Capítulo 2

2. MARCO TEÓRICO

A través de este capítulo, podremos obtener información respecto de las técnicas usadas para la organización, representación gráfica y análisis de un grupo considerado de variables aleatorias.

2.1. **TÉCNICAS MULTIVARIADAS**

Son un conjunto de técnicas que sirven como herramienta para el tratamiento de dos o más variables aleatorias de manera simultánea, siendo estas discretas o continuas, y así realizar análisis del comportamiento de las mismas en su conjunto.

Estas técnicas se refieren a métodos descriptivos debido a que no refieren el uso de ninguna clase de hipótesis probabilística. Por otro lado, estas nos dan una orientación para la interpretación de

ciertos factores que influyen sobre un conjunto determinado de variables. Además, estas técnicas nos permiten tratar toda la información disponible de manera simultánea y efectiva.

El uso de éstas técnicas supone un progreso con relación a los métodos más clásicos en los que sólo hubiéramos podido calcular dos características o calcular un coeficiente de correlación que proporciona una cifra global.

2.1.1. **MATRIZ DE DATOS**

Es un arreglo rectangular al que se lo denominará **X** y que consistirá **n** filas y **p** columnas donde **n** será el número de unidades de investigación o individuos , y **p** >0 será el número de variables aleatorias que serán objeto de nuestra investigación.

Así tendremos la siguiente representación:

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

Donde:

x_{ij} = medida de la j-ésima variable del i-ésimo individuo.

Entonces, el arreglo **X** (matriz de datos) contiene todas las observaciones de todas las variables que serán parte de nuestro estudio.

2.1.2. **VECTOR DE MEDIAS**

Es una representación matricial de **p** filas por una columna en donde se exponen los valores de las medias aritméticas de cada una de las variables aleatorias.

Tenemos la representación de este Vector de la siguiente manera:

$$\overline{X} = \begin{bmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \vdots \\ \overline{X}_p \end{bmatrix}$$

Donde:

$\overline{X}_i =$ es la media de la i - esima variable

$$\overline{X}_i = \frac{1}{n} \sum_{j=1}^{n} X_{j\,i} \qquad i = 1,2,\ldots..,\ p$$

Ésta es la representación matemática de la media para la i-ésima variable que será calculada tomando las n mediciones de cada una de las p variables.

2.1.3. **MATRIZ DE VARIANZAS Y COVARIANZAS**

Este es un arreglo de **p** filas y **p** columnas, es decir es una matriz cuadrada que tiene la propiedad de ser simétrica. Se nota de la siguiente manera:

$$S_{i\,j} = \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1p} \\ S_{21} & S_{22} & \cdots & S_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{p1} & S_{p2} & \cdots & S_{pp} \end{bmatrix}$$

Donde:

$$S_{ij} = \begin{cases} \text{covarianza entre la variable } i \text{ } j, \text{ cuando } i \neq j \\ \text{varianza de la variable } i \text{ , cuando } i = j \end{cases}$$

Y además :

$$S_{ij} = S_{ji}$$

La fórmula matemática para el cálculo de estos valores es la siguiente:

$$S_{ij} = \frac{1}{n} \sum_{k=1}^{n} (X_{ki} - \overline{X}_i)(X_{kj} - \overline{X}_j)$$

En este modo de organización de datos, las varianzas de cada una de las variables de estudio se encontrarán en la diagonal de dicha matriz, y las covarianzas en la matriz triangular superior o inferior, tomando en cuenta que es una matriz simétrica.

2.1.4. MATRIZ DE CORRELACIONES

Este también es un arreglo matricial que consta de **p** filas y **p** columnas, de acuerdo al número de variables. También es una

matriz cuadrada y simétrica que tiene por característica principal el hecho de que su diagonal principal se compone de unos (1), y en la matriz triangular superior o inferior se encuentran los coeficientes de correlación entre las variables en estudio.

Se nota de la siguiente manera:

$$R_{ij} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}$$

Donde :

$r_{ij} =$ es el coeficiente de correlación entre la variable i y la j

Y la representación matemática de este cálculo está dada por la siguiente ecuación:

$$r_{ij} = \frac{S_{ij}}{\sqrt{S_{ii}}\sqrt{S_{jj}}} = \frac{\sum_{k=1}^{n}(X_{ki} - \overline{X}_i)(X_{kj} - \overline{X}_j)}{\sqrt{\sum_{k=1}^{n}(X_{ki} - \overline{X}_i)^2}\sqrt{\sum_{k=1}^{n}(X_{kj} - \overline{X}_j)^2}}$$

Donde **i =** 1,2,....., p y **j =** 1,2,....., p

Cabe indicar que la correlación muestral **r** debe estar entre –1 y +1.

2.1.5. REPRESENTACIONES GRÁFICAS.

Los gráficos son muy importantes debido a que dan una nueva perspectiva en el análisis de datos. Sin embargo, es imposible graficar en forma simultánea todas las mediciones de las variables de estudio. Se pueden hacer gráficos de variables estadísticas en forma individual, pero también pueden ser de mucha utilidad y brindar buena información los gráficos de pares de variables.

