ESTADÍSTICA DESCRIPTIVA

Definiciones Preliminares

La estadística es una ciencia que permite tomar decisiones frente a diferentes problemas que surgen en el proceso de investigación, en los cuales juega un papel importante y decisivo la incertidumbre.

La estadística deriva su nombre debido a que antiguamente se recolectaban datos para diferentes fines. De aquí que la estadística (con minúscula) se entienda actualmente como un conjunto de observaciones o datos, de carácter numérico o no, mientras que la Estadística (con mayúscula) es la ciencia encargada de suministrar las diferentes técnicas y procedimientos que permiten desde organizar la recolección de los datos de una manera coherente, hasta su elaboración, análisis e interpretación.

La diferenciación planteada permite visualizar los diferentes campos de acción de la Estadística y el empleo de estos campos en las diferentes ramas de la ciencia en general, incluyendo las ciencias médicas en particular.

La Estadística puede ser diferenciada en dos grandes áreas: descriptiva e inferencial.

La Estadística Descriptiva está constituida por el conjunto de métodos estadísticos dirigidos a la elaboración primaria de los datos, entendiento esto por el resumen y presentación de la información obtenido en relación con un determinado problema científico.

La Estadística Inferencial es un área de esta ciencia que se dedica al análisis y a la elaboración de los datos, con métodos basados en la teoría de las probabilidades, con el objeto de interpretar sus resultados y tomar decisiones.

De lo expuesto puede derivarse la importancia de las amplias relaciones que se establecen entre la estadística como ciencia y las demás ciencias en general. Con las ciencias médicas en particular, esta relación se caracteriza por los complejos y variados fenómenos que estudia, los cuales en muchos casos solo pueden ser analizados a través de métodos estadísticos.

Por ejemplo, en el campo de la clínica, sólo es posible llegar al diagnóstico de cualquier enfermedad mediante la experiencia obtenida a través del análisis estadístico de un conjunto de síntomas y signos observados en muchos individuos.

Cada nuevo tratamiento propuesto para la terapeútica de determinada enfermedad siempre requiere de un ensayo experimental correctamente elaborado que demuestre su efectividad. Bajo esta circunstancia, la Estadística proporciona la posibilidad de analizar la evidencia recogida y decidir si los efectos observados son debidos al azar o si pueden ser lógicamente atribuidos a efectos del nuevo tratamiento.

La importancia que la Estadística ha alcanzado en nuestros días, tanto como cultura básica, como en el campo profesional y en la investigación, es innegable. Ello es debido a la abundancia de información con la que las personas deben enfrentarse en el trabajo diario. La mayor parte de las veces estas informaciones vienen expresadas en forma de tablas o gráficos estadísticos, por lo que un conocimiento básico de esta ciencia es necesario para la correcta interpretación las mismas.

Sin embargo, nosotros nos ocuparemos de la *bioestadística* que se define de la siguiente manera

> La Bioestadística es la aplicación de los métodos estadísticos a la solución de los problemas biológicos; también se la llama Estadística Biológica o Biometría.

La creciente importancia y aplicación de la estadística a los datos biológicos es evidente incluso al examinar de pasada cualquier revista de Biología. Este incremento tan marcado en el uso de la estadística en biología ha surgido por la comprobación de que en biología, la acción recíproca de variables causa y respuesta obedece a leyes que no están en el modelo clásico de la física del Siglo XIX. En ese siglo biólogos como Roberto Mayer y otros, tratando de demostrar que los procesos biológicos no eran sino fenómenos fisicoquímicos, ayudarona a crear la impresión de que los métodos experimentales y la filosofía natural que habían producido que había producido un progreso tan espectacular en las ciencias físicas, deberían ser imitados plenamente en biología.

De esta manera muchos biólogos habían mantenido hasta entonces la tradición de conceptos de pensamiento estrictamente mecanicistas y deterministas, mientras los físicos, debido a que sus ciencias eran muy refinadas y trataban con partículas "elementales", recurrieron a planteamientos estadísticos. En biología la mayoría de los fenómenos se ven afectados por muchos factores causales incontrolables en su variación y, a menudo no identificables. La estadística es necesaria para medir tales fenómenos variables con un error predecible y para descubrir la realidad de mínimas pero importantes diferencias.

POBLACIONES Y MUESTRAS

Vamos a definir varios términos importantes y necesarios para la comprensión de la bioestadística. Generalmente los datos se basan en observaciones individulaes, que son las observaciones o medidas de la mínima unidad de muestreo, también llamada individuo o unidad de análisis. Con frecuencia estas mínimas unidades de muestreo son también individuos en el sentido biológico ordinario. Si pesamos a 100 recién nacidos, el peso de cada recién nacido es una observación individual; los pesos de los 100 recién nacidos representan la muestra de observaciones, que se define como

> Es un conjunto de observaciones individuales seleccionadas por un procedimiento específico

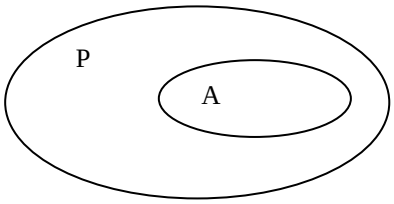
Hemos definido "observación individual" y "muestra de observaciones"; las que sólo definen la estructura pero no la naturaleza de los datos de un estudio. Llamaremos *variable a la propiedad real medida por las observaciones individuales*. En cada unidad mínima de muestreo o individuo pueden medirse más de una variable. A continuación vamos a definir el término *población.*

"La definición de población desde el punto de vista biológico se refiere a todos los individuos de la misma especie que se encuentran en un área limitada en un momento dado".

En estadística, población se define como

> La totalidad de las observaciones individuales sobre las cuales se hacen inferencias, las cuales existen en cualquier parte del mundo o al menos dentro de un área de muestreo claramente especificada, limitada en espacio y tiempo.

Si se toman cinco hombres y se estudia el número de leucocitos en su sangre periférica, con la intención de sacar conclusiones sobre todos los hombres a partir de esta muestra de cinco, en este caso la población de la que se ha extraído la muestra representa los recuentos de leucocitos de todos los varones de la especie Homo sapiens. En cambio si se restringe a cinco varones argentinos de 20 años, la población muestreada estará constituida por los números de leucocitos de todos los varones argentinos de 20 años. En todos los casos la muestra es un subconjunto de la población.



Siendo: P: el conjunto que representa la población y A es un subconjunto de P, $A \subset P$, representa la muestra.

En general representaremos con N al número de elementos de la población, y con n al número de elementos de la muestra.

N: el número de elementos de la población
n: el número de elementos de la muestra.

La muestra debe cumplir ciertas condiciones: representatividad, aleatoriedad e independencia.

- **Representatividad**: la muestra debe revelar las características de la población de la cuál proviene lo más aproximadamente posible. Ha de ser una auténtica representación de toda la población, y por lo tanto no sirve cualquier porción de la misma.

  Para que sea representativa de una población, el tanto por ciento de individuos de la muestra que poseen una propiedad determinada ha de ser el mismo que el tanto por ciento de individuos con esta propiedad en la población.

- **Aleatoriedad**: cada elemento de la población debe tener la misma posibilidad de ser elegido. Sólo si satisface este requisito los métodos estadísticos serán razonables.

- **Independencia**: esto equivale a decir que la probabilidad de que cualquier miembro de la población aparezca en la muestra, no depende de la aparición de los otros miembros de la población en la muestra.

  Para que se cumpla la independencia, la población debe ser infinita. Cuándo la población es finita, y se efectúa el muestreo aleatorio con reemplazo (esto tiene por objeto hacer infinita la población) se cumple la independencia, ya que ante cada selección restablecemos la situación original, este muestreo se llama no exhaustivo o con reposición.

  Cuando la población es finita y se efectúa el muestreo aleatorio sin reemplazo, muestreo llamado exhaustivo o sin reposición, no podemos hablar de que los elementos de la población sean independientes, en este caso decimos que hay dependencia.

Ejemplo 1:

Si para encuestar a 5.000 personas, de las que 2.700 son mujeres y 2.300 son hombres se toma una muestra de 2.000 personas, entre ellas tendrá que haber, para que sea representativa:

$$x: \textbf{mujeres}, \quad \frac{x}{2000} = \frac{2700}{5000}; \quad \Rightarrow \quad x = 1080 \, \textbf{mujeres}$$

$$y: \textbf{hombres}, \quad \frac{y}{2000} = \frac{2300}{5000}; \quad \Rightarrow \quad y = 920 \, \textbf{hombres}$$

EXPERIMENTO ALEATORIO

Es el experimento que proporciona diferentes resultados aún cuando se repita siempre de la misma manera. Los resultados obtenidos sólo dependen del azar

TIPOS DE DATOS Y ESCALAS DE MEDICIÓN

Para poder organizar y representar los distintos tipos de datos vamos primeramente a definir variable, carácter o atributo

DEFINICIÓN DE VARIABLE, CARÁCTER O ATRIBUTO

Es la característica que sintetiza o abrevia conceptualmente lo que se desea conocer acerca de los individuos o unidades de análisis.

A estas características las clasificamos en cualitativas y cuantitativas.

Carácter cualitativo: es aquella característica de la unidad de análisis que no es susceptible de medición cuantitativa, también recibe el nombre de atributo.

Ej. color de ojos, marca de un automóvil, título secundario, nivel social, sexo.

Los atributos toman modalidades.

Ej. Color de ojos: marrones, celestes, verdes, grises, otros.

Marca de autos: ford, fiat, VW, otros

Respuesta a un determinado tratamiento: excelente, muy buena, buena, regular, mala.

Tipo de sangre: A, B.AB, O

Carácter cuantitativo: es aquella característica de la unidad de análisis que si es susceptible de medición, a estas características la vamos a llamar variables. Las variables cuantitativas pueden ser discretas o continuas.

Variables discretas: son aquellas que toman un numero finito o infinito numerable de valores.

Ej. Número de alumnos por aula

Número de fusibles por caja

Número de accidentes por día.

Número de veces que una persona a ingresado a un hospital

Cantidad de veces que se realiza un experimento hasta obtener un éxito.

Así, los experimentos que consisten en el recuento de objeto dan lugar a variables discretas.

Variables continuas: son aquellas que toman un numero infinito no numerable de valores dentro de un intervalo.

Ej: Tiempo de duración de las lámparas
Longitud de tornillos
Tiempo de vida de las personas.
Presión sanguínea de las personas.
Altura de las personas.

Así, al medir magnitudes tales como tiempo, longitud, capacidad, etcétera se obtienen variables continuas.

Las variables pueden ser clasificadas como unidimensionales, bidimensionales o n-dimensionales.

Variable unidimensional, la representaremos con una letra mayúscula imprenta X, se utiliza cuando interesa una sola característica del individuo.

Variable bidimensional: la representaremos con dos letras mayúscula imprenta (X, Y), se utiliza cuando interesa simultáneamente dos características del individuo.

Ejemplo: Diámetro y longitud de un tornillo.
Altura y peso de las personas
Presión sanguínea y edad de las personas

Variable n-dimensional: $(X_1, X_2,.....X_n)$, se utiliza cuando interesa simultáneamente n características del individuo.

ESCALAS DE MEDICIÓN

Puede interesarnos analizar distintas características de un mismo individuo. Estas características dependiendo del tipo de valores que originan, pueden medirse con cuatro tipos distintos de *escalas de medición:*

- ***Escala nominal:***

Una **variable nominal** consiste en categorías a las que se asigna un nombre sin que exista ningún orden implícito entre ellas.

A las variables nominales le asignaremos una ***escala nominal***. Es la escala de nivel más sencilla.

La forma más simple de observación es la clasificación de individuos en categorías que simplemente pueden distinguirse entre sí, no existe orden implícito entre ellas, no pueden compararse ni realizarse entre ellas operaciones aritméticas. En este tipo se incluyen características tales como la profesión, nacionalidad, grupo sanguíneo, provincia de origen, etcétera.

- ***Escala ordinal:***

Una variable ordinal consiste en categorías ordenadas, la diferencia entre categorías puede no ser iguales. En ningún caso sabemos con certeza cuanto "mayor" es una categoría de la variable respecto a otra pues no existe una medición de distancia.

A las variables ordinales asignaremos una *escala ordinal*. La escala ordinal tiene las características de la escala nominal con una relación implícita de orden entre las medidas.

Las distintas calificaciones de un estudiante dadas como *excelente-muy bueno-bueno-regular-aplazado*, tienen cuatro categorías. Difieren de una variable como el color del pelo en el hecho de que existe una ordenación entre estos valores: *excelente* es mejor que *muy bueno* y éste, a su vez, mejor que *bueno*, que es mejor que *regular y que aplazado*. Sin embargo, no podemos suponer que la diferencia entre *excelente o muy bueno*, *bueno, regular y aplazado* sea la misma que la existente entre *excelente* y *aplazado*.

- ***Escala de intervalo:***

Si consideramos una variable cuantitativa, que toma valores reales, a esta variable interválica le corresponde una *escala de intervalo*, donde esta definida la igualdad, orden y distancia, es decir puede indicar cuánto más significa una categoría que otra.

Es necesario que se defina una unidad de medida y un origen, que es por su naturaleza arbitrario. Tal como ocurre con la temperatura, la escala cronológica, el poder adquisitivo etc.

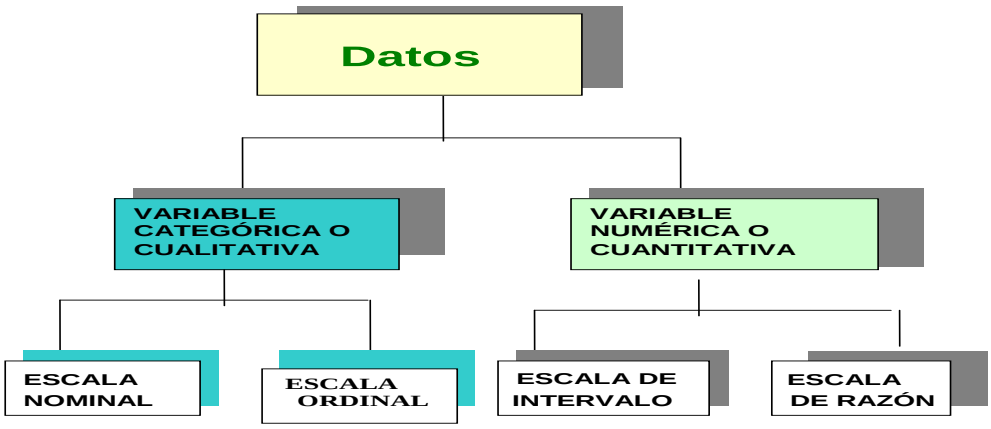
Cuando se dice que una unidad de análisis es “mayor” que otro, se puede especificar cuantas unidades de diferencia hay entre ambos o cuantas unidades una unidad es mayor que el otro. Por ejemplo con la medición de intervalo es posible decir no sólo que Pedro tiene un cociente intelectual mayor que Juan, sino que se puede decir cuanto es la diferencia entre sus cocientes intelectuales. Un cociente intelectual igual a 0 no significa ausencia de inteligencia, sino un problema grave intelectual o de percepción al utilizar los elementos de la prueba.

