Pruebas de hipótesis Multivariadas

Cristian Guarnizo-Lemus
cristianguarnizo@itm.edu.co

Maestria en Automatización y Control Industrial

Problema

- Una variable aleatoria modelada por una normal p-variada, tiene p medias, p varianzas, y $\binom{p}{2}$ covarianzas.
- El numero total de parámetros es

$$p + p + \binom{p}{2} = \frac{1}{2}p(p+3)$$

Por qué es necesario

Multivariado

- Preserva el nivel α.

Univariado

- Realizar p tests univariados infla el error tipo I, α.
- Pro ejemplo, para $p = 10$ test univariados separados a un nivel 0.05, la probabilidad que al menos se rechace uno es mayor que 0.05.
- Si las pruebas son independientes, se tiene para H_0:

$$\begin{aligned} P(\text{al menos un rechazo}) &= 1 - P(\text{todas las pruebas aceptan } H_0) \\ &= 1 - (0.95)^{10} = 0.40 \end{aligned}$$

La tasa de error de 0.40 es inaceptable.

Contenido

Test para $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ con $\boldsymbol{\Sigma}$ conocida

En el caso multivariado se tienen varias variables medidas en cada unidad de muestreo, y queremos asumir la hipótesis que las medias, $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ vs $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$

$$H_0 : \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} = \begin{bmatrix} \mu_{01} \\ \mu_{02} \\ \vdots \\ \mu_{0p} \end{bmatrix}, \quad H_1 : \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} \neq \begin{bmatrix} \mu_{01} \\ \mu_{02} \\ \vdots \\ \mu_{0p} \end{bmatrix}$$

donde cada μ_{0j} son especificados a partir de una experiencia previa o es un valor objetivo.

Test para $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ con $\boldsymbol{\Sigma}$ conocida

Para la prueba de H_0, usamos una muestra aleatoria de n observaciones $\mathbf{y}_1, \ldots, \mathbf{y}_n$ de $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, con $\boldsymbol{\Sigma}$ conocido, y calcular $\bar{\mathbf{y}} = \sum_{i=1}^{n} \mathbf{y}_i / n$. El estadístico se define

$$Z^2 = n(\bar{\mathbf{y}} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{y}} - \boldsymbol{\mu}_0)$$

Si H_0 es verdadera, Z^2 esta distribuido como χ^2_p. Y rechazamos H_0 si $Z^2 > \chi^2_{\alpha,p}$. Si $\boldsymbol{\Sigma}$ es desconocido, podemos usar $\mathbf{S}$ en su lugar, y Z^2 debería tener una distribución χ^2 aproximada.

Test para $H_0 : \mu = \mu_0$ con Σ conocida

Ejemplo: Se tienen los siguientes pesos y alturas.

Person	Height	Weight	Person	Height	Weight
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

Tabla 3.1 [Rencher, p. 45]

Test para $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ con $\boldsymbol{\Sigma}$ conocida

Asumamos que la muestra de la normal bivariada $\mathcal{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, donde

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

Suponga que el prueba $H_0 : \ \boldsymbol{\mu} = [70, 170]^\top$. Para la tabla anterior $\bar{y}_1 = 71.45$ y $\bar{y}_2 = 164.7$.

$$\begin{aligned} Z^2 &= n\left(\overline{\mathbf{y}} - \boldsymbol{\mu}_0\right)^\top \boldsymbol{\Sigma}^{-1} \left(\overline{\mathbf{y}} - \boldsymbol{\mu}_0\right) \\ &= (20) \begin{bmatrix} 71.45 - 70 \\ 164.7 - 170 \end{bmatrix}^\top \begin{bmatrix} 20 & 100 \\ 100 & 1000 \end{bmatrix}^{-1} \begin{bmatrix} 71.45 - 70 \\ 164.7 - 170 \end{bmatrix} \\ &= (20)[1.45, -5.3] \begin{bmatrix} .1 & -.01 \\ -.01 & .002 \end{bmatrix} \begin{bmatrix} 1.45 \\ -5.3 \end{bmatrix} = 8.4026 \end{aligned}$$

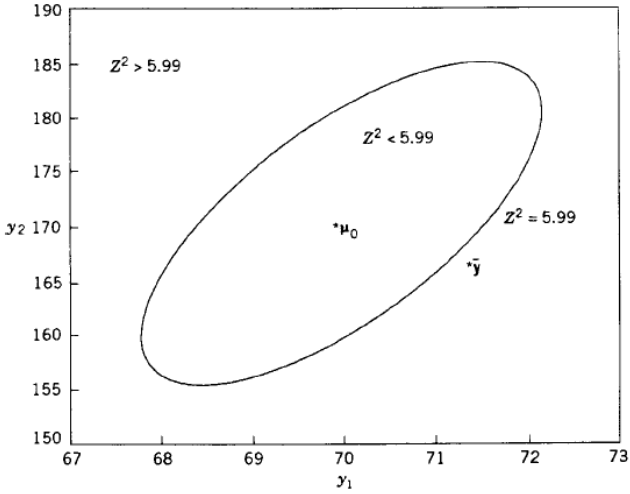
Test para $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ con Σ conocida

