Pontificia Universidad Javeriana
Parcial I de Análisis Multivariado

22 de Agosto de 2024

Tiempo: 1 hora y 45 min

Nombre: _____ CC/TI: _____ Nota: _____

1. En el archivo `DataP1.csv` contiene los datos de una muestra de vectores aleatorios de 5 componentes, $X_1, X_2, ..., X_{60}$.

   a. ***(0.30 pts)*** ¿El vector de medias $\boldsymbol{\mu} = [0.71, -0.65, 0.94, -1.35, 0.83]^\top$ se encuentra en una región de confianza del tipo T^2 del 90 %?

   b. ***(0.60 pts)*** Usando un nivel de confianza 95 %, construya todos los intervalos de confianza de tipo T^2 para todas las diferencias entre dos medias poblacionales, $\mu_i - \mu_k$ con $i \neq k = 1, 2, 3$. Con base en los intervalos contruídos, ¿hay evidencia de que NO existe diferencias significativa entre las medias poblacionales?.

   c. ***(0.40 pts)*** Usando un nivel de confianza 95 %, construya todos los intervalos de confianza de tipo Bonferroni para todas las medias poblacionales, μ_k con $k = 1, 2, 3, 4, 5$.

2. Los datos `Iris` del paquete `datasets` contiene información sobre mediciones de 3 especies de la flor **Iris**. Las especies son: setosa, versicolor y virginica.

   - **sepal length**: Longitud de sepalo en *cm*
   - **sepal width**: Ancho de sepalo en *cm*
   - **petal length**: Longitud de petalo en *cm*
   - **petal width**: Ancho de petalo en *cm*

   Con base en la información dada en **Iris** responda las siguientes preguntas:

   a. ***(0.20 pts)*** Calcule el vector de medias y la matriz de covarianzas muestrales para todos los datos y para cada especie.

   b. ***(0.50 pts)*** Evalue la normalidad multivariada para cada especie usando gráficos normal y chi-cuadrado. De su análisis gráfico qué concluye?

   c. ***(0.50 pts)*** Verifique si hay outliers (multivariados) e identifíquelos para cada especie.

3. Los datos `Protein` del paquete `MultBiplotR` contiene información sobre datos nutricionales de 9 diferentes fuentes de proteínas para los habitantes de 25 países europeos alrededor de 1970:

   - **Comunist**: Sí el país es comunista o no
   - **Region**: Tres regiones Norte Centro Sur
   - **RedMeat**: Consumo de proteínas provenientes de carnes rojas

- **WhiteMeat**: Consumo de proteínas provenientes de carnes blancas;
- **Eggs**: Consumo de proteínas del huevo;
- **Milk**: Consumo de proteínas de la leche;
- **Fish**: Consumo de proteínas provenientes del pescado;
- **Cereals**:Consumo de proteínas procedentes de cereales;
- **Starch**: Consumo de proteínas provenientes de carbohidratos;
- **Nuts**: Consumo de proteínas procedentes de cereales, frutos secos y semillas oleaginosas;
- **FruitVeg**: Consumo de proteínas procedentes de frutas y verduras.

Estos datos fueron colectados incialmente para entender las diferencias nutricionales entre los paises europeos.

a. ***(0.625 pts)*** ¿Existen diferencias estadísticamente significativas en las medias de las variables de las regiones? Plante la hipótesis, escriba y verifique los supuestos requeridos.

b. ***(0.625 pts)*** ¿Existen diferencias estadísticamente significativas entre las medias de los paises comunistas y no comunistas? Plante la hipótesis, escriba y verifique los supuestos requeridos.

4. Manly *et al* (1980) [1] presentaron datos sobre las principales medidas de la mandíbula (en mm): anchura de la mandíbula (X1), altura de la mandíbula debajo del primer molar (X2), longitud del primer molar (X3), anchura del primer molar (X4), longitud del primer al tercer molar (X5), longitud del primer al cuarto premolar (X6); de 7 grupos de perros. El objetivo principal era comparar perros tailandeses prehistóricos (Prehistoric dog) con los otros 6 grupos de animales (Modern dog, Golden jackal, Chinese wolf, Indian wolf, Cuon Dingo).

   a. ***(0.3125 pts)*** En la figura de abajo se presentan las caras de Chernoff para los 7 grupos de caninos. Con base en esta figura, ¿cuál de los grupos de perros podría considerarse como el más parecido a Prehistoric dog? justifique su respuesta.

