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 1. Distribución Normal Multivariante**

  Formas cuadráticas basadas en vectores aleatorios con DNM

- Sea $\mathbf{X} = (X_1, \ldots, X_p)'$ un vector aleatorio y sea A una matriz (cte.) $(p \times p)$-dimensional. Se considera la variable aleatoria dada por la forma cuadrática

$$\mathbf{X}'A\mathbf{X}$$

  Se plantea, en general, el problema de estudiar la distribución de $\mathbf{X}'A\mathbf{X}$ a partir del conocimiento de la distribución de $\mathbf{X}$

  (Como extensión, si $\mathbf{X}$ se reemplaza por una matriz aleatoria de dimensión $p \times q$, la forma cuadrática resultante constituirá una matriz aleatoria de dimensión $q \times q$)

- En particular, se considera el caso en que $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$, que conduce a la *distribución χ^2 no centrada*

  La *distribución χ^2 no centrada* y, definida a partir de ésta, la *distribución F no centrada*, son fundamentales, como distribuciones de diversos estadísticos de interés, en relación con la inferencia basada en la DNM

DISTRIBUCIÓN χ^2 CENTRADA

- Recordemos que la *distribución* χ^2 *centrada* con n grados de libertad se define como la distribución de la suma de cuadrados de n variables aleatorias independientes con distribución normal estándar:

$$\mathbf{Z} = (Z_1, \ldots, Z_n)' \sim N_n(\mathbf{0}, I_n) \quad \longrightarrow \quad Y = \mathbf{Z}'\mathbf{Z} = \sum_{i=1}^{n} Z_i^2 \sim \chi^2_n$$

Función de densidad: $f_Y(y) = \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} y^{\frac{n}{2}-1} e^{-\frac{y}{2}}, \quad y > 0$

Función de distribución: $F_Y(y) = \frac{\gamma\left(\frac{n}{2}, \frac{y}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}, \quad y > 0$

Se definen, $\forall z \in \mathbb{C},\ \Re(z) > 0$,

- Función gamma: $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$
- Funciones gamma incompletas: para $v > 0$,

$$\Gamma(z, v) = \int_v^\infty t^{z-1} e^{-t} dt, \qquad \gamma(z, v) = \int_0^v t^{z-1} e^{-t} dt$$

DISTRIBUCIÓN χ^2 CENTRADA

- Recordemos que la *distribución* χ^2 *centrada* con n grados de libertad se define como la distribución de la suma de cuadrados de n variables aleatorias independientes con distribución normal estándar:

$$\mathbf{Z} = (Z_1, \ldots, Z_n)' \sim N_n(\mathbf{0}, I_n) \quad \longrightarrow \quad Y = \mathbf{Z}'\mathbf{Z} = \sum_{i=1}^{n} Z_i^2 \sim \chi^2_n$$

Función de densidad: $f_Y(y) = \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} y^{\frac{n}{2}-1} e^{-\frac{y}{2}}, \quad y > 0$

Función de distribución: $F_Y(y) = \frac{\gamma\left(\frac{n}{2}, \frac{y}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}, \quad y > 0$

Función característica: $\phi_Y(t) = (1 - 2it)^{-\frac{n}{2}}, \quad \forall t \in \mathbb{R},$

de donde se obtienen, en particular, los momentos

$$E[Y] = n$$
$$\text{Var}(Y) = 2n$$

DISTRIBUCIÓN χ^2 NO CENTRADA

- DEFINICIÓN (RESULTADO): Sea $\mathbf{X} = (X_1, \ldots, X_n)' \sim N_n(\boldsymbol{\mu}, I_n)$. Entonces, la variable aleatoria $Y = \mathbf{X}'\mathbf{X}$ tiene función de densidad

$$f_Y(y) = e^{-\frac{\delta}{2}}\, {}_0F_1\left(;\frac{1}{2}n;\frac{1}{4}\delta y\right)\frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}e^{-\frac{y}{2}}y^{\frac{n}{2}-1}, \quad \text{para } y > 0,$$

siendo $\delta = \boldsymbol{\mu}'\boldsymbol{\mu}$. Se dice que la variable Y tiene *distribución χ^2 no centrada* con n grados de libertad y parámetro de no centralidad δ, denotándose $\chi^2_n(\delta)$

