Tema 1
Resumen: Inferencia en poblaciones normales

Curso 2017-18

1 Distribución normal

1.1 Caso unidimensional: $X \sim N\left(\mu, \sigma^2\right)$

$$f\left(x\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2\sigma^2}\left(x-\mu\right)^2 \right\}, \qquad -\infty < x < \infty$$

1.2 Caso multidimensional: $\underline{X} \sim N_p(\underline{\mu}, \mathbf{\Sigma})$

$$f\left(\underline{x}\right) = f\left(x_1, ..., x_p\right) = \frac{1}{(2\pi)^{p/2}\left|\mathbf{\Sigma}\right|^{1/2}} \quad \exp\left\{ -\frac{1}{2}(\underline{x}-\underline{\mu})'\mathbf{\Sigma}^{-1}(\underline{x}-\underline{\mu}) \right\}, \qquad -\infty < x_i < \infty, i = 1, ..., p$$

2 Distribución ji-cuadrado - Distribución Wishart

2.1 Distribución ji-cuadrado

Sea $X_1, ..., X_n$ una m.a. de $N\left(0, \sigma^2\right)$.

$$\sum_{i=1}^{n} \frac{X_i^2}{\sigma^2} \sim \chi_n^2$$

$$\sum_{i=1}^{n} X_i^2 \sim \sigma^2 \chi_n^2$$

2.2 Distribución Wishart: $\mathbf{W} \sim W_p(n, \mathbf{\Sigma})$

Sea $\underline{X}_1, \ldots, \underline{X}_n$ una m.a. de $N_p(\mathbf{0}, \mathbf{\Sigma})$ y sea $\mathbf{X} = [\underline{X}_1 \ldots \underline{X}_n]'$.

La matriz aleatoria $\mathbf{W} = \mathbf{X}'\mathbf{X} = \sum_{i=1}^{n} \underline{X}_i \underline{X}_i'$ se dice que se distribuye según una **distribución de Wishart** no singular p-dimensional con n grados de libertad y matriz de escala $\mathbf{\Sigma}$.

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

3 Distribución t-student - distribución T^2 de Hotelling

3.1 Distribución t-student

Dadas $Z \sim N(0,1)$ y $V \sim \chi^2_n$, independientes

$$T = \frac{Z}{\sqrt{V/n}} \sim t_n \qquad \text{equivalentemente} \qquad T^2 = n\frac{Z^2}{V} \sim F_{1,n}$$

3.2 Distribución T^2 de Hotelling: $T^2 \sim T^2_{p,n}$.

Sean $\underline{d} \sim N_p(\underline{\mathbf{0}}, \mathbf{I}_p)$ y $\mathbf{W} \sim W_p(n, \mathbf{I}_p)$ independientes.

Se denomina distribución T^2 de Hotelling a la distribución del estadístico

$$T^2 = n\underline{d}^t \mathbf{W}^{-1} \underline{d},$$

Nota. $T^2_{p,n}$ es una distribución unidimensional.

Lema. Sea $\underline{X} \sim N_p(\underline{\mu}, \mathbf{\Sigma})$ y $\mathbf{W} \sim W_p(n, \mathbf{\Sigma})$ independientes.

Entonces

$$T^2 = n(\underline{X} - \underline{\mu})'\mathbf{W}^{-1}(\underline{X} - \underline{\mu}) \sim T^2_{p,n}$$

Lema. Sea $T^2 \sim T^2_{p,n}$. Se verifica

$$\frac{n-p+1}{np} T^2 \sim \mathcal{F}_{p,n-p+1}$$

o equivalentemente

$$T^2 \sim \frac{np}{n-p+1} \mathcal{F}_{p,n-p+1}.$$

4 Estimación de parámetros

Sea $\underline{X}_1, \underline{X}_2, ..., \underline{X}_n$ una muestra aleatoria de una población $N_p(\underline{\mu}, \mathbf{\Sigma})$.

- Vector media muestral: $\overline{\underline{X}} = \frac{1}{n}\sum_{i=1}^{n}\underline{X}_i$
- Matriz suma de cuadrados y productos cruzados (s.c.p.c.)

