# ESTADISTICA Y PROBABILIDAD

## IDEAS PREVIAS

- La ***estadística*** se ocupa básicamente , de la ***presentación e interpretación*** de datos que se dan en un estudio o relevamiento planeado , en una investigación científica , etc.

- La ***presentación*** de datos implica su **procesamiento** y la obtención de algunos **valores característicos**, y su ***interpretación*** debe permitir la **toma de decisiones** frente al tema que miden los datos.

En la mayoría de los problemas hay que tomar decisiones en base a ***experimentos*** :

- Un experimento especifica exactamente qué ensayos y pruebas han de realizarse y qué ha de observarse.

- Estos ensayos que llevan a resultados u observaciones , generalmente se repiten varias veces bajo condiciones uniformes o constantes.

- Los datos que se generan en estos ensayos se llaman ***resultados aleatorios*** pues aún cuando se tiene un gran cuidado en mantener las condiciones del experimento tan uniformes como sea posible, las observaciones individuales presentan una variabilidad intrínseca que no puede eliminarse

Por ejemplo, si un químico realiza varias veces un análisis bajo las mismas condiciones y obtiene diferentes mediciones ello indica la existencia de un elemento de aleatoriedad en el procedimiento experimental. Esta variabilidad inherente en las mediciones se conoce como ***error experimental*** que es un nombre conveniente para una ***fuente de variacion no controlada.*** Resulta que , en todo tipo de experimentos que se repiten bajo condiciones controladas, los resultados de las repeticiones individuales varían; por lo tanto, los resultados de cualquier repetición no pueden predecirse con exactitud.

En vez de ignorar esta variabilidad o tratarla cualitativamente, puede ***incorporarse*** al ***modelo matemático*** del fenómeno físico que se está estudiando; veremos cómo incluir esta variabilidad inherente en el modelo matemático. La estadística puede emplearse para describir y comprender la variabilidad.

En general los especialistas en estadística utilizan la palabra ***experimento*** para describir ***cualquier proceso que genere datos***, sean estos **numéricos** (que representan conteos o mediciones) o **categóricos** ( que brindan una información cualitativa , sexo , color ... ).

En la teoría de probabilidades el término experimento se usa para describir cualquier ***proceso cuyos resultados no se conocen de antemano con certeza.***

Aunque los resultados de cualquier experimento dado no pueden predecirse con exactitud , es posible ***caracterizar*** el conjunto de todos los resultados posibles del experimento.

La ***Estadistica*** actual es el resultado de la unión de dos disciplinas que evolucionaron independientemente hasta confluir , en el siglo XIX :

- El cálculo de las **probabilidades** , que nace en el siglo XVII como **teoría matemática** de los juegos de azar y ;

- La **estadística** ( **o ciencia del estado**, del latín Status) que estudia la descripción de los datos y tiene raíces más antiguas.

La **integración** de ambas líneas de pensamiento da lugar a una disciplina que estudia cómo obtener conclusiones de la investigación empírica mediante el uso de modelos matemáticos. La ***ESTADISTICA*** actúa como puente entre los **modelos matemáticos** y los **fenómenos reales**:

Un modelo matemático es una abstracción simplificada de una realidad más compleja, y siempre existirá cierta discrepancia entre lo observado y lo previsto por el modelo . La Estadística proporciona una metodología para evaluar y juzgar esas discrepancias entre la realidad y la teoría. Por lo tanto, su estudio es básico para aquellos que deseen trabajar en ciencia aplicada ; sea ésta Tecnología, Economía o Sociología; que requiera el análisis de datos y el diseño de experimentos.

*La Estadística es la tecnología del método científico experimental* ( Mood 1.972 ).

Además de su papel instrumental, el estudio de la Estadística es importante para entender las posibilidades y limitaciones de la investigación experimental, para diferenciar las conclusiones que pueden obtenerse de los datos de las que carecen de base empírica y, en definitiva para desarrollar un pensamiento crítico ante la realidad.

Todos los días tenemos a nuestro alcance una amplia diversidad de información numérica que se refiere a fenómenos como la actividad del mercado de valores , las tasas de desempleo, los descubrimientos médicos, los resultados de encuestas de opinión, los pronósticos del clima y la información deportiva. Con frecuencia esa información tiene gran impacto en nuestras vidas.

**La estadística moderna incluye la recopilación , presentación y caracterización de la información a fin de que auxilie tanto en el análisis de datos como en el proceso de toma de decisiones.**

Esto da lugar a que definamos:

- **Estadística Descriptiva** como los métodos que implican recopilación , presentación y caracterización de un conjunto de datos, con el objeto de describir en forma apropiada las diversas características de dicho conjunto.

Pero aunque los métodos de la estadística descriptiva son importantes para presentar y caracterizar información, lo que ha conducido a la amplia aplicación de la estadística en todos los campos de la investigación moderna ha sido el desarrollo de los métodos de la ***inferencia estadística*** como resultado de la teoría de la probabilidad. Surge entonces el concepto:

- **Inferencia Estadística** que se puede definir como los métodos que hacen posible la estimación de una característica de una población , o la toma de una decisión con respecto a una población , con base únicamente en **resultados muestrales.**

Resulta evidente que, para aclarar este concepto , se requieren algunas otras definiciones.

## Población y Muestra

Una *Poblacion* ( o *Universo* ) es la totalidad de elementos o cosas que se consideran o, según otra forma de definirla , una **población consiste en la totalidad de las observaciones en las cuales se está interesado.** El número de observaciones en la población se define como el **tamaño** de la población.

Frecuentemente no es posible estudiar a toda la población pues:

- El estudio puede implicar la destrucción del elemento , como es el caso de ensayos destructivos: por ejemplo estudiar la vida media de una partida de focos o la tensión de rotura de cables.

- Los elementos pueden existir conceptualmente , pero no en la realidad ; es decir que la población no puede ser físicamente listada . Por ejemplo la población de piezas defectuosas que producirá una máquina o la de personas portadoras de SIDA.
  Este es el caso de las llamadas *poblaciones infinitas* .

- Puede ser inviable económicamente estudiar la población.

En estas ocasiones seleccionaremos un conjunto de elementos de la población que llamaremos ***Muestra.***

Una ***Muestra*** es una porción o subconjunto de la población que se selecciona para análisis.

Según DIXON **población puede referirse o a elementos medibles o a las características medibles mismas.**

MENDENHALL , aclara que para la mayoría de las personas la palabra **muestra** tiene dos significados , el conjunto de objetos sobre el cual se hacen las mediciones o se puede referir a las mediciones . Similarmente a la palabra población que se puede usar con doble significado.

En general, usaremos la palabra muestra y población en sentido cotidiano y la mayor parte del tiempo nos referiremos a las mediciones hechas sobre las **unidades experimentales**.

**Unidad Experimental - Datos**

Una unidad experimental o unidad de observación es aquella sobre la cual se efectúan mediciones o se intenta clasificar en categorías . Pueden ser personas, familias , viviendas, células sanguíneas, plantas animales , tornillos , etc.

En el proceso de observación se registra por cada unidad experimental alguna característica y esta observación constituye un **dato.**

Por ejemplo , si el objetivo de una investigación consiste en realizar un estudio de ingresos familiares, la unidad de observación podría ser la familia, entendiéndose por ella a un grupo de personas que habitan juntas y comparten la comida.

El ingreso familiar medido sobre cada unidad de observación ( familia ) es un dato.

El conjunto de datos obtenidos de cada unidad de observación constituirá la base para el análisis estadístico del ingreso familiar.

**Parámetro - Estadístico**

Un *Parámetro* es una medida que se calcula para describir una característica de una población completa.

Un *Estadístico* es una medida que se calcula para medir una característica a partir de sólo una muestra.

Uno de los principales aspectos de la inferencia estadística es el proceso que consiste en utilizar estadísticos para obtener conclusiones acerca de los verdaderos parámetros de la población.

La teoría de la probabilidad proporciona el vínculo, determinando la probabilidad que los resultados provenientes de la muestra reflejen los resultados provenientes de la población.

Se pueden observar con claridad estas ideas en el ejemplo de la encuesta política .Si se desea estimar el porcentaje de votos que un candidato obtendrá en una elección específica, no entrevistará a cada uno de los votantes; más bien

- Seleccionará una muestra de los votantes.
- Con base en el resultado de la muestra obtendría conclusiones acerca de la población, el total de votantes
- A estas conclusiones se le asociaría un planteamiento de probabilidad que especifica la **esperanza** o confianza que se tiene de que los resultados de la muestra reflejen la verdadera conducta de la población.

## DESCRIPCIÓN ESTADÍSTICA DE UNA VARIABLE

Una *Variable* es cualquier característica que varía de una unidad experimental a otra en la población o en la muestra. Dado un conjunto de datos de una **variable X** (número de artículos defectuosos en una línea de producción, nacionalidad de las personas , nivel de ingresos, altura de las personas, etc ) ya hemos dicho que la estadística descriptiva estudia procedimientos para sintetizar la información que estos contienen.

### Tipos de Variables

Los tipos de variables que consideraremos son :

- *VARIABLES CATEGÓRICAS O CUALITATIVAS*

No toman valores numéricos y describen cualidades o atributos. Están definidas por las clases o categorías que las componen.

Son ejemplos : clasificar una pieza como Defectuosa o No Defectuosa, sexo, nacionalidad, el título universitario, distinguir a las personas en empleadas y desocupadas, etc.

- *VARIABLES NUMÉRICAS O CUANTITATIVAS*

Toman valores numéricos. Se pueden clasificar en **discretas y continuas.**

Las variables numéricas ***discretas*** toman valores que surgen de un ***proceso de conteo***

Las variables numéricas ***continuas*** toman valores en un intervalo que surgen de un ***proceso de medición*** de magnitudes continuas ( tiempo , longitud, etc. )

***Las variables numéricas discretas pueden surgir por la asignación de ciertos códigos numéricos a las categorías de las variables categóricas .***

***Una variable continua puede tomar infinitos valores en un rango dado.***

*"El número de revistas al que una persona está suscrita "*, es un ejemplo de una variable cuantitativa discreta ( puede tomar los valores 0 , 1 , 2 , 3... ).

La *"estatura de una persona "* es un ejemplo de variable cuantitativa continua puesto que la respuesta puede tomar cualquier valor dentro de un intervalo dependiendo del instrumento de medición. Podemos establecer la altura de un persona como 1.75; o 1,753; o 1,7532; según la precisión del instrumento de medición . Debido a que los instrumentos de medición que usan los investigadores no resultan lo suficientemente precisos como para percatarse de diferencias pequeñas es frecuente encontrar observaciones iguales en los datos , aún cuando la variable aleatoria sea realmente continua y podamos asegurar ,en teoría, que no existen dos personas que tengan exactamente la misma estatura.

### Escalas de Medición

Los datos recopilados pueden también describirse de acuerdo al **nivel de medición que se logre.**
Una *medición* es establecer números o categorías o códigos a las observaciones mediante *escalas* adecuadas. Las escalas se diferencian por *propiedades de orden y distancia.*

Los cuatro niveles de medición son del más débil al más fuerte: **escala nominal, escala ordinal , escala de intervalo y escala de razón.**

<u>Escala Nominal:</u>

Si los datos que se observan para una **variable cualitativa** simplemente se clasifican en distintas categorías que no implican orden y en consecuencia distancia , se tiene un nivel de medición nominal.

*Ejemplos:* sexo, ocupación, color de cabello.

**Escala Ordinal:**

Si los datos que se observan para una **variable cualitativa** se clasifican en categorías distintas en las que existe algún orden, se obtiene un nivel de medición ordinal. Dice que un valor observado que se clasifica en una categoría posee más la propiedad que se mide que alguna observación que se registra en otra categoría.

Podemos establecer una escala Ordinal : Menor- Mayor o Mayor - Menor.

*Ejemplos:*

Las personas y el hábito de fumar se pueden ordenar en las categorías : Fumadores Empedernidos, Fumadores Moderados y No Fumadores (Mayor - Menor)

Las personas y el nivel de educación podemos ordenarlas en : Educación Primaria, Secundaria, Terciaria y Universitaria.(Menor- Mayor)

En ningún caso sabemos con certeza cuanto mayor es una categoría con respecto a la otra pues no existe una medición de distancia.

Por lo general se supone que los datos que se obtienen para una **variable cuantitativa** se miden en escalas de intervalo o de razón, que constituyen los niveles más elevados de medición porque permiten discernir no sólo cuál de los valores es el mayor , sino por cuánto.

**Escala de Intervalo:**

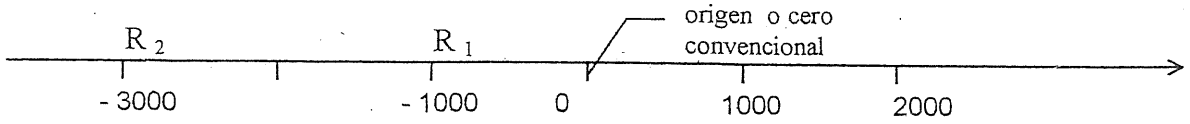
***Una escala de intervalo es una escala ordenada en la cual la diferencia entre mediciones es una cantidad significativa y tiene el mismo significado en cualquier parte de la escala.*** Es decir tiene la propiedad de asignar una medición de distancia entre los valores de la variable.

Una característica de la escala de intervalo es que el punto de origen o punto cero en la escala de medición es un punto de acuerdo o convencional y *es posible hacer comparaciones de distancia o diferencia entre mediciones pero no entre magnitudes proporcionales.*

*Ejemplos :*

Cuando se dice que una montaña mide 1000 mts simplemente se indica que ella se encuentra a 1000mts por encima del nivel del mar o punto origen de esa escala.

Si tomamos en cuenta la línea del tiempo empleada para ubicar los acontecimientos históricos. El tiempo a partir del nacimiento de Cristo se designa como años después de Cristo , con signo positivo y el tiempo anterior se considera con signo menos indicando que es antes del nacimiento de Cristo. El cero de la escala es el nacimiento de Cristo.




Si consideramos el tiempo actual como año 2000 después de Cristo,

Observemos que si $R_1$ es un acontecimiento histórico que sucedió 1000 años antes de Cristo , es decir hace 3.000 años y $R_2$ es otro acontecimiento ocurrido 3000 años antes de Cristo o hace 5.000 años podemos establecer que $R_2$ es anterior en 2000 años a $R_1$ y esto se puede determinar por la distancia entre sus respectivas ubicaciones en la línea del tiempo. No podemos decir , en cambio ,que $R_2$ es tres veces más antiguo que $R_1$ pues la línea del tiempo es una escala de intervalo y no de razón . En este caso ( $R_2 / R_1$ ) = 5000 / 3000 y no ( - 3000 / - 1000 ) ; $R_2$ es 5 / 3 más antiguo que $R_1$ .

Otro ejemplo es la temperatura medida en Grados Centígrados : 40 Grados es más caliente que 20 Grados y 20 Grados es más caliente que 0 Grado . En ambos casos la diferencia es 20 Grados y tiene el mismo significado. Sin embargo , **no** podemos decir que 40 Grados es el doble de caliente que 20 Grados pues el cero adoptado para medir las temperaturas no significa ausencia de temperatura, sino que es un valor arbitrario adoptado por convención.

