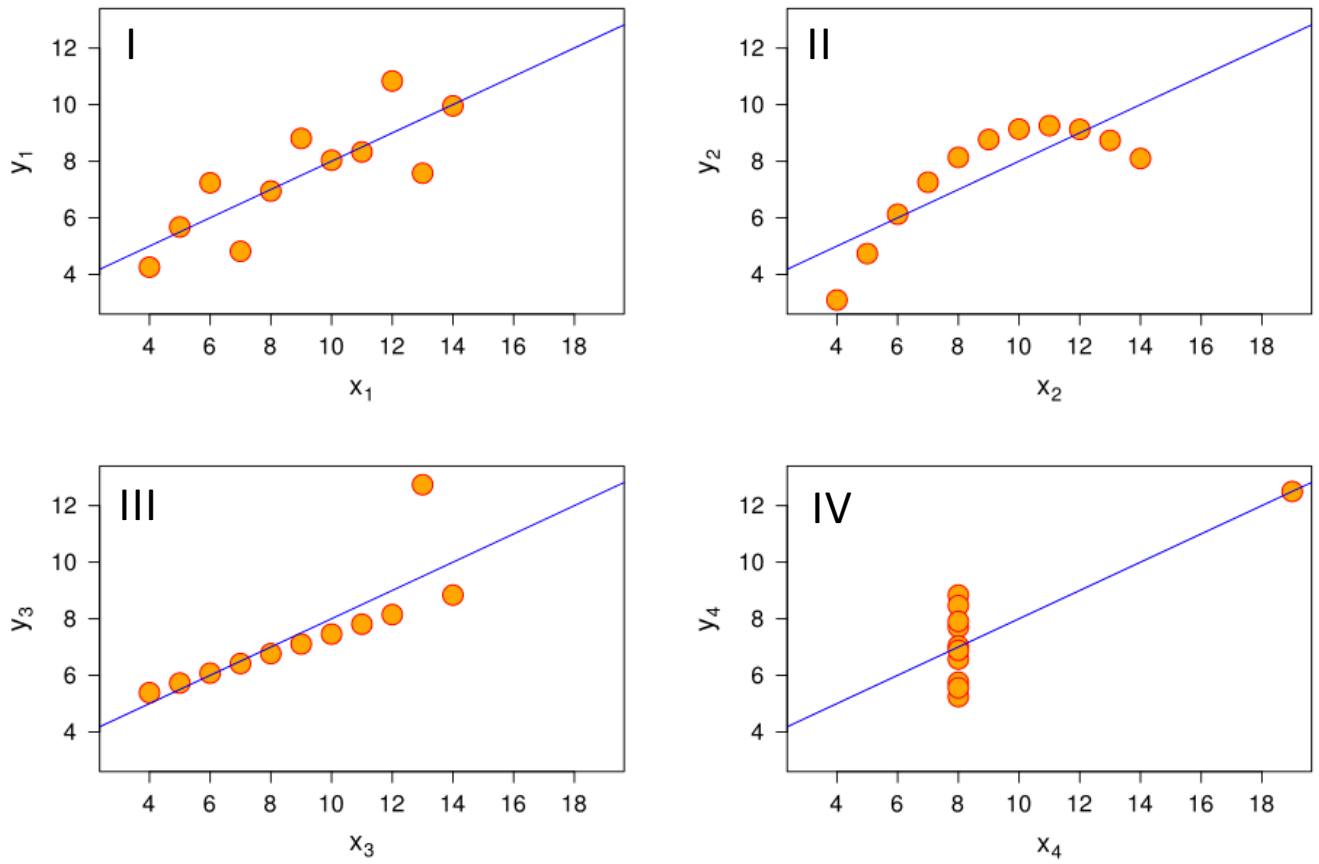
Limitaciones del método y criterios (ejemplo #2)

Tabla de cuatro conjuntos de datos

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Estos datos ajustan exactamente igual de bien a una misma función lineal $y = 3x + 0{,}5$

Los cuatro conjuntos de datos graficados



Adaptado de F.J. Anscombe, "Graphs in Statistical Analysis," American Statistician, 27 (1973), 17-21

Cuantifación de la verosimilitud de un ajuste: χ^2

Construyo un estadístico (una función de los datos) con esta forma:

$$\chi^2 \equiv \sum_{i=1}^{k} \left(\frac{y_i - f(x_i)}{\sigma_i} \right)^2$$

"Chi cuadrado"

Cuantifación de la verosimilitud de un ajuste: χ^2

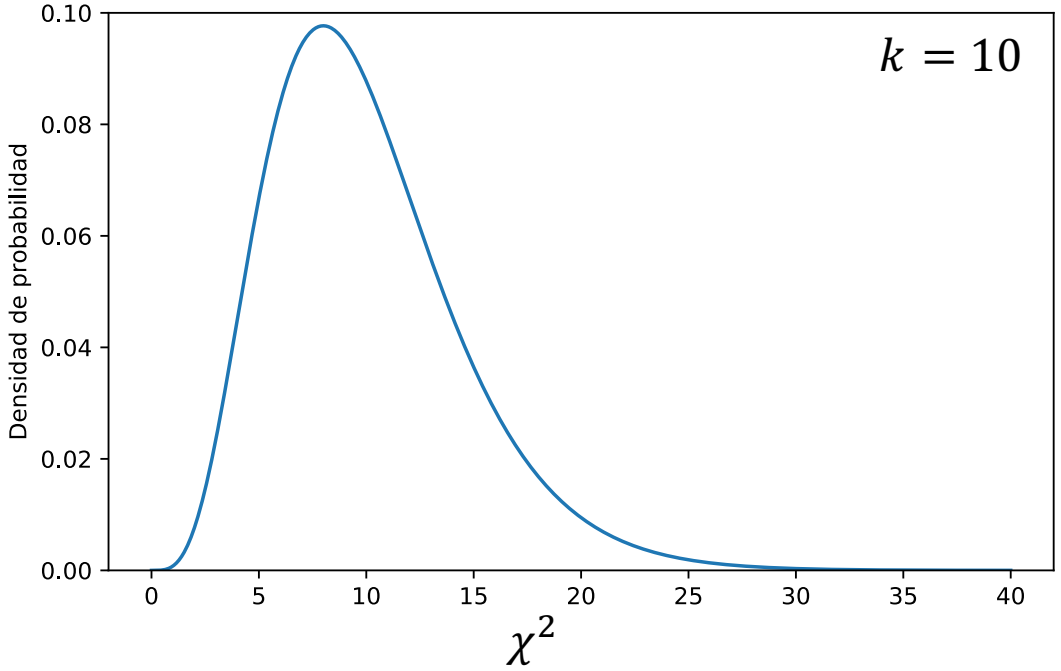
Construyo un estadístico (una función de los datos) con esta forma:

