Taller de Inferencia Multivariante

20582 - Análisis de Datos

Esta práctica consiste en dos partes:

Parte 1:

Resolved el ejercicio que le fue asignado al grupo creando un repositorio en Github con todos los ficheros necesarios (documento .qmd con la respuesta, datos, etc). Esta parte se expondrá el 13/11 o el 16/11 según el orden del número del ejercicio.

- La lista de problemas es la siguiente:

Problema 1:

Sea $\mathbf{X} = (X_1, X_2, X_3)'$ *un vector aleatorio tridimensional que sigue una distribución normal con media* $\boldsymbol{\mu} = (1, 0, -2)'$ *y matriz de varianzas-covarianzas*

$$\boldsymbol{\Sigma} = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 4 & 1 \\ 0 & 1 & 6 \end{pmatrix}.$$

(a) *Escríbase la forma cuadrática* $Q(x_1, x_2, x_3)$ *del exponente de la densidad del vector aleatorio* $\mathbf{X}$.

(b) *Escríbase la matriz de covarianzas cruzadas entre* $\begin{pmatrix} X_1 \\ X_3 \end{pmatrix}$ *y* X_2.

(c) *Encuéntrese la correlación entre* X_1 *y* X_3 *condicionadas por* $X_2 = x_2$.

(d) *Hállese* $\text{var}(X_1 | X_2 = x_2)$ *y compárese con* $\text{var}(X_1)$.

Problema 2:

Sea $\mathbf{X}_1, \ldots, \mathbf{X}_{80}$ *una muestra de una población con media* $\boldsymbol{\mu}$ *y matriz de covarianzas* $\boldsymbol{\Sigma}$.

(a) *¿Cuál es la distribución aproximada de*

$$\bar{\mathbf{X}} = \sum_{i=1}^{80} X_i/80 \ ?$$

(b) *Tómense* $N = 200$ *muestras de tamaño* $n = 80$ *de un vector* $\mathbf{X} = (X_1, X_2)'$ *con distribución uniforme en el cuadrado* $[0, 1] \times [0, 1]$. *Calcúlense las medias* $\bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_N$ *de estas muestras y dibújese el histograma correspondiente a las medias, comprobando si se asemeja a una densidad normal.*

Problema 3:

Sea $\mathbf{X} = (X_1, X_2, X_3)'$ *un vector aleatorio tridimensional que sigue una distribución normal con media* $\boldsymbol{\mu} = (1, 0, -2)'$ *y matriz de varianzas-covarianzas*

$$\boldsymbol{\Sigma} = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 4 & 1 \\ 0 & 1 & 6 \end{pmatrix}.$$

(a) *Escríbase la forma cuadrática* $Q(x_1, x_2, x_3)$ *del exponente de la densidad del vector aleatorio* $\mathbf{X}$.

(b) *Escríbase la matriz de covarianzas cruzadas entre* $\begin{pmatrix} X_1 \\ X_3 \end{pmatrix}$ *y* X_2.

(c) *Encuéntrese la correlación entre* X_1 *y* X_3 *condicionadas por* $X_2 = x_2$.

(d) *Hállese* $\text{var}(X_1 | X_2 = x_2)$ *y compárese con* $\text{var}(X_1)$.

Problema 4:

Una distribución muy relacionada con la ley normal multivariante, y que es el análogo multivariante de la ley χ^2, *es la distribución Wishart. Dados* $\mathbf{X}_1, \ldots, \mathbf{X}_n$ *vectores aleatorios i.i.d.*$\sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$, *la matriz* $p \times p$

$$\mathbf{Q} = \sum_{i=1}^{n} \mathbf{X}_i \, \mathbf{X}_i' \sim W_p(\boldsymbol{\Sigma}, n)$$

sigue una ley Wishart con parámetro de escala $\boldsymbol{\Sigma}$ *y* n *grados de libertad.*
Dadas las variables aleatorias $\mathbf{Z} \sim N_p(\mathbf{0}, \mathbf{I})$ *y* $\mathbf{Q} \sim W_p(\mathbf{I}, n)$ *estocásticamente independientes, la variable aleatoria*