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

- Matriz de varianzas y covarianzas muestrales

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

Teorema

1. $\overline{\underline{X}} \sim N_p(\underline{\mu}, \frac{1}{n}\mathbf{\Sigma})$.
   (a) $\overline{\underline{X}}$ es un estimador insesgado de $\underline{\mu}$..
2. $(n-1)\,\widehat{\mathbf{\Sigma}} = \mathbf{S} \sim W_p(n-1, \mathbf{\Sigma})$
   (a) $\widehat{\mathbf{\Sigma}}$ es un estimador insesgado de $\mathbf{\Sigma}$.
3. $\overline{\underline{X}}$ y $\widehat{\mathbf{\Sigma}}$ son independientes.
4. $\overline{\underline{X}}$ y $\widetilde{\mathbf{\Sigma}} = \frac{1}{n}\mathbf{S}$ son los estimadores de máxima verosimilitud de $\underline{\mu}$ y $\mathbf{\Sigma}$.

5 Contraste de hipótesis sobre el vector de medias de una población.

5.1 Caso univariante

Sea $X_1, ..., X_n$ una m.a de una $N(\mu, \sigma^2)$, con σ desconocida.

- Estadístico del TRV para contrastar $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$

$$\frac{\overline{X} - \mu_0}{S_c}\sqrt{n} \quad \text{o bien} \quad n\left(\frac{\overline{X} - \mu_0}{S_c}\right)^2$$

- Distribución del estadístico bajo H_0

$$\frac{\overline{X} - \mu_0}{S_c}\sqrt{n} \overset{H_0}{\sim} t_{n-1} \quad \text{o bien} \quad n\left(\frac{\overline{X} - \mu_0}{S_c}\right)^2 \overset{H_0}{\sim} \mathcal{F}_{1,n-1}$$

- Región crítica para un test de tamaño α

$$\left|\frac{\overline{X} - \mu_0}{S_c}\right|\sqrt{n} \geq t_{n-1\,1-\alpha/2} \quad \text{o bien} \quad n\left(\frac{\overline{X} - \mu_0}{S_c}\right)^2 \geq \mathcal{F}_{1,n-1,1-\alpha}$$

5.2 Caso multivariante

Sea $\underline{X}_1, ..., \underline{X}_n$ una m.a de una población $N_p\left(\underline{\mu}, \mathbf{\Sigma}\right)$, con $\mathbf{\Sigma}$ desconocida

- Estadístico TRV para contrastar las hipótesis $H_0 : \underline{\mu} = \underline{\mu}_0 \quad vs. \quad H_1 : \underline{\mu} \neq \underline{\mu}_0$

$$n\left(\overline{\underline{X}} - \underline{\mu}_0\right)' \widehat{\mathbf{\Sigma}}^{-1}\left(\overline{\underline{X}} - \underline{\mu}_0\right) = n(n-1)\left[\overline{\underline{X}} - \underline{\mu}_0\right]^t \mathbf{S}^{-1}\left[\overline{\underline{X}} - \underline{\mu}_0\right]$$

- Distribución del estadístico bajo H_0

$$n\left[\overline{\underline{X}} - \underline{\mu}_0\right]^t \widehat{\mathbf{\Sigma}}^{-1}\left[\overline{\underline{X}} - \underline{\mu}_0\right] \overset{H_0}{\sim} T^2_{p,n-1}$$

o

$$\frac{n-p}{p}\, n\left[\overline{\underline{X}} - \underline{\mu}_0\right]^t \mathbf{S}^{-1}\left[\overline{\underline{X}} - \underline{\mu}_0\right] \overset{H_0}{\sim} \mathcal{F}_{p,n-p}$$

- Región crítica para un test de tamaño α

$$\frac{n-p}{p}\, n\left[\overline{\underline{X}} - \underline{\mu}_0\right]^t \mathbf{S}^{-1}\left[\overline{\underline{X}} - \underline{\mu}_0\right] \geq \mathcal{F}_{p,n-p,1-\alpha}$$

- Región de confianza al nivel $(1-\alpha)$ para el vector de medias $\underline{\mu}$

$$\left\{\underline{\mu} \,/ \frac{n-p}{p}\, n\ \left[\overline{\underline{X}} - \underline{\mu}\right]^t \mathbf{S}^{-1}\left[\overline{\underline{X}} - \underline{\mu}\right] \leq \mathcal{F}_{p,n-p,1-\alpha}\right\}$$

Proposición.

