I.- Introducción.

Nota: Estos apuntes se complementará con ejemplos en clases, solo es material conceptual.

1.1 La Estadística.

1.1.1 ¿Qué es la estadística?

La estadística ha existido desde los comienzos de la civilización, en formas muy sencillas y elementales, pues utilizaban símbolos gráficos en pieles, rocas, arcillas, madera, paredes para contar el numero de personas, animales o cosas. Dichos indicios históricos muestran, por ejemplo: los babilonios usaban pequeñas tablillas de arcilla para recopilar datos sobre la producción agrícola y sobre los rubros vendidos o cambiados por trueque, los egipcios, griegos, romanos y chinos realizaban censos de la población y describían las rentas de sus países mediante inventarios de todos sus bienes.

Pero, es solo a partir de los primeros estudios sobre la teoría de probabilidades, en la que comienza a tomar un carácter científico, lo cual se fue desarrollando en el siglo XIX y a comienzos del siglo XX, sufriendo una transformación más rápida en la segunda guerra mundial, cuando por la necesidad del momento se sistematiza la enseñanza de la estadística como ciencia de amplios campos de aplicación.

En gran parte el avance de la estadística comenzó antes de la "revolución de las computadoras", la amplia disponibilidad y su uso, ha acelerado el proceso en gran medida, en particular porque ellas nos permiten clasificar, procesar y analizar grandes cantidades de datos.

El estudio de la Estadística permite, entre otras cosas

- Aprender las reglas y métodos usados en el tratamiento de información
- Evaluar y cuantificar la importancia de los resultados estadísticos obtenidos
- Entender mejor algunos fenómenos de interés (Sociales, Económicos, Biológicos, Educacionales, etc.)
- Dar una visión más clara acerca de la información proveniente de diversas fuentes.

Algunos aspectos estadísticos manejados en la información obtenida de la radio, la televisión u otro medio, influencian fuertemente a gran cantidad de personas pero a veces no proporcionan una descripción cabal de los que pretenden mostrar.

Como una de las tareas de la Estadística es el estudio de fenómenos aleatorios, esto hace muy pertinente el tratar de explicar la manera como se comportan (Variabilidad).

Entre otras cosas la Estadística se ocupa del manejo de la información que pueda ser cuantificada. Implica esto la descripción de conjuntos de datos y la inferencia a partir de la información recolectada de un fenómeno de interés. La función principal de la estadística abarca: Resumir, Simplificar, Comparar, Relacionar, Proyectar.

Entre las tareas que debe enfrentar un estudio estadístico están:

1. Delimitar con precisión la población de referencia o el conjunto de datos en estudio, las unidades que deben ser observadas, las características o variables que serán medidas u observadas.
2. Estrategias de Observación: Censo, Muestreo, Diseño de Experimental.
3. Recolección y Registro de la información.
4. Depuración de la información.
5. Construcción de Tablas.
6. Análisis Estadístico:
   - Producción de resúmenes gráficos y numéricos.
   - Interpretación de resultados.

Cuando los datos comprenden toda la población de referencia, hablamos de un Censo y cuando solo comprometen una parte de ella, hablamos de una muestra. En ambos casos es pertinente un análisis *Descriptivo*. En el segundo caso un análisis *Inferencial*.

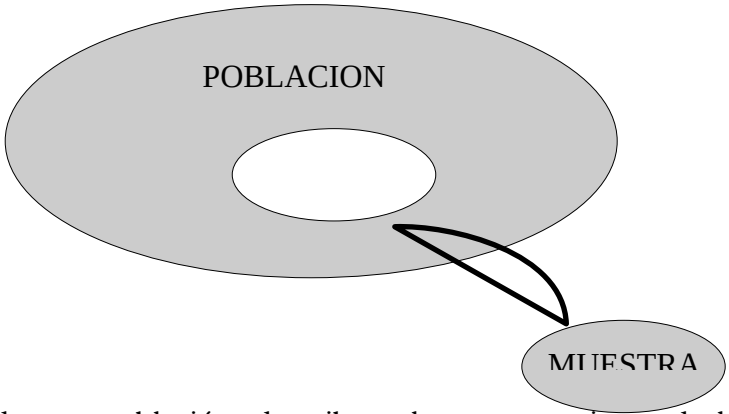
1.1.2 Estadística descriptiva e inferencial.

La estadística se divide en dos partes: a) Estadística descriptiva.
b) Estadística inferencial.

Antes de ver ambos conceptos, definiremos dos conceptos fundamentales para la estadística.

Una *población* es el conjunto de datos o elementos que consta de todas las observaciones concebibles (o hipotéticamente) posibles de un fenómeno determinado y limitado en espacio y tiempo, objeto de estudio que tienen una característica común y observable desde el punto de vista estadístico.

A un subconjunto de la población se le denomina *muestra* y es el número determinado de datos tomados de la población, seleccionada de manera que pueda representar a esa población , es decir debe reflejar las características esenciales de la población de la cual se obtuvo.



Los *parámetros* de una población, describen el comportamiento de la población mediante características constantes (sus valores no varían).

A una función de la muestra se le denomina *estadístico*, esté describe el comportamiento de la muestra mediante características aleatorias (sus valores no son constantes) .

Los parámetros son fijos, pues no dependen de ninguna muestra, los estadísticos dependen siempre de la muestra seleccionada. Las características desconocidas de una población serán llamadas parámetros. Las características calculadas a partir de una muestra son llamadas estadísticas. Una inferencia es una generalización obtenida a partir de una muestra aleatoria.

La *Estadística Descriptiva* la podemos considerar como un conjunto de técnicas y métodos encargados de tabular, organizar, procesar y presentar datos mediante cuadros y gráficas del comportamiento de una población, calculando frecuencias, promedios y porcentajes, es decir; esta diseñada para resumir o describir dichos datos sin factores pertinentes adicionales, sin intentar inferir nada que vaya más allá de los datos como tales. La estadística descriptiva es la que trabaja con todos los elementos de una muestra y los cálculos realizados sólo son validos para dicha muestra.

