1 Análisis Multivariado I - Práctica 2 - Parte 2

Test de Hotelling para una muestra (continuación)

1. Las notas obtenidas por $n = 87$ estudiantes en un examen, el College Level Examination Program (CLEP) para la variable X_1 y el College Qualification Test (CQT) para las variables X_2 y X_3, están dadas en la tabla 2.3, con

   X_1 = ciencias sociales e historia
   X_2 = lengua
   X_3 = ciencias naturales

   (a) Construir Q-Q-plots de las distribuciones marginales de las variables X_1, X_2 y X_3. Construir también los *scatterplots* de todos los posibles pares de variables aleatorias. ¿Se podría decir que tienen distribución normal? (Es decir que $\mathbf{x}_i = (X_{i1}, X_{i2}, X_{i3}) \sim N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$).

   (b) Suponiendo que se contestó afirmativamente la respuesta anterior, hallar intervalos de confianza de nivel simultáneo 0.95 para μ_1, μ_2 y μ_3 por el método de Hotelling.

   (c) Hallar las direcciones principales y las longitudes de los ejes del elipsoide de confianza de nivel 0.95.

   (d) Calcular los intervalos de confianza de Bonferroni de nivel simultáneo 0.95 para μ_1, μ_2 y μ_3. Comparar las longitudes con las de los intervalos hallados en (b).

   (e) Supongamos que $(500, 50, 30)$ representan las notas promedio para miles de estudiantes en los últimos 10 años, testear la hipótesis $H_0 : \boldsymbol{\mu}^{\mathrm{T}} = (500, 50, 30)$ versus $H_1 : \boldsymbol{\mu}^{\mathrm{T}} \neq (500, 50, 30)$ a nivel 0.05. ¿Hay alguna razón para creer que el grupo de estudiantes cuyas notas figuran en la tabla 5.2 tiene un rendimiento distinto? Explicar.

   (f) Testear $H_0 : \boldsymbol{\mu} \in \mathcal{V} = \{\mathbf{z} \in \mathbb{R}^3 : z_1 = 10z_2 \text{ y } 2z_3 = z_2\}$ versus $H_1 : \boldsymbol{\mu} \notin \mathcal{V}$ con nivel 0.05.

   (g) Testear si todas las notas aumentaron (o disminuyeron) en la misma proporción con respecto al $(500, 50, 30)$.

2. Para los siguientes valores de $d = 2, 3, 4$ y nivel de significación $1 - \alpha = 0.95$, buscar el mínimo número de combinaciones lineales necesarias para que el método de Hotelling proporcione intervalos de confianza de nivel simultáneo α más cortos que el método de Bonferroni. Trabajar con $n = 25$ y $n = 100$.

3. Un educador musical llevó a cabo un estudio que involucró a miles de estudiantes en Finlandia. El objetivo del estudio era fijar normas nacionales referidas a la habilidad musical de los finlandeses. En la tabla 2.4 figuran estadísticas que resumen los datos obtenidos. Están basadas en 96 estudiantes en el último año escolar. Aún sin necesidad de suponer normalidad,

(a) Construir intervalos de confianza de nivel simultáneo y aproximado 90% para las medias (μ_i) de cada una de las variables ($1 \leq i \leq 7$).

(b) Basándose en datos muestrales que corresponden a estudiantes estadounidenses, el investigador podría haber supuesto que los escores medios de aptitud musical eran $\boldsymbol{\mu}_0 = (31, 27, 34, 31, 23, 22, 22)^{\mathrm{T}}$.

- ¿Serían estos valores posibles para los correspondientes valores medios finlandeses? Justificar.
- ¿Qué conclusión se hubiese podido sacar si cada componente de $\boldsymbol{\mu}_0$ hubiera pertenecido al intervalo de confianza respectivo calculado en (a)?

4. En EEUU. el gobierno federal exige que el Departamento de Control de Calidad de toda fábrica de hornos microondas monitoree la cantidad de radiación emitida cuando las puertas del horno están cerradas y cuando éstas están abiertas. Se observaron las radiaciones emitidas por 42 hornos elegidos al azar. Los datos aparecen en la tabla 2.5, con la puerta abierta y con la puerta cerrada.

(a) Hacer un Q-Q-plot con los datos univariados y además testear su normalidad.

(b) Una transformación de Box y Cox que mejora la normalidad de los datos para la puerta cerrada se obtiene con $\lambda = 0.25$. Aplicar la transformación a ambas variables $y_{ij} = x_{ij}^{1/4}$ $j = 1, 2$ y comprobarlo a través de nuevos Q-Q-plots

(c) Hallar $\overline{\mathbf{y}}$, $\mathbf{S}$ y $\mathbf{S}^{-1}$ para los datos transformados.

(d) Asumiendo que los datos transformados efectivamente siguen una distribución $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, hallar la elipse de confianza de nivel simultáneo 0.95, dar sus direcciones principales, la longitud de sus ejes y hacer un gráfico aproximado.

