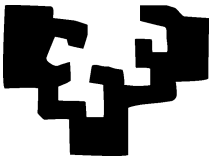
eman ta zabal zazu



Universidad del País Vasco
Euskal Herriko Unibertsitatea

TAREA 2

EJERCICIOS

1. Demuestra que el estadístico T^2 de Hotelling para el contraste de hipótesis sobre el vector de medias de una población es invariante frente a transformaciones lineales no singulares (es decir, demuestra que si en lugar de emplear las observaciones originales $\vec{X}_i$ $(i = 1, \ldots, N)$ para hacer el contraste empleases $\vec{Y}_i = A\vec{X}_i$ siendo A una matriz no singular, el resultado sería exactamente el mismo).

2. En clase se mencionó que el contraste T^2 de Hotelling para la hipótesis $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ es notablemente robusto ante desviaciones de la hipótesis de normalidad multivariante. Haz una pequeña simulación para explorar hasta donde llega esta robustez. Repite cien veces, guardando los resultados cada vez, lo siguiente:

   a) Genera 200 vectores $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{200}$ de dimensión 5 procedentes de una distribución *no* normal.

   b) Obtén vectores transformados linealmente, $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{200}$, de modo que tengan cierta correlación y vector de medias $\boldsymbol{\mu}_0$ conocido.

   c) Estima vector de medias y matriz de covarianzas, y a continuación el estadístico T^2 para contraste de la hipótesis y su transformación en un estadístico con distribución $\mathcal{F}$ de Snedecor. Guarda este último resultado en cada iteración.

   Al finalizar, tendrás 100 valores del estadístico de contraste para la hipótesis $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ (cierta, por construcción de las observaciones). Si las observaciones hubieran sido normales multivariantes, dichos 100 valores procederían de una distribución $\mathcal{F}$ de Snedecor con grados de libertad adecuados. Al fallar el supuesto de normalidad, la distribución no será ya tal. Compara la probabilidad teórica de rechazo de H_0 bajo la hipótesis de normalidad ($= \alpha$, nivel de significación) con el porcentaje empírico de rechazos en las 100 repeticiones del experimento.

3. El conjunto de datos `painters` (forma parte de la librería `MASS`) es una *data frame* con una muestra de pintores, a cada uno de los cuales se ha atribuido[1] cuatro notas (en Composición, Dibujo, Color, Expresión. La quinta columna es un código de la escuela a la que pertenece.

[1] Por un crítico de arte. Haz `help(painters)` si quieres más detalles.

Las respuestas a las siguientes preguntas serán bastante triviales a cualquiera interesado en la pintura, pero haz no obstante los oportunos contrastes de hipótesis formales. Supón normalidad multivariante.

a) ¿Es plausible la hipótesis de que los vectores de notas correspondientes a pintores de distintas escuelas tienen el mismo vector de medias? (En otras palabras: ¿algunas escuelas primaron el dibujo, la composición, etc. sobre otras?)

b) ¿Puede aceptarse que las matrices de covarianzas correspondientes a las diferentes escuelas son iguales? (¿Qué significaría esto?)

c) *Suponiendo* que la matriz de covarianzas de las cuatro notas es la misma para todas las escuelas, ¿puede aceptarse que las cuatro notas que se otorgan a cada pintor están incorreladas?

4. En la *dataframe* `Sitka` (forma parte de la librería `MASS`) tienes datos correspondientes a 79 árboles, medidos en cinco ocasiones. De ellos, 54 fueron cultivados en cámaras enriquecidas en ozono, y 25 fueron controles (cultivados en el medio ambiente, sin tratamiento especial). Los significados de las variables son:

Cuadro 1: Variables de la dataframe `Sitka89`

Variable	Descripción
`size`	Producto altura por diámetro al cuadrado, en escala logarítmica.
`Time`	Momento de la medida (en días desde 1 Enero 89).
`tree`	Identificador del árbol.
`treat`	Tratamiento recibido (ozono o control)

Fuente: Datos en la biblioteca de funciones para R `MASS`, aneja a Venables y Ripley (1994).

Haz un contraste de igualdad de vectores de medias entre las dos poblaciones (árboles tratados y controles).

AYUDAS, SUGERENCIAS Y COMPLEMENTOS

1. Al margen de tus apuntes de clase puedes mirar las secciones relevantes de cualquiera de los muchos manuales a tu disposición en Biblioteca (clasificados en 519.237, segunda planta). En particular puedes consultar Peña (2002), Seber (1984), Hair, Anderson, Tatham, y Black (1992) y Rencher (1995), Johnson y Wichern (1992) o Cuadras (1981). De nivel bastante más alto es Anderson (1984).

   En particular, encontrarás copntrastes de igualdad de matrices de covarianzas y de incorrelación entre variables en las notas de clase y el en Capítulo 10 de Peña (2002).

2. En uno de los problemas tienes datos pertenecientes a varias escuelas, y te interesa repetir los mismos cálculos sobre las observaciones desglosadas por escuela. Hay muchas manera de hacerlo. Una, quizá la más cómoda, recurre al empleo de `split` (que permite construir una lista cada una de cuyas componentes corresponde a observaciones de uno de los grupos),

`lapply` (que permite repetir cálculos sobre cada uno de los componentes de una lista) y `cov.wt` (calcula matrices de covarianzas). Por ejemplo, en el caso de los datos de pintores,

```
x <- painters[,1:4]          #  variables numéricas
g <- painters[,5]            #  indicador de escuela
datos <- split(x,g)          #  Divide la muestra por valores del
                             #  indicador de escuela. Retorna una
                             #  lista.
res <- lapply(datos,cov.wt)  #  Aplica la función cov.wt a cada uno
                             #  de los componentes recién obtenidos.
res                          #  Muestra los resultados.
```

Observa el contenido de `res`; tienes ahí la materia prima para todos los contrastes T^2 y similares que necesites hacer.