Mediante los programas sofisticados de los computadores podemos darnos el lujo de examinar visualmente los datos en una, dos o tres dimensiones con relativa facilidad.

2.1.5.1. DIAGRAMAS DE DISPERSIÓN

Es una representación gráfica en dos dimensiones cuyo eje horizontal **X** corresponde al de una variable, y el eje vertical **Y** le

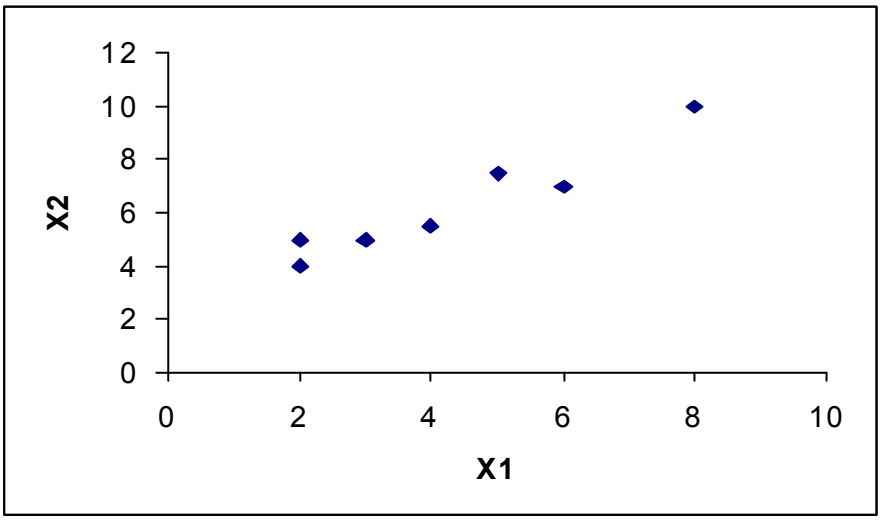
corresponde al de la otra variable con la cual se desea hacer el análisis.

Cada punto del gráfico corresponde a un par ordenado formado por los valores de las dos variables (Variable 1, Variable 2), y los valores que toman estas variables corresponden a las de un individuo diferente en cada punto.

Ponemos a disposición un gráfico como el que se muestra a continuación:

GRÁFICO 2.1.

EJEMPLO DE DIAGRAMA DE DISPERSIÓN



Este tipo de gráfico posee diferentes tipos de información. Según el ejemplo que ponemos a disposición podemos concluir que estas

variables tienen una tendencia lineal positiva, lo que implica una relación directa entre X1 y X2.

2.1.5.2. ESTRELLAS

Esta es una representación gráfica que trata de mostrar una perspectiva diferente de todas las variables objeto de estudio.
Para éste tipo de gráfico no existen restricciones por el número de variables como en el diagrama de dispersión en la que sólo se pueden representar gráficamente hasta tres variables.

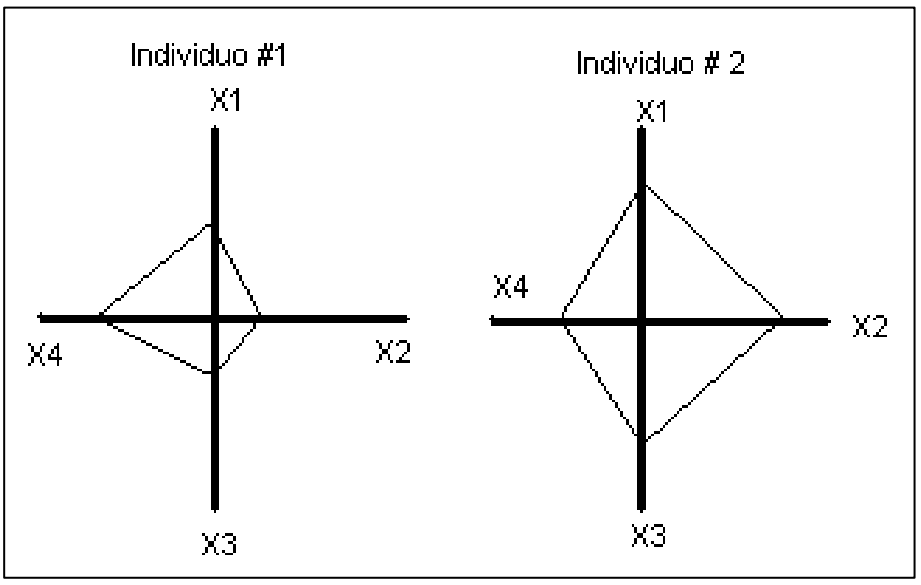
En el gráfico de estrellas se pueden mostrar todas las variables posibles que son de nuestro interés, en donde cada punta de esta estrella representa una variable estadística, y cada gráfico resultante, representa un individuo, es decir si tenemos n=20 tendremos como resultado 20 gráficas de estrellas.
El procedimiento para realizar un gráfico de éste tipo se muestra a continuación:

I. Se estandarizan los datos de cada una de las variables, de acuerdo a su respectiva media y desviación estándar.
II. El mínimo valor observado de estos datos anteriormente estandarizados corresponde a cero.