Usando $\alpha = 0.05$, $\chi^2_{0.05,2} = 5.99$, y entonces rechazamos $H_0 : \boldsymbol{\mu} = [70, 170]^\top$ porque $Z^2 = 8.4026 > 5.99$. Se puede verificar usando la función chi2 de Scipy:

```
from scipy.stats import chi2

prob = 0.95
dof = 2
critical = chi2.ppf(prob, dof)
Z2 = 8.4026
if abs(Z2) >= critical:
   print('Rechazar H0')
else:
   print('Aceptar H0')
```

Test para $H_0 : \mu = \mu_0$ con Σ conocida



Test para $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ con $\boldsymbol{\Sigma}$ conocida

Para el caso univariado, analizando cada variable de manera separada. Tenemos que $z_{\alpha/2} = 1.96$ para $\alpha = 0.05$, entonces

$$z_1 = \frac{\bar{y}_1 - \mu_{01}}{\sigma_1/\sqrt{n}} = 1.450 < 1.96,$$

$$z_2 = \frac{\bar{y}_2 - \mu_{02}}{\sigma_2/\sqrt{n}} = -.7495 > -1.96,$$

aceptamos ambas pruebas de hipótesis. Ninguno de las medias estimadas esta lo suficientemente lejos del valor hipotético para generar el rechazo.

Test para $H_0 : \mu = \mu_0$ con Σ conocida

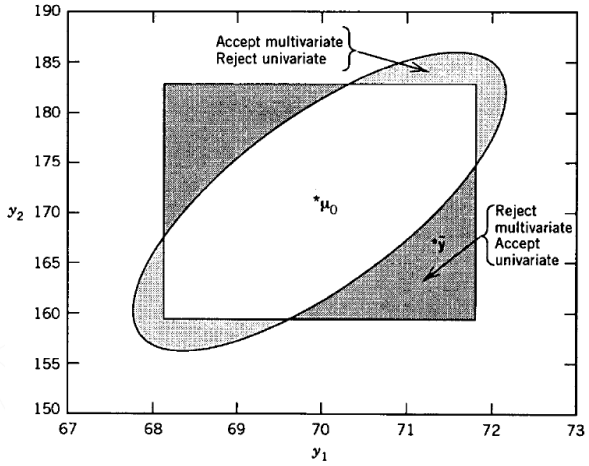


Fig.: Tomado de Rencher (2002), Cap 5.

Contenido

Hotelling's test para la media con Σ desconocida

Asumimos una muestra aleatoria de p variables con n observaciones $\mathbf{y}_1, \ldots, \mathbf{y}_n$ de $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, donde $\mathbf{y}_i$ contiene las p medidas de la i-ésima muestra. Estimamos μ con $\bar{\mathbf{y}}$ y $\boldsymbol{\Sigma}$ con $\mathbf{S}$. El test se realiza con el estadístico

$$T^2 = n(\bar{\mathbf{y}} - \boldsymbol{\mu}_0)^\top \mathbf{S}^{-1} (\bar{\mathbf{y}} - \boldsymbol{\mu}_0).$$

Rechazamos H_0 si $T^2 > T^2_{\alpha,p,n-1}$

Hotelling's test para la media con Σ desconocida

Ejemplo Tabla 3.3 [Rencher] y_1 (calcio disponible), y_2 (calcio intercambiable), y_3 (calcio verde).

Number	y_1	y_2	y_3
1	35	35	280
2	35	49	270
3	40	300	321
4	10	28	273
5	6	27	281
6	20	28	288
7	35	46	290
8	35	109	328
9	35	160	320
10	30	16	13

Hotelling's test para la media con Σ desconocida

En la tabla anterior tenemos $n = 10$ observaciones de $p = 3$ variables. Los valores deseables de y_1 y y_2 son 15.0 y 6.0, y el nivel esperado de y_3 es 2.85. Podemos hacer la prueba de hipótesis

$$H_0 : \boldsymbol{\mu} = \begin{bmatrix} 15.0 \\ 6.0 \\ 2.85 \end{bmatrix}$$

donde

$$\overline{\mathbf{y}} = \begin{bmatrix} 28.1 \\ 7.18 \\ 3.09 \end{bmatrix}, \quad \mathbf{S} = \begin{bmatrix} 140.54 & 49.68 & 1.94 \\ 49.68 & 72.25 & 3.68 \\ 1.94 & 3.68 & .25 \end{bmatrix}$$

