UNIVERSITAT DE BARCELONA

Universitat Oberta de Catalunya

Estadística descriptiva multivariante

Francesc Carmona

18 de febrero de 2019

1. Estadísticos

1. Para la base de datos crabs del paquete MASS de **R**:
   - *a*) Estudiar la base de datos con la ayuda y las funciones `str()` y `summary()`.
   - *b*) Calcular estadísticos descriptivos como la media, la mediana, la varianza, etc. de las variables numéricas, tanto para todos los datos, como para las especies y los sexos por separado.
   - *c*) Hallar los cinco números descriptivos de cada variable numérica. Utilizar la función `fivenum()`.
   - *d*) Hallar las matrices de varianzas y covarianzas de las variables numéricas según especie y sexo.[1]
   - *e*) Comparar cada especie con medidas globales de variabilidad: varianza total $\text{tr}(\mathbf{S})$, varianza media $\text{tr}(\mathbf{S})/p$, varianza generalizada $|\mathbf{S}|$ y varianza efectiva $|\mathbf{S}|^{1/p}$.
2. $(*)$ Para la base de datos huswif del libro de Everitt(2005)[2]
   - *a*) Cargar los datos con la instrucción:

     ```
     huswif <- source("chap1huswif.dat")$value
     ```

     y estudiar la base de datos con las funciones `str()` y `summary()`.
   - *b*) Calcular estadísticos descriptivos como la media, la mediana, la varianza, etc. de todas las variables.
   - *c*) Hallar los cinco números descriptivos de cada variable numérica. Utilizar la función `fivenum()`.
   - *d*) Hallar la matriz de varianzas-covarianzas y la matriz de correlaciones.
   - *e*) Hallar las medidas globales de variabilidad: varianza total $\text{tr}(\mathbf{S})$, varianza media $\text{tr}(\mathbf{S})/p$, varianza generalizada $|\mathbf{S}|$ y varianza efectiva $|\mathbf{S}|^{1/p}$.
3. $(*)$ Para la base de datos airpoll del libro de Everitt(2005).
   - *a*) Cargar los datos con la instrucción:

     ```
     airpoll <- source("chap2airpoll.dat")$value
     ```

     y estudiar la base de datos con las funciones `str()` y `summary()`.
   - *b*) Calcular estadísticos descriptivos como la media, la mediana, la varianza, etc. de todas las variables.
   - *c*) Hallar los cinco números descriptivos de cada variable numérica. Utilizar la función `fivenum()`.
   - *d*) Hallar la matriz de varianzas-covarianzas y la matriz de correlaciones.
   - *e*) Hallar las medidas globales de variabilidad: varianza total $\text{tr}(\mathbf{S})$, varianza media $\text{tr}(\mathbf{S})/p$, varianza generalizada $|\mathbf{S}|$ y varianza efectiva $|\mathbf{S}|^{1/p}$.

[1] Observar que en **R** las funciones matriciales `var()` o `cov()` proporcionan la matriz de varianzas corregida $\widehat{\mathbf{S}} = \frac{n}{n-1}\mathbf{S}$.

[2] Ver la página web `https://www.york.ac.uk/depts/maths/data/everitt/welcome.htm`.

4. Entre los 10 elementos muestrales o parejas casadas (con papeles o no) de la base de datos huswif del libro de Everitt(2005):

   a) Calcular las distancias euclídeas con la función `dist` de **R** y expresarlas en forma de matriz.

   b) Calcular las distancias de K.Pearson y dar la matriz de distancias.

   c) Calcular la distancia de Mahalanobis (sin cuadrado) entre la primera y la última de las parejas (`huswif[1,]` y `huswif[10,]`).

   d) Calcular las distancias al cuadrado de Mahalanobis de cada pareja al vector de medias y su correspondiente matriz de covarianzas. Probar `?mahalanobis` en **R**.

   e) Realizar un `qqplot` entre los cuantiles de una distribución ji-cuadrado con 5 grados de libertad y las distancias al cuadrado de Mahalanobis. ¿Se ajustan?
   También se puede utilizar la función `chisplot` de Everitt.

