CAPÍTULO IV

TEORÍA ESTADÍSTICA

4.1 Matriz de Datos

Una matriz de datos es una matriz **X** compuesta por n filas y p columnas, el número de n filas corresponde al total de unidades investigadas u observadas y las p columnas al número de variables (características de interés) que se investigan; y se representa de la siguiente forma:

$$X = \begin{pmatrix} X_{11} & X_{12} & . & . & . & X_{1p} \\ X_{21} & X_{22} & . & . & . & X_{2p} \\ . & . & . & & & . \\ . & . & & . & & . \\ . & . & & & . & . \\ X_{n1} & X_{n2} & . & . & . & X_{np} \end{pmatrix}$$

Por otra parte si se toma en cuenta sólo las p variables de interés y no tomamos en cuenta el número de observaciones realizadas entonces se tiene lo que se denomina un vector aleatorio; y está compuesto por las p

variables o características de interés, y se representa de la siguiente forma:

$$\mathbf{X^T} = [\, X_1, X_2, \ldots, X_p]$$

4.2 Vector de Medias

Se define al vector de medias & (una matriz de 1 columna y p filas), al vector que contiene las medias o valores esperados de las variables que se investigan.

$$\mu = \mathbf{E(x)} = \begin{bmatrix} \mathbf{E(X_1)} \\ \mathbf{E(X_2)} \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{E(X_p)} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \cdot \\ \cdot \\ \cdot \\ \mu_p \end{bmatrix}$$

4.3 Covarianza o Covariancia

La covarianza es una medida de la dispersión conjunta de un par de variables aleatorias, mide al grado de la desviación de dos variables aleatorias **X** e **Y** de sus respectivas medias, si el valor de la covarianza es alto, entonces se considerará que las variables anteriormente mencionadas son significativamente dependientes, a diferencia de la varianza, la cual siempre deberá tomar valores positivos, si el valor de la covarianza es negativo se infiere que las variables son inversamente dependientes, es decir, a medida que la primera variable incrementa su valor, la segunda disminuye, es decir ; si hay una alta probabilidad de

que los valores de **X** altos vayan con valores de **Y** bajos y viceversa, la covarianza será negativa, por otro lado si hay una alta probabilidad de que los valores de **X** altos vayan con valores de **Y** altos y que valores de **X** bajos vayan con valores de **Y** bajos, entonces la covarianza será positiva. Es en este sentido que la covarianza mide la relación, o asociación, entre los valores de **X** y **Y.** Matemáticamente la covarianza se define de la siguiente manera:

$$\mathbf{Cov(X,Y)=}\ E[(\mathbf{x}-\&_x)\text{-}(\mathbf{y}-\&_y)]$$

Cabe recalcar que cuando se tiene una muestra, el valor de la covarianza se estima mediante el estimador insesgado S_{xy}; el cual se denota de la siguiente manera:

$$\hat{\sigma}_{xy} = S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}).(y_i - \bar{y})$$

$$\hat{\sigma}^2 = S_{xx} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - x)^2$$

4.4 Matriz de Varianzas y Covarianzas

La matriz Σ de Varianzas y Covarianzas es aquella formada en la diagonal principal son las varianzas de las variables y en la posición (i,j) las covarianzas entre la i-ésima y la j-ésima variable, cabe recalcar que i e j son los sub-índices de las columnas correspondientes a las variables en el vector aleatorio $\mathbf{X^T}$, Σ es simétrica con p filas y p columnas.

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdot & \cdot & \cdot & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdot & \cdot & \cdot & \sigma_{2p} \\ \cdot & \cdot & \cdot & & & \cdot \\ \cdot & \cdot & & \cdot & & \cdot \\ \cdot & \cdot & & & \cdot & \cdot \\ \sigma_{p1} & \sigma_{p2} & \cdot & \cdot & \cdot & \sigma_{pp} \end{bmatrix}$$

Por lo general la matriz Σ representa a la matriz de covarianzas, para efecto de esta investigación precisamente Σ es la matriz de varianzas y covarianzas o matriz de covarianzas.