(En la expresión anterior, ${}_0F_1$ representa la 'función hipergeométrica generalizada' de órdenes 0 y 1, también llamada *función hipergeométrica confluente límite*)

DISTRIBUCIÓN χ^2 NO CENTRADA

- DEFINICIÓN (RESULTADO): Sea $\mathbf{X} = (X_1, \ldots, X_n)' \sim N_n(\boldsymbol{\mu}, I_n)$. Entonces, la variable aleatoria $Y = \mathbf{X}'\mathbf{X}$ tiene función de densidad

$$f_Y(y) = e^{-\frac{\delta}{2}}\, {}_0F_1\left(;\frac{1}{2}n;\frac{1}{4}\delta y\right)\frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}e^{-\frac{y}{2}}y^{\frac{n}{2}-1}, \quad \text{para } y > 0,$$

siendo $\delta = \boldsymbol{\mu}'\boldsymbol{\mu}$. Se dice que la variable Y tiene *distribución χ^2 no centrada* con n grados de libertad y parámetro de no centralidad δ, denotándose $\chi^2_n(\delta)$

(Cuando $\boldsymbol{\mu} = \mathbf{0}$, se tiene la distribución χ^2_n centrada)

Función característica: $\phi_Y(t) = (1-2it)^{-\frac{n}{2}}e^{\frac{it\delta}{1-2it}}, \quad \forall t \in \mathbb{R},$

de donde se obtienen, en particular, los momentos

$$E[Y] = n + \delta$$
$$\text{Var}(Y) = 2n + 4\delta \qquad (\square \text{ Probar})$$

DISTRIBUCIÓN χ^2 NO CENTRADA (cont.)

- Un RESULTADO de interés:

  Si Y_1 y Y_2 son variables aleatorias independientes con

  $$Y_1 \sim \chi^2_{n_1}(\delta_1), \qquad Y_2 \sim \chi^2_{n_2}(\delta_2),$$

  entonces se tiene que

  $$Y_1 + Y_2 \sim \chi^2_{n_1+n_2}(\delta_1 + \delta_2) \qquad (\square \text{ Probar})$$

DISTRIBUCIÓN F CENTRADA

- Recordemos que la *distribución F centrada* con (n_1, n_2) grados de libertad se define como la distribución del cociente

$$F = \frac{\frac{Y_1}{n_1}}{\frac{Y_2}{n_2}},$$

con $Y_1 \sim \chi^2_{n_1}$ e $Y_2 \sim \chi^2_{n_2}$ (ambas centradas), independientes. Se denota F_{n_1,n_2}.

Puede verse como la distribución del cociente

$$\frac{\frac{1}{n_1}\mathbf{Z}'_{(1)}\mathbf{Z}_{(1)}}{\frac{1}{n_2}\mathbf{Z}'_{(2)}\mathbf{Z}_{(2)}} = \frac{\frac{1}{n_1}\sum_{k=1}^{n_1} Z_{1k}^2}{\frac{1}{n_2}\sum_{l=1}^{n_2} Z_{2k}^2},$$

con: $\mathbf{Z}_{(1)} = (Z_{11}, \ldots, Z_{1n_1})' \sim N_{n_1}(\mathbf{0}, I_{n_1})$

$\mathbf{Z}_{(2)} = (Z_{21}, \ldots, Z_{2n_1})' \sim N_{n_2}(\mathbf{0}, I_{n_2})$

$\mathbf{Z}_{(1)}$ y $\mathbf{Z}_{(2)}$ independientes;

equivalentemente,

$$\mathbf{Z} = \begin{pmatrix} \mathbf{Z}_{(1)} \\ \mathbf{Z}_{(2)} \end{pmatrix} \sim N_{n_1+n_2}(\mathbf{0}, I_{n_1+n_2})$$

DISTRIBUCIÓN F CENTRADA (cont.)