A la *Estadística Inferencial* es una serie de técnicas y métodos encargados de procesar y analizar datos a partir del conocimiento de la muestra y la incertidumbre, para obtener conclusiones o tomar decisiones. Es la que hace que todas las mediciones hechas a una muestra sean validas para la población de la que se sacó la muestra.

1.2. Estadística Descriptiva.

1.2.1. Definiciones básicas.

Aunque la estadística descriptiva es una rama importante y se continua usando en forma general, casi siempre se deriva de muestras, lo que implica que su análisis requiere de generalizaciones que van más allá de los datos y en consecuencia el avance de la estadística ha sido un cambio en el énfasis de la estadística descriptiva a los métodos de la inferencia estadística.

Variable (*Dato:* Resultado de una medición u observación, que debe estar en forma numérica.)*:* Característica común y observable que se mide con valores cualitativos o cuantitativos y que difieren entre los elementos en una muestra, por lo que se presentan variaciones que sirven para distinguir o describir aspectos específicos, podemos decir, las variables resultan ser aquellas características de interés que desean ser medidas sobre los objetos o individuos seleccionados. Tipos de variables:

___ *Variable cualitativa* son las presentadas en forma de atributos. Ejemplo: sexo, color de auto, tipo de sangre, estado civil, categoría de un profesor, etc.

___ *Variable cuantitativa* son aquellas que se presentan en forma numérica. Ejemplo peso, estatura, presión sanguínea, distancia recorrida, etc. La variables cuantitativas se dividen en *variables discretas* y *variables continuas,* las primeras son aquellas que no pueden tomar todos los valores del recorrido de los datos, solo toma valores enteros, sus datos son contables y las variables continuas son aquellas que pueden tomar cualquier valor del recorrido de los datos, sus datos son medibles.

Nota: Algunos datos numéricos pueden ser clasificados como cuantitativos o cualitativos según su uso. Por ejemplo, la estatura de una persona se mide en centímetros, pies, metros y es entonces una medida cuantitativa. Pero si se mide como Bajo, Medio y Alto, se convierte en una medida cualitativa.

Escala de medición: es el proceso de medición de los diferentes valores de una variable. Tipos de escalas:

__ *Escala nominal*: Usada para variables categóricas o cualitativas, que se convierten en cuantitativas asignando números a las categorías, este tipo de escala solo identifica y clasifica. Por ejemplo, la raza, estado civil, sexo (M o F), religión, tipo de sangre, constituyen datos nominales.

__ *Escala ordinal*: Usada para variables cuantitativas, este tipo de escala clasifica y ordena, es decir, se establecen desigualdades, no tiene sentido realizar operaciones aritméticas con ellas. Por ejemplo, las categorías A, B, C, D, E como calificación o niveles de perfeccionamiento, categoría de un profesor (Ayudante, Profesor de asignatura tipo A, Profesor de asignatura Tipo B, Profesor titular), nivel socioeconómico (alto, medio bajo), son datos ordinales.

___ *Escala de intervalo*: Usada para variables cuantitativas, no solo se ordena, sino que además es posible medir exactamente la intensidad con la que se posee la característica, se puede indicar la distancia entre ellos, exigiendo establecer algún tipo de unidad física de medición, aquí se incluye un origen o un "cero" arbitrario que se establece en base a conveniencias practicas, admite operaciones de suma y resta. Por ejemplo las escalas de temperatura pertenecen a es tipo de escala pues el cero en ellas no está implicando ausencia de temperatura.

___ *Escala de razón*: Usada para variables cuantitativas, tiene el mismo procedimiento que la escala de intervalo, con la diferencia de que se incluye un cero absoluto que si indica ausencia de valores y admite cualquier tipo de operación matemática, así como comparar mediante proporciones o razones. Por ejemplo, número de adultos en un hogar, ingreso mensual en el hogar, etc.

1.2.2 Distribución de frecuencias.

La idea primordial de este proceso es simplificar la forma como se representa la información, se estructuran principalmente para condensar conjuntos numerosos de datos y representarlo en una forma fácil de entender. La información puede mostrarse de dos maneras: No-agrupada y Agrupada.

En aquellos casos donde la cantidad de valores de una o varias variables es muy grande, se hace necesario resumirlos para una presentación más adecuada y en algunos casos agruparlos en clases, rangos o intervalos para facilitar su interpretación.

La *frecuencia* de una medida o de una categoría es el número de veces que esta aparece en una colección de datos. Usualmente denotada **f**. La información que contiene los valores de dichas medidas y sus respectivas frecuencias se llamará *Tabla de Frecuencias*. Por ejemplo, se tienen los datos respecto a las calificaciones de los estudiantes de un curso de Estadística I: 9 8 7 8 4 3 2 1 0 5 3 2 1 1 7 3 2 8 7 6 6 4 3 2 2 0 9 4 6 9 6 9 43 5 7 3 2 1 4 4 2.

Calificación	0	1	2	3	4	5	6	7	8	9
Frecuencia	2	4	7	6	6	2	4	4	3	4

Una agrupación de estos datos puede reducir más la presentación:

Calificación	0	1 – 2	3 – 4	5 – 6	7 – 8	> 9
Frecuencia	2	11	12	6	7	4

La manera como se agrupa la información debe corresponder a algún propósito particular de quien analiza la información o requerimiento del investigador conocedor de la información.

La *frecuencia absoluta* **(fi),** son las frecuencias observadas en una investigación, su suma es siempre igual al número total de observaciones. La *frecuencia acumulada* **(F_i)**, es la suma de las frecuencias absolutas comprendidas hasta un determinado valor de la variable o hasta una determinada clase. La *frecuencia relativa* (**fri**), es el cociente que resulta de dividir cada frecuencia absoluta entre el numero total de observaciones. La *frecuencia relativa acumulada* **(Fr_i)**, es la suma de las frecuencias absolutas comprendidas hasta un determinado valor de la variable o hasta una determinada clase divididas entre el total de las observaciones. El *Intervalo de clase* **(Ic)**, son los fraccionamientos en grupos que se hacen del intervalo total o recorrido de los datos.
Para construir los intervalos de clases se utiliza la metodología del "tanteo":

a) Definir el *rango* o *recorrido* de la variable *R* =Valor max. – Valor min.
b) Calcular el numero de intervalos (k) a construir, utilizando la regla de Sturges
$k = 1 + 3.3\log(n)$ n tamaño de la muestra
c) Determinar la amplitud del intervalo (a_i), $a_i = R / k$

Nota: Se elige el tipo de intervalo a utilizar, se recomienda que sea de amplitud constante.