Hotelling's test para la media con Σ desconocida

Para probar H_0, empleamos

$$
\begin{aligned}
T^2 &= n(\overline{\mathbf{y}} - \boldsymbol{\mu}_0)^\top \mathbf{S}^{-1} (\overline{\mathbf{y}} - \boldsymbol{\mu}_0) \\
&= 10 \begin{bmatrix} 28.1 - 15.0 \\ 7.18 - 6.0 \\ 3.09 - 2.85 \end{bmatrix}^\top \begin{bmatrix} 140.54 & 49.68 & 1.94 \\ 49.68 & 72.25 & 3.68 \\ 1.94 & 3.68 & .25 \end{bmatrix}^{-1} \begin{bmatrix} 28.1 - 15.0 \\ 7.18 - 6.0 \\ 3.09 - 2.85 \end{bmatrix} \\
&= 24.559
\end{aligned}
$$

Test para $H_0 : \mu = \mu_0$ con Σ desconocida

Usando $\alpha = 0.05$, $T^2_{0.05,3,9} = 16.766$, y entonces rechazamos H_0 ya que $T^2 = 24.559 > 16.766$. Se puede verificar usando la función f de Scipy:

```
from scipy.stats import f

prob = 0.95
p = 3
n = 10
v = n-1
critical = f.ppf(prob, p,v-p+1)*(v*p)/(v-p+1) #ver eq. 5.7 Rencher
T2 = 24.559
if abs(T2) >= critical:
   print('Rechazar H0')
else:
   print('Aceptar H0')
```

Contenido

Prueba T^2 de dos muestras multivariadas

Consideramos el caso donde p variables son medidas en cada muestra unitaria en 2 muestras. Deseamos probar

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

Obtenemos un muestra aleatoria $\mathbf{y}_{11}, \mathbf{y}_{12}, \ldots, \mathbf{y}_{1n_1}$ de $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ y una segunda muestra aleatoria $\mathbf{y}_{21}, \mathbf{y}_{22}, \ldots, \mathbf{y}_{2n_2}$ de $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$. Se asumen que las dos muestras son independientes y que $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, con $\boldsymbol{\Sigma}$ desconocido. El estadístico se define

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} (\overline{\mathbf{y}}_1 - \overline{\mathbf{y}}_2)^\top \mathbf{S}_{\text{pl}}^{-1} (\overline{\mathbf{y}}_1 - \overline{\mathbf{y}}_2) \tag{1}$$

La cual esta distribuida como T^2_{p, n_1+n_2-2} cuando $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ es verdadero.

Prueba T^2 de dos muestras multivariadas

Donde S_{pl}

$$\mathbf{W}_1 = \sum_{i=1}^{n_1} (\mathbf{y}_{1i} - \overline{\mathbf{y}}_1)(\mathbf{y}_{1i} - \overline{\mathbf{y}}_1)^\top = (n_1 - 1)\mathbf{S}_1$$

$$\mathbf{W}_2 = \sum_{i=1}^{n_2} (\mathbf{y}_{2i} - \overline{\mathbf{y}}_2)(\mathbf{y}_{2i} - \overline{\mathbf{y}}_2)^\top = (n_2 - 1)\mathbf{S}_2$$

y

$$\mathbf{S}_{\text{pl}} = \frac{1}{n_1 + n_2 - 2}(\mathbf{W}_1 + \mathbf{W}_2)$$

Prueba T^2 de dos muestras multivariadas

Ejemplo Tabla 5.1 [Rencher, p. 125] Cuatro test psicológicos fueron dados a 32 hombres y 32 mujeres. Se toman 4 variables: inconsistencias pictóricas, tablero de papel, herramienta de reconocimiento, vocabulario.

$$\bar{\mathbf{y}}_1 = \begin{bmatrix} 15.97 \\ 15.91 \\ 27.19 \\ 22.75 \end{bmatrix}, \quad \bar{\mathbf{y}}_2 = \begin{bmatrix} 12.34 \\ 13.91 \\ 16.66 \\ 21.94 \end{bmatrix}$$

$$\mathbf{S}_1 = \begin{bmatrix} 5.192 & 4.545 & 6.522 & 5.250 \\ 4.545 & 13.18 & 6.760 & 6.266 \\ 6.522 & 6.760 & 28.67 & 14.47 \\ 5.250 & 6.266 & 14.47 & 16.65 \end{bmatrix}, \mathbf{S}_2 = \begin{bmatrix} 9.136 & 7.549 & 4.864 & 4.151 \\ 7.549 & 18.60 & 10.22 & 5.446 \\ 4.864 & 10.22 & 30.04 & 13.49 \\ 4.151 & 5.446 & 13.49 & 28.00 \end{bmatrix}$$

