CAPÍTULO I

1. LA PÉRDIDA DE DATOS EN UNA INVESTIGACIÓN

1.1 Introducción

El presente capítulo incluye los principios estadísticos relacionados con los Métodos de Imputación que serán parte de esta investigación. Para esto, se presenta, en la sección 1.2 los conceptos relacionados con matrices de datos multivariados, en la siguiente sección se muestra un resumen acerca de la “Pérdida de Datos” en una Investigación y por último se presentan los métodos que emplean toda la información disponible.

1.2 Matriz de Datos Multivariados

Una matriz es un arreglo rectangular de números reales, de n filas y p columnas que contiene información de una muestra aleatoria tomada de una población donde, por ejemplo, a n individuos se le realizan p preguntas. En el Cuadro 1.1, $\boldsymbol{X}$ es la matriz de datos y X_{ij} es el valor de la j-ésima variable investigada al i-ésimo individuo, es decir se miden p características a n individuos.

CUADRO 1.1

Efectos de la Imputación en el análisis de datos multivariados

Matriz de Datos Multivariados

$$\mathbf{X} = \begin{bmatrix} X_{11} & X_{12} & X_{1p} \\ X_{21} & X_{22} & X_{2p} \\ . & . & . \\ . & . & . \\ . & . & . \\ X_{n1} & X_{m2} & X_{np} \end{bmatrix} \quad ; \quad \mathbf{X} \in M_{nxp}$$

Elaborado por: G. Cuenca

1.3 Variables aleatorias Univariadas y Bivariadas

1.3.1 Variables aleatorias univariadas

Sea (Ω, S) un espacio muestral, donde Ω es el conjunto de todos los resultados posibles del experimento y S es el conjunto potencia de Ω, X es una función de valor real definida sobre los elementos de (Ω, S), es decir que: $X : \Omega \rightarrow \Re$, entonces X es una *variable*

aleatoria siendo $\Re$ el conjunto de los Números Reales. Las variables aleatorias pueden ser continuas o discretas.

Variable Aleatoria Discreta

Una Variable Aleatoria Discreta X es, una variable aleatoria para la cual el número de valores $X(w), w \in \Omega$, que puede tomar, es finito o infinito numerable.

Variable Aleatoria Continua

Una Variable Aleatoria Continua X es, una variable aleatoria que toma valores $X(w), w \in \Omega$, en una escala continua, para dos variables cualesquiera siempre se puede encontrar un valor intermedio.

Población Objetivo

Se denomina Población Objetivo al conjunto de todos los elementos acerca de cuyas características deseamos hacer alguna investigación de tipo estadístico.

Población Investigada

La Población Investigada es el conjunto de entes pertenecientes a la Población Objetivo, disponibles al momento de efectuar la investigación, debido a que no siempre se puede acceder a todas

las unidades de investigación que conforman la población objetivo, ya sea por negativas a colaborar, ausencias o cualquier otro tipo de inaccesibilidad. Si todos los entes motivos de la investigación están disponibles, entonces la Población Objetivo es igual a la Población Investigada.

Valores Esperados y Varianza de una Variable Aleatoria

El valor esperado de una función g , dada en términos de X está denotada como $E[g(X)]$ y definida de la siguiente forma:

$$E[g(X)] = \int_{-\infty}^{\infty} g(X) f(x) dx \tag{1.1}$$

Si X es continuo y es tal que su función de densidad $f(x)$ es conocida, la media μ de la población o valor esperado de X es definida como:

$$E(X) = \mu = \int_{-\infty}^{\infty} X \ f(x) dx \tag{1.2}$$

Es simple demostrar que:

a) $E(aX) = aE(X)$ (1.3)

b) $E[g(X)+h(X)] = E[g(X)] + E[h(X)]$ (1.4)

La varianza poblacional $Var(X)$ es definida como:

(1.5)

$$Var(X) = \sigma^2 = E(X - \mu)^2$$

y la función generadora de momentos se define como

$$M_X(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tX} f(x)dx.$$

Utilizando (1.3) y (1.4), la varianza poblacional puede ser expresada como:

$$\sigma^2 = E(X^2) - \mu^2 \qquad (1.6)$$

La raíz cuadrada de la varianza poblacional es llamada como desviación estándar de la población.