**Escala de Razón:**

*Una escala de razón es una escala ordenada en la cual la diferencia entre mediciones son significativas e iguales en todos los puntos de la escala y existe un* cero real *que permita considerar cocientes de mediciones como los factores de proporcionalidad entre una medición y otra.*

Por ejemplo una persona que mide 1,80 m tiene el doble de estatura que una de 90 cm.

Si medimos la temperatura a partir del cero absoluto dado en la escala Kelvin ,la temperatura está dada en una escala de razón , puesto que si se duplica la temperatura , en realidad se duplica la velocidad promedio de las moléculas que componen la sustancia.

Para el caso del rendimiento de la producción de trigo por hectárea el cero es la ausencia de producción y se pueden efectuar comparaciones proporcionales entre distintos rendimientos.

El tipo de escala de medición que se utiliza al medir un variable condiciona el tratamiento estadístico que se efectúa con los datos .

## ESTADISTICA DESCRIPTIVA

Veremos a continuación los **métodos de la estadística descriptiva** que permiten resumir e interpretar los datos recopilados.

### Organización de los Datos Cualitativos

La presentación de datos cualitativos suele hacerse indicando las clases o atributos o categorías consideradas y su frecuencia de aparición como indica la TABLA 1. Esta misma idea se aplica para presentar datos cuantitativos discretos cuando el número de valores posibles es pequeño (menos de 10)

Ejemplo:

TABLA 1

AREAS DE PREFERENCIA ( Alumnos de Ciencias Económicas )

| AREA | Número de alumnos |
|---|---|
| ECONOMÍA | 17 |
| CONTABILIDAD | 12 |
| MATEMATICA | 11 |
| TOTAL | 40 |

La *variable* es *" Area de preferencia "* y no cada alumno. El trabajo lo tenemos que hacer sobre la variable área de preferencia , y no sobre el alumno que es el portador también de otros datos : edad, universidad a la que pertenece, sexo, carrera, etc. La estadística nos brindará elementos para estudiar los datos obtenidos.

La tabla que asocia cada categoría de la variable con el número de veces que se repite dicha categoría se llama **distribución de frecuencias de la variable,** cualitativa en nuestro caso.

A partir de las frecuencias indicadas en la tabla o **frecuencias absolutas** , podemos obtener las **frecuencias relativas** , cuya suma es igual a **uno** , y que se obtienen dividiendo la frecuencia de cada clase por el número total de datos . A las frecuencias relativas las indicamos $f_r$. Las frecuencias absolutas se identifican con **f** .

Otro ejemplo es la temperatura medida en Grados Centígrados : 40 Grados es más caliente que 20 Grados y 20 Grados es más caliente que 0 Grado . En ambos casos la diferencia es 20 Grados y tiene el mismo significado. Sin embargo , **no** podemos decir que 40 Grados es el doble de caliente que 20 Grados pues el cero adoptado para medir las temperaturas no significa ausencia de temperatura, sino que es un valor arbitrario adoptado por convención.

**Escala de Razón:**

*Una escala de razón es una escala ordenada en la cual la diferencia entre mediciones son significativas e iguales en todos los puntos de la escala y existe un* cero real *que permita considerar cocientes de mediciones como los factores de proporcionalidad entre una medición y otra.*

Por ejemplo una persona que mide 1,80 m tiene el doble de estatura que una de 90 cm.

Si medimos la temperatura a partir del cero absoluto dado en la escala Kelvin ,la temperatura está dada en una escala de razón , puesto que si se duplica la temperatura , en realidad se duplica la velocidad promedio de las moléculas que componen la sustancia.

Para el caso del rendimiento de la producción de trigo por hectárea el cero es la ausencia de producción y se pueden efectuar comparaciones proporcionales entre distintos rendimientos.

El tipo de escala de medición que se utiliza al medir un variable condiciona el tratamiento estadístico que se efectúa con los datos .

## ESTADISTICA DESCRIPTIVA

Veremos a continuación los **métodos de la estadística descriptiva** que permiten resumir e interpretar los datos recopilados.

### Organización de los Datos Cualitativos

La presentación de datos cualitativos suele hacerse indicando las clases o atributos o categorías consideradas y su frecuencia de aparición como indica la TABLA 1. Esta misma idea se aplica para presentar datos cuantitativos discretos cuando el número de valores posibles es pequeño (menos de 10)

Ejemplo:

TABLA 1

AREAS DE PREFERENCIA ( Alumnos de Ciencias Económicas )

| AREA | Número de alumnos |
|---|---|
| ECONOMÍA | 17 |
| CONTABILIDAD | 12 |
| MATEMATICA | 11 |
| TOTAL | 40 |

La *variable* es *" Area de preferencia "* y no cada alumno. El trabajo lo tenemos que hacer sobre la variable área de preferencia , y no sobre el alumno que es el portador también de otros datos : edad, universidad a la que pertenece, sexo, carrera, etc. La estadística nos brindará elementos para estudiar los datos obtenidos.

La tabla que asocia cada categoría de la variable con el número de veces que se repite dicha categoría se llama **distribución de frecuencias de la variable,** cualitativa en nuestro caso.

A partir de las frecuencias indicadas en la tabla o **frecuencias absolutas** , podemos obtener las **frecuencias relativas** , cuya suma es igual a **uno** , y que se obtienen dividiendo la frecuencia de cada clase por el número total de datos . A las frecuencias relativas las indicamos $f_r$. Las frecuencias absolutas se identifican con $f$ .

Si a las frecuencias relativas las multiplicamos por cien , obtenemos las **frecuencias relativas porcentuales** , que suman **100** y que indicamos **100 $f_r$ (%) .**

Para nuestra **TABLA 1** , resulta la siguiente **Tabla de Frecuencias o Distribución de frecuencias** para la variable cualitativa en cuestión:

**TABLA DE FRECUENCIAS**

AREAS DE PREFERENCIA ( Alumnos de Ciencias Económicas )

| AREA | Número de alumnos **f (frecuencia absoluta)** | **$f_r$ (frecuencia relativa)** | **100$f_r$(%)($f_r$ porcentual )** |
|---|---|---|---|
| ECONOMIA | 17 | 0,425 | 42,5 |
| CONTABILIDAD | 12 | 0,300 | 30,0 |
| MATEMATICA | 11 | 0,275 | 27,5 |
| TOTAL | 40 | 1,000 | 100,0 |

**Gráficos para las distribuciones de frecuencias :**

Las distribuciones de frecuencias para datos categóricos pueden representarse gráficamente por medio de un **gráfico de barras**, fácil de construir que puede ser interpretado por personas que no tienen una mente organizada hacia los gráficos. Un gráfico de barras presenta las posibles categorías y sus frecuencias.

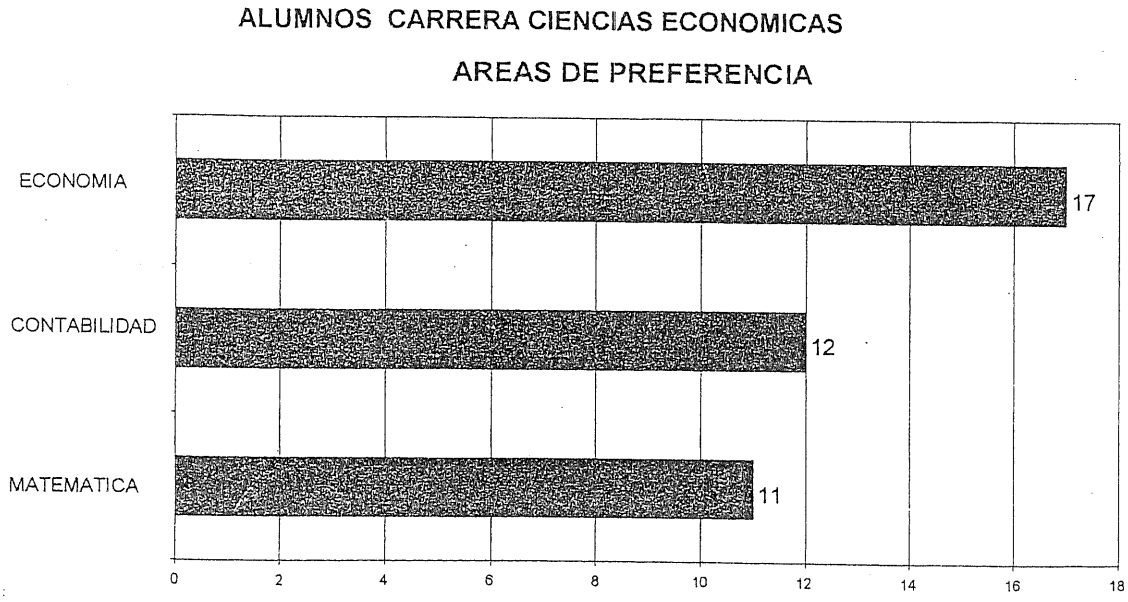
Las distribuciones de frecuencias para datos cualitativos también pueden representarse gráficamente mediante una **gráfica porcentual de tortas** en donde se representan las categorías y las frecuencias relativas porcentuales. Sin embargo se puede utilizar la **gráfica porcentual de puntos** que es la más actual.

El propósito de estas gráficas es mostrar los datos en forma precisa y clara. En nuestro ejemplo en particular , estas figuras pretenden mostrar la misma información respecto a las preferencias de los estudiantes de Ciencias Económicas.

Algunas investigaciones recientes sobre percepción humana de gráficas sugieren que la gráfica de puntos presenta la información de la mejor manera, en tanto que la gráfica de tortas es el tipo más deficiente. No obstante la selección de una gráfica sigue siendo una actividad muy subjetiva y, con frecuencia depende de las preferencias estéticas del investigador.

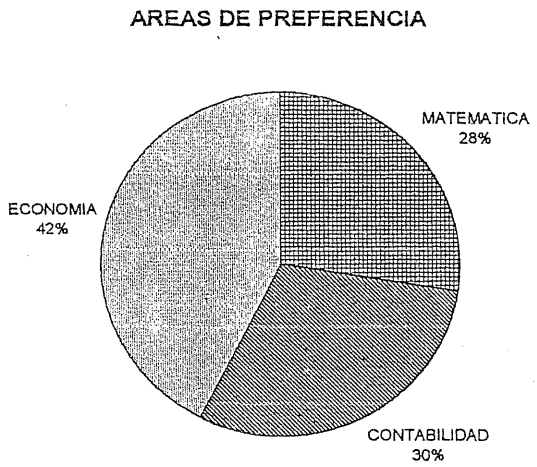
Observemos las gráficas resultantes para nuestro ejemplo:

**Gráfico de barras**



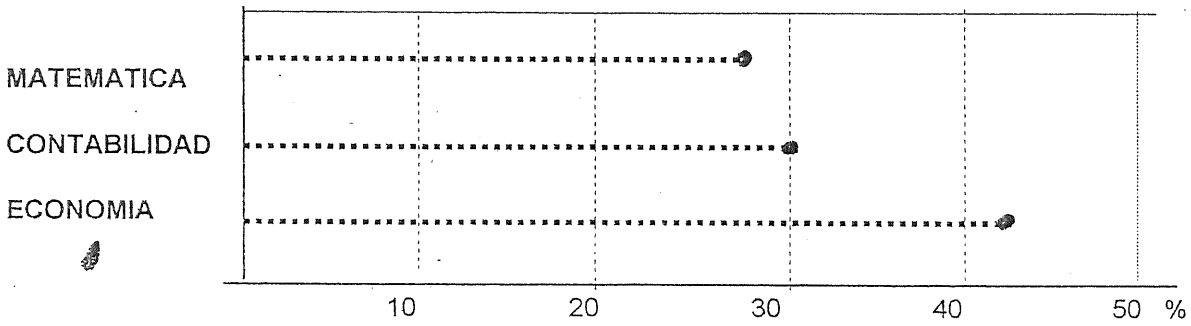
### Gráfico Porcentual de tortas

ALUMNOS CARRERA CIENCIAS ECONOMICAS




### Gráfico porcentual de puntos

AREAS DE PREFERENCIA ( CARRERA CIENCIAS ECONOMICAS )




En los gráficos de barras las magnitudes de frecuencias se muestran mediante las longitudes de las diferentes barras, las que se han trazado con referencia a una escala horizontal. Los valores de la escala se muestran en la parte superior y también en la parte inferior. Se pueden unir o no mediante líneas constituyendo una retícula.

Las barras se identifican mediante títulos en el **talón** de la gráfica. El talón debe mantenerse tan pequeño como sea posible, con el objeto de dejar espacio para la superficie de trazado de la gráfica.

En lo que respecta a las barras en sí, observamos que:

- Las barras difieren en longitud y no en ancho.
- Se ha dejado un espacio entre las barras para que sea fácil identificarla mediante su título.
- Las barras han sido ordenadas por magnitud para facilitar el análisis. El orden puede ser creciente o decreciente. Si se incluye una categoría "todas las demás" o "otras" , generalmente se la muestra en la barra más baja, aún cuando esta posición no coincida con la ordenación por magnitud.
- Para respuestas categóricas o cualitativas las barras se disponen en forma horizontal, mientras que para respuestas numéricas en forma vertical.

- **Importancia de la línea cero :** Una gráfica que ha sido diseñada para mostrar magnitudes absolutas debe tener absolutamente definida la línea cero y una escala ininterrumpida. La impresión percibida si se comienza por algún valor mayor que cero cambia totalmente el mensaje del gráfico.

## DATOS CUANTITATIVOS

### Análisis Exploratorio de Datos ( A E D )

El **A E D** es un grupo de técnicas destinadas a procesar lotes de datos con el objeto de detectar estructuras, sugerir hipótesis y facilitar un posterior análisis "confirmatorio" que se encargará de evaluar sistemáticamente las estructuras o efectos observados. Cabe acotar que el término **lote** se usa en un sentido técnico como un conjunto de números , sin aclarar si estos constituyen una población o una muestra. De las técnicas básicas del **A E D** presentaremos el **Diagrama de Tallos y Hojas**

Previamente vamos a definir el concepto de **lote ordenado** y la **mediana** .

### Lote ordenado - Mediana.

Tomemos el caso del siguiente lote de 54 datos correspondientes a las calificaciones obtenidas por los estudiantes en el curso de Estadística del año 1997. La escala adoptada son los enteros del 1 al 100.

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 01 | 27 | 46 | 40 | 25 | 85 | 75 | 67 | 70 | 23 | 28 | 20 | 44 | 04 | 08 |
| 61 | 37 | 68 | 68 | 41 | 25 | 48 | 78 | 15 | 69 | 05 | 98 | 70 | 70 | 37 |
| 26 | 49 | 60 | 62 | 45 | 53 | 55 | 54 | 10 | 15 | 43 | 43 | 15 | 72 | 63 |
| 61 | 60 | 60 | 60 | 60 | 50 | 27 | 19 | 12 | | | | | | |

Estos datos se pueden ordenar de menor a mayor. La sencilla operación de ordenar por magnitud un lote de datos, implica, por lo menos los siguiente:

- Apreciar mejor algunas cualidades del lote, por ejemplo, entre qué valores oscilan, dónde se concentran , etc.-
- Despojar a los números de la asociación que pueden tener con la fuente que los originó y comenzar a analizarlos como meros números.