   b. ***(0.3125 pts)*** Con base en la matriz de covarianzas muestrales,

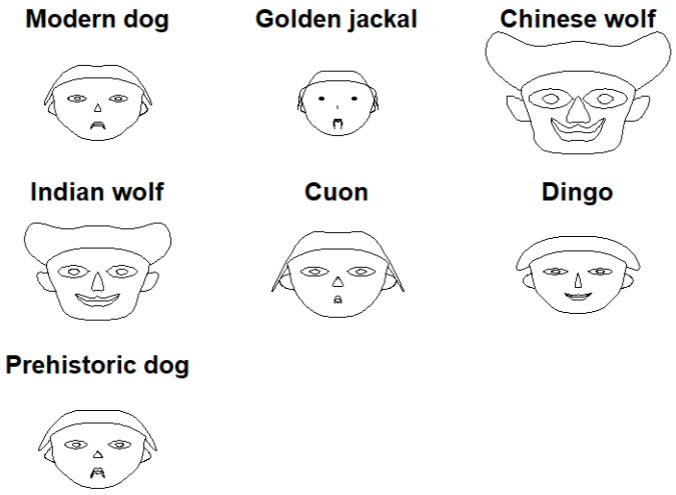
$$S = \begin{bmatrix} 2.88 & 5.25 & 4.85 & 1.93 & 6.53 & 7.74 \\ 5.25 & 10.56 & 8.90 & 3.59 & 11.46 & 15.58 \\ 4.85 & 8.90 & 9.61 & 3.51 & 13.43 & 16.31 \\ 1.93 & 3.59 & 3.51 & 1.36 & 4.86 & 5.92 \\ 6.53 & 11.46 & 13.43 & 4.86 & 24.36 & 24.68 \\ 7.74 & 15.58 & 16.31 & 5.92 & 24.68 & 31.52 \end{bmatrix}$$

   ¿usted diria que los perros son homogéneos o heterogeneos? en caso de ser heterogéneos ¿en qué variable sería más predominante dicha heterogeneidad?

   c. ***(0.3125 pts)*** Los resultados en (b) son concluyentes? en caso de no serlo, ¿como verificaría la homogeneidad/heterogeneidad? Plantee el procedimiento adecuado en conjunto con los supuestos requeridos.

[1] Higham, C. F. W.; Kinjingam, A. e Manly, B. F. (1980), An analysis of prehistoric canid remains from Thailand, J. Archaelogical Sci., 7: 140-165.

d. ***(0.3125 pts)*** ¿Que metodología emplearía para determinar si existe diferencias entre las medias de las variables de los 7 grupos? Argumente su respuesta.



Fórmulas:

$$n\left(\overline{\mathbf{x}}-\boldsymbol{\mu}\right)^{\top}\mathrm{S}^{-1}\left(\overline{\mathbf{x}}-\boldsymbol{\mu}\right)\leq\frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)$$

$$\overline{x}_i-\overline{x}_k\pm\sqrt{\frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)}\sqrt{\frac{s_{ii}+s_{kk}-2s_{ik}}{n}}$$

$$\overline{x}_k\pm t_{(n-1)}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{kk}}{n}}$$

$$T^2=\left(\overline{\mathbf{x}}_1-\overline{\mathbf{x}}_2-\boldsymbol{\delta}_0\right)^{\top}\left[\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\mathrm{S}_{pooled}\right]^{-1}\left(\overline{\mathbf{x}}_1-\overline{\mathbf{x}}_2-\boldsymbol{\delta}_0\right)$$

$$>\frac{(n_1+n_2-2)p}{n_1+n_2-p-1}F_{p,n_1+n_2-p-1}(\alpha)$$

$$\overline{\mathbf{x}}_1=\frac{1}{n_1}\sum_{j=1}^{n_1}\mathbf{x}_{1j}$$

$$\overline{\mathbf{x}}_2=\frac{1}{n_2}\sum_{j=1}^{n_2}\mathbf{x}_{2j}$$

$$\mathrm{S}_{pooled}=\frac{n_1-1}{n_1+n_2-2}\mathrm{S}_1+\frac{n_2-1}{n_1+n_2-2}\mathrm{S}_2$$

$$\mathrm{S}_1=\frac{1}{n_1-1}\sum_{j=1}^{n_1}\left(\mathbf{x}_{1j}-\overline{\mathbf{x}}_1\right)\left(\mathbf{x}_{1j}-\overline{\mathbf{x}}_1\right)^{\top}$$

$$\mathrm{S}_2=\frac{1}{n_2-1}\sum_{j=1}^{n_2}\left(\mathbf{x}_{2j}-\overline{\mathbf{x}}_2\right)\left(\mathbf{x}_{2j}-\overline{\mathbf{x}}_2\right)^{\top}$$