$$T^2 = n \, \mathbf{Z}' \, \mathbf{Q}^{-1} \, \mathbf{Z} \sim T^2(p, n)$$

sigue una ley T^2 *de Hotelling con* p *y* n *grados de libertad. Si* $p = 1$, *entonces* $T^2(1, n)$ *es el cuadrado de una variable aleatoria con ley* t *de Student y* n *grados de libertad. En general,* $T^2(p, n)$ *es proporcional a una* F *de Fisher*

$$\frac{n - p + 1}{n \, p} \, T^2(p, n) = F(p, n - p + 1).$$

La variable T^2 *se utiliza de manera análoga a la ley* t *de Student, en contrastes sobre medias multivariantes.*
Para p *y* n *fijos, genérese una muestra de tamaño* N *de una ley* $T^2(p, n)$ *de Hotelling. Represéntense los resultados mediante un histograma.*

Problema 5:

Si $\mathbf{A} \sim W_p(\mathbf{\Sigma}, a)$ *y* $\mathbf{B} \sim W_p(\mathbf{\Sigma}, b)$ *son independientes,* $\mathbf{\Sigma}$ *es regular y* $a \geq p$*, la variable aleatoria*

$$\Lambda = \frac{|\mathbf{A}|}{|\mathbf{A} + \mathbf{B}|}$$

tiene una ley Lambda de Wilks , $\Lambda(p, a, b)$*, con parámetros* p*,* a *y* b*.*
La ley Λ *no depende del parámetro* $\mathbf{\Sigma}$ *de* $\mathbf{A}$ *y* $\mathbf{B}$*, por lo que es suficiente considerarla para* $\mathbf{\Sigma} = \mathbf{I}$*. Tiene la misma distribución que un producto de* b *v.a. independientes con distribución Beta, es decir, si* $L \sim \Lambda(p, a, b)$ *entonces*

$$L = \prod_{i=1}^{b} u_i, \quad \text{donde } u_i \sim \text{Beta}\left(\frac{a+i-p}{2}, \frac{p}{2}\right).$$

Genérese una muestra de tamaño N *de una ley* Λ *de Wilks. Represéntense los resultados mediante un histograma.*

Problema 6:

La Tabla 3.1 contiene las medidas de 5 variables biométricas sobre gorriones hembra, recogidos casi moribundos después de una tormenta. Los primeros 21 sobrevivieron mientras que los 28 restantes no lo consiguieron. Las variables son X_1 = longitud total, X_2 = extensión del ala, X_3 = longitud del pico y de la cabeza, X_4 = longitud del húmero y X_5 = longitud del esternón.
Realícense comparaciones de medias y de covarianzas entre el grupo de supervivientes y el de no supervivientes.

La tabla 3.1 está disponible en Aula digital, pestaña Actividades, con el nombre "gorriones.xlsx".

Ayuda:

Comparación de covarianzas. Supondremos que X es una muestra aleatoria simple de tamaño n_X de una ley normal multivariante $\mathbf{X} \sim N_5(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X)$ y que Y es otra muestra aleatoria simple independiente de la anterior y de tamaño n_Y de una ley normal multivariante $\mathbf{Y} \sim N_5(\boldsymbol{\mu}_Y, \boldsymbol{\Sigma}_Y)$. Queremos contrastar la hipótesis de igualdad de covarianzas, es decir:

$$H_0 : \boldsymbol{\Sigma}_X = \boldsymbol{\Sigma}_Y = \boldsymbol{\Sigma}$$

Utilizaremos el contraste de la razón de verosimilitudes, cuyo estadístico es

$$\lambda_R = \frac{|\mathbf{S}_X|^{n_X/2}\,|\mathbf{S}_Y|^{n_Y/2}}{|\mathbf{S}|^{n/2}},$$

donde $\mathbf{S}_X$ y $\mathbf{S}_Y$ son las matrices de covarianzas muestrales de cada grupo, $n = n_X + n_Y$ y S es la matriz de covarianzas común, que se obtiene mediante la siguiente ponderación:

$$\mathbf{S} = \frac{n_X\,\mathbf{S}_X + n_Y\,\mathbf{S}_Y}{n_X + n_Y}.$$

Bajo la hipótesis nula dada por (3.6), tenemos que

$$-2\log(\lambda_R) \sim \chi^2_q,$$

donde

$$q = (g-1)p(p+1)/2\,,$$

g es el número de grupos y p es el número de variables.

