1 Análisis Multivariado I - Práctica 2 - Parte 2

Test de Hotelling para una muestra (continuación)

Los ejercicios marcados en rojo no son para elegir para exponer, aunque deben hacerse.

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 3 y en el archivo `P2-2-ej1-2019.txt`, 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 3 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 4 y archivo `P2-2-ej3-2019.txt` 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 5 y en el archivo `P2-2-ej4-2019.txt`, 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

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

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 6 y en el archivo `P2-2-ej6-2019.txt`. 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. Dentro de la familia de distribuciones elípticas, las llamadas mezcla de normales juegan un rol importante. Dichas distribuciones se definen 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 (1), v es tal que $k/v \sim \chi^2_k$, $\mu = \mathbf{0}$ y $\mathbf{\Sigma} = \mathbf{I}_p$, entonces $\mathbf{x} \sim \mathcal{T}_{p,k}$.

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

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

(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 $\mathbf{\Sigma} = \mathbf{C}\mathbf{C}^{\text{T}}$.

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

(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 $\mathbf{\Sigma} = \mathbf{C}\mathbf{C}^{\text{T}}$ y $\mathbf{z} \sim \mathcal{T}_{p,k}$ entonces

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

8. El siguiente es un experimento para saber qué sucede con $\overline{\mathbf{x}}$ y $\mathbf{S}$ como estimadores de $\boldsymbol{\mu}$ y $\mathbf{\Sigma}$ 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 $D = \|\overline{\mathbf{x}} - \boldsymbol{\mu}\|^2$, $D_1 = \text{tr}(\mathbf{S}\,\mathbf{\Sigma}^{-1})$ y $D_2 = \text{tr}(\mathbf{S}\,\mathbf{\Sigma}^{-1})/\,(\mathbb{E}v)$. 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 $\mathbf{x}_i \sim \mathcal{T}_{p,k}(\boldsymbol{\mu}, \mathbf{\Sigma})$ usando el estadístico

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

(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.

10. El archivo `P2-2-ej10-2019.txt` contiene una muestra de 20 vectores de dimensión 2. Los resultados se muestran en la Tabla 7.

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

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

Alumno	X1	X2	X3	Alumno	X1	X2	X3
1	468	41	26	45	494	41	24
2	428	39	26	46	541	47	25
3	514	53	21	47	362	36	17
4	547	67	33	48	408	28	17
5	614	61	27	49	594	68	23
6	501	67	29	50	501	25	26
7	421	46	22	51	687	75	33
8	527	50	23	52	633	52	31
9	527	55	19	53	647	67	29
10	620	72	32	54	647	65	34
11	567	63	31	55	614	59	25
12	541	59	19	56	633	65	28
13	561	53	26	57	448	55	24
14	468	62	20	58	408	51	19
15	614	65	28	59	441	35	22
16	527	48	21	60	435	60	20
17	507	32	27	61	501	54	21
18	580	64	21	62	507	42	24
19	507	59	21	63	620	71	36
20	521	54	23	64	415	52	20
21	574	52	25	65	554	69	30
22	587	64	31	66	348	28	18
23	488	51	27	67	468	49	25
24	488	62	18	68	507	54	26
25	587	56	26	69	527	47	31
26	421	38	16	70	527	47	26
27	481	52	26	71	435	50	28
28	428	40	19	72	660	70	25
29	640	65	25	73	733	73	33
30	574	61	28	74	507	45	28
31	547	64	27	75	527	62	29
32	580	64	28	76	428	37	19
33	494	53	26	77	481	48	23
34	554	51	21	78	507	61	19
35	647	58	23	79	527	66	23
36	507	65	23	80	488	41	28
37	454	52	28	81	607	69	28
38	427	57	21	82	561	59	34
39	521	66	26	83	614	70	23
40	468	57	14	84	527	49	30
41	587	55	30	85	474	41	16
42	507	61	31	86	441	47	26
43	574	54	31	87	607	67	32
44	507	53	23				

Table 3: Tabla de datos de notas escolares. Corresponde a la Tabla 5.2 de Johnson y Wichern (1982)