- ***Escala de razón o de proporción***:

Una variable cuantitativa a la cual se le puede asignar una escala de intervalos pero que además esta definido el cero absoluto, se denomina variable de razón o de proporción y dicha escala se denomina de ***escala de razón o de proporción.*** El cero absoluto representa la ausencia de la característica bajo estudio.

Ejemplos de este tipo de variables son: el volumen de ventas, los costos de producción, la cotización de un cierto tipo de acciones, etcétera

Resumen



ORGANIZACIÓN Y PRESENTACIÓN DE DATOS

Existen tres formas para presentar los datos relativos a una población ó a una muestra, ya organizados y procesados de cualquier estudio estadístico: texto, cuadros o tablas y gráficos.

TEXTO

Es una combinación de cifras y texto. Esta forma de presentación permite llamar la atención sobre las comparaciones de importancia y destacar ciertas cifras. Sin embargo, sólo puede utilizarse cuando los datos por presentar son pocos.

CUADROS O TABLAS

Este tipo de presentación de la información permite volcar un gran número de datos en forma resumida, con lo que hace fácil y clara su lectura. Es más breve, puesto que los encabezados de las columnas y los títulos de las filas evitan repetir explicaciones, Fundamentalmente, facilita las comparaciones de los datos.

GRÁFICOS

La representación gráfica de los datos contenidos en un estudio estadísticos tiene como finalidad ofrecer una visión de conjunto del fenómeno sometido a investigación, más rápidamente perceptible que la observación directa de los datos numéricos. De aquí que las representaciones gráficas sean un medio eficaz para el análisis de la información estadística, ya que las magnitudes y las regularidades se aprecian y recuerdan con más facilidad cuando se examinan gráficamente. Hay que advertir, sin embargo, que la representación gráfica no es más que una herramienta de la investigación estadística, la cual es básicamente numérica.Las representaciones gráficas pueden hacerse utilizando un sistema geométrico de representación, en cuyo caso gozan de rigurosidad y precisión, o bien pueden utilizarse símbolos alusivos al tema en estudio, por ejemplo: casas, árboles figuras humanas, etcétera. Mediante este último sistema de representación no se persigue una rigurosa exactitud, sino lograr efectos visuales en quien está leyendo la información.Existe una gran variedad de gráficos. Su elección depende de las variables en estudio y de las características que se quieren destacar. Para la construcción de gráficos no hay reglas únicas. Siempre se debe tener presente que un gráfico de información más rápida pero menos precisa que la tabla.

ORGANIZACIÓN DE LOS DATOS

Cuando se comienza a analizar una variable estamos interesados en saber los valores que puede tomar, el número total de datos con que contamos y cuántas veces aparecen los diferentes valores. Para presentar una variable es útil representarla mediante una tabla o cuadro.Cuando se dispone de gran número de datos, es útil distribuirlos en clases o categorías y determinar el número de individuos pertenecientes a cada clase, que es la frecuencia de clase. Una ordenación tabular de los datos, con las frecuencias correspondientes, se conoce como una distribución de frecuencias. A continuación presentamos una distribución de frecuencias correspondientes a una muestra de tamaño n.

TABLA DE DISTRIBUCIÓN DE FRECUENCIAS

X	f_i	f_{ri}	F_i	F_{ri}
x_1	f_1	f_{r1}	$F_1 = f_1$	$F_{r1} = f_{ri}$
x_2.			$F_2 = f_1+f_2$	$F_{r2} = f_{r1}+f_{r2}$
.				
x_i	f_i	f_{ri}	$F_i = f_1+..+f_i$	$F_{ri} = f_{r1}+..+ f_{ri}$
.				
.				
x_m	f_m	f_{rm}	$F_m = \sum_{j=1}^{m} f_j = n$	$F_{rm} = \sum_{j=1}^{m} f_{rj} = 1$
Total	$\sum_{j=1}^{m} f_j = n$	$\sum_{j=1}^{m} f_{rj} = 1$		

- ⇨ En la 1era columna aparecen los distintos valores x_i de la característica X (atributo o variable)
- ⇨ En la 2da columna aparece la ***frecuencia absoluta*** f_i , representa el número de individuos que presentan los distintos valores de la variable.
- ⇨ En la 3er columna aparecen las ***frecuencias relativas*** f_{ri}, que representa la frecuencia con respecto al total de la muestra.
- ⇨ En la cuarta columna aparece la ***frecuencia acumulada*** F_i, de los distintos valores de la variable
  Es la frecuencia total de todos los valores menores o iguales al valor de x_i, correspondiente.
- ⇨ En la quinta columna aparece la ***frecuencia relativa acumulada*** F_{ri}, que representa la frecuencia acumulada respecto al total de la muestra.
  Se calcula:

$$Fr_i = F_i / n$$

FRECUENCIA ABSOLUTA

Frecuencia absoluta
Definición: es el número de veces que se presenta cada valor de la variable.

La suma de todas las frecuencias absolutas nos da el tamaño de la muestra**, n.**

$$\sum_{i=1}^{m} f_i = n$$

Siendo n: el número de elementos de la muestra
m: el número de valores distintos que toma la variable X
N: tamaño de la población

FRECUENCIA RELATIVA

Frecuencia relativa
Definición: Frecuencia relativa es el cociente entre la frecuencia absoluta f_i y el número total de elementos n de la muestra.

$$f_{ri} = \frac{f_i}{n}$$

La frecuencia relativa indica la proporción de individuos que poseen ese valor de la variable con respecto al total de individuos

$$0 \leq f_{ri} \leq 1$$

La frecuencia relativa toma como valor mínimo el cero y como máximo el valor 1.
Una propiedad muy importante es que: "La suma de todas las frecuencias relativas nos da siempre 1".

$$\sum_{i=1}^{m} f_{ri} = 1$$

I: VARIABLES

La presentación de los datos de estas variables puede realizarse mediante tablas con datos sin agrupar o con datos agrupados por intervalos.

a) DATOS SIN AGRUPAR

Realizaremos primeramente una presentación tabular de los datos sin agrupar. Se enumeran todos los valores observados de la variable, con sus respectivas frecuencias absolutas o frecuencias relativas:

TABLA DE DISTRIBUCIÓN DE FRECUENCIAS

X	f
x_1	f_1
x_2	f_2
.	.
.	.
.	.
x_k	f_k

	$\sum_{i=1}^{k} f_i$ =n

$k \leq n$

El gráfico que le corresponde, cuando queremos graficar la variable aleatoria con relación a su frecuencia absoluta o frecuencia relativa es el ***gráfico de bastones***, cuando la variable es cualitativa.

El gráfico que corresponde para representar las frecuencias absolutas acumuladas o frecuencias relativas acumuladas, en variables cuantitativas es el ***gráfico de escaleras***.

Ejemplo 2:

Sea X: ***"Número de cuadras caminadas por 14 alumnos de una escuela rural, para llegar cada mañana".***

La muestra observada es

5 5 5 6 8 4 4 2 1 8 6 6 4 5

siendo n =14

Se pide: organizar los datos en una tabla de frecuencias y luego graficar la variable X con relación a su frecuencia absoluta

Solución: Debemos realizar primeramente una presentación tabular de los datos, para tal fin lo primero que realizamos es un ordenamiento de los datos observados.

1 2 4 4 4 5 5 5 5 6 6 6 8 8

TABLA DE DISTRIBUCIÓN DE FRECUENCIAS

X	f	F
1	1	1
2	1	2
4	3	5
5	4	9
6	3	12
8	2	14
Total	$\sum_{i} f_i$ =14	

GRÁFICO DE BARRAS

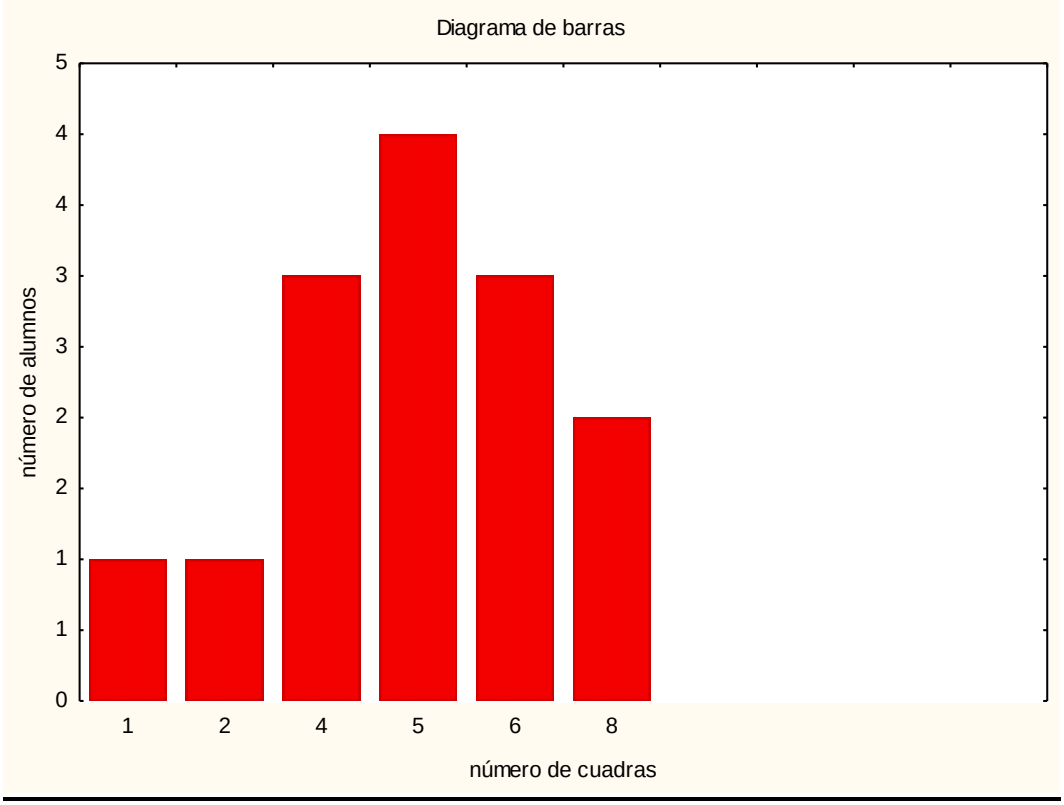


GRÁFICO DE FRECUENCAS ABSOLUTAS ACUMULADAS: GRÁFICO DE ESCALERAS

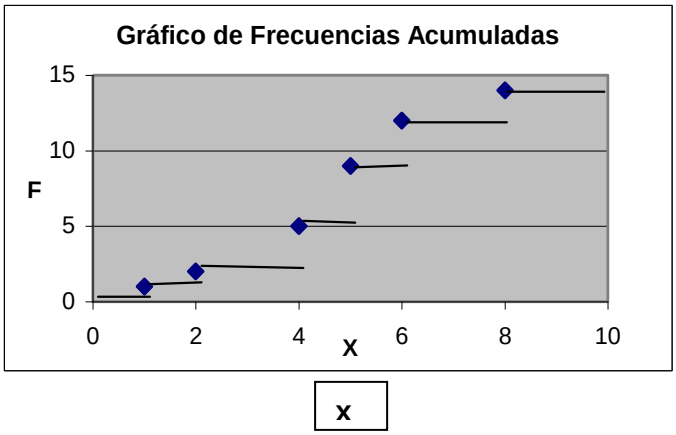


GRÁFICO DE TALLO Y HOJAS

Es interesante conocer simultáneamente el valor individual de cada una de las observaciones. El gráfico de tallo y hojas fue descripto por Tukey.

Para realizar este gráfico, basta seguir los siguientes pasos:

1. Primero se ordenan los datos de menor a mayor
2. Se apartan uno o más dígitos de cada dato, según el número de filas que se desea obtener, en general no más de 12 ó 15, empezando por la izquierda. Cada valor diferente de estos dígitos apartados, se lista uno debajo del otro, trazando a la derecha de los mismos una línea vertical. Éste es el tallo.
3. Para cada dato original se busca la línea en la que aparece su tallo. Los dígitos que nos quedaban los vamos escribiendo en la fila correspondiente de forma ordenada.
4.

Ejemplo 3:

A continuación se listan los tiempos (en minutos) que requieren 30 médicos en atender a sus pacientes. Realice con estos datos un gráfico de tallo y hoja.

20 24 15 33 17 16 32 14 12 13
35 18 20 14 14 30 15 31 23 11
22 31 27 26 12 31 36 12 24 15

1. Los datos ordenados de menor a mayor son:

11 12 12 12 13 14 14 14 15 15
15 16 17 18 20 20 22 23 24 24
26 27 30 31 31 31 32 33 35 36

2. Observamos que en todos los datos, los dígitos de la izquierda son los números 1,2 y 3. Listamos estos números de arriba abajo y dibujamos una línea vertical

1 |
2 |
3 |

3. A continuación, para cada dato original, vamos escribiendo, ordenadamente, el dígito que nos queda, en su fila correspondiente. Si alguno se repite, se escribe tantas veces como aparezca. Completando de esta forma el gráfico de tallo y hojas

GRÁFICO DE TALLO Y HOJAS

1	1 2 2 2 3 4 4 4 5 5 5 6 7 8
2	0 0 2 3 4 4 6 7
3	0 1 1 1 2 3 5 6

GRÁFICO DE PUNTOS

Otro gráfico que podemos utilizar es el ***gráfico de puntos o puntigramas***, que nos permiten distinguir claramente las variables y su frecuencia.

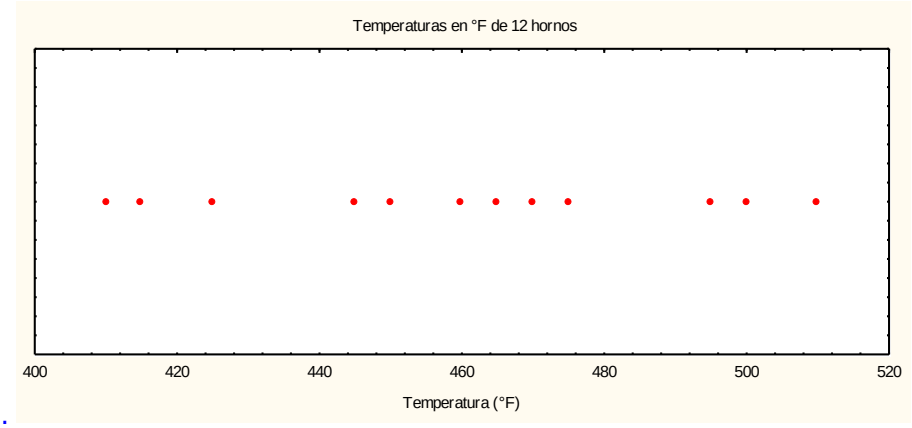
Ejemplo 4:

Los siguientes datos son doce lecturas de temperatura en varios puntos de un gran horno (en grados Fahrenheit):

445 410 470 460 415 510 450 495 465 500 475 425

Realice un diagrama de puntos.

Temperaturas en varios puntos de un horno (en ºF)



b) **DATOS AGRUPADOS**

Para resumir la información y adquirir una visión global y sintética de variables cuantitativas con un gran número de valores, se suelen agrupar los datos en intervalos de clases al elaborar las tablas de frecuencias y se registra el número de valores observados que corresponde a cada clase.

