Datos Categóricos: Clase 7

Juan Carlos Correa Morales

Universidad Nacional de Colombia
Sede Medellín
Facultad de Ciencias
Escuela de Estadística

Distribución Normal Multivariable

Un modelo matemático para representar la relación entre un conjunto de k variables aleatorias es la normal multivariable.

Si $\mathbf{x}$ es un vector aleatorio con función de densidad de probabilidad

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{k/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

decimos que $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, donde $\boldsymbol{\mu}$ es el vector de medias de la distribución y $\boldsymbol{\Sigma}$ es la matriz de varianzas y covarianzas de la población.

Distribución Normal Multivariable Un vector aleatorio $\mathbf{x}$ tiene una distribución normal nultivariable si $\mathbf{l}'\mathbf{x}$ se distribuye normalmente para todo vector conformable $\mathbf{l}$.

Resultado Si $\mathbf{x}$ se distribuye $N(\mu, \mathbf{\Sigma})$ y si

$$\mathbf{z} = \mathbf{\Sigma}^{-1/2}(\mathbf{x} - \mu)$$

donde $\mathbf{\Sigma}^{-1/2}$ es la raíz cuadrada simétrica y definida positiva de $\mathbf{\Sigma}^{-1}$, entonces $z_1, z_2, \cdots, z_d$ son variables aleatorias independientes.

Resultado Si $\mathbf{x}$ se distribuye $N(\mu, \mathbf{\Sigma})$, entonces

$$\begin{aligned} \mathcal{E}(\mathbf{x}) &= \mu \\ \mathcal{D}(\mathbf{x}) &= \mathbf{\Sigma} \end{aligned}$$

Resultado Si $\mathbf{x}$ se distribuye $N(\mu, \mathbf{\Sigma})$, entonces

$$\mathbf{y} = (\mathbf{x} - \mu)' \mathbf{\Sigma}^{-1} (\mathbf{x} - \mu) \sim \chi^2_d$$

Resultado Si $\mathbf{x}$ se distribuye $N(\mu, \mathbf{\Sigma})$ y si $\mathbf{w} = \mathbf{A}\mathbf{y} + \mathbf{c}$ donde $\mathbf{A}_{q \times d}$ y $\mathbf{c}$ un vector de q constantes, entonces $\mathbf{w}$ se distribuye $N(\mathbf{A}\mu + \mathbf{c}, \mathbf{A}\mathbf{\Sigma}\mathbf{A}')$

Resultado Cualquier subconjunto de un vector multinormal tiene una distribución multinormal.

En particular los elementos individuales tienen distribuciones normales univariables.

1. Muchas variables reales se pueden aproximar mediante este modelo.

2. La mayoría de técnicas estadísticas tradicionales asumen este modelo.

3. El Teorema Central del Límite Multivariable nos garantiza que si n es grande entonces $\bar{\mathbf{x}}$ tiene una distribución muestral con vector de medias μ y matriz de varianzas y covarianzas $\frac{1}{n}\mathbf{\Sigma}$.

4. El exponente de la distribución tiene una distribución muestral conocida

$$(\mathbf{x}-\mu)'\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu) \sim \chi^2_k$$

- Resultados teóricos útiles
  - Si $\mathbf{\Sigma}$ es una matriz diagonal, entonces las variables son independientes y viceversa.
  - Si $\mathbf{C}$ es una matriz de constantes de dimensión $p \times k$, entonces $\mathbf{Cx} \sim N(\mathbf{C}\mu, \mathbf{C\Sigma C}')$
  - Si $\mathbf{a}$ es un vector de constantes de dimensión $1 \times k$, entonces $\mathbf{a}'\mathbf{x} \sim N(\mathbf{a}'\mu, \mathbf{a}'\mathbf{\Sigma a})$

Herramientas Asintóticas

En esta sección presentamos unos elementos muy básicos de asintótica que permiten mostrar muchos de los resultados posteriores de una manera directa.

Convergencia en Distribución *Una sucesión de variables aleatorias* $\{X_n\}$ *converge en distribución a una* $N\left(\mu, \sigma^2\right)$*, con* $\sigma > 0$*, si equivalentemente la sucesión* $\left\{\frac{X_n-\mu}{\sigma}\right\}$ *converge a una* $N(0,1)$.

