8 feb

Varianza muestral generalizada y Normal multivariada

$$S_n = \frac{1}{n}\sum_j (x_{ji} - \bar{x}_i)$$

$$S = \frac{1}{n-1}\sum_j (x_{ji} - \bar{x}_i) \to \text{Insesgado}$$

con esta se puede hacer una matriz de varianzas y covarianzas muestral.

→ ¿Qué tan dispersos son nuestros datos en general?

Varianza muestral generalizada $= |S|$ → esto es el determinante de la matriz S

→ mat de covs y vars

es una suerte de volumen de los vectores de desviación

- $|S| = 0 \iff$ hay al menos un vector desviación que es combinación lineal de los otros.

- también si $n \le p$ y $X = \left[\quad \right] \Rightarrow |S| = 0$.
  - p → n variables
  - n datos
  - matriz de datos

  (deben haber más observaciones que variables)

→ Varianza generalizada determinada por $|R|$ → matriz de coefs. de correlación.

→ Si s_{ii} es muy grande (o muy pequeña) geométricamente el vector de desviación correspondiente $d_i = (y_i - \bar{x}_i 1_n)$ será muy largo o muy corto y afectará el volumen.

Para evitar eso es útil escalar todos los valores de desviación para que tengan la misma longitud esto es aplicar a todos los datos originales:

$$x_{ji} = (x_{ji} - \bar{x}_i)/s_{ii}$$

entonces la matriz de varianzas y covarianzas es R.

luego $|R|$ = Varianza muestral generalizada de las variables estandarizadas

También existe la varianza total de la muestra

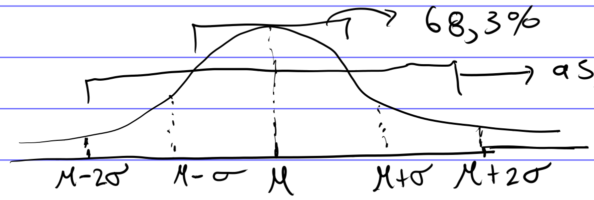
$$VTM = \sum_i S_{ii}$$

Normal multivariada

Recordemos: la PDF de una v.a. normal con media μ y var σ^2 es:

$$f_x(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-\mu)}{2\sigma^2}}, \quad x \in \mathbb{R}$$

ley empírica:



- $X \sim N(\mu, \sigma^2)$
- $\frac{(x-\mu)^2}{\sigma^2} \longrightarrow$ Distancia estadística entre X y μ

$= (x-\mu)(\sigma^2)^{-1}(x-\mu)$

Generalizando $\vec{x}$ vector de obs. de X, una fila de la matriz X.

distancia estadística de $\vec{x}$ a $\vec{\mu}$

$$(\vec{x}-\vec{\mu})' \Sigma^{-1} (\vec{x}-\vec{\mu})$$

entonces podemos generalizar y obtener la pdf de la normal multivariada.

$$f(x) = \frac{1}{(2\pi)^{p/2} \cdot |\Sigma|^{1/2}} \cdot e^{-(\vec{x}-\vec{\mu})' \Sigma^{-1} (\vec{x}-\vec{\mu})/2}$$

vector aleatorio $\leftarrow x = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix} \rightarrow$ v.a.

Denotamos la normal multivariada como:

$$x \sim N_p(\vec{\mu}, \Sigma)$$

Si $\overset{\#vars}{p} = 2$ y $\rho_{12} = 0 = cov(x_1, x_2) \Rightarrow x_1$ y x_2 indep.

- Si son v.a.s normales y sus correlaciones son cero entonces son indep.

Teorema: Si Σ es def positiva (Σ^{-1} existe) entonces si $\vec{e}$ es un vector propio de Σ con valor propio asociado λ, $\vec{e}$ es un vector propio de Σ^{-1} con val propio asociado $\frac{1}{\lambda}$ además Σ^{-1} def positiva.

obs: 1) $(\vec{x}-\vec{\mu})' \Sigma^{-1} (\vec{x}-\vec{\mu}) \leq \chi^2_p(\alpha)$

p → gl

Proposición 1) Si $\vec{X}$ vector aleatorio normal multivariado $X \sim N_p(\vec{\mu}, \Sigma)$ entonces

Si $\vec{a} \in \mathbb{R}^p$, $\vec{a}'\vec{x} = \sum a_i x_i$
$\approx N(\vec{a}'\vec{\mu}, \vec{a}'\Sigma\vec{a})$

Prop 2) Si $a'\vec{X} \sim N(\vec{a}'\vec{\mu}, \vec{a}'\Sigma\vec{a})$, $\forall a \in \mathbb{R}^p$ entonces $X \sim N_p(\vec{\mu}, \Sigma)$

Prop 3) $X \sim N_p(\vec{\mu}, \Sigma)$ sea $\underset{q\times p}{A} \in \mathbb{R}^{q\times p}$

$$AX \sim N_q(A\vec{\mu}, A\Sigma A')$$

Prop 4) $X \sim N_p(\vec{\mu}, \Sigma)$ $\vec{d} \in \mathbb{R}^p$

$$X + \vec{d} \sim N_p(\vec{\mu} + \vec{d}, \Sigma)$$

Prop 5) todas las particiones de vectores aleatorios normales resultan en vect. aleatorios normales.

