UNIVERSIDAD TECNOLÓGICA NACIONAL
FACULTAD REGIONAL CÓRDOBA

PROFESOR TITULAR

MBA ROBERTO P. FUGIGLANDO

Año Lectivo 2021

(PRIMERA PARTE)

Dado que el objetivo perseguido en la elaboración de las presentes **NOTAS de PROBABILIDAD Y ESTADÍSTICA,** para alumnos de las distintas Carreras que se dictan en esta FRC – UTN, ***no es la formación de estadísticos,*** se omiten en los distintos temas, desarrollos matemáticos de la Estadística Matemática, que el alumno olvida fácilmente, y que en ***nada contribuyen,*** para lograr un ***aprendizaje significativo*** de la materia.

Lo importante a mi criterio, es que el estudiante de esta Facultad, alcance a comprender el "Qué", el "Porqué", el "Dónde", y el "Cuándo", aplicar los **métodos, procedimientos** y **herramientas estadísticas**, más que el "Cómo" de las mismas, cuando se enfrente en su vida como profesional, en situaciones de toma de decisiones ante situaciones de incertidumbre, o bien como usuario de información estadística.

Por lo expuesto, en el presente curso se impartirán procedimientos y técnicas estadísticas, con el fin de ***motivar*** a los estudiantes a:

1) Conocer la aplicación de las mismas,
2) Interpretar las distintas herramientas de la Estadística Descriptiva.
3) Adquirir destreza en la interpretación del Cálculo de Probabilidad.
4) Comprender la importancia de la aplicación del análisis estadístico, en situaciones problemáticas de la sociedad, en la que se va a desempeñar como profesional.

MBA Roberto P. Fugiglando
Profesor Titular

UNIDAD Nº I

"METODOLOGÍA ESTADÍSTICA"

SIGNIFICADO DE ESTADÍSTICA

Podemos decir que la **ESTADÍSTICA,** está dada por *"El conjunto de métodos* y *procedimientos, utilizados en la recopilación, organización, presentación, análisis e interpretación de datos, para extraer conclusiones y tomar decisiones razonables en base a las mismas, en situaciones de incertidumbre".* También se dice, que es: *"La tecnología del método científico que proporciona instrumentos útiles para la toma de decisiones, en situaciones de incertidumbre".*

Se la *clasifica* en:

1) ***Descriptiva:*** se corresponde con la metodología utilizada, para describir un conjunto de datos, a través de métodos numéricos o gráficos. Por ejemplo, la evolución de las ventas en una empresa, en un cierto período de tiempo.

2) ***Inferencial:*** se corresponde con la metodología aplicada, cuando se quiere estudiar a una cierta población estadística, tomando en consideración sólo a una parte **representativa** de los elementos que la conforman. Por ejemplo, cuando se quiere estudiar si un proceso de producción está operando "bajo control".

POBLACIÓN ESTADÍSTICA

Está formada por la **totalidad** de los elementos, acerca de los cuales se realiza una cierta investigación estadística. Por ejemplo: el total de empresas de una cierta rama de actividad, o el agua contenida en un cierto dique.

Se la ***clasifica*** en:

1) ***Finita:*** cuando está conformada, por un número **finito numerable** de componentes. Por ejemplo: la totalidad de los empleados de una empresa.

2) ***Infinita:*** cuando está conformada por un número **finito no numerable,** o **infinitos** componentes. Por ejemplo: cuando se desea estudiar el grado de contaminación del agua de un lago.

DATO ESTADÍSTICO

Es toda información susceptible de ser sometida al análisis estadístico. Es decir, que debe poder ser comparada, analizada e interpretada, razón por la cual se dice que los datos estadísticos, presentan relaciones significativas. De lo expuesto, se deduce que un valor aislado, no constituye dato estadístico.

RECOPILACIÓN DE DATOS

Los datos estadísticos, pueden ser generados por el investigador mismo, o bien haber sido recopilados por un tercero. En el primer caso, se dice que la ***FUENTE DE DATOS es PRIMARIA***, mientras que en el segundo caso, se dice que ***es SECUNDARIA***.

