Probabilidad multidimensional

Análisis estadístico de datos

2021

1. Graficar las superficies de nivel 1σ de las distribuciones binormales con parámetros $\mu_1 = 1.3$, $\mu_2 = 0.5$, $\sigma_1 = 1.7$, y $\sigma_2 = 2.3$ para tres diferentes correlaciones $\rho = -0.9$, 0 y 0.5.

2. Considerar dos variables independientes X_1 y X_2 distribuidas con la función de densidad de probabilidad binormal,

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2} \exp\bigg(- \frac{q(x_1, x_2)}{2} \bigg),$$

   dónde

$$q(x_1, x_2) = \left(\frac{x_1 - \mu_1}{\sigma_1} \right)^2 + \left(\frac{x_2 - \mu_2}{\sigma_2} \right)^2.$$

   Transformar a una variable χ^2 para calcular la probabilidad que la variable aleatoria bidimensional (X_1, X_2) caiga en las regiones 1σ, 2σ y 3σ definidas por las elipses $q = 1$, $q = 4$ y $q = 9$ respectivamente. Sin hacer cuentas, pensar como cambiarían las elipses si X_1 y X_2 estuvieran correlacionadas y cuánta probabilidad contienen las elipses 1σ, 2σ y 3σ.

3. Considerar la distribución binormal con parámetros $\mu_1 = \mu_2 = 0$, $\sigma_1 = \sigma_2 = 1$ y correlación $\rho \neq 0$. Escribir la matriz Hessiana $\boldsymbol{A}$ correspondiente a estos cinco parámetros y encontrar analíticamente sus autovalores y autovectores. Graficar las regiones 1σ, 2σ y 3σ correspondientes a la forma cuadrática $q(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x}$ para $\rho = -0.9$.

4. La eficacia de una vacuna (VE) se puede medir en un test clínico en base a la proporción de pacientes contagiados vacunados con respecto al número total de pacientes contagiados (p),

$$\mathrm{VE} = \frac{1 - 2\,p}{1 - p}$$

   .

En el test de la vacuna covid de Astra-Zeneca se estimó que la media de la variable aleatoria p es $\mathrm{E}(p) = 0.28$ y su desviación estándar es $\sigma(p) = 0.05$. Aplicando la fórmula de propagación de la varianza calcular la media y la desviación estándar de la eficacia VE.

5. Considerar dos variables independientes X_1 y X_2 con varianzas σ_1^2 y σ_2^2. Calcular la matriz de covarianza de las nuevas variables $Y_1 = X_1 + X_2$ y $Y_2 = X_1 - X_2$ en términos de σ_1 y σ_2. Calcular la correlación entre Y_1 e Y_2 y decidir si estas dos variables son independientes.

6. En un experimento se miden los siguientes datos que se ajustan con una regresión lineal:

i	x_i	y_i
1	1	1.85
2	2.5	2.72
3	3.1	5.15
4	4	5.7
5	5.5	6.9

Los datos y_i se consideran como un vector $\boldsymbol{y}$ de cinco variables normales. Se asume además las variables y_i tienen la misma desviación estándar $\sigma = 0.5$ (homocedasticidad). La pendiente ($\hat{m}$) y la ordenada al origen ($\hat{y}_0$) de la recta que ajusta los datos son una nueva variable aleatoria de dos dimensiones:

$$\hat{\boldsymbol{\theta}} = \begin{pmatrix} \hat{y}_0 \\ \hat{m} \end{pmatrix}$$

Sus valores se calculan con la multiplicación de matrices $\hat{\boldsymbol{\theta}} = \mathbf{B}\,\boldsymbol{y}$ dónde

$$\boldsymbol{B} = \begin{pmatrix} 0.834 & 0.406 & 0.234 & -0.023 & -0.451 \\ -0.197 & -0.064 & -0.011 & 0.069 & 0.202 \end{pmatrix}$$

.

Calcular la ordenada al origen y la pendiente de la recta que ajusta los datos. Graficar los datos y el ajuste. Expresar la matriz de covarianza de $\boldsymbol{y}$. Considerando que la variable aleatoria en dos dimensiones $\hat{\boldsymbol{\theta}}$ es una combinación linear de la variable aleatoria en cinco dimensiones $\boldsymbol{y}$ vía la matriz $\boldsymbol{B}$, calcular la matriz de covarianza de $\hat{\boldsymbol{\theta}}$. Calcular las desviaciones estándares de $\hat{y}_0$, $\hat{m}$ y su correlación. Identificar la distribución que sigue $\hat{\boldsymbol{\theta}}$ y los valores de sus parámetros.

7. **(Para entregar)** Simular una variable aleatoria $X = (X_1, X_2)$ que sigue una distribución binormal con parámetros $\mu_1 = 2.3$, $\mu_2 = 1.5$, $\sigma_1 = 1.2$, $\sigma_2 = 0.5$ y correlación $\rho = 0.7$. Usar que la distribución conjunta $f(x_1, x_2) = h_2(x_2|\, x_1)\, g_1(x_1)$, con $g_1(x_1)$ la distribución marginal de X_1 y $h_2(x_2|\, x_1)$ la distribución condicional de X_2 dado X_1. Esta distribución condicional se puede expresar como $h_2(x_2|x_1) = N(\mu_2', \sigma_2')$ con $\mu_2' = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x_1 - \mu_1)$ y ${\sigma_2'}^2 = \sigma_2^2\,(1 - \rho^2)$. Repetir la simulación 1000 veces. Calcular la fracción de eventos caen en la elipse 1σ y comparar con la probabilidad contenida dentro de dicha región. Graficar los datos simulados junto a la elipse 1σ. *Nota: un punto* (x_1, x_2) *pertenece a la elipse 1*σ *si la forma cuadrática asociada* $q(x_1, x_2) \leq 1$.