El **lote de datos ordenados**, en este caso resulta

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 01 | 04 | 05 | 08 | 10 | 12 | 15 | 15 | 15 | | | | | | |
| 19 | 20 | 23 | 25 | 25 | 26 | 27 | 27 | 28 | | | | | | |
| 37 | 37 | 40 | 41 | 43 | 43 | 44 | 45 | **46** | | | | | | |
| **48** | 49 | 50 | 53 | 54 | 55 | 60 | 60 | 60 | | | | | | |
| 60 | 60 | 61 | 61 | 62 | 63 | 67 | 68 | 68 | | | | | | |
| 69 | 70 | 70 | 70 | 72 | 75 | 78 | 85 | 98 | | | | | | |

La **mediana** de un lote de datos ordenados lo divide en dos partes iguales : los datos de una parte son menores o iguales a la mediana , y los de la otra son mayores o iguales a la mediana.

Para todo lote de una cantidad impar de datos ordenados la mediana es el valor central y para todo lote de una cantidad par de datos ordenados , la mediana es el promedio de los dos valores centrales.

En el ejemplo, la mediana corresponde al promedio del par de datos centrales . En este caso los datos que ocupan las posiciones veintisiete y veintiocho, los números 46 y 48 respectivamente . Resulta:

mediana = ( 46 + 48 ) / 2 = 47 $\Rightarrow$ $M_E = 47$

**Fórmula de posicionamiento de la mediana = ( N + 1 ) / 2**

**Diagrama de tallos y hojas**

Un lote de datos se puede organizar gráficamente mediante un diagrama de tallo y hojas. Este sencillo diagrama que puede construirse manualmente, facilita la observación del lote completo; además, a partir de él se puede ver, entre otras cosas, el recorrido de los valores de los datos, la simetría del lote , la presencia de valores distantes , dónde se concentran los datos, etc.

Para el lote de datos ordenados ( 54 calificaciones ), en una inspección rápida , se ve que hay valores del orden de los 10 , 20 , . . . ,90. El primer dígito de cada valor, *dígito principal*, debe usarse como *TALLO* y el segundo , *dígito secundario* , como *HOJA*

Si hubiera tres dígitos, por ej. para el número 342 podrían tomarse: 34 como dígitos principales y 2 como dígito secundario. También es usual construir un diagrama de tallos y hojas truncado: 3 es el dígito principal ( tallo ) y 4 el dígito secundario truncando la tercera cifra del número.

Se escriben en una columna todos los valores posibles para los tallos ordenados de menor a mayor y luego se indican las hojas en las líneas correspondientes de acuerdo al tallo:

| | |
|---|---|
| 0 | 1 4 5 8 |
| 1 | 0 2 5 5 5 9 |
| 2 | 0 3 5 5 6 7 7 8 |
| 3 | 7 7 |
| 4 | 0 1 3 3 4 5 6 8 9 |
| 5 | 0 3 4 5 |
| 6 | 0 0 0 0 0 1 1 2 3 7 8 8 9 |
| 7 | 0 0 0 2 5 8 |
| 8 | 5 |
| 9 | 8 |

Completamos entonces el diagrama en donde las hojas ya están ordenadas de menor a mayor:

Diagrama de tallo y hojas: CALIFICACIONES alumnos del curso de ESTADISTICA en el año 1997.

| Profundidades | ( Unidad = 1 ) | |
|---|---|---|
| 4 | 0 | 1 4 5 8 |
| 6 | 1 | 0 2 5 5 5 9 |
| 8 | 2 | 0 3 5 5 6 7 7 8 |
| 2 | 3 | 7 7 |
| ( 9 ) | 4 | 0 1 3 3 4 5 6 8 9 |
| 4 | 5 | 0 3 4 5 |
| 13 | 6 | 0 0 0 0 0 1 1 2 3 7 8 8 9 |
| 6 | 7 | 0 0 0 2 5 8 |
| 1 | 8 | 5 |
| 1 | 9 | 8 |
| n = 54 | | |

Se ha indicado la **unidad empleada para las hojas** e incluido una columna de profundidades a la izquierda de los tallos .

Las profundidades indican el número de hojas que hay desde el extremo más próximo del lote en esa línea , excepto en la línea que contiene a la mediana donde se indican entre paréntesis la cantidad de hojas que hay en ella .Esta distinción no se hace en el caso en que el número de los datos es par y la mediana se encuentra entre dos líneas.

Cuando se observan muchas hojas en cada línea , existe la posibilidad de dividir las líneas repitiendo los tallos .

- Se pueden considerar dos líneas por cada tallo : en la primera línea , que se indica con " * ", se colocan las hojas 0 , 1 , 2 , 3 y 4 ; en la segunda , señalada mediante un " . " , los dígitos entre 5 y 9.

En este caso el ancho del intervalo es 5 veces una potencia de 10.

- Otra opción es considerar cinco líneas por cada tallo. Para este caso una notación propuesta es , si tomamos el tallo 1 :

| | |
|---|---|
| 1 * | en la línea con " * " se ubican las hojas 0 y 1 . |
| t | en la línea con la letra " t " , se ubican las hojas 2 y 3 ( two, three ) |
| f | en la línea con la letra " f " , se ubican las hojas 4 y 5 ( four, five ) |
| s | en la línea con la letra " s " , se ubican las hojas 6 y 7 ( six, seven ) |
| 1 · | en la línea con " .", se ubican las hojas 8 y 9. |

Aquí el ancho del intervalo es 2 veces una potencia de 10.

También podemos emplear el diagrama de tallo y hojas *para ayudarnos a ordenar* un lote. Indicamos el tallo (dígito principal) y luego le agregamos las hojas correspondientes en el orden que aparecen en el lote. Luego solo resta ordenar las hojas de menor a mayor y reconstruir, si fuera necesario el lote ordenado.

## ORGANIZACIÓN DE DATOS PARA VARIABLES CUANTITATIVAS

### Distribución de frecuencias para *datos NO agrupados*

Los datos cuantitativos **discretos**, cuando el número de valores posibles es pequeño ( <10 ), pueden presentarse mediante una TABLA de **Distribución de Frecuencias de la Variable.**

En esta tabla se indican las Frecuencias Absolutas ( $f$ ), Relativas ( $f_r$ ) y Relativas Porcentuales ( $100\, f_r$ ( % ) ) que ya hemos estudiado. Se ha agregado una columna de **Frecuencias Acumuladas** ( $f_a$ ) a continuación de la columna de Frecuencias Absolutas y una columna con **Frecuencias Acumuladas Porcentuales** ( $100\, f_a$ ( % ) ) a continuación de la columna correspondiente a la de Frecuencias Relativas Porcentuales .

Las frecuencias acumuladas se obtienen para cada valor de la variable sumando las frecuencias absolutas correspondientes a los valores menores o iguales al valor considerado.
Así , para el valor $x_j$ de la variable $X$ , donde $X$ toma los $k$ valores ordenados $x_1 , x_2 , \ldots x_J , \ldots x_K$ , con frecuencias absolutas $f_1 , f_2 , \ldots f_J , \ldots , f_k$ respectivamente , la **frecuencia acumulada** para $x_J$ es

$$f_{aj} = \sum_{i \le j} f_i = f_1 + f_2 + \ldots + f_J$$

Las frecuencias acumuladas porcentuales se obtienen sumando las frecuencias relativas porcentuales correspondientes a los valores de la variable menores o iguales al valor considerado. También pueden obtenerse multiplicando por 100 las frecuencias acumuladas correspondientes, a cada valor de la tabla.

Es decir la **frecuencia acumulada porcentual** para $x_J$ es : $f_{aj}\,(\%) = 100\, f_{aj} = \sum_{i \le j} 100\, f_{ri}$

*Ejemplo*: Para 20 familias clasificadas por el número de hijos, resulta la siguiente TABLA:

**DISTRIBUCION DE FRECUENCIAS**
CANTIDAD DE HIJOS POR FAMILIA

| NÚMERO DE HIJOS | f | $f_a$ | $f_r$ | 100 $f_r$ (%) | 100 $f_a$ (%) |
|---|---|---|---|---|---|
| 0 | 4 | 4 | 0,2 | 20 | 20 |
| 1 | 8 | 12 | 0,4 | 40 | 60 |
| **2** | 4 | **16** | 0,2 | 20 | **80** |
| 3 | 2 | 18 | 0,1 | 10 | 90 |
| 4 | 2 | 20 | 0,1 | 10 | 100 |
| TOTAL | 20 | | 1,0 | 100 | |

En la a columna de **Frecuencias Acumuladas**, para el renglón resaltado , se interpreta " hasta **2** hijos **16** familias. En la columna **Frecuencias Acumuladas Porcentuales ( 100 $f_a$ ( % ) )** , para el renglón resaltado se interpreta " hasta **2** hijos el **80 %** de las familias "

## Distribución de Frecuencias para datos AGRUPADOS

Cuando el número de valores posibles de una *variable DISCRETA* sea grande o cuando la variable sea *CONTINUA* conviene agrupar los datos en **clases o categorías**. Es decir, independientemente de la selección de un lote ordenado o de un diagrama de tallo y hojas para organizar los datos , al incrementarse el número de observaciones , se hace necesario condensar los datos en tablas apropiadas de resumen . Para ello se acomodan los datos en *grupos de clases* , es decir categorías , dividiendo en forma conveniente las observaciones. A este arreglo de datos en forma de tabla se le denomina , al igual que para variables cualitativas y variables discretas para k pequeño, **Distribución de frecuencias .**

Una **Distribución de Frecuencias para Datos Agrupados** es una tabla resumen en la que se disponen los datos divididos en grupos ordenados numéricamente que se denominan clases o categorías.

Cuando se agrupan datos, o se los condensa en tablas de Distribución de Frecuencias, es más manejable y significativo el proceso de análisis e interpretación de datos. En esa forma resumida es muy sencillo aproximar las principales características de los datos y de esta manera se compensa el hecho de que al agrupar los datos se pierde alguna información inicial referente a las observaciones individuales.

Al construir una tabla de Distribución de Frecuencias, se debe prestar atención a lo siguiente:

- Seleccionar el número adecuado de clases para cada tabla.
- Obtener un intervalo de clase apropiado para cada clase.
- Seleccionar los límites de clase que definen los intervalos, de manera que las clases sean de la misma longitud y cada observación se clasifique sin ambigüedad en una sola clase.

### Criterios para la elección de la amplitud de los intervalos

Para determinar la amplitud de los intervalos, comenzamos por determinar el número de clases o categorías a considerar. Damos a continuación algunos criterios empíricos

- Hay una vieja fórmula para ello; si **n** es el tamaño del lote , se trata de hallar **K** , tal que:

$$[K] = \text{Nro de intervalos} \quad , \quad \text{si } K \text{ satisface la relación: } \quad n \sim 2^{(K-1)}$$

De esta expresión resulta: $K \sim 1 + 3{,}322 \log n$ . ( $[K]$ representa la parte entera de $K$ )

- Un criterio generalizado es tomar un número de intervalos de clase entre 5 y 20 ,dependiendo de los datos.

- Una regla frecuentemente usada consiste en tomar el número de clases igual al entero más próximo a $2\sqrt{n}$ , siendo n el número de datos.

- Podemos emplear el diagrama de tallos y hojas ya que cada uno de los tallos define una clase. En este caso obtenemos la distribución de frecuencias para los datos contando simplemente las hojas correspondientes a cada tallo. Se debe tener la precaución que el diagrama de tallos y hojas empleado no presente los datos demasiados concentrados ni demasiado dispersos para que el números de categorías a proponer sea adecuada.

Para determinar la amplitud $h$ de cada intervalo hacemos $h = \text{OSCILACION} / K$

$K$ es el número de categorías y la **OSCILACION ( o RANGO )** es la distancia entre el valor máximo y el valor mínimo.

Siempre hay que tener presente que la determinación de las categorías debe condicionarse a que la los límites de cada categoría puedan indicarse con números o expresiones sencillas.

**Marca de clase**

Definimos cada clase o categoría mediante un intervalo de clase expresado en la forma

$$x_i - h/2 \quad - \quad x_i + h/2$$

Donde el punto medio $x_i$, se llama *MARCA DE CLASE* . Este valor es el centro del intervalo que define la clase y es el valor numérico representativo de los datos de la clase.

$x_i - h/2$ es el *límite inferior* de la clase y $x_i + h/2$ es el *límite superior* .

Las maneras de determinar la clase definida por $x_i$ son:

- Desde $x_i - h/2$ **inclusive , hasta menos de** $x_i + h/2$ . Diremos que **el dato** $x_j$ **pertenece a esta clase sii** $x_i - h/2 \leq x_j < x_i + h/2$ . Como vemos , en cada intervalo de clase se incluye al límite inferior.

- **Desde más de** $x_i - h/2$ **hasta menos o inclusive** $x_i + h/2$ . Diremos que **el dato** $x_j$ **pertenece a esta clase sii** $x_i - h/2 < x_j \leq x_i + h/2$ . En este caso , en cada intervalo de clase se incluye al límite superior.

Puede adoptarse cualquiera de estos dos criterios para ubicar los valores que coinciden con los límites de cada clase pero éste debe mantenerse a través de todo el proceso de agrupamiento.

Para las **CALIFICACIONES de los alumnos del curso de ESTADISTICA en el año 1.997** con el criterio de computar en cada clase los valores correspondientes a su límite superior se obtiene la siguiente **Tabla de Frecuencias .**

Este ejemplo corresponde a una *VARIABLE CUANTITATIVA DISCRETA* pero dado que el número de registros diferentes es supera ampliamente el valor 10 recurrimos al agrupamiento en intervalos de clases. De este modo la variable se trabaja de la misma forma que emplearíamos si fuera una variable continua.

Con una flecha indicamos que el límite superior de cada clase está incluido . Otra forma de indicar cada clase es mediante la expresión mas de .... y hasta .....

TABLA DE FRECUENCIAS
CALIFICACIONES

| INTERVALO DE CLASE | MARCA DE CLASE | F Frecuencia Absoluta | $f_a$ Frecuencia acumulada | $f_r$ Frecuencia relativa | $f_r$ ( % ) Frec. Rel. porcentual | $F_a$ ( % ) Frec. Ac. porcentual |
|---|---|---|---|---|---|---|
| 00 – 10 | 5 | 5 | 5 | 0,093 | 9,3 | 9,3 |
| 10 – 20 | 15 | 6 | 11 | 0,111 | 11,1 | 20,4 |
| 20 – 30 | 25 | 7 | 18 | 0,130 | 13 | 33,4 |
| 30 – 40 | 35 | 3 | 21 | 0,055 | 5,55 | 38,95 |
| 40 – 50 | 45 | 9 | 30 | 0,167 | 16,7 | 55,65 |
| 50 – 60 | 55 | 8 | 38 | 0,148 | 14,8 | 70,45 |
| 60 – 70 | 65 | 11 | 49 | 0,204 | 20,4 | 90,85 |
| 70 – 80 | 75 | 3 | 52 | 0,055 | 5,55 | 96,40 |
| 80 – 90 | 85 | 1 | 53 | 0,018 | 1,80 | 98,20 |
| 90 – 100 | 95 | 1 | 54 | 0,018 | 1,80 | 100 |
| TOTAL | | 54 | | 1,000 | 100 | |

## Gráficos para las distribuciones de frecuencias

### Histogramas

La información que proporciona una distribución o tabla de frecuencias para este tipo de datos agrupados es más fácil de entender si se presenta en forma gráfica mediante un diagrama llamado **histograma**.