4.5 Coeficiente de Correlación

El coeficiente de correlación es una medida de asociación lineal entre las variables X e Y. Se representa por ρ_{xy}:

$$\rho_{xy} = \frac{Cov(X,Y)}{\sigma_x \sigma_y} \qquad \text{donde} \quad Cov(X,Y) = \tau_{xy}$$

σ_x, σ_y son las desviaciones típicas de las variables X e Y respectivamente, y σ_{xy} es la covarianza muestral de X e Y, que se define como la media de los productos de las desviaciones correspondientes de X e Y y de sus medias muestrales.

$$\sigma_{xy} = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

Propiedades

- $\rho_{x,y}$ está siempre comprendido entre -1 y 1.
- Si $\rho_{x,y}$ = 1 ó $\rho_{x,y}$ = -1 entonces los puntos de la muestra están situados en línea recta (correlación lineal perfecta).
- Si $\rho_{x,y}$ está próximo a 1 ó a -1, habrá una asociación lineal fuerte entre ambas variables.
- Si $\rho_{x,y}$ es cercano a 0, habrá una asociación lineal muy débil.
- $\rho_{x,y}$ no varía cuando en las variables se realiza un cambio de escala o de origen. Esto demuestra que $\rho_{x,y}$ no tiene dimensión.

Dos consideraciones sobre el coeficiente de correlación.

1. Se trata de una medida matemática que luego hay que interpretar. Aunque un alto grado de correlación indique buena aproximación a un modelo matemático lineal, su interpretación puede no tener ningún sentido, ya que en muchos casos puede existir una alta correlación entre dos variables, pero a su vez ambas variables pueden estar claramente disociadas entre sí.

2. Aunque el grado de correlación sea cercano a cero (pobre aproximación al modelo lineal) eso no significa que no haya relación entre las dos variables. Puede ser que dicha relación sea no lineal.

4.6 Matriz de Correlaciones

En esta matriz podemos ordenar los diferentes coeficientes de correlación de cada variable con el resto y consigo misma, obteniendo una matriz con cada elemento igual a :

$$\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i . \sigma_j}$$

El resultado es una matriz simétrica, con la diagonal principal igual a 1 y se denota como ρ

$$\rho = \begin{bmatrix} 1 & \rho_{12} & . & . & . & \rho_{1p} \\ \rho_{21} & 1 & . & . & . & \rho_{2p} \\ . & . & . & & & . \\ . & . & & . & & . \\ . & . & & & . & . \\ \rho_{p1} & \rho_{p2} & . & . & . & 1 \end{bmatrix}$$

4.7 Análisis de Componentes Principales

El Análisis de Componentes Principales es una técnica estadística multivariada en la que se estudian p variables de interés, las cuales constituyen un vector aleatorio, cuyas componentes son variables aleatorias discretas o continuas, esto es:

$$\underline{X} = (x_1, x_2, x_3, \ldots x_p)$$

El análisis de componentes principales comprende un procedimiento matemático que transforma un conjunto de variables correlacionadas de respuestas en un conjunto menor de variables no correlacionadas.

Razones para usar el Análisis de Componentes Principales

1. **Cribado de los datos**

   Los análisis de seguimiento sobre las componentes principales son útiles para comprobar hipótesis que el investigador podrá establecer acerca de un conjunto de datos multivariados y para identificar y localizar los datos *outliers.*

2. **Agrupación**

   Útil siempre que desee agrupar las unidades experimentales en subgrupos semejantes. Ayuda a verificar los resultados de los programas de agrupación.

3. **Análisis Discriminante**

   Usando las nuevas variables (componentes principales) como variables de entrada a un programa de análisis discriminante.

4. **Regresión**

   Puede ayudar a determinar si ocurre multicolinealidad entre las variables predictoras.

   Para casi todas las situaciones de análisis de datos se puede recomendar el Análisis de Componentes Principales como un primer paso, este se debe realizar sobre un conjunto de datos, antes de realizar cualquier clase de análisis discriminante.

5. **Sinterización de la información**

   El análisis de componentes principales ayuda a la sinterización de los datos, o reducción de la dimensión (número de variables). Es decir, ante un banco de datos con muchas variables, el objetivo será

reducirlas a un menor número perdiendo la menor cantidad de información posible.

Objetivos del Análisis de Componentes Principales

1. Reducir la dimensionalidad del conjunto de datos.
2. Identificar nuevas variables subyacentes.

Las *nuevas variables* (componentes principales) presentan un orden decreciente de importancia, el cual es el siguiente:

1. No están correlacionadas.
2. La primera componente principal explica el mayor porcentaje de variabilidad presente en los datos.
3. Cada componente subsiguiente toma en cuenta el % de variabilidad restante como sea posible.

