Estadística en Física Experimental (Verano 2019)

Guía de Problemas N° 4 | Distribuciones multidimensionales – Teorema Central del Límite – Covarianza y Correlación

Distribuciones multidimensionales

1. Sea X una variable aleatoria con densidad de probabilidad simétrica alrededor de cero. Muestre que X e $Y{=}X^2$ tienen correlación nula a pesar de no ser independientes.

2. La suma de dos distribuciones uniformes es una distribución triangular:
   (a) Si X e Y son variables independientes con distribución uniforme en [0,1], halle la distribución conjunta $g(U,V)$ de $U \equiv X+Y$ y $V \equiv X-Y$.
   (b) Tomando la correspondiente distribución marginal, muestre que U es una variable aleatoria con distribución triangular:
   $$f_U(t) = \begin{cases} t & 0 < t < 1 \\ 2-t & 1 < t < 2 \\ 0 & \text{en otro caso} \end{cases}$$
   (c) Encuentre la distribución de V y determine si U y V son independientes.
   (d) Calcule la varianza de U via $\int_0^2 (t-1)^2 f_U(t)\mathrm{d}t$.
   (e) Confirme que se obtiene el mismo resultado usando la propiedad $\mathrm{Var}(a_1X_1 + a_2X_2 + \ldots + a_NX_N) = \sum_{i=1}^N a_i^2 \mathrm{Var}(X_i) + 2\sum_{i=1}^N \sum_{j>i}^N a_i a_j \mathrm{Cov}(X_i, X_j)$.

3. La suma de gaussianas es gaussiana:
   (a) Probar que si X e Y son variables independientes con distribución normal de parámetros (μ_1, σ_1) y (μ_2, σ_2), entonces $Z = X + Y$ es una gaussiana de parámetros $(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2})$.
   (b) Ahora bien, si la suma de gaussianas es gaussiana, ¿cómo es que en el problema 9b de la guía 3 mostró que con la suma de dos gaussianas consigue aproximar una distribución de Cauchy, que claramente no es gaussiana?

4. Muestre que el cociente $Z \equiv X/Y$ de dos variables independientes con distribución normal canónica tiene distribución de Cauchy, $f_Z(t) = 1/[\pi(1+t^2)]$.

5. Sean X e Y dos variables independientes con distribución uniforme en [0,1], a partir de las cuales se definen $U = \sqrt{-2a\ln X}\cos(2\pi Y)$ y $V = \sqrt{-2a\ln X}\sin(2\pi Y)$. Encuentre la distribución conjunta $g(U,V)$, identifique qué distribución es, indique el significado del parámetro a y determine si U y V son independientes. Nota: este mapeo de $X, Y \to U, V$ se llama transformación de Box-Muller.

Teorema Central del Límite

6. Estudie el grado de validez del teorema central del límite dibujando las distribuciones siguientes, y superponiendo sobre ellas la gaussiana con el μ y σ correspondiente.
   (a) B_k(5,0.2), B_k(30,0.4)
   (b) P_n(4), P_n(10), P_n(40)

7. El teorema central del límite permite evaluar probabilidades binomiales sin necesidad de sumar muchos términos que involucran factoriales de grandes números, a partir de la distribución acumulativa normal canónica $\Phi(x)$,
   $$\sum_{k=a}^{b} B_k(n,p) = \sum_{k=a}^{b} \binom{n}{k} p^k q^{n-k} \simeq \Phi\left(\frac{b - np + \frac{1}{2}}{\sqrt{npq}}\right) - \Phi\left(\frac{a - np - \frac{1}{2}}{\sqrt{npq}}\right)$$
   Discuta el origen de esta fórmula y utilícela para calcular la probabilidad de aprobar un examen multiple choice con 100 preguntas de tres opciones cada una, si se contesta al azar y se aprueba con 4 (40% de respuestas correctas). [Rta: 0.0966 con la suma exacta, y 0.0951 con la fórmula aproximada.]

8. Utilizando el teorema central del límite escribir un generador aproximado de números gaussianos N(0,1), a partir de variables aleatorias independientes $\{X_i\}$ con distribución uniforme en [0,1], como una función $f(Z)$ siendo $Z = \sum_i^n X_i$.
   (a) Si se elige n=50, ¿cuál debe ser $f(Z)$?
   (b) ¿En qué rango de la abscisa seguro falla la aproximación a la normal?
   (c) Genere de este modo 10000 números con la computadora, haga un histograma de su distribución, y grafique N(0,1) sobre éste.
   (d) Muestre que el promedio de N variables independientes con distribución de Cauchy tiene a su vez distribución de Cauchy. ¿Por qué falla en este caso el teorema central del límite?

9. ¿Cuánta gente deberá encuestarse en Argentina si se desea conocer dentro de un 1% la intención de voto a un candidato con un nivel de confianza de 95%, sabiendo que aproximadamente (a) el 45% (b) el 5% del electorado votará por él? Discuta intuitivamente por qué obtiene distintos resultados para los casos (a) y (b). [Rta: 9900 y 1900]
   Sugerencia: considerar que la población tiene muchos más individuos que cualquiera de estas muestras y usar la aproximación gaussiana.

Covarianza y correlación

10. Para cada uno de los cuatro pares de datos de la tabla:
    (a) Calcular la media muestral de X y de Y,
    (b) Calcular la varianza muestral de X y de Y,
    (c) Calcular la correlación entre X e Y: $\rho(X,Y) = \dfrac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}} = \dfrac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$;
    (d) Graficar cada par de puntos.