Aunque con el proceso de agrupamiento, generalmente se pierde parte del detalle original de los datos, tiene la importante ventaja de presentarlos todos en un sencillo cuadro que facilita el hallazgo de las relaciones que pueda haber entre ellos.

Para poder identificar los patrones en un conjunto de datos agrupamos las observaciones en un número relativamente pequeño de clases que no se superpongan entre sí.

TABLA DE DISTRIBUCIÓN DE FRECUENCIAS PARA DATOS AGRUPADOS

X	f	fr
$[x_0\text{-}x_1)$	f_1	f_{r1}
$[x_1\text{-}x_2)$	f_2	f_{r2}
:	:	:
:	:	:
$[x_{k-1}\text{-}x_k]$	f_k	f_{rk}
	$\sum_{i=1}^{k} f_i = n$	$\sum_{i=1}^{k} f_{ri} = 1$

Para proceder a la construcción de una tabla de distribución de frecuencias con datos agrupados es preciso tener en cuenta las siguientes nociones:

FRECUENCIA ABSOLUTA DE CLASE: es el número de observaciones de una clase f.

FRECUENCIA RELATIVA DE CLASE: es el cociente entre la frecuencia de clase y el número de observaciones. Representa la proporción de observaciones contenidas en cada clase f_r.

Clase: se llama clase a cada uno de los intervalos en que podemos dividir el recorrido de la variable estadística. Los intervalos pueden o no ser de la misma amplitud, en general trabajaremos con intervalos de igual amplitud.

Las fronteras de clase se denominan ***límites*** son los puntos específicos de la escala de medición que sirven para separar clases adyacentes. ***Límite superior de la clase***, es el máximo valor del intervalo. ***Límite inferir de la clase*** es el mínimo valor del intervalo.

Si calculamos el rango del intervalo de clases obtenemos ***la longitud del intervalo de clase.*** El ***Rango de una clase*** puede ser determinado restando al límite exacto superior de clase el límite exacto inferior

Para ciertos propósitos, los valores de una clase se representan a menudo por ***el punto medio de clase o marca de clase,*** que es el punto medio del intervalo de clase, puede ser determinado calculando la media aritmética entre los límites superior e inferior.

Al graficar los intervalos de clase relacionado con sus frecuencias relativas, o frecuencias absolutas, se produce lo que se conoce como ***histograma de frecuencias relativas, o histogramas de frecuencias absolutas*** respectivamente.

Un histograma o histograma de frecuencias consiste en una serie de rectángulos que tiene: a) sus bases sobre un eje horizontal (eje X) con centros en las marcas de clase y longitud igual al tamaño de los intervalos de clase, b) si los intervalos de clase tiene todos igual tamaño, las alturas de los rectángulos son proporcionales a las frecuencias de clase. Si los intervalos de clase no son de igual tamaño, estas alturas deberán ser calculadas.

También puede graficarse los datos agrupados mediante un ***polígono de frecuencias,*** que es un gráfico de línea que se obtiene uniendo las frecuencias relativas o absolutas de los puntos medios de las bases superiores de los rectángulos de un histograma de frecuencias. Además completaremos la poligonal, uniendo los puntos medios del intervalo anterior al primero de nuestra muestra y posterior al último de nuestra muestra.

Ejemplo 5:

En el siguiente ejemplo aplicaremos un método para construir la tabla de distribución de frecuencias de la variable en estudio.

A continuación se registran los pesos de 40 estudiantes con aproximación de una libra. Se pide:

a) Construir una tabla de frecuencia con datos agrupados
b) Graficar.

138	164	150	132	144	125	149	157
146	158	140	147	136	148	152	144
168	126	138	176	163	119	154	165
146	173	142	147	135	153	140	135
161	145	135	142	150	156	145	128

El número de clases que se emplea depende del total de observaciones. Si el número de observaciones es muy pequeño el número de clases será cercano a 5 generalmente no menor de este valor, si el número de observaciones es grande se utilizaran 8 ó 12, pero no más de 15 clases. No existe una regla fija, es una relación entre la pérdida de la información que supone el agrupamiento y la visión global y sintética que se persigue. Esta flexibilidad para la selección de la cantidad de intervalos puede provocar dudas o confusiones, es por eso que Sturguess da una fórmula para quien no quiera o no pueda decidir la cantidad de clases a utilizar

Se puede utilizar la ***fórmula de Sturguess*** para calcular el número de intervalos:

$k = 1 + 3{,}3 \log n$ siendo n: número de muestras consideradas.

Un número muy pequeño de clases puede ocultar la distribución real del conjunto de datos, mientras que un número muy grande puede dejar sin observaciones algunas clases.

Una buena práctica es la creación de clases de igual longitud.

Esto se obtiene tomando la diferencia entre los dos valores extremos del conjunto de datos y dividiéndola entre el número de clases. El resultado será aproximadamente la longitud del intervalo de cada clase. Hay casos en que este método no puede aplicarse, y se deberá tomar intervalos de clases diferentes.

Para establecer las fronteras de clase es necesario considerar la unidad más cercana con respecto a la cual se miden las observaciones, en nuestro ejemplo está redondeado a la libra más cercana, tomamos entonces la unidad decimal para establecer las fronteras, ej. 126,5.

Estas fronteras se conocen como límites verdaderos.

Dado que los pesos están tomados a la libra más cercana, pueden tomarse los límites de las clases: (118--126) (127--135), de esta forma las clases no se superponen. Esta forma de tomar las fronteras se conoce como límites de escritura.

Utilizaremos para realizar nuestro ejemplo los límites de tal forma que el límite superior de cada clase coincida con el límite inferior de la siguiente, adoptaremos como criterio que los intervalos se suponen cerrados por izquierda y abiertos por la derecha, es decir, en cada clase se incluyen los valores de la variable que sean mayores o iguales al límite superior, pero estrictamente menores que el límite superior.

Como excepción al criterio adoptado, en la última clase, el intervalo será cerrado en ambos extremos, si no fuera así, el valor máximo quedaría fuera del intervalo.

El método consta de los siguientes pasos:

1. Ordenar los datos de menor a mayor

119	125	126	128	132	135	135	135
136	138	138	140	140	142	142	144
144	145	145	146	146	147	147	148

149	150	150	152	153	154	156	157
158	161	163	164	165	168	173	176

2. Determinar el tamaño de la muestra

n = 40

3. Determinar el valor máximo y el valor mínimo de la variable

El peso mayor es 176 libras y el menor 119 libras,

4. Calcular el Rango o Recorrido

Recordemos que en general ***Rango*** es la diferencia entre el valor máximo y el valor mínimo de la variable en una serie estadística:

$$R = x_{max} - x_{min}$$

R =176-119=57 libras.

5. Calcular la cantidad de intervalos a utilizar

$k = 1 + 3{,}3 \log n$

$k = 1 + 3.3 \log 40 = 6{,}29 \approx 7$

6. Calcular la longitud de cada intervalo

En general:

Longitud de cada intervalo de clase =R/k

Si utilizamos 7 intervalos de clase, la longitud de cada uno será:

$L = 57/7 \approx 8$ aproximadamente.

La tabla de frecuencia correspondiente a nuestro ejemplo es:

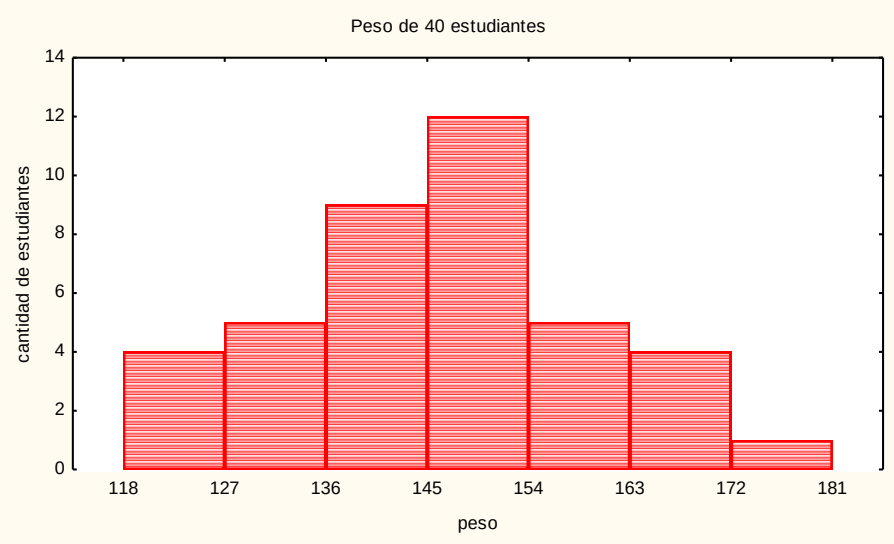
TABLA DE DISTRIBUCIÓN DE FRECUENCIAS

Para la variable: X: Pesos de 40 estudiantes (en libras)

Limite de escritura de la clase	Punto medio	fi
[118,127)	122,5	3
[127,136)	131,5	5
[136,145)	140,5	9
[145,154)	149,5	12
[154,163)	158,5	5
[163,172)	167,5	4
[172,181]	176,5	2
Total		Σfi=40

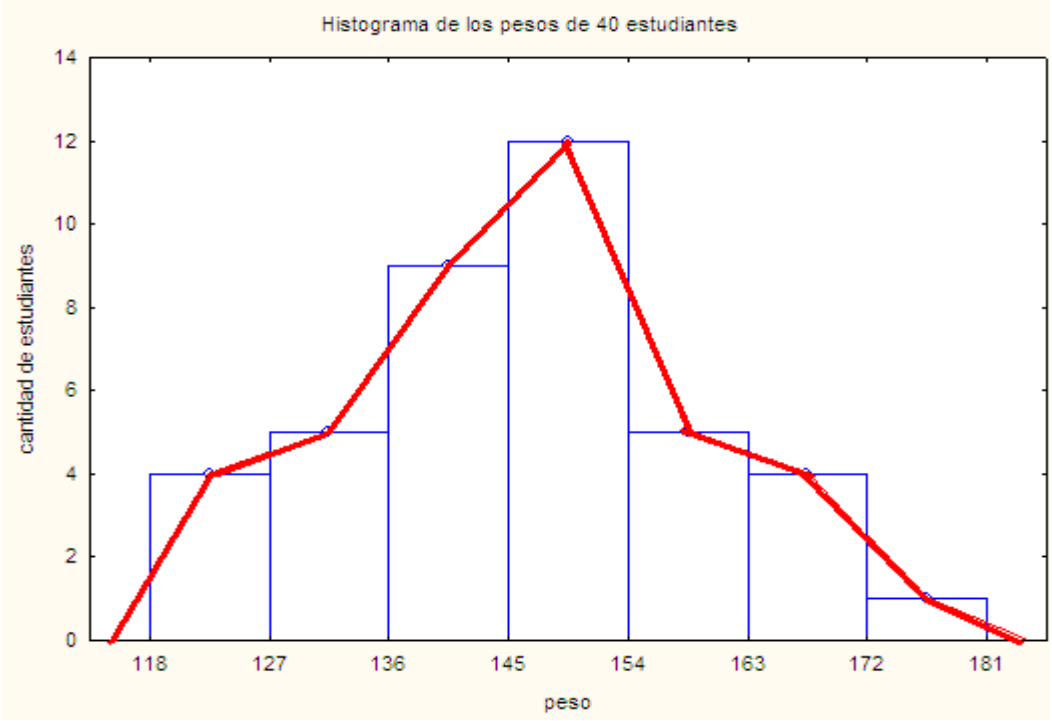
HISTOGRAMA DE FRECUENCIAS ABSOLUTAS

X: Pesos de 40 estudiantes (en libras)



POLÍGONO DE FRECUENCIAS

X: Pesos de 40 estudiantes (en libras)



Otras distribuciones útiles que consideraremos son: la ***distribución de frecuencia acumulada y la distribución de la frecuencia relativa acumulada u ojiva.***

FRECUENCIA ACUMULADA: es la frecuencia total de todos los valores menores que el límite verdadero superior de clase de un intervalo de clase.

Es decir que este gráfico muestra directamente cuántos de los elementos son menores que, la marca de clase.

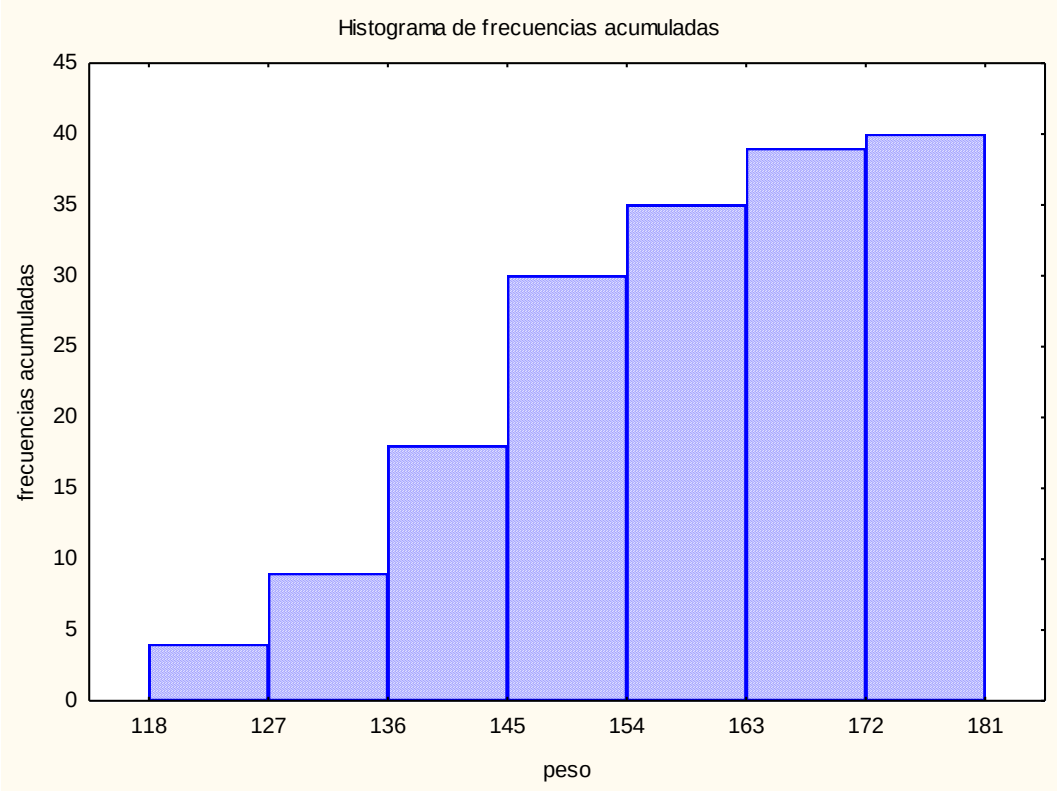
FRECUENCIA RELATIVA ACUMULADA: es la frecuencia acumulada dividia por la frecuencia total. Representa la proporción de observaciones cuyos valores son menos o iguales al límite superior de la clase.