III. Se realiza una resta en valor absoluto de los datos estandarizados menos el mínimo valor encontrado que correspondía a cero, por cada individuo.

IV. Con estos resultados procedemos a realizar la gráfica correspondiente. Se muestra un ejemplo:

GRÁFICO 2.2.
EJEMPLO DE GRÁFICO DE ESTRELLAS



En el ejemplo expuesto, tratamos un caso hipotético de p=4 y n=2, es decir cuatro variables y dos individuos. Mediante este gráfico se pueden hacer comparaciones entre individuos, de acuerdo a los valores que toman en cada una de las variables.

2.1.6. ANÁLISIS DE COMPONENTES PRINCIPALES (ACP).

El método de Análisis de Componentes Principales, también conocido como ACP, es utilizado para describir una matriz **X** (matriz de datos), es decir una matriz de variables continuas del tipo individuos x variables.

Es también un objetivo del ACP, reducir el número de variables, además de buscar correlaciones entre grupos de variables llamados **factores**, y no solamente entre variables como lo muestra la matriz de correlaciones expuesta en la sección **2.1.4.** y en la que se indica la relación que existe entre una y otra variable, y no entre grupos de variables, como por ejemplo la correlación que pueda existir entre factores como nivel socioeconómico, nivel de vida, nivel educacional, entre otros.

La matriz de datos **X** puede ser muy disimétrica, y las variables muy heterogéneas, tanto en media como en desviación, por ejemplo, una variable puede medir los activos de una empresa, y otra los tipos de rendimiento, con lo cual las diferencias de medias serían enormes.

Por esta razón, antes de realizar el análisis general, el ACP realiza una transformación de la matriz, es decir, trabaja con variables estandarizadas.

2.1.6.1. ESQUEMA PARA EL DESARROLLO DE UN ACP

a) Estandarización de la Matriz de Datos

Para evitar que las variables que toman valores muy altos, tengan un peso muy importante en la determinación de los ejes de componentes principales, se realiza una transformación que consiste en centrar los datos de la siguiente forma:

$$x'_{ij} = x_{ij} - \overline{X}_j$$

Donde :

$$\overline{X}_j = \text{media de la variable j}$$

De esta manera se elimina la influencia del nivel general de las variables. Equivale a una traslación del origen al centro de gravedad de la nube de datos.

Si además, las dispersiones de las variables son muy diferentes, se hará necesario realizar otra transformación en los datos de partida. Se tipifican dividiendo para su desviación.

$$x'_{ij} = \frac{X_{ij} - X_j}{S_j}$$

Donde:

S_j = es la desviación estandar de la variable j

Una vez realizada la transformación de la tabla mediante la estandarización, se procede al análisis usando la nueva tabla.

b) Obtención de la Matriz de Varianzas y Covarianzas

Para el efecto, se toma la matriz de datos estandarizada y se procede según el esquema utilizado en la sección **2.1.3**.

En esta sección se especifican los pasos para la obtención de la matriz en cuestión. También se puede utilizar un esquema matricial para obtener dicha matriz, mediante nuestro grupo de datos.

$$S = E\left[\begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \\ \vdots \\ x_p - \mu_p \end{pmatrix}\begin{pmatrix} x_1 - \mu_1 & x_2 - \mu_2 & \cdots & x_p - \mu_p \end{pmatrix}\right]$$

c) Valores y Vectores Propios

Este proceso para hallar los valores y los vectores propios desde la Matriz de Varianzas y Covarianzas se lo conoce como diagonalización de la matriz en mención.

Se define de la siguiente manera:

- Sea **S** una matriz cuadrada de orden p . El número **λ** es llamado valor propio de S si existe un vector X que pertenece al conjunto de los vectores que se encuentran en el espacio n-dimensional, no nulo tal que SX = λX en tal caso, al vector X se lo llama vector propio correspondiente o asociado al valor propio λ.

- Sea **S** una matriz pxp, el número λ es un valor propio de S si y solamente si:

$$|S-\lambda I| = 0$$

Expuesto en forma matricial obtenemos lo siguiente:

$$\det\left[\begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1p} \\ S_{21} & S_{22} & \cdots & S_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{p1} & S_{p2} & \cdots & S_{pp} \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 1 \end{bmatrix}\right] = 0$$

Desarrollando el determinante de la matriz resultante, obtendremos los valores propios con sus correspondientes vectores propios.

El cálculo de la determinante da como resultado un polinomio llamado polinomio característico de la matriz S.

Para el desarrollo de un ACP, los vectores característicos obtenidos anteriormente deben tener la característica de ser perpendiculares entre sí. En caso contrario deberán pasar por un proceso de ortonormalización.

Cada componente resultante de un ACP será una combinación lineal de cada uno de los vectores característicos ortonormalizados.

2.1.6.2. ELECCIÓN DEL NÚMERO DE COMPONENTES

Existen diferentes métodos para elegir el número de Componentes que se han de retener, aunque ninguno de ellos es determinante.

a) Se puede fijar el porcentaje mínimo de inercia que se quiere conservar y retener el número de ejes necesario para ello.