$$\chi^2 \equiv \sum_{i=1}^{k} \left(\frac{y_i - f(x_i)}{\sigma_i} \right)^2$$

"Chi cuadrado"

- χ^2 es una variable aleatoria, por lo tanto tiene una distribución de probabilidad asociada

Distribución de probabilidad de χ^2



Cuantifación de la verosimilitud de un ajuste: χ^2

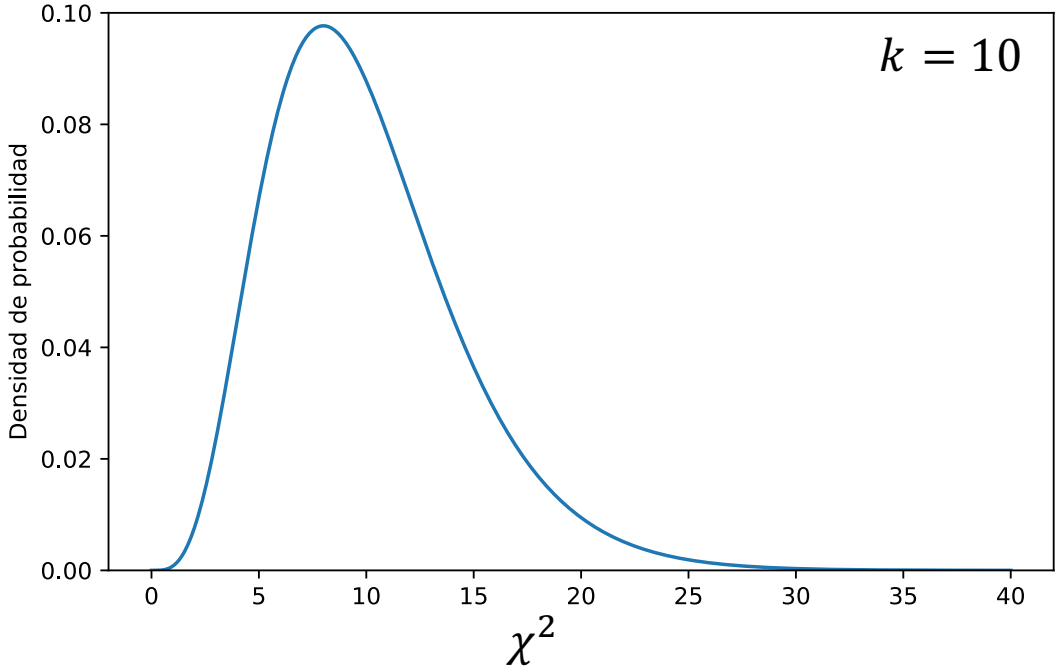
Construyo un estadístico (una función de los datos) con esta forma:

$$\chi^2 \equiv \sum_{i=1}^{k} \left(\frac{y_i - f(x_i)}{\sigma_i} \right)^2$$

"Chi cuadrado"

- χ^2 es una variable aleatoria, por lo tanto tiene una distribución de probabilidad asociada
- k es la cantidad de mediciones del experimento (la forma de la distribución de χ^2 va a depender de k)

Distribución de probabilidad de χ^2



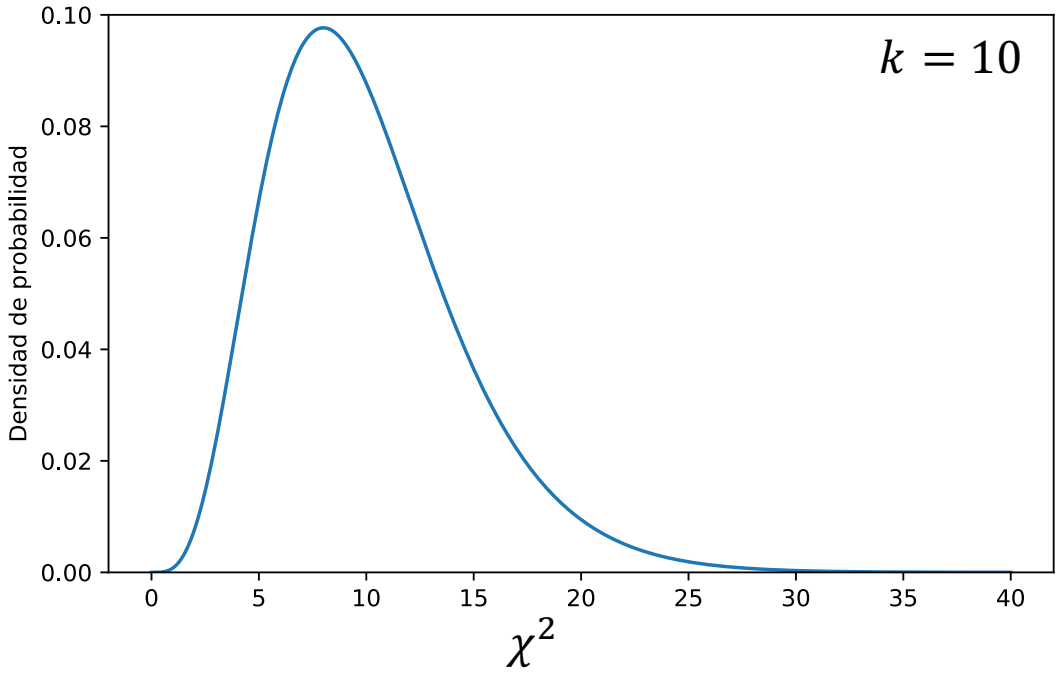
Cuantifación de la verosimilitud de un ajuste: χ^2

