ESTADÍSTICA MULTIVARIANTE

UGR, GRADO EN MATEMÁTICAS

Curso Académico 2023-2024

José Miguel Angulo Ibáñez (*jmangulo@ugr.es*)

Departamento de Estadística e Investigación Operativa
Universidad de Granada

- **TEMA 2. Inferencia en la Distribución Normal Multivariante**

  Introducción: media muestral y matriz de covarianzas muestral

‘Muestra aleatoria simple’ vectorial

- Sea $\mathbf{X} = (X_1, \ldots, X_p)'$ un vector aleatorio con DNM no singular,

$$\mathbf{X} \sim N_p(\boldsymbol{\mu}, \Sigma) \quad (\Sigma > 0)$$

(Por ejemplo, consideremos que las variables representan, en una situación práctica, p magnitudes relativas a ‘características’ de los ‘individuos’ de una ‘población’, con esa distribución de probabilidad conjunta)

- Se trata de realizar la inferencia (bajo distintas formas) sobre la distribución, supuestos desconocidos total o parcialmente sus parámetros, a partir de una *muestra aleatoria simple* obtenida de ésta (*i. e.*, un conjunto de ‘observaciones’ dadas en términos de *variables aleatorias independientes e idénticamente distribuidas*, con igual distribución que $\mathbf{X}$), que denotaremos

$$\{\mathbf{X}_\alpha : \ \alpha = 1, \ldots, N\} \qquad (N = \text{‘tamaño muestral’}),$$

siendo $X_{\alpha i}$ la componente i-ésima (correspondiente a la variable X_i) de la observación (vectorial) α-ésima, $\mathbf{X}_\alpha$

‘Muestra aleatoria simple’ vectorial (cont.)

- Se utilizará la misma notación pero en minúsculas cuando las variables de una muestra aleatoria simple tomen valores numéricos concretos, es decir, se tenga una realización de ésta:

$$\{\mathbf{x}_\alpha : \ \alpha = 1, \ldots, N\} \qquad (N = \text{‘tamaño muestral’}),$$

- A efectos algebraicos y computacionales, es frecuente disponer las variables componentes de la muestra en forma de ‘matriz de datos’ (aleatoria),

$$\mathsf{X} = \begin{pmatrix} \mathbf{X}'_1 \\ \vdots \\ \mathbf{X}'_\alpha \\ \vdots \\ \mathbf{X}'_p \end{pmatrix} = \begin{pmatrix} X_{11} & \cdots & X_{1i} & \cdots & X_{1p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{\alpha 1} & \cdots & X_{\alpha i} & \cdots & X_{\alpha p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{N1} & \cdots & X_{Ni} & \cdots & X_{Np} \end{pmatrix}$$

(y análogamente, usando minúsculas, en el caso de las correspondientes realizaciones de la muestra)

Medidas muestrales de primer y segundo orden

- En primer lugar, como paso previo a otros aspectos de la inferencia, se abordará en el apartado siguiente la *estimación puntual* de los parámetros $\boldsymbol{\mu}$ y Σ. Se adoptará el criterio basado en la *maximización de la función de verosimilitud*

  Como se verá, los *estimadores máximo-verosímiles* de $\boldsymbol{\mu}$ y Σ se expresarán en términos de medidas descriptivas de la muestra dadas por el *vector de medias muestral* y la *matriz de dispersiones muestral* (respecto del vector de medias muestral)

- VECTOR DE MEDIAS MUESTRAL ($p \times 1$)

$$\bar{\mathbf{X}} := \frac{1}{N}\sum_{\alpha=1}^{N}\mathbf{X}_\alpha = \begin{pmatrix} \frac{1}{N}\sum_{\alpha=1}^{N} X_{\alpha 1} \\ \vdots \\ \frac{1}{N}\sum_{\alpha=1}^{N} X_{\alpha p} \end{pmatrix} =: \begin{pmatrix} \bar{X}_1 \\ \vdots \\ \bar{X}_p \end{pmatrix}$$

  (cada componente $\bar{X}_i = \frac{1}{N}\sum_{\alpha=1}^{N} X_{\alpha i}$ representa la media muestral para la variable correspondiente X_i, $i = 1, \ldots, p$)

  $\bar{\mathbf{X}}$ puede verse como el 'centroide' (aleatorio) de la muestra en $\mathbb{R}^p$