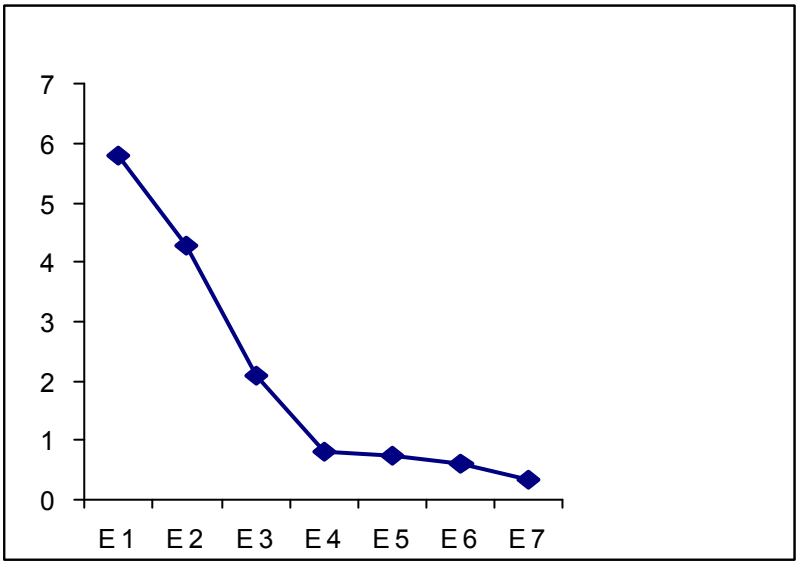
b) Otras reglas son empíricas.

I. Si la nube de datos inicial no tiene ninguna dirección privilegiada, los valores propios serán próximos: no diferirán mucho. Se conservará entonces un eje cuyo porcentaje de varianza sea netamente superior a 1/ p * 100, ya que éste es el valor que le corresponderá si todos los ejes explicasen exactamente la misma cantidad.

II. La inercia explicada por cada eje debe ser decreciente. Esto quiere decir que si representamos gráficamente los valores

propios obtenidos de la matriz de Varianzas y Covarianzas, nos deberíamos quedar con el número de ejes que se encuentran anteriores al “codo “ que se produce en la curva de la gráfica en el plano de las ordenadas x abscisas. Podemos mostrar un ejemplo:

GRÁFICO 2.3.
EJEMPLO DE GRÁFICA DE LOS VALORES PROPIOS.



Según éste gráfico, deberíamos retener los tres primeros ejes que se encuentran anteriores al “codo”.

Como se puede apreciar, se puede adoptar el criterio de escoger el número de componentes, según el número de valores propios que sean mayores a 1. En el gráfico, los tres valores propios que se grafican son mayores a uno.

En la práctica, estas reglas dan el mismo resultado.

2.1.6.3. INTERPRETACIÓN DE UN ACP

Para interpretar el análisis se representan las nubes de puntos sobre planos formados por parejas de ejes de Componentes.

Así entonces, se podrá representar gráficamente C1xC2, C1xC3, C2xC3, , etc., sucesivamente hasta lograr una representación suficiente.

Según criterios de expertos, se consideran importantes solamente las gráficas de C1 vs C2 y de C1 vs C3.

Es útil representar sobre el mismo gráfico las nubes de puntos individuo y puntos variables superpuestos (para ello debe multiplicarse por un coeficiente para que estén en la misma escala). La disposición de los puntos variables en proyección permite interpretar la nube de puntos individuo.

En la interpretación de un Análisis de Componentes Principales es conveniente seguir el siguiente orden:

1. Interpretación de los factores en función de su correlación con las variables.

2. Interpretación de la nube de variables.
3. Interpretación de la nube de individuos.
4. Interpretación de la representación simultánea.

Son de importancia para nuestro trabajo, las interpretaciones de las nubes de variables y de individuos. Para ello presentaremos un análisis de estas interpretaciones.

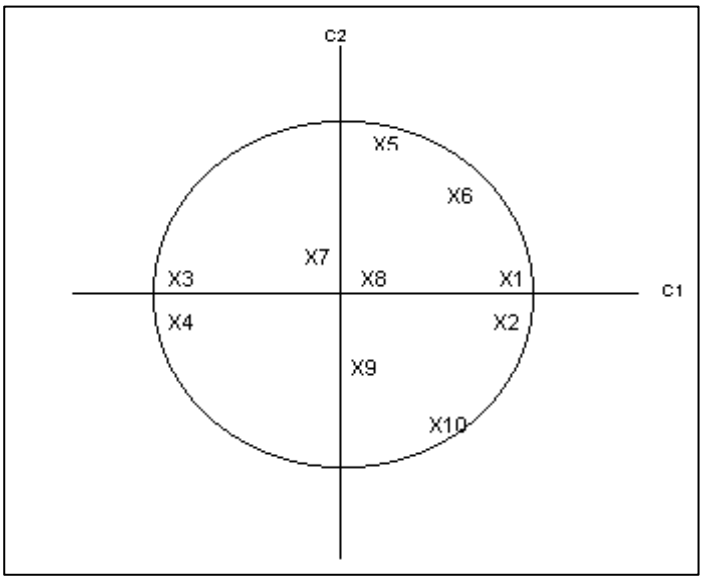
a) Interpretación de la nube de variables

Sobre los planos factoriales los puntos variables están situados en el interior de un círculo de radio unidad . Los puntos variables están mejor representados en el plano cuando están próximos al borde del círculo.