Normalidad Asintótica *Una sucesión de variables aleatorias* $\{X_n\}$ *es asintóticamente normal con "media"* μ_n *y varianza* σ_n^2 *si*

- $\sigma_n > 0$ *para todo* n *suficientemente grande y*
-
$$\frac{X_n - \mu_n}{\sigma_n} \xrightarrow{D} N(0,1)$$

$\{\mu_n\}$ *y* $\left\{\sigma_n^2\right\}$ *son sucesiones de constantes.*

No es necesario que μ_n y σ_n^2 sean la media y la varianza de X_n, ni aún que X_n posea tales momentos.

Resultado *Si* X_n *es* $AN\left(\mu_n, \sigma_n^2\right)$*, entonces también* X_n *es* $AN\left(\tilde{\mu}_n, \tilde{\sigma}_n^2\right)$*, si y sólo si*

$$\begin{aligned} \frac{\tilde{\sigma}_n^2}{\sigma_n^2} &\rightarrow 1 \text{ y} \\ \frac{\tilde{\mu}_n - \mu_n}{\sigma_n} &\rightarrow 0 \end{aligned}$$

Resultado *Si* X_n *es* $AN\left(\mu_n, \sigma_n^2\right)$*, entonces también* $a_n X_n + b_n$ *es* $AN\left(\mu_n, \sigma_n^2\right)$*, si y sólo si*

$$\begin{aligned} a_n &\rightarrow 1 \text{ y} \\ \frac{\mu_n\left(a_n - 1\right) - b_n}{\sigma_n} &\rightarrow 0 \end{aligned}$$

Ejemplo Si X_n es $AN\left(n, 2n\right)$ entonces $\frac{n-1}{n}X_n$ es AN pero $\frac{\sqrt{n}-1}{\sqrt{n}}X_n$ no lo es.

Resultado Importante I *Suponga que* X_n *es* $AN\left(\mu, \sigma_n^2\right)$ *con* $\sigma_n \to 0$*. Sea* g *una función de valor real diferenciable en* $X = \mu$ *con* $g'(\mu) \neq 0$*. Entonces*

$$g\left(X_n\right) \sim AN\left(g(\mu), \left[g'(\mu)\right]^2 \sigma_n^2\right)$$

Resultado Importante II *Sea* $\mathbf{X}_n = (X_{n1}, X_{n2}, \cdots, X_{nk})'$ *y además asuma que*

$$\mathbf{X}_n \sim AN\left(\mu, b_n^2 \mathbf{\Sigma}\right)$$

con $\mathbf{\Sigma}$ *matriz de covarianzas y* $b_n \to 0$*.*

Sea $g(\mathbf{x}) = (g_1(\mathbf{x}), g_2(\mathbf{x}), \cdots, g_m(\mathbf{x}),)'$*, donde* $\mathbf{x} = (x_1, x_2, \cdots, x_k)'$*, una función con argumento un vector y donde cada componente es una función de valor real y tiene un diferencial no cero* $g_i(\mu; \mathbf{t})$*,* $\mathbf{x} = (t_1, t_2, \cdots, t_k)'$*, en* $\mathbf{x} = \mu$*. Haga*

$$\mathbf{D} = \left[\frac{\partial g_i}{\partial x_j} \mid_{\mathbf{x}=\mu}\right]_{m \times k}$$

Entonces $g\left(\mathbf{X}_n\right) \sim AN\left(g(\mu), b_n^2 \mathbf{D}\mathbf{\Sigma}\mathbf{D}'\right)$

Teorema Central del Límite Multivariable *1. Sean* $\{\mathbf{X}_i\}$ *vectores aleatorios i.i.d. con vector de medias* μ *y matriz de covarianzas* $\mathbf{\Sigma}$*. Entonces*

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

esto es

$$\bar{\mathbf{X}}_n \sim AN\left(\mu, \frac{1}{n}\mathbf{\Sigma}\right)$$

2. Sean $\{\mathbf{X}_i\}$ *vectores aleatorios independientes con medias* $\{\mu_i\}$ *y matrices de covarianzas* $\{\mathbf{\Sigma}_i\}$ *y funciones de distribución* $\{F_i\}$*. Suponga que*

$$\frac{\mathbf{\Sigma}_1 + \cdots + \mathbf{\Sigma}_n}{n} \to \mathbf{\Sigma},$$

para $n \to \infty$ *y que*

$$\frac{1}{n}\sum_{i=1}^{n}\int_{\|\mathbf{x}-\mu_i\|>\epsilon\sqrt{n}} \|\mathbf{x} - \mu_i\|^2 \, dF_i(\mathbf{x}) \to 0,$$

para $n \to \infty, \ \forall \epsilon > 0$.

