Cap. 3 Muestreo aleatorio:

Jueves 30 de Agosto 2007

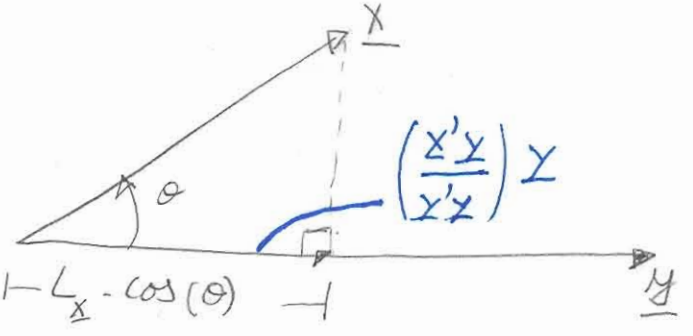
Una muestra de 1 obsn. multivar. es el conjunto de mediciones de 'p' variables distintas en el mismo sujeto ó ensayo. Si se obtienen 'n' observaciones, los datos se resumen en la matriz $\underset{(n\times p)}{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & & x_{np} \end{bmatrix}$ ← 1 obs. 4D. con 'p' componentes

## 1) Representación Geométrica de la muestra:

Re-escribamos $X$ así: $\underset{(n\times p)}{X} = [\underline{Y}_1 \,|\, \underline{Y}_2 \,|\, \cdots \,|\, \underline{Y}_p]$, donde

$\underline{Y}_i' = [x_{1i}\ x_{2i} \cdots x_{ni}]$ son las 'n' mediciones hechas para los 'n' individuos, respecto de la 'i'-ésima variable

Apartado: La proyección (o sombra) del vector $\underline{X}$ sobre el vector $\underline{Y}$ es: $\dfrac{(\underline{X}'\underline{Y})}{\underline{Y}'\underline{Y}}\underline{Y} = \dfrac{(\underline{X}'\underline{Y})}{L_{\underline{Y}}}\dfrac{1}{L_{\underline{Y}}}\underline{Y}$, donde $L_{\underline{Y}} = \sqrt{\underline{Y}'\underline{Y}}$




$\cos(\theta) = \dfrac{\underline{X}'\underline{Y}}{L_{\underline{X}} L_{\underline{Y}}}$

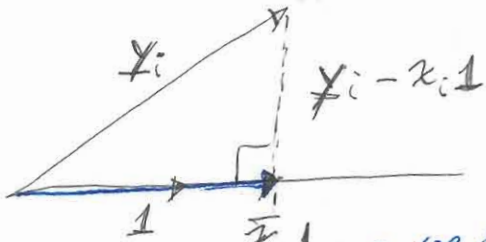
$L_{\underline{X}} = \sqrt{\underline{X}'\underline{X}}$

(ver p. 53 cap. 2)

Consideren el vector $\frac{1}{\sqrt{n}}\begin{bmatrix}1\\1\\\vdots\\1\end{bmatrix}_{(n\times 1)} = \frac{1}{\sqrt{n}}\cdot\mathbf{1}$. Entonces,

$\frac{1}{\sqrt{n}}\cdot\mathbf{1} = 1$. La proyección del vector $\underline{y_i}$ en el vector unitario $\left(\frac{1}{\sqrt{n}}\right)\mathbf{1}$ es igual a: $\underline{y_i}'\left(\frac{1}{\sqrt{n}}\mathbf{1}\right)\cdot\frac{1}{\sqrt{n}}\cdot\mathbf{1}$, y

$\underline{y_i}'\left(\frac{1}{\sqrt{n}}\mathbf{1}\right)\frac{1}{\sqrt{n}}\cdot\mathbf{1} = \frac{x_{1i}+x_{2i}+\cdots+x_{ni}}{n}\cdot\mathbf{1} = \bar{x}_i\mathbf{1}$. Esto es, la media en la muestra de la variable '$i$' $\left(\bar{x}_i = \frac{\underline{y}'\mathbf{1}}{n}\right)$, es la cantidad por la cual hay que multiplicar el vector $\mathbf{1}$ para obtener la proyección de $\underline{y_i}$ sobre la línea determinada por $\mathbf{1}$:




$\bar{x}_i\mathbf{1}$ → vector proyección de $\underline{y_i}$ sobre $\left(\frac{1}{\sqrt{n}}\right)\mathbf{1}$

Entonces,

↳ $\underline{d_i} = \begin{bmatrix}x_{1i}-\bar{x}_i\\x_{2i}-\bar{x}_i\\\vdots\\x_{ni}-\bar{x}_i\end{bmatrix}$ es entonces el vector que contiene las desviaciones de la media $\bar{x}_i$ para cada observación, para la variable '$i$'.

