UNIVERSITAT DE BARCELONA

Universitat Oberta de Catalunya

Prueba de evaluación continua 2
Estadística Multivariante

Francesc Carmona

15 de junio de 2024

Ejercicio 1 (40 pt)

Lubischew[1] consideró el problema de discriminación entre tres especies de escarabajuelos (*flea beetles*) llamadas *Chaetocnema concinna*, *Chaetocnema heikertingeri* y *Chaetocnema heptapotamica*, basada en varias medidas físicas. Los datos se pueden hallar en el *data.frame* `flea` del paquete `GGally`, donde podemos ver las variables con detalle.
De éstas, vamos a estudiar las tres últimas:

- `aede1` the maximal width of the aedeagus in the fore-part in microns
- `aede2` the front angle of the aedeagus (1 unit = 7.5 degrees)
- `aede3` the aedeagus width from the side in microns

Figura 1: Chaetocnema concinna.

(a) Estudiar gráficamente y con algún test la normalidad univariante en cada una de las tres especies. Resumir los resultados en la medida de lo posible.

(b) Estudiar la normalidad multivariante con el test de Mardia que se explica en el apéndice de este documento.

Calcular los estadísticos de asimetría $b_{1,p}$ y de kurtosis $b_{2,p}$. Calcular también los estadísticos estandarizados $\gamma_{1,p}$ y $\gamma_{2,p}$ y sus correspondientes p-valores y decidir si admitimos la normalidad en las tres especies.

Acompañar el test con un gráfico qq-plot de ajuste a la distribución ji-cuadrado de las distancias de Mahalanobis de los datos a la media en cada especie.

(c) Repetir el estudio de la normalidad multivariante con la asimetría y la kurtosis multivariante de Mardia gracias a la función `mvn` del paquete `MVN`. La misma función permite el estudio univariante y también hace los gráficos.

Observar la coincidencia con los resultados del apartado anterior.

(d) Con el mismo fin que los dos apartados anteriores se puede utilizar la función `mardia` del paquete `psych`, sin embargo los resultados son distintos ya que esta función utiliza la matriz de covarianzas insesgada.

(e) Comparar las matrices de covarianzas de los tres especies con el test de la razón de verosimilitudes.

También se puede utilizar el test M de Box[1].

(f) Si suponemos que se verifican todas las condiciones, realizar un MANOVA de un factor y contrastar si hay diferencias entre ellas.

(g) Realizar contrastes de comparación de medias dos a dos. ¿Debemos utilizar la corrección por contrastes múltiples?

Ejercicio 2 (25 pt)

Con los mismos datos del ejercicio 1, vamos a estudiar un posible análisis discriminante entre las dos especies de insectos del género *Chaetocnema*, *Ch. concinna* (especie a) y *Ch. heikertlingeri* (especie b). Para ello, utilizaremos únicamente las tres primeras variables para cada insecto:

- `tars1` width of the 1st joint of the tarsus (legs)
- `tars2` width of the 2nd joint of the tarsus
- `head` the maximal width of the head

Además, nos limitaremos a los diez primeros individuos para cada especie que utilizaremos como datos de entreno (*training data*). Estos datos se utilizarán para estimar los parámetros del modelo que usaremos en la función discriminante. El objetivo es obtener una regla de clasificación que identifique la especie de los insectos a partir de estas tres variables.

(a) Reducir el conjunto de datos según lo explicado y realizar un gráfico con la función `pairs()` de las variables con distintos símbolos o colores para cada especie.

(b) Contrastar la homogeneidad de las matrices de varianzas-covarianzas de las dos especies con el test de Bartlett[2]. ¿Está justificado hacer un análisis lineal discriminante?

(c) Realizar un análisis discriminante lineal con la función `lda()` del paquete `MASS` y probabilidades a priori iguales a 0.5.

(d) Hallar las probabilidades a posteriori.

(e) Predecir la especie del resto de insectos de las dos especies consideradas como si desconociéramos su valor. Calcular la tabla de validación.

(f) Utilizar la función `partimat()` del paquete `klaR` para mostrar los resultados de la clasificación lineal y cuadrática con **todos** los insectos de las dos especies.

