

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

Distribución T^2 de Hotelling. Estadístico T^2 de Hotelling asociado al muestreo de una población con DNM. Aplicación al problema de contraste sobre el vector de medias μ , con Σ desconocida

Distribución T^2 de Hotelling

- **DEFINICIÓN:** Sea \mathbf{X} un vector aleatorio p -dimensional con distribución $N_p(\mu, \Sigma)$ ($\Sigma > 0$). Sea B una matriz aleatoria $(p \times p)$ -dimensional con distribución $W_p(n, \Sigma)$, $n \geq p$. Supongamos, además, que \mathbf{X} y B son independientes. Entonces, se define la *distribución T^2 de Hotelling no centrada*, con p y n grados de libertad y parámetro de no centralidad $\delta = \mu' \Sigma \mu$, como la distribución de la variable aleatoria (unidimensional) dada por la forma cuadrática

$$T^2 = n\mathbf{X}'B^{-1}\mathbf{X}$$

- La distribución anterior es conocida, y se relaciona con la distribución F de Snedecor no centrada del siguiente modo:

RESULTADO: Bajo las condiciones de la definición anterior, se tiene que

$$\frac{T^2}{n} \frac{n-p+1}{p} \sim F_{p,n-p+1}(\delta),$$

con parámetro de no centralidad $\delta = \mu' \Sigma^{-1} \mu$ (□ Probar)

- Si $\mu = \mathbf{0}$, se tiene entonces una $F_{p,n-p+1}$ centrada
- En el caso $p = 1$, se tiene que $T^2 \equiv F_{1,n}(\delta)$, con $\delta = (\frac{\mu}{\sigma})^2$

Estadístico T^2 de Hotelling (inferencia multivariante)

- Originalmente, el denominado 'estadístico T^2 de Hotelling' fue introducido por Harold Hotelling (1931) en el contexto de la inferencia estadística a partir del muestreo de una población con DNM (puede verse como una generalización, multivariante y no estandarizada, del estadístico t de Student).

Este uso se explica a continuación, como aplicación de la definición y el resultado anteriores:

- DEFINICIÓN:** Sea $\mathbf{X}_1, \dots, \mathbf{X}_N$ una muestra aleatoria simple de una distribución $N_p(\boldsymbol{\mu}, \Sigma)$ ($\Sigma > 0$), y sean $\bar{\mathbf{X}}$, A , S_N y S_n (con $n = N - 1$) los correspondientes estadísticos vector de medias, matriz de dispersiones y matrices de covarianzas y de quasi-covarianzas muestrales. Se define el *estadístico T^2 de Hotelling*, basado en la muestra, como

$$T^2 := n\bar{\mathbf{X}}' S_N^{-1} \bar{\mathbf{X}} = N\bar{\mathbf{X}}' S_n^{-1} \bar{\mathbf{X}} = nN\bar{\mathbf{X}}' A^{-1} \bar{\mathbf{X}}$$

(Formalmente, se escribiría $T^2(\mathbf{X}_1, \dots, \mathbf{X}_N)$, con $T^2 : \mathbb{R}^{Np} \rightarrow \mathbb{R}_0^+$ medible, aunque se suele sobreentender de forma implícita el argumento)

Estadístico T^2 de Hotelling (inferencia multivariante) (cont.)

- Por el resultado visto al inicio, se tiene el siguiente, donde se caracteriza convenientemente la distribución del estadístico T^2 en términos de la distribución F de Snedecor no centrada:

RESULTADO: Bajo las condiciones anteriores, el estadístico T^2 se distribuye como sigue:

$$\frac{T^2}{n} \frac{n-p+1}{p} \sim F_{p,n-p+1}(\delta),$$

o, equivalentemente,

$$\frac{T^2}{N-1} \frac{N-p}{p} \sim F_{p,N-p}(\delta),$$

con parámetro de no centralidad $\delta = N\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$

- **PLANTEAMIENTO:** Sea $\mathbf{X}_1, \dots, \mathbf{X}_N$ una muestra aleatoria simple extraída de una población con distribución $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ ($\boldsymbol{\Sigma} > 0$), siendo $N > p$.

Denotaremos $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_N)'$ (matriz correspondiente a la muestra aleatoria) y, correspondientemente, $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_N)'$ (matriz de datos muestrales observados).

PROBLEMA DE CONTRASTE: Se consideran las hipótesis

$$\begin{array}{ll} H_0 : & \boldsymbol{\mu} = \boldsymbol{\mu}_0 \quad \text{('hipótesis nula')} \\ H_1 : & \boldsymbol{\mu} \neq \boldsymbol{\mu}_0 \quad \text{('hipótesis alternativa')}, \end{array}$$

donde $\boldsymbol{\mu}_0 \in \mathbb{R}^p$ es un vector fijo dado.