La varianza de la muestra para la variable 'i' es proporcional al (largo)$^2$ de $\underline{d}_i$:

$$\left(L_{\underline{d}_i}\right)^2 = \underline{d}_i' \underline{d}_i = \sum_{j=1}^{n} (x_{ji} - \bar{x}_i)^2 \quad \text{ec. (*)}$$

Así mismo, el largo de dichas desviaciones es proporcional a la desviación estándar de la muestra (para la variable 'i').

Para dos vectores de desviaciones $\underline{d}_i$ y $\underline{d}_k$,

$$\underline{d}_i' \underline{d}_k = \sum_{j=1}^{n} (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k) \quad \text{ec. (**)}$$

Si $\theta_{ik}$ denota el ángulo entre $\underline{d}_i$ y $\underline{d}_k$, entonces

$$\underline{d}_i' \underline{d}_k = L_{\underline{d}_i} L_{\underline{d}_k} \cos(\theta_{ik}) \quad , \text{ y usando las ecs. (*) y (**):}$$

$$\underbrace{\sum_{j=1}^{n} (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)}_{= s_{ik}} = \underbrace{\sqrt{\sum_{j=1}^{n} (x_{ji} - \bar{x}_i)^2}}_{\sqrt{s_{ii}}} \underbrace{\sqrt{\sum_{j=1}^{n} (x_{jk} - \bar{x}_k)^2}}_{\sqrt{s_{kk}}} \times \cos(\theta_{ik})$$

Por lo que:

$$\cos(\theta_{ik}) = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}} = r_{ik} \quad //$$

El coseno del ángulo entre $\underline{d}_i$ y $\underline{d}_k$ es la correlación entre la muestra 'i' y 'k'. Entonces, si dos vectores desviación tienen casi la misma orientación (i.e. $\theta_{ik} \approx 0$), entonces la correlación va a ser casi 1. Si dos vectores son perpendiculares, la correlación de la muestra será 0 y si están orientados de manera opuesta, -1.

## 2) Muestreo aleatorio y momentos de la muestra.

Antes del muestreo, asumimos que cada uno de los vectores $\underline{X}_j$ que tendrá mediciones en 'p' variables es un vector aleatorio, y que juntos, los 'n' vectores conforman una matriz aleatoria:

$$\underset{(n\times p)}{X} = \begin{bmatrix} \underline{X}_1' \\ \underline{X}_2' \\ \vdots \\ \underline{X}_n' \end{bmatrix} \qquad \text{ec (***)}$$

**Def.:** *Muestra aleatoria:* Si los vectores $\underline{X}_1', \underline{X}_2', \dots, \underline{X}_n'$ en (***) representan observaciones independientes de una misma distribución conjunta con densidad $f(\underline{x}) = f(x_1, x_2, \dots, x_p)$, entonces se dice que $\underline{X}_1, \underline{X}_2, \dots, \underline{X}_n$ forman una muestra aleatoria de $f(\underline{x})$. La distribución conjunta de la muestra está dada por

$$f(\underline{x}_1) f(\underline{x}_2) \cdots f(\underline{x}_n), \text{ donde } f(\underline{x}_j) = f(x_{j1}, x_{j2}, \dots, x_{jp}).$$

**Resultado 2.1:** Sea $\underline{X}_1, \underline{X}_2, \dots, \underline{X}_n$ una muestra aleatoria de una distribución conjunta con vector media $\underline{\mu}$ y matriz de var-cov: $\Sigma$. Entonces $\overline{\underline{X}}$ es un vector estimador insesgado de $\underline{\mu}$ y su matriz de var-cov es $\frac{1}{n}\Sigma$. Además, $\frac{n}{n-1} S_n$ es un estimador insesgado de $\Sigma$, i.e.:

$$E\left[\left(\frac{n}{n-1}\right) S_n\right] = \Sigma$$

Demo/ $\mathbb{E}(\bar{X}) = \mathbb{E}\left( \frac{1}{n} X_1 + \frac{1}{n} X_2 + \cdots + \frac{1}{n} X_n \right)$

$$= \frac{1}{n} \mathbb{E} X_1 + \cdots + \frac{1}{n} \mathbb{E} X_n = \sum_{j=1}^{n} \frac{1}{n} \mathbb{E}(X_j)$$

