TEMA 1
INTRODUCCIÓN A LA ESTADÍSTICA MULTIVARIANTE.
INFERENCIA EN POBLACIONES NORMALES

Curso 2017-18

1 Introducción. Conceptos básicos de inferencia estadística.

- Muestra aleatoria
- Estadístico
- Estimador
- Intervalos de Confianza
- Test de hipótesis. Test de Razón de Verosimilitudes
- Contrastes en poblaciones normales

  Una muestra - Dos muestras independientes - Dos muestras relacionadas

1.1 Características poblaciones y muestra aleatoria:

Caso unidimensional: X

- Características (parámetros) poblacionales
  - Esperanza: $E(X) = \mu$
  - Varianza: $Var(X) = \sigma^2$

- Muestra aleatoria de X
  - $X_1, ..., X_n$ independientes
  - $X_1, ..., X_n$ idénticamente distribuidas a X
  - $E\left(\overline{X}\right) = \mu; \quad var\left(\overline{X}\right) = \dfrac{\sigma^2}{n}$
  - $E\left(S_c^2\right) = \sigma^2$

2 Características poblaciones y muestra aleatoria

Caso $p-$dimensional: $\underline{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix} = [X_1, ..., X_p]'$

2.1 Características (parámetros) poblacionales

- Vector de medias

$$E[X_i] = \mu_i,$$

$$\begin{bmatrix} E[X_1] \\ \vdots \\ E[X_p] \end{bmatrix} = \underline{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_p \end{bmatrix}$$

- Matriz de covarianzas (matriz de varianzas y covarianzas o matriz dispersión)

$$\sigma_{jk} = Cov\left(X_j, X_k\right) = E\left(\left(X_j - \mu_j\right)(X_k - \mu_k)\right)$$

$$\sigma_j^2 = \sigma_{jj} = Cov\left(X_j, X_j\right) = Var\left(X_j\right)$$

$$\mathbf{\Sigma} = Cov(\underline{X}) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix} = \left[\sigma_{jk}\right]_{j,k=1,\ldots,p}$$

- Matriz de correlaciones

$$\rho_{jk} = \frac{\sigma_{jk}}{\sigma_j \sigma_k}, \qquad \rho_{jj} = 1$$

$$\mathbf{R} = (\rho_{jk})_{j,k=1,\ldots,p}$$

$$\mathbf{R} = \mathbf{\Delta}^{-1}\mathbf{\Sigma}\mathbf{\Delta}^{-1}$$

con $\mathbf{\Delta} = diag\{\sigma_1, \ldots, \sigma_p\}$

Vector de variables estandarizadas

$$\underline{X}_{(s)} = \begin{bmatrix} \frac{X_1-\mu_1}{\sigma_1} \\ \vdots \\ \frac{X_p-\mu_p}{\sigma_p} \end{bmatrix} = \mathbf{\Delta}^{-1}\left(\underline{X} - \underline{\mu}\right)$$

$$\mathbf{R} = \mathbf{\Sigma}_{(s)} = \mathbf{R}_{(s)}$$

Notación

$$\underline{X} \sim \left(\underline{\mu}, \mathbf{\Sigma}\right)$$

2.2 Muestra aleatoria de $\underline{X} \sim \left(\underline{\mu}, \mathbf{\Sigma}\right)$

- $\underline{X}_1,\ \underline{X}_2, ..., \underline{X}_n$ independientes e idénticamente distribuidas a $\underline{X}$

  – Elementos muestrales

$$\underline{X}_i = \begin{bmatrix} X_{i1} \\ X_{i2} \\ \vdots \\ X_{ip} \end{bmatrix}, \quad \underline{X}_i^t = \begin{pmatrix} X_{i1} & X_{i2} & \ldots & X_{ip} \end{pmatrix}, \quad i = 1, 2, ..., n$$

- **Matriz (aleatoria) de datos muestrales**

$$\underset{n \times p}{\mathbf{X}} = \begin{bmatrix} \underline{X}_1^t \\ \underline{X}_2^t \\ \vdots \\ \underline{X}_n^t \end{bmatrix} = \begin{bmatrix} X_{11} & X_{12} & \ldots & X_{1p} \\ X_{21} & X_{22} & \ldots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \ldots & X_{np} \end{bmatrix}$$

