1 Análisis Multivariado - Práctica 2

1.1 Test de Hotelling para una muestra

1. Supongamos que los datos de la tabla 2.1 son una muestra aleatoria normal multivariada.

   (a) Testear la hipótesis de que el peso medio es 63 kg y la altura media es 1,60m.

   (b) Hallar un elipsoide de confianza de nivel 95% para el peso y la altura medios de los indios peruanos.

2. Dada $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n \sim N_d(\mu, \Sigma)$ muestra aleatoria sea $T^2_{\mathbf{x}}$ el estadístico de Hotelling para testear $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$. Consideremos la siguiente transformación de los datos, sean $\mathbf{y}_i = A\mathbf{x}_i + \mathbf{b}$ con $A \in \mathbb{R}^{d\times d}$ inversible fija y $\mathbf{b} \in \mathbb{R}^{d\times 1}$, sea $T^2_{\mathbf{y}}$ el estadístico de Hotelling para testear $H_0 : \mu_Y = A\mu_0 + \mathbf{b} = \mathbf{0}$, probar que entonces $T^2_{\mathbf{x}} = T^2_{\mathbf{y}}$.

3. Sea $\mathbf{x}_1, \ldots, \mathbf{x}_n$ una m.a. con distribución $N_d(\mu, \Sigma)$, dada $C \in \mathbb{R}^{q\times d}$ con $rg(C) = q$ y $\mathbf{b} \in \mathbb{R}^{q\times 1}$, encontrar un test de Hotelling para testear $H_0 : C\mu = \mathbf{b}$.

4. Sea $\mathbf{x}_1, \ldots, \mathbf{x}_n$ una m.a. con distribución $N_d(\mu, \Sigma)$ y sea $K \in \mathbb{R}^{d\times q}$ con $q < d$ y $rg(K) = q$. Se quiere testear la hipótesis $H_0 : \exists\, \beta \in \mathbb{R}^{q\times 1}$ tal que $\mu = K\beta$. Interpretar el significado de esta hipótesis. Mostrar que H_0 puede escribirse como $A\mu = \mathbf{0}$ para cierta matriz A y por lo tanto puede testearse usando un test de Hotelling.

5. Sean $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ i.i.d. $N_2(\mu, \Sigma)$ con $\mu' = (\mu_1, \mu_2)$ y sea $y_i = x_{i1} - x_{i2}$ $(1 \leq i \leq n)$. Mostrar que el test T^2 de Hotelling para testear $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$ es equivalente al test usual para muestras apareadas basado en el estadístico $\bar{y}\sqrt{n}/s_y$.

6. Consideremos los primeros 14 datos de la tabla 2.2. Suponiendo que son una muestra normal multivariada:

   (a) Testear $H_0 : \mu = (45, 42, 45, 42)'$.

   (b) Testear la hipótesis de que las medias de los pesos son iguales en las cuatro direcciones.

   (c) Testear la hipótesis de que las medias de los pesos en las direcciones norte y sur son iguales y en las direcciones este y oeste también. Comparar con el resultado obtenido en el inciso (a).

   (d) Encontrar intervalos de confianza simultáneos de nivel 95% para $\mu_1 - \mu_3$ y $\mu_2 - \mu_4$. ¿Qué método conviene usar?

   (e) Construir intervalos de confianza simultáneos de nivel 95% para todos los contrastes $\mathbf{c}'\mu$, $\sum_{i=1}^4 c_i = 0$. Probar que existe un $\mathbf{c}$ tal que 0 no pertenece al intervalo de confianza para $\mathbf{c}'\mu$; intentar hallarlo explícitamente.

7. Sean $\phi_i = \mu_i - \mu_d$ $(1 \leq i \leq d-1)$. Mostrar que el conjunto de todas las combinaciones lineales $\sum_{i=1}^{d-1} h_i\phi_i$ es equivalente al conjunto de contrastes $\sum_{i=1}^d c_i\mu_i$ $\left(\sum_{i=1}^d c_i = 0\right)$.

8. 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 $\mathbf{x}_i = (X_{i1}, X_{i2}, X_{i3}) \sim N_3(\mu, \Sigma)$).