Construyo un estadístico (una función de los datos) con esta forma:

$$\chi^2 \equiv \sum_{i=1}^{k} \left(\frac{y_i - f(x_i)}{\sigma_i} \right)^2$$

"Chi cuadrado"

Distribución de probabilidad de χ^2



- χ^2 es una variable aleatoria, por lo tanto tiene una distribución de probabilidad asociada
- k es la cantidad de mediciones del experimento (la forma de la distribución de χ^2 va a depender de k)
- $E(\chi^2) = k$
- $VAR(\chi^2) = 2k \; ; \; SD(\chi^2) = \sqrt{2k}$

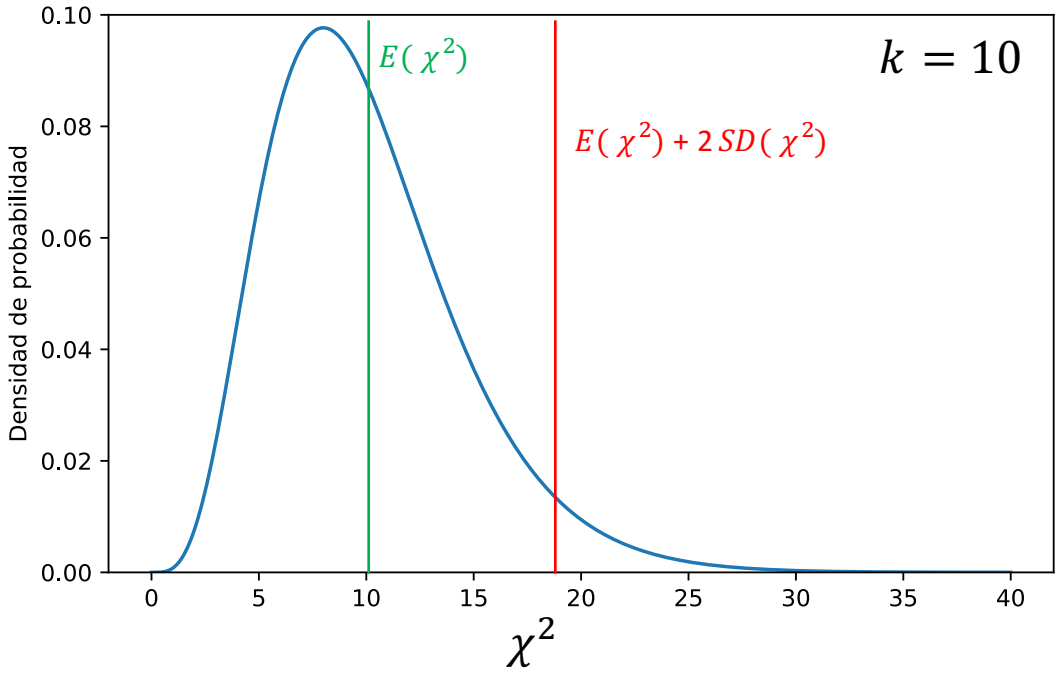
Cuantifación de la verosimilitud de un ajuste: χ^2

Construyo un estadístico (una función de los datos) con esta forma:

$$\chi^2 \equiv \sum_{i=1}^{k} \left(\frac{y_i - f(x_i)}{\sigma_i} \right)^2$$

"Chi cuadrado"

Distribución de probabilidad de χ^2



- χ^2 es una variable aleatoria, por lo tanto tiene una distribución de probabilidad asociada
- k es la cantidad de mediciones del experimento (la forma de la distribución de χ^2 va a depender de k)
- $E(\chi^2) = k$
- $VAR(\chi^2) = 2k\ ;\ SD(\chi^2) = \sqrt{2k}$

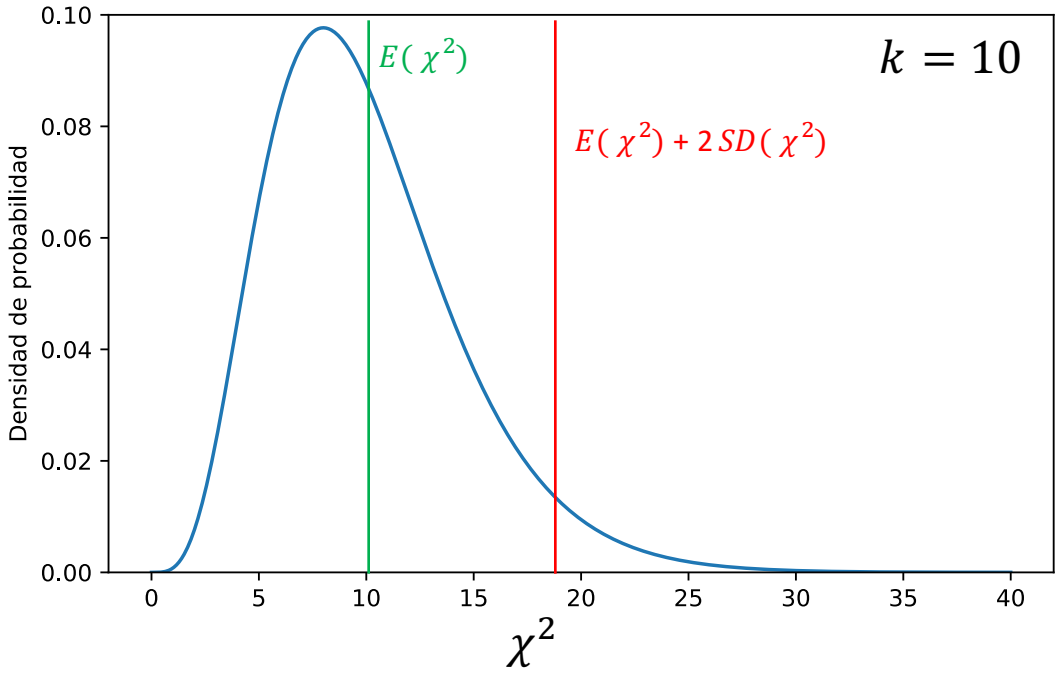
Cuantifación de la verosimilitud de un ajuste: χ^2