Prueba T^2 de dos muestras multivariadas

Con los datos anteriores calculamos

$$\mathbf{S}_{\text{pl}} = \frac{1}{32+32-2}\left[(32-1)\mathbf{S}_1 + (32-1)\mathbf{S}_2\right] = \begin{bmatrix} 7.164 & 6.047 & 5.693 & 4.701 \\ 6.047 & 15.89 & 8.492 & 5.856 \\ 5.693 & 8.492 & 29.36 & 13.98 \\ 4.701 & 5.856 & 13.98 & 22.32 \end{bmatrix}$$

Obtenemos

$$T^2 = \frac{n_1 n_2}{n_1+n_2}\left(\overline{\mathbf{y}}_1 - \overline{\mathbf{y}}_2\right)^\top \mathbf{S}_{\text{pl}}^{-1}\left(\overline{\mathbf{y}}_1 - \overline{\mathbf{y}}_2\right) = 97.6015$$

Prueba T^2 de dos muestras multivariadas

Usando $\alpha = 0.01$, $T^2_{0.01,4,62} = 15.373$, y entonces rechazamos H_0 porque $T^2 = 97.6015 > 15.373$. Se puede verificar usando la función f de Scipy:

```
from scipy.stats import f

prob = 0.99
p = 4
n1 = 32
n2 = 32
critical = f.ppf(prob, p,n1+n2-p-1)*p*(n1+n2-2)/(n1+n2-p-1) #ver eq 5.11 Rencher
T2 = 97.6015
if abs(T2) >= critical:
   print('Rechazar H0')
else:
   print('Aceptar H0')
```

Tests emparejados Multivariados

Asumimos que queremos relizar un test de emparejamiento entre $\mathbf{y}_i$ de la primera muestra y $\mathbf{x}_i$ de la segunda muestra, con $i = 1, \ldots, n$.

Numero de pareja	Tratamiento 1	Tratamiento 2	Diferencia $\mathbf{d}_i = \mathbf{y}_i - \mathbf{x}_i$
1	$\mathbf{y}_1$	$\mathbf{x}_1$	$\mathbf{d}_1$
2	$\mathbf{y}_2$	$\mathbf{x}_2$	$\mathbf{d}_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	$\mathbf{y}_n$	$\mathbf{x}_n$	$\mathbf{d}_n$

Tests emparejados Multivariados

Asumimos que **x** y **y** estan correlacionados y tienen una distribución normal multivariada.

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{x} \end{pmatrix} \quad \text{es} \quad N_{2p}\left[\begin{pmatrix} \boldsymbol{\mu}_y \\ \boldsymbol{\mu}_x \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yx} \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{pmatrix}\right]$$

Para probar que $H_0: \boldsymbol{\mu}_d = \mathbf{0}$, lo que equivale a $H_0: \boldsymbol{\mu}_y = \boldsymbol{\mu}_x$, calculamos

$$\overline{\mathbf{d}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{d}_i \quad \text{y} \quad \mathbf{S}_d = \frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{d}_i - \overline{\mathbf{d}}\right)\left(\mathbf{d}_i - \overline{\mathbf{d}}\right)^{\top}$$

Tests emparejados Multivariados

Finalmente tenemos,

$$T^2 = \overline{\mathbf{d}}^\top \left(\frac{\mathbf{S}_d}{n}\right)^{-1} \overline{\mathbf{d}} = n\overline{\mathbf{d}}^\top \mathbf{S}_d^{-1} \overline{\mathbf{d}}$$

Bajo H_0, este estadístico T^2 de comparación emparejada esta distribuido como $T^2_{p,n-1}$.
Rechazamos H_0 sí $T^2 > T^2_{\alpha,p,n-1}$.

Tests emparejados Multivariados

Finalmente tenemos,

$$T^2 = \overline{\mathbf{d}}^\top \left(\frac{\mathbf{S}_d}{n}\right)^{-1} \overline{\mathbf{d}} = n\overline{\mathbf{d}}^\top \mathbf{S}_d^{-1} \overline{\mathbf{d}}$$

Bajo H_0, este estadístico T^2 de comparación emparejada esta distribuido como $T^2_{p,n-1}$.
Rechazamos H_0 sí $T^2 > T^2_{\alpha,p,n-1}$.

Lecturas recomendadas

- Capitulo 7 de Härdle and Simar. Hypothesis Testing.
- Capitulo 5.9 de Rencher. Análisis de Perfiles.

Referencias

- Alvin C. Rencher. “Methods of Multivariate Analysis”, 2002.
- W. K. Härdle and L. Simar. “Applied Multivariate Statistical Analysis”, 2019.