1 Análisis Multivariado I - Práctica 2 - Parte 1

Test de Hotelling para una muestra

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

Definición: Se define el elipsoide como el conjunto de puntos $\mathbf{x} \in \mathbb{R}^d$ tales que

$$(\mathbf{x}-\mathbf{b})^{\mathrm{T}}\mathbf{A}(\mathbf{x}-\mathbf{b}) = c^2 \tag{1}$$

donde $\mathbf{A}$ es una matriz definida positica y $\mathbf{b}$ es el centro del elipsoide.
Si $d = 2$ se llama elipse.

1. En este ejercicio graficaremos elipses. Consideremos la definición dada en (1).

   (a) Supongamos que $\mathbf{b} = (0,0)^{\mathrm{T}}$ y $\mathbf{A} = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$. Desarrollar (1) en este caso y graficar en `R` el caso $\lambda_1 = 2$, $\lambda_2 = 3$ y $c = 1$ usando dos funciones (función superior y función inferior) o coordenadas polares.

   (b) Considerar el caso $\mathbf{A}$ simétrica y definida positiva. Usar la descomposición espectral para transformar el problema $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} = c^2$ en

   $$\mathbf{y}^{\mathrm{T}} \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \mathbf{y} = c^2$$

   Usar la transformación para graficar en `R` la elipse que tiene $\mathbf{A} = \begin{pmatrix} 2.5 & -0.5 \\ -0.5 & 2.5 \end{pmatrix}$ y $c = 2$.

   (c) ¿Cómo se modifica el ítem anterior si ahora tenemos $\mathbf{b}$ genérico? Graficar en `R` la elipse anterior pero ahora centrada en $\mathbf{b} = (1,3)^{\mathrm{T}}$.

2. Supongamos que los datos de la Tabla 1 son una muestra aleatoria normal multivariada (ver archivo `P2-1-ej2-2019.txt`).

   (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.

   (c) Use el test de normalidad de Shapiro–Wilks multivariado para ver si el supuesto de normalidad es razonable. ¿Si rechaza la hipótesis de normalidad a qué se lo atribuiría? ¿Qué haría?

3. Dada una muestra aleatoria $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, sea $T^2_{\mathbf{x}}$ el estadístico de Hotelling para testear $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ versus $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$.

Consideremos la siguiente transformación de los datos. Sean

$$\mathbf{y}_i = \mathbf{A}\mathbf{x}_i + \mathbf{b}$$

con $\mathbf{A} \in \mathbb{R}^{d\times d}$ inversible fija y $\mathbf{b} \in \mathbb{R}^{d\times 1}$, sea $T^2_{\mathbf{y}}$ el estadístico de Hotelling para testear $H_0 : \boldsymbol{\mu}_{\mathbf{y}} = \mathbf{A}\boldsymbol{\mu}_0 + \mathbf{b}$, donde $\boldsymbol{\mu}_{\mathbf{y}} = \mathbb{E}(\mathbf{y}_1)$. Probar que $T^2_{\mathbf{x}} = T^2_{\mathbf{y}}$.

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

5. Sea $\mathbf{x}_1, \ldots, \mathbf{x}_n$ una m.a. con distribución $N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ y sea $\mathbf{K} \in \mathbb{R}^{d\times q}$ con $q < d$ y rango $(\mathbf{K}) = q$. Se quiere testear la hipótesis

$$H_0 : \exists\, \boldsymbol{\beta} \in \mathbb{R}^{q\times 1} \quad \text{tal que} \quad \boldsymbol{\mu} = \mathbf{K}\boldsymbol{\beta}\,.$$

   (a) Interpretar el significado de esta hipótesis.

   (b) Mostrar que H_0 puede escribirse como $\mathbf{A}\boldsymbol{\mu} = \mathbf{0}$ para cierta matriz $\mathbf{A}$ y por lo tanto, puede testearse usando un test de Hotelling.

6. Sean $\mathbf{x}_1, \ldots, \mathbf{x}_n$ i.i.d. $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ con $\boldsymbol{\mu}^{\mathrm{T}} = (\mu_1, \mu_2)$ y sea $y_i = x_{i1} - x_{i2}$ $(1 \le i \le n)$. 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 $\bar{y}\sqrt{n}/s_y$.

7. Sean $\phi_i = \mu_i - \mu_d$ $(1 \le i \le d-1)$. 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. Consideremos los 28 datos de la Tabla 2 dados en el archivo `P2-1-ej8-2019.txt` que corresponden al peso del espesor de la capa de corcho medido para los cuatro puntos cardinales en 28 árboles.

   (a) Realizar un test para estudiar el supuesto de normalidad multivariada de los datos.