Para implementar este contraste

$$-2\log(\lambda_R) = n\,\log|\mathbf{S}| - (n_X \log|\mathbf{S}_X| + n_Y \log|\mathbf{S}_Y|).$$

Problema 7:

En una fábrica de zumos se diseña el siguiente procedimiento de control de calidad. Se toma una muestra piloto (véase la Tabla 3.2) de $n = 50$ extracciones de zumo cuando el proceso de fabricación funciona correctamente y en ella se mide la concentración de $p = 11$ aminoácidos, $\mathbf{X} = (X_1, \ldots, X_{11})'$. Supóngase que $\mathbf{X}$ sigue una distribución normal. A continuación cada día se observan estas mismas variables con objeto de detectar algún cambio significativo en la calidad del proceso (véase Tabla 3.3). Supóngase que estas sucesivas observaciones, $\mathbf{y}_i$, $i = 1, \ldots, 10$, son independientes de la muestra piloto y entre sí.

Constrúyase un gráfico de control para estos nuevos diez días como se indica a continuación. En primer lugar calcúlense la media $\bar{\mathbf{x}}$ y la matriz de covarianzas $\mathbf{S}$ para la muestra piloto. A continuación para la observación $\mathbf{y}_i$ constrúyase el estadístico

$$T^2(i) = \frac{n}{n+1}(\mathbf{y}_i - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{y}_i - \bar{\mathbf{x}})$$

que debería seguir una $T^2(p, n-1)$ si la distribución de $\mathbf{y}_i$ es la misma que la de la muestra piloto.

Represéntense secuencialmente los valores de $T^2(i)$ en un gráfico y márquese en él un límite de control $LC = \frac{(n-1)p}{n-p} F^{\alpha}(p, n-p)$, siendo α el nivel de significación que deseemos fijar ($\alpha = 0.05$, por ejemplo). Párese el proceso de fabricación el primer día i que una observación $\mathbf{y}_i$ esté fuera de la región de control, es decir, $\mathbf{y}_i > LC$.

Problema 8:

Con algunos programas de ordenador sólo se pueden generar muestras normales univariantes. Supongamos, sin embargo, que deseamos generar una muestra de un vector bidimensional $\mathbf{Y} = (Y_1, Y_2)'$ *con distribución* $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, *donde*

$$\boldsymbol{\mu} = (\mu_1, \mu_2)',$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sqrt{\sigma_{11}}\sqrt{\sigma_{22}}\rho \\ \sqrt{\sigma_{11}}\sqrt{\sigma_{22}}\rho & \sigma_{22} \end{pmatrix}$$

y ρ *denota la correlación entre* Y_1 *e* Y_2. *Entonces podemos recurrir al procedimiento que explicamos a continuación.*

(a) *genera observaciones normales univariantes e independientes entre sí, y para un tamaño muestral* n *a elegir, genérese una muestra*

$$\begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{pmatrix}$$

de un vector $\mathbf{X} = (X_1, X_2)'$ *con distribución* $N_2(\mathbf{0}, \mathbf{I})$.

(b) *Ahora consideremos las siguientes transformaciones lineales de* $\mathbf{X}$

$$\begin{aligned} Y_1 &= \mu_1 + \sqrt{\sigma_{11}} X_1 \\ Y_2 &= \mu_2 + \sqrt{\sigma_{22}}(\rho X_1 + \sqrt{1-\rho^2} X_2). \end{aligned}$$

Demuéstrese que $\mathbf{Y} = (Y_1, Y_2)'$ *sigue una distribución* $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Los ficheros de datos *gorriones.xlsx*, *tabla_3_2.txt* y *tabla_3_3.txt*, están disponibles en Aula Digital.

Parte 2:

Revisad si las variables cuantitativas del conjunto de datos que habéis seleccionado para trabajar normal multivariante. En caso de que lo fuese, realizad una comparación de medias que sea importante para vuestro problema. Esta actividad, debe subirse en el repositorio de Github de la entrega que habéis expuesto.