La nube de variables no está centrada en el origen, sino que las variables pueden estar situadas todas al mismo lado del origen si se correlacionan positivamente. En el gráfico que mostraremos a continuación, podremos observar que las variables X1, X2, X3, X4, X5, X6 y X10 están bien representadas sobre el plano (1,2) mientras que X7, X8, X9, lo están mal (alejadas del círculo).

- X1 y X2 están muy correlacionadas positivamente entre sí con la componente 1.
- X3 y X4 están muy correlacionadas positivamente entre sí y negativamente con la componente 1.
- X5 y X2 están incorrelacionadas.

GRÁFICO 2.4.
EJEMPLO DE GRÁFICO DE LA NUBE DE VARIABLES.



El primer factor (componente) opone las variables X1 y X2 a X3 y X4 será aproximadamente función lineal creciente de X1 y X2, y decreciente de X3 y X4. Puede interpretarse como ligado a estas variables.

Se puede sustituir las variables X1 a X4 por C1 sin que se pierda mucha información.

El segundo factor (componente) recoge información que no recogía el primer factor, por ser perpendiculares. Esta componente está ligada a la variable X5 y se interpretaría en función de ella.
La interpretación de las variables X7, X8 y X9 habrá que buscarla en otros ejes, pues están muy mal representados sobre el plano (1,2). No sirven para la interpretación de estos factores, pues su correlación es muy pequeña.

b) Interpretación de la nube de individuos.

Una vez interpretada la nube de variables, se puede pasar a la de individuos, representados por sus coordenadas sobre los factores (componentes).
La nube de individuos estará centrada en el origen por la transformación que hemos realizado en los datos iniciales.

Si dos individuos están bien representados en el plano, su proximidad se interpreta como comportamiento semejante, es decir que si se observa un grupo de individuos, y están bien

representados, significa que toman valores próximos para todas las variables medidas.

Es necesario tener en cuenta que los ejes factoriales se obtienen a partir de la información suministrada por todos los individuos, y uno puede estar mal representado sobre este plano; por eso es necesario comprobar siempre su calidad de representación.

2.1.6.4 APLICACIONES MÁS IMPORTANTES DEL ACP.

El ACP puede ser utilizado para realizar gran variedad de trabajos y se puede complementar con otros métodos. Las aplicaciones más importantes son:

I. Obtención e interpretación de factores.

El ACP permite combinar variables con el objeto de obtener unos factores o variables latentes, que si bien no se pueden medir directamente, sí tienen un significado. Permite establecer si existe una estructura latente.

II. Para seleccionar y reducir el número de variables.

El análisis se puede aplicar a una prueba piloto con gran número de variables para seleccionar un grupo reducido de ellas, las que más importancia tienen en la descripción del fenómeno, y para un análisis más amplio se utilizan las variables más correlacionadas con los primeros factores.

III. Utilización de los factores (componentes) como nuevas variables. Se puede reducir el número de variables a considerar en un estudio si se utilizan los factores resultantes de ACP. Los análisis posteriores son más sencillos al ser menor el número de variables; además éstas están incorrelacionadas, lo que evita algunos problemas que surgen en análisis como el de regresión cuando las variables están correlacionadas.

IV. Agrupación de variables que tienen comportamientos análogos.

2.1.7. ANÁLISIS DE DISCRIMINACIÓN.

Este Análisis de Discriminación es una técnica multivariada que se enfoca en la separación de grupos de observaciones.

Los objetivos principales del análisis de discriminación son los siguientes:

- Describir gráficamente (en tres o algunas dimensiones) o algebraicamente, las diferentes características de observaciones de varias poblaciones consideradas conocidas. Trataremos de encontrar una función que dependa de valores numéricos de tal forma que los grupos de datos se puedan separar tanto como sea posible.

- Clasificar observaciones dentro de dos o más grupos etiquetados anteriormente. El énfasis en obtener una regla que pueda ser usada para asignar de manera óptima, nuevas observaciones en los grupos clasificados.

Una función que clasifica observaciones puede servir algunas veces como un distribuidor, y una regla que distribuye observaciones sugiere un procedimiento discriminatorio. En la práctica, los objetivos anteriormente expuestos, coinciden frecuentemente y la diferencia entre separación y distribución se vuelve confusa. Los valores observados de **X** difieren en algunos puntos de una población respecto de la otra.

Entonces, se puede pensar que los valores de **x** pueden definir la población

2.1.7.1. SEPARACIÓN Y CLASIFICACIÓN DE DOS POBLACIONES.

Establecer ideas, nos permiten hacer una lista de situaciones en las cuales puede ser interesante separar dos clases de poblaciones o asignar una nueva observación a una de dos poblaciones.