$$\overset{iid}{=} \frac{1}{n} n \mu = \mu \quad //$$

Ahora bien:

$$(\bar{X} - \mu)(\bar{X} - \mu)' = \left( \frac{1}{n} \sum_{j=1}^{n} (X_j - \mu) \right) \left( \frac{1}{n} \sum_{\ell=1}^{n} (X_\ell - \mu) \right)'$$

$$\overset{(\text{¿Porque?})}{=} \frac{1}{n^2} \sum_{j=1}^{n} \sum_{\ell=1}^{n} (X_j - \mu)(X_\ell - \mu)' .$$

Chequenla!

Entonces $\mathbb{E}\left[ (\bar{X} - \mu)(\bar{X} - \mu)' \right] = \frac{1}{n^2} \sum_{j=1}^{n} \sum_{\ell=1}^{n} \mathbb{E}\left[ (X_j - \mu)(X_\ell - \mu)' \right]$. (ec. Δ)

Para todo $j \neq \ell$, $\mathbb{E}\left[ (X_j - \mu)(X_\ell - \mu)' \right] = 0$ (¿Porque?),

entonces $Cov(\bar{X}) = \mathbb{E}\left[ (\bar{X} - \mu)(\bar{X} - \mu)' \right] \overset{(ec.\Delta)}{=} \frac{1}{n^2} \sum_{j=1}^{n} \mathbb{E}\left[ (X_j - \mu)(X_j - \mu)' \right]$.

$$\overset{X_j \text{ iid}}{=} \frac{1}{n^2} \times n \cdot \Sigma = \frac{1}{n} \Sigma .$$

Ahora calculemos $\mathbb{E}[S_n]$. Para esto, notemos que:

$$\sum_{j=1}^{n}(\underline{X}_j-\bar{\underline{X}})(\underline{X}_j-\bar{\underline{X}})' = \sum_{j=1}^{n}(\underline{X}_j-\bar{\underline{X}})\underline{X}_j' + \Big(\sum_{j=1}^{n}(\underline{X}_j-\bar{\underline{X}})\Big)(-\bar{\underline{X}})' \quad (\text{el paréntesis} = 0)$$

$$= \sum_{j=1}^{n}\underline{X}_j\underline{X}_j' - \sum_{j=1}^{n}\bar{\underline{X}}\underline{X}_j'$$

$$= \sum_{j=1}^{n}\underline{X}_j\underline{X}_j' - \bar{\underline{X}}\sum_{j=1}^{n}\underline{X}_j'$$

$$= \sum_{j=1}^{n}\underline{X}_j\underline{X}_j' - \bar{\underline{X}}\,n\,\bar{\underline{X}}' = \sum_{j=1}^{n}\underline{X}_j\underline{X}_j' - n\bar{\underline{X}}\bar{\underline{X}}'$$

$$\therefore \quad \mathbb{E}\left[\sum_{j=1}^{n}(\underline{X}_j-\bar{\underline{X}})(\underline{X}_j-\bar{\underline{X}})'\right] = \sum_{j=1}^{n}\mathbb{E}(\underline{X}_j\underline{X}_j') - n\,\mathbb{E}\left(\bar{\underline{X}}\bar{\underline{X}}'\right) \quad (\text{ec. } \boxtimes)$$

Ahora bien, para todo vector aleatorio $\underline{V}$ con media y covarianza $\mu_V$ y $\Sigma_V$ respectivamente, $\mathbb{E}(VV') = \Sigma_V + \mu_V\mu_V'$, por lo tanto:

$$\text{L.D. de ec. } \boxtimes \;(\text{lado derecho}) = \sum_{j=1}^{n}\left(\Sigma + \underline{\mu}\underline{\mu}'\right) - n\left[\frac{1}{n}\Sigma + \underline{\mu}\underline{\mu}'\right]$$

$$\underline{X}_j \overset{iid}{=} n\Sigma + n\underline{\mu}\underline{\mu}' - \Sigma - n\underline{\mu}\underline{\mu}'$$

$$= (n-1)\Sigma. \quad \text{Como } S_n = \frac{1}{n}\left(\sum_{j=1}^{n}\underline{X}_j\underline{X}_j' - n\bar{\underline{X}}\bar{\underline{X}}'\right)$$

resulta que $\mathbb{E}(S_n) = \dfrac{(n-1)}{n}\Sigma$ $\square$

: Matriz Var-Cov (insesgada):

$$S = \left(\frac{n}{n-1}\right) S_n = \frac{1}{n-1} \sum_{j=1}^{n} (\underline{X}_j - \bar{\underline{X}})(\underline{X}_j - \bar{\underline{X}})'$$

En lo que queda del curso, $S$ reemplazará a $S_n$ en la mayoría de las pruebas estadísticas que veremos.