Aparte de $\left.\frac{\delta M}{\delta t}\right|_{t=0} = E(X)$ y en general la $\left.\frac{\delta^r M}{\delta t^r}\right|_{t=0} = E(X^r)$

Si cada valor de X es multiplicado por una constante a, la varianza de la población de X se multiplica por a^2, es decir:

$$Var(aX) = a^2 \sigma^2 \qquad (1.7)$$

Muestra

Una muestra $X_1, X_2, ..., X_n$, tomada de una población X, que es discreta, es aleatoria si y solo si, es escogida de tal forma que cada subconjunto de tamaño n en la población, tiene igual probabilidad

de constituir la muestra. La probabilidad de escoger una muestra de tamaño n de una población de tamaño N es $\frac{1}{\binom{N}{n}}$.

Una muestra $X_1, X_2, ..., X_n$, tomada de una población X, que es contìnua, es aleatoria, si y solo si $X_1, X_2, ..., X_n$ son variables aleatorias independientes e idénticamente distribuìdas.

La media aritmética $\overline{X}$ de una muestra aleatoria de tamaño n, $X_1, X_2, ..., X_n$ es definida por:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{1.8}$$

Si $X_1, X_2, ..., X_n$ es una muestra aleatoria de una población que tiene media μ y varianza σ^2, entonces la media de la muestra $\overline{X}$ es un estimador insesgado de la media poblacional μ, esto es:

$$E(\overline{X}) = \mu . \tag{1.9}$$

La media muestral tiene una propiedad similar a la que definimos en (1.3). Si el $Z_i = aX_i$ para $i = 1,2,3,...,n$, entonces $\overline{Z} = a\overline{X}$; veamos:

$$\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i = \frac{1}{n}\sum_{i=1}^{n} aX_i = \frac{1}{n}a\sum_{i=1}^{n} X_i$$

$$\bar{Z} = a\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = a\bar{X}$$

$$\bar{Z} = a\bar{X} \qquad (1.10)$$

Para una muestra de n observaciones, la varianza muestral se define como:

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} \qquad (1.11)$$

La que también es igual a:

$$s^2 = \frac{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}{n-1} \qquad (1.12)$$

Si $X_1, X_2, \ldots, X_n$ es una muestra aleatoria de una población con media μ y varianza σ^2, entonces la varianza muestral s^2 es un estimador insesgado de la varianza poblacional σ^2; esto es:

$$E(s^2) = \sigma^2 \qquad (1.13)$$

La cual se demuestra de la siguiente forma:

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

$$E(s^2) = E\left[\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}\right]$$

$$= \frac{1}{n-1}E\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right]$$

$$= \frac{1}{n-1}E\left(\sum_{i=1}^{n}X_i^2 - 2\bar{X}\sum_{i=1}^{n}X_i + \sum_{i=1}^{n}\bar{X}^2\right)$$

$$= \frac{1}{n-1}E\left(\sum_{i=1}^{n}X_i^2 - 2n\bar{X}^2 + n\bar{X}^2\right)$$

$$= \frac{1}{n-1}\left(\sum_{i=1}^{n}E(X_i^2) - nE(\bar{X}^2)\right)$$

$$= \frac{1}{n-1}\left(\sum_{i=1}^{n}(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right)$$

$$= \frac{1}{n-1}(n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2)$$

$$= \frac{1}{n-1}(n\sigma^2 - \sigma^2) = \frac{n-1}{n-1}\sigma^2 = \sigma^2$$

Similarmente, si definimos $Z_i = aX_i$, i=1,2,…,n, entonces la varianza muestral de Z es dada por $s_Z^2 = a^2 s^2$, la cual demostraremos a continuación:

$$s_Z^2 = \frac{\sum_{i=1}^{n}(Z_i - \bar{Z})^2}{n-1} = \frac{\sum_{i=1}^{n}(aX_i - a\bar{X})^2}{n-1}$$

$$= \frac{\sum_{i=1}^{n}[a(X_i - \bar{X})]^2}{n-1} = \frac{a^2\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

$$= a^2 s^2 \qquad (1.14)$$

1.3.2 Variables Aleatorias Bivariadas

Un vector aleatorio bivariado $\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ surge cuando dos características X_1 y X_2 son medidas de manera simultánea en cada ente que se investiga.