La *marca de clase* se obtiene una vez obtenido los intervalos de clase (**Ic**), la marca de clase del i-ésimo intervalo de clase se obtiene con:

mi = (Valor máx. $\mathbf{Ic}_i$ + Valor min. $\mathbf{Ic}_i$)/ 2

1.2.4 Medidas descriptivas

A menudo es necesario resumir los datos por medio de un número único que describe a su modo el conjunto entero. El tipo de numero que se selecciona depende de la característica particular que se quiere describir, y pertenecen a un conjunto de medidas numéricas descriptivas.

1.2.3.1 Medidas de tendencia central.

Las medidas de tendencia central también se les conoce como medidas de localización o de tendencia; entre estas describen el centro o punto medio de los datos y las más empleadas son: la media, mediana y moda.

La *media aritmética* **(μ)**, representa el centro o punto en el cual giran los valores de la variable por debajo o por encima . La popularidad de la media como medida del punto medio de un conjunto

de datos no es una coincidencia; además del hecho de ser una medida simple y común, estas son algunas de las características importantes que hay que recordar:

1. Un conjunto de datos tiene una y solo una media, entonces esta siempre es única
2. Lleva a un tratamiento estadístico más a fondo
3. Es relativamente confiable en el sentido de que las medias de muchas muestras obtenidas a partir de la misma población no fluctúan o varían tanto como otras medidas estadísticas empleadas para estimar la media de una población
4. Toma en cuenta todos los elementos de un conjunto de datos.
5. Es sensible a valores extremos.

Para *datos simples o no agrupados*, la media poblacional se calcula de la siguiente forma

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

donde: μ es la media poblacional,
x_i es la *i*-ésima observación,
N es el total de observaciones de la población.

La media muestral se obtiene de la siguiente manera

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

donde: $\overline{X}$ es la media nuestral,
x_i es la *i*-ésima observación,
n es el total de observaciones de la muestra.

Para *datos agrupados* , la media muestral se obtiene de la siguiente manera

$$\overline{X} = \frac{\sum_{i=1}^{k} \mathbf{mifi}}{n}$$

donde: **mi** es la marca de clase,
fi es la frecuencia absoluta del i-ésimo intervalo,
k es el total de intervalos de clases,
n es el total de observaciones de la muestra.

La poblacional se obtiene de manera análoga, sólo el tamaño de la muestra *n* por el de la población *N*.

La *mediana* **(Me)** es una medida de localización que se utiliza cuando en un conjunto de datos hay valores extremos y se necesita describir el "centro" de estos; pues esta medida no es sensible a dichos valores, ella es el valor de la variable en donde se dividen *n* observaciones en partes iguales y para su ubicación es necesario que los datos se ordenen en forma creciente o ascendente. Como se podrá observar, la mediana divide la información en dos partes porcentualmente iguales

Se toma como mediana **Me** la observación de posición $l(\mathbf{Me}) = (n + 1)/2$

Para *datos simples*: a) si *n* es impar la **Me** es el valor central.
b) Si *n* es par es el promedio de los dos valores centrales.

Para *datos agrupados*

$$\mathbf{Me} = L_i + a_i\,(n/2 - \mathbf{F_{i-1}}) \,/\, \mathbf{fi}$$

Primero se ubica la *clase medianal* que se ubica en la columna de la frecuencia acumulada **Fi** $\geq n/2$ y donde:
L_i es el limite inferior de la clase medianal
a_i es la amplitud del intervalo de la clase medianal
$\mathbf{F_{i-1}}$ es la frecuencia acumulada anterior a la clase medianal
fi la frecuencia absoluta de la clase medianal

La *moda* **(Mo)** se define como el valor que más se repite, no se encuentra afectada por valores extremos y destaca los valores individuales. En caso de existir dos datos que se repitan más, se dirá que la distribución de los datos es bimodal. Para *datos simples*, sólo se toma el valor más frecuente.
Para datos agrupados se realiza los siguiente :

$$\mathbf{Mo} = L_i + a_i\,(\Delta 1 \,/\, \Delta 1 + \Delta 2)$$

Primero se ubica la *clase modal* que es aquella a la cual corresponde la mayor frecuencia absoluta y donde:
Li es el limite inferior de la clase modal
a_i la amplitud de la clase modal
$\Delta 1$ es la frecuencia absoluta modal menos la frecuencia absoluta premodal, $\Delta 1 = \mathbf{fi} - \mathbf{fi_{-1}}$
$\Delta 2$ es la frecuencia absoluta modal menos la frecuencia absoluta postmodal, $\Delta 2 = \mathbf{fi} - \mathbf{fi_{+1}}$

1.2.3.2 Medidas de posición

Indican que porcentaje de datos dentro de una distribución de frecuencias superan estas expresiones, estas medidas reciben el nombre de fractiles, los cuales fraccionan los datos en n partes iguales y entonces tenemos los cuartiles, deciles y porcentiles.

Los *Cuartiles* **($\mathbf{Q_k}$),** dividen al conjunto de observaciones ordenadas en cuatro partes, Q_1 es aquel valor que supera al 25% de los datos y es superado por el 75% restante, Q_2 supera y es superado por el 50% de los datos, Q_2 es igual a la mediana y Q_3 es aquel valor que supera al 75% de los datos y es superado por el 25% de los datos restantes. Veamos como calcular el cuartil inferior Q_1 y el cuartil superior Q_3.
Para obtener el cuartil inferior se calcula primero

$$l(q_1) = \frac{[l(\mathbf{Me})]+1}{2}$$

donde $[l(\mathbf{Me})]$ es la parte entera de la posición que ocupa la mediana, una vez calculado $l(q_1)$, se obtiene la observación de la posición $l(q_1)$ contando de menor a mayor si fue entero (los datos deben estar ordenados, o el promedio de las observaciones alrededor de $l(q_1)$ si fue fraccionario.

