1 Análisis Multivariado I - Práctica 3

1.1 Test de Hotelling para dos muestras

1. En la primera fase de un estudio sobre el costo de transporte de la leche desde las granjas hasta las lecherías, se tomó una muestra de empresas de transporte vinculadas al transporte de lácteos. En la tabla 3.1 se presentan los datos de costos por milla de

$$\begin{aligned} X_1 &= \text{combustible} \\ X_2 &= \text{reparaciones} \\ X_3 &= \text{capital} \end{aligned}$$

para $n_1 = 36$ camiones nafteros y $n_2 = 23$ camiones a diesel.

   (a) Testear si hay diferencias entre los vectores de costos medios. Tomar $\alpha = 0.01$.

   (b) Si la hipótesis de igual vector de costos medios es rechazada en la parte (a), hallar la combinación lineal de las componentes de las medias que es más responsable del rechazo.

   (c) Construir intervalos de confianza de nivel simultáneo 0.99 para los pares de costos medios. Si los hay ¿Qué costos aparecen como muy distintos?

   (d) Comentar la validez de los supuestos realizados.

2. Grogan y Smith (1981) describen dos especies de mosquitos (midges): *Amerohelea fasciata (Af)* y *A. pseudofasciata (Apf)* recientemente descubiertas. En la tabla 3.2 aparecen los datos correspondientes a las mediciones de la longitud de las antenas y de las alas de nueve insectos Af y seis Apf. Definamos dos nuevas variables Y_1=*longitud de las antenas + longitud de las alas* e Y_2=*longitud de las alas.*

   (a) Calcular el estadístico T^2 de Hotelling para testear la igualdad entre los vectores de medias de ambos grupos, basado en los $\mathbf{y}$. Verificar que se obtiene el mismo valor que el del estadístico de Hotelling si uno usaba las variables originales. (Comparar con el ejercicio 2 de la práctica 2). Concluir si se acepta o rechaza la hipótesis.

   (b) Mostrar que en los tests de t univariados de nivel de significación 0,05 realizados sobre cada variable por separado se aceptaría la hipótesis de igualdad de medias.

   (c) Realizar un *scatterplot* de Y_1 versus Y_2 para los datos, marcando los puntos correspondientes a cada grupo con símbolos distintos, y explicar cómo puede suceder que en los tests univariados la hipótesis de la igualdad de medias es aceptada y sin embargo, en el caso multivariado, la hipótesis de igualdad de medias es claramente rechazada.

   (d) Dibujar la elipse de confianza de nivel 98% para el vector de diferencia de medias y mostrar que no cubre al vector $\mathbf{0}$. En el mismo gráfico, dibujar un rectángulo correspondiente a los intervalos de confianza univariados de nivel 99% ($0.99^2 \cong 0.98$) para las diferencias entre las medias de Y_1 e Y_2 . ¿Cómo se puede ver en este gráfico que es mejor realizar un test de Hotelling multivariado que dos tests de t univariados?

3. Como parte de un estudio sobre el amor y el matrimonio a una muestra de maridos y esposas se les pidió que respondieran a estas preguntas:

   i. ¿Cuál es el nivel de amor apasionado que siente por su pareja?

   ii. ¿Cuál es el nivel de amor apasionado que su pareja siente por ud.?

   iii. ¿Cuál es el nivel de sentimiento de compañerismo que siente por su pareja?

   iv. ¿Cuál es el nivel de sentimiento de compañerismo que su pareja siente por ud.?

   Las respuestas se registraron en una escala de 5 puntos:
   1. nada
   2. poco
   3. algo
   4. bastante
   5. mucho

   Las respuestas de 30 parejas figuran en la tabla 3.3, donde X_i = respuesta (en la escala de 1-5) para la pregunta i-ésima.

   (a) Graficar los valores medios de hombres y mujeres como perfiles muestrales.

   (b) ¿Es el perfil para hombres paralelo al perfil para mujeres? Testear si los perfiles son paralelos con $\alpha = 0.05$. Si no se rechaza la hipótesis de paralelismo, testear si los perfiles son coincidentes, para el mismo nivel de significancia. Finalmente, si los perfiles son coincidentes, testear si el perfil común está nivelado, es decir si todas las variables tienen la misma media (siempre con $\alpha = 0.05$.)

4. Consideremos los datos que aparecen en la tabla 3.4 que corresponden a los resultados de tomar un test de habilidad sicolingual a dos grupos de 27 chicos de edades 8-9 años. Los primeros corresponden a chicos con una enfermedad neonatal (TNT) y los segundos a chicos normales que forman el grupo control. Interesa estudiar las siguientes situaciones:
   H_{01}: los dos perfiles son similares
   H_{02}: los dos perfiles están al mismo nivel
   H_{03}: no hay diferencias entre las medias de los tests

   (a) Expresar las hipótesis anteriores matemáticamente.

   (b) Graficar (en un mismo gráfico) las medias de cada grupo en función de las variables, es decir los perfiles.

   (c) Testear H_{01}.

   (d) En caso de no rechazarse H_{01}, testear H_{03} e interpretar el significado de $H_{01} \cap H_{03}$. En el caso de rechazarse H_{01}, testear H_{02} e interpretar el significado de testear ésto.

1.2 Inferencia para matrices de covarianza

1. Consideremos el modelo de regresión lineal

$$y_i = \theta_0 + \mathbf{x}_i^{\mathrm{T}}\theta + \varepsilon_i$$

donde $\mathbf{x}_i$ m.a. $N_d(\mu, \Sigma)$ independientes de ε_i m.a. $N(0, \sigma^2)$.

   (a) ¿Qué distribución tiene $\mathbf{z}_i = (\mathbf{x}_i^{\mathrm{T}}, y_i)^{\mathrm{T}}$?