La covarianza poblacional es definida como:

$$\operatorname{cov}(X_i, X_j) = \sigma_{ij} = \mathrm{E}\left[(X_i - \mu_i)(X_j - \mu_j)\right] \tag{1.15}$$

donde μ_i y μ_j son las medias de X_i y X_j respectivamente. Se puede demostrar que:

$$\sigma_{ij} = \mathrm{E}(X_i, X_j) - \mu_i \mu_j \tag{1.16}$$

Para una muestra (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n) la covarianza muestral se define como:

$$s_{XY} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{n-1} = \hat{\sigma}_{XY} \tag{1.17}$$

La que es equivalente a:

$$s_{XY} = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{n-1} \tag{1.18}$$

La covarianza muestral s_{XY} es un estimador insesgado para la covarianza poblacional σ_{XY} es decir:

$$E(s_{XY}) = \sigma_{XY} \tag{1.19}$$

Puesto que la covarianza depende de la escala de la medida de X y Y, es difícil para comparar covarianzas entre diversos pares de variables. Por ejemplo, si cambiamos una medida de pulgadas a centímetros, la covarianza cambiará. Para encontrar una medida de la relación lineal que sea invariante a los cambios de escala, podemos estandardizar la covarianza dividiéndola para las desviaciones estándar de las dos variables. Esta covarianza estandardizada se llama usualmente coeficiente de correlación. La correlación poblacional de dos variables aleatorias X y Y es:

$$\rho_{XY} = corr(X,Y) = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sqrt{E(X-\mu_X)^2}\sqrt{E(Y-\mu_Y)^2}} \tag{1.20}$$

Y la correlación muestral se da por:

$$r_{XY} = \frac{s_{XY}}{s_X s_Y} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \tag{1.21}$$

El coeficiente de correlación poblacional y muestral es un valor entre -1 y 1.

1.3.3 Vectores Media y Matriz de Covarianza para Vectores Aleatorios

Supongamos que se tiene una muestra aleatoria multivariada de n vectores observados $\mathbf{X_1}, \mathbf{X_2}, ..., \mathbf{X_n}$, tomada de una población

p-variada $\mathbf{X}$. Dos vectores $\mathbf{X_1}$ y $\mathbf{X_2}$ son independientes, si cada variable X_{1j} en $\mathbf{X_1}$ es independiente de cada variable X_{2j} en $\mathbf{X_2}$. Ya que $\mathbf{X_1, X_2, ..., X_n}$ constituye una muestra aleatoria, entonces sus n vectores son independientes.

Los n vectores observados son transpuestos y listados como filas en la matriz de datos $\mathbf{X} \in \Re^p$:

$$\mathrm{X} = \begin{pmatrix} \mathrm{X}_1^{\mathrm{T}} \\ \mathrm{X}_2^{\mathrm{T}} \\ \cdot \\ \cdot \\ \cdot \\ \mathrm{X}_{\mathrm{i}}^{\mathrm{T}} \\ \cdot \\ \cdot \\ \cdot \\ \mathrm{X}_{\mathrm{n}}^{\mathrm{T}} \end{pmatrix} = \begin{pmatrix} X_{11} & X_{12} \ldots X_{1\mathrm{j}} & \ldots X_{1\mathrm{p}} \\ X_{21} & X_{22} \ldots X_{2\mathrm{j}} & \ldots X_{2\mathrm{p}} \\ \cdot & \cdot \quad \cdot & \cdot \\ \cdot & \cdot \quad \cdot & \cdot \\ \cdot & \cdot \quad \cdot & \cdot \\ X_{\mathrm{i}1} & X_{\mathrm{i}2} \ldots X_{\mathrm{ij}} & \ldots X_{\mathrm{ip}} \\ \cdot & \cdot \quad \cdot & \cdot \\ \cdot & \cdot \quad \cdot & \cdot \\ \cdot & \cdot \quad \cdot & \cdot \\ X_{\mathrm{n}1} & X_{\mathrm{n}2} \ldots X_{\mathrm{nj}} & \ldots X_{\mathrm{np}} \end{pmatrix} \tag{1.22}$$

En la matriz $\mathbf{X}$, el primer subíndice representa unidades de investigación o individuos, y el segundo subíndice corresponde a las variables o características, donde en general $n > p$.