5. (∗) Estudiar el ajuste de las distancias al cuadrado de Mahalanobis para los datos de `airpoll` del libro de Everitt a una distribución ji-cuadrado.

   Utilizar un `qqplot` o un `chisplot` de Everitt. También podemos utilizar el gráfico de la función `outlier` del paquete `psych`.

6. Con la base de datos huswif del libro de Everitt(2005):

   a) Realizar una estandarización univariante del tipo

   $$\mathbf{y} = \mathbf{D}^{-1/2}(\mathbf{x} - \bar{\mathbf{x}})$$

   donde $\mathbf{D}^{-1/2} = \text{diag}(s_1^{-1}, \ldots, s_p^{-1})$ y obtener la base de datos transformada.

   b) Realizar una estandarización multivariante del tipo

   $$\mathbf{y} = \mathbf{S}^{-1/2}(\mathbf{x} - \bar{\mathbf{x}})$$

   c) Observar que la distancia euclídea entre las filas de datos tras la estandarización multivariante coinciden con la distancia de Mahalanobis de los datos originales.

7. Comprobar que la desviación típica generalizada en el caso de dos variables es

   $$|\mathbf{S}|^{1/2} = s_x s_y \sqrt{1 - r^2}$$

   donde s_x^2 y s_y^2 son las varianzas muestrales de las dos variables y r el coeficiente de correlación.

8. Medidas de dependencia lineal con las variables numéricas de la base de datos crabs del paquete MASS de **R**, únicamente para los datos de la especie *Blue* y sexo *Macho*:

   a) Hallar la correlación entre las variables CL i CW.
   Calcular la matriz de correlaciones **R** de todas las variables numéricas.

   b) Hallar la correlación múltiple entre la variable BD y el resto.

   c) El coeficiente de correlación parcial es una medida de dependencia directa entre dos variables.
   Se sabe que $r_{ij.k_1 \ldots k_q} = -\frac{s^{ij}}{\sqrt{s^{ii} s^{jj}}}$ donde s^{ij} son los elementos de la matriz $\mathbf{S}^{-1}$, llamada matriz de precisión. Calcular la matriz de correlaciones parciales

   $$\mathbf{P} = (-1)^* \text{diag}(\mathbf{S}^{-1})^{-1/2} \mathbf{S}^{-1} \text{diag}(\mathbf{S}^{-1})^{-1/2}$$

   donde $(-1)^*$ es el producto por -1 de todos los elementos excepto los de la diagonal. Comprobar que **no** es la inversa de la matriz de correlaciones $\mathbf{R}^{-1}$.

2. Gráficos

9. Con las variables CL y CW de la base de datos crabs del paquete MASS de **R**:

   a) Crear el diagrama de dispersión de las dos variables y sus histogramas marginales.

   b) Crear el diagrama de dispersión de las dos variables y sus diagramas de caja marginales.

10. Con la base de datos crabs del paquete MASS de **R**, realizar los siguientes diagramas de caja múltiples:

    a) Para comparar las variables según especie.

    b) Para comparar las variables según sexo.

    c) Para comparar las variables según especie y sexo.

11. Con la base de datos crabs del paquete MASS de **R**, realizar un matriz de diagramas de dispersión según especie y sexo con la instrucción `pairs()`.

    a) Añadir los histogramas de cada variable a la diagonal.

    b) Añadir la recta de regresión a cada diagrama de dispersión.

12. Con la base de datos huswif del libro de Everitt(2005):

    a) Calcular algún gráfico de dispersión 3D con la función `scatterplot3d()` del paquete scatterplot3d.

    b) Realizar un gráfico de dispersión en dos dimensiones de las variables edad del marido y edad de la mujer. Añadir con la función `symbols()` la variable edad del marido en el primer matrimonio.