Función de densidad: con $n_1, n_2 \in \mathbb{N} - \{0\}$,

$$g_F(f) = \frac{1}{fB(\frac{n_1}{2}, \frac{n_2}{2})} \left(\frac{n_1 f}{n_1 f + n_2}\right)^{\frac{n_1}{2}} \left(1 - \frac{n_1 f}{n_1 f + n_2}\right)^{\frac{n_2}{2}}, \quad f \geq 0$$

Función de distribución: $G_F(f) = I_{\frac{n_1 f}{n_1 f + n_2}}\left(\frac{n_1}{2}, \frac{n_2}{2}\right), \quad f \geq 0$

Se definen, $\forall a, b \in \mathbb{C}$, con $\Re(a), \Re(b) > 0$,

- Función beta: $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \int_0^1 t^{a-1}(1-t)^{b-1}dt$
- Función beta incompleta: para $x \in [0,1]$,

  $B(x;a,b) = \int_0^x t^{a-1}(1-t)^{b-1}dt$

  (para $x = 1$ se tiene la función beta completa)
- Función beta incompleta regularizada: $I_x(a,b) = \frac{B(x;a,b)}{B(a,b)}$

DISTRIBUCIÓN F CENTRADA (cont.)

Función de densidad: con $n_1, n_2 \in \mathbb{N} - \{0\}$,

$$g_F(f) = \frac{1}{fB(\frac{n_1}{2}, \frac{n_2}{2})} \left(\frac{n_1 f}{n_1 f + n_2}\right)^{\frac{n_1}{2}} \left(1 - \frac{n_1 f}{n_1 f + n_2}\right)^{\frac{n_2}{2}},$$

Función de distribución: $G_F(f) = I_{\frac{n_1 f}{n_1 f + n_2}}\left(\frac{n_1}{2}, \frac{n_2}{2}\right) \quad f \geq 0$

Función característica (versión corregida de Phillips (1982)):

$$\phi_F(t) = \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_2}{2}\right)} U\left(\frac{n_1}{2}, 1 - \frac{n_2}{2}; -\frac{n_2}{n_1} it\right)$$

(En la expresión anterior, U representa la *función hipergeométrica confluente de segunda especie*,

$$U(a, b; z) = \frac{\Gamma(1-b)}{\Gamma(a+1-b)} {}_1F_1(a; b; z) + \frac{\Gamma(b-1)}{\Gamma(a)} z^{1-b} {}_1F_1(a+1-b; 2-b; z),$$

con ${}_1F_1$ la 'función hipergeométrica generalizada' de órdenes 1 y 1, también llamada *función de Kumar de primera especie*)

DISTRIBUCIÓN F CENTRADA (cont.)

Función de densidad: con $n_1, n_2 \in \mathbb{N} - \{0\}$,

$$g_F(f) = \frac{1}{fB(\frac{n_1}{2}, \frac{n_2}{2})} \left(\frac{n_1 f}{n_1 f + n_2}\right)^{\frac{n_1}{2}} \left(1 - \frac{n_1 f}{n_1 f + n_2}\right)^{\frac{n_2}{2}},$$

Función de distribución: $G_F(f) = I_{\frac{n_1 f}{n_1 f + n_2}}\left(\frac{n_1}{2}, \frac{n_2}{2}\right) \quad f \geq 0$

Función característica (versión corregida de Phillips (1982)):

$$\phi_F(t) = \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_2}{2}\right)} U\left(\frac{n_1}{2}, 1 - \frac{n_2}{2}; -\frac{n_2}{n_1} it\right)$$

Se obtienen, en particular, los momentos

$$E[F] = \frac{n_2}{n_2 - 2}, \quad \text{para } n_2 > 2 \quad (= \infty \quad \text{si } n_2 \in (0, 2])$$

$$\text{Var}(F) = \frac{2n_2^2(n_1 + n_2 - 2)}{n_1(n_2 - 2)^2(n_2 - 4)}, \quad \text{para } n_2 > 4 \quad (= \infty \quad \text{si } n_2 \in (2, 4])$$