Un histograma es un conjunto de rectángulos o barras verticales cada una de los cuales representa un intervalo de agrupación o clase . Sus bases están centradas en la marca de clase y son iguales a la amplitud del intervalo y las alturas se determinan de manera que su área sea proporcional a la frecuencia ( o frecuencia relativa , o frecuencia relativa porcentual ). Sobre el eje vertical puedo indicar frecuencias , frecuencias relativas o frecuencias relativas porcentuales y obtendré los correspondientes histogramas de frecuencias, frecuencias relativas o porcentual.

A continuación representamos el *histograma de frecuencias* asociado al ejemplo de datos agrupados y el histograma correspondiente al ejemplo de datos no agrupados ( 20 familias clasificadas por el número de hijos ) para una variable discreta. Para este caso puede resultar apropiada la representación de un diagrama por puntos .

### El Diagrama de Tallo y Hojas como Distribución de Frecuencias e Histograma.

El Diagrama de Tallo y Hojas organiza los datos de forma tal que permite simultáneamente realizar el análisis más detallado y brinda una presentación en forma tanto tabular como gráfica. El diagrama reúne tres aspectos un arreglo ordenado ,una distribución de frecuencias y un histograma sin sacrificar la información original que se refiere a las observaciones individuales mismas.

- Si se juntaran las hojas de los diferentes tallos en forma consecutiva se obtiene un arreglo ordenado
- Si sólo se registraran las hojas correspondientes a cada tallo, se construirá una distribución de frecuencias
- Si se girara 90 ° el diagrama de tallo y hojas , se ilustraría un diagrama de frecuencias , polígono o histograma, de manera que los puntos o barras verticales estarían representados por las hojas individuales de cada tallo.

### Polígonos de Frecuencias

Se obtienen polígonos de frecuencia , si asumimos que el punto medio de cada clase representa los datos de esa clase ( número de observaciones , proporción o porcentaje ) y conectamos los puntos correspondientes a cada clase secuencialmente mediante segmentos de recta.

Estos polígonos son particularmente útiles cuando se comparan dos o más conjuntos de datos. Por ejemplo para las calificaciones de los distintos años de los alumnos de ESTADISTICA . En este caso convendrá emplear proporciones o porcentajes ( frecuencias relativas o frecuencias relativas porcentuales ) para independizarnos del número de calificaciones que dependerá del número de alumnos en cada curso o del número de evaluaciones rendidas.

A continuación obtenemos el polígono correspondiente al ejemplo de las Calificaciones.

**Polígono Porcentual Acumulado ( Ojiva )**

Se obtiene para variables cuantitativas con datos agrupados.

Para construirlo el fenómeno de interés se representa sobre el eje horizontal, en tanto que los porcentajes ( o frecuencias relativas ) sobre el eje vertical.
En cada uno de los límites superiores se busca el valor del porcentaje acumulado correspondiente, después se conectan esos puntos por segmentos de líneas rectas.

Si efectuamos esta representación para las calificaciones obtenemos para cada punto del gráfico el porcentaje de notas menores o iguales al valor en el eje de abscisas. Para determinar cuál es el valor mediano de estas calificaciones, es decir el valor para el que las calificaciones que lo superen resultan iguales a las que son inferiores , buscamos el valor correspondiente a un porcentaje acumulado del 50% de calificaciones.

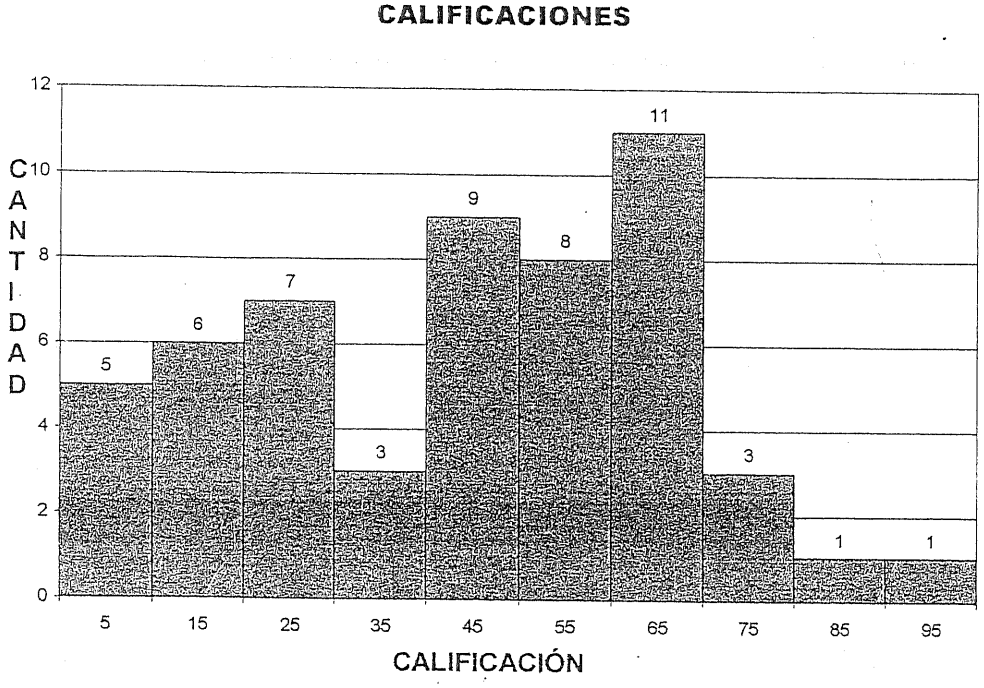
Se agrega la representación correspondiente del polígono porcentual acumulado para el ejemplo.

**Gráfico a Escalones**

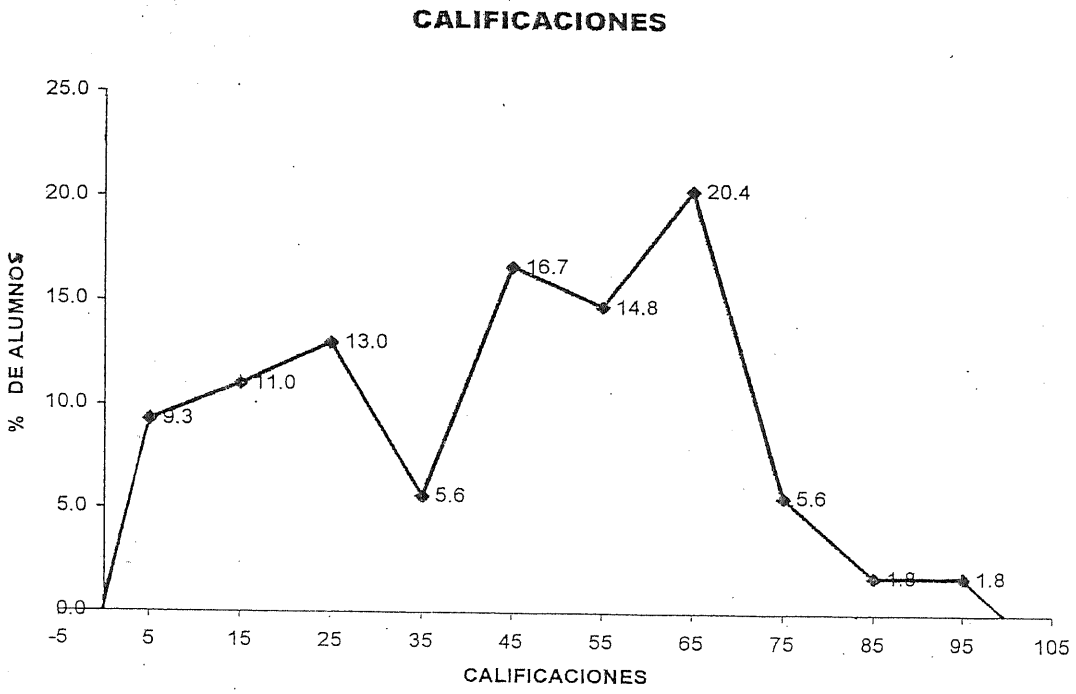
En el caso de variables cuantitativas discretas no agrupadas , si representamos las frecuencias relativas acumuladas obtenemos un gráfico a escalones . Esto lo observamos en la representación correspondiente al ejemplo de 20 familias agrupadas por el número de hijos .

CALIFICACIONES ALUMNOS ESTADISTICA – 1997

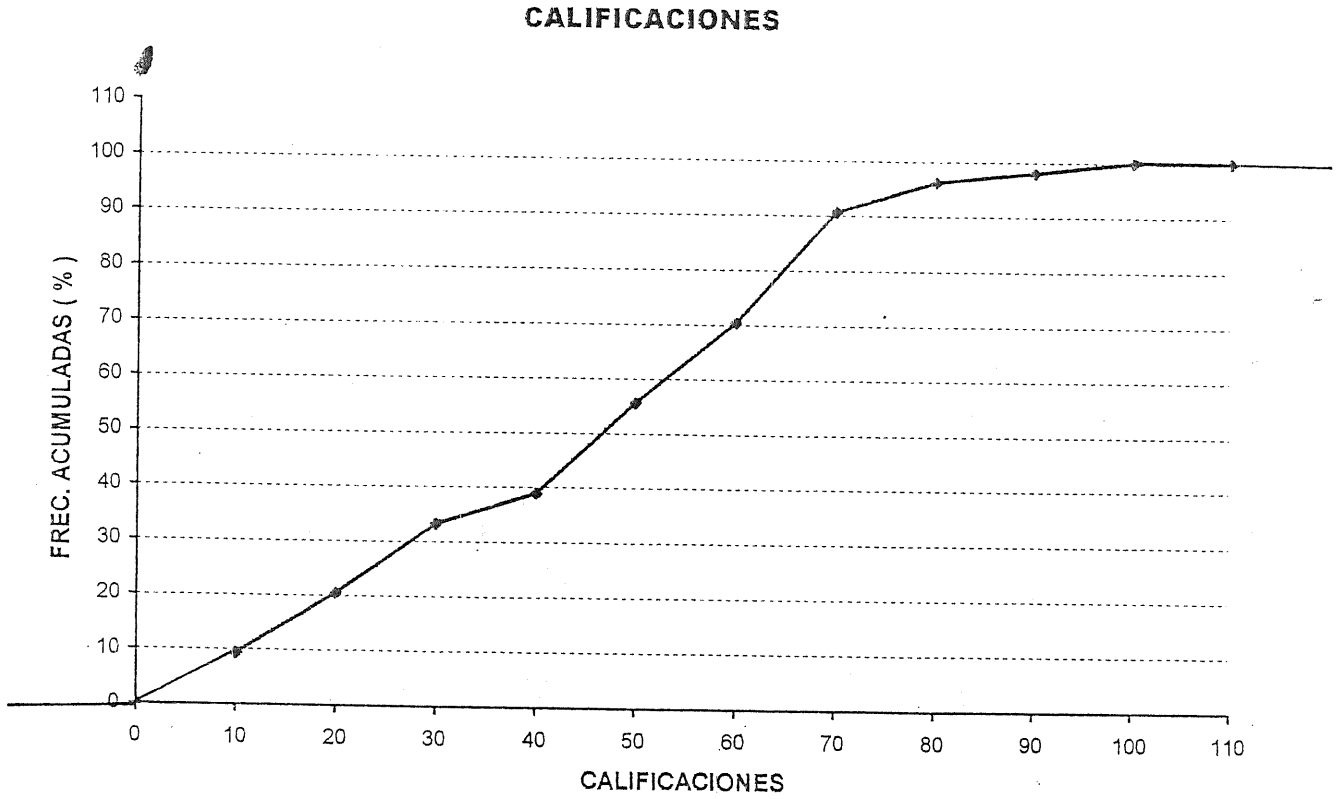
HISTOGRAMA DE FRECUENCIAS ABSOLUTAS



## POLIGONO DE FRECUENCIAS RELATIVAS PORCENTUALES

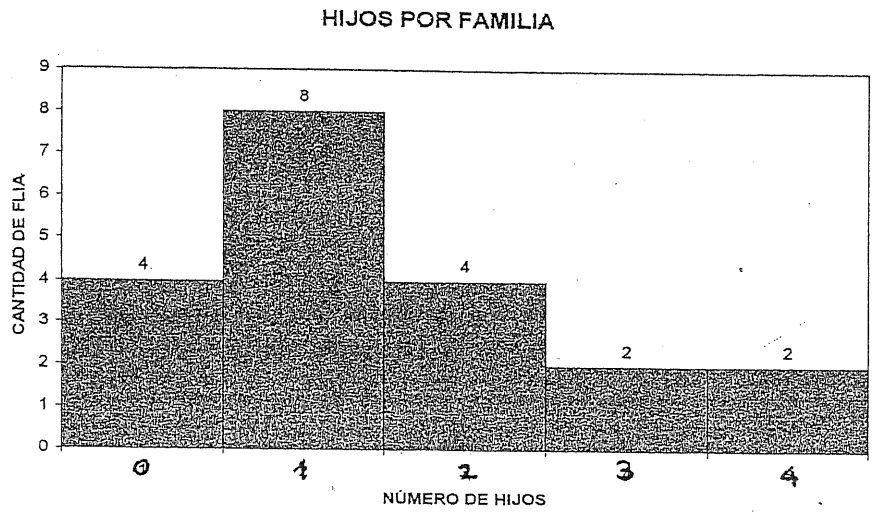



## OJIVA : FRECUENCIAS ACUMULADAS PORCENTUALES

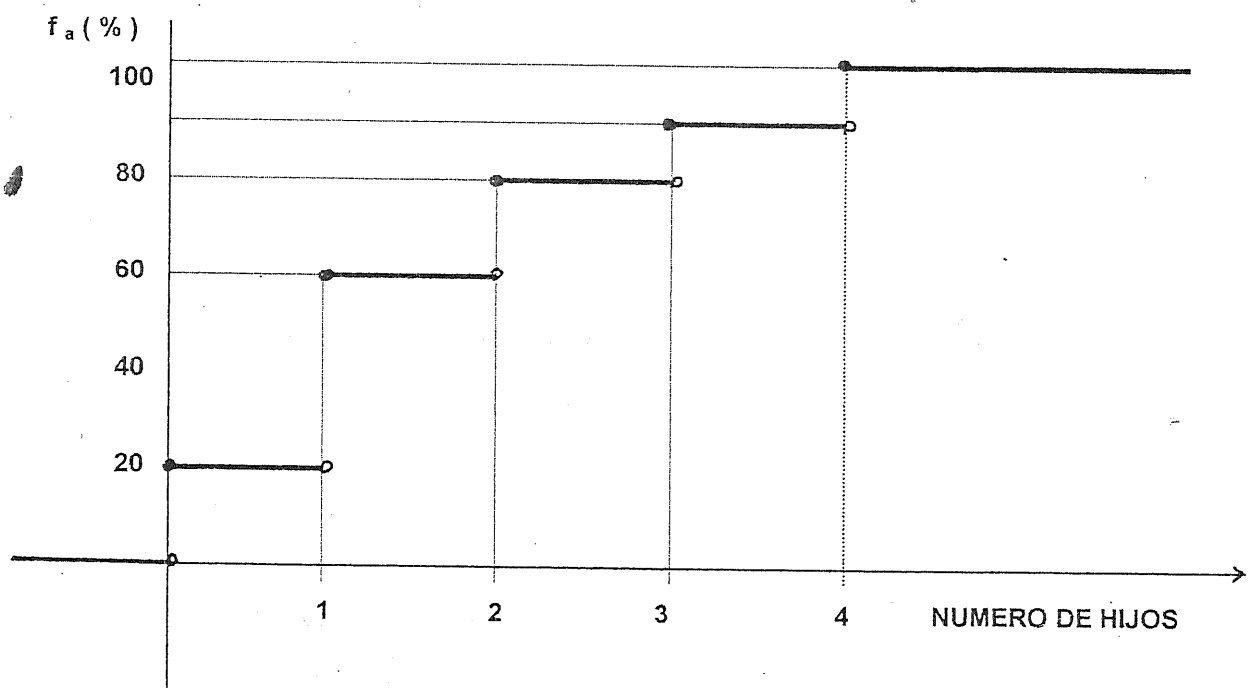


## NÚMERO DE HIJOS POR FAMILIA

## HISTOGRAMA DE FRECUENCIAS ABSOLUTAS




## GRAFICO A ESCALONES - FRECUENCIAS RELATIVAS ACUMULADAS




Número de hijos por familia : variable cuantitativa discreta –datos sin agrupar

## PROPIEDADES DE LOS DATOS CUANTITATIVOS

Las tres propiedades que describen un conjunto de datos numéricos son :