(e) Testear $H_0 : \boldsymbol{\mu} = (0.562, 0.589)^{\mathrm{T}}$ versus $H_1 : \boldsymbol{\mu} \neq (0.562, 0.589)^{\mathrm{T}}$ con nivel 0.05.

(f) Testear $H_0 : \boldsymbol{\mu} = (0.55, 0.60)^{\mathrm{T}}$ versus $H_1 : \boldsymbol{\mu} \neq (0.55, 0.60)^{\mathrm{T}}$ con nivel 0.05.

(g) Testear $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$ con nivel 0.05.

(h) Hallar intervalos de confianza simultáneos para μ_1, μ_2 y $\mu_1 - \mu_2$. Interpretarlos gráficamente a partir de la elipse.

5. Sean $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ una muestra aleatoria donde

$$\boldsymbol{\Sigma} = \sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix} = (1-\rho)\,\sigma^2 \mathbf{I}_d + \rho\sigma^2 \mathbf{1}_d \mathbf{1}_d^{\mathrm{T}}$$

con $\sigma^2 > 0$ y $\rho \in (-1, 1)$. Probar que los estimadores de máxima verosimilitud de σ^2 y ρ cumplen lo siguiente:

$$\begin{aligned} \widehat{\sigma}^2 &= \frac{tr(\mathbf{Q}/n)}{d} = \frac{1}{nd}\sum_{i=1}^{d} Q_{ii} \\ \widehat{\sigma}^2\widehat{\rho} &= \frac{\mathbf{1}_d^{\mathrm{T}}\left[\mathbf{Q}/n\right]\mathbf{1}_d - tr(\mathbf{Q}/n)}{d(d-1)} = \frac{1}{d(d-1)}\sum_{j=1}^{d}\sum_{i=1, i\neq j}^{d}\frac{Q_{ij}}{n} \end{aligned}$$

donde

$$\mathbf{Q} = \sum_{j=1}^{n} (\mathbf{x}_j - \overline{\mathbf{x}}) (\mathbf{x}_j - \overline{\mathbf{x}})^{\mathrm{T}} \ .$$

Sugerencias: Escribir la función de verosimilitud en términos de $a = (1-\rho)\sigma^2$ y $b = (1-\rho)\sigma^2 + d\rho\sigma^2$, hallar los EMV de a y b y luego obtener los EMV de σ^2 y ρ. Además, en caso de necesitarlo, recordar que

$$\begin{aligned}
(\mathbf{A} - \mathbf{v}\mathbf{v}^{\mathrm{T}})^{-1} &= \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{v}\mathbf{v}^{\mathrm{T}}\mathbf{A}^{-1}}{1 - \mathbf{v}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{v}} \\
(\mathbf{A} + \mathbf{v}\mathbf{v}^{\mathrm{T}})^{-1} &= \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{v}\mathbf{v}^{\mathrm{T}}\mathbf{A}^{-1}}{1 + \mathbf{v}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{v}} \\
\det(\mathbf{A} + \mathbf{v}\mathbf{v}^{\mathrm{T}}) &= \det(\mathbf{A}) \left(1 + \mathbf{v}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{v}\right)
\end{aligned}$$

6. Las alturas (en milímetros) de un hueso de la mandíbula de 20 chicos fue medida a los 8, $8\frac{1}{2}$, 9 y $9\frac{1}{2}$ años, y los resultados figuran en la tabla 2.6. El objetivo principal del estudio era establecer una tabla de crecimiento estándar para uso de los ortodoncistas.

   (a) Graficar las alturas medias muestrales en función de la edad. ¿Qué curva proporciona un aparente buen ajuste?

   (b) Suponiendo que los datos son una muestra $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \sim N_4(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, testear con nivel $\alpha = 0.05$ la hipótesis sugerida por (a) $H_0 : \mu_i = \beta_0 + \beta_1 t_i$, $1 \leq i \leq 4$, siendo $\mathbf{t} = (t_1, ..., t_4)^{\mathrm{T}}$ las edades a las cuales fueron tomadas las mediciones.

7. Una familia de distribuciones elípticas constituyen la mezcla de normales que se define como sigue: *Se dice que* $\mathbf{x} \sim \mathcal{MN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *si*

$$\mathbf{x} = \boldsymbol{\mu} + v^{1/2}\mathbf{z} \tag{1}$$

donde $v \geq 0$ *es una variable aleatoria independiente de* $\mathbf{z}$ *y* $\mathbf{z} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$.