- **DOS ESCENARIOS:**

- ▶ $\boldsymbol{\Sigma}$ conocida → Se resuelve a partir de la distribución $\mathcal{X}^2(\delta)$
- ▶ $\boldsymbol{\Sigma}$ desconocida → Se resuelve a partir de la distribución T^2

- Caso Σ conocida

Estadístico de contraste:

$$U := N(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)' \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{X}} - \boldsymbol{\mu}_0) \sim \mathcal{X}_p^2(\delta),$$

$$\text{con } \delta = N(\boldsymbol{\mu} - \boldsymbol{\mu}_0)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0)$$

Función ‘test’: Con $\phi : \mathbb{R}^{Np} \rightarrow \{0, 1\}$,

$$\phi(\mathbf{x}) = \begin{cases} 1 & \text{si } u := U(\mathbf{x}) > \mathcal{X}_{p,\alpha}^2 \\ 0 & \text{si } u := U(\mathbf{x}) \leq \mathcal{X}_{p,\alpha}^2 \end{cases}$$

para todo $\mathbf{x} \in \mathbb{R}^{Np}$, siendo $\mathcal{X}_{p,\alpha}^2$ el valor de una distribución \mathcal{X}^2 (centrada) con p grados de libertad que deja a su derecha una masa de probabilidad igual a α .

- Caso Σ desconocida

Estadístico de contraste:

$$T^2 := N(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)' S_n^{-1} (\bar{\mathbf{X}} - \boldsymbol{\mu}_0),$$

verificándose que

$$\frac{T^2}{n} \frac{n-p+1}{p} \sim F_{p,n-p+1}(\delta),$$

con $\delta = N(\boldsymbol{\mu} - \boldsymbol{\mu}_0)' \Sigma^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0)$

Función ‘test’: Con $\phi : \mathbb{R}^{Np} \longrightarrow \{0, 1\}$,

$$\phi(\mathbf{x}) = \begin{cases} 1 & \text{si } t := T^2(\mathbf{x}) > \frac{np}{n-p+1} F_{p,n-p+1,\alpha} \\ 0 & \text{si } t := T^2(\mathbf{x}) \leq \frac{np}{n-p+1} F_{p,n-p+1,\alpha} \end{cases}$$

para todo $\mathbf{x} \in \mathbb{R}^{Np}$, siendo $F_{p,n-p+1,\alpha}$ el valor de una distribución F (centrada) con $(p, n-p+1)$ grados de libertad que deja a su derecha una masa de probabilidad igual a α .

Ejemplo

[Rencher y Christensen (2012), Tabla 3.1]

'Altura' (pulgadas) y 'Peso' (libras) para una muestra de $N = 20$ individuos

Individuo	Altura (X_1)	Peso (X_2)	Individuo	Altura (X_1)	Peso (X_2)
1	69	153	11	72	140
2	74	175	12	79	265
3	68	155	13	74	185
4	70	135	14	67	112
5	72	172	15	66	140
6	67	150	16	71	150
7	66	115	17	74	165
8	70	137	18	75	185
9	76	200	19	75	210
10	68	130	20	76	220

Estadísticos muestrales básicos:

- Vector de medias muestral:

$$\bar{\mathbf{x}} = \begin{pmatrix} 71,45 \\ 164,7 \end{pmatrix}$$

- Matriz de cuasi-covarianzas muestral ($n = N - 1 = 19$)

$$S_n = \begin{pmatrix} 14,576 & 128,88 \\ 128,88 & 1441,2653 \end{pmatrix}$$

- Coeficiente de correlación muestral:

$$r_{12} = 0,889$$

Se quiere contrastar las hipótesis

$$\left| \begin{array}{l} H_0 : \mu = \begin{pmatrix} 70 \\ 170 \end{pmatrix} \\ H_1 : \mu \neq \begin{pmatrix} 70 \\ 170 \end{pmatrix} \end{array} \right.$$

Ejemplo (cont.)

Contraste sobre el vector de medias μ , con Σ conocida:

Supongamos que se conoce que, para la población de referencia, la matriz de covarianzas es

$$\Sigma = \begin{pmatrix} 20 & 100 \\ 100 & 1000 \end{pmatrix}$$

- Valor del estadístico de contraste:

$$u = (20) \begin{pmatrix} 71,45 - 70 \\ 164,7 - 170 \end{pmatrix}' \begin{pmatrix} 20 & 100 \\ 100 & 1000 \end{pmatrix}^{-1} \begin{pmatrix} 71,45 - 70 \\ 164,7 - 170 \end{pmatrix} = 8,4026$$

- Comparación con el valor teórico bajo H_0 a un nivel de significación $\alpha = 0,05$, $\chi^2_{2,0,05} = 5,99$

$$u = 8,4026 > 5,99 = \chi^2_{2,0,05} \quad \rightarrow \quad \boxed{\text{Se rechazaría } H_0}$$

- Comparación con el valor teórico bajo H_0 a un nivel de significación $\alpha = 0,01$, $\chi^2_{2,0,01} = 9,21$

$$u = 8,4026 < 9,21 = \chi^2_{2,0,01} \quad \rightarrow \quad \boxed{\text{No se rechazaría } H_0}$$

Ejemplo (cont.)