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

Nota: estos son cuatro conjuntos de datos que F.J. Anscombe generó para mostrar que hacer buenos gráficos de los datos son una parte esencial del análisis de regresión lineal. F. J, Anscombe, (1973). "Graphs in Statistical Analysis". Am Stat, Vol. 27, No. 1, 17-21

11. La *Distribución Multinormal* es la generalización a n dimensiones de la normal (la gaussiana) y, al igual que ésta, juega un rol preponderante en probabilidades y estadística. Dadas n variables aleatorias correlacionadas $\{X_i\}$, con esperanza $E(X_i) = \mu_i$ y matriz de covarianza $\mathbb{V}$, ésto es $\mathrm{Cov}(X_i,X_j)=V_{ij}$, se dice que su densidad de probabilidad conjunta $f(\underline{x})$ es multinormal si todas las distribuciones marginales $f(x_i)$ y todas las distribuciones condicionales unidimensionales $f(x_i|x_j, j \neq i)$ son gaussianas. La densidad de probabilidad conjunta $f(\underline{x})$ viene dada por

$$f(\underline{x}) = \frac{1}{\sqrt{(2\pi)^n |\mathbb{V}|}} \exp\left[-\frac{1}{2}(\underline{x}-\underline{\mu})^T \mathbb{V}^{-1} (\underline{x}-\underline{\mu})\right]$$

donde $\underline{x}$ y $\underline{\mu}$ son vectores columna de tamaño n, $\underline{x}^T$ y $\underline{\mu}^T$ los respectivos vectores traspuestos (vectores fila) y $\mathbb{V}$ es cuadrada (de $n \times n$), simétrica y definida positiva, con $|\mathbb{V}| \equiv \det(\mathbb{V})$.

   (a) Verifique que para $n = 1$, $f(\underline{x})$ es una gaussiana.

(b) En el caso $n = 2$ (multinormal bivariada) la matriz de covarianza de una multinormal depende de tres parámetros (¿por qué?). Elijamos σ_1, σ_2 y el coeficiente de correlación ρ, ésto es, $V_{11} = \sigma_1^2$, $V_{22} = \sigma_2^2$ y $V_{12} = \rho\,\sigma_1\sigma_2$. Muestre entonces que

$$f(x_1, x_2) = \left(2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}\right)^{-1} \exp\left(-\frac{Q}{2}\right)$$

con

$$Q = \frac{1}{1-\rho^2}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)\right]$$

(c) Compruebe que cuando $\rho = 0$, $f(x_1, x_2) = N(\mu_1, \sigma_1)N(\mu_2, \sigma_2)$. Esto es, para la multinormal, correlación nula implica que las variables son independientes.

En adelante, puede trabajar con $\mu_1 = \mu_2 = 0$ para simplificar las cuentas.

(d) Muestre que la distribución marginal $f(x_2)$ es la gaussiana $N(\mu_2, \sigma_2)$, independientemente del valor del nivel de correlación ρ.

(e) Una manera de visualizar la forma de una multinormal con $n = 2$ es dibujar curvas de nivel de f en el plano x_1, x_2. Considere las correspondientes a $Q = 1$, y muestre que son elipses centradas en (μ_1, μ_2), denominadas *elipses de covarianza*. Para $\mu_1 = \mu_2 = 0$, verifique que éstas están contenidas en el rectángulo $(\pm\sigma_1, \pm\sigma_2)$, que son tangentes a dicho rectángulo en los puntos $(\sigma_1, \rho\,\sigma_2)$ y $(\rho\,\sigma_1, \sigma_2)$.

(f) Muestre que $f(x_2|x_1)$ es gaussiana, con $N(\mu_2 + \rho(\sigma_2/\sigma_1)(x_1 - \mu_1), \sigma_2\sqrt{1-\rho^2})$. Discuta cómo varía la esperanza de x_2 en función de x_1 según el signo de ρ, y analice cómo varía el ancho de la distribución condicional con el grado de correlación. Interprete estos resultados cortando con líneas x_1=cte las elipses dibujadas a mano alzada en el item anterior. ¿Qué ocurre en el caso límite $\rho = 1$?

12. Aplicando los resultados del ejercicio anterior para el caso de $\underline{x}$ bidimensional,

(a) Dibuje a mano alzada elipses de covarianza con distintos ρ para el caso $\sigma_1 = \sigma_2$. Discuta la diferencia entre tomar como error para X_1 el rango máximo cubierto por la elipse sobre el eje x_1, o el segmento entre los puntos de intersección de la elipse con el eje x_1.

(b) ¿Por qué tiene más sentido considerar la elipse como rango de confianza, que el propio rectángulo $(\pm k\sigma_1, \pm k\sigma_2)$?

(c) Considere las elipses de covarianza encerradas dentro del rectángulo $(\pm k\sigma_1, \pm k\sigma_2)$ alrededor de (μ_1, μ_2). Muestre que la probabilidad conjunta de que (x_1, x_2) se encuentre dentro de una de estas elipses con k=1 es 39.3%, independientemente del valor de la correlación ρ (este resultado es el equivalente al 68.3% obtenido para el caso n=1). Sugerencia: pensar en otro suceso que tenga la *misma* probabilidad que el suceso "(x_1, x_2) se encuentra dentro de una de estas elipses" y que involucre a la variable aleatoria Q.

(d) ¿Cuánto debería ser k para que la elipse corresponda a un nivel de confianza de 95%? Verifique que este resultado puede obtenerse también analíticamente (para el caso bidimensional), además de usando las tablas. [Rta: k=2.448]