Sea $\underline{X}_1, \underline{X}_2, ..., \underline{X}_n$ una muestra aleatoria de $\underline{X} \sim N_p[\underline{\mu}, \mathbf{\Sigma}]$.

Sea $\mathbf{A}$ una matriz $m \times p \quad (m \leq p)$, de rango m, y $\underline{d}$ un vector m-dimensional .

El estadístico del test de razón de verosimilitudes para el contraste:

$$H_0 : \mathbf{A}\underline{\mu} = \underline{d} \quad vs. \quad H_1 : \mathbf{A}\underline{\mu} \neq \underline{d}$$

viene dado por:

$$T^2 = (n-1)n \ \left[\mathbf{A}\overline{\underline{X}} - \underline{d}\right]^t \left[\mathbf{ASA}^t\right]^{-1} \left[\mathbf{A}\overline{\underline{X}} - \underline{d}\right]$$

y

$$T^2 \overset{H_0}{\sim} T^2_{m,n-1}.$$

- Consideremos n elementos de una población, sobre los que se realizan p medidas de una misma magnitud en p situaciones experimentales distintas.
- Puede considerarse la siguiente modelización:
  - $\underline{X} = (X_1, ..., X_p)'$ medidas de la misma magnitud en p situaciones experimentales distintas.
  - $\underline{X} \sim N_p\left(\underline{\mu}, \mathbf{\Sigma}\right)$ donde $\underline{\mu}= \left[\mu_1, ..., \mu_p\right]'$.
  - $\underline{X}_1, ..., \underline{X}_n$ m.a. de $\underline{X}$
- Hipótesis básica : igualdad de medias de la magnitud en todas las situaciones experimentales

$$H_0 : \mu_1 = ... = \mu_p$$

  - Equivalentemente

$$H_0 : \mu_1 - \mu_p = ... = \mu_{p-1} - \mu_p = 0 \Leftrightarrow H_0 : \mathbf{C}\underline{\mu}= \mathbf{0}$$

siendo

$$\mathbf{C} = \begin{bmatrix} 1 & 0 & .... & 0 & -1 \\ 0 & 1 & .... & 0 & -1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & .... & 1 & -1 \end{bmatrix}_{(p-1)\times p}$$

- **Solución del problema**
  - Transformar los datos $\underline{Y}_1 = \mathbf{C}\underline{X}_1, ..., \underline{Y}_n = \mathbf{C}\underline{X}_n$
  - Realizar el contraste

$$H_0 : \mu_{\underline{Y}} = \mathbf{0}$$

5.3.2 Test para la simetría

- Consideremos n individuos de una población, sobre los que se consideran p magnitudes o características en una situación A y las mismas magnitudes en otra situación B.
- Puede considerarse la siguiente modelización:
  - $\underline{X} = (X_1, ..., X_{2p})^{'}$, de forma que $X_1, ..., X_p$ medidas de p magnitudes en la situación A y $X_{p+1}, ..., X_{2p}$ medidas de las mismas magnitiudes en la situación B.
    * X_i y X_{i+p}, $i = 1, ..., p$ corresponden a la misma magnitud en situaciones A y B.
  - $\underline{X} \sim N_{2p}[\underline{\mu}, \mathbf{\Sigma}]$ donde

$$\underline{\mu} = \left[\underbrace{\mu_1, ..., \mu_p}_{A}, \underbrace{\mu_{p+1}, ..., \mu_{2p}}_{B}\right]^{'}.$$

  - $\underline{X}_1, ..., \underline{X}_n$ m.a. de $\underline{X}$
- La hipótesis de interés es

$$H_0 : \mu_i = \mu_{i+p} \; i = 1, ..., p$$

- Equivelentemente

$$H_0 : \mathbf{A}\underline{\mu} = \mathbf{0}$$

  donde $\mathbf{A} = [\mathbf{I}_p \;\; -\mathbf{I}_p]$.
- **Solución del problema**
  - Transformar los datos $\underline{Y}_1 = \mathbf{A}\underline{X}_1, ..., \underline{Y}_n = \mathbf{A}\underline{X}_n$
  - Realizar el contraste

$$H_0 : \mu_{\underline{Y}} = \mathbf{0}$$

6 Contraste de hipótesis sobre la igualdad de medias de dos poblaciones

6.1 Caso unidimesional

- Muestras independientes: $X_1, ..., X_n$ de $N\left(\mu_1, \sigma_1^2\right)$ y $Y_1, ..., Y_m$ de $N\left(\mu_2, \sigma_2^2\right)$
- Hipótesis básica