- *TENDENCIA CENTRAL*
- *DISPERSION*
- *FORMA*

En todo análisis se pueden utilizar diversas *medidas descriptivas* de tendencia central, dispersión y forma para *extraer y resumir* las principales características de los datos. Si se calculan a partir de los datos de una *muestra* se les denomina *estadísticos* , si se calculan a partir de una *población* se les denomina *parámetros*.

### Medidas de Tendencia Central y otras Medidas de Posición

La mayor parte de un conjunto de datos muestran una tendencia a agruparse alrededor de un punto, al que se llama central y, por lo general, es posible elegir algún valor, al que se llama promedio, que describa todo el conjunto de datos. Aunque la palabra promedio se refiere a cualquier medida de resumen de tendencia central, se utiliza con mayor frecuencia como sinónimo de media.

Con frecuencia se utilizan los siguientes tipos de promedio como **medidas de tendencia central** :

- ***Media Aritmética***
- ***Mediana***
- ***Moda***
- ***Rango Medio u Oscilación Media***
- ***Eje Medio***

Otras medidas de posición son los ***Cuartiles*** y ***Percentiles***

#### *Media Aritmética*

La *Media Aritmética* o *Media* es la medida de posición que se usa con más frecuencia. Se calcula sumando todas las observaciones de un conjunto y dividiendo después ese resultado entre el número total de elementos involucrados. O sea:

Dado un conjunto de **n** datos numéricos ; $x_1 , x_2 , \ldots , x_n$ ; se define la **media aritmética** como :

$$\overline{x} = [\sum_{i=1}^{n} x_i] / n$$

*Ejemplo :* Un conjunto de datos consiste en los siguientes valores 6 - 3 - 8 - 4 - 6 - 3 - 6

En este caso es **n** = 7 y la media aritmética resulta:

$$\overline{x} = (6+3+8+4+6+3+6) / 7 = 36/7 \quad \Rightarrow \quad \overline{x} = 5{,}14$$

También podemos calcular la media, mediante el *uso directo de la distribución de frecuencias.* Resulta:

Dados k valores de la variable ; $x_1 , x_2 , \ldots , x_k$ ; con **frecuencias absolutas** ; $f_1 , f_2 , \ldots , f_k$ ; respectivamente, la **media aritmética** se define :

$$\overline{x} = [\sum_{i=1}^{k} x_i f_i] / [\sum_{i=1}^{k} f_i] \qquad \sum_{i=1}^{k} f_i = n$$

k : número de valores distintos de la variable

$f_i$ frecuencia absoluta correspondiente al valor $x_i$

Para el *ejemplo*, confeccionamos la tabla de frecuencias en donde tenemos cuatro valores $x_i$ distintos de la variable $X$ ( $k = 4$ ó $i = 1, 2, 3, 4$ ) :

| $x_i$ | $f_i$ |
|---|---|
| 3 | 2 |
| 4 | 1 |
| 6 | 3 |
| 8 | 1 |
| | 7 |

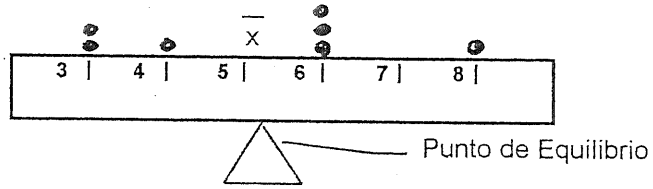
Obtenemos la media aritmética mediante el uso de la tabla de frecuencias :

$$\overline{x} = ( 3.2 + 4.1 + 6.3 + 8.1 ) / 7 = (6 + 4 + 18 + 8) / 7 = 36 / 7 \Rightarrow \overline{x} = 5{,}14$$

Este cálculo se puede sistematizar en la siguiente tabla:

| $x_i$ | $f_i$ | $x_i . f_i$ |
|---|---|---|
| 3 | 2 | 6 |
| 4 | 1 | 4 |
| 6 | 3 | 18 |
| 8 | 1 | 8 |
| TOTAL | 7 | 36 |

Se puede tener una representación de la media $\overline{x}$ si se piensa en una regla numérica equilibrada sobre un punto de apoyo, sobre la cual se colocan pesas iguales en el número correspondiente a cada observación. La media actúa como el punto de apoyo que mantiene el equilibrio de las pesas.




Para ***datos agrupados*** en intervalos , llamando $x_i$ *al centro del intervalo ó* ***marca de clase*** y $f_i$ a la frecuencia de la clase, el cálculo de la media se efectúa suponiendo que en cada intervalo todas las observaciones son iguales a la marca de clase según la siguiente expresión :

$$\overline{x} = [\sum_{i=1}^{k} x_i f_i ] / [ \sum_{i=1}^{k} f_i ] \qquad x_i : \text{marca de clase} \qquad k : \text{número de clases}$$

*Ejemplo :* Sea la siguiente tabla de datos agrupados del *ejemplo* de las "Calificaciones ... " con $n = 54$ y $k = 10$.

Calificaciones de alumnos de Estadística ( 1997) - Datos Agrupados

Tabla de Frecuencias absolutas y marca de clase para obtener la media

| CLASE | INTERVALO | MARCA $x_i$ | $f_i$ | $x_i . f_i$ |
|---|---|---|---|---|
| 1 | 00 –10 | 5 | 5 | 25 |
| 2 | 10 – 20 | 15 | 6 | 95 |
| 3 | 20 – 30 | 25 | 7 | 175 |
| 4 | 30 – 40 | 35 | 3 | 105 |
| 5 | 40 – 50 | 45 | 9 | 405 |
| 6 | 50 – 60 | 55 | 8 | 440 |
| 7 | 60 – 70 | 65 | 11 | 715 |
| 8 | 70 – 80 | 75 | 3 | 225 |
| 9 | 80 – 90 | 85 | 1 | 85 |
| 10 | 90 – 100 | 95 | 1 | 95 |
| TOTAL | | | 54 | 2.365 |

Resulta, si empleamos esta tabla de datos agrupados:

$$\tilde{x} = [\sum_{i=1}^{k} x_i f_i] / [\sum_{i=1}^{k} f_i] = 2.365 / 54 \cong 44$$

Si calculamos la media a partir del lote de datos ; resulta

$$\bar{x} = [\sum_{i=1}^{n} x_i] / n = 2427 / 54 \cong 45$$

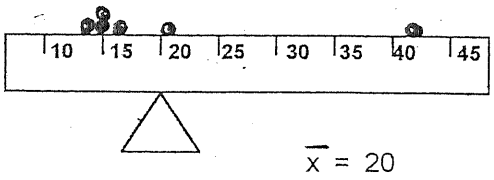
El cálculo de la media se basa en todas las observaciones del conjunto de datos . Ninguna otra medida de posición posee estas características.

Como el cálculo se basa en todas las observaciones resulta muy afectada por valores extremos . En tales casos la media aritmética presenta una información muy distorsionada respecto a la que contienen los datos realmente y no resultará la mejor medida de posición para describir o resumir ese conjunto de datos .

*Ejemplo :* Tomemos dos muestras de una misma población de tamaño n = 6

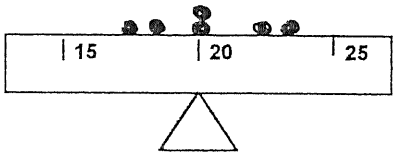
MUESTRA 1 : 13 - 14 - 14 - 16 - 21 - 42

$\bar{x}$ = 120 / 6 = 20




MUESTRA 2 : 17 - 18 - 20 - 20 - 22 - 23

$\bar{x}$ = 120 / 6 = 20




Estas dos figuras ilustran diagramas de puntos de dos muestras de igual media , $\bar{x}$ = 20 . Para la MUESTRA 1 la media aritmética da una información distorsionada de la información que contienen los datos y no es la mejor medida de posición que se pueda usar ; mientras que para la MUESTRA 2 la media es la medida descriptiva apropiada para caracterizar ese conjunto de datos puesto que no se dan observaciones muy diferentes.

***Mediana***

**Es el valor que se encuentra en el centro de un lote ordenado** ; es decir , la mediana en un lote ordenado lo divide en dos partes iguales. Los datos de una parte son menores o iguales que la mediana y los de la otra parte son mayores o iguales que la mediana.

Si n ( número de datos ) es impar , la mediana es el dato central y si n es par , la mediana es la media aritmética de los dos valores centrales.

En el *ejemplo* de la MUESTRA 1 , la mediana es $M_E$ = ( 14 + 16 ) / 2 = 15 y resulta una medida descriptiva apropiada para ese conjunto de datos .

La mediana *no se ve afectada por las observaciones extremas* en un conjunto de datos. Por ello cuando se presenta alguna observación extrema resulta apropiado utilizar la mediana y no la media para representar el conjunto de datos.

En el ejemplo de las " , hemos obtenido , empleando el lote ordenado , $M_E$ = 47

También podemos obtener la mediana a partir de la tabla de frecuencias para datos agrupados mediante interpolación.

- Ubicamos en la tabla la clase mediana teniendo en cuenta que la mediana es el dato que corresponde a la mitad de la frecuencia total , es decir que deja tantos datos por encima como por debajo , en este ejemplo ; 27 ;5 .
- Como para la clase 4 correspondiente al intervalo 40 - 50 le corresponde una frecuencia acumulada de 30 , en esa clase estará la mediana .
- La frecuencia absoluta de la clase mediana es 9 y la suma de las frecuencias absolutas de las clases inferiores a la mediana es 21 ;
- La mediana estará a [ ( 27,5 - 21 ) / 9 ] ( 50 - 40 ) del límite inferior de esta clase y resulta:

$$M_E = 40 + (65/9) \cong 47$$

La **fórmula de interpolación** para obtener la mediana a partir de una tabla de frecuencias para **datos agrupados** es la siguiente :

$$M_E = L_2 + \frac{(n/2) - (\sum f)_2}{f_{ME}} C$$

$L_2$ : Límite inferior de la clase mediana n : número de datos

$(\sum f)_2$ suma de las frecuencias de todas las clases inferiores a la mediana

$f_{ME}$ : frecuencia de la clase mediana C : tamaño del intervalo de la clase mediana

***Cuartiles***

**Son los valores de la variable que dividen al conjunto ordenado de datos en cuatro subconjuntos que contienen la misma cantidad de datos.**

Para calcular los cuartiles de una distribución de frecuencias se procede del mismo modo que en el caso de la mediana , salvo que ahora dividiremos a la distribución de la variable en cuatro partes iguales en lugar de dos.

Los cuartiles se simbolizan con la letra Q. La mediana coincide con el segundo cuartil ; $M_E = Q_2$. El primer cuartil , $Q_1$ , debe dividir a la primera mitad de la serie en dos partes iguales. Del mismo modo , el tercer cuartil , $Q_3$ , divide a la segunda mitad de la serie en dos partes iguales.

*Ejemplo :* Supongamos que un veterinario ha registrado los pesos de 8 pollos de seis semanas de vida y los ordenó de menor a mayor obteniendo :

150 - 151 - 152 - 154 - 155 - 156 - 157 - 159 [ gramos ]

La mediana estará ubicada entre el cuarto y quinto valor de la serie , siendo : $M_E = Q_2 = 154{,}5$ g
El primer cuartil se ubicará entre el segundo y el tercer valor de la serie : $Q_1 = 151{,}5$ g
El tercer cuartil es : $Q_3 = 156{,}5$ g

*Ejemplo :* Para la siguiente tabla de frecuencias correspondiente al número de hijos de 56 familias urbanas obtendremos los cuartiles:

Tabla de Frecuencias
Número de hijos de familias urbanas

| NÚMERO DE HIJOS | NÚMERO DE FAMILIAS - f - | FRECUENCIA ACUMULADA – f a - |
|---|---|---|
| 0 | 5 | 5 |
| 1 | 11 | 16 |
| 2 | 35 | 51 |
| 3 | 2 | 53 |
| 4 | 2 | 55 |
| 5 | 1 | 56 |
| TOTAL | 56 | |

Para calcular $Q_1$ , se establece el orden $n / 4 = 56 / 4 = 14$ . Buscamos en la columna de frecuencias acumuladas el primer valor que supera a 14 y obtenemos $Q_1 = 1$ para $f_a = 16$.

