UNIVERSITAT DE BARCELONA

Universitat Oberta de Catalunya

Prueba de evaluación continua 1
Estadística Multivariante

Francesc Carmona

23 de abril de 2024

Ejercicio 1 (55 pt.)

El *data.frame* `urine` del paquete `boot` de **R** contiene un conjunto de 79 muestras de orina con intención de determinar si ciertas características físicas están relacionadas con la formación de cristales de oxalato de calcio. Estos datos se han obtenido de la tabla 44.1 del libro de Andrews and Herzberg (1985).
Las variables registradas son:

r Indicador de la presencia de oxalato de calcio: 0 ausencia y 1 presencia.

gravity Gravedad específica de la orina.

ph pH de la orina.

osmo Osmolaridad de la orina.

cond Conductividad de la orina.

urea Concentración de urea en la orina.

calc Concentración de calcio (milimoles por litro).

(a) Identificar los datos faltantes (variables y observaciones).

Imputar a esos datos faltantes los valores que resulten del siguiente procedimiento[1].

- Utilizar para el conjunto de variables numéricas el método `pmm` de la función `mice` del paquete del mismo nombre.
- Fijaremos la semilla en 123.
- Realizaremos $m = 50$ imputaciones y calcularemos la mediana del conjunto de valores.
- Finalmente, rellenaremos los valores faltantes con las medianas obtenidas.

Nota: Si no se sabe hacer la imputación, se puede seguir el ejercicio suprimiendo las observaciones que contienen algún dato faltante.

(b) Obtener los vectores de medias y las matrices de varianzas-covarianzas para cada grupo según el indicador de oxalato de calcio. Mejor si no se utiliza ninguna instrucción en bucle (`for`).

Calcular la variación total y la varianza generalizada de los dos grupos definidos según el indicador.

A simple vista, ¿podemos decir que la variabilidad de ambos grupos es similar?

[1] Se puede ver un ejemplo en el apartado *Imputing the missing data* de la página web Imputing missing data with R; MICE package

(c) La mediana geométrica es un indicador robusto de centralidad y se define[2] como el punto $\mathbf{m}$ tal que la suma de distancias euclídeas a los puntos de la muestra es mínima:

$$\sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{m}\|$$

Calcular la mediana geométrica, también llamada mediana espacial (*spatial median*) o mediana L1, de cada grupo según el indicador.

Comparar la distancia euclídea entre los puntos medios de los dos grupos y la distancia euclídea entre las medianas geométricas.

Nota: **R** dispone de varias funciones en diversos paquetes que calculan la mediana geométrica.

(d) Estudiar la presencia de datos atípicos en cada uno de los dos grupos por separado.

Utilizar la distancia de Mahalanobis con estimadores robustos del vector de medias y la matriz de covarianzas.

Recordemos que hay que fijar la semilla con el valor 123.

(e) En el ANOVA multivariante, que estudiaremos más adelante, es necesario calcular algunas matrices de covarianzas como la matriz "dentro" del grupo (*within-group covariances*) o la matriz de covarianzas combinadas (*pooled covariance matrix*). Los estadísticos para el contraste de la hipótesis de igualdad de medias se basan en estas matrices. En particular, la llamada lambda de Wilks.

Calcular la matriz de covarianzas combinadas (*pooled*) de estos datos y los dos grupos que define el indicador `r`:

$$\mathbf{S}_P = \frac{1}{n-k} \sum_{i=1}^{k} (n_i - 1)\mathbf{S}_i$$

donde k es el número de grupos, $n = \sum n_i$ es el número total de observaciones, n_i el tamaño del i-ésimo grupo y $\mathbf{S}_i$ la matriz de covarianzas muestral de cada grupo.

Calcular la lambda de Wilks en la forma:

$$\Lambda = \frac{|\mathbf{W}|}{|\mathbf{T}|}$$

donde $\mathbf{W} = (n-k)\mathbf{S}_P$, $\mathbf{T} = (n-1)\mathbf{S}$ y $\mathbf{S}$ es la matriz de covarianzas del conjunto de datos.

(f) Realizar un análisis de componentes principales con las variables numéricas.