13. (∗) Representar con caras de Chernoff las parejas de la base de datos huswif del libro de Everitt(2005).[3]

14. (∗) Representar con un gráfico de estrellas los vectores de medias de todas las variables numéricas de la base de datos crabs del paquete MASS de **R** en los cuatro grupos, según especie y sexo.[4]

15. Señalar los datos atípicos en la base de datos crabs, únicamente para los datos de la especie *Blue* y sexo *Macho*:

    a) Según un criterio univariante como

    $$\frac{|x_i - \text{med}(x)|}{\text{MAD}(x)} > 5$$

    donde $\text{med}(x)$ es la mediana de las observaciones y $\text{MAD}(x)$ la mediana de las desviaciones en valor absoluto respecto a la mediana.

    b) Con otro criterio univariante como

    $$x_i < Q_1 - 1.5 \cdot \text{IQR} \quad \text{o} \quad x_i > Q_3 + 1.5 \cdot \text{IQR}$$

    donde Q_1,Q_3 son los cuartiles y IQR el rango intercuartílico.

    c) Con un criterio multivariante sencillo como

    1) Se busca el 50 % de los datos con menor distancia de Mahalanobis al centro $\bar{\mathbf{x}}$, media de todos los datos.
    2) Se calculan la media $\bar{\mathbf{x}}_R$ y la matriz de covarianzas $\mathbf{S}_R$ con ese conjunto reducido de datos.

[3] Investigar la función `faces()` del paquete `aplpack`.
[4] Consultar la función `stars()`.

3) Se calculan las distancias de Mahalanobis d_M^2 de todos los datos a la media $\bar{\mathbf{x}}_R$ con la matriz de covarianzas $\mathbf{S}_R$.
4) Se consideran atípicos los datos tales que $d_M^2 > p + 3\sqrt{2p}$, donde p es el número de variables.

d) También podemos utilizar la distancia de Mahalanobis d_M^2 calculada con estimadores robustos como los que se obtienen con la función `cov.rob` del paquete `MASS`. Uno de los métodos más utilizados es el *minimum covariance determinant* o MCD que calcula los estimadores de centro y covarianza con un grupo de datos reducido, precisamente el que minimiza el determinante de la covarianza (una de las medidas de variabilidad multivariante).

Con estas distancias robustas, podemos dibujar un gráfico para observar por inspección los puntos más extremos. Algunos autores sugieren como punto de corte la cola superior al 97.5 % de la distribución ji-cuadrado asociada a la distancia de Mahalanobis.

Otro gráfico con estas distancias robustas es el ajuste a la distribución ji-cuadrado para observar los valores máximos.

Otros criterios pueden verse en Peña(2002) 4.5.3 y 11.3.

16. Realizar un boxplot bivariante con la función `bvbox()` de Everitt para las variables CL y CW de la base de datos `crabs` del paquete MASS de **R**.

¿Cuales son las medidas de dispersión robustas que se utilizan en el gráfico?

17. (∗) Un grupo muy importante de técnicas de representación de datos que no hemos explicado en los documentos de esta asignatura son los **gráficos en panel** (trellis) que podemos hallar en el paquete `lattice` de **R**.

Uno de los lugares donde informarse sobre los gráficos trellis es:

`http://www.stat.auckland.ac.nz/~ihaka/787/lectures-trellis.pdf`

Escribir un breve informe (menos de una página) que describa las características de los gráficos trellis y discutir sus méritos.

¿Qué hace diferentes a los gráficos trellis de otras técnicas de representación de datos multivariantes? ¿Pueden ser útiles para ti? ¿Cuando crees que es mejor usar gráficos en trellis que otras técnicas?

Mejor si se incluyen referencias a las fuentes.

18. (∗) El paquete `ggplot2`, creado por Hadley Wickham, ofrece un poderoso lenguaje gráfico para crear gráficos elegantes y complejos. Su popularidad en la comunidad R ha explotado en los últimos años. Originalmente basado en la gramática de los gráficos de Leland Wilkinson, `ggplot2` permite crear gráficos que representan datos numéricos y categóricos univariantes y multivariantes de manera directa. La agrupación se puede representar por color, símbolo, tamaño y transparencia. La creación de gráficos en panel (es decir, con condiciones) es relativamente simple.

Repetir el ejercicio 10 con los gráficos del paquete `ggplot2`.