  – X_{ij} es independiente de X_{kt}, para $i \neq k$

  – X_{ij} no es independiente de X_{it}

3 Estadísticos muestrales

3.1 Vector de medias

$$\underline{\overline{X}} = \frac{1}{n}\sum_{i=1}^{n} \underline{X}_i = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} X_{i1} \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} X_{ip} \end{bmatrix} = \begin{bmatrix} \overline{X}_1 \\ \vdots \\ \overline{X}_p \end{bmatrix}$$

$$\underline{\overline{X}} = \frac{1}{n}\mathbf{X}^t \underline{1}_n \qquad \left(\underline{1}_n = (1, ..., 1)^t\right)$$

- Propiedad

$$\sum_{i=1}^{n} \left(\underline{X}_i - \underline{\overline{X}}\right) = \underline{0}$$

3.1.1 Matriz de datos centrados

$$\widetilde{\mathbf{X}} = \begin{bmatrix} X_{11} - \overline{X}_1 & X_{12} - \overline{X}_2 & \dots & X_{1p} - \overline{X}_p \\ X_{21} - \overline{X}_1 & X_{22} - \overline{X}_2 & \dots & X_{2p} - \overline{X}_p \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} - \overline{X}_1 & X_{n2} - \overline{X}_2 & \dots & X_{np} - \overline{X}_p \end{bmatrix} = \begin{bmatrix} \underline{X}_1^t - \underline{\overline{X}}^t \\ \underline{X}_2^t - \underline{\overline{X}}^t \\ \vdots \\ \underline{X}_n^t - \underline{\overline{X}}^t \end{bmatrix}$$

$$\widetilde{\mathbf{X}} \;=\; \mathbf{X} - \underline{1}_n \underline{\overline{X}}^t = \mathbf{X} - \frac{1}{n} \underbrace{\underline{1}_n \underline{1}^t{}_n}_{\mathbf{E}_n} \mathbf{X} = \mathbf{X} - \frac{1}{n}\, \mathbf{E}_n \; \mathbf{X} = \left[\underbrace{\mathbf{I}_n - \frac{1}{n}\, \mathbf{E}_n}_{\mathbf{H}} \right] \mathbf{X}$$

$\widetilde{\mathbf{X}} \;=\; \mathbf{HX},$ siendo $\mathbf{H}$ simétrica e idempotente $\left(\mathbf{H}^2 = \mathbf{H}\right)$.

3.1.2 Matriz de suma de cuadrados y productos cruzados (s.c.p.c.)

Suma de cuadrados

$$S_{jj} = \sum_{i=1}^{n} \left(X_{ij} - \overline{X}_j \right)^2 \quad 1 \leq j \leq p$$

Suma de productos cruzados

$$S_{jk} = \sum_{i=1}^{n} \left(X_{ij} - \overline{X}_j \right) \left(X_{ik} - \overline{X}_k \right) \quad 1 \leq j \neq k \leq p$$

Matriz s.c.p.c.

$$\mathbf{S} = \begin{bmatrix} S_{11} & S_{12} & \dots & S_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{p1} & S_{p2} & \dots & S_{pp} \end{bmatrix} = \left[S_{jk} \right]$$

$$\begin{aligned} \mathbf{S} &= \sum_{i=1}^{n} \left(\underline{X}_i - \underline{\overline{X}} \right) \left(\underline{X}_i - \underline{\overline{X}} \right)^t = \sum_{i=1}^{n} \underline{X}_i \, \underline{X}_i^t - n \underline{\overline{X}} \, \underline{\overline{X}}^t = \\ \mathbf{S} &= \widetilde{\mathbf{X}}^t \widetilde{\mathbf{X}} = \mathbf{X}^t \mathbf{H} \mathbf{X} \end{aligned}$$