¿Cuantas componentes parecen necesarias para tener una buena representación de los datos? Para contestar esta pregunta los cuatro métodos más utilizados son:

1. El criterio de Kaiser: se eligen las primeras componentes cuyos valores propios son superiores a 1.
2. Varianza explicada: se eligen las primeras componentes que explican como mínimo el 70 o 80 % de la varianza total.
3. Scree plot de Cattel (1966): Éste es un método gráfico por el que se eligen las componentes hasta el codo del gráfico.
4. El criterio de la interpretabilidad: se trata de elegir todas las componentes que mejor recojan la esencia del significado de las variables y verificar que esa interpretación tiene sentido en los términos conocidos del problema bajo investigación.

[2] Ver su definición en la Wikipedia.

Seguramente, la solución es una combinación de los cuatro criterios.

(g) Dibujar un gráfico con las dos primeras componentes y representar en él los puntos en función de la presencia de oxalato de calcio.

Repetir el mismo gráfico con las tres primeras componentes.

Evaluar las componentes respecto a las variables y discutir el resultado.

Ejercicio 2 (45 pt.)

En el trabajo de Hunt et al.[2] se estudió la capacidad reproductiva de cinco especies de aves marinas en dos colonias en el sureste del mar de Bering. Además, el apéndice de este estudio resume las colonias y los tamaños de las poblaciones de otros trabajos. El archivo `seabirds.csv` recoge los datos (número de pájaros) de 23 especies en 9 colonias en el área del norte polar y subpolar.

El principal interés de este ejercicio es representar las colonias de diversas formas.

(a) Calcular las frecuencias relativas, las frecuencias relativas marginales y la matriz de perfiles. El resultado debería ser la tabla 12.6 del libro de Krebs[3] que reproducimos al final de este documento.

(b) Calcular la matriz de distancias ji-cuadrado entre los perfiles de las columnas y su inercia total.

(c) Con la matriz de distancias ji-cuadrado entre los perfiles realizar un escalado multidimensional. Dibujar las coordenadas principales para las columnas.

(d) Realizar un análisis de correspondencias y calcular las inercias principales (en %) y la inercia total con los valores propios.

Dibujar una representación simétrica del CA. A pesar de la confusión de nombres, ¿cuales son las especies que caracterizan a la colonia SI (Skomer Island, Irish Sea)?

(e) Dada la gran cantidad de ceros en la tabla 12.6, en el libro de Krebs[3] se sugiere la utilización de la distancia de Canberra entre las columnas de la tabla 12.6. La distancia de Canberra no tiene una única definición y, además, ha cambiado a lo largo de la historia. Una posible definición entre dos vectores $\mathbf{p} = (p_1, p_2, \ldots, p_k)'$ y $\mathbf{q} = (q_1, q_2, \ldots, q_k)'$ de la misma longitud es

$$d_C(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{k} \frac{|p_i - q_i|}{|p_i| + |q_i|}$$

Cuando el denominador es cero, el cociente es NaN, y el sumando se elimina.

Comprobar que esta definición no sirve para calcular la matriz de similaridades de la tabla 12.7 del libro de Krebs[3] que se reproduce al final de este documento.

Figura 1: Dos ejemplares de frailecillo corniculado (Fratercula corniculata), *horned puffin*.

Una modificación de la distancia anterior es considerar la distancia

$$d_C(\mathbf{p}, \mathbf{q}) = \frac{1}{k} \sum_{i=1}^{k} \frac{|p_i - q_i|}{|p_i| + |q_i|}$$

Igual que antes, cuando el denominador es cero, el sumando se elimina.

Comprobar que con esta definición se puede obtener[3] la tabla 12.7.

Finalmente, se puede comprobar que ésta tampoco es la definición que utiliza **R** para calcular la distancia de Canberra. Tras una ardua investigación, se comprueba que la definición de **R** es

$$d_C(\mathbf{p}, \mathbf{q}) = \frac{k}{k - n_z} \sum_{i=1}^{k} \frac{|p_i - q_i|}{|p_i| + |q_i|}$$

donde n_z es el número de denominadores cero.

Comprobar que ésta es la definición de la distancia de Canberra según **R**.

(f) Realizar un MDS con la distancia de Canberra de **R**. Comprobar que se trata de una distancia euclídea. Dibujar el mapa.

Comparar el resultado con el obtenido con la distancia ji-cuadrado. Utilizar la función `procrustes()` del paquete `vegan`.

