Universidad de la República **Estadística Multivariada Computacional**
Facultad de Ingeniería - IMERL **Primer Semestre 2021**

Práctico 2

1. Sean $X_1 \sim \mathcal{N}(0,1)$ y $U \in \{-1,1\}$ tal que $\mathbb{P}(U=1) = \mathbb{P}(U=-1) = 0{,}5$ independiente de X_1. Consideramos la variable aleatoria $X_2 = UX_1$.

   a) Pruebe que $X_2 \sim \mathcal{N}(0,1)$ pero que $\mathbf{X} = (X_1, X_2)$ no es un vector con distribución normal.

   b) Simule $n = 100$ datos del vector (X_1, X_2) y analizar gráficamente las distribuciones univariadas (mediante histogramas y densidades normales, por ejemplo) y conjunta (diagrama de dispersión).

2. Recordar que (X, Y) tiene distribución normal estándar bivariada con correlación ρ si existe $Z \sim \mathcal{N}(0,1)$ independiente de X tal que $Y = \rho X + \sqrt{1-\rho^2} Z$.

   a) Probar que si $X = x$ entonces $Y \sim \mathcal{N}(\rho x, 1-\rho^2)$

   b) Usando que $f(y|x) = f(x,y)/f(x)$, probar que la densidad conjunta es

$$f(x,y) = \frac{1}{(2\pi)\sqrt{1-\rho^2}} \, e^{-\frac{1}{2}\begin{pmatrix} x & y \end{pmatrix}\Sigma^{-1}\begin{pmatrix} x \\ y \end{pmatrix}}$$

   siendo $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$

3. Consideramos un vector aleatorio $\mathbf{X} \in \mathbb{R}^p$ con distribución normal de parámetros (μ, Σ), que particionamos como $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2}) \in \mathbb{R}^{p_1} \times \mathbb{R}^{p_2}$, $p_1 + p_2 = p$. El vector de esperanzas y la matriz de varianzas-covarianzas también queda partida por bloques como $\mu = (\mu_{\mathbf{1}}, \mu_{\mathbf{2}})'$ y $Var(\mathbf{X}) = \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$

   donde $\Sigma_{11} = Var(\mathbf{X_1}) \in \mathcal{M}_{p_1 \times p_1}$, $\Sigma_{22} = Var(\mathbf{X_2}) \in \mathcal{M}_{p_2 \times p_2}$, $\Sigma_{12} = Cov(\mathbf{X_1}, \mathbf{X_2}) \in \mathcal{M}_{p_1 \times p_2}$ y $\Sigma_{21} = \Sigma_{12}'$
   La distribución condicionada de $\mathbf{X_1}$ condicionada a $\mathbf{X_2} = X_2^0$ es normal con vector de esperanzas:

$$\mathbb{E}(\mathbf{X_1} | \mathbf{X_2} = X_2^0) = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2^0 - \mu_{\mathbf{2}})$$

   y matriz de varianzas-covarianzas:

$$Var(\mathbf{X_1} | \mathbf{X_2} = X_2^0) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

   Sea $\mathbf{X} = (X_1, X_2, X_3)'$ un vector normal con media $\mu = (-1, 1, 0)'$ y matriz $\Sigma = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 3 & 1 \\ 1 & 1 & 2 \end{pmatrix}$

   a) Encuentre la distribución de la variable aleatoria $Y = X_1 + 2X_2 - 3X_3$

   b) Encuentre un vector $\mathbf{a} = (a_1, a_2)$ tal que las variables aleatorias X_1 y $X_1 - a' \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ sean independientes

   c) Calcule la distribución de X_3 condicionada a $(X_1, X_2)' = (x_1^0, x_0^2)$.

4. *a*) Pruebe que si $X \sim \mathcal{N}(\mu, \sigma^2)$ entonces $Y = \left(\frac{X-\mu}{\sigma}\right)^2 \sim \chi^2(1)$.

   b) Pruebe que si $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ y $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ son variables aleatorias independientes, entonces $\left(\frac{X_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{X_2-\mu_2}{\sigma_2}\right)^2 \sim \chi^2(2)$ (se sugiere utilizar la función generatriz de momentos).

   c) Generalizar para n.

