Distribución normal multivariada

Propiedades e inferencias

Mario J. P. López

Departamento de Matemáticas
Universidad El Bosque

2020

f.d.p. normal p-variada

Un vector aleatorio $\boldsymbol{X}$ $(p \times 1)$ tiene distribución normal multivariada, con vector de medias $\boldsymbol{\mu} \in \mathbb{R}^p$ y matriz de covarianzas $\Sigma > 0$, si su f.d.p. está dada por

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right\}, \ \boldsymbol{x} \in \mathbb{R},$$

y escribimos $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$. El término

$$\Delta = \left[(\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right]^{1/2}$$

se conoce como distancia de Mahalanobis y $|\Sigma|$ es la varianza generalizada de $\boldsymbol{X}$.

Función generadora de momentos

f.g.m.

Un vector aleatorio $\boldsymbol{X} \sim N_p\left(\boldsymbol{\mu}, \Sigma\right)$ tiene función generadora de momentos, f.g.m.

$$M_{\boldsymbol{X}}\left(\boldsymbol{t}\right) := E\left[\exp\left(\boldsymbol{t}^\top \boldsymbol{X}\right)\right] = \exp\left(\boldsymbol{\mu}^\top \boldsymbol{t} + \frac{1}{2}\boldsymbol{t}^\top \Sigma \boldsymbol{t}\right),$$

con $\boldsymbol{t} \in \mathbb{R}^p$.

1. Combinaciónes lineales de v.a. normales multivariadas

- Sea $\boldsymbol{a}$ un vector $(p \times 1)$ de constantes y $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$, entonces $\boldsymbol{a}^\top \boldsymbol{X} \sim N\left(\boldsymbol{a}^\top \boldsymbol{\mu}, \boldsymbol{a}^\top \Sigma \boldsymbol{a}\right)$.
- Si A es una matriz $(q \times p)$ de constantes, de rango $q \leq p$, y $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$, entonces $A\boldsymbol{X} \sim N_q\left(A\boldsymbol{\mu}, A\Sigma A^\top\right)$.

2. Estandarización multivariada

Dado $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$, un vector estandarizado $\boldsymbol{Z}$, con distribución $N_p\left(\mathbf{0}, I_p\right)$, se puede obtener como

$$\boldsymbol{Z} = \left(\Gamma^\top\right)^{-1} (\boldsymbol{X} - \boldsymbol{\mu}),$$

donde $\Sigma = \Gamma^\top \Gamma$.

3. Normalidad de distribuciones marginales

Sean $\boldsymbol{X}_1 = (X_1, \ldots, X_r)^\top$ y $\boldsymbol{X}_2 = \left(X_{r+1}, \ldots, X_p\right)^\top$, con $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$ y

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

entonces $\boldsymbol{X}_1 \sim N_r(\boldsymbol{\mu}_1, \Sigma_{11})$.

- En particular, $X_j \sim N\left(\mu_j, \sigma_j^2\right)$, $j = 1, \ldots, p$
- lo contrario no es necesariamente cierto

4. Independencia

Sean $\boldsymbol{X}$ $(p \times 1)$ y $\boldsymbol{Y}$ $(q \times 1)$, tal que

$$\begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{pmatrix} \sim N_{p+q}\left(\begin{pmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{pmatrix}, \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}\right),$$

entonces los sub-vectores $\boldsymbol{X}$ y $\boldsymbol{Y}$ son independientes si $\Sigma_{xy} = \mathbf{0}$.

Propiedades

5. Distribución condicional

En la propiedad anterior, si $\boldsymbol{X}$ y $\boldsymbol{Y}$ no son independientes, entonces $\Sigma_{xy} \neq \mathbf{0}$ y la distribución condicional de $\boldsymbol{X}$ dado $\boldsymbol{Y}$ es normal multivariada con

$$E\left[\boldsymbol{Y} \mid \boldsymbol{X}\right] = \boldsymbol{\mu}_y + \Sigma_{yx}\Sigma_{xx}^{-1}\left(\boldsymbol{X} - \boldsymbol{\mu}_x\right)$$
$$C\left[\boldsymbol{Y} \mid \boldsymbol{X}\right] = \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}$$