   (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 $H_0 : \mu' = (500, 50, 30)$ versus $H_1 : \mu' \neq (500, 50, 30)$ a nivel 0.05. ¿Hay alguna razon para creer que el grupo de estudiantes cuyas notas figuran en la tabla 5.2 tiene un rendimiento distinto? Explicar.

   (f) Testear $H_0 : \mu \in V = \{\alpha \in \mathbb{R}^3 : \alpha_1 = 10\alpha_2 \text{ y } 2\alpha_3 = \alpha_2\}$ versus $H_1 : \mu \notin V$ con nivel 0.05

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

9. Para los siguientes valores de $d = 2, 3, 4$ y nivel de significación $\alpha = 0.95$, 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 $n = 100$.

10. 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 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 $\mu_0 = (31, 27, 34, 31, 23, 22, 22)'$. ¿Serían estos valores posibles para los correspondientes valores medios finlandeses? Justificar. ¿Qué conclusión se hubiese podido sacar si cada componente de μ_0 hubiera pertenecido al intervalo de confianza respectivo calculado en (a)?

11. 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 $\bar{\mathbf{y}}$,S, S^{-1} para los datos transformados.

    (d) Asumiendo que los datos transformados efectivamente siguen una distribución $N_2(\mu, \Sigma)$,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 $H_0 : \mu' = (0.562, 0.589)$ versus $H_1 : \mu' \neq (0.562, 0.589)$ con nivel 0.05.

    (f) Testear $H_0 : \mu' = (0.55, 0.60)$ versus $H_1 : \mu' \neq (0.55, 0.60)$ 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.

12. Sean $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n \sim N_d(\mu, \Sigma)$ una muestra aleatoria con Σ satisfaciendo:

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}$$

    con $\sigma^2 > 0$ y $\rho \in (-1, 1)$. Sea $Q_H = \sum_{i=1}^{n} \sum_{j=1}^{d} (\bar{x}_{.j} - \bar{x}_{..})^2$ y
    $Q_E = \sum_{i=1}^{n} \sum_{j=1}^{d} (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}_{..})^2$, donde x_{ij} es el j-ésimo elemento de $\mathbf{x}_i$. Mostrar que $H_0 : \mu_1 = \mu_2 = \cdots = \mu_d$ se puede testear usando $F = (n-1) Q_H / Q_E$, donde $F \sim F_{d-1,(n-1)(d-1)}$ cuando H_0 es verdadera.
    Sugerencia: usar que si $\mathbf{y} \sim N_d(\mu, \Sigma)$, $A \in \mathbb{R}^{d \times d}$ simétrica de rango r, tenemos la siguiente equivalencia: $\mathbf{y}' A \mathbf{y} \sim \chi_r^2 \iff A\Sigma A = A$).

13. Sean $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n \sim N_d(\mu, \Sigma)$ una muestra aleatoria con Σ satisfaciendo:

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix} = (1-\rho)\sigma^2 I_d + \rho\sigma^2 \mathbf{1}_d \mathbf{1}_d'$$

    con $\sigma^2 > 0$ y $\rho \in (-1, 1)$ y $\mathbf{1}_d$ es un vector de unos. Probar que los estimadores de máxima verosimilitud de σ^2 y ρ cumplen lo siguiente:

$$\hat{\sigma}^2 = \frac{tr(Q/n)}{d} = \frac{1}{nd} \sum_{i=1}^{d} Q_{ii}$$

$$\hat{\sigma}^2\hat{\rho} = \frac{\mathbf{1}_d'\left[Q/n\right]\mathbf{1}_d - tr(Q/n)}{d(d-1)} = \frac{1}{d(d-1)}\sum_{j=1}^{d}\sum_{i=1,i\neq j}^{d}\frac{Q_{ij}}{n}$$

donde $Q = \sum_{j=1}^{n}\left(\mathbf{x}_j - \bar{\mathbf{x}}\right)\left(\mathbf{x}_j - \bar{\mathbf{x}}\right)'$.
Sugerencias:

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

14. 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 $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n \sim N_4\left(\mu, \Sigma\right)$, 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 $\mathbf{t} = (t_1, ...t_4)'$ las edades a las cuales fueron tomadas las mediciones.