5. Considere un vector aleatorio $\mathbf{X}$ con distribución normal estándar en $\mathbb{R}^p$ ($\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathrm{I}_\mathbf{p})$), y para cada $p = 1, 2, \ldots, 16$ estime la probabilidad $\mathbb{P}(\mathbf{X}'\mathbf{X} \leq 1)$.

   Se sugiere reproducir el mismo esquema de simulación por Monte Carlo del documento sobre la maldición de la dimensionalidad.

6. Probar que si $\mathbf{z}$ es normal estándar en $\mathbb{R}^p$, entonces su función generatriz de momentos $M : \mathbb{R}^p \to \mathbb{R}$ definida como $M(t) = \mathbb{E}(e^{t'\mathbf{z}})$ es $e^{t't/2}$. Deducir que la función generatriz de momentos de $\mathbf{x} \sim \mathcal{N}(\mu, \mathbf{\Sigma})$ es $e^{t'\mu + t'\Sigma t/2}$.

7. *a*) Consideramos en $\mathbb{R}^p$ las n variables independientes $\mathbf{y_i} \sim \mathcal{N}(\mu_i, \Sigma_i)$. Pruebe que

$$\sum_{i=1}^{n} a_i \mathbf{y_i} \sim \mathcal{N}\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \Sigma_i\right)$$

   Sugerencia: utilizar el resultado del ejercicio anterior.

   b) Mostrar que la condición de que las variables sean independientes es necesaria para que la combinación lineal sea normal.

   c) Mostrar que la combinación lineal de densidades es una densidad pero que la combinación lineal de densidades de variables aleatorias normales no es la densidad de una variable aleatoria normal.

8. Sea $\mathbf{x}$ un vector aleatorio con *distribución mezclada* cuya densidad está dada por $f(\mathbf{x}) = \sum_{i=1}^{G} \pi_i f_i(\mathbf{x})$.

   Si (μ_i, Σ_i) es el vector de medias y la matriz de varianzas y covarianzas, respectivamente, de la componente i-éstima de la mezcla, probar que

$$\mu = \mathbb{E}(\mathbf{x}) = \sum_{\mathbf{i=1}}^{\mathbf{G}} \pi_\mathbf{i} \mu_\mathbf{i} \qquad \text{y} \qquad \mathbf{\Sigma} = \mathbb{V}\mathbf{ar}(\mathbf{x}) = \sum_{\mathbf{i=1}}^{\mathbf{G}} \pi_\mathbf{i} \mathbf{\Sigma_i} + \sum_{\mathbf{i=1}}^{\mathbf{G}} \pi_\mathbf{i} (\mu_\mathbf{i} - \mu)(\mu_\mathbf{i} - \mu)'$$

   Sugerencia: utilizar la fórmula de esperanza condicionada.

9. Genere y represente 5000 datos normales con media $(0,0)'$ y matriz de varianzas-covarianzas

$$\begin{pmatrix} 1 & 0{,}8 \\ 0{,}8 & 1 \end{pmatrix}$$

   a) Represente la nube de puntos generada y estime con estos datos el vector de medias y la matriz de varianzas-covarianzas.

   b) Investigue sobre la función `mahalanobis` del paquete `mvtnorm`.

   Aplique a las distancias de Mahalanobis obtenida sobre una muestra test de 1000 datos, un test de bondad de ajuste de Kolmogorov-Smirnov para ver si se ajusta a una $\chi^2(2)$.

   c) Usando la función `quantile`, calcule una distancia d tal que el $10\,\%$ de los puntos dista del centro más que d. Represente la nube de puntos marcando de color rojo el $10\,\%$ de puntos más distantes del centro.

   d) Añade a los puntos generados un dato atípico. Repita las preguntas anteriores para la nueva muestra comparando los resultados obtenidos en ambos casos.