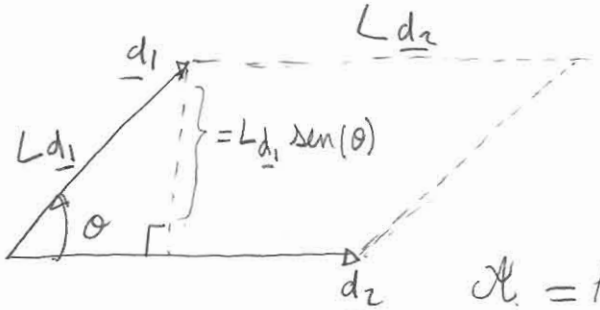
## 3) Varianza Generalizada:

Con una sóla variable, la varianza de la muestra es suficiente para describir la variabilidad en los datos. Cuando se observan $p$ variables, se puede representar la variabilidad total de la muestra con un sólo número: el determinante de $S$. A este se le llama "Varianza generalizada".

### 3.1 Varianza generalizada cuando $p=2$:

Sean $S = \begin{bmatrix} S_{11} & S_{12} \\ S_{12} & S_{22} \end{bmatrix}$ y $\begin{cases} \underline{d}_1 = \underline{X}_1 - \bar{X}_1 \underline{1}, \; \underline{d}_2 = \underline{X}_2 - \bar{X}_2 \underline{1}, \\ L_{\underline{d}_1} = \sqrt{\sum_{j=1}^{n} (x_{j1} - \bar{x}_1)^2} = \sqrt{(n-1) S_{11}} \\ L_{\underline{d}_2} = \sqrt{\sum_{j=1}^{n} (x_{j2} - \bar{x}_2)^2} = \sqrt{(n-1) S_{22}} \end{cases}$

Entonces,




$\mathcal{A}$ = Area paralelogramo = $L_{\underline{d}_2} \times (L_{\underline{d}_1} \operatorname{sen}(\theta))$

$$\mathcal{A} = \| d_2 \| \| d_1 \| \operatorname{sen}(\theta) = \| d_2 \| \| d_1 \| \sqrt{1 - \cos^2(\theta)}$$

$$= \sqrt{(n-1) S_{22}} \sqrt{(n-1) S_{11}} \sqrt{1 - r_{12}^2} \quad (\text{porque } \cos(\theta) \; r_{12})$$

$$= (n-1) \sqrt{S_{11}} \sqrt{S_{22}} \sqrt{1 - r_{12}^2} .$$

Teniendo en cuenta que $|S| = S_{11} S_{22} - S_{12})^2$

$$= S_{11} S_{22} - \left(\sqrt{S_{11}} \sqrt{S_{22}} \, r_{12}\right)^2$$

$$= S_{11} S_{22} \left(1 - r_{12}^2\right) ,$$

entonces $|S| = \mathcal{A}^2 / (n-1)^2$ y en $p$ dimensiones:

Varianza generalizada $= |S| = (n-1)^{-p} (\text{Volumen})^2$

este es el significado geométrico para los $p$ valores de las variables.

Considerando la interpretación de una forma cuadrática como una distancia, y sustituyendo la matriz $A$ en la def. por $S^{-1}$ y $\bar{x}$ como el punto fijo, entonces las coordenadas $x' = [x_1, x_2, \ldots x_p]$ de los puntos $q$ están a una distancia cte. '$c$' de $\bar{x}$ satisfacen

$$(x - \bar{x})' S^{-1} (x - \bar{x}) = c^2 .$$

La hyper-elipsoide definida por $(\underline{X}-\bar{\underline{X}})' S^{-1} (\underline{X}-\bar{\underline{X}}) = c^2$ está relacionada a la variánza generalizada por la ecuación

$$\text{Volúmen de } \left\{ \underline{X} : (\underline{X}-\bar{\underline{X}})' S^{-1} (\underline{X}-\bar{\underline{X}}) \leq c^2 \right\} = \frac{2\pi^{p/2} \cdot c^p \cdot |S|^{1/2}}{p\,\Gamma\left(\frac{p}{2}\right)}$$