TABLA DE DISTRIBUCIÓN DE FRECUENCIAS

Límite de escritura de la clase	punto medio	f	fr	Fa	Fra
[118 ,127)	122,5	3	0,075	3	0,075
[127 ,136)	131,5	5	0,125	8	0,200
[136 ,145)	140,5	9	0,225	17	0,425
[145 ,154)	149,5	12	0,300	29	0,725
[154 ,163)	158,5	15	0,125	34	0,850
[163, 172)	167,5	4	0,100	38	0,950
[172 ,181]	176,5	2	0,050	40	1,000
		Σf_i=40	Σf_{ri} = 1		

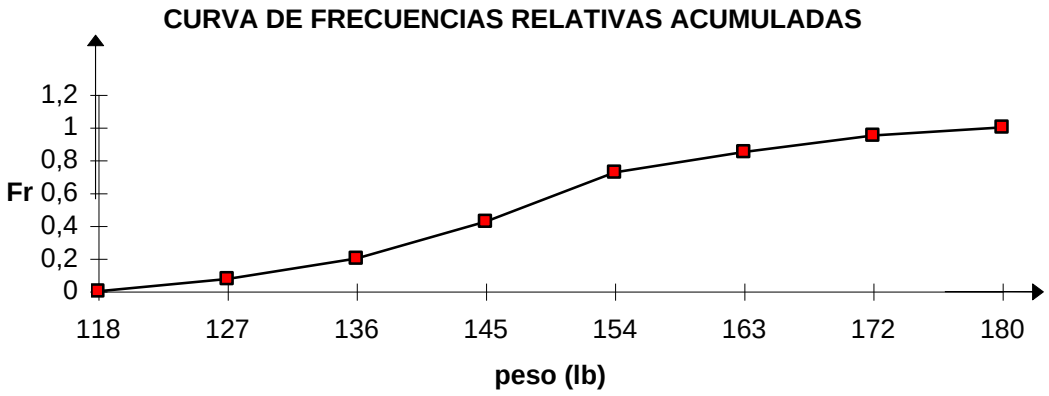
HISTOGRAMA DE FRECUENCIAS ACUMULADAS

Graficaremos la columna con las frecuencias acumuladas, obteniendo de esta forma un ***Histograma de frecuencias acumuladas.***



CURVA DE FRECUENCIAS RELATIVAS ACUMULADAS: Curva Ojiva

Llamamos ***ojiva*** al polígono de frecuencias acumuladas. Se obtiene uniendo con segmentos los puntos cuyas coordenadas son las abscisas correspondientes a los extremos superiores de cada clase y las ordenadas correspondientes a las frecuencias acumuladas (relativa o absoluta) hasta dicha clase.



II - ATRIBUTOS:

Recordemos que llamamos atributo a aquella característica de la unidad de análisis que no es susceptible de medición cuantitativa, en estos casos realizamos una clasificación en categorías.

Ejemplos:

Color de ojos, marca de un automóvil, título secundario, nivel social, sexo, grupo étnico que pertenece una persona, nivel de gravedad que posee una persona una determinada enfermedad.

Para presentar un atributo lo podemos hacer mediante una tabla o cuadro, que ofrece una visión numérica, sintética,y global de dicho atributo.

Estas tablas o cuadros constan de las siguientes partes:

1-**EL TITULO**: que debe responder a las ¿qué?, ¿dónde?, ¿cuándo?

2-**EL CUERPO**: que consta de encabezamiento de columnas, columna matriz, columna secundaria.

3-**EL PIE DEL CUADRO**: que consta de fuente de los datos, alguna nota o algún dato importante.

Los gráficos más utilizados para representar variables cualitativas o atributos son: el diagrama circular o gráfico de sectores, los gráficos de barras, que pueden ser verticales u horizontales, dentro de estos gráficos de barras debemos nombrar el diagrama de Pareto, muy utilizado hoy en día en calidad, y los pictogramas.

Ejemplo 6:

En la siguiente tabla se presenta el personal ocupado en la industria maquiladora de exportación, durante los años 2004 al 2008.

X: "***Personal ocupado en la industria maquiladora de exportación, durante los años 2004 al 2008"***

TABLA DE DISTRIBUCIÓN DE FRECUENCIAS

Año	Frecuencia absoluta Total f_i	Frecuencia Absoluta de Obreros	Frecuencia Absoluta de Técnicos	Frecuencia Absoluta de Empleados
2004	199	166	22	11
2005	211	173	25	13
2006	249	204	30	15
2007	305	248	36	21
2008	369	301	44	24
Total	1333	1092	157	84

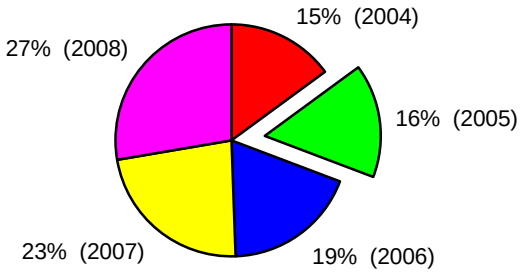
Fuente: Revista Expansión, 25 de Octubre de 2009.

DIAGRAMA CIRCULAR O DIAGRAMA DE SECTORES

Se representará el total del personal ocupado en la industria maquiladora de exportación durante los años 20044 al 2008. También podría haberse representado para cada año como estaba compuesto ese personal.

DIAGRAMA CIRCULAR

Total del Personal de 2004 a 2008



Fuente: Revista Expansión, 25 de Octubre de 2009

Nota: No deben ponerse los porcentajes dentro del gráfico.

DIAGRAMA DE PARETO

Una variante importante en los diagramas de barras es el diagrama de Pareto. Este diagrama tiene un uso muy amplio sobretodo por su valor para realizar comparaciones.

Las categorías están ordenadas de mayor a menor frecuencia, modo tal que en la parte izquierda del gráfico aparezca la categoría con mayor frecuencia, seguida por la segunda mayor frecuencia y así, sucesivamente. En este tipo de variables generalmente aparece una categoría denominada otros, en la cual van todos aquellas modalidades con poca frecuencia absoluta. Esta categoría siempre irá en el último lugar aunque aparezca con un poco más de frecuencia que alguna otra categoría.

Este tipo de diagramas debe se nombre al economista italiano V. Pareto

Ejemplo 7:

De 2000 circuitos de computadora revisados por el fabricante se obtuvieron los siguientes datos:

conexiones defectuosas	31
agujeros demasiado grandes	55
agujeros sin abrir	182
circuitos de tamaño incorrecto	5
otros	7

a) Confeccione una tabla de frecuencias.
b) Realice un diagrama de Pareto

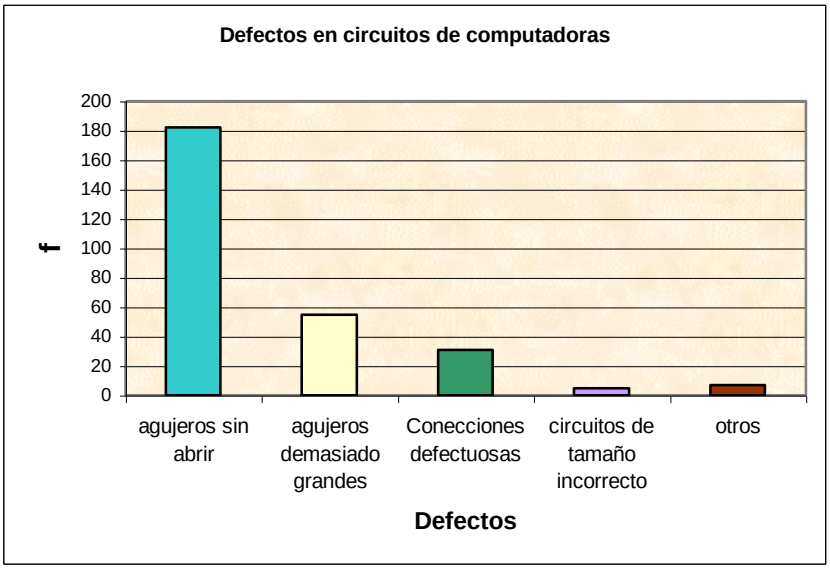
TABLA DE FRECUENCIAS

DEFECTO	f_i	F_i	fr_i	Fr_i
Conexiones defectuosas	*31*	*31*	*0,11*	*0,11*
agujeros demasiado grandes	*55*	*86*	*0,20*	*0,31*
agujeros sin abrir	*182*	*268*	*0,65*	*0,96*
circuitos de tamaño incorrecto	*5*	*273*	*0,02*	*0,98*
otros	*7*	*280*	*0,02*	*1,00*

Para realizar el diagrama de Pareto ordenamos las modalidades de mayor frecuencia a menor frecuencia y luego realizamos el diagrama de barras, denominado de Pareto.

DEFECTO	f_i	F_i	fr_i	Fr_i
agujeros sin abrir	*182*	*182*	*0,65*	*0,11*
agujeros demasiado grandes	*55*	*237*	*0,20*	*0,31*
Conexiones defectuosas	*31*	*268*	*0,11*	*0,96*
circuitos de tamaño incorrecto	*5*	*273*	*0,02*	*0,98*
otros	*7*	*280*	*0,02*	*1,00*

DIAGRAMA DE PARETO



FUENTE: Revista Expansión, 25 de Octubre de 2009.

GRÁFICO DE CAJA

Este Gráfico es también denominado ‘box and whiskers”, lo utilizaremos para representar variables cuantitativas cuyos datos están presentados en forma individual, ya que este gráfico permite identificar valores atípicos o outliers.

Utilidades del gráfico de caja:

- Nos proporciona la posición relativa de la mediana, los cuartiles y los extremos de la distribución.
- Nos proporciona información sobre valores atípicos, sugiriendo la necesidad o no de utilizar determinados estadísticos.
- Nos informa de la simetría o asimetría de la distribución.
- Se puede utilizar para comparar la misma variable en dos muestras distintas.

Explicaremos su construcción paso a paso para poder interpretarlo, aunque se puede realizar con cualquier software estadístico.

1. Se traza una línea horizontal de longitud proporcional al recorrido de la variable, que llamaremos eje. Los extremos del eje serán los valores mínimo y máximo de la distribución.
2. Paralelamente al eje se construye una caja rectangular con altura arbitraria y cuya base abarca desde el primer cuartil hasta el tercer cuartil. Como vemos, esta caja indica gráficamente el intervalo de variación del 50 % de valores centrales de la distribución.
3. La caja se divide en dos partes, trazando una línea a la altura de la mediana. Cada una de estas partes indica, pues, el intervalo de variabilidad de una cuarta parte de los datos.
4. El recorrido entre el valor mínimo y el primer cuartil, y el recorrido entre el tercer cuartil y el valor máximo se denominan bigotes. Si algún dato esta muy alejado de los valores centrales, estos bigotes se alargarían señalándonos así los valores atípicos u outliers.

Realizaremos ahora un ejemplo para ver como se interpreta el gráfico y para verificar sus utilidades.

Ejemplo 14:

La siguiente muestra estadística, contiene los pesos, en kilogramos de un grupo de sesenta personas.

Varones	55	64	70	74	75	70	62	93	60	62	70	71
	70	80	61	60	62	68	65	66	68	71	72	65
Mujeres	60	49	52	54	56	66	45	52	48	54	56	61
	46	50	52	53	56	68	47	50	53	57	60	64

X : ***“Pesos de 60 personas”(en kg.)***

GRÁFICO DE CAJA

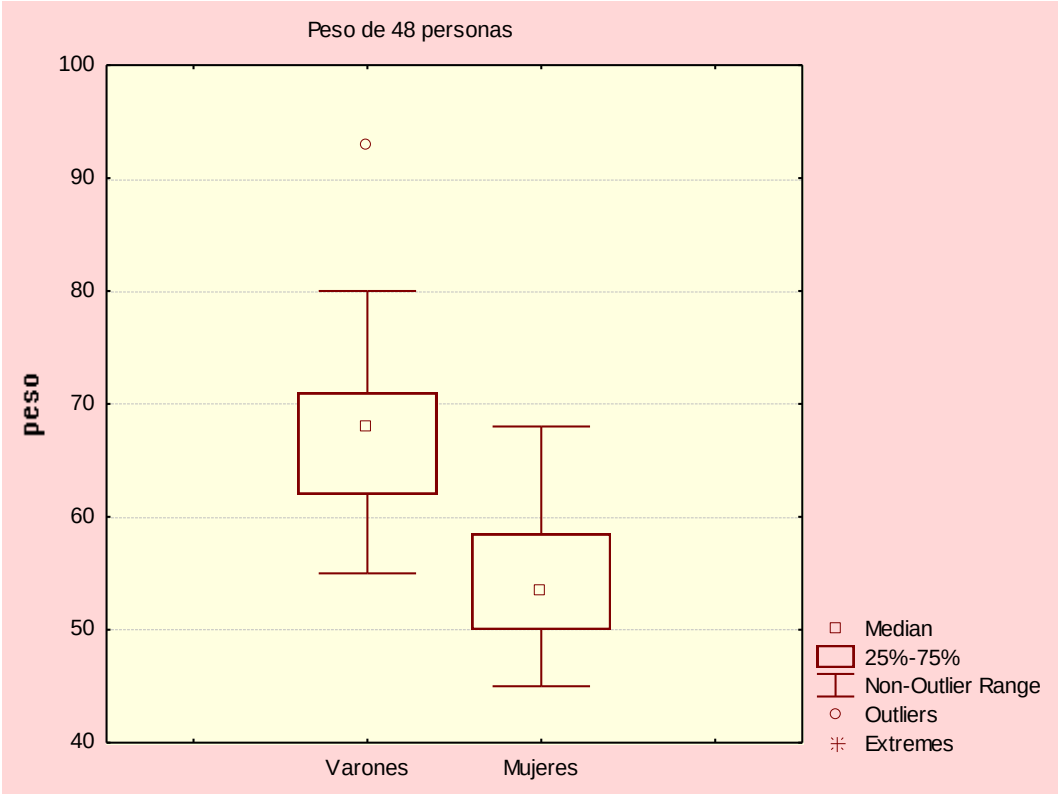
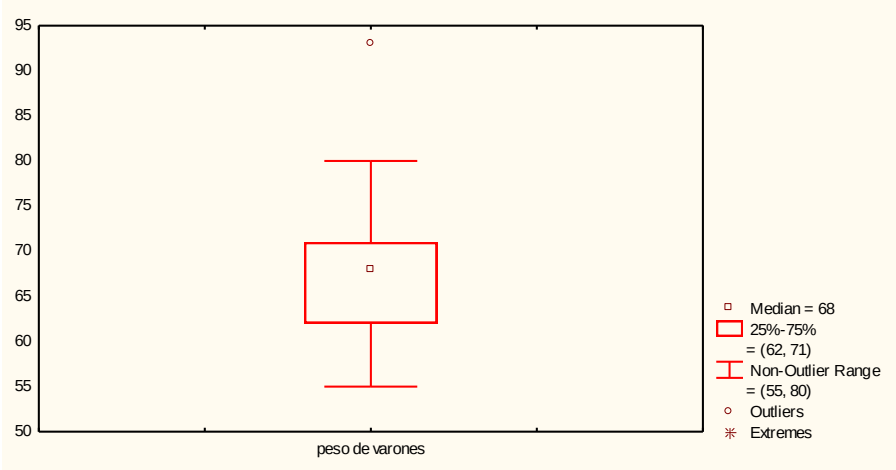


DIAGRAMA DE CAJA PARA LOS PESOS DE LOS VARONES



Se puede observar que el bigote superior es más largo que el bigote inferior por lo que podemos decir que la distribución es asimétrica. Además se observa un punto, que representa un valor atípico (93).