Para obtener el cuartil superior se realiza el mismo procedimiento que el que se hizo para obtener el cuartil inferior pero ahora contando de mayor a menor .

Para datos agrupados los cuarteles se obtienen de la siguiente manera

$$Q_k = L_i + a_i\,(\,kn/4 - \mathbf{F_{i-1}})/\,\mathbf{fi}\,, \qquad \mathbf{Fi} \geq kn/4 \qquad k = 1,2,3$$

1.2.3.3 Medidas de dispersión

En la mayoría de los conjuntos de datos, no todos son valores iguales, el grado en el que varían es de suma importancia en la estadística inferencial. El concepto de variabilidad permite conocer el grado de homogeneidad o heterogeneidad que presentan los datos en estudio con respecto a la media elegida en las medidas de tendencia central. Mientras menor sea el grado de dispersión, mucho más concentrados están los datos con respecto a la media y más representativo serán; de lo contrario, mientras mayor sea la variación menos representativos serán. Las medidas más empleadas son el rango, la varianza, la desviación típica y el coeficiente de variación.

El *Rango* **(R)** es un indicador de dispersión absoluta, que no refleja el comportamiento de los datos, solo define la diferencia entre la mayor y menor observación de los datos.

$$\mathbf{R} = \text{Valor max.} - \text{Valor min.}$$

Rango intercuartílico (**Rq**), indica información acerca del 50% de los datos,

$$\mathbf{Rq} = Q_3 - Q_1$$

Es una estadística resistente ya que su valor no se verá afectado en presencia de observaciones aberrantes.

La *Varianza* **(σ^2)** es el promedio cuadrático de las desviaciones de los valores de la variable con respecto a la media. Se usa principalmente para estimar la varianza de la población en problemas de inferencia. Si la varianza de un conjunto de datos es pequeña los valores se concentran cercano a la media y si está, es amplia los valores se acumulan en forma esparcida alrededor de la media.

Para datos simples se obtiene de la siguiente manera:
La varianza poblacional

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

La varianza muestral

$$S^2 = \frac{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2}{n}$$

Nota: Si de lugar de dividir entre n dividimos entre n–1, obtenemos la cuasivarianza o varianza de Cochran.

Para datos agrupados, la varianza poblacional y la varianza muestral son respectivamente

$$\sigma^2 = \frac{\sum_{i=1}^{k}\left(\mathbf{mi} - \mu\right)^2 \mathbf{fi}}{N}$$

y

$$S^2 = \frac{\sum_{i=1}^{k}\left(\mathbf{mi} - \overline{X}\right)^2 \mathbf{fi}}{n}$$

La *desviación típica o estándar (σ)*, es la raíz cuadrada de la varianza.

Todas las medidas de dispersión nombradas anteriormente son medidas de dispersión absoluta, la única medida de dispersión elativa es el *coeficiente de variación*.

El *Coeficiente de Variación* **(CV)** es un medida de dispersión relativa, que se obtiene del cociente de la desviación típica sobre la media, con la gran ventaja de que este coeficiente es independiente de las unidades de medición, debido a que tanto la media como la desviación estándar se miden en las unidades originales. Por esta razón resulta sumamente útil para comparar la variabilidad de dos o más conjuntos de datos.

1.2.3.4 Otras medidas

__ Medidas de asimetría.

Permiten conocer la forma como los datos se alejan o se acercan con respecto a la media en una distribución , deduciéndose así , hacia donde se encuentran concentrados la mayoría de los datos. De acuerdo al tipo de distribución se clasifica en: Asimetría Positiva, Negativa o Simétrica.

Asimetría Positiva (*Sesgada a la derecha o positiva*), los datos están ubicados en la cola derecha de la curva.
Asimetría Negativa (*Sesgada a la izquierda o negativa*), los datos están concentrados en la cola izquierda de la curva.
Simétrica, los datos se encuentran proporcionalmente distribuidos en ambos lados de la curva

Una forma de saber la asimetría de una distribución es:

$$\gamma_1 = \frac{\mu_3}{\sigma^3}$$

donde $\mu_3 = E\left[\left(x-\mu\right)^3\right]$ es el tercer momento alrededor de μ,

Cuando $\gamma_1 = 0$ es simétrica .

Cuando $\gamma_1<0$ la asimetría es derecha o positiva.
Cuando $\gamma_1>0$ la asimetría es por la izquierda o negativa.

En caso de tener una muestra de la población se puede utilizar la muestral

$S^2= \sqrt{\frac{\sum_{i=1}^{n}\left(x_i-\overline{X}\right)^2}{n}}$ para σ y $\hat{\mu}_3 = \frac{\sum_{i=1}^{n}\left(x_i-\overline{X}\right)^3}{n}$ de lugar de μ_3.

Otra forma de saber que tipo de asimetría tienen los datos es por medio del siguiente coeficiente de asimetría:

$$K_A = \frac{\overline{X}-Me}{S}$$

Cuando la media es mayor que la mediana, la asimetría es positiva
Cuando la media es menor que la mediana, la asimetría es negativa
Cuando la media es igual a la mediana, es simétrica

__ Medidas de apuntamiento (Kurtosis)

Determinan el grado de apuntamiento o picudez de una curva con respecto a la distribución o curva normal, se usa para mostrar el grado de concentración (curva con gran apuntamiento) de los datos o dispersión de los mismos (curva achatada).

Los tipos de Kurtosis que podemos encontrar son:

Curva Mesokúrtica, estas curvas presenta el mismo apuntamiento que la curva normal.