Para calcular $Q_3$ , el orden se busca haciendo : $3n / 4 = 42$ y obtenemos para $f_a = 51$ , $Q_3 = 2$

En el caso de la variable cantidad de hijos en familias urbanas, se tiene: $Q_1 = 1$ hijo $Q_3 = 2$ hijos

Para calcular los **cuartiles para datos agrupados** se aplican las **fórmulas de interpolación** :

**Primer Cuartil ( $Q_1$ )** para **datos agrupados :**

$$Q_1 = L_1 + \frac{(n/4) - (\sum f)_1}{f_{Q1}} C$$

$L_1$ : Límite inferior de la clase primer cuartil $n$ : número de datos

$(\sum f)_1$ suma de las frecuencias de todas las clases inferiores a la del primer cuartil

$f_{Q1}$ : frecuencia de la clase primer cuartil $C$ : tamaño del intervalo

**Tercer Cuartil ( $Q_3$ )** para **datos agrupados :**

$$Q_3 = L_3 + \frac{(3n/4) - (\sum f)_3}{f_{Q3}} C$$

$L_3$ : Límite inferior de la clase tercer cuartil $n$ : número de datos

$(\sum f)_3$ suma de las frecuencias de todas las clases inferiores a la del tercer cuartil

$f_{Q3}$ : frecuencia de la clase tercer cuartil $C$ : tamaño del intervalo

*Ejemplo :* Calcularemos ahora el primer y el tercer cuartil en el siguiente ejemplo correspondiente a la longitud de tornillos :

Tabla de Frecuencias
Longitud de tornillos fabricados por una máquina

| INTERVALOS | f | f a |
|---|---|---|
| 6 – 7 | 11 | 11 |
| 7 – 8 | 9 | 20 |
| 8 – 9 | 14 | 34 |
| 9 – 10 | 11 | 45 |
| 10 – 11 | 22 | 67 |
| 11 – 12 | 14 | 81 |
| 12 – 13 | 7 | 88 |
| 13 – 14 | 5 | 93 |
| 14 – 15 | 4 | 97 |
| 15 – 16 | 3 | 100 |
| TOTAL | 100 | |

Como el primer cuartil $Q_1$ está en el primer cuarto de la distribución la posición del dato correspondiente surge de dividir el número total de observaciones por 4 ; $n / 4 = 100 / 4 = 25$

A continuación, se debe buscar en la columna de la frecuencia absoluta acumulada , a la menor de dichas frecuencias que supera el valor $n / 4$ , en nuestro ejemplo 34 ; el límite inferior que contiene a esta frecuencia es $L_1 = 8$ y la frecuencia de la clase primer cuartil es 14 . La amplitud de dicho intervalo es $C = 1$. Reemplazando en la fórmula de interpolación se obtiene :

$$Q_1 = 8 + (25 - 20) / 14 = 8{,}36 \text{ mm.}$$

Esta medida resumen indica que el 25 % de los tornillos producidos mide menos de 8,36 mm y el 75 % mide más de 8,36 mm,

Para obtener el tercer cuartil se utiliza el mismo procedimiento pero calculando el indicador 3 n / 4 para localizar la posición del dato que corresponde al tercer cuarto de la distribución :

Posición del Tercer Cuartil : 3 n / 4 = 3 100 / 4 = 75 y el tercer cuartil resulta :

$$Q_3 = 11 + (75 - 67) / 14 = 11{,}57 \text{ mm.}$$

***Percentiles***

**Los percentiles son los valores de la variable que dividen al conjunto de datos ordenados en cien partes iguales.**

Los percentiles tienen el mismo significado y la misma forma de cálculo que los cuartiles. Así , cuando se habla del percentil 15, se quiere expresar que es el valor de la variable que deja el 15% de los datos a su izquierda y el 85% de los mismos a su derecha.

***Moda***

**Es el valor de un conjunto de datos que aparece con mayor frecuencia.**

- Se obtiene fácilmente a partir de un arreglo ordenado .
- A diferencia de la media aritmética, la moda no se afecta ante la ocurrencia de valores extremos .
- Es más variable para distintas muestras que las demás medidas de tendencia central.
- Puede no existir y en caso de existir no ser única.
- Se puede emplear para datos o variables categóricas.

En la MUESTRA 1 la moda es 14. En la MUESTRA 2 la moda es 20.

Cuando se tienen *datos agrupados* , existe un *intervalo modal* que es el intervalo de *mayor frecuencia*.

Hemos visto que al trabajar con datos agrupados se pierde la información original y como los valores de la variable estás representados por intervalos , no se puede identificar un único valor de ese intervalo para definir la moda. Este valor puede calcularse recurriendo a alguna de las siguientes formas:

- Como el punto medio o **marca de clase del intervalo modal.**
- A través de la siguiente **fórmula de interpolación** :

$$M_O = L_1 + C\,[f_1 / (f_1 + f_2)]$$

$L_1$ : Límite inferior del intervalo modal
C : Amplitud del intervalo modal
$f_1$ : Frecuencia del intervalo modal menos frecuencia del intervalo anterior
$f_2$ : Frecuencia del intervalo modal menos frecuencia del intervalo posterior

En el *ejemplo* de las "Calificaciones . . . " el intervalo modal es el 60 - 70 . La moda es $M_o = 65$ si empleamos la marca de clase de este intervalo .

Si empleamos la fórmula de interpolación, reemplazando , resulta : $M_o = 60 + 10\,[3 / (3 + 8)] \cong 63$

***Rango Medio u Oscilación Media***

El Rango Medio ( R.M ) es el promedio de las observaciones mayor y menor de un conjunto de datos :

$$R.M = (x_{MIN} + x_{MAX}) / 2$$

A pesar de su sencillez se debe usar con cautela ya que sólo se emplean las observaciones mayor y menor de un conjunto de datos y si hay observaciones extremas se distorsiona como medida de tendencia central.

***Eje Medio***

Una medida de resumen , similar en formato al Rango Medio y que no se ve afectada por las observaciones extremas es el Eje Medio ( E.M )

El Eje Medio ( E.M ) es el promedio del Primer y Tercer Cuartil de una serie de datos

$$E.M = (Q_1 + Q_3) / 2$$

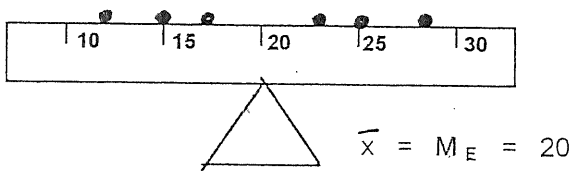
Se emplea para sortear los problemas potenciales introducidos por los valores extremos de los datos

**Medidas de Dispersión**

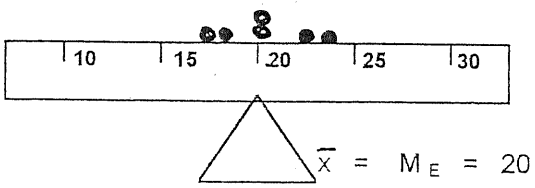
La ***dispersión*** es el ***grado de variación o diseminación*** de los datos .

Dos conjuntos de datos pueden diferir tanto en tendencia central como en dispersión ; ó los conjuntos de datos pueden tener las mismas medidas de posición pero diferir mucho en términos de dispersión como sucede en el siguiente *ejemplo* :

MUESTRA A : 12 - 15 - 17 - 23 - 25 - 28




MUESTRA B : 17 - 18 - 20 - 20 - 22 - 23




La MUESTRA B es mucho menos variable que la A.

Las **medidas de dispersión** para un lote de datos son :

- **Rango**
- **Rango Intercuartílico**
- **Varianza Muestral**
- **Desviación Estándar Muestral**
- **Coeficiente de variación**

***Rango***

El **rango** es la diferencia entre las observaciones mayor y menor de un conjunto de datos:

$$RANGO\ (R) = x_{MAX} - x_{MIN}$$

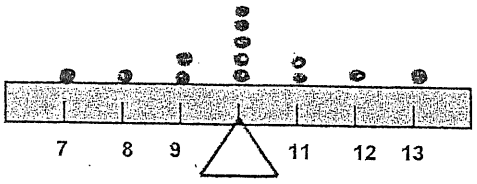
Para la MUESTRA A , R = 28 - 12 = 16 y para la MUESTRA B : R = 23 - 17 = 6

El rango mide la *dispersión total* del conjunto de datos. Es una medida de dispersión simple que se calcula con facilidad pero su debilidad preponderante es que no toma en consideración la forma en que se distribuyen los datos entre los valores más pequeños y los más grandes.

En la siguiente figura podemos comparar tres conjuntos de datos con el mismo rango

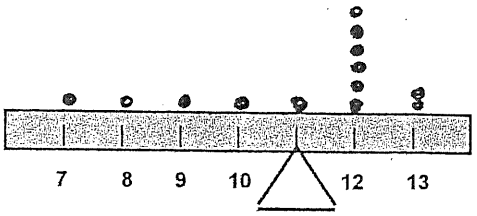
$$R = 13 - 7 = 6$$

A )



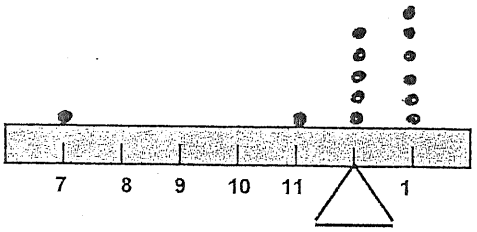

$\bar{x} = 10$

B )




$\bar{x} = 11$

C )




$\bar{x} = 12$

Como se comprueba en C ) no sería apropiado utilizar el rango como medida de dispersión cuando una o ambas de sus componentes son observaciones extremas

*Rango Intercuartílico*

El Rango Intercuartílico o propagación media es la diferencia entre el tercer y primer cuartil en una serie de datos : RI = Q 3 - Q 1

Esta medida considera la propagación en 50 % de los datos centrales y por tanto no se influenciada de ninguna manera por valores extremos de posible ocurrencia.

*Varianza y Desviación Estándar Muestral*

Dos medidas de dispersión que se utilizan con frecuencia y que *toman en consideración la forma en que se distribuyen todos los valores* son la varianza y su raíz cuadrada la desviación estándar. Estas medidas establecen la forma en que los valores fluctúan con respecto a la media.

La **Varianza Muestral** es ***casi*** **el promedio de los cuadrados de las diferencias entre cada una de las observaciones de un conjunto de datos y la media**.

Para una muestra que contiene N observaciones $x_1, x_2, \ldots, x_n$ ; la **varianza muestral**, representada por $S^2$ , se obtiene de la siguiente manera :

$$S^2 = [(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_N - \bar{x})^2] / (n - 1) \text{ ; o sea}$$

$$S^2 = [\sum_{i=1}^{n} (x_i - \bar{x})^2] / (n - 1)$$

También puede emplearse la siguiente **Fórmula Abreviada :**

$$S^2 = \frac{\sum_{i=1}^{n} x_i^2 - [(\sum_{i=1}^{n} x_i)^2 / n]}{n - 1}$$

Esta fórmula abreviada, se deduce algebraicamente de la anterior y en ella *la media no interviene en los cálculos*

Si el denominador hubiera sido $n$ en lugar de $n - 1$ , se hubiera obtenido el promedio de las diferencias al cuadrado en torno a la media. Sin embargo, se utiliza $n - 1$ , debido a ciertas propiedades matemáticas que tiene el estadístico $S^2$ y que lo hacen apropiado para realizar inferencias estadísticas.

Si el tamaño de la muestra es grande , la diferencia entre dividir por $n$ o por $n - 1$ no es significativa.

Si $k$ valores de la variable; $x_1 , x_2 , \ldots , x_k$ ; se presentan con **frecuencias** $f_1 , f_2 , \ldots , f_k$ respectivamente, la **varianza** es :

$$S^2 = \left[ \sum_{i=1}^{k} f_i \, (x_i - \bar{x})^2 \right] / (n - 1)$$

y la **Fórmula Abreviada** equivalente resulta :

$$S^2 = \frac{\sum_{i=1}^{k} f_i \, x_i^2 - \left[ \left( \sum_{i=1}^{k} f_i \, x_i \right)^2 / n \right]}{n - 1}$$

- Estas dos últimas fórmulas también son adecuadas para **datos agrupados** donde $x_i$ ***representa las marcas de clase*** , $f_i$ ***las correspondientes frecuencias de cada clase y*** $k$ ***es el número de intervalos de clase.***

La **desviación estándar muestral** , cuya notación es $S$ , es simplemente **la raíz cuadrada de la varianza muestral.** Es decir :

$$S = \sqrt{S^2}$$

*Ejemplos:*

- Para la muestra 12 , 15 , 17 , 23 , 25 , 28 ; $\bar{x} = 20$ .

$S^2 = ( 8^2 + 5^2 + 3^2 + 3^2 + 5^2 + 8^2 ) / 5 = 196 / 5 = 39,2$ unidades cuadradas

$S = \sqrt{S^2} = 6,26$ unidades

- Para el ejemplo de las "Calificaciones. . . " calculemos $S^2$ y $S$ para los **datos agrupados** y empleando la fórmula abreviada:

$\sum_{i=1}^{10} f_i \, x_i = 2.365$ $\quad n = 54$ $\quad n - 1 = 53$

$\sum_{i=1}^{10} f_i \, x_i^2 = 131.550$ , reemplazando obtenemos $S^2 = 527,77$ y $S = 22,97 \cong 23$

Ni la varianza ni la desviación estándar pueden ser negativas. En el único caso que son nulas es cuando no hay variación en los datos , es decir si todas las observaciones de la muestra tienen exactamente el mismo valor. En este caso, muy poco común, el rango también sería cero. Si embargo, los datos son valores de variables por naturaleza , no constantes. Debido a que los datos son

inherentemente variables , es tan importante estudiar no sólo medidas de tendencia central que resumen los datos sino también medidas de dispersión que reflejan la forma en que varían los datos.

***Qué indican la varianza y la desviación estándar ?***

Miden la dispersión promedio en torno a la media, es decir como fluctúan las observaciones mayores por encima de la media y cómo se distribuyen las observaciones menores por debajo de ella.

En el ejemplo de las calificaciones la desviación estándar es 23 . Se observa que la mayor parte de las calificaciones de esa muestra se agrupan dentro de 23 unidades por encima y por debajo de la media, $\bar{x} = 44$ ; es decir en el intervalo $( \bar{x} - S , \bar{x} + S ) = ( 21, 67 )$ .En este caso contamos 32 datos de un total de 54 ( 60 % del total )

La varianza tiene ciertas propiedades matemáticas útiles . Sin embargo , al calcularla , se obtienen unidades al cuadrado ( pesos al cuadrado , cm al cuadrado , etc ) ; por ello en la práctica , la principal medida de dispersión que se utiliza es la desviación estándar cuyo valor está dado en las unidades originales de los datos.

***Porqué se elevan las desviaciones al cuadrado ?***

Pues $\sum_{i=1}^{n} ( x_i - \bar{x} ) = 0$ dado que la media actúa como punto de equilibrio para las obser - vaciones que son mayores y menores que ella.

***Coeficiente de variación***

Se denomina **coeficiente de variación** al **cociente desviación estándar S sobre media aritmética** $\bar{x}$ Es una medida relativa de variabilidad y se expresa en porcientos . En símbolos:

$$C.V = ( S / \bar{x} ) \ 100 \ \%$$

Es independiente de las unidades utilizadas . El C.V mide la dispersión de los datos respecto a la media. A medida que el coeficiente de variación disminuye , se observa una mayor homogeneidad en los datos o , lo que es lo mismo , los datos están más concentrados alrededor de la media.

Se utiliza para comparar la variabilidad de dos o más conjuntos de datos expresados en diferentes unidades de medición. También es útil cuando se comparan dos o más conjuntos de datos que se miden en las mismas unidades pero difieren en tal medida que una comparación de las respectivas desviaciones estándar no resulta muy útil.