Referencias

[1] Andrews, D.F. and Herzberg, A.M. (1985) *Data: A Collection of Problems from Many Fields for the Student and Research Worker*. Springer-Verlag.

[2] George L. Hunt, Zoe A. Eppley and David C. Schneider, Reproductive Performance of Seabirds: The Importance of Population and Colony Size, The Auk 103: 306-317, April 1986.

[3] Krebs, C.J., *Ecological Methodology*, 3rd ed. (in prep) Chapters revised to date (14 March 2014).

[4] Manly, Bryan F.J. and Navarro Alberto, Jorge A., *Multivariate Statistical Methods: A Primer*, Fourth Edition, Chapman & Hall, 2016.

[3] La tabla 12.7 contiene dos o tres erratas.

TABLE 12.6 RELATIVE ABUNDANCES (PROPORTIONS) OF 23 SPECIES OF SEABIRDS ON 9 COLONIES IN NORTHERN POLAR AND SUBPOLAR AREAS[a]

	Cape Hay, Bylot Island	Prince Leopold Island, eastern Canada	Coburg Island, eastern Canada	Norton Sound, Bering Sea	Cape Lisburne, Chukchi Sea	Cape Thompson, Chukchi Sea	Skomer Island, Irish Sea	St. Paul Island, Bering Sea	St. George Island, Bering Sea
Northern fulmar	0	.3422	0	0	0	0	.0007	.0028	.0278
Glaucous-winged gull	.0005	.0011	.0004	.0051	.0004	.0007	0	0	0
Black-legged kittiwake	.1249	.1600	.1577	.1402	.1972	.0634	.0151	.1221	.0286
Red-legged kittiwake	0	0	0	0	0	0	0	.0087	.0873
Thick-billed murre	.8740	.4746	.8413	.0074	.2367	.5592	0	.4334	.5955
Common murre	0	0	0	.7765	.5522	.3728	.0160	.1537	.0754
Black guillemot	.0006	.02200.	.0005	0	.0013	.00001	0	0	0
Pigeon guillemot	0	0	0	0	0	.00003	0	0	0
Horned puffin	0	0	0	.0592	.0114	.0036	0	.0173	.0111
Tufted puffin	0	0	0	.0008	.0002	0	0	.0039	.0024
Atlantic puffin	0	0	0	0	0	0	.0482	0	0
Pelagic cormorant	0	0	0	.0096	.0006	.0001	.0001	0	0
Red-faced cormorant	0	0	0	0	0	0	0	.0099	.0020
Shag	0	0	0	0	0	0	.0001	0	0
Parakeet auklet	0	0	0	.0012	0	0	0	.1340	.0595
Crested auklet	0	0	0	0	0	0	0	.0236	.0111
Least auklet	0	0	0	0	0	0	0	.0906	.0992
Razorbill	0	0	0	0	0	0	.0130	0	0
Manx shearwater	0	0	0	0	0	0	.7838	0	0
Storm petrel	0	0	0	0	0	0	.0389	0	0
Herring gull	0	0	0	0	0	0	.0229	0	0
Great black-backed gull	0	0	0	0	0	0	.0001	0	0
Lesser black backed gull	0	0	0	0	0	0	.0603	0	0

[a] Data from Hunt et al. (1986).

TABLE 12.7 MATRIX OF SIMILARITY COEFFICIENTS FOR THE SEABIRD DATA IN TABLE 12.6. ISLANDS ARE PRESENTED IN SAME ORDER AS IN TABLE 12.6[a]

	CH	PLI	CI	NS	CL	CT	SI	SPI	SGI
CH	1.0	0.88	0.99	0.66	0.77	0.75	0.36	0.51	0.49
PLI		1.0	0.88	0.62	0.70	0.71	0.36	0.51	0.49
CI			1.0	0.66	0.78	0.75	0.36	0.50	0.48
NS				1.0	0.73	0.64	0.28	0.53	0.50
CL					1.0	0.76	0.29	0.51	0.49
CT						1.0	0.34	0.46	0.45
SI							1.0	0.19	0.20
SPI								1.0	0.80
SGI									1.0

[a] The complement of the Canberra metric (1.0 - *C)* is used as the index of similarity. Note that the matrix is symmetrical about the diagonal.