$X \sim N_p(\vec{\mu}, \Sigma)$, $\quad X = \begin{pmatrix} X_{(1)} \\ \hline X_{(2)} \end{pmatrix} \begin{matrix} \}\, q \\ \}\, p-q \end{matrix}$

$\mu = \begin{pmatrix} \vec{\mu}_{(1)} \\ \hline \vec{\mu}_{(2)} \end{pmatrix}$

$X_{(1)} \sim N_q(\vec{\mu}_{(1)}, \Sigma_{11})$

$X_{(2)} \sim N_{p-q}(\vec{\mu}_{(2)}, \Sigma_{22})$ $\quad \Sigma = \left(\begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{array}\right)$

Prop 6) ① Si $\underset{q_1\times 1}{X_{(1)}}$, $\underset{q_2\times 1}{X_{(2)}}$ normales multivariadas e independientes entonces $cov(X_{(1)}, X_{(2)})=0$

$q_1\times q_2$ matriz

② si $X = \begin{pmatrix} X_{(1)} \\ \hdashline X_{(2)} \end{pmatrix} \sim N_{q_1+q_2}\left(\begin{pmatrix} \mu_{(1)} \\ \hdashline \mu_{(2)} \end{pmatrix}, \left(\begin{array}{c:c} \Sigma_{11} & \Sigma_{12} \\ \hdashline \Sigma_{21} & \Sigma_{22} \end{array}\right)\right)$

$X_{(1)}$ y $X_{(2)}$ indep sii $\Sigma_{12}=0$

Covarianza cero sí implica independencia cuando son Normales

③ $X_{(1)}^{q_1\times 1}$, $X_{(2)}^{q_2\times 1}$ normales indep sii

$$X = \begin{pmatrix} X_{(1)} \\ \hdashline X_{(2)} \end{pmatrix} \sim N_{q_1+q_2}\left(\begin{pmatrix} \mu_{(1)} \\ \hdashline \mu_{(2)} \end{pmatrix}, \left(\begin{array}{c:c} \Sigma_{11} & 0 \\ \hdashline 0 & \Sigma_{22} \end{array}\right)\right)$$

prop 7) sea $X \sim N_p(\vec{\mu}, \Sigma)$, $|\Sigma| > 0$

entonces:

① $(X - \vec{\mu})' (\Sigma^{-1}) (X - \vec{\mu}) \sim \chi^2_p$

la distancia de los valores a la media tiene distr. χ^2

$$\{(x - \vec{\mu})' (\Sigma^{-1}) (X - \vec{\mu}) \leq \chi^2_p(\alpha)\}$$

es $1-\alpha$

Muestreo de la normal multivariada

suponga que $\vec{X}_1, \dots, \vec{X}_n$ son una muestra aleatoria de una población normal multivariada con media $\vec{\mu}$ y cov Σ

Densidad conjunta de $\vec{X}_1, \dots, \vec{X}_n$

$$\prod_{j=1}^{n} \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}[(\vec{x}_j - \vec{\mu})' \Sigma^{-1} (\vec{x}_j - \vec{\mu})]}$$

La función de verosimilitud (función de $\vec{\mu}$ y Σ)

Dadas las observaciones $x_1, \dots, x_n : L(\mu, \Sigma)$

Método de máxima verosimilitud:

Utilizar como estimaciones de parámetros poblacionales desconocidos los valores que maximizan $L(\vec{\mu}, \Sigma)$ los valores que "mejor" explican los datos.

Teorema: Sean $X_1, \dots, X_n$ una muestra aleatoria de una población normal con media $\vec{\mu}$ y cov. Σ entonces el MLE:

$$\hat{\vec{\mu}} = \overline{X} \quad , \quad \hat{\Sigma} = \frac{1}{n} \sum_{i=1} (X_i - \overline{X})(X_i - \overline{X})'$$

$$= S_n = \frac{n-1}{n} S$$

obs: $\hat{\vec{\mu}}$, $\hat{\Sigma}$ son vectores aleatorios hasta el momento en que hay data.

② $$L(\hat{\vec{\mu}}, \hat{\Sigma}) = \frac{1}{(2\pi)^{\frac{np}{2}}} \cdot e^{-\frac{np}{2}} \times \frac{1}{|\underbrace{\hat{\Sigma}}_{\frac{n-1}{n} S}|^{n/2}}$$

la varianza generalizada determina el máximo de la función de verosimilitud