- MATRIZ DE DISPERSIONES MUESTRAL ($p \times p$)

$$A := \sum_{\alpha=1}^{N} (\mathbf{X}_\alpha - \bar{\mathbf{X}})(\mathbf{X}_\alpha - \bar{\mathbf{X}})'$$

$$= \begin{pmatrix} \sum_{\alpha=1}^{N}(x_{\alpha 1} - \bar{x}_1)^2 & \sum_{\alpha=1}^{N}(x_{\alpha 1} - \bar{x}_1)(x_{\alpha 2} - \bar{x}_2) & \cdots & \sum_{\alpha=1}^{N}(x_{\alpha 1} - \bar{x}_1)(x_{\alpha p} - \bar{x}_p) \\ \sum_{\alpha=1}^{N}(x_{\alpha 2} - \bar{x}_2)(x_{\alpha 1} - \bar{x}_1) & \sum_{\alpha=1}^{N}(x_{\alpha 2} - \bar{x}_2)^2 & \cdots & \sum_{\alpha=1}^{N}(x_{\alpha 2} - \bar{x}_2)(x_{\alpha p} - \bar{x}_p) \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{\alpha=1}^{N}(x_{\alpha p} - \bar{x}_p)(x_{\alpha 1} - \bar{x}_1) & \sum_{\alpha=1}^{N}(x_{\alpha p} - \bar{x}_p)(x_{\alpha 2} - \bar{x}_2) & \cdots & \sum_{\alpha=1}^{N}(x_{\alpha p} - \bar{x}_p)^2 \end{pmatrix}$$

A partir de la matriz de dispersiones muestral, se definen:

- MATRIZ DE COVARIANZAS MUESTRAL: $S_N = \frac{A}{N} = \left(s_{ij}\right)_{i,j=1,\ldots,p}$
- MATRIZ DE CUASI-COVARIANZAS MUESTRAL: $S_{N-1} = \frac{A}{N-1}$
- MATRIZ DE CORRELACIONES MUESTRAL: $R = \left(\frac{s_{ij}}{s_{ii}^{\frac{1}{2}} s_{jj}^{\frac{1}{2}}}\right)_{i,j=1,\ldots,p}$

  (puede escribirse como $R = D^{-\frac{1}{2}} S_N D^{-\frac{1}{2}}$ con $D = \mathrm{diag}(s_{11}, \ldots, s_{pp})$)

Fórmula de momentos multivariante muestral

- LEMA: Sea $\{\mathbf{X}_\alpha : \alpha = 1, \ldots, N\}$ una muestra aleatoria de una distribución p-dimensional, y sea $\bar{\mathbf{X}}$ el vector de medias muestral. Entonces, para cualquier vector $\mathbf{b} \in \mathbb{R}^p$, se verifica que

$$\sum_{\alpha=1}^{N} (\mathbf{X}_\alpha - \mathbf{b})(\mathbf{X}_\alpha - \mathbf{b})' = \sum_{\alpha=1}^{N} \left(\mathbf{X}_\alpha - \bar{\mathbf{X}}\right)\left(\mathbf{X}_\alpha - \bar{\mathbf{X}}\right)' + N(\bar{\mathbf{X}} - \mathbf{b})(\bar{\mathbf{X}} - \mathbf{b})'$$
$$= A + N(\bar{\mathbf{X}} - \mathbf{b})(\bar{\mathbf{X}} - \mathbf{b})'$$

($\square$ Probar)

En particular, se tiene:

- $\boxed{\text{Con } \mathbf{b} = \boldsymbol{\mu}}$ (vector de medias poblacional, si existe)

$$\sum_{\alpha=1}^{N} (\mathbf{X}_\alpha - \boldsymbol{\mu})(\mathbf{X}_\alpha - \boldsymbol{\mu})' = A + N(\bar{\mathbf{X}} - \boldsymbol{\mu})(\bar{\mathbf{X}} - \boldsymbol{\mu})'$$

- $\boxed{\text{Con } \mathbf{b} = \mathbf{0}}$

$$\sum_{\alpha=1}^{N} \mathbf{X}_\alpha \mathbf{X}_\alpha' = A + N\bar{\mathbf{X}}\bar{\mathbf{X}}'$$

('fórmula de momentos multivariante muestral')