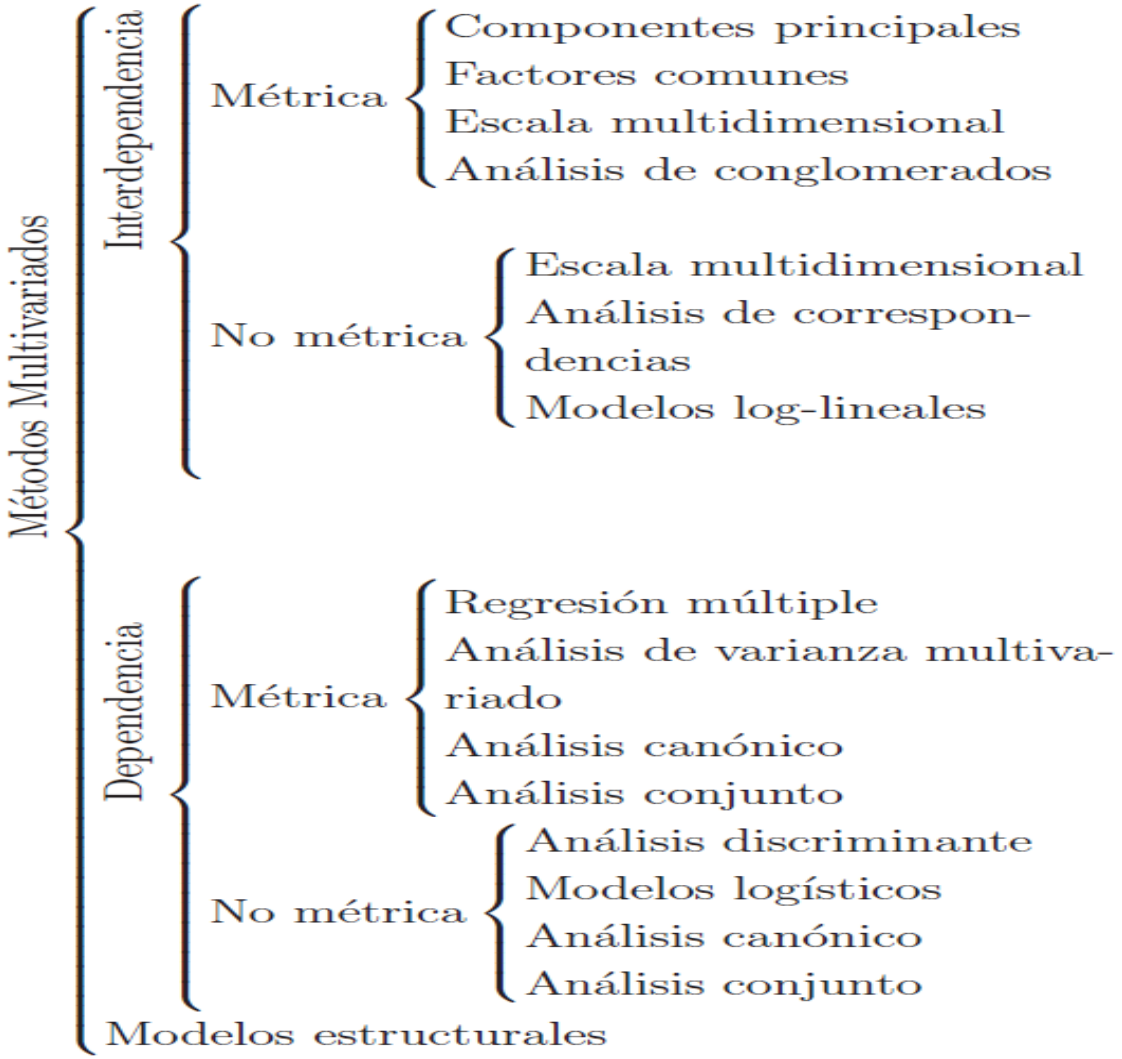
TECNICAS MULTIVARIADAS



Dado un vector aleatorio $\boldsymbol{X}$, como el definido en (1.1), el *valor esperado* de $\boldsymbol{X}$, notado $\mathrm{E}(\boldsymbol{X})$, es el vector de valores esperados de cada una de las variables aleatorias, así:

$$\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{X}) = \begin{pmatrix} \mathrm{E}(X_1) \\ \vdots \\ \mathrm{E}(X_p) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix}$$

La *matriz de varianzas y covarianzas* de $\boldsymbol{X}$, la cual notaremos por $\boldsymbol{\Sigma}$, está dada por:

$$\boldsymbol{\Sigma} = \mathrm{cov}(\boldsymbol{X}) = \mathrm{E}\left\{(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})'\right\} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix} \quad (1.2)$$

Donde σ_{ij} denota la covarianza entre la variable X_i y la variable X_j, la cual se define como:

$$\sigma_{ij} = \mathrm{E}[(X_i - \mu_i)(X_j - \mu_j)].$$

Al desarrollar el producto y aplicar las propiedades del valor esperado, se obtiene una expresión alterna para la matriz de varianzas y covarianzas; ésta es

$$\boldsymbol{\Sigma} = \mathrm{cov}(\boldsymbol{X}) = \mathrm{E}(\boldsymbol{X}\boldsymbol{X}') - \boldsymbol{\mu}\boldsymbol{\mu}' \quad (1.3)$$

Los elementos de la diagonal de la matriz (1.2) corresponden a las varianzas de cada una de las variables, los elementos fuera de la diagonal son las covarianzas entre las variables correspondientes de la fila y la columna.