Si deseamos discutir ambas columnas y filas de $\mathbf{X}$, las columnas son denotadas de la siguiente manera:

$$\mathrm{X} = (\mathrm{X}_{(1)}, \mathrm{X}_{(2)}, \ldots, \mathrm{X}_{(p)}) \tag{1.23}$$

Así, por ejemplo $\mathbf{X}_2$ es el vector p-dimensional de las variables medidas en la segunda unidad investigada, mientras $\mathbf{X}_{(2)}$ es el *n*-vector de observaciones en la segunda variable.

El vector muestral es definido como:

$$\overline{\mathbf{X}} = \begin{pmatrix} \overline{\mathbf{X}}_1 \\ \overline{\mathbf{X}}_2 \\ . \\ . \\ . \\ \overline{\mathbf{X}}_p \end{pmatrix} \tag{1.24}$$

Así el promedio de los n vectores produce el promedio de cada variable.

Podemos calcular $\overline{\mathbf{X}}$ directamente de $\mathbf{X}$:

$$\overline{\mathbf{X}} = \frac{1}{n}\mathbf{X}'\mathbf{j} \text{ donde } \mathbf{j} \text{ es un vector } n\text{x1 de unos } \quad \mathbf{j} = \begin{bmatrix} 1 \\ 1 \\ . \\ . \\ . \\ 1 \end{bmatrix} \tag{1.25}$$

La media poblacional o valor esperado del vector aleatorio $\mathbf{X}$ es definido como el vector de valores esperados de p variables,

$$E(\mathbf{X}) = E\begin{pmatrix} X_1 \\ X_2 \\ . \\ . \\ . \\ X_p \end{pmatrix} = \begin{pmatrix} E(X_1) \\ E(X_2) \\ . \\ . \\ . \\ E(X_p) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ . \\ . \\ . \\ \mu_p \end{pmatrix} = \boldsymbol{\mu}, \tag{1.26}$$

donde $E(X_j) = \mu_j$. Ya que $E(\bar{X}_j) = \mu_j$, entonces:

$$\mathrm{E}(\bar{X}) = \begin{pmatrix} \mu_1 \\ \mu_2 \\ . \\ . \\ . \\ \mu_p \end{pmatrix} = \boldsymbol{\mu} \tag{1.27}$$

lo cual significa que $\bar{\mathbf{X}}$ es un estimador insesgado de $\boldsymbol{\mu}$.

La Matriz Muestral de Varianzas y Covarianzas es simétrica:

$$\mathbf{S} = (s_{jk}) = \begin{pmatrix} s_{11} & s_{12} & \cdot & \cdot & \cdot & s_{1p} \\ s_{21} & s_{22} & \cdot & \cdot & \cdot & s_{2p} \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & & & & \cdot \\ s_{p1} & s_{p2} & \cdot & \cdot & \cdot & s_{pp} \end{pmatrix}, \; s_{ij} = s_{ji} \tag{1.28}$$

Y por tanto diagonalizable ortogonalmente

La matriz de varianzas y covarianzas de la población es definida como:

$$\sum = \mathrm{E}\left[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^{T} \right] \tag{1.29}$$

Donde resulta que $\sum$ es una matriz cuadrada simétrica por lo tanto, diagonalizable ortogonalmente,

$$\sum = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}$$

El valor σ_{ij} es la covarianza entre X_i y X_j. Para el caso en que i sea igual a j, σ_{ij} es la varianza de la i-ésima variable X_i, σ_i^2, esto es $\sigma_{ii} = \sigma_i^2$.

1.3.4 Matriz de Correlación

La matriz de correlación poblacional está definida como:

$$\mathbf{P}_\rho = (\rho_{jk}) = \begin{pmatrix} 1 & \rho_{12}\ldots\ldots\ldots\rho_{1p} \\ \rho_{21} & 1\ldots\ldots\ldots\ldots\rho_{2p} \\ & \\ & \\ \rho_{p1} & \rho_{p2}\ldots\ldots\ldots\ldots 1 \end{pmatrix} \tag{1.30}$$

donde $\rho_{jk} = \dfrac{\sigma_{jk}}{\sigma_j \sigma_k}$. El subíndice ρ en $\mathbf{P}_\rho$ es usado como recordatorio de que $\mathbf{P}$ es la versión mayúscula de ρ.