6. Distribución χ^2

Si $\boldsymbol{X} \sim N_p\left(\boldsymbol{\mu}, \Sigma\right)$ y $\boldsymbol{Z} = \left(\Gamma^\top\right)^{-1}\left(\boldsymbol{X} - \boldsymbol{\mu}\right)$ tal que $\boldsymbol{Z} \sim N_p\left(\mathbf{0}, I_p\right)$ entonces

$$\boldsymbol{Z}^\top\boldsymbol{Z} = \left(\boldsymbol{x} - \boldsymbol{\mu}\right)^\top \Sigma^{-1}\left(\boldsymbol{x} - \boldsymbol{\mu}\right) \sim \chi^2_p$$

7. Distribución de la suma de dos sub-vectores

Sean $\boldsymbol{X}$ $(p \times 1)$ y $\boldsymbol{Y}$ $(p \times 1)$, independientes, tal que

$$\begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{pmatrix} \sim N_{2p}\left(\begin{pmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{pmatrix}, \begin{pmatrix} \Sigma_{xx} & \mathbf{0} \\ \mathbf{0} & \Sigma_{yy} \end{pmatrix}\right),$$

entonces

$$\boldsymbol{X} \pm \boldsymbol{Y} \sim N_p\left(\boldsymbol{\mu}_x \pm \boldsymbol{\mu}_y, \Sigma_{xx} + \Sigma_{yy}\right)$$

Función de (log)verosimilitud

Sean $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n$ una muestra aleatoria de una v.a. $\boldsymbol{X} \sim N_p\left(\boldsymbol{\mu},\Sigma\right)$, la función de máxima verosimilitud de la muestra está dada por

$$
\begin{aligned}
L\left(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n;\boldsymbol{\mu},\Sigma\right) &= \prod_{i=1}^{n} f_{\boldsymbol{X}_i}\left(\boldsymbol{x}_i\right) \\
&= \prod_{i=1}^{n} \frac{1}{(2\pi)^{p/2}\left|\Sigma\right|^{1/2}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{x}_i-\boldsymbol{\mu}\right)^{\top}\Sigma^{-1}\left(\boldsymbol{x}_i-\boldsymbol{\mu}\right)\right\} \\
&= \frac{1}{(2\pi)^{np/2}\left|\Sigma\right|^{n/2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left(\boldsymbol{x}_i-\boldsymbol{\mu}\right)^{\top}\Sigma^{-1}\left(\boldsymbol{x}_i-\boldsymbol{\mu}\right)\right\}
\end{aligned}
$$

y la función de log-verosimilitud

$$
\ell\left(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n;\boldsymbol{\mu},\Sigma\right) = -\frac{np}{2}\log\left(2\pi\right) - \frac{n}{2}\log\left|\Sigma\right| - \frac{1}{2}\sum_{i=1}^{n}\left(\boldsymbol{x}_i-\boldsymbol{\mu}\right)^{\top}\Sigma^{-1}\left(\boldsymbol{x}_i-\boldsymbol{\mu}\right)
$$

Estimación

MLE

Los estimadores de máxima verosimilitud de $\boldsymbol{\mu}$ y Σ son

$$\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i$$

$$\hat{\Sigma} = S_n = \frac{1}{n}\sum_{i=1}^{n} \left(\boldsymbol{x}_i - \bar{\boldsymbol{x}}\right)\left(\boldsymbol{x}_i - \bar{\boldsymbol{x}}\right)^{\top}$$

respectivamente.

Distribución de $\overline{\boldsymbol{X}}$

- Cuando $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ es una muestra aleatoria de una v.a. $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$, entonces $\bar{\boldsymbol{X}} \sim N(\boldsymbol{\mu}, \Sigma/n)$
- Cuando $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ es una muestra aleatoria de una distribución no normal con vector de medias $\boldsymbol{\mu}$ y matriz de covarianzas Σ, entonces conforme $n \longrightarrow \infty$, $\bar{\boldsymbol{X}}$ tiene una distribución aproximada $N_p(\boldsymbol{\mu}, \Sigma)$. Este resultado se conoce como el teorema del límite central multivariado. En particular

$$\sqrt{n}\left(\bar{\boldsymbol{X}} - \boldsymbol{\mu}\right) \sim N(\mathbf{0}, \Sigma)$$

conforme $n \longrightarrow \infty$.

Distribución Wishart

Una matriz aleatoria $\boldsymbol{W}$ $(p \times p)$, definida positiva, tiene una distribución Wishart $\boldsymbol{W} \sim W_p(n, V)$ con parámetro $V > 0$, $(p \times p)$ y grados de libertad $n > p - 1$ si su f.d.p. está dada por

$$f_W(\boldsymbol{w}) = \frac{|\boldsymbol{w}|^{(n-p-1)/2} e^{-\mathrm{tr}\left(V^{-1}\boldsymbol{w}\right)/2}}{2^{np/2} |V|^{n/2} \Gamma_p\left(\frac{n}{2}\right)},$$

donde $\Gamma(\cdot)$ es la función gama multivariada.