3. Mira bien los datos de `Sitka`. No están tal cual los necesitas utilizar, pero no te costará transformarlos a lo que necesitas. Te dan los datos correspondientes a cada árbol en varios registros. Tú quieres los datos de cada árbol como una observación multivariante (una fila de la matriz X). Es decir, te dan

```
> Sitka
   size Time tree treat
1  4.51  152    1 ozone
2  4.98  174    1 ozone
3  5.41  201    1 ozone
4  5.90  227    1 ozone
5  6.15  258    1 ozone
6  4.24  152    2 ozone
7  4.20  174    2 ozone
8  4.68  201    2 ozone
9  4.92  227    2 ozone
10 4.96  258    2 ozone
11 3.98  152    3 ozone
12 4.36  174    3 ozone
   ...
```

y tu querrías:

```
    Time
tree  152  174  201  227  258
  1  4.51 4.98 5.41 5.90 6.15
  2  4.24 4.20 4.68 4.92 4.96
  3  3.98 4.36 4.79 4.99 5.03
  ...
```

Una forma fácil de pasar de lo uno a lo otro es el uso de `xtabs` así (mira la documentación):

```
datos   <- split(Sitka[,1:3],Sitka[,4])
ozono   <- xtabs(size ~ tree + Time,data=datos$ozone)
control <- xtabs(size ~ tree + Time,data=datos$control)
```

Esto todavía no es lo que quieres, porque `ozono` y `control` son objetos de tipo `xtabs` y tu quieres matrices. Puedes hacer la conversión facilmente así:

```
ozono   <- as.matrix(ozono)
control <- as.matrix(control)
```

No necesitas hacer uso de él aquí, pero para reorganizar datos de todas las maneras imaginables, el paquete `reshape2` puede serte de utilidad.

4. Hay una diferencia notable entre los datos de pintores y los de crecimiento de árboles. En estos últimos hay más estructura: un mismo árbol es observado en diferentes momentos de tiempo, lo que introduce restricciones (por ejemplo, los árboles no "encogen", de manera que las medidas correspondientes a un mismo árbol debieran ser monótonas: observarás sin embargo alguna anomalía en los datos cuyo motivo desconozco.). Son *datos longitudinales*.

   Esta mayor estructura sugiere la posibilidad de una parametrización más parca. En el ejemplo de los árboles, en que además se tienen los momentos de las medidas y el intervalo entre ellos por consiguiente, podríamos modelizar la matriz de covarianzas entre medidas con menos parámetros —por ejemplo, con una única tasa de crecimiento—. No es éste el objeto de la tarea. Sobre datos longitudinales encontrarás bastante bibliografía.

5. En el problema referido a las notas de pintores, habrás recurrido a hacer contrastes de igualdad de vector de medias por parejas. Esto es muy insatisfactorio, porque la hipótesis era: "vector de medias común", y tu has tenido que contrastar muchas "subhipótesis" que sólo involucran a dos escuelas cada vez.

   En la asignatura **Estadística: Modelos Lineales** dedicamos bastante atención al problema de la inferencia simultánea: lo que aprendiste entonces es de aplicación aquí, y las limitaciones del análisis que has realizado ponen de manifiesto la necesidad de una herramienta mejor para este tipo de problemas (que estudiaremos con el nombre de MANOVA).

6. Si el número de observaciones no es abrumador, un paso previo a cualquier análisis estadístico es examinar los datos cuidadosamente. Puedes calcular sus estadísticos de posición y dispersión, hacer algunos gráficos de caja (que pondrán sobre aviso de *outliers* univariantes), dibujar los histogramas de cada variable y quizá hacer gráficos QQ (lo que ilustrará acerca de la normalidad de las marginales). Los ficheros `prevSitka.R` y `prevPainters.R` ilustran cosas que puedes querer mirar. Asegúrate de entender lo que hace cada instrucción e incorpórala a tu acervo para uso posterior.

7. El fichero `Ttest.R` contiene una función que toma la salida de `cov.wt` y realiza todos los posibles contrastes de igualdad de medias por parejas, utilizando la distribución T^2 de Hotelling. De nuevo, asegúrate de entender lo que hace, sin utilizarla como una caja negra.

Referencias

Anderson, T. W. (1984). *An introduction to multivariate statistical analysis.* Wiley.

Cuadras, C. M. (1981). *Métodos de análisis multivariante.* Barcelona: Eunibar.

Hair, J. J., Anderson, R. E., Tatham, R. L., y Black, W. C. (1992). *Multivariate data analysis.* New York: Maxwell MacMillan.

Johnson, R. A., y Wichern, D. W. (1992). *Applied multivariate statistical analysis.* Prentice-Hall International.

Krzanowski, W. J. (1988). *Principles of multivariate analysis: A user's perspective.* Oxford. (Signatura: 519.23 KRZ)

Peña, D. (2002). *Análisis de datos multivariantes.* McGraw-Hill.

Rencher, A. C. (1995). *Methods of multivariate analysis.* Wiley.

Seber, G. A. F. (1984). *Multivariate observations.* New York: Wiley.

Venables, W. N., y Ripley, B. D. (1994). *Modern applied statistics with s-plus.* New York: Springer-Verlag. (Signatura: 681.03.068 VEN)