	Media	Desvo estándar
x_1 = melodía	28.1	5.76
x_2 = harmonía	26.6	5.85
x_3 = tempo	35.4	3.82
x_4 = metro	34.2	5.12
x_5 = fraseo	23.6	3.76
x_6 = balance	22.0	3.93
x_7 = estilo	22.7	4.03

Table 4: Perfil de aptitud musical. Medias y desvos estándar para 96 estudiantes finlandeses que participaron en el programa de estandarización. Corresponde a la Tabla 5.5 de Johnson y Wichern (1982)

Horno	Radiación	Horno	Radiación	Horno	Radiación
Datos con la puerta abierta					
1	.30	16	.20	31	.10
2	.09	17	.04	32	.10
3	.30	18	.10	33	.10
4	.10	19	.01	34	.30
5	.10	20	.60	35	.12
6	.12	21	.12	36	.25
7	.09	22	.10	37	.20
8	.10	23	.05	38	.40
9	.09	24	.05	39	.33
l0	.10	25	.15	40	.32
11	.07	26	.30	41	.12
12	.05	27	.15	42	.12
13	.01	28	.09		
14	.45	29	.09		
15	.12	30	.28		
Datos con la puerta cerrada					
1	.15	16	.10	31	.10
2	.09	17	.02	32	.20
3	.18	18	.10	33	.11
4	.10	19	.01	34	.30
5	.05	20	.40	35	.02
6	.12	21	.10	36	.20
7	.08	22	.05	37	.20
8	.05	23	.03	38	.30
9	.08	24	.05	39	.30
l0	.10	25	.15	40	.40
11	.07	26	.10	41	.30
12	.02	27	.15	42	.05
13	.01	28	.09		
14	.10	29	.08		
15	.10	30	.18		

Table 5: Datos de radiación de hornos microondas. Corresponden a las Tablas 4.1 y 4.4 de Johnson y Wichern (1982)

	Edad en años			
Individuo	8	81/2	9	91/2
1	47.8	48.8	49.0	49.7
2	46.4	47.3	47.7	48.4
3	46.3	46.8	47.8	48.5
4	45.1	45.3	46.1	47.2
5	47.6	48.5	48.9	49.3
6	52.5	53.2	53.3	53.7
7	51.2	53.0	54.3	54.5
8	49.8	50.0	50.3	52.7
9	48.1	50.8	52.3	54.4
10	45.0	47.0	47.3	48.3
11	51.2	51.4	51.6	51.9
12	48.5	49.2	53.0	55.5
13	52.1	52.8	53.7	55.0
14	48.2	48.9	49.3	49.8
15	49.6	50.4	51.2	51.8
16	50.7	51.7	52.7	53.3
17	47.2	47.7	48.4	49.5
18	53.3	54.6	55.1	55.3
19	46.2	47.5	48.1	48.4
20	46.3	47.6	51.3	51.8
Media	48.655	49.625	50.570	51.450
D.E.	2.52	2.54	2.63	2.73

Table 6: Altura del hueso ramus de la mandíbula correspondiente a 20 chicos (en milímetros). Corresponden a la Tabla 3.4 de Seber (1984)

x_1	x_2
-0.241090036638898	-1.59608226009888
-3.7952633327004	2.98214882424911
-1.46016022598351	-1.38046349953979
0.379559191399193	-1.15452396785592
1.79055559112919	-3.48881466126336
0.280127907438186	-0.285273181282812
0.176867562641972	-0.28012743418403
-0.177515661157421	-0.850541001915607
-0.120468207289594	-0.0879665469155482
-0.292745498643899	0.346069084861567
-0.620660642729617	-0.71619202678777
-3.32643826698319	-2.07719367296236
0.480169735133903	-0.691875087102133
1.0052013805365	0.0112228044982784
0.777867607179416	-0.31831792510449
0.476982713741645	-0.691634068730137
-0.203653252841699	0.881124629352115
0.376822809921803	0.2675900508025
-0.573386456123506	0.364223460417591
0.981494910901135	-0.286601788598483

Table 7: Datos generados.