Distribución de S_{n-1}

- La v.a.

$$W = \sum_{i=1}^{n} (\boldsymbol{X}_i - \boldsymbol{\mu})(\boldsymbol{X}_i - \boldsymbol{\mu})^\top$$

  tiene distribución $W_p(n, \Sigma)$

- La v.a.

$$W = \sum_{i=1}^{n} \left(\boldsymbol{X}_i - \bar{\boldsymbol{X}}\right)\left(\boldsymbol{X}_i - \bar{\boldsymbol{X}}\right)^\top = (n-1)\, S_{n-1}$$

  tiene distribución $W_p(n-1, \Sigma)$.

- Se puede demostrar además que $\bar{\boldsymbol{X}}$ y S_n son estadísticamente independientes.

q-q plot univariado

- Sean $X_{(1)}, X_{(2)}, ..., X_{(n)}$ las estadísticas de orden de una m.a.
- Encontrar los valores q_i tal que $\Phi(q_i) = P(Z \leq q_i) = \frac{i-0.5}{n}$
- Graficar la nube de puntos conformada por las parejas $(q_i, X_{(i)})$
- Agregar la línea $X_{(i)} = a + bq_i$ que "mejor" se ajusta a la nube de puntos.

Verificando normalidad

q-q plot multivariado

- Sea $\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_n$ una m.a. de v.a. p-variadas.
- Si $\boldsymbol{X}_i \sim N_p(\boldsymbol{\mu}, \Sigma)$, $i = 1, ..., n$, entonces

$$u_i = \frac{nD_i^2}{(n-1)^2} \sim Beta\left(\frac{p}{2}, \frac{n-p-1}{2}\right)$$

con

$$D_i = \left(\boldsymbol{X}_i - \bar{\boldsymbol{X}}\right)^\top S_{n-1}^{-1}\left(\boldsymbol{X}_i - \bar{\boldsymbol{X}}\right)$$

Verificando normalidad

q-q plot multivariado

- Calcular los cuantiles de la distribución beta, v_i, $i = 1, \dots, n$, como solución a

$$\int_0^{v_i} \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{n-p-1}{2}\right)} y^{\frac{p}{2}-1}(1-y)^{\frac{n-p-1}{2}-1}\,dy = \frac{i-a}{n-a-b+1}$$

  con $a = \frac{(p-2)}{2p}$ y $b = \frac{n-p-3}{2(n-p-1)}$.
- Graficar la nube de puntos $\left(v_i, u_{(i)}\right)$
- Agregar la línea $u_{(i)} = a + bv_i$.

Test de Mardia

Se basa en la asimetría y la curtosis de poblaciones normales multivariadas

$$\beta_{1,p} = E\left[(\boldsymbol{X}-\boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{X}-\boldsymbol{\mu})\right]^3 = 0$$
$$\beta_{2,p} = E\left[(\boldsymbol{X}-\boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{X}-\boldsymbol{\mu})\right]^2 = p\,(p+2)$$

los cuales se pueden estimar como

$$\hat{\beta}_{1,p} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} g_{ij}^3, \qquad \hat{\beta}_{2,p} = \frac{1}{n}\sum_{i=1}^{n} g_{ii}^2$$

con $g_{ij} = \left(\boldsymbol{X}_i - \bar{\boldsymbol{X}}\right)^\top S_n^{-1}\left(\boldsymbol{X}_j - \bar{\boldsymbol{X}}\right)$.

Verificando normalidad

Test de Mardia

Luego, se calculan las estadísticas

$$Z_1 = \frac{(p+1)(n+1)(n+3)}{6\left[(p+1)(n+1)-6\right]}\hat{\beta}_{1,p} \sim \chi^2_{p(p+1)(p+2)/6}$$

$$Z_2 = \frac{\hat{\beta}_{2,p} - p(p+2)}{\sqrt{8p(p+2)/n}} \sim N(0,1)$$

Así, se rechaza la hipótesis de normalidad multivariada

- en términos de asimetría si $Z_1 \geq \chi^2_{0.05,p(p+1)(p+2)/6}$
- en términos de curtosis si $|Z_2| \geq Z_{0.05}$

Datos atípicos

Detectando *outliers*

Un simple procedimiento para detectar *outliers* multivariados consiste en calcular la distancia de Mahalanobis muestral para cada individuo i-ésimo,

$$MD_i = \left[\left(\boldsymbol{X}_i - \bar{\boldsymbol{X}} \right)^\top S_n^{-1} \left(\boldsymbol{X}_i - \bar{\boldsymbol{X}} \right) \right]^{1/2},$$

y compararla con $\chi^2_{p,0.975}$.