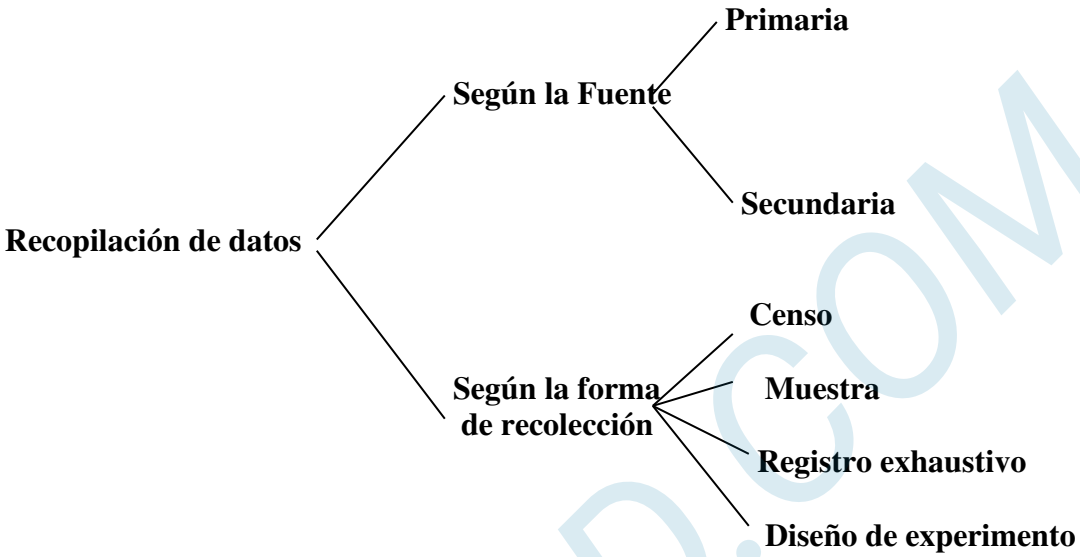
El proceso de generación de los datos estadísticos, puede efectuarse de distintas formas dadas por:

1) ***CENSO:*** es el proceso consistente en estudiar características de ***cada uno*** de ***todos*** los elementos de una cierta población estadística, debiendo ser ésta ***finita,*** siendo un relevamiento ***estático*****.** Por ejemplo: un censo de población, o un inventario de todos los productos elaborados existentes en un depósito. Cabe destacar que, en muchas situaciones prácticas, no es posible acceder *a cada uno de todos* los componentes de una población estadística, por imposibilidad, problemas de costo, practicidad, o bien cuando el proceso de obtención del dato es destructivo, como ocurre cuando se quiere estudiar la resistencia de un cierto material.

2) ***MUESTRA:*** consiste en estudiar características de una cierta población estadística, tomando en consideración sólo una *parte representativa* de la misma, a través de las denominadas *técnicas de muestreo,* de forma tal que, las conclusiones obtenidas de esa parte, puedan ser generalizadas al resto de los elementos de la población, a través de la metodología de la ***inferencia estadística***. Resulta aplicable tanto a poblaciones finitas como infinitas, siendo el único método de investigación en este último caso, así como cuando el proceso de investigación es destructivo. Por ejemplo, estudiar la contaminación del agua de un río, o bien la vida útil de lámparas de luz de una determinada marca.

3) ***REGISTRO EXHAUSTIVO:*** la información estadística se obtiene por un acto administrativo, como ocurre por ejemplo, cuando se quiere estudiar la procedencia de los alumnos de una Facultad, la que puede obtenerse cuando el alumno cumplimentó su inscripción en esa Facultad, así como la oficina de Registro Civil de un lugar, cuando se desea saber el número de nacimientos inscriptos en el mismo, en un cierto período de tiempo.