## FORMA

La forma, es la propiedad que describe la manera en que se distribuyen los datos .Una distribución de datos puede ser **simétrica o insesgada** o **sesgada o asimétrica.**

Para describir la forma , lo que se requiere , es *comparar la media y la mediana.*:

Si estas dos medidas son iguales, $M_E = \bar{x}$ , en general , decimos que los datos son **simétricos** ( o con sesgo cero )

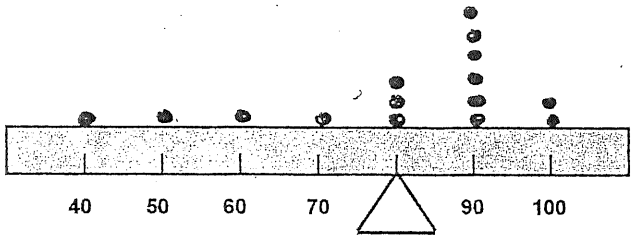
Si la media es superior a la mediana, $\bar{x} > M_E$ , en general , se dice que los datos tienen **sesgo positivo o hacia la derecha.**

Si la mediana es mayor que la media, $M_E > \bar{x}$ , se dice que los datos tienen **sesgo negativo o hacia la izquierda.**

El sesgo positivo se presenta cuando la media se ve afectada por algunos valores extraordinariamente grandes; el sesgo negativo ocurre cuando la media se ve reducida por algunos valores extremadamente bajos.

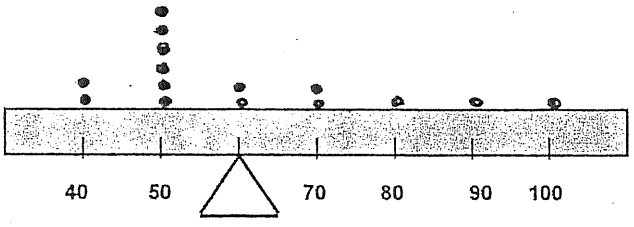
*Ejemplos :*

$\bar{x} = 80 < M_E = 90$



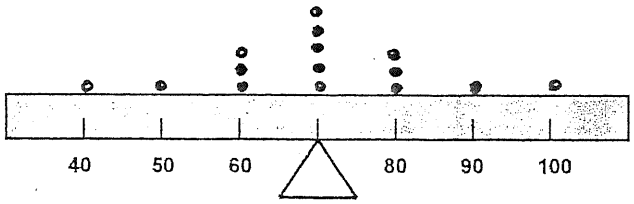

Datos sesgados hacia la izquierda. Ilustra el buen desempeño de un grupo de 15 estudiantes ; la distorsión hacia la izquierda es causada por valores extremadamente pequeños

$\bar{x} = 60 > M_E = 50$



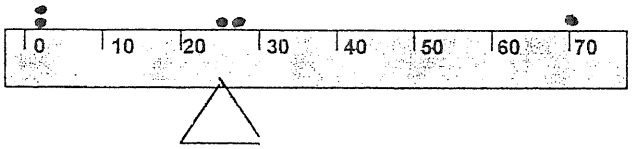

Distribución de datos sesgada hacia la derecha . Ilustra el mal desempeño de un grupo de 15 estudiantes en un examen; el sesgo a la derecha es causado por valores extremadamente grandes.

$\bar{x} = M_E = 70$



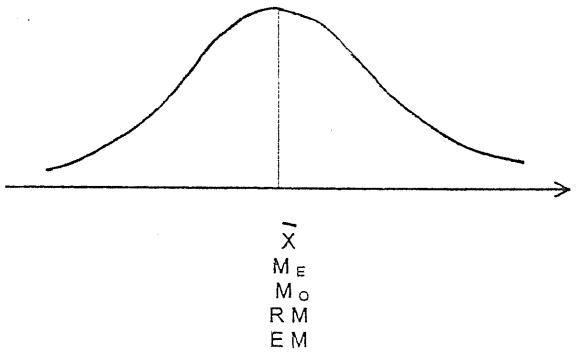

Datos simétricos . Muestran el desempeño normal de 15 estudiantes en un examen ; los valores altos y bajos de la escala se equilibran y $x = M_E$ e igual a la moda y al rango medio.

Para los datos 2 - 2 - 25 - 26 - 70 , resulta $x = M_E = 25$ y sin embargo la distribución no es simétrica.

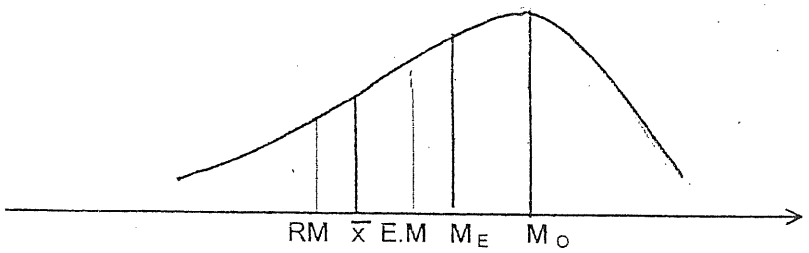



Se usan *los polígonos de frecuencias, los histogramas y los diagramas de tallos y hojas para determinar la forma* en que se distribuyen los datos .Si aproximamos estas distribuciones a una curva sin quiebres , pueden surgir las siguientes situaciones

**Distribución simétrica ( normal )** en forma de campana :



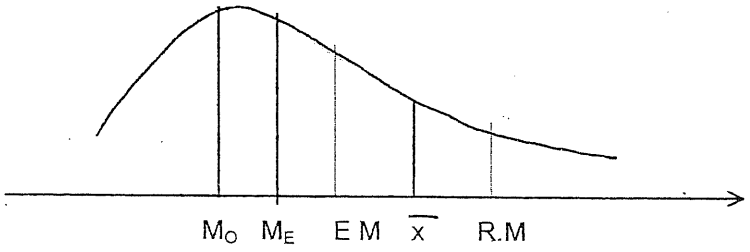
**Distribución con sesgo hacia la izquierda**




**RANGO MEDIO < $\bar{x}$ < EJE MEDIO < $M_E$ < $M_O$**

Las pocas observaciones extremadamente pequeñas distorsionan el Rango Medio y la media hacia la cola del lado izquierdo y se esperaría que la Moda fuera el valor más alto y el rango medio el menor.

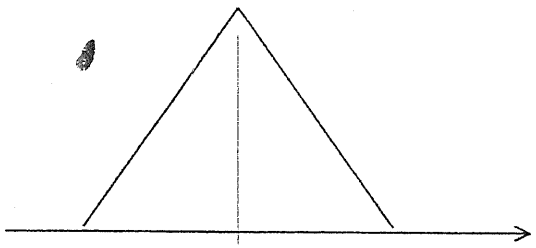
**Distribución con sesgo hacia la derecha**




**$M_O$ < $M_E$ < EJE MEDIO < $\bar{x}$ < RANGO MEDIO**

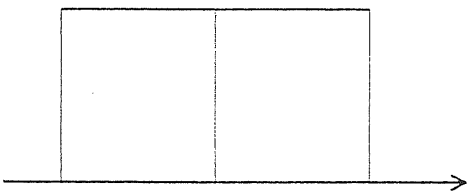
Pocas observaciones de gran magnitud distorsionan el Rango Medio y la Media hacia la cola del lado derecho . Por ello se espera que el rango medio exceda o esté a la derecha de todas las otras medidas.

***Otras distribuciones simétricas*** **:** La Mediana, la Media , el Rango Medio y el Eje Medio coinciden



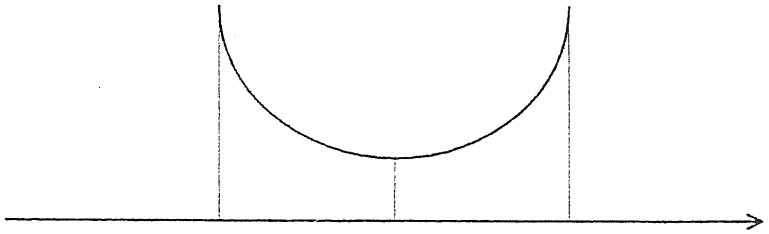
$\bar{x}$
$M_E$
$M_O$
R M
E M

Curva en forma triangular



$\bar{x}$
$M_E$
R M
E M
No existe Moda

Curva en forma rectangular



PRIMERA MODA $\bar{x}$ SEGUNDA MODA
$M_E$
$M_O$
R M
E M

En la mayor parte de los conjuntos de datos , gran parte de las observaciones tienden a aglutinarse en torno a la mediana del siguiente modo :

- En los conjuntos de datos sesgados hacia la izquierda este aglutinamiento ocurre hacia valores mayores que la mediana.
- En los conjuntos de datos sesgados hacia la derecha este aglutinamiento ocurre hacia valores menores que la mediana.
- En los conjuntos simétricos de datos las observaciones tienden a distribuirse en forma homogénea alrededor de la mediana o la media.

**La desviación estándar y la forma**

***Regla empírica para distribuciones simétricas***

Cuando no se da un sesgo extremo , es decir , los datos son simétricos , se puede utilizar la siguiente regla empírica para examinar la propiedad de variabilidad de los datos y lograr una mayor comprensión de lo que la desviación estándar mide:

Se encontrará que aproximadamente **dos de cada tres observaciones** , es decir el 67 % , están comprendidas dentro de una distancia de **una desviación estándar en torno a la media , en el intervalo** $( \bar{x} - S , \bar{x} + S )$ y que , aproximadamente **entre el 90 y el 95 %** de las observaciones están comprendidas en una distancia de **dos desviaciones estándar en torno a la media , en el intervalo** $( \bar{x} - 2S , \bar{x} + 2S )$.

***Regla de Chebyschev para una distribución cualquiera o desconocida***

Sin importar cómo se distribuye un conjunto de datos , el **porcentaje de observaciones** que están contenidas dentro de k **desviaciones estándar en torno a la media** , debe ser , cuando menos :

$$[ 1 - ( 1 / k^2 ) ] \ 100 \ \%$$

Por lo tanto , para una distribución de datos de cualquier forma , resulta :

- Por lo menos el $[ 1 - ( 1 / 2^2 ) ] \ 100 \ \% = 75 \ \%$ de las observaciones debe estar contenidas dentro de **2 desviaciones estándar en torno a la media** ( **k = 2** ) , es decir estarán en el intervalo $( \bar{x} - 2S , \bar{x} + 2S )$

- Por lo menos el $[ 1 - ( 1 / 3^2 ) ] \ 100 \ \% = 88,89 \ \%$ de las observaciones debe estar en el intervalo $( \bar{x} - 3S , \bar{x} + 3S )$

- Por lo menos el $[ 1 - ( 1 / 4^2 ) ] \ 100 \ \% = 93,75 \ \%$ de las observaciones debe estar en el intervalo $( \bar{x} - 4S , \bar{x} + 4S )$

El teorema de Chebyschev es aplicable a observaciones de cualquier distribución de y por esta razón los resultados son generalmente débiles .

El valor dado por el teorema es una *cota inferior* ; es decir se sabe que por lo menos el 75 % de las observaciones están dentro de dos desviaciones estándar de la media , o dicho de otra forma hay una probabilidad , por lo menos del 0,75 de que una observación esté dentro de dos desviaciones estándar de la media .Nunca se sabe cuánto más podría ser . Sólo cuando se conoce la *distribución de probabilidad* pueden determinarse las probabilidades exactas.

## Diagrama de cajas o box-plot

El diagrama de cajas es un formato resumido que describe las características de los datos : tendencia central, dispersión y forma. Esta forma del análisis exploratorio de datos (AED) se desarrolla a partir de un resumen de cinco números: el dato de menor valor , el primer cuartil, la mediana, el tercer cuartil y el dato de mayor valor.

$$x_{menor} \quad Q_1 \quad M_E \quad Q_3 \quad x_{mayor}$$

En estos números están involucradas :

- Tres medidas de tendencia central : la mediana, el eje medio y el rango medio

$$M_E \qquad EM = ( Q_1 + Q_3 ) / 2 \qquad RM = ( x_{menor} + x_{mayor} ) / 2$$

- Dos medidas de variación : el rango y el rango intercuartílico

$$R = x_{mayor} - x_{menor} \qquad RI = Q_3 - Q_1$$

De este modo, a partir del resumen de estos cinco números, podemos obtener conclusiones respecto de la forma de la distribución de los datos

Si los datos son Simétricos se observaría que

- La distancia de $Q_1$ a la $M_E$ sería igual a la distancia de $Q_3$ a la $M_E$
- La distancia de $x_{menor}$ a $Q_1$ resultaría igual a la distancia de $Q_3$ a $x_{mayor}$
- La mediana el eje medio y el rango medio serían todos iguales ( estas mediciones también serían iguales a la media aritmética)

Si la distribución de datos es No Simétrica y sesgada a la derecha se observaría que

- La distancia de $Q_3$ a $x_{mayor}$ excede en gran medida a la distancia de $x_{menor}$ a $Q_1$
- $M_E < EM < RM$

Si la distribución de datos es No Simétrica y sesgada a la izquierda se observaría que

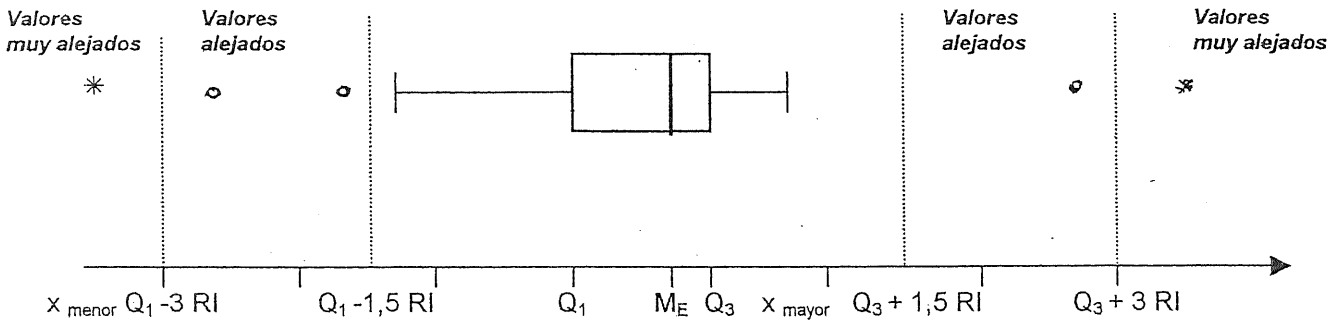
- La distancia de $x_{menor}$ a $Q_1$ excede en gran medida a la distancia de $Q_3$ a $x_{mayor}$
- $RM < EM < M_E$

Estos cinco números intervienen en la gráfica de cajas y bigotes ( box -plot ) que consiste en construir un rectángulo tal que uno de los lados tiene como longitud el Rango Intercuartílico mientras que la longitud del otro es arbitraria. El rectángulo se separa en dos partes trazando una línea vertical en donde se ubica la $M_E$ . De este modo, este rectángulo o caja contiene el 50% de las observaciones centrales de la distribución.
Si el resto de las observaciones resulta menor que $Q_3 + 1,5$ RI o mayor que $Q_1 - 1,5$ RI se representan mediante una línea que conecta respectivamente el lado derecho de la caja con el $x_{mayor}$ o el lado izquierdo de la caja con el $x_{menor}$ respectivamente construyendo los sesgos o bigotes del diagrama. La ubicación de $x_{menor}$ y de $x_{mayor}$ se indica mediante una patilla o cota definida por una barra vertical.
Si en cambio hay observaciones que superan $Q_3 + 1,5$ RI pero que son menores que $Q_3 + 3$ RI o que son menores que $Q_1 - 1,5$ RI pero mayores que $Q_1 - 3$RI , se consideran valores alejados ( atípicos o outliers ) y se representan mediante O ( una circunferencia ) .
Las observaciones que superan $Q_3 + 3$ RI o que son menores que $Q_1 - 3$ RI , si las hubiere, se consideran valores muy alejados ( muy atípicos o far outliers ) y se representan mediante * ( un asterisco ) .