[1] La función `boxM()` del paquete `heplots` nos puede ayudar

[2] La generalización multivariante del test de Bartlett es el test de la razón de verosimilitudes

Ejercicio 3 (35 pt)

Vamos a realizar un análisis de conglomerados con diversos métodos utilizando los datos ecológicos de Woodyard Hammock, un bosque de hayas y magnolias en el norte de Florida. Los datos incluyen las frecuencias del número de árboles de cada especie en $n = 72$ lugares. Se identificaron y contabilizaron un total de 31 especies, sin embargo, solo $p = 13$ de las especies más comunes son el objeto de nuestro estudio.

`carcar`	*Carpinus caroliniana*	Ironwood
`corflo`	*Cornus florida*	Dogwood
`faggra`	*Fagus grandifolia*	Beech
`ileopa`	*Ilex opaca*	Holly
`liqsty`	*Liquidambar styraciflua*	Sweetgum
`maggra`	*Magnolia grandiflora*	Magnolia
`nyssyl`	*Nyssa sylvatica*	Blackgum
`ostvir`	*Ostrya virginiana*	Blue Beech
`oxyarb`	*Oxydendrum arboreum*	Sourwood
`pingla`	*Pinus glabra*	Spruce Pine
`quenig`	*Quercus nigra*	Water Oak
`quemic`	*Quercus michauxii*	Swamp Chestnut Oak
`symtin`	*Symplocus tinctoria*	Horse Sugar

La primera columna proporciona un código de 6 letras para identificar las especies, la segunda columna da el nombre científico y la tercera el nombre común de cada especie. De éstas, lógicamente, las especies que se hallaron con más frecuencia fueron las hayas y las magnolias. Tenemos estos datos en el archivo adjunto `wood.txt`[3].

(a) Realizar un análisis jerárquico aglomerativo al estilo SAS: con la distancia euclídea y *complete linkage*. Dibujar el dendograma con 6 clusters.

(b) Hacer un Análisis de la Varianza para cada especie (columna) en función del factor cluster obtenido en el apartado anterior y señalar los que resultan significativos. Una tabla con la especie, el estadístico F y su p-valor[4] será clarificadora.

Calcular una tabla con las medias para cada cluster de cada una de las especies donde las diferencias son significativas y señalar en esa tabla las medias más altas (mayor abundancia de esa especie en el cluster). Eso puede caracterizar los clusters.

(c) Repetir el apartado anterior con el método de Ward y un total de 4 clusters que podemos implementar con la función `agnes()` del paquete `cluster`.

[3] Cuidado con la columna adicional que indica el lugar de muestreo.

[4] Para ello habrá que hacer múltiples comparaciones, de forma que debe aplicarse una corrección de Bonferroni. En este caso, $p = 13$ significa que el nivel de significación será $\alpha = 0.05/13$.

(d) Aplicar el procedimiento de particionado de las k-medias con el algoritmo de Hartigan-Wong para 4 grupos.

Dibujar los lugares como puntos de un PCA o CA con un símbolo distinto para cada grupo[5]. Hallar los centros de cada grupo y dibujarlos en el gráfico anterior. Hallar las sumas de cuadrados dentro de cada grupo y los tamaños de los conglomerados.

(e) Realizar un particionado alrededor de los medoides o k-medoides o PAM con la función `pam()` del paquete `cluster` de R y hallar los medoides.

Calcular y representar la silueta como medida de la calidad de los conglomerados formados.

Referencias

[1] Lubischew, A.A. (1962) *On the Use of Discriminant Functions in Taxonomy*, Biometrics 18, pp. 455-477.

[2] Mardia, K.V. (1970) *Measures of multivariate skewness and kurtosis with applications*, Biometrika 57. pp. 519-530.

[5] La función `clusplot()` del paquete `cluster` nos puede ayudar.

Apéndice: Asimetría y kurtosis multivariantes

La asimetría y la kurtosis multivariantes son generalizaciones de la asimetría y la kurtosis univariantes, es decir, los momentos tercero y cuarto estandarizados tanto para distribuciones multivariantes como para muestras. Otras medidas de asimetría univariante, como (media - moda)/σ^2 también se llama asimetría y puede generalizarse del mismo modo.
Sea F una distribución p-dimensional arbitraria, $\boldsymbol{\mu}$ su vector $p \times 1$ de medias (poblacional) y $\boldsymbol{\Sigma}$ su matriz $p \times p$ de covarianzas (poblacional). Sea $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ una muestra de $p \times 1$ observaciones de forma que la matriz de datos es

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix}$$

cuyo vector de medias muestrales es

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$$

y matriz de covarianzas muestral (sesgada)

$$\mathbf{S} = \frac{1}{n} (\mathbf{x}_i - \bar{\mathbf{x}})'(\mathbf{x}_i - \bar{\mathbf{x}})$$

En las fórmulas anteriores, los vectores de observaciones $\mathbf{x}_i$ son vectores columna que forman las filas de la matriz de datos $\mathbf{X}$.