4) ***DISEÑO DE EXPERIMENTO:*** es el proceso que se presenta cuando el investigador genera sus propios datos, diseñando un experimento, a la medida de sus necesidades. Por ejemplo, probar la eficiencia de determinadas drogas para tratar una determinada enfermedad.



UNIDAD ESTADÍSTICA Y UNIDAD DE RELEVAMIENTO

Se denomina ***unidad estadística o unidad elemental,*** a cada componente de una población estadística. Por ejemplo, si se está estudiando el modelo de auto de una cierta marca más vendido, cada auto de esa marca, sería la ***unidad estadística***.

Se denomina ***unidad de relevamiento*** donde la información estadística va a ser recopilada, o por quien proporciona la misma. Por ejemplo, si queremos saber de la producción diaria de una empresa, cuántos artículos fallados fueron producidos, y existe en la misma un área de Control de Calidad, este sector sería la *unidad de relevamiento*. En un censo de población, la *unidad de relevamiento* es el *hogar*.

Cabe destacar que, existen situaciones en las que ambas unidades coinciden. Así por ejemplo, si queremos saber las edades de los alumnos de este curso y le consultamos a cada uno de los alumnos su edad, son coincidentes, mientras que si se obtiene el dato de los legajos de los alumnos, no son coincidentes.

ESTUDIOS Y MÉTODOS ESTADÍSTICOS

Acorde a la ***intervención del investigador*** los estudios estadísticos pueden ser:

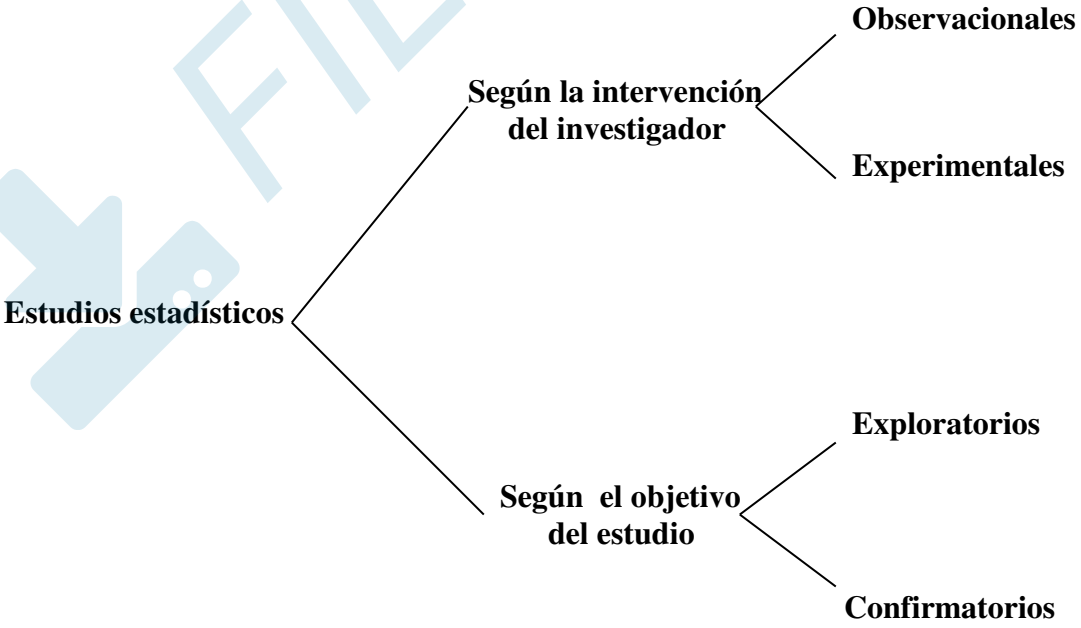
- ***Observacionales.***
- ***Experimentales.***

Se dice que se realizan ***estudios observacionales*** cuando los datos se generan a partir de un censo, muestra o registro exhaustivo, y consisten en medir u observar las unidades elementales de una cierta población midiendo, observando y registrando las características bajo estudio.