Construyo un estadístico (una función de los datos) con esta forma:

$$\chi^2 \equiv \sum_{i=1}^{k} \left(\frac{y_i - f(x_i)}{\sigma_i} \right)^2$$

"Chi cuadrado"

Distribución de probabilidad de χ^2

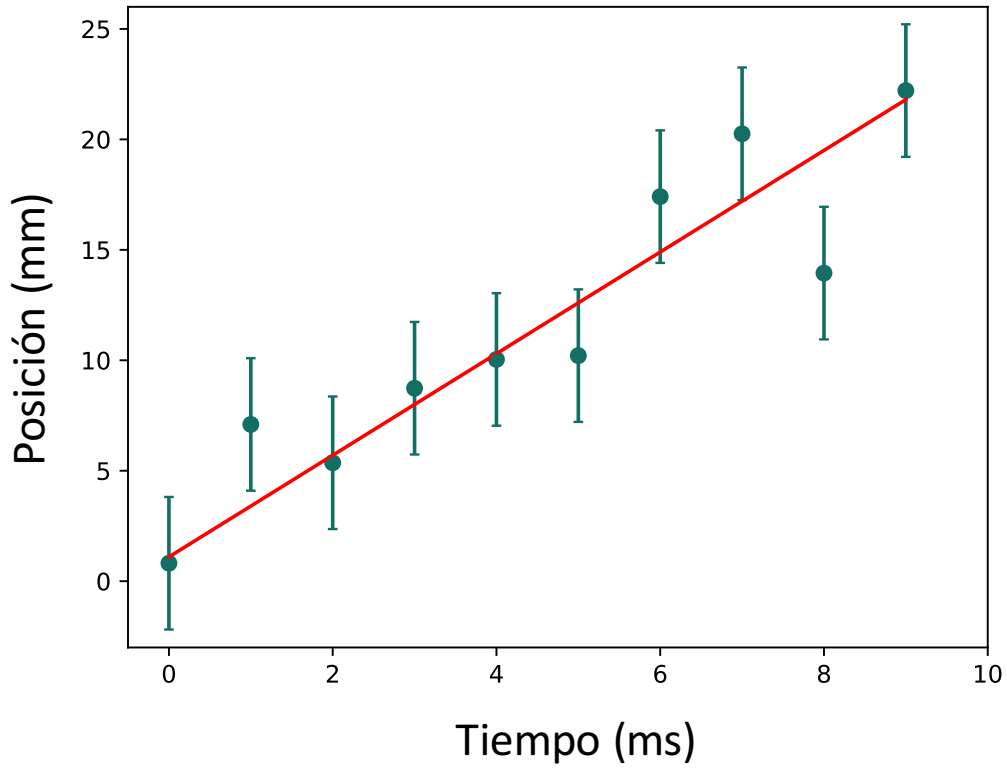


- χ^2 es una variable aleatoria, por lo tanto tiene una distribución de probabilidad asociada
- k es la cantidad de mediciones del experimento (la forma de la distribución de χ^2 va a depender de k)
- $E(\chi^2) = k$
- $VAR(\chi^2) = 2k\ ;\ SD(\chi^2) = \sqrt{2k}$
- El valor de χ^2 es una medida (no la única) de la probabilidad de que los datos $\{y_i\}$ provengan de un fenómeno físico con modelo teórico $y = f(x)$, es decir, χ^2 es una medida de la **verosimilitud** del modelo o ajuste

Cuantifación de la verosimilitud de un ajuste: χ^2

¿Cuándo rechazo un modelo (teoría) o cuándo la acepto?