Se observa un símbolo □ adentro de la caja, éste me indica donde está ubicada la mediana.

Se deja al lector comparar la muestra con los pesos de las mujeres con la muestra de los pesos de los hombres.

MEDIDAS DE TENDENCIA CENTRAL DE UNA POBLACIÓN

La representación gráfica de los datos es una primera incursión en el análisis de datos, pero tiene sus limitaciones, si se desea describir los datos no es fácil hacerlo a partir del gráfico, e incluso, no es fácil comparar, por ejemplo, serie de datos para esto es fundamental resumir los datos

Las medidas de tendencia central dan un valor típico o representativo de un conjunto de datos(población ó muestra). La tendencia central de un conjunto de datos es la disposición de estos para agruparse ya sea alrededor del centro de la distribución de los mismos.

Se consideran principalmente 3 medidas de tendencia central: Moda, Mediana y Media. Si estas medidas se calculan considerando todas las observaciones correspondientes a una población reciben el nombre de **parámetros**.

MODA, MODO O VALOR MODAL

La moda de un conjunto de observaciones correspondientes a una población es el valor de la observación que ocurre con mayor frecuencia en la población.

Dado un conjunto de observaciones, estas deben ser ordenadas siempre de menor a mayor.

Una moda es un valor x_i tal que:

$$f(x_i) \geq f(x_{i-1}) \qquad f(x_i) \geq f(x_{i+1})$$

$f(x_i)$ es un máximo relativo

□□La moda muestra hacia que valor tienden los datos a agruparse. Puede suceder que una serie de datos halla más de una moda. En tal caso se denomina unimodal, bimodal, trimodal, etc. según el número de modas que presente.

□□Se puede considerar en algunos casos el valor de frecuencia mayor como la moda del conjunto de datos, en otros no podemos tomar dichos valores como representativos, será en estos casos el estadístico que analiza la situación quien decidirá.

□□El cálculo de la moda para datos individuales es sencillo, basta con buscar el valor de la variable que presente la máxima frecuencia absoluta.

□

Ejemplo 8:

De acuerdo con la revista Informes al Consumidor en su número de febrero de 2008, las cuotas anuales de las 40 compañías argentinas para un seguro de $25000 para hombre de 35 años de edad son las siguientes (en pesos), constituye la siguiente población:

82	85	86	87	87	89	89	90	91	91	92	93	94	95
95	95	95	95	97	98	99	99	100	100	101	101	103	103
103	104	105	105	106	107	107	107	109	110	110	111		

TABLA DE DISTRIBUCIÓN DE FRECUENCIAS

OBSERVACIONES	f_i	F_i	F_{ri}
82	1	1	0,025
85	1	2	0,050
86	1	3	0,075
87	2	5	0,125
89	2	7	0,175
90	1	8	0,200
91	2	10	0,250
92	1	11	0,275
93	1	12	0,300
94	1	13	0,325
95	5	18	0,450
97	1	19	0,475
98	**1**	**20**	**0,500**
99	2	22	0,550
100	2	24	0,600
101	2	26	0,650
103	3	29	0,725
104	1	30	0,750
105	2	32	0,800
106	1	33	0,825
107	3	36	0,900
109	1	37	0,925
110	2	39	0,975
111	1	40	1
	$\sum_{i=1}^{24} f_i = 40$		

Mayor frecuencia Moda: $\mathbf{X_{MO} = \$\ 95}$

Interpretación: La cuota anual para un seguro de $25000 para un hombre de 35 años de edad, que se presenta con mayor frecuencia es de $95

MEDIANA

Es como su nombre lo indica el valor central del conjunto de las observaciones que constituyen una población. Cuando todas las observaciones se ordenan en forma creciente, el 50% de las observaciones son menores o iguales que la mediana y el otro 50% son mayores o iguales que la mediana. La mediana se representa con la notación $x_{0,5}$, tal que $F(x_{0,5})= 0,5$.

Ejemplo 9: La mediana es $ 98 porque F(98)=0,50

Interpretación: El 50% de las cuotas anuales observadas por el pago de un seguro de $25.000 para un hombre de 35 años son de $98 o menos y el otro 50% son $98 o mayores

MEDIA O MEDIA ARITMÉTICA

La media de las observaciones $x_1, x_2,..., x_N$ de una población es el promedio aritmético de estas y se denota por:

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

Si tenemos datos que se repiten podemos utilizar la fórmula

$$\mu = \frac{\sum_{i=1}^{k} x_i \cdot f_i}{n}$$

Siendo k la cantidad de valores distintos que toma la variable X.

Es el valor que tomaría la variable si estuviera uniformemente repartida entre todos los individuos que forman la población (corresponde al concepto de centro de gravedad en física).

La media es una medida de tendencia central que utiliza todos los datos. Sin embargo debido a que cualquier observación se emplea para el cálculo, el valor de la media puede afectarse de manera desproporcionada por la existencia de algunos valores extremos.

Ejemplo 10 : Consideremos la población de todos los estudiantes de Medicina de la provincia de Mendoza que cursan el segundo año de la carrera en el presente año 2012 y vamos a considerar la variable “cantidad de materias aprobadas”

Cantidad de materias aprobadas	Cantidad de estudiantes (frecuencia)
4	10
5	30
6	42
7	66
8	62

9	90

$$\mu = \frac{10.4 + 30.5 + 42.6 + 66.7 + 62.8 + 90.9}{300} = 7{,}37 \text{ materias}$$

MEDIDAS DE TENDENCIA CENTRAL DE UNA MUESTRA

Consideraremos las tres medidas de tendencia central: moda, mediana y media de una muestra. Estas medidas calculadas sobre una muestra reciben el nombre de **estadísticos**.

Moda Muestral

Es el valor de la variable que tienen mayor frecuencia.

Consideremos una muestra aleatoria de tamaño 10 de la población considerada en el ejemplo 10.

Muestra 1= $\{7,7,7,8,8,8,8,8,9,9\}$

En este caso la moda es Mo= 8

Si consideramos otra muestra aleatoria de tamaño 10 de la población del ejemplo 10.

Muestra 2= $\{6,6,6,6,7,7,8,8,9,9\}$

En este caso la moda es Mo= 6

Se puede observar que la moda es una variable, que toma diferentes valores en cada una de las muestras.

Ventajas y desventajas de la moda:

La moda presenta como ventaja que se puede utilizar como una medida de tendencia central para cualquier tipo de datos, ya sean estos de tipo cualitativos, como datos cuantitativos.

Otra ventaja muy importante es que igual que la mediana, no es afectada indebidamente por valores extremos.

A pesar que tiene las mismas ventajas que la mediana, no es tan utilizada como la media aritmética o mediana y esto se debe a que muchas veces la distribución de los datos no tiene moda, y otras contiene más de una moda y resulta difícil interpretar o sacar una conclusión a partir de distribuciones multimodales.

Mediana Muestral

Si el número de observaciones es impar, la mediana es el valor de la observación que se encuentra a la mitad del conjunto ordenado. Si n es impar la mediana ser el valor de la observación que ocupa el lugar (n+1)/2.

$$\text{Me} = x_{\left(\frac{n+1}{2}\right)}$$

Si el número de observaciones es par se considera la mediana como el promedio aritmético de los valores de las dos observaciones que ocupan el lugar n/2 y (n+1)/2 del conjunto ordenado.

Es decir: $$\text{Me} = \frac{x_{\frac{n}{2}} + x_{\frac{n+1}{2}}}{2}$$

En el cálculo de la mediana la existencia de valores extremos no afecta su valor.

Por lo tanto si un conjunto contiene valores extremos y un número alto de observaciones, la mediana puede ser una medida de tendencia central mucho más deseable que la media.

En la Muestra 1 la mediana es:

Me = $\frac{x_{(5)} + x_{(6)}}{2} = \frac{8+8}{2} = 8$

En la Muestra 2 la mediana es:

Me= $\frac{x_{(5)} + x_{(6)}}{2} = \frac{7+7}{2} = 7$

Si consideramos nuevamente el ejemplo 8 donde hemos calculado la moda, como n es par, para saber la posición del valor de la mediana, buscamos en la columna de la frecuencia acumulada las posiciones n/2 y (n+1)/2, luego se ve los valores de la variable correspondientes y se calcula el promedio entre ellos, obteniendo así el valor de la mediana que deja por encima y por debajo de él el 50% de las observaciones.

Ventajas y desventajas de la mediana

La mediana se puede obtener en un conjunto de datos cuantitativos, pero también, en un conjunto de datos cualitativos, donde se utiliza una escala ordinal. Si se trabaja con estas clases de observaciones, se puede enunciar como ventaja, que siempre existe y es única, al igual que en el caso de la media, existe una sola mediana para un conjunto de observaciones.

La determinación de la mediana no es afectada por la existencia de valores extremos. Por lo tanto, si un conjunto de datos contiene valores extremos, la mediana puede ser una medida de tendencia central mucho más representativa que la media.

Como desventaja se puede decir que para su cálculo se utiliza un solo valor del conjunto de datos o como máximo dos, en comparación con la media que se utiliza todos los datos para su determinación.

Media Muestral

Es el promedio aritmético de las observaciones de una muestra.

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

Consideremos nuevamente la población del ejemplo 10 y dos muestras aleatorias de esta población:

Muestra 1= $\{7,7,7,8,8,8,8,8,9,9\}$ y Muestra 2= $\{6,6,6,6,7,7,8,8,9,9\}$

Calcularemos la media de cada una de estas muestras:

$$\overline{X}_1 = \frac{3.7 + 5.8 + 2.9}{10} = 7{,}9 \qquad \overline{X}_2 = \frac{4.6 + 2.7 + 2.8 + 2.9}{10} = 7{,}2$$

Se puede observar que cada una de las dos medias muestrales no coinciden con la media poblacional.

Ventajas y desventajas de la media aritmética

La media aritmética es una medida de tendencia central que tiene importantes ventajas. En todo conjunto de datos cuantitativos se puede encontrar siempre su media, es una medida que puede calcularse y es única.

Otra importante propiedad es que la media aritmética utiliza todos los valores del conjunto de datos para su cálculo, esta es una propiedad deseable. Sin embargo debido a que todas las observaciones se emplean para su obtención, el valor de la media puede ser afectada de manera desproporcionada por la existencia de valores extremos, que no son representativos del resto de los datos. En este caso no sería la medida de resumen más conveniente, se podría utilizar otras como la mediana o la moda.

Otra desventaja que presenta es que puede emplearse como medida de resumen sólo para variables numéricas o cuantitativas, pero no resultará adecuada para variables categóricas o cualitativas, en sus dos escalas de medición, nominal u ordinal. Este tipo de variables toman categorías con las cuales no se pueden realizar sumas ni cocientes.

Ejemplo 11

Una muestra de 11 pacientes admitidos para diagnóstico y evaluación en un departamento psiquiátrico en un hospital general experimentó los siguientes tiempos de permanencia, en días.

X: "***Tiempo de permanencia de un paciente en el departamento psiquiátrico, en días"***

Paciente	1	2	3	4	5	6	7	8	9	10	11
Tiempo en días	14	11	12	14	13	32	12	12	11	13	14

$$\overline{\mathbf{x}} = \frac{\sum_{i=1}^{n} \mathbf{x_i}.\mathbf{f_i}}{\mathbf{n}} = \frac{11\bullet 2 + 12\bullet 3 + 13\bullet 2 + 14\bullet 3 + 32}{11} = 14{,}36$$

El tiempo medio de permanencia de un paciente en el departamento psiquiátrico es de 14,36 días. Sin embargo si se calcula el tiempo medio y excluimos el valor 32 días, nos da 12,6 días. Se observa la existencia de un valor extremo, 32, el cual

distorsiona el valor obtenido para la media. Este es un ejemplo en el cual si se utiliza otra medida de tendencia central el resultado sería más representativo del conjunto de observaciones.

CUANTILES MUESTRALES

Dada una muestra ordenada en forma creciente, el valor que divide al conjunto de datos en dos partes iguales es la mediana.

Por extensión, si preferimos tener una descripción más detallada de la variabilidad de los valores individuales, se puede dividir los datos en otras cantidades de partes iguales. Por ejemplo, en cien, en diez o en cuatro partes iguales, llamando a estas medidas ***percentiles, deciles y cuartiles*** respectivamente.

PERCENTILES

Al dividir los datos en cien partes iguales quedan definidos los percentiles: que se representan por p_1, p_2, ..., p_{99}, la mediana muestral es el percentil de orden 50.

La fórmula para obtener el ***lugar*** k-ésimo percentil, siendo n: el número de observaciones:

$$L_{pk} = k\,\frac{(n+1)}{100}$$ Fórmula para obtener el **lugar del k-ésimo percentil**

Así buscando en la lista ordenada de los valores o en la columna de la frecuencia acumulada, se ve el valor de la variable correspondiente.

En caso de no ser un valor entero se calcula por interpolación lineal el valor del percentil. En la práctica, tomaremos el valor de la variable correspondiente a la frecuencia acumulada inmediata superior al valor obtenido de l_{pk}

Es decir el valor del percentil será: $\mathrm{P_k} = X_{\left(k.\frac{n+1}{100}\right)}$

DECILES

Análogamente los valores que dividen los datos en 10 partes iguales quedan definidos los deciles y se representan por D1, D2, ..., D9.

La fórmula para obtener el lugar del k-ésimo decil, siendo n el número de observaciones, es:

$$L_{Dk} = k\,\frac{(n+1)}{10}$$ Fórmula para obtener **el lugar del k-ésimo decil**

Así buscando en la lista ordenada, de los valores o en la columna de la frecuencia acumulada, se ve el valor de la variable correspondiente.

En caso de no ser un valor entero se calcula por interpolación lineal el valor del decil. En la práctica, tomaremos el valor de la variable correspondiente a la frecuencia acumulada inmediata superior al valor obtenido de lDk.

Es decir el valor de los deciles será:

$\mathrm{D_k} = X_{\left(k.\frac{n+1}{10}\right)}$

CUARTILES

Análogamente si se dividen los datos en cuatro partes iguales, quedan definidos los cuartiles.

Estos valores se representan Q_1, Q_2, Q_3, se llaman primero, segundo y tercer cuartil, respectivamente, el valor correspondiente al segundo cuartil, Q_2, coincide con el valor de la mediana.

La fórmula para obtener el ***lugar del k-ésimo cuartil***, siendo n el número de observaciones, es:

$$LQk = k \frac{(n+1)}{4}$$ Fórmula para obtener **el lugar del k-ésimo cuartil.**

Así buscando en la lista ordenada, de los valores o en la columna de la frecuencia acumulada, se ve el valor de la variable correspondiente.