Un caso particular de dichas distribuciones la constituyen las distribuciones $\mathcal{T}$ multivariada. Indicaremos $\mathbf{x} \sim \mathcal{T}_{p,k}$, si $\mathbf{x}$ tiene densidad

$$\frac{\Gamma\left(\frac{1}{2}(p+k)\right)}{\Gamma\left(\frac{1}{2}k\right)(k\pi)^{p/2}} \frac{1}{\left(1 + \frac{1}{k}\|\mathbf{x}\|^2\right)^{\frac{p+k}{2}}} \ .$$

Efectivamente, si en (**??**), v es tal que $k/v \sim \chi^2_k$, $\mu = \mathbf{0}$ y $\boldsymbol{\Sigma} = \mathbf{I}_p$, entonces $\mathbf{x} \sim \mathcal{T}_{p,k}$.

   (a) Mostrar que si $\mathbf{x} \sim \mathcal{MN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ y $\mathbb{E}v < \infty$ entonces

$$\mathbb{E}\mathbf{x} = \mu \qquad \text{y} \qquad \mathrm{Cov}(\mathbf{x}) = \mathbb{E}(v)\boldsymbol{\Sigma}$$

   (b) Si $\mathbf{x} \sim \mathcal{T}_{p,k}$, ¿qué condición debe cumplir k para que exista $\mathbb{E}v$?

   (c) Armar un programa en `R` para generar vectores con distribución $\mathcal{T}_{p,k}$.

   (d) Sean $\mathbf{z}_1, \ldots, \mathbf{z}_n \sim \mathcal{T}_{p,k}$. Defina

$$\mathbf{x}_i = \mu + \mathbf{C}\mathbf{z}_i$$

   donde $\boldsymbol{\Sigma} = \mathbf{C}\mathbf{C}^{\mathrm{T}}$.

Mostrar que $\mathbf{x}_i \sim \mathcal{MN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. En este caso diremos que $\mathbf{x}_i \sim \mathcal{T}_{p,k}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

(e) Armar una rutina en `R` para generar observaciones como en (d).

(f) Mostrar que si $\mathbf{x} = \boldsymbol{\mu} + \mathbf{C}\mathbf{z}$ donde $\boldsymbol{\Sigma} = \mathbf{C}\mathbf{C}^{\mathrm{T}}$ y $\mathbf{z} \sim \mathcal{T}_{p,k}$ entonces

$$\frac{1}{p}(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \mu) \sim \mathcal{F}_{p,k}$$

8. El siguiente es un experimento para saber qué sucede con $\overline{\mathbf{x}}$ y $\mathbf{S}$ como estimadores de $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$ cuando $\mathbf{x} \sim \mathcal{T}_{p,k}$ con $p = 3$.

   (a) i. Fije la semilla
   ii. genere $\mathbf{x}_i \sim \mathcal{T}_{p,k}$ $1 \leq i \leq n$ independientes.
   iii. Calcular $\overline{\mathbf{x}}$ y $\mathbf{S}$ para $k = 1, 2, 4, 10$ y $n = 10, 20, 50, 100, 200, 300, 400, 500$.
   iv. Calcular $D = \|\overline{\mathbf{x}} - \boldsymbol{\mu}\|^2$, $D_1 = \mathrm{tr}(\mathbf{S}\,\boldsymbol{\Sigma}^{-1})$ y $D_2 = \mathrm{tr}(\mathbf{S}\,\boldsymbol{\Sigma}^{-1}) / (\mathbb{E}v)$. Armar una tabla de doble entrada con los resultados obtenidos para las combinaciones de k y n consideradas en el ítem anterior.
   Observación: recuerde el resultado obtenido en el ejercicio 7 b).
   v. Para cada k, ¿qué observa?

   (b) Realice ahora 1000 replicaciones del experimento descripto en a), guarde los resultados obtenidos para cada muestra.
   i. Reporte el promedio de los mismos en un tabla para las distintas combinaciones de k y n.
   ii. Para cada uno de los valores de k, haga un plot donde grafique la evolución con n del promedio sobre las replicaciones de D_1.

9. Se quiere ahora testear $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ con $\boldsymbol{\mu}_0 = \mathbf{0}$ cuando $\mathbf{x}_i \sim \mathcal{T}_{p,k}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ usando el estadístico

$$T = n(\overline{\mathbf{x}} - \boldsymbol{\mu}_0)^{\mathrm{T}} \mathbf{S}^{-1} (\overline{\mathbf{x}} - \boldsymbol{\mu}_0)$$

   (a) Hacer una simulación para decidir qué resultado obtendría si se usara el percentil de T como si las observaciones fueran normales.
   Use $k = 1, 2, 4, 10$, $p = 2, 4$ y $n = 20$.

   (b) Armar un mecanismo bootstrap para testear H_0 en este caso.

   (c) Si $\boldsymbol{\Sigma}$ es conocida, ¿qué estadístico podría usar?

   Hint: Usar el ítem f) del ejercicio 7.

10. El archivo `ej10-P2-2.txt` contiene una muestra de 20 vectores de dimensión 2.

   (a) Testear la normalidad de este conjunto de datos. ¿A qué conclusión llega?

   (b) Basándonse en la conclusión anterior, realizar un test para testear $H_0 : \boldsymbol{\mu} = \mathbf{0}$.