Curva Platikúrtica, tienen menor apuntamiento que la normal, es más achatada o aplastada , encontrándose que los datos aunque proporcionalmente distribuidos están mucho más dispersos.

*Curva Leptokúrtica***,** presentan mayor apuntamiento que la curva normal, existiendo mayor concentración de los datos alrededor de la media, menos dispersión.

$$K = \frac{\mu_4}{\sigma^4}$$

en caso de tener una muestra de la población se utilizá las muestrales.

Para

K< 3, Platikurtica
K > 3, Leptokurtica
K = 3, Mesokurtica.

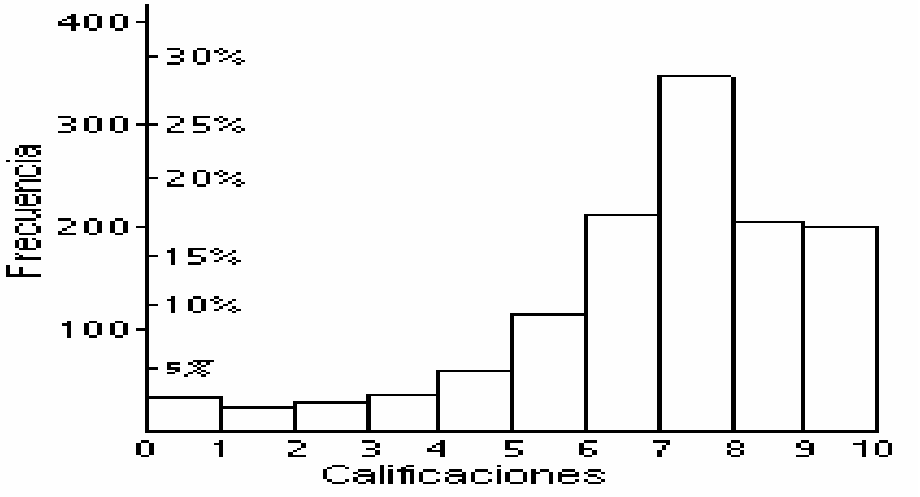
1.2.4 Representación gráfica de las distribuciones de frecuencias.

Para las variables cualitativas se pueden usar, *diagrama de sectores o circular*, *diagrama de barras*, *diagramas de caja* (*Boxplot*), *diagrama de punto*, *diagrama de tallos y hojas* .

Para las **v**ariables cuantitativas se suelen usar , *histogramas*, o*jiva*, *polígono de frecuencia*, *Boxplot*, *diagrama de tallos y hojas*.

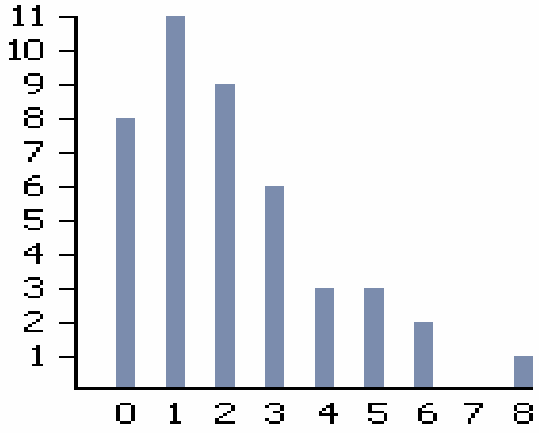
A continuación se dará una breve explicación sobre los tipos de gráficas más usales y la forma de graficarlos.

El *Histograma,* se define como rectángulos que tiene por base el intervalo de clase de la distribución de frecuencias y por altura la frecuencia absoluta o acumulada correspondiente a cada clase según sea su tipo.

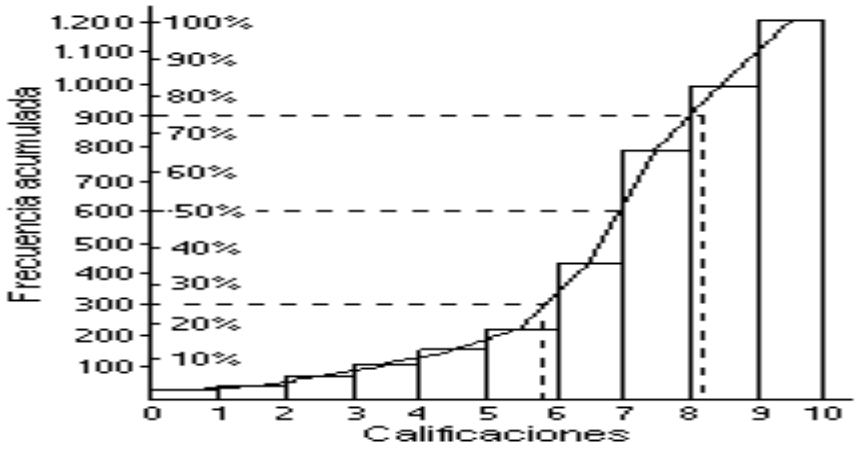


Nota. El histograma no representa los datos del ejemplo visto en la definición de frecuencia.

El *Diagrama de Barras***,** son parecidas a los histogramas , las alturas de los rectángulos o barras representan las frecuencias de clases, pero no hay motivo para tener una escala horizontal continua. Sirven principalmente para representar el total de una cierta cantidad para cada año o para cada categoría presentada.



El *Polígono de Frecuencia*, es un gráfico de líneas continuas que se traza sobre las marcas de clases, geométricamente el área bajo la curva representa el 100% de las observaciones.



La *Ojiva***:,** llamado también polígono de frecuencia acumulada, es igual al poligono de frecuencia definido anteriormente sólo que las líneas se trazan en las fronteras de clase en lugar de hacerlo en las marcas de clase.