Los ***estudios observacionales,*** pueden ser ***retrospectivos*** o ***prospectivos.*** Los ***retrospectivos*** evalúan situaciones pasadas, en busca de una explicación de acontecimientos actuales. Así por ejemplo estudiar el grado de formación profesional, de los gerentes exitosos de distintas empresas. Los ***prospectivos,*** realizan un seguimiento en el futuro de los efectos de algún acontecimiento observado, en el presente o en el pasado. Por ejemplo registrar sistemáticamente la penetración en el mercado en los próximos años de un cierto producto, resultante de la implementación de una nueva estrategia de ventas.

Los ***estudios experimentales*** son aquellos en los que el investigador, diseña un experimento, consistente en aplicar uno o más tratamientos a las *unidades elementales* de una cierta *población*, analizando los resultados obtenidos.

Considerando el ***objetivo*** perseguido por el investigador, los estudios pueden ser ***exploratorios*** o ***confirmatorios.*** Los primeros, son aquellos aplicados cuando se quiere determinar un modelo, que describa o explique el comportamiento de distintas variables de los componentes de una población. Por ejemplo, tratar de encontrar un modelo econométrico, asociado al comportamiento de las variables relacionadas con la demanda de productos de las empresas, de una cierta rama de actividad. Los ***confirmatorios*** en cambio tratan de determinar si el comportamiento de distintas variables, de los componentes de una población, responden a un modelo supuesto. Por ejemplo, evaluar si realmente un modelo econométrico propuesto, se corresponde con el comportamiento, de las variables asociadas a la demanda de los productos de las empresas de una cierta rama de actividad.



VARIABLES

De las unidades estadísticas de una determinada población podemos estudiar una o más características dadas por la o las ***variables.***

Se **clasifican** a las ***variables*** de dos formas:

1) **Cualitativas o Categóricas.**
2) **Cuantitativas o Numéricas.**

1) Son **Cualitativas,** aquellas en la característica bajo estudio es un atributo o categoría. Por ejemplo: la profesión de los empleados de una empresa.

De acuerdo a la **forma de medición** las variables **Cualitativas** pueden ser clasificadas en **escalas** dadas por:

a) **Escala Nominal:** es aquella, en la que no es posible establecer una relación de magnitud, entre los distintos atributos, en que puede ser expresada la variable. Por ejemplo: nacionalidad de las personas.

b) **Escala Ordinal:** es el caso en que es posible, establecer un orden de importancia entre los distintos atributos en que puede ser expresada la variable. En esta escala de medición, es posible establecer relaciones del tipo A > B, A = B, o A < B. Por ejemplo, la opinión de las personas, acerca del servicio brindado por una empresa, siendo las posibles respuestas: Excelente, Muy Bueno, Bueno, Regular, Malo. Se observa, que no es posible decir que los que responden Bueno están doblemente satisfechos, en relación a los que respondieron Malo.

2) Son **Cuantitativas,** aquellas en la característica bajo estudio es **numérica**. Según sea la forma de obtener el valor numérico, se las clasifica en:

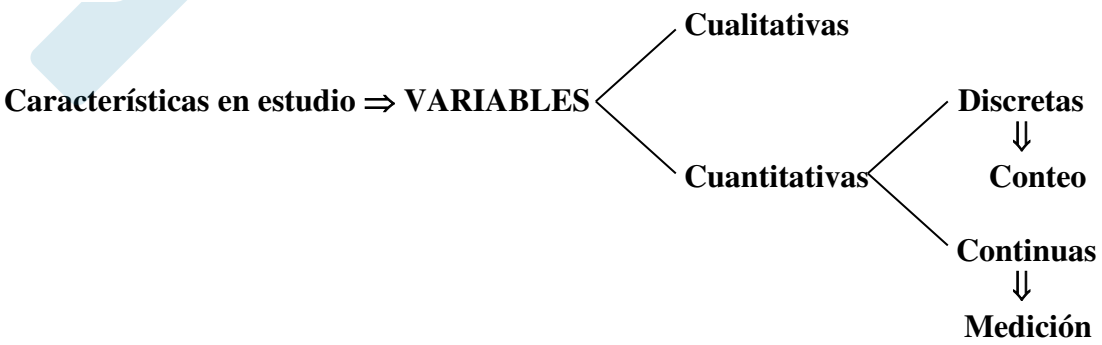
- **Discretas:** cuando el valor surge de un ***proceso de conteo***, asumiendo por lo tanto valores enteros. Por ejemplo: el número de empleados de una empresa.