Tengo 10 mediciones y quiero evaluar la verosimilitud de mi modelo teórico

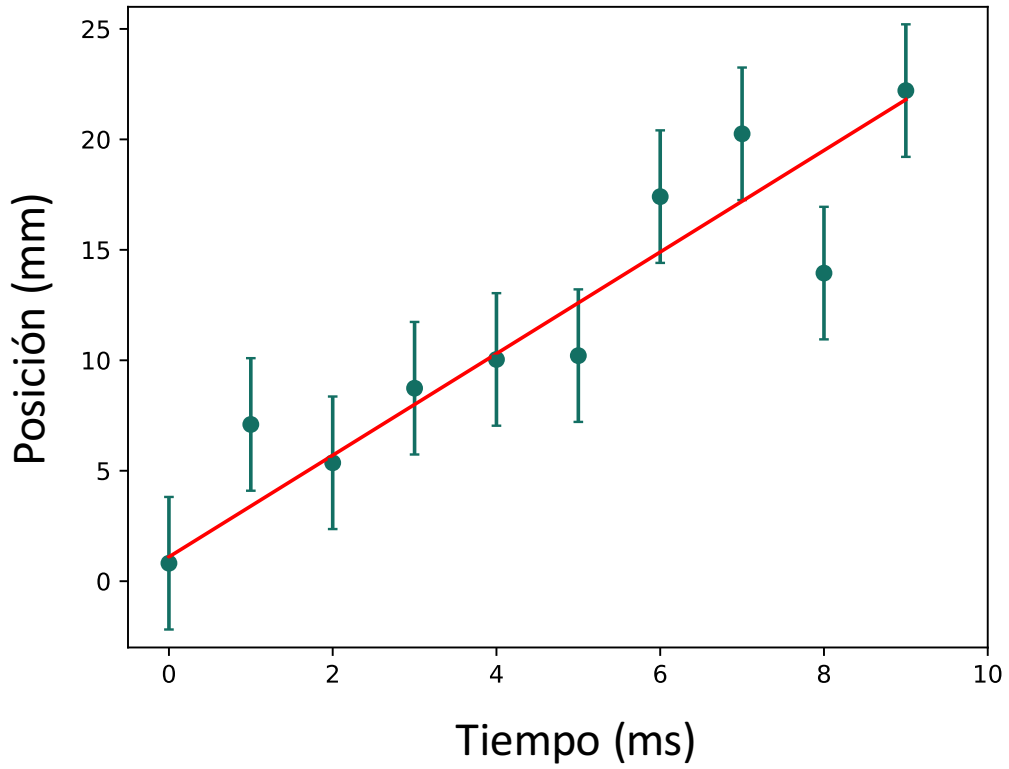


$$\chi^2 = \sum_{i=1}^{k} \left(\frac{y_i - f(x_i)}{\sigma_i} \right)^2$$

Cuantifación de la verosimilitud de un ajuste: χ^2

¿Cuándo rechazo un modelo (teoría) o cuándo la acepto?