Si definimos $\mathbf{D}_\sigma = diag(\sigma_1, \sigma_2, ..., \sigma_p)$ será una matriz diagonal de la desviación de la población estándar análoga para D_S, luego:

$$\mathbf{P}_\rho = \mathbf{D}_\sigma^{-1} \sum \mathbf{D}_\sigma^{-1} \tag{1.31}$$

$$\sum = \mathbf{D}_\sigma \mathbf{P}_\rho \mathbf{D}_\sigma \tag{1.32}$$

Mientras $\overline{\mathbf{X}}$ y $\mathbf{S}$ son estimadores insesgados de μ y $\sum$, este no es el caso con $\mathbf{R}$.

Por (1.25) la correlación muestral entre las *j*-ésimas y *k*-ésimas variables está dada por:

$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj} s_{kk}}} = \frac{s_{jk}}{s_j s_k} \tag{1.33}$$

La matriz de correlación muestral es también una matriz de covarianzas definida como:

$$\mathbf{R} = (r_{jk}) = \begin{pmatrix} 1 & r_{12} \ldots\ldots r_{1p} \\ r_{21} & 1 \ldots\ldots r_{2p} \\ . & . & . \\ . & . & . \\ . & . & . \\ r_{p1} & r_{p2} \ldots\ldots 1 \end{pmatrix} \tag{1.34}$$

La cual es simétrica ya que $r_{jk} = r_{kj}$

$\mathbf{R}$ es una matriz de varianzas y covarianzas para datos estandarizados.

Para relacionar $\mathbf{R}$ (matriz de correlación muestral) y $\mathbf{S}$ (matriz de varianzas y covarianzas muestrales), se define la matriz diagonal:

$$\mathbf{D}_S = [diag(\mathbf{S})]^{1/2} = diag(s_1, s_2, \ldots, s_p) \tag{1.35}$$

Es posible probar que:

$$\mathbf{R} = \mathbf{D}_S^{-1}\ \mathbf{S}\ \mathbf{D}_S^{-1} \tag{1.36}$$

$$\mathbf{S} = \mathbf{D}_S\ \mathbf{R}\ \mathbf{D}_S \tag{1.37}$$

Si la matriz $\mathbf{X} = X_{ij}$ es estandarizada para $\mathbf{Z} = Z_{ij}$ donde $Z_{ij} = (X_{ij} - \overline{X}) / s_j$ luego la matriz de covarianza de las zetas es igual a la matriz de correlación de las equis:

$$\mathbf{S}_Z = \frac{1}{n-1}\mathbf{Z}'\mathbf{Z} = \mathbf{R} \tag{1.38}$$

1.4 La Pérdida de Datos en una Investigación

En el análisis de datos reales es habitual encontrarse con matrices que tienen sus datos incompletos ya sea por inconvenientes en la recolección de la información, por la negativa a cooperar, incapacidad de contestar de los entrevistados, ausencia temporal del entrevistado, pérdida de formularios, errores de digitación, etc.

Esta situación dificulta el tratamiento y análisis de los datos así como también la utilización de los procedimientos estadísticos estándares ya que estamos dentro de un problema de falta de datos, lo cual puede introducir sesgo en la estimación e incrementar o disminuir la varianza muestral debido a la reducción del tamaño de la muestra, y afectar a los valores de la matriz de varianzas y covarianzas y correlaciones.

En décadas anteriores era habitual, a la hora de analizar datos, ignorar aquellos registros que poseían datos faltantes. Por un lado las estimaciones pueden estar sesgadas, ya que la eliminación de estos registros, supone que la no-respuesta se distribuye de forma aleatoria

entre los distintos tipos de entrevistados. En el mejor de los casos, aquel en el que la no-respuesta se distribuye de forma aleatoria, estamos perdiendo una cantidad importante de información al eliminar los datos que estos individuos proporcionan a otras preguntas o proposiciones del cuestionario.

1.5 Métodos que emplean toda la información disponible

Los métodos que emplean toda la información disponible consisten en considerar para los sucesivos análisis únicamente la información completa de las variables investigadas. Existen dos métodos que se comentan a continuación:

1.5.1 Eliminación por Filas

El método de eliminación por filas consiste en emplear solamente los registros que tengan respuesta en todas las variables de estudio, es decir solo para los entrevistados que contesten todas las preguntas o cuyos datos fueron íntegramente digitados. Las ventajas de este método son su simplicidad pero se desperdicia información que se conoce. [6]

Para ilustrar este método, se tiene una matriz de datos cuyas columnas son muestras tomadas de tres poblaciones todas ellas