Para $\boldsymbol{\mu}$ con Σ conocida

- Se desea probar $H_0 : \mu = \mu_0$ contra $H_1 : \mu \neq \mu_0$, dada una m.a. de una $N(\mu, \Sigma)$.
- Estadístico de prueba:

$$Z^2 = n\left(\bar{\boldsymbol{X}} - \mu_0\right)^\top \Sigma^{-1}\left(\bar{\boldsymbol{X}} - \mu_0\right)$$

  que bajo H_0 sigue una distribución χ^2_p.
- Rechazamos H_0 si $Z^2 > \chi^2_{\alpha,p}$

Para $\boldsymbol{\mu}$ con Σ desconocida

- Se desea probar $H_0 : \mu = \mu_0$ contra $H_1 : \mu \neq \mu_0$, dada una m.a. de una $N(\mu, \Sigma)$.
- Estadístico de prueba:

$$T^2 = n\left(\bar{\boldsymbol{X}} - \mu_0\right)^\top S_{n-1}^{-1}\left(\bar{\boldsymbol{X}} - \mu_0\right)$$

  que bajo H_0 sigue una distribución $T^2_{p,n-1}$.
- Rechazamos H_0 si $T^2 > T^2_{\alpha,p,n-1}$.
- Tener en cuenta que si $w \sim T^2_{p,m}$ entonces

$$\frac{m-p+1}{pm} w \sim F_{p,m-p+1}$$

Para $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ *con* $\Sigma_1 = \Sigma_2$ *desconocidas*

- Se desea probar $H_0 : \mu_1 = \mu_2$ contra $H_1 : \mu_1 \neq \mu_2$, dadas un par de m.a. de una $N(\mu_1, \Sigma)$ y una $N(\mu_2, \Sigma)$.
- Estadístico de prueba:

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} \left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right)^\top S_c^{-1} \left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right)$$

  que bajo H_0 sigue una distribución T^2_{p,n_1+n_2-2}.
- Rechazamos H_0 si $T^2 > T^2_{\alpha,p,n_1+n_2-2}$.
- Donde

$$S_c = \frac{(n_1 - 1) S_1 + (n_2 - 1) S_2}{n_1 + n_2 - 2}$$

Para $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ *con* $\Sigma_1 \neq \Sigma_2$ *desconocidas*

- Se desea probar $H_0 : \mu_1 = \mu_2$ contra $H_1 : \mu_1 \neq \mu_2$, dadas un par de m.a. de una $N(\mu_1, \Sigma_1)$ y una $N(\mu_2, \Sigma_2)$.
- Estadístico de prueba:

$$T^2 = \left(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2\right)^\top \left[\frac{S_1}{n_1} + \frac{S_2}{n_2}\right]^{-1} \left(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2\right)$$

que bajo H_0 sigue una distribución χ^2_p.
- Rechazamos H_0 si $T^2 > T^2_{\alpha,p}$.

Homogeneidad de matrices de covarianza

- Se desea probar $H_0 : \Sigma_1 = \Sigma_2 = \cdots = \Sigma_k$ contra $H_1 : \Sigma_i \neq \Sigma_j$, $i \neq j$, dadas k m.a. de $N(\mu_i, \Sigma_i)$.
- Prueba de *Bartlett*: Se basa en la estadística

$$MC^{-1} \sim \chi^2_{(k-1)(p+1)p/2}$$

donde

$$C^{-1} = 1 - \frac{2p^2 + 3p - 1}{6(p+1)(k-1)} \left(\sum_{i=1}^{k} (n_i - 1)^{-1} - \left(\sum_{i=1}^{k} (n_i - 1) \right)^{-1} \right)$$

$$M = \sum_{i=1}^{k} (n_i - 1) \ln |S_c| - \sum_{i=1}^{k} (n_i - 1) \ln |S_i|$$

Referencias

Rencher, A.
Methods of Multivariate Analysis.
2nd ed. Willey, 2002.