3.2 Matriz de covarianzas muestrales

- (Cuasi-)varianza muestral de la j-ésima componente

$$\widehat{\sigma}_j^2 = \widehat{\sigma}_{jj} = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_{ij} - \overline{X}_j\right)^2 = \frac{1}{n-1} S_{jj} \qquad 1 \leq j \leq p$$

- (Cuasi-)covarianza muestral entre las componentes j y k

$$\widehat{\sigma}_{jk} = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_{ij} - \overline{X}_j\right)\left(X_{ik} - \overline{X}_k\right) = \frac{1}{n-1} S_{jk} \quad 1 \leq j \neq k \leq p$$

- **Matriz de varianzas y covarianzas muestrales**

$$\widehat{\mathbf{\Sigma}} = \left(\widehat{\sigma}_{jk}\right)_{j,k=1,\ldots,p}$$

  - $\underset{p\times p}{\widehat{\mathbf{\Sigma}}}$ es simétrica semidefinida positiva (con probabilidad 1)
  -

$$\hat{\mathbf{\Sigma}} = \frac{1}{n-1}\mathbf{S}$$

3.3 Matriz de correlaciones muestrales

Correlaciones poblacionales

$$\rho_{jk} = \frac{\sigma_{jk}}{\sigma_j \sigma_k}, \qquad \rho_{jj} = 1$$

$$\mathbf{R} = (\rho_{jk})_{1 \leq j,k \leq p} = \mathbf{\Delta}^{-1} \mathbf{\Sigma} \mathbf{\Delta}^{-1}$$

Correlaciones muestrales

$$\widehat{\rho}_{jk} = \frac{\widehat{\sigma}_{jk}}{\widehat{\sigma}_j \widehat{\sigma}_k}, \quad \widehat{\rho}_{jj} = 1$$

$$\widehat{\mathbf{R}} = \widehat{\mathbf{\Delta}}^{-1} \widehat{\mathbf{\Sigma}}\, \widehat{\mathbf{\Delta}}^{-1}$$

con $\widehat{\mathbf{\Delta}} = diag\{\widehat{\sigma}_1, ..., \widehat{\sigma}_p\}$

Vector de variables estandarizadas

$$\underline{X}_{(s)} = \begin{bmatrix} \frac{X_1-\mu_1}{\sigma_1} \\ \vdots \\ \frac{X_p-\mu_p}{\sigma_p} \end{bmatrix} = \mathbf{\Delta}^{-1}\left(\underline{X} - \underline{\mu}\right)$$

Vector de datos estandarizados

$$\underline{X}_{i(s)} = \begin{bmatrix} \frac{X_{i1}-\overline{X}_1}{\widehat{\sigma}_1} \\ \vdots \\ \frac{X_p-\overline{X}_p}{\widehat{\sigma}_p} \end{bmatrix} = \widehat{\mathbf{\Delta}}^{-1}\left(\underline{X_i} - \underline{\overline{X}}\right)$$

Matriz de datos estandarizados

$$\mathbf{X}_{(s)} = \begin{bmatrix} \frac{X_{11}-\overline{X}_1}{\widehat{\sigma}_1} & \frac{X_{12}-\overline{X}_2}{\widehat{\sigma}_2} & \dots & \frac{X_{1p}-\overline{X}_p}{\widehat{\sigma}_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{X_{n1}-\overline{X}_1}{\widehat{\sigma}_1} & \frac{X_{n2}-\overline{X}_2}{\widehat{\sigma}_2} & \dots & \frac{X_{np}-\overline{X}_p}{\widehat{\sigma}_p} \end{bmatrix} = \begin{bmatrix} \underline{X}_1^t - \underline{\overline{X}}^t \\ \underline{X}_2^t - \underline{\overline{X}}^t \\ \vdots \\ \underline{X}_n^t - \underline{\overline{X}}^t \end{bmatrix} \widehat{\mathbf{\Delta}}^{-1} = \widetilde{\mathbf{X}}\widehat{\mathbf{\Delta}}^{-1}$$

$$\widehat{\mathbf{R}} = \widehat{\mathbf{\Sigma}}_{(s)} = \widehat{\mathbf{R}}_{(s)}$$