   (b) ¿Qué expresión tiene (en términos de $\mathbf{x}_i^{\mathrm{T}}$ e y_i) el estadístico del test del cociente de verosimilitud para la independencia de bloques aplicado a las $\mathbf{z}_i$?

   Comentario: el test de F usual para $H_0 : \theta = \mathbf{0}$ es equivalente al test del cociente de verosimilitud anterior.

2. Dada un muestra $N_d(\mu, \Sigma)$ mostrar que el estadístico del cociente de verosimilitud para testear $H_0 : \Sigma = \sigma^2 \mathrm{I}$ cumple

$$l^{2/(nd)} = \frac{|Q|^{1/d}}{tr(Q)/d}.$$

   Escribirlo en función de los autovalores e interpretar. ¿Cuál es la distribución asintótica de $-2\ln\ell$ bajo H_0?

3. Consideremos una muestra $\mathbf{x}_1, \ldots, \mathbf{x}_n \sim N_d(\mu, \Sigma)$. Hallar el estadístico del test del cociente de verosimilitud para las hipótesis $H_0 : \Sigma = \sigma^2(1-\rho)\mathrm{I} + \sigma^2\rho\mathbf{1}_d\mathbf{1}_d^{\mathrm{T}}$ versus $H_1 : \nexists\sigma^2$ ó $\rho \in (0,1)$ tal que $\Sigma = (1-\rho)\mathrm{I} + \sigma^2\rho\mathbf{1}_d\mathbf{1}_d^{\mathrm{T}}$.

4. Sea $\mathbf{x}_1, \ldots, \mathbf{x}_n$ una muestra de una $N_d(\mu, \Sigma)$. Encontrar el test basado en el principio de unión–intersección para $H_0 : \Sigma = \Sigma_0$. Explicar cómo se podrían calcular los valores críticos del test.

5. A partir de una muestra $N_d(\mu, \Sigma)$ encontrar intervalos de confianza para $\mathbf{v}^{\mathrm{T}}\Sigma\mathbf{v}$ de nivel simultáneo $1-\alpha$ para todo $\mathbf{v} \neq \mathbf{0}$.

6. La tabla 3.5 contiene información fisonómica de hermanos. Las variables son X_1 = longitud de la cabeza del primer hijo, X_2 = ancho de la cabeza del primer hijo, X_3 = longitud de la cabeza del segundo hijo, y X_4 = ancho de la cabeza del segundo hijo; referidas a 25 familias distintas. Consideremos el siguiente vector aleatorio:

$$\mathbf{y} = (X_1 + X_3, X_2 + X_4, X_1 - X_3, X_2 - X_4)$$

   Testear si las primeras dos coordenadas de $\mathbf{y}$ son independientes de las segundas dos, a nivel 0.05 con el test de máxima verosimilitud y con el test basado en el principio de unión-intersección. Comparar las conclusiones.

7. Sea $\mathbf{x}_1, \ldots, \mathbf{x}_n$ una m.a. $N_d(\mu, \Sigma)$. Se desea testear que las d variables son independientes, es decir

$$H_0 : \Sigma = \mathrm{diag}(\Sigma_{11}, \Sigma_{22}, \ldots, \Sigma_{dd}).$$

Consideremos la matriz de correlación muestral R cuyos elementos son

$$R_{jk} = \frac{\hat{\Sigma}_{jk}}{\left(\hat{\Sigma}_{jj}\hat{\Sigma}_{kk}\right)^{1/2}} = \frac{Q_{jk}}{\left(Q_{jj}Q_{kk}\right)^{1/2}}.$$

(a) Probar que el estadístico del test del cociente de verosimilitud es $\Lambda = |R|^{n/2}$.

(b) ¿Cuál es la distribución asintótica de $-2\ln\Lambda$ bajo H_0?

1.3 Test para igualdad de covarianza entre dos poblaciones

1. En el ejercicio 1 de la Sección 1.1 se estudiaba el costo de transporte de la leche desde las granjas hasta las lecherías para $n_1 = 36$ camiones nafteros y $n_2 = 23$ camiones a diesel.

   (a) Para testear si haba diferencias entre los vectores de costos medios, se supuso que las dos poblaciones tenían igual matriz de covarianza. Es este supuesto razonable? Tomar $\alpha = 0.01$.

   (b) Si la hipótesis de igualdad de matrices de covarianzas es rechazada en la parte (a), como testearía la igualdad de vectores medios?

2. En el ejercicio 4 de la Sección 1.1 se estudiaban los resultados de tomar un test de habilidad sicolingual a dos grupos de 27 chicos de edades 8-9 años. Nos interesa estudiar las siguientes situaciones:
   H_{01}: los dos perfiles son similares
   H_{02}: los dos perfiles están al mismo nivel
   H_{03}: no hay diferencias entre las medias de los tests

   (a) Para cada una de las hipótesis anteriores, es necesario suponer que las matrices de covarianza del vector $\mathbf{x} = (x_1, \ldots, x_{10})^{\mathrm{T}}$ donde
   x_1 = recepción auditiva
   x_2 = recepción visual
   x_3 = memoria visual
   x_4 = asociación auditiva
   x_5 = memoria auditiva
   x_6 = asociación visual
   x_7 = oclusión visual
   x_8 =expresión oral
   x_9 = oclusión gramatical
   x_{10} = destreza manual
   en las dos poblaciones es la misma o puede hacer un supuesto más débil? Exprese matemáticamente la hipótesis sobre la igualdad de covarianzas que sea necesaria en cada caso.

   (b) Para cada una de las situaciones de interés estudie si el supuesto de igualdad de covarianzas es razonable? Tomar $\alpha = 0.01$.