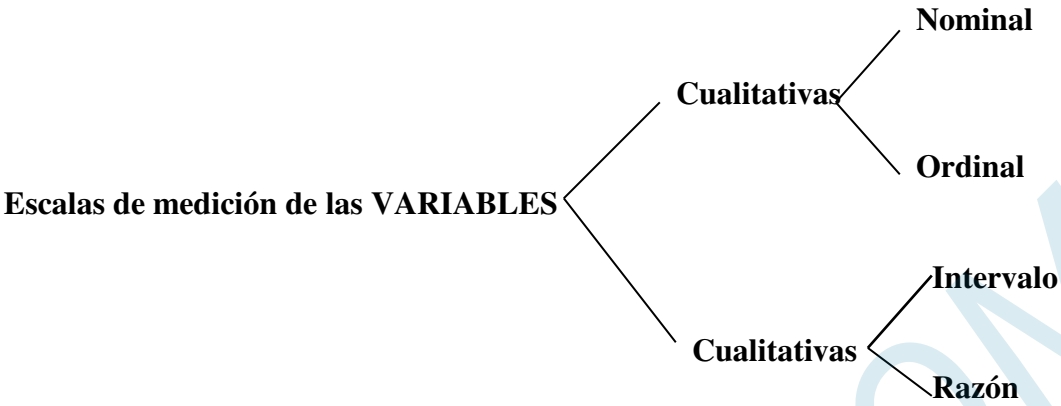
- **Continuas:** cuando el valor, surge de un ***proceso de medición***, siendo la diferencia entre un valor y otro un infinitésimo, asumiendo por lo tanto cualquier valor real. Por ejemplo: la altura de los empleados de una empresa.

De acuerdo a la **forma de medición** las variables **Cuantitativas** pueden ser clasificadas en **escalas** dadas por:

a) Escala de Intervalos: representa un nivel superior de medición respecto a la **escala ordinal**, permitiendo establecer relaciones del tipo A > B, A = B, o A < B. De la misma forma, es posible comparar intervalos de valores como consecuencia de que, a lo largo de toda la escala, dos valores adyacentes representan siempre la misma diferencia de magnitud, siendo posible realizar comparaciones tal como, A > B. Una característica de esta escala, es que el valor cero es arbitrario. Un ejemplo de una variable en esta escala de medición, puede estar dado por el horario de ingreso de los empleados a su trabajo en una fábrica. Entre las 8 y las 9 de la mañana existe la misma diferencia horaria que entre las 4 y 5 de la tarde, pero no es posible afirmar que quien ingresó a las 8 de la noche, ha ingresado el doble de tarde, de aquel que ingresó a las 8 de la mañana. Cabe destacar que, la hora cero no indica particularmente nada.

b) Escala de Razón: representa el nivel más alto de medición. Además de tener las propiedades de las otras escalas, tiene un cero absoluto, pudiéndose calcular proporciones entre los valores de la escala. Así por ejemplo podemos afirmar que, una persona de 38 años tiene el doble de edad que una de 19, o que la diferencia entre una persona de 38 años y otra de 39 es la misma que la existente entre una persona de 19 años y otra de 20, o que una persona de 30 años es mayor que otra de 22. Sobre esta escala, es posible realizar todas las operaciones matemáticas asociadas a los números (suma, resta, multiplicación, división).





Cabe destacar que de **acuerdo al número de variables en estudios,** los métodos de análisis estadísticos, pueden ser ***Univariados*** o ***Multivariados***. Los primeros consideran una *única variable*, mientras que los segundos *dos o más*.