Es conveniente etiquetar las poblaciones como π_1 y π_2 . Las observaciones son comúnmente separadas o clasificadas en base a mediciones asociadas a **p** variables aleatorias **X' = [** $X_1, X_2, ...X_p$**].**

Los valores observados de **X** difieren en algunos puntos de una población respecto de la otra. Entonces, se puede pensar que los valores de **x** pueden definir la población π_1 o π_2. Estas dos poblaciones pueden entonces ser descritas por funciones de densidad $f_1(\mathbf{x})$ y $f_2(\mathbf{x})$.

Aquí presentamos algunos ejemplos que sugieren análisis de separación o clasificación de observaciones.

TABLA I
EJEMPLOS DE CASOS QUE SUGIEREN UN ESTUDIO MEDIANTE ANÁLISIS DISCRIMINANTE.

Poblaciones $\pi 1$ y $\pi 2$	Variables de medida **X**
1. Compradores potenciales de un nuevo producto y rezagados (referido a personas que adquieren un producto de manera lenta)	Educación, bienes que posee, tamaño de familia.
2. Masculinos y femeninos.	Mediciones Antropológicas tales como circunferencia y volumen de cráneos ancianos
3. Compañías de riesgo financiero o compañías estables económicamente.	Total de Activos, capital privado, capital prestado, índices financieros, ingresos por ventas.

Como podemos apreciar, en el primer ítem que se propone como ejemplo para análisis discriminante, se muestran dos poblaciones etiquetadas (“compradores” y “rezagados o compradores

de última hora"), y las variables relevantes para la discriminación podrían ser: educación, bienes que posee, tamaño de la familia de la que procede, entre otras . En todo caso, lo que se tratará de clasificar serán observaciones de la forma **X'** = [X_1 (Educación), X_2 (Bienes que posee), X_3 (Tamaño de familia)] como población π_1 , compradores, o población π_2, rezagados.

Se concentrará en una clasificación para dos poblaciones. Usualmente, las reglas de clasificación son obtenidas desde las muestras .

2.1.7.2. CLASIFICACIÓN CON DOS POBLACIONES NORMALES MULTIVARIADAS. FUNCIÓN DE DISCRIMINACIÓN DE FISHER

Los procedimientos de clasificación basados en poblaciones normales predominan en la práctica estadística.

Se asume que $f_1(\mathbf{x})$ y $f_2(\mathbf{x})$ son funciones de densidad normales multivariadas, la primera con vector de medias μ_1 y matriz de

covarianzas $\mathbf{S}_1$ y la segunda con vector de medias μ_2 y matriz de covarianzas $\mathbf{S}_2$.

El caso de matrices de covarianza iguales es muy usual y está relacionado con una estadística de clasificación lineal simple. Entonces en este caso asumimos $S_1 = S_2 = \mathbf{S}$.

Se toman dos tamaños de muestra $\mathbf{n_1}$ y $\mathbf{n_2}$ de poblaciones normales con parámetros especificados anteriormente. Deseamos construir una **función lineal** que discrimine entre observaciones de dos poblaciones por algún tipo de medida de separación.

Para el efecto se debe determinar **a** que proporciona la máxima razón cuadrática media :

$$t^2(a) = \left\{ \left[a^T (\mu_1 - \mu_2) \right]^2 / \; a^T S \; a \right\}$$

Efectuando la maximización con

$$a^T S \; a = 1$$

resulta lo siguiente:

$$a = S^{-1} (\mu_1 - \mu_2)$$

y la función lineal discriminante :

$$y = (\mu_1 - \mu_2)^T (S_{pooled})^{-1} X$$

Entonces adoptamos la siguiente regla de clasificación:

- Asigne la observación **X = [** $X_1, X_2, \ldots, X_p$**]** a la población π_1

si:

$$(\mu_1 - \mu_2)^T (S_{pooled})^{-1} X > \frac{1}{2}(\mu_1 - \mu_2)^T (S_{pooled})^{-1}(\mu_1 + \mu_2)$$

-Asigne la observación **X = [** $X_1, X_2, \ldots, X_p$**]** a la población π_2

$$(\mu_1 - \mu_2)^T (S_{pooled})^{-1} X \leq \frac{1}{2}(\mu_1 - \mu_2)^T (S_{pooled})^{-1}(\mu_1 + \mu_2)$$

si:

Y donde **S pooled** es :

$$S_{pooled} = \left[\frac{n_1 - 1}{(n_1 - 1) + (n_2 - 1)}\right] S_1 + \left[\frac{n_2 - 1}{(n_1 - 1) + (n_2 - 1)}\right] S_2$$

En el caso de una diferencia entre matrices de covarianza de las poblaciones, es decir $S_1 \neq S_2$, la razón de verosimilitud nos conduce a una función discriminante cuadrática:

$$g(X) = (X - \mu_2)^T S_2^{-1} (X - \mu_2) - (X - \mu_1)^T S_1^{-1} (X - \mu_1) - \ln\left(\frac{\det S_1}{\det S_2}\right)$$

Y se adopta la siguiente regla de clasificación:

- Asigne **X** a la población π_1 si:

$g(X) > 0$

- Asigne **X** a la población π_2 si:

$g(X) \leq 0$

En resumen, la idea de Fisher era transformar las observaciones de la matriz multivariada **X** a observaciones univariadas Y, tales que las Y provenientes de las poblaciones π_1 y π_2 tanto como sea posible.