- Nota

  - Matriz de datos $\mathbf{X}_{n\times p}$

  - Vector de medias: $\underline{\overline{X}}_{p\times 1}$

  - Matriz de s.c.p.d., covarianzas y correlaciones:

$$\mathbf{S}_{p\times p},\ \widehat{\boldsymbol{\Sigma}}_{p\times p} \text{ y } \widehat{\mathbf{R}}_{p\times p}$$

- **Transformaciones lineales:** Dada $\mathbf{A}_{m\times p}$, sea $\underline{Y}_{m\times 1} = \mathbf{A}\ \underline{X}$

  Muestra transformada $\underline{Y}_1, ..., \underline{Y}_n$ con $\underline{Y}_i = \mathbf{A}\ \underline{X}_i$.

  -

$$\underline{\overline{Y}} = \mathbf{A}\underline{\overline{X}}$$

  -

$$\begin{aligned}\mathbf{S}_Y &= \mathbf{A}\mathbf{S}_X\mathbf{A}^t \\ \widehat{\boldsymbol{\Sigma}}_Y &= \mathbf{A}\ \widehat{\boldsymbol{\Sigma}}_X\mathbf{A}^t\end{aligned}$$

4 Matrices aleatorias

- Sea $\left\{Z_{ij}, i = 1, ..., m; j = 1, ..., q\right\}$ un conjunto de variables aleatorias

- **Matriz aleatoria**

$$\mathbf{Z} = \begin{bmatrix} Z_{11} & Z_{12} & \dots & Z_{1q} \\ Z_{21} & Z_{22} & \dots & Z_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{m1} & Z_{m2} & \dots & Z_{mq} \end{bmatrix} = \left[Z_{ij} \right]$$

4.1 Operador esperanza. Propiedades

Definición 1 *(Supuesto que existe* $E\left(Z_{ij}\right)$*)*

Matriz esperanza o matriz de medias

$$E[\mathbf{Z}] = \left[\begin{array}{cccc} E[Z_{11}] & E[Z_{12}] & \dots & E[Z_{1q}] \\ \vdots & \vdots & \ddots & \vdots \\ E[Z_{m1}] & E[Z_{m2}] & \dots & E[Z_{mq}] \end{array}\right] = \left[E\left(Z_{ij}\right)\right]$$

Lema 1 *Sean* $\mathbf{Z}$ *e* $\mathbf{Y}$ *son dos matrices aleatorias de la misma dimensión. Sean* $\mathbf{A}$,$\mathbf{B}$ *y* $\mathbf{C}$ *matrices constantes de dimensiones adecuadas*

- $E[\mathbf{AZB}+\mathbf{C}] = \mathbf{A}\ E[\mathbf{Z}]\ \mathbf{B}+\mathbf{C}$

- $E[\mathbf{Z}+\mathbf{Y}] = E[\mathbf{Z}]+E[\mathbf{Y}]$

Lema 2 *Sea* $\underline{X} \sim \left(\underline{\mu}, \mathbf{\Sigma}\right)$

1. $\mathbf{\Sigma} = E\left(\left(\underline{X} - E\left(\underline{X}\right)\right)\left(\underline{X} - E\left(\underline{X}\right)\right)^t\right) = E[\underline{X}\,\underline{X}^t] - \underline{\mu\mu}^t$

2. $E\left(\mathbf{A}\underline{X} + \underline{b}\right) = \mathbf{A}\ E[\underline{X}] + \underline{b}$

3. $Cov\left(\mathbf{A}\underline{X} + \underline{b}\right) = \mathbf{\Sigma}_{\mathbf{A}\underline{X}+\underline{b}} = \mathbf{A\Sigma A}^t$

4. En particular para $\underline{c} \in \mathbb{R}^p$

$$\begin{aligned} E[\underline{c}^t\underline{X}] &= \underline{c}^t\underline{\mu} \\ Var[\underline{c}^t\underline{X}] &= \underline{c}^t\mathbf{\Sigma}\underline{c} \end{aligned}$$