En fin, los nuevos componentes principales o factores serán una combinación lineal de las variables originales, y además serán independientes entre sí.

4.7.1 Análisis en R^p

Mediante este análisis podemos describir una matriz A (matriz de datos) de variables continuas, la cual tiene la siguiente forma:

$$A = \begin{bmatrix} a_{11} & a_{12} & . & . & . & a_{1p} \\ a_{21} & a_{22} & . & . & . & a_{2p} \\ . & . & . & & & . \\ . & . & & . & & . \\ . & . & & & . & . \\ a_{p1} & a_{2p} & . & . & . & a_{mp} \end{bmatrix}$$

A ϵ $M_{mxp,}$ es decir m filas (individuos) y columnas (p variables). Los elementos de A pueden ser heterogéneos, tanto en su media como en su desviación.

En muchas ocasiones las variables toman valores muy altos y obviamente tienen un peso muy importante, por ello se realiza una transformación que consiste en centrar los datos, esto es restar la media de cada uno de los elementos, es decir:

$$X_{ij} = a_{ij} - \bar{a}_j$$

Donde $\bar{a}_j = \sum_i \frac{a_{ij}}{n}$ es la media de la variable j.

De esta manera se elimina la influencia del nivel general de las variables.

Claro que si las dispersiones de las variables son muy diferentes, se hará necesario dividir las variables para su desviación correspondiente, así las variables se encontrarán estandarizadas.

Después del procedimiento anterior, tendremos una matriz X obtenida a partir de la matriz A con la siguiente forma:

$$X = \begin{bmatrix} \frac{a_{11} - \bar{a}_1}{s_1\sqrt{n}} & \frac{a_{12} - \bar{a}_2}{s_2\sqrt{n}} & \cdots & \frac{a_{1p} - \bar{a}_p}{s_p\sqrt{n}} \\ \vdots & \vdots & & \vdots \\ \frac{a_{m1} - \bar{a}_1}{s_1\sqrt{n}} & \frac{a_{m1} - \bar{a}_2}{s_2\sqrt{n}} & \cdots & \frac{a_{mp} - \bar{a}_p}{s_p\sqrt{n}} \end{bmatrix}$$

Donde s_i=1…p es la desviación de cada una de las variables y el término $\frac{1}{\sqrt{n}}$ se introduce en la transformación con el objetivo de que el producto de las matrices X^TX coincida con la matriz de correlación $\sum$.

El análisis consiste en obtener los vectores propios de la matriz $\sum = X^TX$. Las proyecciones de los individuos sobre estos vectores propios son los componentes principales, los cuales se obtienen de la siguiente manera:

Supongamos que los valores propios característicos de $\sum$ son:

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_p$$

Con sus respectivos vectores propios, es decir:

$$\mu_1 = \begin{bmatrix} \beta_{11} \\ \beta_{21} \\ . \\ . \\ . \\ \beta_{p1} \end{bmatrix}, \quad \mu_2 = \begin{bmatrix} \beta_{12} \\ \beta_{22} \\ . \\ . \\ . \\ \beta_{p2} \end{bmatrix}, \ldots \quad \mu_p = \begin{bmatrix} \beta_{1p} \\ \beta_{2p} \\ . \\ . \\ . \\ \beta_{pp} \end{bmatrix}$$

Luego, se definen q variables no observadas Y_1, Y_2, Y_3,…Y_q como una combinación lineal de X_1, X_2, X_3,…X_q, entonces:

$$Y_1 = \beta_{11}X_1 + \beta_{21}X_2 + \beta_{31}X_3 + \ldots + \beta_{p1}X_p$$

$$Y_2 = \beta_{12}X_1 + \beta_{22}X_2 + \beta_{32}X_3 + \ldots + \beta_{p2}X_p$$

$$Y_3 = \beta_{13}X_1 + \beta_{23}X_2 + \beta_{33}X_3 + \ldots + \beta_{p3}X_p$$

.
.
.