El *Diagrama de Sectores o Circular (o de Pay)*, se divide un círculo en secciones que sean proporcionales en tamaño con las frecuencias o porcentajes correspondientes, a cada categoría. Se pueden obtener los ángulos para cada una de las proporciones de la siguiente manera:

ángulo de la proporción del círculo = 360x**fri** (**fri** frecuencia relativa)

Diagrama de puntos. En estos diagramas se puede apreciar el número de veces en que se presneta cada medición en el conjunto de datos. La construcción de los diagramas de puntos se lleva a cabo colocando en el eje horizontal las diferentes observaciones de la variable y sobre cada valo se anotan tantos puntos como veces se repiten estos valores.

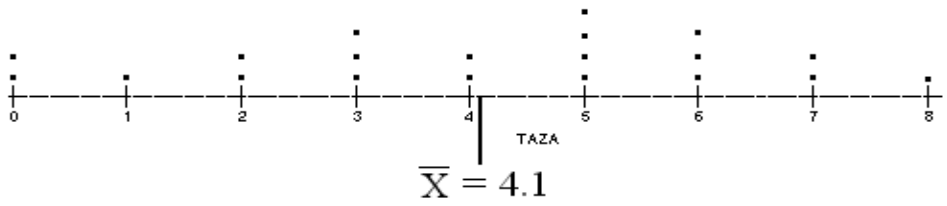


Diagrama de tallo y hoja, este procedimiento combina dos metodos: uno gráfico y otro de ordenación, en donde los valores de los datos se utilizan par efectuar dicha ordenación. El tallo del diagram se forma con el (los) primer(os) dígito(s) del dato, mientras que la hoja se forma con los dígitos restantes.

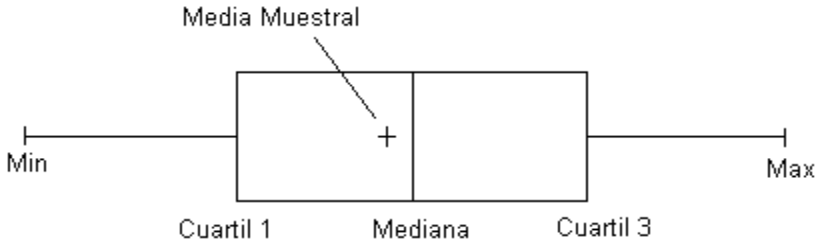
Diagrama de Cajas (Boxplot), los diagramas de cajas son herramientas gráficas muy útiles para describir características importantes en un conjunto de datos, como son centro, simetría o asimetría, valores atípicos (raros), etc. La construcción de este diagrama emplea medidas

descriptivas que son poco sensibles a datos extremos y por lo tanto presentan una descripción más clara de la información. Básicamente empleamos para su construcción los tres cuartiles, los valores mínimos y máximos y la media muestral solo como medida de localización en el gráfico.

Una observación se dice *Atípica o Inusual* si está a más de 1.5 veces el rango intercualtil de alguno de los cuartiles Q_1 o Q_3. Una observación se dice *atípica extrema* si está a más de 3 veces el rango intercuartil de alguno de los cuartiles Q_1 o Q_3.

El diagrama está conformado por una caja la cual se construye con ayuda del primer y tercer cuartil. La mediana es dibujada en el interior de la caja al igual que la media muestral. Los "bigotes" se extienden desde los cuartiles a la derecha y a la izquierda. Su longitud depende de si hay o no datos atípicos.

Sin valores atípicos ni extremos:



Con valores atípicos y/o extremos:

1.3. Análisis Exploratorio Multivariado.

1.3.1 Introducción.

Existen varios métodos para producir rápidamente y visualizar resúmenes simples de conjunto de datos de dos dimensiones, tal que nos permiten obtener información importante acerca de ellos; por ejemplo, el análisis por medio de la obtención de la mediana, los cuantiles o histogramas.

Tales métodos simples son muy útiles para analizar conjuntos de datos de una dimensión no mayor a dos, pero cuando la dimensión de las variables respuesta crece, su habilidad para visualizar relaciones interdimensionales entre variables decrece, se presentan a continuación varios procedimientos que nos permitirán hacer análisis de datos multivariados, por medio de gráficos.

En la mayoría de los casos hacer análisis sobre datos sin que se pueda ver visualmente algún resultado analítico, resulta un tanto complicado, por lo cual es importante recurrir a realizar presentaciones visuales de los datos multivariados. Dos de los principales objetivos al hacer esto son los siguientes:

a) Permite a los investigadores localizar e identificar anormalidades, tales como datos atípicos (o conocidos como *outliers*, es decir observaciones que al parecer se generaron de

manera distinta al resto de las observaciones) que podrían existir en los datos y los cuales es casi imposible descubrir en un conjunto grande de datos sin situarlos en una gráfica

b) Percibir relaciones de dependencia, no necesariamente lineales, entre variables.
c) Detectar posibles agrupaciones entre los individuos u objetos.

La utilidad de aplicar técnicas exploratorias como las que se explicaran en este capitulo es que sirven de apoyo en el análisis de resultados obtenidos en la aplicación de otras técnicas multivariadas como el análisis de componentes principales o análisis de cluster, principalmente al detectar las posibles agrupaciones entre observaciones y la localización de outliers.

Debido a la complejidad de realizar graficas en mas de tres dimensiones, existen diferentes técnicas que permiten hacer representaciones graficas utilizando todas las variables al mismo tiempo, además de ser sencillas de realizar.

Antes de describir la aplicación de las técnicas, es importante señalar el porque se hace hincapié en la detección de *outliers* en un conjunto de datos. Como se menciono arriba, los *outliers* o datos atípicos son aquellas observaciones que parecen haberse generado de forma distinta al resto de los datos. Pueden ser causadas por mala captura de los datos, errores de medición, algún cambio de instrumento de medición o alguna heterogeneidad intrínseca de los elementos observados. Por ejemplo, supongamos que estamos estudiando las características de vivienda de una zona urbana, donde en su mayoría son casas-habitación pero en la muestra se incluyo una vivienda unifamiliar que cuenta con jardín. En este caso el dato *outlier* es la vivienda unifamiliar y la cual corresponde a una heterogeneidad real de los datos, por lo que es importante separar ambos tipos de vivienda para poder obtener una mejor descripción de los datos.