5. $\mathbf{\Sigma}$ es semidefinida positiva $(\mathbf{\Sigma} \geq \mathbf{0})$

Definición 2 *Sean* $\underline{X} = (X_1, ..., X_p)^t$ *e* $\underline{Y} = (Y_1, ..., Y_q)^t$ *dos vectores aleatorios.*

Se define la **matriz de covarianzas** *entre* $\underline{X}$ *e* $\underline{Y}$*:*

$$Cov\left(\underline{X}, \underline{Y}\right) = \left[Cov\left(X_i, Y_j\right)\right]$$

Lema 3 *Propiedades*

- *Si* $\underline{X}$ *e* $\underline{Y}$ *son independientes entonces* $Cov\left(\underline{X}, \underline{Y}\right) = \mathbf{0}_{p\times q}$
- $Cov\left(\underline{X}\right) = Cov\left(\underline{X}, \underline{X}\right) = \mathbf{\Sigma}_{\underline{X}}$
- $Cov\left(\underline{X}, \underline{Y}\right) = E\left[\left(\underline{X} - E\left(\underline{X}\right)\right)\left(\underline{Y} - E\left(\underline{Y}\right)\right)^t\right]$
- $Cov\left(\underline{X}, \underline{Y}\right) = Cov\left(\underline{Y}, \underline{X}\right)^t$

- $Cov\left(\mathbf{A}\underline{X}+\underline{c},\mathbf{B}\underline{Y}+\underline{d}\right)=\mathbf{A}Cov\left(\underline{X},\underline{Y}\right)\mathbf{B}^{t}$

- $Cov\left(\underline{X}_1+\underline{X}_2,\underline{Y}\right)=Cov\left(\underline{X}_1,\underline{Y}\right)+Cov\left(\underline{X}_2,\underline{Y}\right)$

- *Si* $p=q$, *entonces* $\mathbf{\Sigma}_{\underline{X}+\underline{Y}}=\mathbf{\Sigma}_{\underline{X}}+Cov\left(\underline{X},\underline{Y}\right)+Cov\left(\underline{Y},\underline{X}\right)+\mathbf{\Sigma}_{\underline{Y}}$

- $Cov\left(\sum_{i=1}^{n}\underline{X}_i\right)=\sum_{i=1}^{n}\mathbf{\Sigma}_{\underline{X}_i}+\sum_{i\neq j}Cov\left(\underline{X}_i,\underline{X}_j\right)$

5 Propiedades de los estadísticos muestrales

Proposición 4 *Sea* $\underline{X}_1,\ \underline{X}_2, ..., \underline{X}_n$ *una m.a. de un vector aleatorio* $\underline{X}$ p*-dimensional,* $\underline{X} \sim (\mu, \Sigma)$.

Se verifica:

$$(1)\ E[\overline{\underline{X}}] = \underline{\mu} \quad ; \quad (2)\ Cov\left[\overline{\underline{X}}\right] = \frac{1}{n}\boldsymbol{\Sigma} \quad ; \quad (3)\ E\left(\widehat{\boldsymbol{\Sigma}}\right) = E[\frac{1}{n-1}\mathbf{S}] = \boldsymbol{\Sigma}$$

- El elemento j de $\overline{\underline{X}}$ es un estimador insesgado de μ_j.

- El elemento (j, k) de $\widehat{\boldsymbol{\Sigma}}$ es un estimador insesgado de σ_{jk}.

Definición 3

$$\textit{Varianza muestral generalizada} \ : \ \left|\widehat{\boldsymbol{\Sigma}}\right| = \left|\frac{1}{n-1}\mathbf{S}\right|$$

$$\textit{Varianza muestral total} \ : \ tr(\widehat{\boldsymbol{\Sigma}}) = \frac{1}{n-1}tr(\mathbf{S})$$

$$\textit{Varianza muestral efectiva} \ : \ \left|\widehat{\boldsymbol{\Sigma}}\right|^{1/p} = \frac{1}{n-1}\left|\mathbf{S}\right|^{1/p}$$