Fisher sugirió el empleo de combinaciones lineales de **X** para crear Y.

2.1.7.3. DISTANCIA CUADRÁTICA DE MAHALANOBIS.

La función discriminante para la clasificación con parámetros conocidos es:

$$y = (\mu_1 - \mu_2)^T (S_{pooled})^{-1} X$$

entonces:

$$\mu_y = E[y] = (\mu_1 - \mu_2)^T (S_{pooled})^{-1} \mu_i$$

$$\sigma_y = VAR[y] = (\mu_1 - \mu_2)^T (S_{poolde})^{-1} (\mu_1 - \mu_2) = \partial^2$$

y de donde δ es la distancia cuadrática de Mahalanobis. Mientras más grande sea ésta, será menos probable que la función discriminante clasifique observaciones de manera errónea.

2.1.7.4. SELECCIÓN DE VARIABLES PARA LA DISCRIMINACIÓN.

En este punto se examinará algún tipo de método para la selección de una o más variables independientes cuantitativas.

Si se utiliza un método de selección de variables para la discriminación, los resultados deberían ser interpretados con precaución. No existen garantías de que el sub-grupo de variables seleccionado, es el mejor, indiferentemente de los criterios usados para dicha selección.

Los problemas asociados con procedimientos para la selección de variables se incrementan cuando existe una gran correlación entre las variables o entre combinaciones lineales de las mismas.

En el análisis discriminante, se puede seleccionar aquel subconjunto de las variables independientes que más discrimine los grupos establecidos por los valores de la variable dependiente. En este caso se puede utilizar el criterio Lambda de Wilks.

LAMBDA DE WILKS.

Según este criterio, para elegir el subconjunto de variables independientes más discriminantes, sería adecuado considerar aquel tal que, al representar el conjunto de toda la muestra en el subespacio generado por los valores de las variables, por un lado, los centros de los grupos estuvieran muy separados entre sí y, por

otro, dentro de cada grupo el comportamiento fuera muy homogéneo, con valores poco dispersos y cercanos al centro.

La Lambda de Wilks para un conjunto de *p* variables independientes mide las desviaciones dentro de cada grupo respecto a las desviaciones totales sin distinguir grupos, en el espacio *p*-dimensional generado por los valores de las *p* variables. Si su valor es pequeño , la variabilidad total será debida a las diferencias entre grupos y, por tanto, el conjunto de variables correspondiente discriminará los grupos. Por el contrario, si su valor es próximo a 1 los grupos estarán mezclados y el conjunto de variables independientes no será adecuado para construir las funciones discriminantes.

El hecho de que una variable sea la candidata a ser seleccionada no implica que vaya a serlo. Es decir, que la Lambda de Wilks tome el mínimo valor no implica que éste sea pequeño. Después, habrá que establecer un criterio para determinar si la información aportada por la variable candidata a ser seleccionada en un paso es significativa.

2.1.8. ANÁLISIS DE CONTINGENCIA

El análisis de contingencia es una técnica estadística que permite determinar dependencia entre dos variables aleatorias.

Su análisis se realiza a partir de la tabla de contingencia que es un arreglo matricial de r filas y c columnas, donde r es el número de niveles del factor 1 o de la variable X y c el número de niveles del factor 2 o de la variable Y. La dependencia o independencia se concluye mediante un contraste de hipótesis del siguiente tipo:

$$Ho: X \; y \; Y \quad son \; independientes$$
$$vs$$
$$H_1: \neg Ho$$

El procedimiento para verificar la validez del siguiente contraste es el siguiente:

- Realizar la tabla bivariada de c filas x r columnas.
- Bajo la hipótesis nula, determinar los valores esperados para cada celda i j de la tabla bivariada.

- Se halla el estadístico de prueba para el contraste mencionado anteriormente. Este estadístico se lo halla de la siguiente manera:

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(X_{ij} - E_{ij}\right)^2}{E_{ij}}$$

donde:

X_{ij} : valor observado en la celda i j.

E_{ij} : valor esperado de la celda i j.

r : # de filas de la tabla.

c : # de columnas de la tabla.

- Se rechaza *Ho* en favor de H_1 con un (1-∝)100% si:

$$X^2 > \chi_{\alpha}^{2}(r-1)(c-1)$$

2.1.9. FUNDAMENTOS DE ANÁLISIS FINANCIERO.

Mediante este punto daremos a conocer algunos conceptos que están íntimamente relacionados con las variables numéricas a utilizarse en nuestro análisis estadístico, por lo cual considero importante mencionarlos.