Tengo 10 mediciones y quiero evaluar la verosimilitud de mi modelo teórico

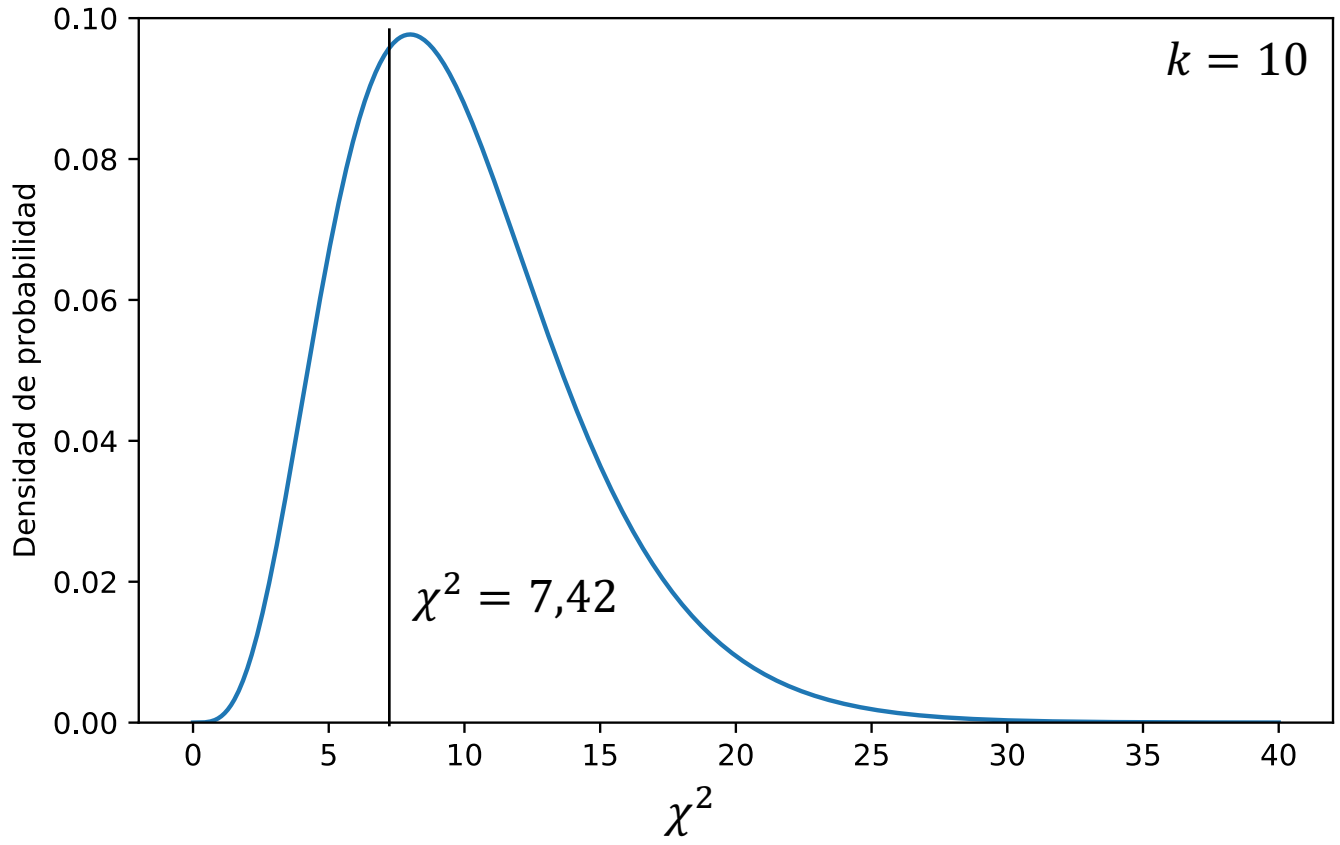


$$\chi^2 = \sum_{i=1}^{k} \left(\frac{y_i - f(x_i)}{\sigma_i} \right)^2$$

$$\boxed{\chi^2 = 7{,}42}$$

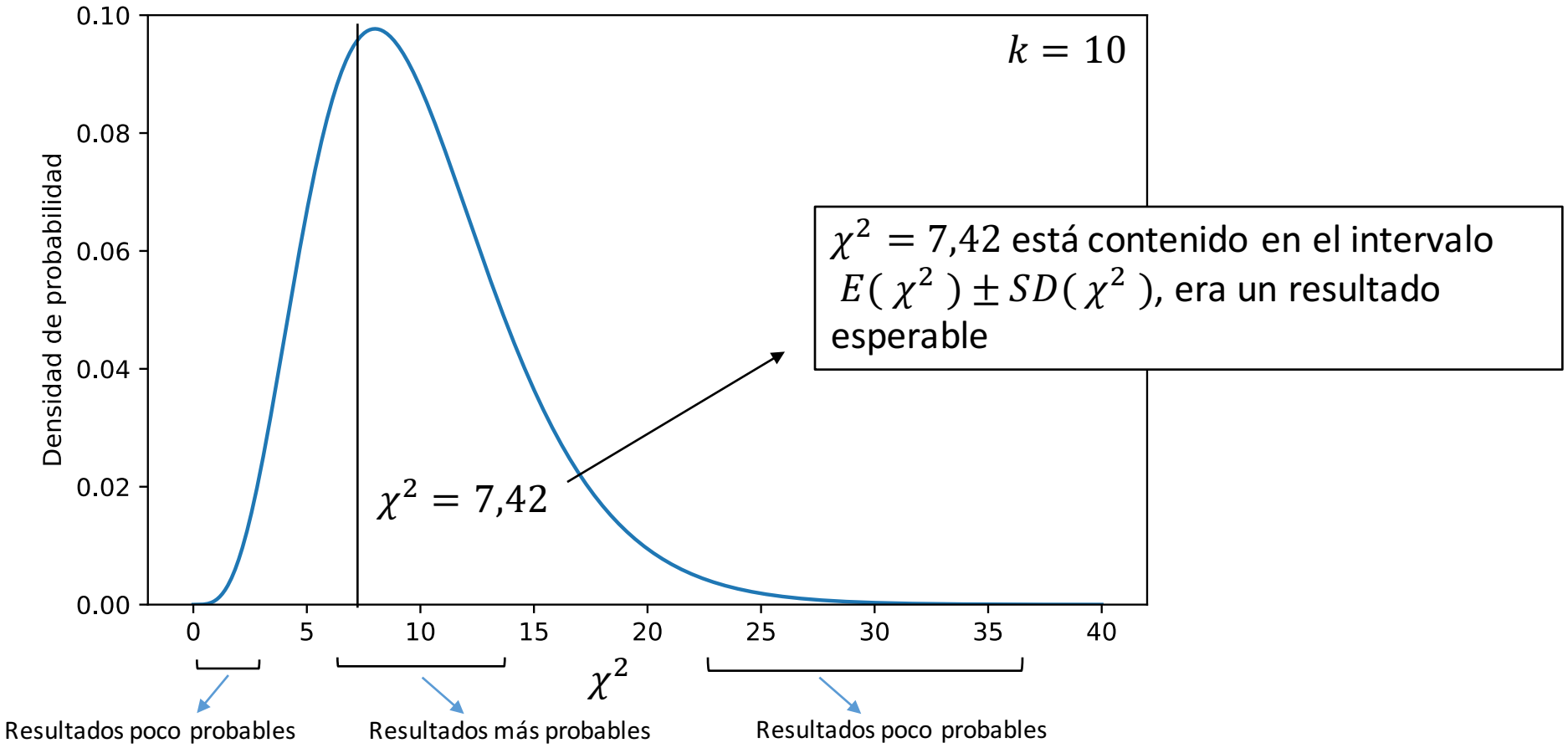
Cuantifación de la verosimilitud de un ajuste: χ^2

¿Cuándo rechazo un modelo (teoría) o cuándo la acepto?



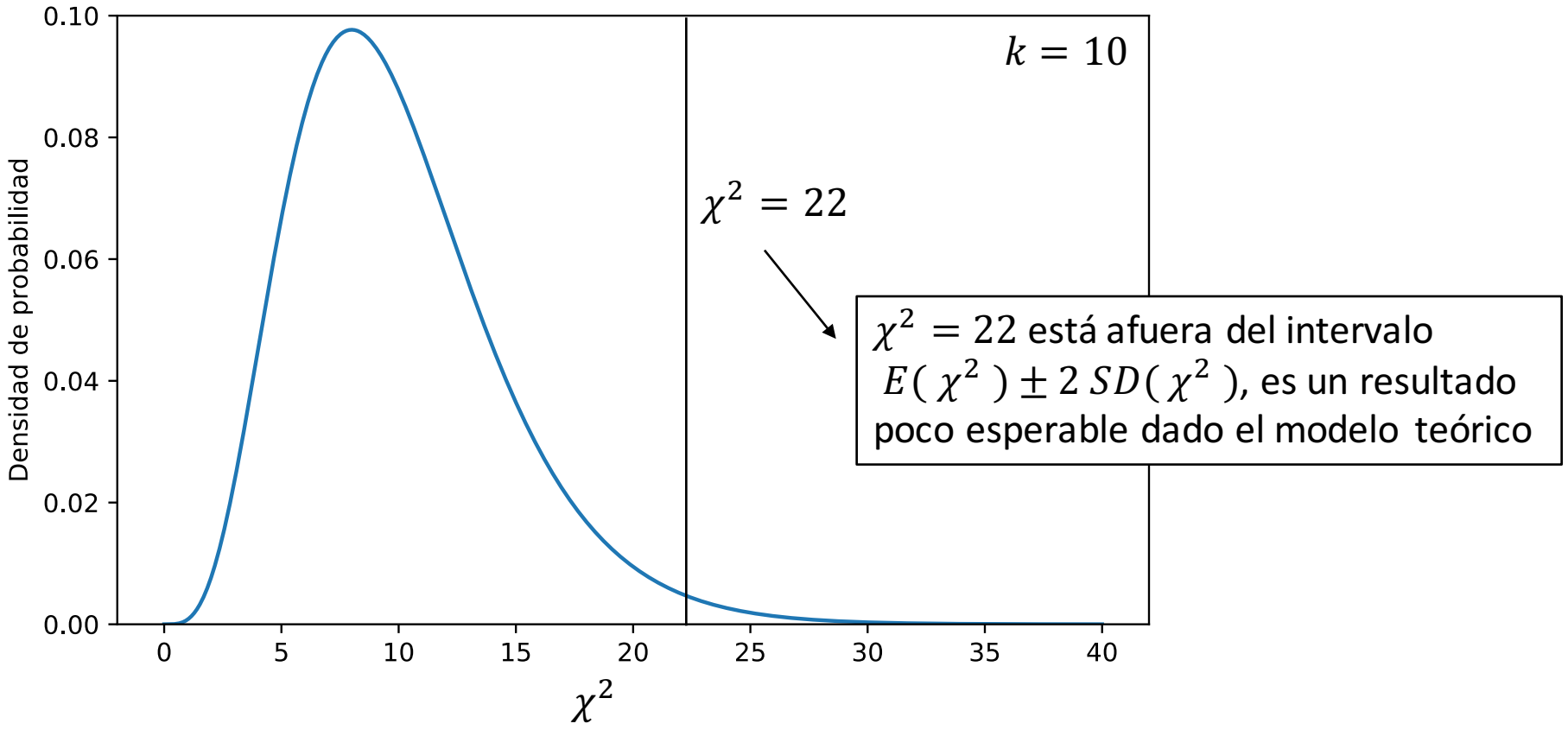
Cuantifación de la verosimilitud de un ajuste: χ^2