En caso de no ser un valor entero se calcula por interpolación lineal el valor del cuartil. En la práctica, tomaremos el valor de la variable correspondiente a la frecuencia acumulada inmediata superior al valor obtenido de lQk.

Es decir el valor del los cuartiles será:

$$Q_k = x_{k\,\frac{(n+1)}{4}}$$

Si queremos determinar un valor debajo del cuál se halle el 25% de los datos, calculamos el 1er cuartil Q_1, si queremos calcular un valor debajo del cuál se halle el 50% de los datos calculamos el segundo cuartil Q_2 y si nos interesa calcular un valor debajo del cuál se halle el 75% de los datos, calculamos el tercer cuartil Q_3.

Si se obtiene valores fraccionados, hacemos una interpolación lineal entre los dos valores correspondientes a las dos observaciones de la muestra.

Ejemplo 11:

Las siguientes cifras son el importe del consumo de 15 personas en un restaurant, en orden ascendente, en unidad pesos, 100, 100, 250, 250, 250, 350, 400, 530, 900, 1250, 1350, 2450, 2750, 3090, y 4100. Determinar a) el primer quartil, b) el primer decil, c) el 40-ésimo percentil

$$a) L_{Q_1} = \frac{n+1}{4} = \frac{15+1}{4} = 4$$

$$Q_1 = x_{\frac{n+1}{4}} = x_4 = 250$$

$$b) L_{D_1} = \frac{n+1}{10} = \frac{16}{10} = 1{,}6 \cong 2$$

$D_1 = 100$

¿Qué significa que el primer cuartil tome el valor $250?

Significa que el 25 % de los valores son inferiores o iguales a $250 y el 75% restante es mayor o igual a $250.

¿Qué significa que el primer decil tome el valor $100?

Significa que el 10 % de los valores son inferiores o iguales a $100 y el 90 % restante es mayor o igual a $100.

c) $L_{P_{40}} = \frac{p.n+p}{100} = \frac{40.(n+1)}{100} = \frac{2}{5}.(n+1) = 6,4 \cong 7$

$P_{40} = x_{(7)} = 400$

MEDIDAS DE TENDENCIA CENTRAL PARA DATOS AGRUPADOS

MODA

Para calcular la moda es necesario identificar primero, la clase modal, que es aquella que tenga mayor frecuencia absoluta.

Una vez identificada la clase modal el valor de la moda dentro del intervalo se halla mediante la fórmula:

$$x_{Mo} = L_m + \left[\frac{\Delta_1}{\Delta_1 + \Delta_2} \right].c$$

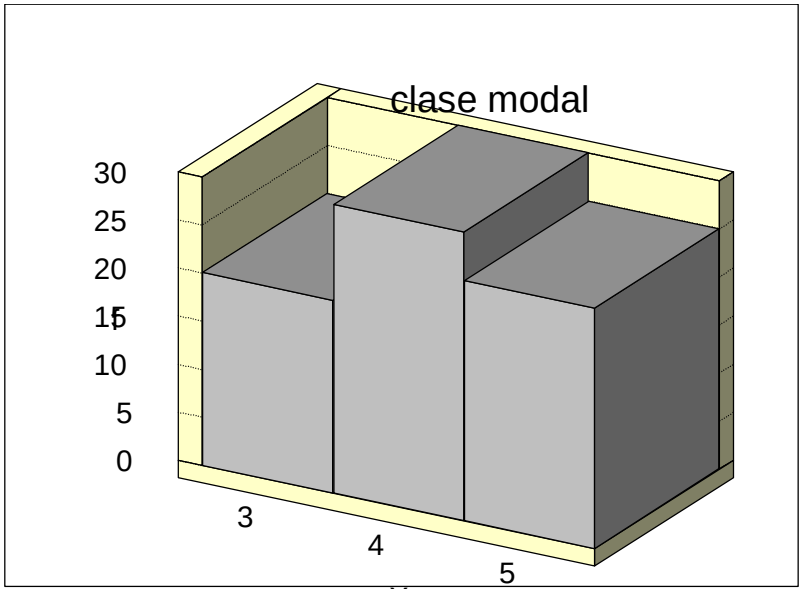
Siendo:

L : límite inferior de la clase modal

Δ_1 : valor absoluto de la diferencia entre la frecuencia de la clase modal y la premodal.

Δ_2: valor absoluto de la diferencia entre la frecuencia de la clase modal y la posmodal

c: longitud de la clase modal



MEDIA ARITMÉTICA MUESTRAL

Para calcular la media para datos agrupados utilizaremos la formula:

$$\bar{x} = \frac{f_1 .m_1 + ... + f_k .m_k}{f_1 + f_2 + ... + f_k} = \frac{1}{n} . \sum_{i=1}^{k} f_i .m_i$$

Siendo:
mi: punto medio de la clase i
fi . mi : valor total de observaciones que corresponde a la clase i
k: número de clases

MEDIANA MUESTRAL

Hay dos métodos para localizar la mediana, el método gráfico y el método de interpolación algebraico.

En el método gráfico, el valor de la mediana se halla por interpolación de la curva ojiva de la distribución.

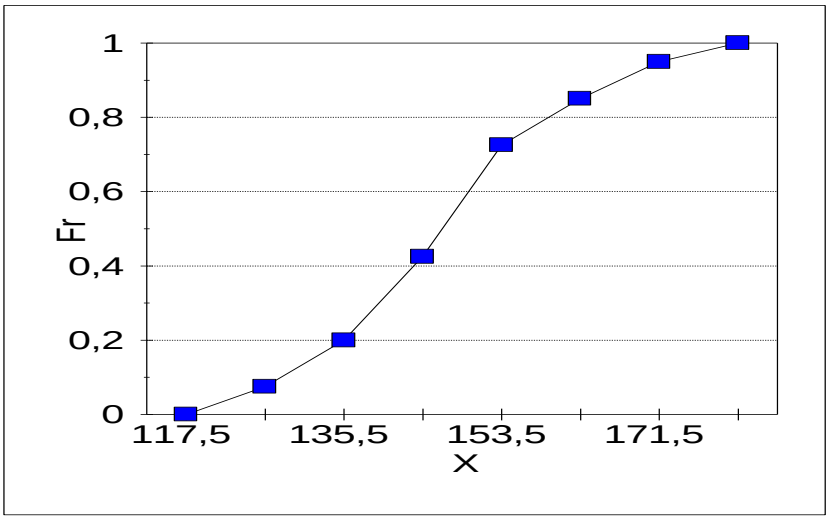
Trazamos una línea horizontal partiendo del eje vertical por n/2 ó 0,5 según sea la escala de frecuencia absoluta o frecuencia relativa hasta la ojiva, luego trazamos una recta perpendicular a la escala horizontal para localizar el valor de la mediana.

Ejemplo 12:

Hallar la mediana de los pesos de 40 estudiantes de la State University, dados por el siguiente cuadro:

Peso (libra)	fi	Fi	fri	Fri
117,5 - 126,5	3	3	0,075	0,075
126,5 -135,5	5	8	0,125	0,200
135,5 - 144,5	9	17	0,225	0,425
144,5 - 153,5	12	29	0,300	0,725
153,5 - 162,5	5	34	0,125	0,850
162,5 - 171,5	4	38	0,100	0,950
171,5 - 180,5	2	40	0,050	1
	$\sum_{i=1}^{n} f_i = 40$			

CURVA OJIVA Y MEDIANA



Método Algebraico, el valor de la mediana se obtiene ubicando primero la clase mediana. La fórmula para obtener dicho lugar, siendo n el número de observaciones es:

$$L_{Me} = \frac{(n+1)}{2}$$

así, buscando el valor obtenido en la columna de la frecuencia acumulada, se ve el intervalo correspondiente a la clase mediana. Si el número obtenido en L_{Me} no existe, se toma el inmediato superior.

Una vez identificada la clase mediana el valor de la mediana dentro del intervalo se halla mediante la fórmula:

$$Me = L_{0,5} + \left[\frac{\frac{n}{2} - F_p}{f_{0,5}} \right].c$$

Siendo

$L_{0,5}$: límite inferior de la clase mediana.
F_p : frecuencia acumulada de la clase anterior a la clase mediana.
$f_{0,5}$: frecuencia absoluta de la clase mediana.
c: longitud de la clase mediana
n: tamaño de la muestra

Hallar la mediana en nuestro ejemplo 12:

$$Me = 144,5 + \left[\frac{20 - 17}{12} \right].8 = 146,8 libras$$

Conclusión el 50% de los alumnos pesan 146,8 libras o menos, y el otro 50% pesan 146,8 libras o más.

CUANTILES MUESTRALES PARA DATOS AGRUPADOS

PERCENTILES

Hay dos métodos para localizar el **percentil k**, el método gráfico y el método algebraico.

En método gráfico, el valor del percentil k se halla mediante la observación de la ojiva de la distribución.

En el método algebraico, el valor del percentil k se obtiene ubicando primero, la clase del percentil k. La fórmula para obtener dicho lugar, siendo n el número de observaciones es:

$$LPk = k \frac{(n+1)}{100}$$ Fórmula para obtener el **lugar del k-ésimo percentil**

Así, buscando el valor obtenido, LPk, en la columna de la frecuencia acumulada, se puede obtener el intervalo correspondiente a la clase k, en la columna de los valores de la variable. Si el número obtenido en LPk, no existe, se toma el inmediato superior.

Una vez identificada la clase del percentil k el valor del percentil k dentro del intervalo se halla mediante la fórmula:

$$P_k = L_{\inf .Pk} + \left[\frac{\frac{k.n}{100} - F_{ant.Pk}}{f_{pk}} \right].c$$

Siendo:

L_{infPk} : límite inferior de la clase del percentil k.

$F_{ant\ Pk}$: frecuencia acumulada de la clase anterior a la clase del percentil k.

f_{Pk} : frecuencia absoluta de la clase del percentil k.

c: longitud de la clase del percentil k.

n: tamaño de la muestra.

DECILES:

Hay dos métodos para localizar el ***decil k***, el método gráfico y el método algebraico.

En método gráfico, el valor del decil k se halla mediante la observación de la ojiva de la distribución.

En el método algebraico, el valor del decil k se obtiene ubicando primero, la clase del decil k. La fórmula para obtener dicho lugar, siendo n el número de observaciones es:

$$LDk = k \frac{(n+1)}{100}$$ Fórmula para obtener el **lugar del k-ésimo decil**

Así, buscando el valor obtenido, LDk, en la columna de la frecuencia acumulada, se puede obtener el intervalo correspondiente a la clase k, en la columna de los valores de la variable. Si el número obtenido en LDk, no existe, se toma el inmediato superior. Una vez identificada la clase del decil k el valor del decil k dentro del intervalo se halla mediante la fórmula:

$$D_k = L_{\inf} + \left[\frac{\frac{k.n}{10} - F_{ant}}{f_{Dk}} \right].c$$

Siendo:

L_{inf}: límite inferior de la clase del decil k.

F_{ant} : frecuencia acumulada de la clase anterior a la clase del decil k.

f_{Dk} : frecuencia absoluta de la clase del decil k.

c: longitud de la clase del decil k.

n: tamaño de la muestra

CUARTILES:

Como la mediana es el 2do cuartil o el 5to decil o el 50-ésimo percentil. De forma análoga que calculamos la mediana calcularemos los cuartiles.

Por lo tanto hay dos métodos para localizar el ***cuartil k***, el método gráfico y el método algebraico.

En método gráfico, el valor del cuartil k se halla mediante la observación de la ojiva de la distribución.

En el método algebraico, el valor del cuartil k se obtiene ubicando primero, la clase del cuartil k. La fórmula para obtener dicho lugar, siendo n el número de observaciones es:

$$LQk = k\,\frac{(n+1)}{4}$$

Fórmula para obtener el **lugar del k-ésimo cuartil.**

Así, buscando el valor obtenido en la columna de la frecuencia acumulada, se puede obtener el intervalo correspondiente a la clase k, en la columna de los valores de la variable. Si el número obtenido en LQk, no existe se toma el inmediato superior.

Una vez identificada la clase del cuartil k el valor del cuartil k dentro del intervalo se halla mediante la fórmula:

$$Q_k = L_{inf} + \left[\frac{\frac{k.n}{4} - F_{ant}}{f_{Q_k}}\right] c$$

Siendo:

L_{inf} : límite inferior de la clase del cuartil k.
F_{ant} : frecuencia acumulada de la clase anterior a la clase del cuartil k.
f_{Qk} : frecuencia absoluta de la clase del cuartil k.
c: longitud de la clase del cuartil k.
n: tamaño de la muestra.

A los cuartiles, deciles, percentiles y otros valores obtenidos por subdivisiones análogas de los datos se los llama en forma general cuantiles.

MEDIDAS DE DISPERSIÓN DE UNA POBLACIÓN

Las medidas de tendencia central nos indican los valores alrededor de los cuales se distribuyen los datos.

Las medidas de dispersión nos proporcionan una medida del mayor o menor agrupamiento de los datos respecto a los valores de tendencia central.

Todas las medidas de dispersión son valores mayores o iguales a cero, indicando un valor cero, la ausencia de dispersión.

Un promedio puede ser engañoso a menos que vaya acompañado de otra información que nos diga la amplitud o sus desviaciones con relación al promedio.

♦ RECORRIDO O RANGO

El rango de las observaciones en un conjunto de datos es la diferencia entre el valor más grande y el más pequeño.

$$R = x_{máx} - x_{mín}$$

El rango proporciona una rápida indicación de la variabilidad existente entre las observaciones de un conjunto de datos. Sin embargo, debe usarse con precaución ya que su valor es función únicamente de dos valores extremos pertenecientes al conjunto, debe evitarse el uso del rango como medida de variabilidad, cuando el número de observaciones en un conjunto es grande o cuando éste contenga algunas observaciones cuyo valor sea relativamente grande

Para muchos problemas tiene una mayor utilidad determinar el recorrido entre dos valores cuantiles que entre dos valores extremos.

La diferencia entre el tercer y primer percentil, recibe el nombre de ***recorrido intercuantil***, sólo incluye el 50% central de la distribución.

La diferencia entre los percentiles noveno y décimo recibe el nombre de ***recorrido interdecil***, toma el 80% central de la distribución.

♦ VARIANZA POBLACIONAL

La varianza de las observaciones $x_1, x_2, ..., x_N$ de una población es el promedio del cuadrado de las distancias entre cada observación y la media poblacional.

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

♦ DESVIACIÓN ESTÁNDAR POBLACIONAL

La raíz cuadrada de la varianza se denomina ***desviación estándar***.

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

La varianza y la desviación estándar no son medidas de variabilidad distintas, debido a que esta última no puede determinarse a menos que se conozca la varianza.

El valor de la varianza puede sufrir un cambio muy desproporcionado aún más que la media, por la existencia de algunos valores extremos en el conjunto de datos.

A menudo se prefiere la desviación estándar con relación a la varianza, porque se expresa en las mismas unidades físicas de las observaciones.

COEFICIENTE DE VARIACIÓN POBLACIONAL

A menudo nos interesa comparar la variabilidad entre dos o más poblaciones.

Puede hacerse esto con sus respectivas varianzas o desviaciones estándar, cuando las variables se dan en las mismas unidades, y sus medias son aproximadamente iguales.