Poisson, independientes e idénticamente distribuidas con parámetro conocido $\lambda = 5$, $\mathbf{X} \in \mathrm{M}_{5x3}$, *i= 1,2,3,4,5* y *j= 1,2,3* y se supone que tiene el 13% de datos faltantes, es decir dos datos, los que recayeron en las variables X_2 y X_3 y son: el $X_{2,2}$=4 y $X_{4,3}$=7. Nótese que el 13% de datos faltantes en la matriz, constituye el 20% de datos faltantes en la columna que corresponde a X_2 y 20% de datos faltantes en la columna X_3. (Ver Tabla 1.1)

Tabla 1.1
Efectos de la imputación en el análisis de datos multivariados
Matriz de datos de variables aleatorias independientes con distribución Poisson
$\lambda = 5$
Tamaño de muestra n=5

X_1	X_2	X_3
8	4	6
4	**4**	5
3	5	6
1	7	**7**
6	5	2

Elaborado por: G. Cuenca

El vector de medias de los datos originales es:

$$\overline{\mathbf{X}} = \begin{pmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \overline{X}_3 \end{pmatrix} = \begin{pmatrix} 4.400 \\ 5.000 \\ 5.200 \end{pmatrix}$$

Como tenemos dos datos faltantes entonces se procede a prescindir de las dos filas que contienen los mismos y la matriz de datos ahora de datos resultante es (Ver Tabla 1.2)

Tabla 1.2
Efectos de la imputación en el análisis de datos multivariados
Matriz de datos de variables aleatorias independientes con distribución Poisson
$\lambda = 5$
Método de Eliminación por Filas
Tamaño de muestra n=5, 13% de datos faltantes en la matriz

X_1	X_2	X_3
8	4	6
3	5	6
6	5	2

Elaborado por: G. Cuenca

El vector de medias para las tres filas restantes es:

$$\overline{\mathbf{X}} = \begin{pmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \overline{X}_3 \end{pmatrix} = \begin{pmatrix} 5.667 \\ 4.667 \\ 4.667 \end{pmatrix}$$

Como era de esperarse el vector de medias de los datos originales y de los datos con filas eliminadas no coincide.

Ahora analicemos el efecto que causa en la matriz de varianzas y covarianzas, la eliminación de dos filas, con un tamaño de muestra *n*= 5.

CUADRO 1.2
Efectos de la Imputación en el Análisis de Datos Multivariados
Variables aleatorias independientes con distribución Poisson $\lambda = 5$
Método de eliminación por Filas
Tamaño de muestra n=5, 13% de datos faltantes en la matriz

Matriz de Varianzas y Covarianzas (Datos Originales)

	X_1	X_2	X_3
X_1	7.300		
X_2	-2.500	1.500	
X_3	-2.350	0.750	3.700

Matriz de Varianzas y Covarianzas (Dos Filas Eliminadas)

	X_1	X_2	X_3
X_1	6.333		
X_2	-1.167	0.333	
X_3	-0.667	-0.667	5.333

Elaborado por: G. Cuenca

Analizando el Cuadro 1.2 se puede apreciar que las covarianzas entre las variables disminuyeron, en la matriz con dos filas eliminadas, tal es el caso de la covarianza entre X_1 y X_3, la que disminuye de 0.750 a 0.667.

1.5.2 Eliminación por Pares

El método de eliminación por pares emplea todas las observaciones que tienen valores válidos para las variables de interés en cada momento, es decir usa todas las observaciones disponibles cuando calculamos $\overline{X}$ y todos los pares disponibles de valores en el cálculo de la matriz de correlación $\mathbf{R}$ y la matriz de covarianzas $\mathbf{S}$. [6]

Para ilustrar consideraremos la siguiente matriz de datos:

$$\mathbf{X}= \begin{bmatrix} X_{11} & X_{12} & X_{13} \\ X_{21} & - & X_{23} \\ X_{31} & X_{32} & X_{33} \\ X_{41} & X_{42} & - \\ X_{51} & X_{52} & X_{53} \end{bmatrix} \quad X \in M_{5x3}$$

Para obtener $\overline{X}_1$ se tienen cinco observaciones; para $\overline{X}_2$ y $\overline{X}_3$ se tienen cuatro observaciones disponibles. Para s_{12} y s_{13}, hay cuatro pares de observaciones; para s_{23}, solo tres pares están disponibles.