La circunferencia o el asterisco se dibujan de modo que coincidan con la observación y a la altura de la línea que representa el bigote.

En el caso que aparezcan observaciones alejadas o muy alejadas la patilla está dada por el dato más próximo a $Q_3 + 1{,}5$ RI pero menor que él y / ó por el dato más próximo a $Q_1 - 1{,}5$ RI pero mayor que él.

Veamos un esquema que representa la descripción dada:

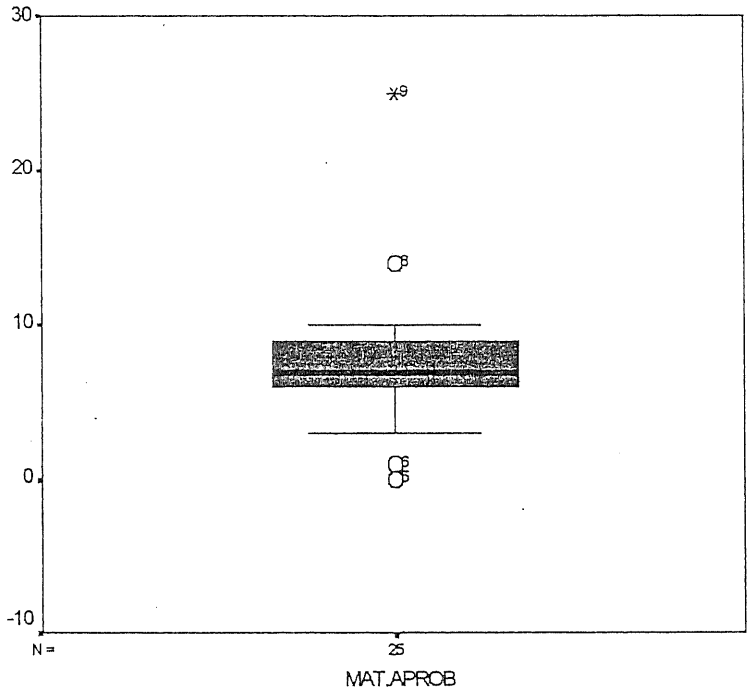



Este diagrama de cajas indica una distribución con sesgo hacia la izquierda con dos registros alejados y un registro muy alejado.

La caja también puede representarse en forma vertical.

El siguiente es el diagrama de cajas que se obtiene empleando el software de estadística SPSS para el siguiente lote de 25 observaciones correspondientes al número de materias aprobadas por los alumnos de la Facultad de Ingeniería que asisten a un curso de Estadística de carácter optativo.

0 - 1 - 3 - 4 - 5 - 6 - 6 - 6 - 7 - 7 - 7 - 7 - 7 - 7- 7- 8 - 8 - 9 - 9 - 9 - 9 - 10 - 10 - 14 - 25




El número que aparece indicado junto a los valores alejados o muy alejados corresponde al orden en que se ingresaron las observaciones para ser procesadas . En este caso los datos no se ingresaron como un lote ordenado.

Para este ejemplo el resumen de cinco números es: $x_{menor} = 0$ , $Q_1 = 6$ , $M_E = 7$ , $Q_3 = 9$ , $x_{mayor} = 25$

Se obtiene : RI = 3 1,5 RI = 4,5 3 RI = 9

$Q_1$ - 3 RI = -3 $Q_1$ - 1,5 RI = 1,5 $Q_3$ + 1,5 RI = 13,5 $Q_3$ + 3 RI = 18

Los valores alejados a la izquierda son las observaciones 0 y 1 . La observación 14 es alejada a la derecha y el valor 25 corresponde a una observación muy alejada hacia la derecha .

***Ejemplo de Salida de SPSS***

Para la variable Total de materias aprobadas observada en un grupo de 169 alumnos de un curso de Estadística de la Carrera de Ciencias Económicas se muestran los resultados obtenidos empleando el software SPSS.

La siguiente tabla proporciona las medidas descriptivas . También provee otra información como la de los intervalos de confianza para le media que aprenderemos a interpretar y usar mas adelante cuando trabajemos con la inferencia estadística

**Descriptives**

| | | | Statistic | Std. Error |
|---|---|---|---|---|
| Total de materias aprobadas | Mean | | 3.59 | .19 |
| | 95% Confidence Interval for Mean | Lower Bound | 3.21 | |
| | | Upper Bound | 3.97 | |
| | 5% Trimmed Mean | | 3.46 | |
| | Median | | 3.00 | |
| | Variance | | 6.160 | |
| | Std. Deviation | | 2.48 | |
| | Minimum | | 0 | |
| | Maximum | | 10 | |
| | Range | | 10 | |
| | Interquartile Range | | 3.00 | |
| | Skewness | | .593 | .187 |
| | Kurtosis | | -.140 | .371 |

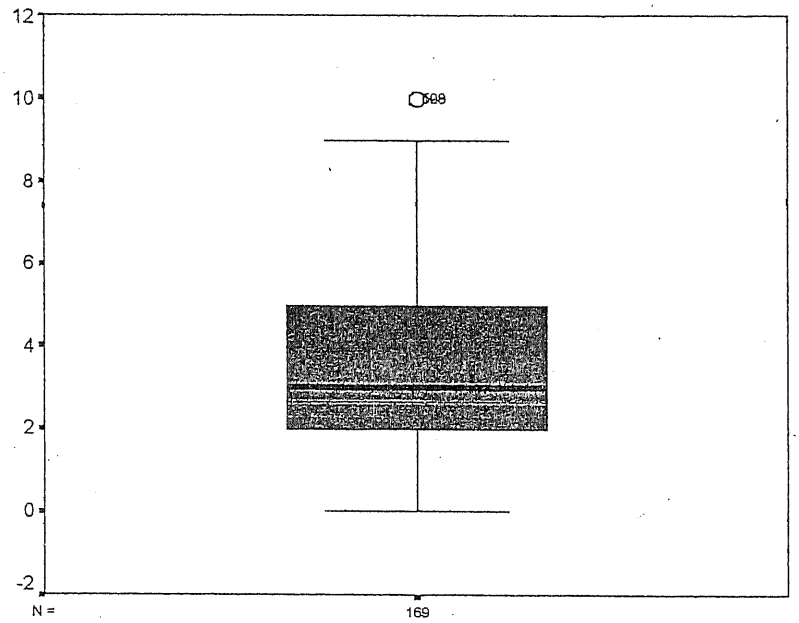
Diagrama de Tallos y Hojas

```
Frequency     Stem &  Leaf

    16.00       0 .  0000000000000000
      .00       0 .
    23.00       1 .  00000000000000000000000
      .00       1 .
    23.00       2 .  00000000000000000000000
      .00       2 .
    28.00       3 .  0000000000000000000000000000
      .00       3 .
    22.00       4 .  0000000000000000000000
      .00       4 .
    24.00       5 .  000000000000000000000000
      .00       5 .
     10.00      6 .  0000000000
      .00       6 .
    11.00       7 .  00000000000
      .00       7 .
     3.00       8 .  000
      .00       8 .
     5.00       9 .  00000
    4.00 Extremes    (>=10.0)

 Stem width:         1
 Each leaf:       1 case(s)
```

Diagrama de Caja




El diagrama de tallos y hojas y el diagrama de caja muestran que la distribución presenta un sesgo hacia la derecha. Los casos con 10 materias aprobadas se consideran alejados y así se indican en el diagrama de tallo y hojas mientras que en la caja están fuera del brazo o bigote. ( Superan la media más 1,5 R.I ). El 50 % de los valores de la variable están entre 2 y 5 materias aprobadas. El 25 % de la muestra tiene dos o menos materias aprobadas e igual porcentaje tiene 5 o más materias aprobadas.

Medidas Descriptivas por Sexo

**Descriptives**

| | Sexo | | | Statistic | Std. Error |
|---|---|---|---|---|---|
| Total de materias aprobadas | Femenino | Mean | | 3.17 | .23 |
| | | 95% Confidence Interval for Mean | Lower Bound | 2.71 | |
| | | | Upper Bound | 3.63 | |
| | | 5% Trimmed Mean | | 3.07 | |
| | | Median | | 3.00 | |
| | | Variance | | 4.837 | |
| | | Std. Deviation | | 2.20 | |
| | | Minimum | | 0 | |
| | | Maximum | | 9 | |
| | | Range | | 9 | |
| | | Interquartile Range | | 4.00 | |
| | | Skewness | | .438 | .254 |
| | | Kurtosis | | -.302 | .503 |
| | Masculino | Mean | | 4.08 | .30 |
| | | 95% Confidence Interval for Mean | Lower Bound | 3.47 | |
| | | | Upper Bound | 4.68 | |
| | | 5% Trimmed Mean | | 3.97 | |
| | | Median | | 4.00 | |
| | | Variance | | 7.302 | |
| | | Std. Deviation | | 2.70 | |
| | | Minimum | | 0 | |
| | | Maximum | | 10 | |
| | | Range | | 10 | |
| | | Interquartile Range | | 4.00 | |
| | | Skewness | | .545 | .271 |
| | | Kurtosis | | -.445 | .535 |

Diagramas de Tallos y hojas por Sexo

```
Total de materias and-Leaf Plot for     SEXO= Femenino
 Frequency     Stem &  Leaf

    10.00        0 .  0000000000
    17.00        1 .  00000000000000000
     8.00        2 .  00000000
    17.00        3 .  00000000000000000
    12.00        4 .  000000000000
    14.00        5 .  00000000000000
     6.00        6 .  000000
     3.00        7 .  000
     1.00        8 .  0
     2.00        9 .  00

 Stem width:          1
 Each leaf:       1 case(s)

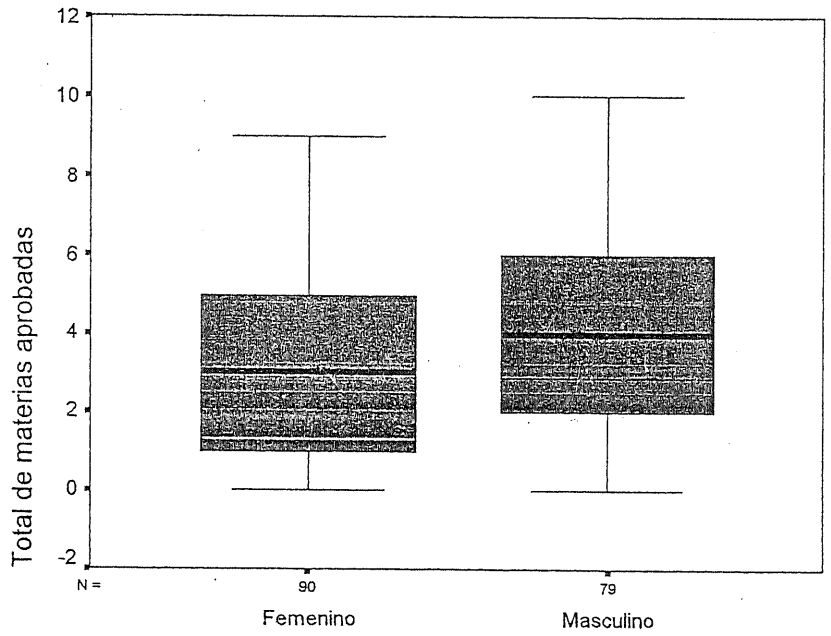
Total de materias aprobadas Stem-and-Leaf Plot for    SEXO= Masculino

 Frequency     Stem &  Leaf

    12.00        0 .  000000111111
    26.00        0 .  22222222222222233333333333
    20.00        0 .  44444444445555555555
    12.00        0 .  666677777777
     5.00        0 .  88999
     4.00        1 .  0000

 Stem width:         10
 Each leaf:       1 case(s)
```

Diagrama de cajas por Sexo




El análisis efectuado discriminando los datos por sexo hacen evidente que las distribuciones mantienen las características de la distribución general . Sin embargo se evidencia un mejor rendimiento del sexo masculino donde la media supera en una materia a la media del sexo femenino. El valor 10 no resulta alejado en este caso y aparece en el sexo masculino.

Como se observa los diagramas de cajas son útiles para hacer comparaciones gráficas entre conjuntos de datos ya que tienen gran impacto visual y son fáciles de comprender

## Medidas Descriptivas de la Población

Hasta ahora hemos examinado estadísticos, medidas descriptivas que se utilizan para resumir la información numérica que se obtiene de una muestra o porción de la población. Las mediciones resultantes calculadas a partir de la población o el universo con el objeto de describir y resumir las propiedades de tendencia central , variación y son las que llamamos parámetros y se describen a continuación.

### *Medidas de Tendencia Central de la población*

La media de la población se indica con el símbolo $\mu_X$, la letra minúscula griega $\mu$ con el subíndice X

Resulta

$$\mu_X = \left( \sum_{i=1}^{N} x_i \right) / N$$

N es el tamaño de la población y $x_i$ es el iésimo valor de la variable aleatoria
La mediana, moda, el rango medio y el eje medio de una población de tamaño N se obtienen respectivamente como se describió para el caso de una muestra de tamaño n.

### *Medidas de Variación de la población*

El rango y el rango intercuartílico para una población de tamaño N se obtienen respectivamente como se describió para una muestra de tamaño n

La varianza de la población está dada por el símbolo $\sigma_X^2$ , la letra minúscula griega sigma con el subíndice X elevada al cuadrado y la desviación estándar está dada por el símbolo $\sigma_X$

Esto es

$$\sigma_X^2 = \sum_{i=1}^{N} ( x_i - \mu_X )^2 / N$$

$$\sigma_X = \sqrt{\sum_{i=1}^{N} ( x_i - \mu_X )^2 / N}$$

N es el tamaño de la población ,
$x_i$ es el i-ésimo valor de la variable aleatoria y $\mu_X$ es la media de la población

Observamos que la varianza y la desviación estándar de la población difieren de las de la muestra en que n - 1 en el denominador de $S^2$ y S , se reemplaza por N en el denominador de $\sigma_X^2$ y $\sigma_X$

El coeficiente de variación de la población dado por el símbolo $CV_{pob}$ ' mide la dispersión en los datos relativa a la media .

Puede calcularse mediante

$$CV_{pob} = (\sigma_X / \mu_X) \ 100 \ \%$$

$\sigma_X$ = desviación estándar de la población y $\mu_X$ = media de la población