Cuando no sucede esto utilizamos una medida relativa de variabilidad llamada coeficiente de variabilidad.

El ***coeficiente de variabilidad*** es la razón entre la desviación estándar y la media.

$$C.V. = \frac{\sigma}{\mu}$$

Esta medida es independiente de las unidades utilizadas. Por esta razón es útil para comparar distribuciones donde las unidades pueden ser diferentes.

Un inconveniente del coeficiente de variación es que deja de ser útil cuando x esta próxima a cero.

MEDIDAS DE DISPERSIÓN DE UNA MUESTRA

Las medidas de dispersión que se calcularon para una población se pueden calcular en una muestra. Estas son: rango, varianza, desviación estándar y coeficiente de variación. Estas medidas calculadas en una muestra reciben el nombre es **estadísticos.**

Rango de una muestra

$R = x_{máx} - x_{mín}$

VARIANZA MUESTRAL $$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

DESVIACIÓN ESTÁNDAR MUESTRAL $$S = \sqrt{\frac{\sum_{i}(x_i - \bar{x})^2}{n-1}}$$

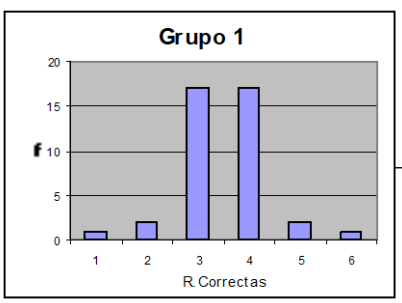
COEFICIENTE DE VARIACIÓN $$C.V. = \frac{S}{\bar{x}}$$

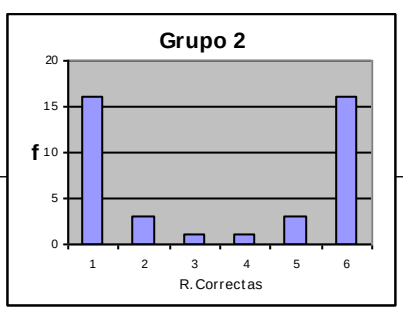
Ejemplo 13: Para ver sus aplicaciones analizaremos tres muestras de 40 alumnos cada una, a los que se les tomó una evaluación de seis preguntas. Los x_i indican el número de respuestas correctas y los f_i indican la cantidad de alumnos que obtuvieron dicho valor.

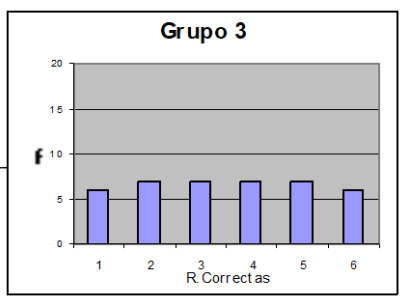
x_i	f_i
1	1
2	2
3	17
4	17
5	2
6	1

x_i	f_i
1	16
2	3
3	1
4	1
5	3
6	16

x_i	f_i
1	6
2	7
3	7
4	7
5	7
6	6







Las tres distribuciones tienen la misma media aritmética, 2,5 puntos ¿pero podemos afirmar que hay homogeneidad entre los grupos?. Gráficamente vemos que el valor de la media aritmética no es suficiente para describir cada una de las situaciones.

Para precisar mejor lo que denominamos como "dispersión" podemos calcular unos estadísticos que nos den más información, sin necesidad de representar los datos.

VARIANZA PARA DATOS AGRUPADOS

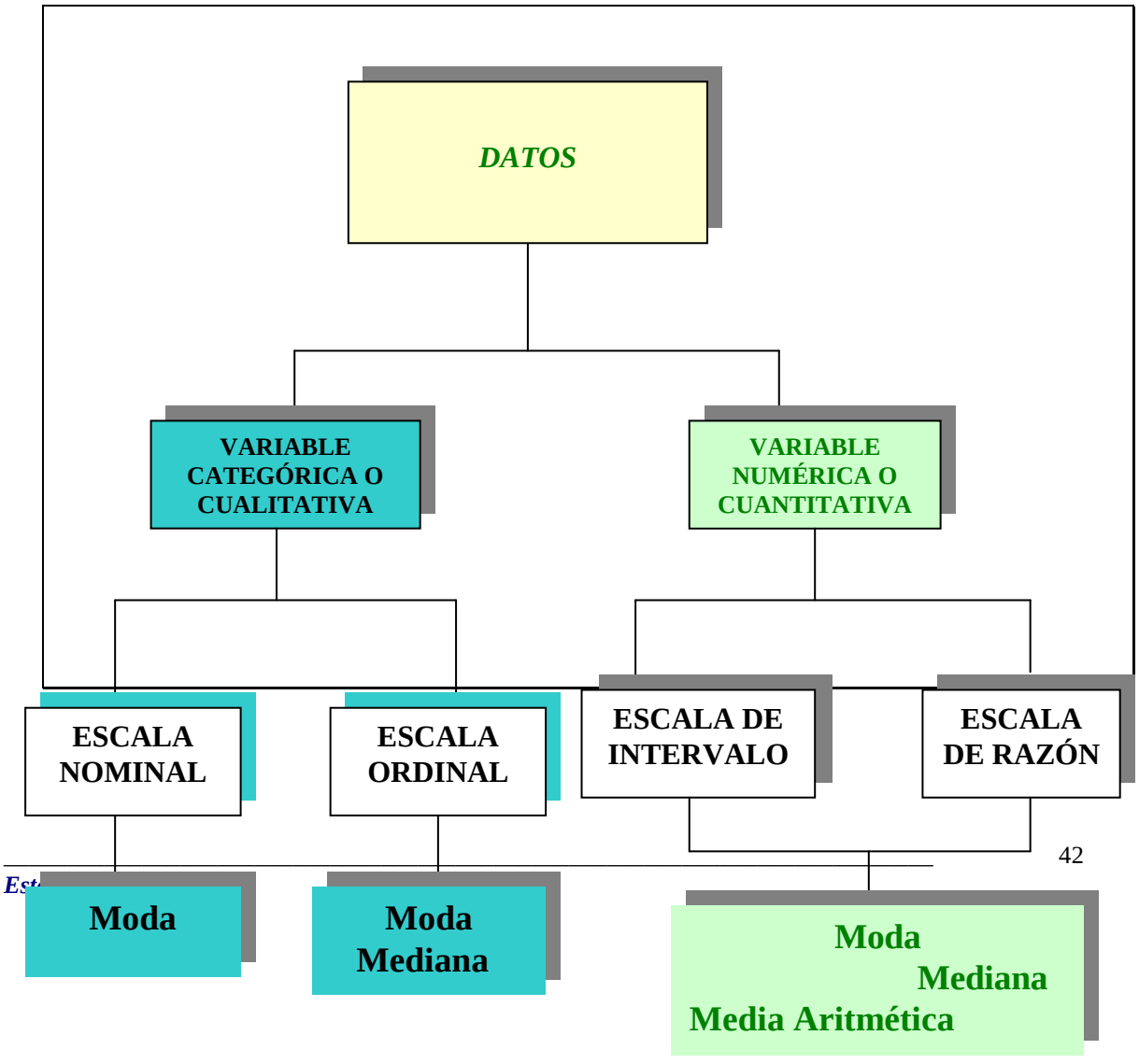
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{k} f_i(m_i - \bar{x})^2$$

Siendo:
m_i : punto medio de la clase
k: número de clases

DESVIACIÓN ESTÁNDAR PARA DATOS AGRUPADOS:

$$s = \sqrt{\frac{\sum_{i=1}^{k} f_i(m_i - \bar{x})^2}{n-1}}$$

***Resumen**: Estamos en condiciones de resumir en un gráfico que escala debo utilizar con cada variable y que medidas de tendencia central se pueden obtener con cada escala.*



Ejemplo 14:

Se ha medido la vida, en horas, de cincuenta lámparas incandescentes, obteniendo la siguiente muestra:

1067	919	1196	785	1126	936	918	1156	920	948
855	1092	1162	1170	929	950	905	972	1035	1045
1157	1195	1195	1240	1122	938	970	1237	956	1102
1022	978	832	1009	1157	1151	1009	765	958	902
923	1233	811	1217	1085	896	958	1211	1037	702

a) Agrupe los datos en una tabla de distribución de frecuencias con intervalos de clase de 100 horas, comenzando con la clase [699,5 ; 799,5).

Consideraremos la variable en estudio:

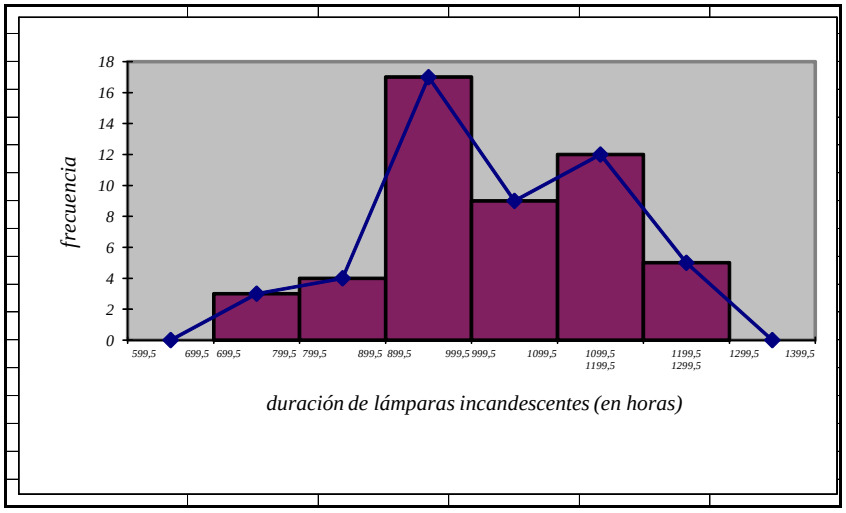
X: ***"Duración de lámparas incandescentes" (en horas)***

TABLA DE DISTRIBUCIÓN DE FRECUENCIAS

Intervalos	f_i	F_i	fr_i	Fr_i
[699,5 ; 799,5)	3	3	0,06	0,06
[799,5 ; 899,5)	4	7	0,08	0,14
[899,5 ; 999,5)	17	24	0,34	0,48
[999,5 ; 1099,5)	9	33	0,18	0,66
[1099,5 ; 1199,5)	12	45	0,24	0,90
[1199,5 ; 1299,5)	5	50	0,10	1,00
	50			

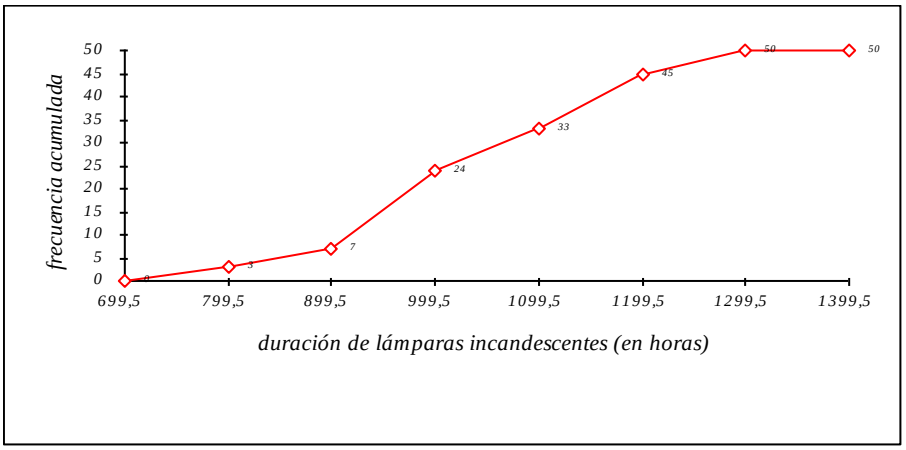
b) Dibuje un histograma y el polígono de frecuencias correspondiente

HISTOGRAMA Y POLÍGONO DE FRECUENCIAS



c) Grafique la función de frecuencia acumulativa (curva ojiva).

CURVA OJIVA



d) Calcule la media, la mediana, el modo y la desviación estándar de los datos brutos.

- **Media aritmética:**

$$\bar{x} = \frac{1}{n}\sum_i x_i f_i = \frac{1}{50}.\ 51054 = \textbf{1021,08 horas}.$$

\+ En promedio, las lámparas duran 1021,08 horas

- **Mediana:**

El orden de la mediana es (n+1)/2 = 25,5, lo que indica que la mediana será el promedio entre los números que ocupen las posiciones n/2 y (n+2)/2, o sea, el promedio entre 1009 horas y 1009 horas, es decir:

Me = 1009 horas.

+ Esto indica que el 50% de las lámparas duran 1009 horas o menos y el otro 50% de las lámparas duran 1009 horas o más.

- **Moda:**

Como hay tres valores que tienen máxima frecuencia, decimos que la distribución es trimodal, siendo:

Mo_1 = 1009 horas **Mo_2 = 1157 horas** **Mo_3 = 1195horas**

+ Los valores con mayor frecuencia son 1009 horas, 1157 horas y 1195 horas.

- **Desviación estándar:**

s = $\sqrt{s^2}$ = **137,484818 horas.**

+ En promedio la duración de las lámparas se apartan de la media en aproximadamente 137,48 horas.

e) Calcule la media, la mediana, el modo y la desviación estándar de los datos agrupados.

- **Media aritmética:**

$\bar{x} = \frac{1}{n} \sum_i m_i f_i = \frac{1}{50} . 51275$ = **1025,5 horas**.

+ En promedio, las lámparas duran 1025,5 horas

- **Mediana:**

El orden de la mediana es (n+1)/2 = 25,5, luego, observando la frecuencia acumulada, elegimos la primera que sea mayor o igual a 25,5, entonces, Me $\in$ [999,5 ; 1099,5) que es la **clase mediana**.

El valor de la mediana dentro del intervalo está dado por:

$$\mathbf{Me} = Li_{Me} + \frac{\mathbf{c} . \left(\frac{\mathbf{n}}{\mathbf{2}} - \mathbf{F_{ant\,Me}} \right)}{\mathbf{f_{Me}}} = 999,5 + \frac{100 . (25 - 24)}{9} = \mathbf{1010,61\ horas}.$$

+ Esto indica que el 50% de las lámparas duran 1010,61 horas o menos y el otro 50% de las lámparas duran 1010,61 horas o más.

*** Moda:** La mayor frecuencia se da en el intervalo [899,5 ; 999,5) que es la **clase modal**. Se observa que con los datos agrupados tenemos una distribución unimodal.

El valor del modo dentro del intervalo está dado por:

Mo = Li_{Mo} + l . $\frac{\Delta_1}{\Delta_1 + \Delta_2}$ = 899,5 + $100 . \frac{13}{13+8}$ = **961,40 horas**.

(siendo $\Delta_1 = f_{Mo} - f_{ant.\ Mo}$ y $\Delta_2 = f_{Mo} - f_{post.\ Mo}$)

+ El valor que representa la mayor frecuencia es 961,40 horas.