Gran número de las metodologías señaladas en la primera parte de este capítulo se basan en la estructura y propiedades de $\boldsymbol{\Sigma}$; se destacan entre otras las siguientes propiedades:

1. La matriz $\boldsymbol{\Sigma}$ es simétrica; es decir, $\boldsymbol{\Sigma}' = \boldsymbol{\Sigma}$, puesto que $\sigma_{ij} = \sigma_{ji}$.

2. Los elementos de la diagonal de $\boldsymbol{\Sigma}$ corresponden a la varianza de las respectivas variables ($\sigma_{ii} = \sigma_i^2$).

3. Toda matriz de varianzas y covarianzas es *definida no negativa* ($|\boldsymbol{\Sigma}| \geq 0$). Y es definida positiva, cuando el vector aleatorio es continuo.

4. Si $\mathrm{E}(\boldsymbol{X}) = \boldsymbol{\mu}$ y $\mathrm{cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}$, entonces:

$$\mathrm{E}(\boldsymbol{AX} + \boldsymbol{b}) = \boldsymbol{A\mu} + \boldsymbol{b} \quad \text{y} \quad \mathrm{cov}(\boldsymbol{AX} + \boldsymbol{b}) = \boldsymbol{A\Sigma A}',$$

con $\boldsymbol{A}$ matriz de constantes de tamaño $(q \times p)$ y $\boldsymbol{b}$ vector $(q \times 1)$ también de constantes.

Se dice que un conjunto de datos es una *muestra aleatoria* multivariada si ésta tiene la misma probabilidad de extraerse que cualquier otra del mismo tamaño. A cada individuo (objeto) seleccionado de manera aleatoria de la población de individuos, se le registran una serie de atributos u observaciones (valores de las variables aleatorias). Sea x_{ij} la observación de la j-ésima variable en el i-ésimo individuo, se define la *matriz de datos multivariados* como el arreglo

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

La matriz $\boldsymbol{X}$ también puede definirse como el arreglo de vectores fila o vectores columna. El i-ésimo vector fila se nota por $\boldsymbol{X}_{(i)}$ y el j-ésimo vector columna se nota por $\boldsymbol{X}^{(j)}$. Así cada uno denota el i-ésimo individuo o la j-ésima variable respectivamente.
Se define la *media muestral* de la j-ésima variable por

$$\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}, \quad \text{con } j = 1, \ldots, p.$$

El vector formado por las p–medias muestrales, es el vector de promedios o de medias (centroide de los datos)

$$\overline{\boldsymbol{X}}' = \frac{1}{n}\mathbf{1}'\boldsymbol{X} = (\overline{x}_1, \ldots, \overline{x}_p)$$

donde $\mathbf{1}$ es el vector columna de n unos.
Se define la covarianza muestral entre la variable columna j y la variable columna k como:

$$s_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)(x_{ik} - \overline{x}_k); \quad j,k = 1, \ldots, p.$$

nótese que si $j = k$, se obtiene la varianza muestral asociada a la variable j-ésima. La matriz constituida por las covarianzas s_{ij}, es la *matriz de varianzas y covarianzas muestral*, ésta es:

$$\boldsymbol{S} = \frac{1}{n}\mathbb{X}'(\boldsymbol{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}')\mathbb{X} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{pmatrix}$$

En algunas circunstancias se necesita disponer de un solo número que señale la dispersión de los datos; la *varianza generalizada* y la *variabilidad total* son dos de tales parámetros. La varianza generalizada se define como el determinante de la matriz $\boldsymbol{S}$, y se nota $|\boldsymbol{S}|$; es decir,

$$VG = |\boldsymbol{S}|$$

La varianza total se define como la traza de la matriz $\boldsymbol{S}$; téngase presente que los elementos de la diagonal de $\boldsymbol{S}$ son las varianzas de cada una de las variables:

$$VT = \mathrm{tr}(\boldsymbol{S}) = \sum_{j=1}^{p} s_j^2.$$

Aunque a mayor variabilidad, los valores de VG y de VT aumentan, se debe tener cuidado por la influencia de valores extremos en la varianza. Su raíz cuadrada se denomina la *desviación típica generalizada*. Nótese que si $p = 1$; $VG = VT = s^2$.

También a partir de la matriz $\boldsymbol{S}$ se puede obtener la matriz de correlación $\boldsymbol{R}$, cuyos elementos son los coeficientes de correlación entre cada par de variables. Cada elemento r_{jk} de $\boldsymbol{R}$ es de la forma:

$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}s_{kk}}},$$

donde r_{jk} es el *coeficiente de correlación lineal* entre la variable j y la variable k.

$$\boldsymbol{R} = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{12} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix} = \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{S}\boldsymbol{D}^{-\frac{1}{2}}, \tag{1.11}$$

donde $\boldsymbol{D}^{-\frac{1}{2}}$ es la matriz diagonal con los inversos de las desviaciones estándar sobre la diagonal; es decir, $\boldsymbol{D}^{-\frac{1}{2}} = \text{diag}(1/s_j)$.

Ejemplo 1.4.1. Los siguientes datos se refieren a la altura de una planta X_1 (en m.), su longitud radicular X_2 (en cm), su área foliar X_3 (en cm^2) y su peso en pulpa X_4 (en gm.), de una variedad de manzano. Los datos (matriz $\boldsymbol{X}$) se presentan en la tabla 1.3.

Tabla 1.3: Medidas sobre manzanos

Obs.	X_1	X_2	X_3	X_4
1	1.38	51	4.8	115
2	1.40	60	5.6	130
3	1.42	69	5.8	138
4	1.54	73	6.5	148
5	1.30	56	5.3	122
6	1.55	75	7.0	152
7	1.50	80	8.1	160
8	1.60	76	7.8	155
9	1.41	58	5.9	135
10	1.34	70	6.1	140