Mardia (1960?) define la asimetría $\beta_{1,p}$ y la kurtosis $\beta_{2,p}$ multivariantes de una distribución F como

$$\beta_{1,p} = E\{[(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu})]^3\}$$
$$\beta_{2,p} = E\{[(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})]^2\}$$

donde $\mathbf{x}$ y $\mathbf{y}$ son vectores aleatorios $p \times 1$ independientes con esa distribución. También definió la asimetría $b_{1,p}$ y la kurtosis $b_{2,p}$ muestrales multivariantes de un conjunto de observaciones $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ como

$$b_{1,p} = \frac{1}{n^2} \sum_{i,j=1}^{n} [(\mathbf{x}_i - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_j - \bar{\mathbf{x}})]^3$$
$$b_{2,p} = \frac{1}{n} \sum_{i=1}^{n} [(\mathbf{x}_i - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})]^2$$

Para cualquier matriz $\mathbf{A}$ $p \times p$ no singular y cualquier vector $\mathbf{d}$ $p \times 1$, $b_{1,p}$ y $b_{2,p}$ son invariantes para transformaciones afines $\mathbf{Ax} + \mathbf{d}$ de las muestras. Los coeficientes poblacionales $\beta_{1,p}$ y $\beta_{2,p}$ también son invariantes para este tipo de transformaciones.
Cuando la dimensión p es 1, $\beta_{1,p}$ y $b_{1,p}$ se reducen al cuadrado de los coeficientes univariantes de asimetría habituales $\sqrt{\beta_1}$ y b_1. Además, $\beta_{2,p}$ y $b_{2,p}$ se reducen a los coeficientes de kurtosis univariantes β_2 y b_2.
La asimetría de cualquier distribución simétrica sobre su media es $\beta_{1,p} = 0$. Luego la distribución normal multivariante p-dimensional $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ tiene asimetría $\beta_{1,p} = 0$. Para una muestra aleatoria de esta distribución, el estadístico

$$\gamma_{1,p} = \frac{n}{6} b_{1,p} \sim \chi_f^2$$

sigue asintóticamente ($n \to \infty$) una distribución ji-cuadrado con $f = p(p+1)(p+2)/6$ grados de libertad. Se puede hallar una versión mejorada de este resultado gracias al cálculo de la esperanza de este estadístico $MS = \gamma_{1,p}$ en el caso normal

$$E(MS) = \frac{p(p+2)[(n+1)(p+1) - 6]}{(n+1)(n+2)}$$

La distribución normal multivariante p-dimensional $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ tiene kurtosis $\beta_{2,p} = p(p+2)$. Para una muestra aleatoria de esta distribución el estadístico $b_{2,p}$ se puede estandarizar

$$\gamma_{2,p} = \frac{b_{2,p} - p(p+2)}{[8p(p+2)n^{-1}]^{1/2}}$$

o su versión mejorada

$$MK = \frac{b_{2,p} - p(p+2)(n-1)(n+1)^{-1}}{[8p(p+2)n^{-1}]^{1/2}}$$

tienen asintóticamente una distribución $N(0,1)$.

Mardia[2] propuso utilizar la asimetría muestral multivariante y la kurtosis muestral multivariante para contrastar la normalidad. Un test basado en la asimetría rechazará la hipótesis de normalidad multivariante si $b_{1,p}$ (o bien $\gamma_{1,p}$) es grande. Un test basado en la kurtosis se puede hacer si rechazamos la hipótesis de normalidad cuando $b_{2,p}$ (o bien $\gamma_{2,p}$) es muy pequeño o muy grande. En caso contrario, con los dos estadísticos con valores moderados, aceptaremos la normalidad multivariante si las distribuciones marginales son en general normales.