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 \\ H_1 &: \mu_1 \neq \mu_2 \end{aligned}$$

  - Estadístico (supuesto $\sigma_1^2 = \sigma_2^2$) y distribución bajo H_0

$$T = \frac{\overline{X} - \overline{Y}}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}} \overset{H_0}{\sim} t_{n+m-2}$$

    con

$$S_p^2 = \frac{(n-1)\, S_{c1}^2 + (m-1)\, S_{c2}^2}{n+m-2}$$

  - Región crítica a un nivel α

$$|T| \geq t_{n+m-2, 1-\alpha/2}$$

- Dos **muestras independientes** procedentes de dos poblaciones normales p-variantes con la misma matriz de varianzas y covarianzas

$$\underline{X}_1, \underline{X}_2, ..., \underline{X}_{n_1} \text{ i.i.d. } N_p\left(\underline{\mu}_1, \mathbf{\Sigma}\right)$$

$$\underline{Z}_1, \underline{Z}_2, ..., \underline{Z}_{n_2} \text{ i.i.d. } N_p\left(\underline{\mu}_2, \mathbf{\Sigma}\right)$$

- El problema fundamental a abordar es el contraste $H_0 : \underline{\mu}_1 = \underline{\mu}_2$.

- **Resultados previos y notación**
  - $\overline{\underline{X}} \sim N_p[\underline{\mu}_1, \frac{1}{n_1}\mathbf{\Sigma}]$,
  - $\overline{\underline{Z}} \sim N_p[\underline{\mu}_1, \frac{1}{n_2}\mathbf{\Sigma}]$
  - $\mathbf{S}_1 = (n_1 - 1)\,\widehat{\mathbf{\Sigma}}_1^{-1} = \sum_{j=1}^{n_1} \left(\underline{X}_j - \overline{\underline{X}}\right)\left(\underline{X}_j - \overline{\underline{X}}\right)^t \sim W_p(n_1 - 1, \mathbf{\Sigma})$
  - $\mathbf{S}_2 = (n_2 - 1)\,\widehat{\mathbf{\Sigma}}_2^{-1} = \sum_{j=1}^{n_2} \left(\underline{Z}_j - \overline{\underline{Z}}\right)\left(\underline{Z}_j - \overline{\underline{Z}}\right)^t \sim W_p(n_2 - 1, \mathbf{\Sigma})$
  - Sea
  $$\mathbf{S}_p = \mathbf{S}_1 + \mathbf{S}_2$$
  - La matriz
  $$\widehat{\mathbf{\Sigma}} = \frac{1}{n_1 + n_2 - 2}\,\mathbf{S}_p$$
  es un estimador insesgado de $\mathbf{\Sigma}$.

Teorema

1. $\overline{\underline{X}} - \overline{\underline{Z}} \sim N_p\left(\underline{\mu}_1 - \underline{\mu}_2, \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\mathbf{\Sigma}\right)$
2. $\mathbf{S}_1 + \mathbf{S}_2 \sim W_p(n_1 + n_2 - 2, \mathbf{\Sigma})$
3. $\overline{\underline{X}} - \overline{\underline{Z}}$ es independiente de $\mathbf{S}_1 + \mathbf{S}_2$
4. $\frac{n_1 n_2 (n_1+n_2-2)}{n_1+n_2}\left(\overline{\underline{X}} - \overline{\underline{Z}} - \left(\underline{\mu}_1 - \underline{\mu}_2\right)\right)^t \mathbf{S}_p^{-1}\left(\overline{\underline{X}} - \overline{\underline{Z}} - \left(\underline{\mu}_1 - \underline{\mu}_2\right)\right)^t \sim T^2_{p,n_1+n_2-2}$

Teorema En las condiciones anteriores,

1. El estadístico del TRV para el contraste:

$$H_0 : \underline{\mu}_1 = \underline{\mu}_2 \quad vs. \quad H_1 : \underline{\mu}_1 \neq \underline{\mu}_2$$

viene dado por

$$T^2 = \frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2} \quad \left[\overline{\underline{X}} - \overline{\underline{Z}}\right]^t \mathbf{S}_p^{-1} \left[\overline{\underline{X}} - \overline{\underline{Z}}\right]$$