*** Desviación estándar:**

s = $\sqrt{s^2}$ = **134,8619854 horas.**

+ En promedio la duración de las lámparas se apartan de la media en aproximadamente 134,86 horas.

f) Compare los resultados obtenidos en los incisos d y e.

	Datos no agrupados	Datos agrupados
Media	1021,08 horas	1025,5 horas
Mediana	1009 horas	1010,61 horas
Modo	1009 horas - 1157 horas -1195 horas	961,40 horas
Desviación estándar	137,48 horas	134,86 horas

Comentarios:

La media se mantiene parecida porque la muestra es bastante homogénea.

La mediana es la que menos varía.

El modo es muy distinto porque al ser agrupados las frecuencias cambian.

El desvío se mantiene bastante parecido.

g) Calcule R, $Q_3 - Q_1$ y C.V..

*** Rango o amplitud muestral:**

R = $x_{máx}$ - $x_{mín}$ = 1240 horas - 702 horas = **538 horas.**

La amplitud de la muestra es de 538 horas.

- **Recorrido intercuartílico:**

Recordemos que llamamos recorrido intercuartílico a la amplitud entre el primer cuartil y el tercer cuartil. Se utiliza cuando nos interesa trabajar con el 50 % central de la distribución alrededor de la mediana.

$Q_3 - Q_1$ = 1137 horas - 931,85 horas = **205,15 horas**.

El 50% de los valores alrededor de la mediana se hallan entre 931,85 horas y 1137 horas, o sea, hay una diferencia de 205,15 horas entre los cuartiles 1 y 3.

Como lo obtuvimos:

El orden del primer cuartil es (n+1) .1 / 4 = 12,75

Luego, observando la frecuencia acumulada, elegimos la primera que sea mayor o igual a 12,75, entonces:

$Q_1 \in [\ 899,5\ ;\ 999,5\ )$ que es la clase del cuartil 1.

El valor del cuartil 1 en el intervalo está dado por

$$Q_1 = Li_{Q1} + \frac{\mathbf{c}.\left(\frac{\mathbf{n}}{\mathbf{4}}\mathbf{.1} - \mathbf{F_{ant\,Q_1}}\right)}{\mathbf{f_{Q_1}}}$$

$$Q_1 = 899,5 + \frac{100.\left(12,5 - 7\right)}{17} = 931,85 \text{ horas.}$$

El 25% de las lámparas duran 931,5 horas o menos y el 75 % restantes duran 931,5 horas o más.

El orden del tercer cuartil es (n+1) .3 / 4 = 38,25

Luego, observando la frecuencia acumulada, elegimos la primera que sea mayor o igual a 38,25; entonces:

$Q_3 \in [\ 1099,5;\ 1199,5\ )$ que es la clase del cuartil 3.

El valor del cuartil 3 en el intervalo está dado por

$$Q_3 = Li_{Q3} + \frac{\mathbf{c}.\left(\frac{\mathbf{n}}{\mathbf{4}}\mathbf{.3} - \mathbf{F_{ant\,Q_3}}\right)}{\mathbf{f_{Q_3}}}$$

$$Q_3 = 1099,5 + \frac{100\ .\ \left(37,5\ -\ 33\right)}{12} = 1137 \text{ horas.}$$

\+ El 75% de las lámparas duran 1137 horas o menos y el 25 % restantes duran 1137 horas o más.

* **Coeficiente de variación:**

$$\textbf{C.V.} = \frac{s}{\bar{x}} = \frac{134,86}{1025,5} = \mathbf{0,1315}.$$

Tenemos un coeficiente de variación de 0,1315 o del 13,15%.
Esto significa que el desvío estándar s es un 13,15 % de la media.

h) ¿Cuál es y qué representa el percentil 45?.

El percentil de orden 45 es (n+1) . 45 / 100 = 22,95

Luego, observando la frecuencia acumulada, elegimos la primera que sea mayor o igual a 22,95; entonces:

$P_{45} \in$ [899,5 ; 999,5) que es la **clase del percentil 45**.

El valor del percentil 45 en el intervalo está dado por

$$P_{45} = Li_{P45} + \frac{c \cdot \left(\frac{n}{100} .45 - F_{ant\,P45} \right)}{f_{P45}}$$

$$\mathbf{P_{45}} = 899,5 + \frac{100 \cdot (22,5 - 7)}{17} = \mathbf{990,68\ horas}.$$

El 45% de las lámparas duran 990,68 horas o menos y el 55% restante duran 990,68 horas o más.

i) ¿Por debajo de qué valor se halla el 25% de las horas de vida de estas lámparas?.

El valor que deja por debajo de él el 25% es el cuartil 1, Q_1 = 931,85 horas .

El 25% de las lámparas duran 931,5 horas o menos.

j) ¿Qué porcentaje de lámparas duró hasta 1140 horas?.

El rango percentil de x está dado por

$$\text{RP de x} = \frac{F_{ant\ P\#} + \frac{x - Li_{P\#}}{c} .f_i}{n} .100\%$$

$$\mathbf{RP\ de\ 1140} = \frac{33 + \frac{1140 - 1099,5}{100} .12}{50} . 100\% = \mathbf{75,72\%}\ .$$

El 75,72% de las lámparas duran 1140 horas o menos.

k) ¿Qué porcentaje de lámparas duró menos de 1000 horas?.

$$\mathbf{RP\ de\ \textit{1000}} = \frac{24 + \frac{1000 - 999,5}{100} .33}{50} . 100\% = \mathbf{\textit{48,33\%}}.$$

El 48,33% de las lámparas duran 1000 horas o menos.

l) ¿Qué porcentaje de lámparas duró más de 900 horas?.

$$\textbf{RP de 900} = \frac{7 + \frac{900 - 899,5}{100}.17}{50} . 100\% = \textbf{14,17\%}.$$

El 14,17% de las lámparas duran 900 horas o menos, por lo que 100-14,17=85,83 %, es decir el 85,83% de las lámparas duran más de 900 horas.

CARACTERÍSTICAS DE FORMA DE UNA POBLACIÓN

Cuando conocemos las medidas de tendencia central, las medidas de dispersión es conveniente conocer también la forma de la distribución, es decir analizar su simetría y su curtosis.

MEDIDAS DE ASIMETRÍA

Decimos que una ***distribución es simétrica*** cuando las medidas de tendencia central, media aritmética, mediana y moda coinciden:

$$\mu = \mathrm{X}_{\mathrm{Mo}} = X_{0,5}$$

El sesgo es el grado de asimetría, de una distribución. Si la curva de frecuencia de una distribución tienen una "cola" más larga a la derecha del máximo central que a la izquierda se dice que la distribución es sesgada a la derecha o que tiene sesgo positivo. Si es al contrario, se dice que es sesgada a la izquierda o que tiene sesgo negativo.

La medida Pearsoniana de asimetría se basa en las relaciones entre la media, mediana y la moda. Para una distribución unimodal simétrica, estas tres medidas son de idéntico valor, pero para una distribución asimétrica, la media se aleja de la moda con la mediana entre ellas.

La distancia entre la media y la moda podría usarse para medir la asimetría.

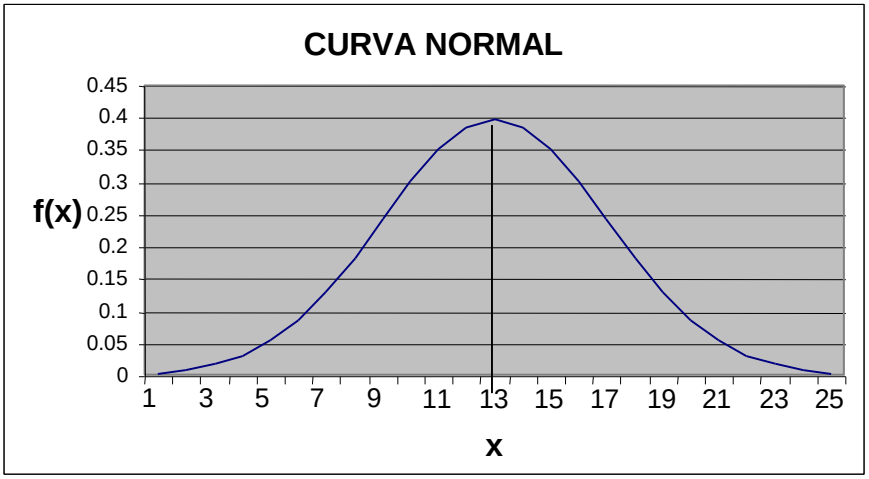
$$\text{ASIMETRÍA} = \mu - \mathrm{X}_{\mathrm{Mo}}$$

Cuanto mayor es esa distancia, tanto más asimétrica, es la distribución.

Esta medida puede adimensionarse dividiéndola por una medida de dispersión, tal como la desviación típica, esta medida se conoce como ***coeficiente de asimetría.***

Un ejemplo típico de una distribución simétrica es la distribución normal.

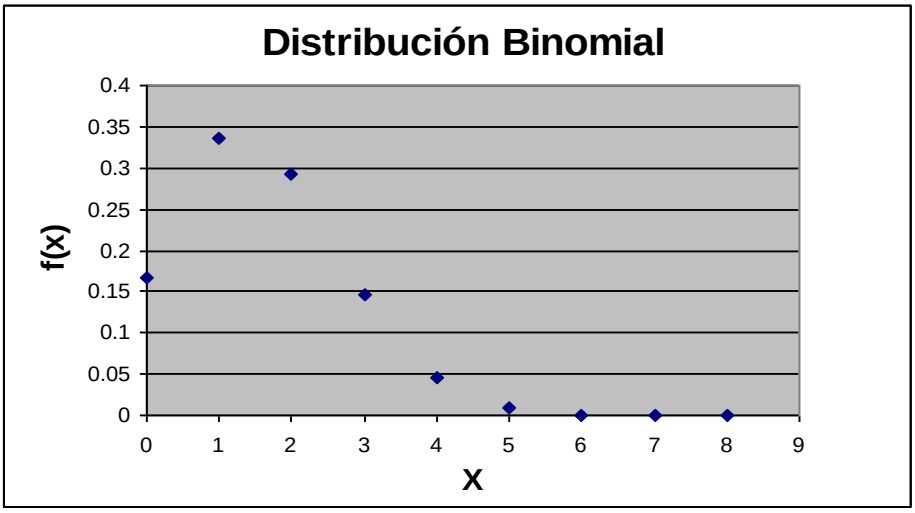
DISTRIBUCIÓN NORMAL: **función densidad**



Se observa que en esta distribución la media aritmética, la mediana y la moda coinciden.

Curva simétrica: $\mu = \mathrm{X}_{\mathrm{Mo}} = X_{0,5}$

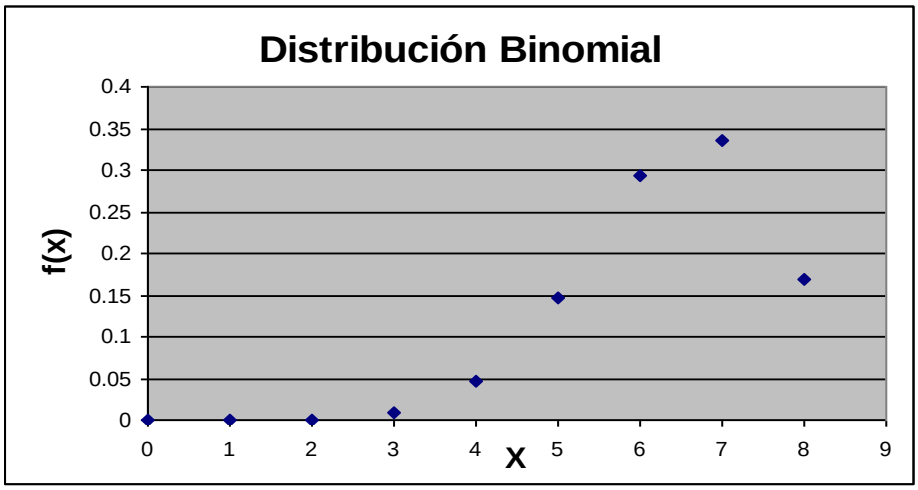
Un ejemplo de una **distribución asimétrica a derecha o con sesgo positivo, es la Distribución Binomial con parámetros: n=8, p=0,20**



Se observa que la media aritmética es mayor que la mediana y la moda, se observa que la mediana se encuentra entre la media aritmética y la moda.

Curva asimétrica positiva: $X_{Mo} < X_{0,5} < \mu$

Un ejemplo de una **distribución asimétrica a izquierda o con sesgo negativo: Binomial** con parámetros n = 8, p= 0,8



Se observa que la Moda es mayor que la mediana y la media aritmética. La mediana se encuentra entre la moda y la media aritmética.

Curva asimétrica negativa: $\mu < X_{0,5} < X_{Mo}$

Conclusión

Curva simétrica: $\mu = X_{Mo} = X_{0,5}$

Curva asimétrica positiva: $X_{Mo} < X_{0,5} < \overline{\mu}$

Curva asimétrica negativa: $\mu < X_{Me} < X_{Mo}$

COEFICIENTE DE ASIMETRÍA

Para saber si una distribución con una sola moda, es asimétrica a la derecha o a la izquierda, sin necesidad de representarla gráficamente, podemos utilizar el coeficiente de asimetría de Pearson que calcularemos:

$$A_P = \frac{\mu - x_{MO}}{d}$$

- En una *distribución simétrica* unimodal, la media, la mediana y la moda coinciden. En este tipo de distribuciones los datos se encuentran repartidos a lo largo del recorrido de forma que todas las medidas de tendencia central están justo en el centro del conjunto de datos. Si la distribución es simétrica, Ap=0, ya que la media aritmética es igual al modo.

- Si la distribución unimodal *es asimétrica a la derecha*, entonces, $A_p > 0$, ya que la media aritmética es mayor que el modo.

- Si la distribución unimodal *es asimétrica a la izquierda,* entonces, $A_p < 0$, ya que la media aritmética es menor que el modo.

CURTOSIS O APUNTAMIENTO

La curtosis mide como se concentran los datos alrededor de su media. Es una medida de qué tan puntiaguda es la distribución de probabilidad.

COEFICIENTE DE CURTOSIS

Cuando una distribución es simétrica, a veces es interesante saber si es más o menos apuntada que la curva normal. Existe un coeficiente ideado por Fisher, que mide el apuntamiento y se calcula:

$$C_F = \frac{\frac{1}{N}\sum_i (x_i - \mu)^4 . f_i}{\left(\frac{1}{N}\sum_i (x_i - \mu)^2 . f_i\right)^2} - 3$$

Observemos que las tres curvas que se ven a continuación son simétricas, pero se diferencian en cuanto que unas son más planas que otras, esta característica es conocida como curtosis.

La curva A, muestra la curva ideal, o distribución normal, llamada ***mesocúrtica***, Si calculamos su curtosis su valor será tres.

La curva B, es más picuda, nos referimos a esta distribución como ***leptocúrtica***, si calculamos su valor nos dará un valor positivo, mayor que tres. Los datos se encuentran muy concentran alrededor de su media.

En contraposición la curva C, es más plana que la curva ideal o normal que se toma como referencia, se la llama ***platocúrtica***. Si calculamos su curtosis nos dará un valor menor que 3. Los datos estan alejados de su media.