DISTRIBUCIÓN F NO CENTRADA

- DEFINICIÓN (RESULTADO): Sean $Y_1 \sim \chi^2_{n_1}(\delta)$ e $Y_2 \sim \chi^2_{n_2}$, independientes. Entonces, la variable aleatoria

$$F = \frac{\frac{Y_1}{n_1}}{\frac{Y_2}{n_2}}$$

tiene función de densidad

$$g_F(f) = e^{-\frac{\delta}{2}}\, {}_1F_1\left(\frac{1}{2}(n_1+n_2); \frac{1}{2}n_1; \frac{-\frac{1}{2}\frac{n_1}{n_2}\delta f}{1+\frac{n_1}{n_2}f}\right) \times \frac{\Gamma\left(\frac{1}{2}(n_1+n_2)\right)}{\Gamma\left(\frac{1}{2}n_1\right)\Gamma\left(\frac{1}{2}n_2\right)} \frac{f^{\frac{n_1}{2}-1}\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}}}{\left(1+\frac{n_1}{n_2}f\right)^{\frac{(n_1+n_2)}{2}}}, \quad \text{para } f > 0$$

Se dice que F tiene *distribución F no centrada* con n_1 y n_2 grados de libertad y parámetro de no centralidad δ, denotándose $F_{n_1,n_2}(\delta)$.

DISTRIBUCIÓN F NO CENTRADA (cont.)

Función característica (versión corregida de Phillips (1982)):

$$\phi_F(t) = \frac{e^{-\frac{1}{2}\delta}}{\Gamma\left(\frac{1}{2}n_2\right)} \sum_{j=0}^{\infty} \frac{\left(\frac{\delta}{2}\right)^j}{j!} \Gamma\left(\frac{1}{2}n_1 + \frac{1}{2}n_2 + j\right) U\left(\frac{n_1}{2} + j, 1 - \frac{n_2}{2}; -\frac{n_1}{n_2}it\right)$$

Se obtienen, en particular, los momentos

$$E[F] = \frac{n_2(n_1+\delta)}{n_1(n_2-2)}, \quad \text{para } n_2 > 2 \quad (= \infty \quad \text{si } n_2 \in (0,2])$$

$$\text{Var}(F) = 2\left(\frac{n_2}{n_1}\right)^2 \frac{(n_1+\delta)^2 + (n_1+2\delta)(n_2-2)}{(n_2-2)^2(n_2-4)}, \quad \text{para } n_2 > 4$$

$$(= \infty \ \text{ si } n_2 \in (2,4])$$

Formas cuadráticas $\mathbf{X}'A\mathbf{X}$, con $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$ $(\Sigma > 0)$

A continuación veremos algunos resultados relativos a la distribución de formas cuadráticas del tipo

$$\mathbf{X}'A\mathbf{X},$$

con $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$ $(\Sigma > 0)$ y A una matrix (cte.) $(p \times p)$-dimensional simétrica.

- De forma preliminar, se puede hacer un **estudio directo de los momentos de primer y segundo orden**:

  Sea $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$ $(\Sigma > 0)$ y A una matrix (cte.) $(p \times p)$-dimensional simétrica. Entonces,

  (a) $E[\mathbf{X}'A\mathbf{X}] = \text{tr}(A\Sigma) + \boldsymbol{\mu}'A\boldsymbol{\mu}$

  (b) $\text{Var}(\mathbf{X}'A\mathbf{X}) = 2\text{tr}\left((A\Sigma)^2\right) + 4\boldsymbol{\mu}'A\Sigma A\boldsymbol{\mu}$

  (OBSERVACIÓN: El resultado (a) no requiere la hipótesis de normalidad, solo la existencia de la esperanza)

  (□ Probar (a))

- **RESULTADO 1**:

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

  1. $(\mathbf{X} - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{X} - \boldsymbol{\mu}) \sim \chi^2_p$
  2. $\mathbf{X}'\Sigma^{-1}\mathbf{X} \sim \chi^2_p(\delta)$, con $\delta = \boldsymbol{\mu}'\Sigma^{-1}\boldsymbol{\mu}$

  (□ Probar)

- **RESULTADO 2**:

  Sea $\mathbf{X} \sim N_p(\boldsymbol{\mu}, I_p)$ y sea B una matriz (cte.) $(p \times p)$-dimensional simétrica. Entonces, $\mathbf{X}'B\mathbf{X}$ tiene una distribución χ^2 no centrada si y solo si B es idempotente (*i. e.* $B^2 = B$), en cuyo caso los grados de libertad y el parámetro de no centralidad son, respectivamente, $k = \text{rango}(B) = \text{tr}(B)$ y $\delta = \boldsymbol{\mu}'B\boldsymbol{\mu}$.