La correcta clasificación de las variables en estudio, el número de ellas, así como la definición de la escala de medición, orientará al investigador, sobre las técnicas estadísticas más adecuadas para analizarlas.

PARÁMETROS Y ESTADÍGRAFOS

Se denomina ***Parámetro*****,** a toda medida estadística calculada considerando a **todos los elementos,** de una población bajo estudio. Por ejemplo, calcular las ventas promedio de las empresas, de una determinada rama de actividad, considerando las ventas de ***cada una*** de **todas** esas empresas.

Se denomina ***Estadígrafo, Estadístico o Estimador,*** a toda medida estadística, calculada considerando una **muestra** de los elementos de una determinada población. Esta medida, servirá de ***estimador,*** del verdadero valor del ***parámetro*** de esa población (**Inferencia Estadística**).

UNIDAD Nº II

"ORGANIZACIÓN Y PRESENTACIÓN DE DATOS ESTADÍSTICOS"

DISTRIBUCIONES UNIDIMENSIONALES

Cuando de las *unidades estadísticas* de una población bajo estudio, se considera una sola variable, se dice que se está en presencia de una investigación **unidimensional**. Si son dos las variables, **bidimensional**, y así sucesivamente.

SERIE SIMPLE

Se denomina de esta forma, al conjunto de datos obtenidos en bruto, luego de realizado el relevamiento pertinente. En forma genérica, se denota con $\mathbf{x_i}$ a cada observación, variando el subíndice **i** desde **1** hasta **n,** siendo **n** el ***total de observaciones***. En consecuencia, simbólicamente una Serie Simple de **n** elementos, en una investigación unidimensional, adoptará la forma:

$$x_1 , x_2 , x_3 , \ldots\ldots , x_n \Rightarrow x_i \ / \ i = \overline{1 , n}$$

Por ejemplo, si de los 14 empleados de una empresa se quiere saber el número de hijos que cada uno de ellos tiene, una **serie simple,** podrá adoptar la forma:

$x_1=3$; $x_2=2$; $x_3=3$; $x_4=2$; $x_5=1$; $x_6=0$; $x_7=5$; $x_8=2$; $x_9=3$; $x_{10}=3$; $x_{11}=1$; $x_{12}=0$; $x_{13}=1$; $x_{14}=3$

En el ejemplo considerado, con $\mathbf{x_i}$ se quiere representar el *número de hijos del empleado entrevistado*, que genéricamente, se los individualiza con el subíndice **i**.

DISTRIBUCIONES DE FRECUENCIAS
(Datos agrupados)

Pueden presentarse dos situaciones:

1) Variable Discreta
2) Variable Continua

DISTRIBUCIÓN DE FRECUENCIA DE VARIABLE DISCRETA

Si los datos de una serie simple se corresponden con una *variable discreta*, y sus *valores* ***son homogéneos***, pueden agruparse los que son iguales, definiendo **una nueva variable**, que la denotaremos con $\mathbf{y_i}$ **,** variando el subíndice **i,** desde **1** hasta **k**, siendo **k** el número de valores **distintos de la variable** observados en la serie simple. Con los valores de la *nueva variable*, y *el número de veces* que ese valor aparece repetido, en la serie simple, denominada **frecuencia absoluta**, que la denotaremos con $\mathbf{n_i}$, se construye una tabla, denominada ***Distribución de Frecuencia o Datos Agrupados de Variable Discreta.***

Considerando la serie simple anterior, correspondiente al número de hijos de los 14 empleados de una empresa, la ***distribución de frecuencia de variable discreta*** asociada, responderá a la forma:

Nº de hijos y_i	**Nº de empleados** n_i	h_i	**Frecuencias Acumuladas** N_i	H_i
0	2	0,15	2	0,15
1	3	0,21	5	0,36
2	3	0,21	8	0,57
3	5	0,36	13	0,93
5	1	0,07	14	1,00
$\sum$	14	1,00		

Las ***frecuencias absolutas*** **($\mathbf{n_i}$)** representativas del número de veces, que el valor de la variable aparece repetido en la serie simple, tienen las siguientes propiedades:

a) Son números enteros comprendidos entre **0** y **n**, siendo **n** el total de observaciones.

b) $\sum \mathbf{n_i} = \mathbf{n}$, (Nº de observaciones).