Entonces

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i \sim AN\left(\frac{1}{n}\sum_{i=1}^{n}\mu_i, \frac{1}{n}\mathbf{\Sigma}\right)$$

El T.C.L. y la Distribución Multinomial

Sea $X_1, X_2, \cdots, X_n$ una muestra aleatoria de una distribución multinomial con vector de probabilidades

$$\pi = \begin{bmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \\ . \\ . \\ . \\ \pi_R \end{bmatrix}$$

Cada vector X_i puede está compuesto de ceros y un uno que nos indica a cuál categoría pertence el i-ésimo elemento de la muestra.

Su matriz de varianzas y covarianzas es

$$\Sigma_X = \begin{bmatrix} \pi_1(1-\pi_1) & -\pi_1\pi_2 & \cdots & -\pi_1\pi_R \\ -\pi_1\pi_2 & \pi_2(1-\pi_2) & \cdots & -\pi_2\pi_R \\ \vdots & \vdots & & \vdots \\ -\pi_1\pi_R & -\pi_2\pi_R & \cdots & \pi_R(1-\pi_R) \end{bmatrix}$$

El estimador de máxima verosimilitud en el caso multinomial de π es

$$\widehat{\pi} = \sum_{i=1}^{n} X_i = \begin{bmatrix} \widehat{\pi}_1 \\ \widehat{\pi}_2 \\ \vdots \\ \widehat{\pi}_R \end{bmatrix}$$

Y la matriz de varianzas y covarianzas del estimador es

$$\Sigma_{\widehat{\pi}} = \frac{1}{n} \begin{bmatrix} \pi_1(1-\pi_1) & -\pi_1\pi_2 & \cdots & -\pi_1\pi_R \\ -\pi_1\pi_2 & \pi_2(1-\pi_2) & \cdots & -\pi_2\pi_R \\ \vdots & \vdots & & \vdots \\ -\pi_1\pi_R & -\pi_2\pi_R & \cdots & \pi_R(1-\pi_R) \end{bmatrix}$$

Este estimador es asintóticamente normal con vector de medias π y matriz de varianzas y covarianzas $\Sigma_{\widehat{\pi}}$.

Distribución Asintótica del Producto Multinomial

Si X vector aleatorio que sigue una distribución producto multinomial, entonces $\widehat{\pi}$ se distribuye asintóticamente normal multivariable con vector de medias π y matriz de varianzas y covarianzas dado por

$$\boldsymbol{\Sigma}_{SR\times SR} = \begin{bmatrix} \boldsymbol{\Sigma}_1 & 0 & \cdots & 0 \\ 0 & \boldsymbol{\Sigma}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{\Sigma}_S \end{bmatrix}$$

$$\widehat{\pi} \sim AN\left(\pi, \boldsymbol{\Sigma}_{SR\times SR}\right)$$

Si consideramos $\log(\pi)$, entonces el e.m.v., digamos $\log(\widehat{\pi})$ se distribuye, asintóticamente

$$\log\left(\widehat{\pi}\right) \sim AN\left(\log\left(\pi\right), \mathbf{D}\boldsymbol{\Sigma}_{SR\times SR}\mathbf{D}\right)$$

donde

$$\mathbf{D} = \begin{bmatrix} \frac{1}{\pi_{11}} & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & \frac{1}{\pi_{12}} & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\pi_{1R}} & 0 & \vdots & \cdots & 0 \\ 0 & 0 & \cdots & 0 & \frac{1}{\pi_{21}} & 0 & \cdots & 0 \\ 0 & 0 & 0 & \vdots & 0 & \frac{1}{\pi_{22}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & \cdots & \cdots & \frac{1}{\pi_{SR}} \end{bmatrix}$$