   (b) Suponiendo que son una muestra normal multivariada:

      i. Testear $H_0 : \boldsymbol{\mu} = (45, 42, 45, 42)^{\mathrm{T}}$.

      ii. Testear la hipótesis de que las medias de los pesos son iguales en las cuatro direcciones.

      iii. 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).

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

      v. Construir intervalos de confianza simultáneos de nivel 95% para todos los contrastes $\mathbf{c}^{\mathrm{T}}\boldsymbol{\mu}$ donde $\sum_{i=1}^{4} c_i = 0$. Probar que existe $\mathbf{c}$ tal que 0 no pertenece al intervalo de confianza para $\mathbf{c}^{\mathrm{T}}\boldsymbol{\mu}$. Intentar hallarlo explícitamente.

9. Un proceso industrial fabrica elementos cuyas características de calidad se miden por un vector $\mathbf{x} \in \mathbb{R}^3$ que suponemos tiene distribución normal $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Cuando el proceso está en control, los valores esperados de las variables deben ser $\boldsymbol{\mu}_0 = (12, 4, 2)^{\mathrm{T}}$. Para comprobar si el proceso funciona adecuadamente, se tomó una muestra de 20 elementos y se midieron las 3 características de calidad, obteniendose un promedio de

$$\overline{\mathbf{x}} = (11.5, 4.3, 1.2)^{\mathrm{T}}$$

mientras que la matriz de covarianza muestral y su inversa son

$$\mathbf{S} = \begin{pmatrix} 10.53 & 4.21 & -5.26 \\ 4.21 & 12.63 & -3.16 \\ -5.26 & -3.16 & 4.21 \end{pmatrix} \qquad \mathbf{S}^{-1} = \begin{pmatrix} 0.2531 & -0.0065 & 0.3113 \\ -0.0065 & 0.0976 & 0.0652 \\ 0.3113 & 0.0652 & 0.6755 \end{pmatrix}$$

donde redondeamos los números por simplicidad.

(a) Si observa cada variable por separado puede determinar si el proceso está fuera de control con nivel 5%, es decir, testee $H_{0,j} : \mu_j = \mu_{0,j}$ versus $H_{1,j} : \mu_j \neq \mu_{0,j}$ para $1 \leq j \leq 3$. Describa claramente

i) cuales son los supuestos que está haciendo,

ii) que hipótesis está testeando,

iii) que estadístico está utilizando,

iv) el $p-$valor obtenido y la conclusión que saca.

(b) Consideremos ahora que analiza el proceso mirando conjuntamente las 3 variables, es decir interesa es decir, $H_0 : \boldsymbol{\mu} = (12, 4, 2)^{\mathrm{T}}$. Qué test utilizaría? Indique claramente

i) cuales son los supuestos que está haciendo,

ii) que hipótesis está testeando,

iii) que estadístico está utilizando,

iv) el $p-$valor obtenido y la conclusión que saca

v) Si rechaza la hipótesis nula, cual sería la dirección que permite observar dicha diferencia?

(c) Qué concluye de a) y b)? Son contradictorias las conclusiones? Explique porqué responde por si o por no?

Para explicar calcule la matriz de correlación.

Peso	Altura	Peso	Altura
71.0	1629	56.5	1569
56.0	1561	61.0	1619
65.0	1566	62.0	1639
53.0	1494	53.0	1568
65.0	1540	57.0	1530
66.5	1622	59.1	1486
64.0	1578	69.5	1645
64.0	1648	56.5	1521
57.0	1547	55.0	1505
57.0	1473	58.0	1538
59.5	1513	61.0	1653
57.0	1566	57.5	1580
74.0	1647	72.0	1620
62.5	1637	68.0	1528
63.4	1647	68.0	1605
69.0	1625	73.0	1615
64.0	1640	65.0	1610
71.0	1572	60.2	1534
55.0	1536	70.0	1630
87.0*	1542*		

Table 1: Tabla de Peso y altura de 39 indgenas Peruanos. El dato marcado con * se sospecha como outlier. Corresponde a la Tabla 3.1 de Seber (1984)

N	E	S	O	N	E	S	O
72	66	76	77	91	79	100	75
60	53	66	63	56	68	47	50
56	57	64	58	79	65	70	61
41	29	36	38	81	80	68	58
32	32	35	36	78	55	67	60
30	35	34	26	46	38	37	38
39	39	31	27	39	35	34	37
42	43	31	25	32	30	30	32
37	40	31	25	60	50	67	54
33	29	27	36	35	37	48	39
32	30	34	28	39	36	39	31
63	45	74	63	50	34	37	40
54	46	60	52	43	37	39	50
47	51	52	43	48	54	57	43

Table 2: Peso (en centigramos) del espesor de la capa de corcho medido para los cuatro puntos cardinales en 28 árboles. Corresponde a Tabla 3.3 de Seber (1984).