$$Y_q = \beta_{1q}X_1 + \beta_{2q}X_2 + \beta_{3q}X_3 + \ldots + \beta_{pq}X_p$$

La cual también puede ser expresada de la siguiente forma:

$$Y_i = \beta_{1i}X_1 + \beta_{2i}X_2 + \beta_{3i}X_3 + \ldots + \beta_{pi}X_p = \mu_i^T \mu$$

donde i=1,2,3,…,p

Y la varianza de Y_p queda denotada por VAR(Y_p)

VAR(Y_p)= VAR ($u_i^T X$)= $u_i^T \sum u_i$ donde i=1,2,3,…,p

Y la covarianza entre Y_i y Y_k , donde i,k=1,2,3,…,p ; es:

$$\text{Cov}\ (Y_i, Y_k) = u_i^T \sum u_k$$

Por lo tanto, las Componentes Principales de X son aquellas combinaciones lineales construidas de la forma anteriormente descrita, que son *no correlacionadas* y cuyas varianzas son tan grandes como sea posible.

La **primera componente principal** es la primera combinación lineal con mayor varianza, es decir la que maximiza VAR(Y_i)= VAR ($u_i^T X$)= $u_i^T \sum u_i$, dicha varianza puede incrementarse en su a_1 multiplicando por alguna constante, para eliminar estas indeterminaciones, se utilizan solamente vectores unitarios.

Después la primera componente principal se define como la combinación lineal que maximiza VAR(Y_1) sujeta a $u_1^T u_1$=1.

La **segunda componente principal** se define como la combinación lineal que maximiza VAR(Y_2) sujeta a $u_2^T u_2=1$; y además $Cov(u_1^T X, u_2^T X)=0$

Siguiendo esta frecuencia y patrón, tenemos que la **k-ésima componente principal** se define como la combinación lineal que maximiza VAR(Y_k) sujeta a $u_k^T u_k=1$; y además $Cov(u_i^T X, u_i^T X)=0$ para k<i.

Definición de Traza de una Matriz

La traza de una matriz A, denotada por tr(A), está definida como la suma de los elementos de su diagonal principal.

Traza de La Matriz de Varianzas y Covarianzas

Para el caso de la matriz de varianzas y covarianzas $\sum$, en donde los elementos de la diagonal principal son las varianzas de cada una de las variables, la traza quedará establecida de la siguiente manera:

$$tr(\sum)= \sigma_{11} + \sigma_{22} + \sigma_{33} + \ldots + \sigma_{pp} = \lambda_1 + \lambda_2 + \ldots + \lambda_p$$

La tr(∑) nos ayudara a establecer, la proporción de variabilidad total en las variables originales que es explicada por la k-ésima componente principal:

Proporción de la variabilidad total en las variables originales que es explicada por la k-ésima componente principal $= \frac{\lambda}{tr(\sum}$ k=1,2,3,…p

4.7.2 Determinación del número de componentes principales

Si varios de los valores propios de $\hat{\sum}$ son cero o cercanos a cero, entonces la dimensionalidad real de los datos es la del número de valores propios diferentes de cero.

Consideremos a ***d*** como la dimensionalidad real de los datos.

Existen varios métodos para la determinación del número de componentes principales, a continuación se detallan dos de los métodos más utilizados:

Método 1

Consiste en tomar σ100% de la variabilidad total en las variables originales.

Sea $(\lambda_1 + \lambda_2 + .. + \lambda_k)/tr(\sum)$

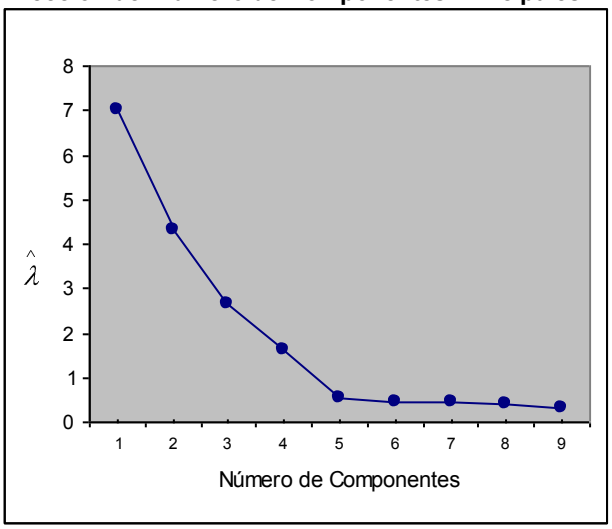
Donde k=1,2,…,p y $tr(\sum)= \lambda_1 + \lambda_2 + .. + \lambda_p$

Entonces ***d*** será el menor de los valores de k en el que, por primera vez, se sobrepasa σ.