Entre los preceptos más importantes, podemos mencionar a :

2.1.9.1. LAS CUENTAS

El dispositivo básico del resumen de contabilidad es la cuenta. Ésta es el registro detallado de los cambios que han ocurrido en un activo, pasivo o patrimonio, durante un período. Las cuentas se agrupan en tres amplias categorías de acuerdo con la ecuación contable:

ACTIVOS = PASIVOS + PATRIMONIO

ACTIVOS

Activos son los recursos económicos que benefician al negocio y continuarán haciéndolo en el futuro. La mayoría de las empresas usan las siguientes cuentas de activos:

-Efectivo. -Documentos por Cobrar.

-Cuentas por Cobrar -Edificios.

Entre otros.

PASIVOS

El Pasivo es una deuda, es parte de la empresa que le pertenece a terceras personas. Una empresa tiene generalmente menos cuentas de pasivos que cuentas de activos, porque los pasivos del negocio pueden resumirse bajo relativamente pocas categorías. Entre las más importantes tenemos:

-Documentos por Pagar

-Cuentas por Pagar.

-Pasivos acumulados.

PATRIMONIO

Es el derecho del propietario a los activos de un negocio o empresa. También se lo conoce como “capital contable”. En una compañía de propietario único o una sociedad , el Patrimonio se divide a menudo en cuentas separadas: para el lado del capital del propietario y para los retiros del propietario.

INGRESOS POR VENTAS

Es considerada una cuenta del Patrimonio. Es el aumento en el capital contable debido a la entrega de bienes o servicios a los clientes. Si un negocio presta dinero a una persona ajena al negocio, necesitará una cuenta por intereses para el interés ganado por el préstamo.

2.1.9.2. TIPOS DE RAZONES FINANCIERAS

Cada tipo de análisis tiene un propósito o un uso que determina las diferentes relaciones que destaca. Un analista puede ser un banquero, y los banqueros están preocupados en la liquidez a corto plazo de una empresa a la que se podría otorgar un préstamo a corto plazo. En contraste, los acreedores a largo plazo dan mucho más énfasis al poder de generación de utilidades y a la eficacia de operación. Ellos saben que las operaciones no rentables erosionan los valores de los activos. Los inversionistas del Capital Contable están similarmente interesados en la rentabilidad a largo plazo y en la eficacia. Por esto, es muy útil clasificar a las razones financieras dentro de seis

tipos fundamentales:

1.- Razones de Liquidez.

Miden la capacidad de la empresa para satisfacer las operaciones que venzan a corto plazo.

2.- Razones de Apalancamiento Financiero.

Miden el grado en que la empresa ha sido financiada mediante deudas a terceros.

3.- Razones de Actividad.

Mide la eficacia con la cual la empresa usa sus recursos.

4.- Razones de Rentabilidad.

Miden la eficacia de la administración y cómo se muestra en los rendimientos que se han generado con las ventas, la inversión y los activos.

5.- Razones de Crecimiento.

Miden la capacidad de la empresa para mantener su posición económica en el crecimiento de la economía y de la industria.

6.- Razones de Valuación.

Miden la capacidad que tiene la administración para crear valores de mercado en exceso de los desembolsos del costo de la inversión. Las razones de valuación son la medida más completa del desempeño, en tanto que reflejan las razones de riesgo y las razones de rendimiento.

Por el hecho de que en nuestra matriz de datos encontramos razones de rentabilidad, pondremos énfasis en los conceptos respecto de estas razones financieras.

2.1.9.3. RAZONES DE RENTABILIDAD

La rentabilidad es el resultado neto de un buen número de políticas y decisiones. Las razones de rentabilidad dan las respuestas finales acerca de la eficacia con que se maneja la empresa.

Rendimiento sobre los Activos Totales.

El rendimiento sobre los activos totales pretende medir la eficacia con la cual la empresa ha empleado sus recursos totales; alguna veces se denomina rendimiento sobre la inversión. Se calcula mediante la siguiente ecuación:

Rendimiento = (Utilidad o Pérdida / Activos Totales.)*100.

Si tenemos un rendimiento del 8%, podemos entender que por cada 100 unidades monetarias que la empresa invierte en la compra de activos, gana 8 unidades monetarias sobre estos.

Rendimiento sobre el Patrimonio

La razón de la utilidad neta después de las impuestos al patrimonio o capital contable mide la tasa de rendimiento sobre la inversión de los accionistas. Se calcula así:

Rendimiento = (Utilidad o Pérdida / Capital Contable)*100.

Si se obtuviera un rendimiento del 12% se podrá entender que por cada 100 unidades que invierten los accionistas de la empresa, ganan 12 unidades.

Rendimiento sobre los Ingresos.

También se lo conoce como “margen de utilidad” sobre las ventas. Se lo calcula dividiendo las Utilidades Netas después de Impuestos entre los Ingresos por ventas , y nos da como resultado la utilidad o pérdida por 100 unidades monetarias de ventas.

Rendimiento = (Utilidad o Pérdida / Ingresos por ventas)*100.

Si tuviéramos como resultado de esta ecuación un porcentaje del 5%, y este se por debajo del promedio de la industria, podríamos

decir que los precios de la empresa son relativamente bajos o que sus costos son relativamente altos, o ambas cosas a la vez.