2. Distribución bajo H_0

$$T^2 \overset{H_0}{\sim} T^2_{p,n_1+n_2-2}.$$

o equivalentemente

$$F = \frac{(n_1 + n_2 - p - 1)}{p} \frac{n_1 n_2}{n_1 + n_2} \quad \left[\overline{\underline{X}} - \overline{\underline{Z}}\right]^t \mathbf{S}_p^{-1} \left[\overline{\underline{X}} - \overline{\underline{Z}}\right] \overset{H_0}{\sim} \mathcal{F}_{p,(n_1+n_2-p-1)},$$

3. Región crítica para un test de tamaño α

$$F \geq \mathcal{F}_{p,(n_1+n_2-p-1),1-\alpha}.$$

Proposición. Sean dos muestras independientes:

$$\underline{X}_1, \underline{X}_2, ..., \underline{X}_{n_1} \sim N_p[\underline{\mu}_1, \Sigma] \; ; \; \underline{Z}_1, \underline{Z}_2, ..., \underline{Z}_{n_2} \sim N_p[\underline{\mu}_2, \Sigma]$$

y sea $\mathbf{A}$ una matriz de dimensiones $m \times p \, (m \leq p)$, de rango m.

El estadístico del test de razón de verosimilitudes para el contraste:

$$H_0 : \mathbf{A}\underline{\mu}_1 = \mathbf{A}\underline{\mu}_2 \quad vs. \quad H_1 : \mathbf{A}\underline{\mu}_1 \neq \mathbf{A}\underline{\mu}_2$$

viene dado por:

$$F = \frac{(n_1 + n_2 - m - 1)}{m(n_1 + n_2 - 2)} \frac{n_1 n_2}{n_1 + n_2} \left[\mathbf{A}(\overline{\underline{X}} - \overline{\underline{Z}})\right]^t (\mathbf{A}\mathbf{S}_p\mathbf{A}^t)^{-1} \left[\mathbf{A}(\overline{\underline{X}} - \overline{\underline{Z}})\right]$$

y

$$F \overset{H_0}{\sim} \mathcal{F}_{m,(n_1+n_2-m-1)}$$

Consideremos una batería de p magnitudes medidas sobre muestras de tamaño n_1 y n_2, obtenidas de dos poblaciones con vectores medias $\underline{\mu}_1 = \left(\mu_{11}, ..., \mu_{1p}\right)^t$ y $\underline{\mu}_2 = \left(\mu_{21}, ..., \mu_{2p}\right)^t$, respectivamente.

El gráfico obtenido por la unión de los puntos $(1, \mu_{11}), (2, \mu_{12}), ..., (p, \mu_{1p})$ se denomina *perfil de la primera población.*

Análogamente, la unión de los puntos $(1, \mu_{21}), (2, \mu_{22}), ..., (p, \mu_{2p})$ constituye el perfil de la segunda población.

- Cuestiones de interés sobre los perfiles
  - ¿Son paralelos?
  - ¿Tienen idéntico perfil medio?
- **Paralelismo de perfiles**

  La hipótesis de paralelismo se puede expresar como

$$H_0 : \mu_{1k} - \mu_{1(k-1)} = \mu_{2k} - \mu_{2(k-1)} \quad \forall k = 2, ..., p$$

o equivalentemente

$$H_0 : \mathbf{A}\underline{\mu}_1 = \mathbf{A}\underline{\mu}_2$$

donde

$$\mathbf{A} = \begin{bmatrix} -1 & 1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -1 & 1 \end{bmatrix} \in \mathcal{M}_{(p-1)\times p}$$

- **Mismo perfil medio**

  La hipótesis se puede expresar como

$$H_0 : \frac{1}{p}\sum_{i=1}^{p}\mu_{1i} = \frac{1}{p}\sum_{i=1}^{p}\mu_{2i} \Leftrightarrow H_0 : 1_p\underline{\mu}_1{=}1_p\underline{\mu}_2$$

  donde

$$1_p = [1\dots 1] \in \mathcal{M}_{1\times p}$$