  (□ Probar)

- **RESULTADO 3**:

  Sea $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$ $(\Sigma > 0)$. Supongamos el particionamiento

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_{(1)} \\ \mathbf{X}_{(2)} \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_{(1)} \\ \boldsymbol{\mu}_{(2)} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{(11)} & \Sigma_{(12)} \\ \Sigma_{(21)} & \Sigma_{(22)} \end{pmatrix},$$

  con $\mathbf{X}_{(1)}$ y $\boldsymbol{\mu}_{(1)}$ subvectores q-dimensionales y $\Sigma_{(11)}$ submatriz $(q \times q)$-dimensional. Entonces,

$$Q := (\mathbf{X} - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{X} - \boldsymbol{\mu}) - (\mathbf{X}_{(1)} - \boldsymbol{\mu}_{(1)})'\Sigma_{(11)}^{-1}(\mathbf{X}_{(1)} - \boldsymbol{\mu}_{(1)}) \sim \chi^2_{p-q}$$

  (□ Probar)

- **RESULTADO 4**:

  Sea $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$ $(\Sigma > 0)$ y sea B una matriz (cte.) $(p \times p)$-dimensional simétrica. Entonces, $\mathbf{X}'B\mathbf{X}$ tiene una distribución $\chi^2_k(\delta)$, con $k = \text{rango}(B)$ y $\delta = \boldsymbol{\mu}'B\boldsymbol{\mu}$, si y solo si $B\Sigma$ es idempotente (*i. e.* $(B\Sigma)^2 = B\Sigma$; equivalentemente, en este caso, $B\Sigma B = B$).

  (□ Probar)

APÉNDICE: Funciones hipergeométricas generalizadas

- DEFINICIÓN: Se denomina *función hipergeométrica generalizada* de órdenes p y q a

$$ {}_pF_q(a_1,\ldots,a_p;b_1,\ldots,b_q;z) := \sum_{k=0}^{\infty} \frac{(a_1)_k \cdots (a_p)_k}{(b_1)_k \cdots (b_p)_k} \frac{z^k}{k!}, $$

  donde $(a)_k = a(a+1)\cdots(a+k-1)$, siendo $a_1,\ldots,a_p,b_1,\ldots,b_q$ parámetros (posiblemente complejos) y $z \in \mathbb{C}$ el argumento de la función.

- Algunas OBSERVACIONES:
  - Ningún parámetro b_j puede ser 0 o un entero negativo (en este caso, uno de los denominadores de la serie sería 0 a partir de un cierto k)
  - Si algún parámetro en el numerador es 0 o un entero negativo, los términos de la serie se anulan a partir de un cierto k y queda un polinomio en z

APÉNDICE: Funciones hipergeométricas generalizadas (cont.)

- Algunas OBSERVACIONES (cont.):
  - La serie
    - converge para todo z finito si $p \leq q$
    - converge para $|z| < 1$ y diverge para $|z| > 1$ si $p = q + 1$
    - diverge para todo $z \neq 0$ si $p > q + 1$
  - El término '*generalizada*' se refiere a que ${}_pF_q$ es una generalización de la *función hipergeométrica* clásica (o *gaussiana*), ${}_2F_1$.
  - ${}_1F_1$ se denomina *función hipergeométrica confluente*
  - ${}_0F_1$, definida como

$$ {}_0F_1(;b;z) := \lim_{a\to\infty} {}_1F_1\left(a;b;\frac{z}{a}\right), $$

    se denomina *función hipergeométrica confluente límite*