El no detectar *outliers* en un conjunto de datos nos puede conducir a distorsionar los resultados de aplicar las técnicas multivariadas. Aunque parezca insignificante, la aparición de un solo dato *outlier* en los datos, puede afectar al valor del vector de medias y a todas las varianzas y covarianzas entre las variables, además de que puede distorsionar arbitrariamente los coeficientes de correlación entre las variables, todo esto nos llevaría a no hacer un análisis real de los datos y por lo tanto inferir en resultados no verdaderos.

Pero ¿cómo identificamos *outliers* en un conjunto de datos?. Gráficamente estos datos se identifican cuando en un grupo de datos hay algunos puntos que se encuentran alejados del resto. Pero cuando hay mas de tres datos *outliers*, se pueden formar grupos de estos datos, de manera que aunque detectemos alguno de ellos y lo eliminemos, los otros se pueden ocultar entre si.

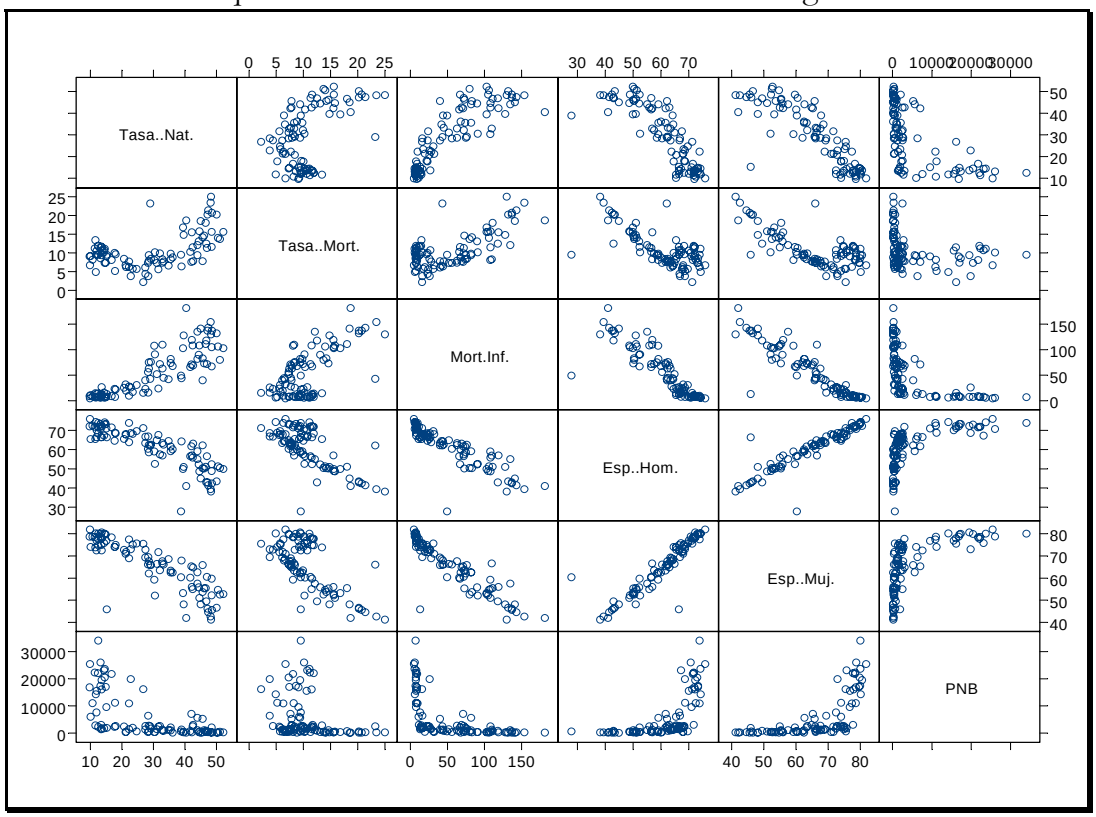
1.3.2 Graficas de dispersión.

El primer paso de cualquier análisis multivariado es representar gráficamente las variables de manera individual, mediante un histograma o algún diagrama alternativo, no entraremos en detalle en este tipo de representaciones, además de que este tipo de graficas por lo general es útil solo cuando tenemos una dimensión menor a tres.

En general las graficas de dispersión son muy útiles para detectar asimetrías, heterogeneidad, *outliers*, correlación entre las variables u otra característica de los datos.

La aplicación de esta técnica consiste en que una vez graficadas las variables de manera individual, es conveniente construir diagramas de dispersión de las variables por pares, los cuales se pueden representar en forma matricial para facilitar el análisis, pues permite visualizar el tipo de relación existente entre tales pares, además de que es mas fácil identificar *outliers* en la relación bivariante. En S-plus se puede representar matricialmente las graficas de dispersión, además que permite darle una mejor presentación de los datos de tal manera que se identifiquen y señalen los datos outliers, así como identificar posible correlación entre variables.

En particular estos gráficos son importantes para apreciar si existen relaciones no necesariamente lineales, en cuyo caso la matriz de covarianzas puede no ser un buen resumen de la dependencia entre las variables. Esta manera de construir graficas de dispersión a menudo es muy útil debido a que seria imposible graficar datos de mas de 3 dimensiones a la vez. Aunque en un estudio donde solo se involucren 3 variables es muy practico realizar graficas tridimensionales. Así mismo, para variables discretas, podemos construir diagramas de barras tridimensionales y para variables continuas los equivalentes multidimensionales de los histogramas.