A simple vista, esta forma de aproximarse al problema es atractiva porque usa toda la información disponible, pero el procedimiento generalmente no se recomienda ya que para el estudio de la correlación o covarianza entre las distintas variables el número de elementos variará según el número de registros que no tengan valores faltantes en dichas variables.

Se ilustra este método utilizando los mismos datos del ejemplo anterior, es decir, una matriz de datos cuyas columnas son muestras tomadas de tres poblaciones todas ellas Poisson, independientes e idénticamente distribuidas con parámetro conocido $\lambda = 5$, $\mathbf{X} \in \mathrm{M}_{5\mathrm{x}3}$, *i= 1,2,3,4,5* y *j= 1,2,3* y se supone que tiene el 13% de datos faltantes, dos datos, los que recayeron en las variables X_2 y X_3 y son: el $X_{2,2}$=4 y $X_{4,3}$=7.

Tabla 1.3
Efectos de la imputación en el análisis de datos multivariados
Matriz de datos de variables aleatorias independientes con distribución Poisson
$\lambda = 5$
Método de Eliminación por Pares
Tamaño de muestra n=5, 13% de datos faltantes en la matriz

X_1	X_2	X_3
8	4	6
4	**4**	5
3	5	6
1	7	**7**
6	5	2

Elaborado por: G. Cuenca

Entonces para obtener $\overline{X}_1$ se tienen cinco observaciones, en cambio para $\overline{X}_2$ y $\overline{X}_3$ se tienen solo cuatro observaciones. Para s_{12}

y s_{13}, hay cuatro pares de observaciones; para s_{23}, solo tres pares están disponibles y estos son:

Para s_{12} los pares de observaciones disponibles son: (8,4),(3,5),(1,7) y (6,5), ya que aquí se elimina un par de observaciones. (Ver Cuadro 1.3)

CUADRO 1.3

Efectos de la Imputación en el Análisis de Datos Multivariados

Variables aleatorias independientes con distribución Poisson $\lambda = 5$

Método de eliminación por Pares

Tamaño de muestra n=5, 13% de datos faltantes en la matriz

Pares de observaciones disponibles para s_{12}

X_1	X_2
8	4
3	5
1	7
6	5

Matriz de Varianzas y Covarianzas para s_{12}

Variables	X_1	X_2
X_1	9.670	
X_2	-3.500	1.580

Elaborado por: G. Cuenca

Para s_{13} los pares de observaciones disponibles son: (8,6),(4,5),(3,6) y (6,2).

CUADRO 1.4

Efectos de la Imputación en el Análisis de Datos Multivariados

Variables aleatorias independientes con distribución Poisson $\lambda = 5$

Método de eliminación por Pares

Tamaño de muestra n=5, 13% de datos faltantes en la matriz

Pares de observaciones disponibles para s_{13}

X_1	X_3
8	6
4	5
3	6
6	2

Matriz de Varianzas y Covarianzas para s_{13}

Variables	X_1	X_2
X_1	4.920	
X_2	-0.580	3.580

Elaborado por: G. Cuenca

Para s_{23} los pares de observaciones disponibles son: (4,6),(5,6) y (5,2)

CUADRO 1.5

Efectos de la Imputación en el Análisis de Datos Multivariados

Variables aleatorias independientes con distribución Poisson $\lambda = 5$

Método de eliminación por Pares

Tamaño de muestra n=5, 13% de datos faltantes en la matriz

Pares de observaciones disponibles para s_{23}

X_1	X_3
4	6
5	6
5	2

Matriz de Varianzas y Covarianzas para s_{23}

Variables	X_1	X_3
X_1	0.330	
X_3	-0.670	5.330

Elaborado por: G. Cuenca

Donde la matriz de correlaciones es de la forma:

Tabla 1.4

Efectos de la imputación en el análisis de datos multivariados

Variables aleatorias independientes con distribución Poisson $\lambda = 5$

Método de eliminación por Pares

Tamaño de muestra n=5, 13% de datos faltantes en la matriz

Matriz de Varianzas y Covarianzas

Variables	X_1	X_2	X_3
X_1	1		
X_2	-3.500	1	
X_3	-0.580	-0.670	1

Elaborado por: G. Cuenca

Este método tiene la desventaja de no poder asegurar que la matriz de correlaciones sea definida positiva, condición indispensable para invertir la matriz de correlaciones. Esta situación es debido a que se emplean distintas submuestras para el cálculo de las distintas correlaciones.