Método 2

Como la varianza explicada por cada eje sucesivo debe ser decreciente, se puede representar el histograma de los valores propios $\left\{\hat{\lambda}_k\right\}$, con los números de los ejes en ordenadas y los porcentajes de inercia explicadas en las abscisas, obteniendo las parejas $\left\{1, \hat{\lambda}_1\right\}$, $\left\{2, \hat{\lambda}_2\right\}$, $\left\{3, \hat{\lambda}_3\right\}$, … , $\left\{p, \hat{\lambda}_p\right\}$, se pueden eliminar los ejes cuyo número de orden es posterior al “codo” que se produce en la curva. Por ejemplo en el siguiente gráfico, sólo se tomarían los cuatro primeros ejes.

Gráfico 4.1
Elección del Número de Componentes Principales



4.7.3 Interpretación de los factores

Las variables iniciales pueden tener redundancias y estar midiendo en parte la misma característica. Como se mencionó el factor es un agrupamiento de estas variables y se interpreta a partir de su correlación, que es la proyección de la variable sobre el factor.

Si una variable está muy correlacionada con un factor, tendrá una coordenada muy alta próxima a ± .

- Si $F_\alpha(1) = +$ se puede interpretar al factor como una clasificación de los individuos a lo largo de él en orden de valores crecientes de la variable j.

- Si $F_\alpha(j) = -$ los individuos están clasificados sobre el eje en orden decreciente de los valores de j.
- Si $F_\alpha(j) = 0$ entonces no existe relación entre el factor y la variable.

Mientras mayor sea el valor absoluto de $F_\alpha(j)$ más alta es la relación entre j y el factor α.

4.8 Análisis de Tablas de Contingencia

Dada dos variables X y Y esta técnica estadística multivariada utiliza arreglos matriciales *(rxc)*, donde *r* filas indican la cantidad de niveles de la variable X y las *c* columnas indican el número de niveles de la variable Y o sus posibles resultados. El objetivo es verificar la independencia entre las dos variables X y Y, a través del siguiente contraste de hipótesis:

H_o: Las variables X y Y son independientes.

Vs.

H_1: Las variables X y Y no son independientes.

Estableciendo la frecuencia observada de la i-ésima fila y la j-ésima columna (x_{ij}); los totales por fila (X_i) y columna (Y_j) y la suma de todas

las frecuencias de las celdas (n), la Tabla de Contingencia (rxc) se la construye como se presenta a continuación:

Gráfico 4.2
Tabla de Contingencia entre las variables X y Y

		Variable Y						***Total***
		1	2	…	j	…	c	
Variable X	1	x_{11}	x_{12}		x_{1j}		x_{1c}	$X_{1.}$
	2	x_{21}	x_{22}		x_{2j}		x_{2c}	$X_{2.}$
	⋮							⋮
	i	x_{i1}	x_{12}		x_{ij}		x_{ic}	$X_{i.}$
	⋮							⋮
	r	x_{r1}	x_{r2}		x_{rj}		x_{rc}	$X_{r.}$
Total		$Y_{.1}$	$Y_{.2}$	…	$Y_{.j}$	…	$Y_{.c}$	***n***

Donde: $X_{i.} = \sum_{j=1}^{c} x_{ij}$, $Y_{.j} = \sum_{i=1}^{r} x_{ij}$ y $E_{ij} = \frac{X_{i.} Y_{.j}}{n}$; i=1,2,…r; j=1,2,…c

Entonces se define el estadístico de prueba para esta técnica multivariada:

$$\chi^{2} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(x_{ij} - E_{ij})^2}{E_{ij}}$$

Se puede probar que $\chi^{2} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(x_{ij} - E_{ij})^2}{E_{ij}}$ puede ser modelada como una distribución ji-cuadrado (χ^{2}) con (r-1)(c-1) grados de libertad. Con $(1 - \alpha)100\%$ de confianza, se rechaza H_o a favor de H_1 si: $\chi^{2} > \chi^{2}_{(r-1)(c-1)}$.