Contraste sobre el vector de medias μ , con Σ desconocida:

- Valor del estadístico de contraste:

$$t = (20) \begin{pmatrix} 71,45 - 70 \\ 164,7 - 170 \end{pmatrix}' \begin{pmatrix} 14,576 & 128,88 \\ 128,88 & 1441,2653 \end{pmatrix}^{-1} \begin{pmatrix} 71,45 - 70 \\ 164,7 - 170 \end{pmatrix} = \dots$$

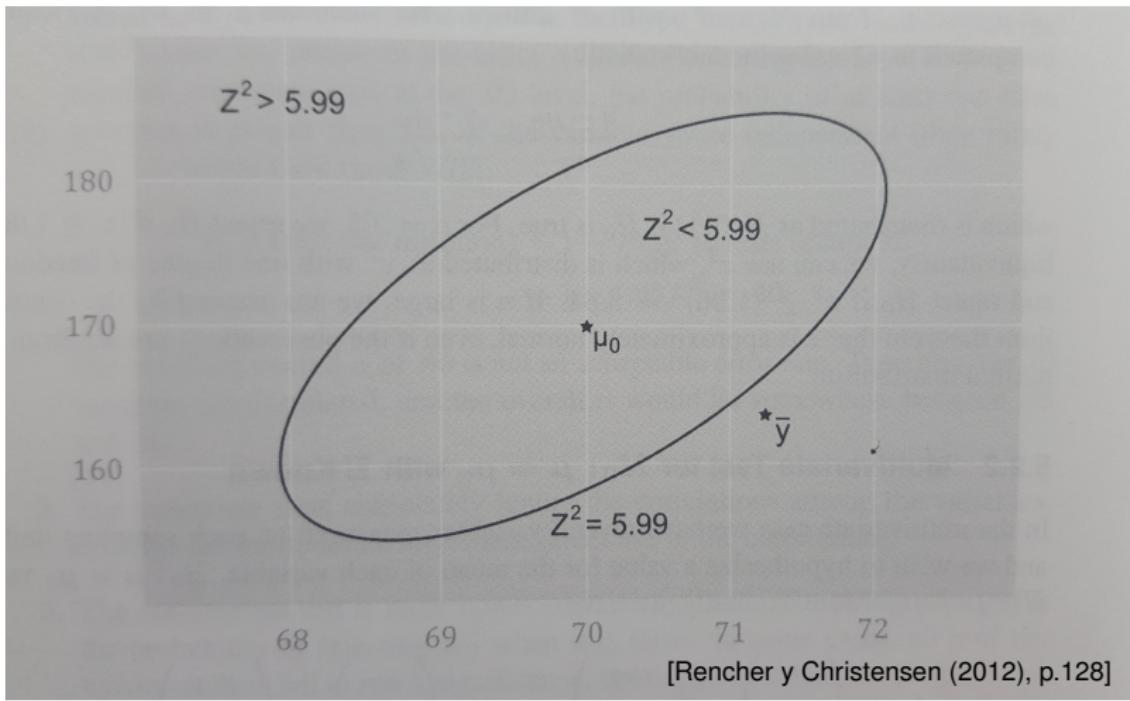
- Valor de comparación teórico bajo H_0 a un nivel de significación α :

$$\frac{(19)(2)}{19 - 2 + 1} F_{2, 19-2+1, \alpha} = \frac{38}{18} F_{2, 18, \alpha} = \dots$$

■ COMPLETAR:

- Decisión sobre H_0 , con Σ desconocida, al nivel $\alpha = 0,1$, $\alpha = 0,05$, $\alpha = 0,01$
- Gráficos de 'regiones de aceptación' en torno a μ_0 , con Σ conocida y desconocida, para distintos valores del 'nivel de significación' α
- Gráficos de 'regiones de confianza' en torno a \bar{x} , con Σ conocida y desconocida, para distintos valores del 'nivel de confianza' $1 - \alpha$

Ejemplo (cont.)



Región elíptica de aceptación en torno a μ_0 , con Σ conocida, para un nivel de significación $\alpha = 0,05$