¿Cuándo rechazo un modelo (teoría) o cuándo la acepto?



Cuantifación de la verosimilitud de un ajuste: χ^2

¿Cuándo rechazo un modelo (teoría) o cuándo la acepto?



χ^2 reducido

Se puede definir en forma análoga al χ^2, un χ^2 *reducido* :

$$\chi^2_{red} \equiv \frac{1}{k}\sum_{i=1}^{k}\left(\frac{y_i - f(x_i)}{\sigma_i}\right)^2 \longrightarrow \begin{cases} E\left(\chi^2_{red}\right) = 1 \\ VAR\left(\chi^2_{red}\right) = \frac{2}{k} \\ SD\left(\chi^2_{red}\right) = \sqrt{\frac{2}{k}} \end{cases}$$

- Sirve para independizar el valor medio de la cantidad de mediciones k
- El programa Origin computa $\chi^2{}_{red}$ en vez del χ^2

Resumen

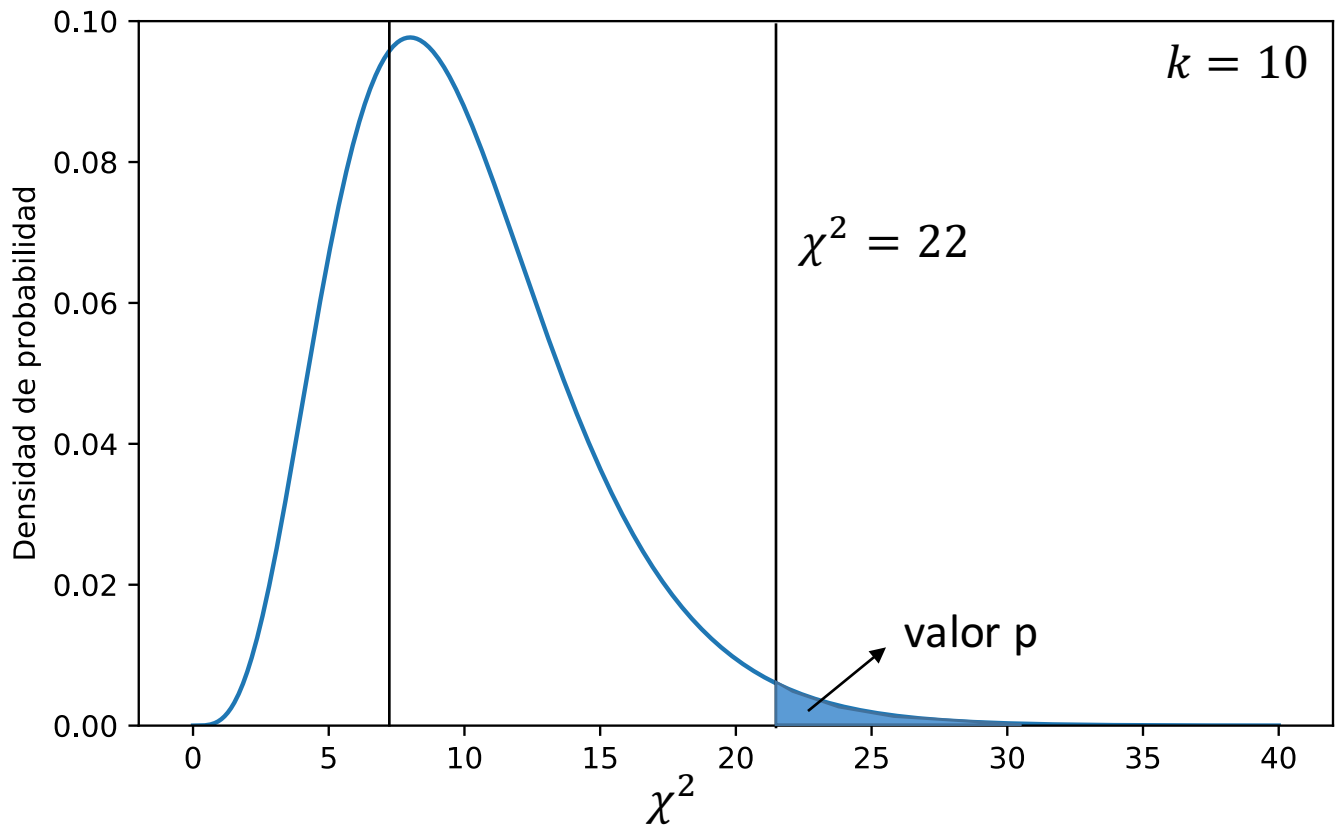
- “Cuadrados mínimos” es un método (no el único) que sirve para encontrar los parámetros que mejor se ajustan a los datos experimentales dado un modelo teórico