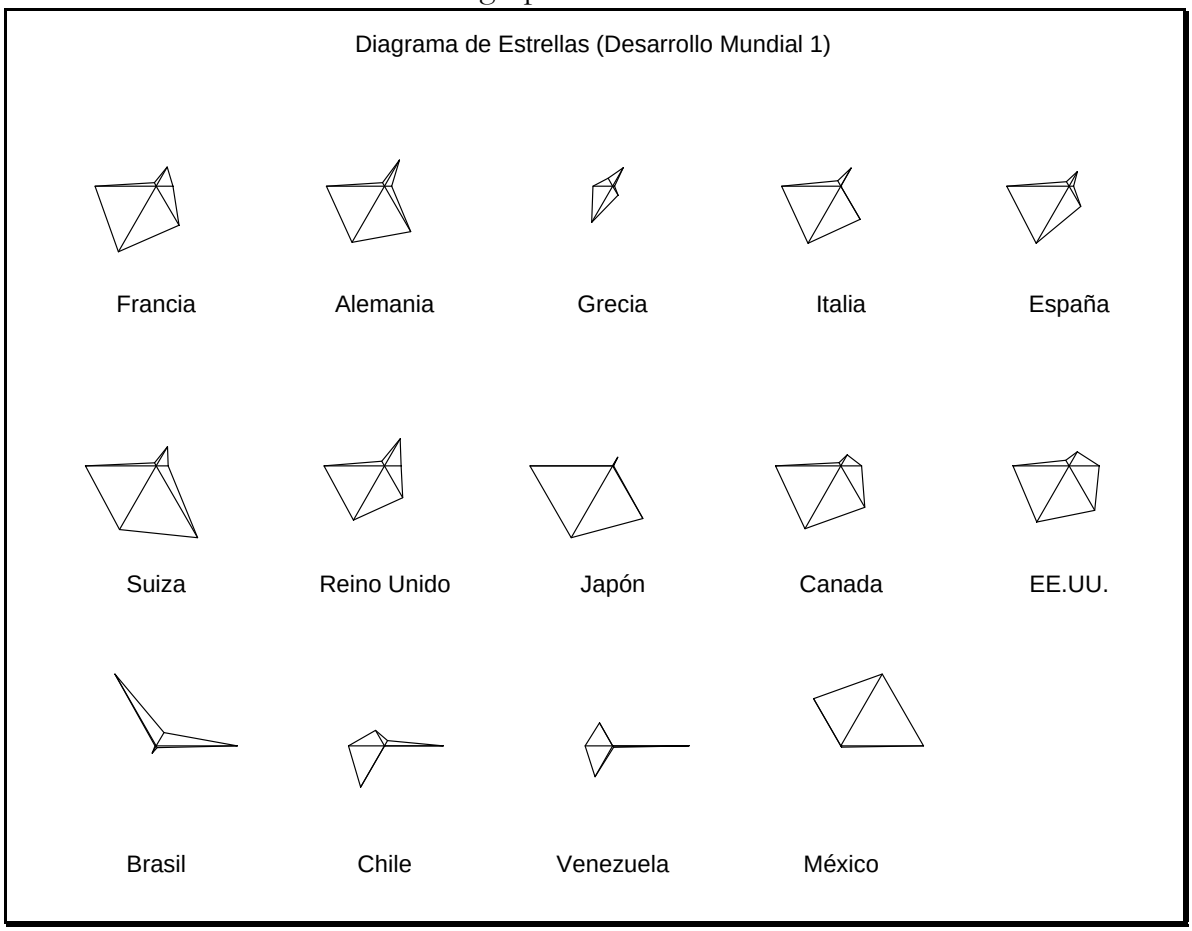
1.3.3 Diagrama de estrellas.

Las graficas de estrellas o grafica de rayos muestran las características de los datos de manera individual. Es decir, se hace una representación de cada observación respecto a las diversas variables mediante rayos que parten de un mismo origen. Este tipo de diagramas se aplica cuando las variables toman valores positivos.

Estas graficas se construyen al representar la distancia a la que se encuentra cada variable sobre rayos o ejes que irradia de un punto central.

El método consiste en graficar, dadas p variables, a partir de un rayo que apunta hacia el norte (correspondiente a una variable) el resto de las variables en sentido de las manecillas del reloj y manteniendo el mismo espacio entre uno y otro. Cada diagrama tendrá p rayos, uno por cada variable.

Estas graficas son de utilidad para detectar datos *outliers* y también sirven de apoyo en el análisis de resultados de otras técnicas de agrupación.



1.3.4 Caras de Chernoff.

Otra técnica grafica para explorar datos es la diseñada por Herman Chernoff en 1973, quien con el objeto de estudiar datos multidimensionales y debido a las limitaciones y dificultades que implicaba la representación gráfica, construyo una técnica con la cual se permite graficar datos con un numero grande de variables. Donde cada observación es convertida en una cara, y cada variables asignada a una característica de la misma. A pesar de que Chernoff logro representar 18 características en una cara, el S-Plus solo puede representar 15 características.

El procedimiento es simple pero efectivo. Este método consiste en graficar un conjunto multivariado en forma de caras de manera que cada característica facial se asocia con variables diferentes. Es decir, la forma y el tamaño de la cara, la posición de los ojos, la longitud y el ancho de la nariz, la longitud y ancho de las cejas, etc. son las que describen cada variable sobre una observación. Por su puesto entre más variables se empleen, son más las características de la cara las que se formaran. En particular en el S-Plus asigna las variables a las siguientes características: 1-

área de la cara; 2-forma de la cara; 3-longitud de la nariz; 4-localización de la boca; 5-curva de la sonrisa; 6-grosor de la boca; 7, 8, 9, 10, 11- localización, separación, ángulo, forma y grosor de los ojos; 12-localización de la pupila; 13, 14, 15-localización, ángulo y grosor de las cejas.

A menudo esta técnica es muy útil para identificar anormalidades en los datos, en particular datos atípicos (*outliers*), así como identificar posibles agrupaciones de datos, las cuales pueden servir de ayuda en la aplicación de otros técnicas de agrupación; es decir se busca dividir las unidades experimentales de un conjunto de datos en subgrupos, denominados *agrupamientos*, de tal manera que los individuos de dicho agrupamiento sean semejantes entre si.

Es importante mencionar que por lo general antes de aplicar esta técnica es preciso estandarizar los datos, ya que por lo general las variables se muestran correlacionadas y por medio de esta transformación pasamos de variables correlacionadas a variables incorrelacionadas con matriz de varianzas identidad.