- Si bien es un método poderoso y muy útil, hay que tener cuidado y evaluar los resultados de los ajustes con criterio y sin perder de vista la física del experimento

- χ^2 es un *estadístico* (y una variable aleatoria) que da una medida (no la única) de la verosimilitud del ajuste efectuado a los datos experimentales

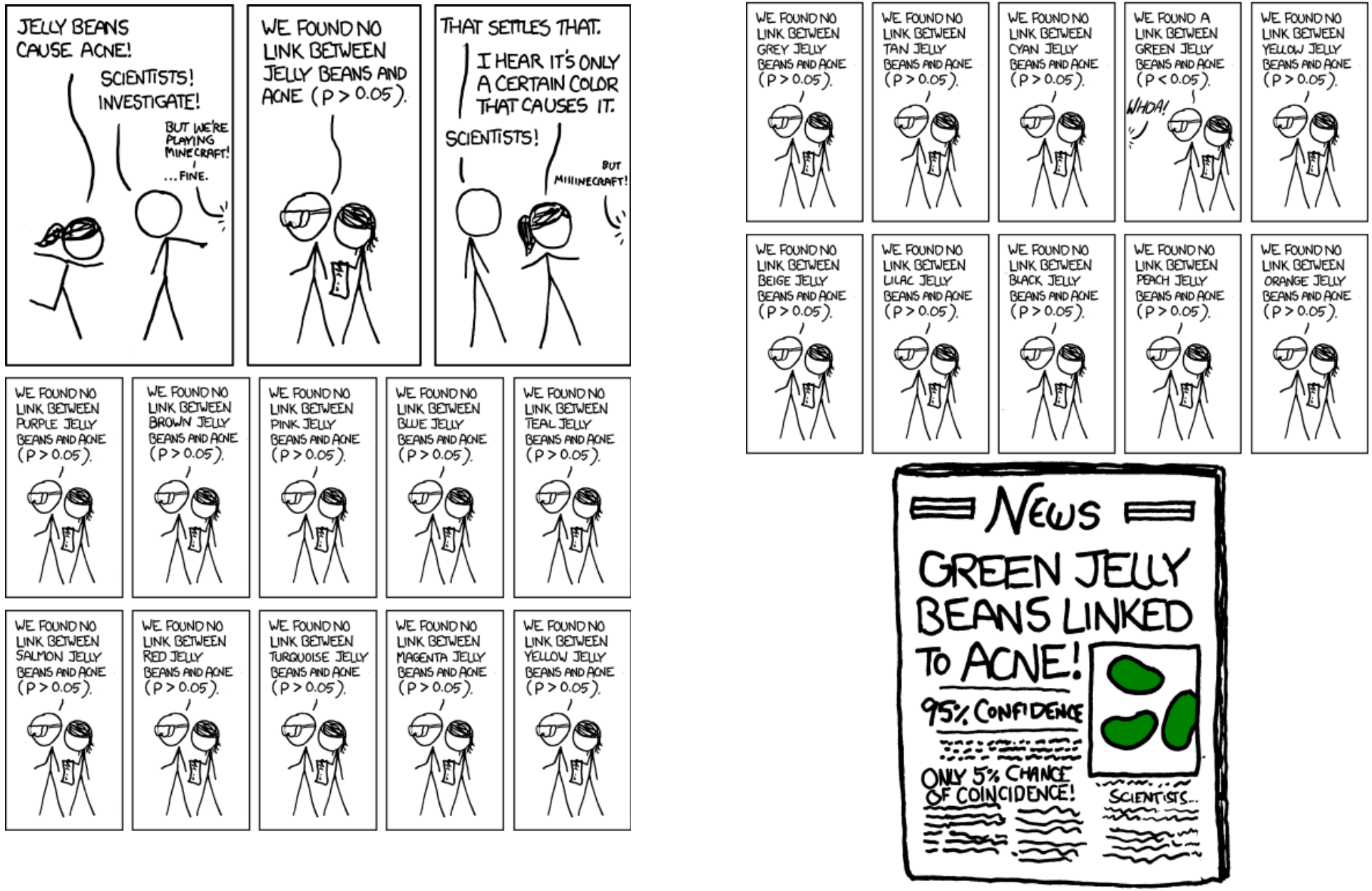
Extra: Test de hipótesis y p-valor

¿Cuándo rechazo un modelo (teoría) o cuándo la acepto?



- Recordemos que la integral de la densidad de probabilidad da 1
- Se define el **valor p** como la probabilidad de obtener un conjunto de mediciones como el que obtuve o más extremo (menos probable aún) dada una cierta hipótesis o modelo
- Se suele* tomar **valor p** < 0,05 (que a su vez define un χ^2 umbral) para rechazar un modelo (o aceptar el contrario)
- El valor p determina el nivel de **significación** de la discrepancia del modelo con los datos

* La elección es arbitraria y de hecho actualmente hay un fuerte debate en el ámbito de las ciencias biomédicas, ver por ejemplo: S. Wellek "A critical evaluation of the current 'p-value controversy'", *Biometrical Journal*, **59**, 5 (2017)


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...