Frecuencias relativas

Se las denota con $\mathbf{h_i}$, y se obtienen del ***cociente*** *entre las* ***frecuencias absolutas*** *asociadas a* ***cada valor de la variable****, y el* ***total de observaciones***. Es decir:

$$\mathbf{h_i = n_i / n}$$

Las ***frecuencias relativas,*** indican *la importancia relativa, de cada valor de la variable en relación al total de observaciones*, pudiéndoselas interpretar en términos porcentuales. Verifican las siguientes propiedades:

a) Son números fraccionarios, comprendidos en el intervalo: $\mathbf{0 \leq h_i \leq 1}$.

b) $\mathbf{\sum h_i = 1}$

Frecuencias Acumuladas

Son representativas *del total acumulado de frecuencias absolutas o relativas*, ***hasta un determinado valor de la variable***. Se las denota con $\mathbf{N_i}$, a las ***absolutas acumuladas***, y con $\mathbf{H_i}$, a las ***relativas acumuladas.***

Las ***últimas frecuencias acumuladas*** son iguales a:

- $\mathbf{N_k = n}$ (Total de observaciones)
- $\mathbf{H_k = 1}$

Expresión genérica de una Distribución de Frecuencias de Variable Discreta

Suponiendo **k** valores **distintos de la variable,** observados en la serie simple, genéricamente, una ***distribución de frecuencias de variable discreta,*** responde a la estructura:

$\mathbf{y_i}$	$\mathbf{n_i}$	$\mathbf{h_i}$	$\mathbf{N_i}$	$\mathbf{H_i}$
y_1	n_1	h_1	N_1	H_1
y_2	n_2	h_2	N_2	H_2
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
y_k	n_k	h_k	N_k	H_k
$\sum$	**n**	**1**		

Representación gráfica de una Distribución de Frecuencias de Variable Discreta

Se presentan dos situaciones:

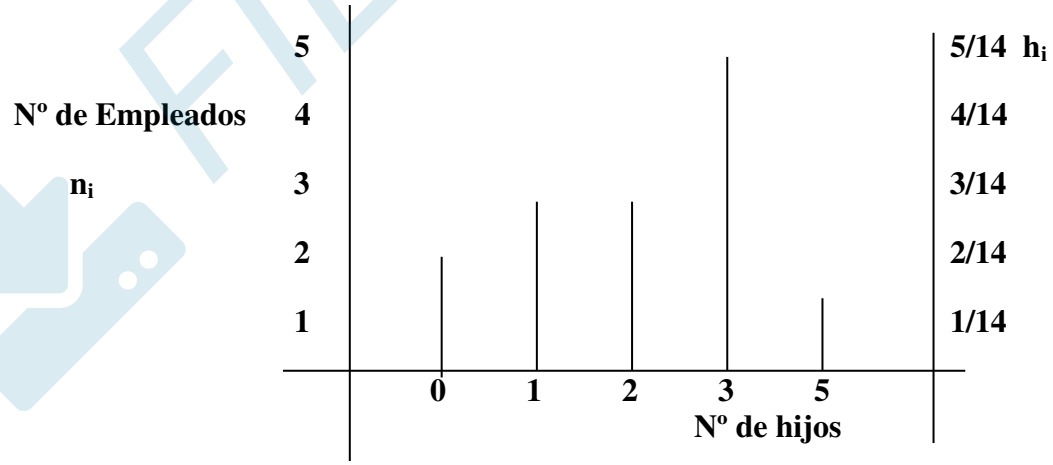
a) Frecuencias Simples ($\mathbf{n_i}$ y $\mathbf{h_i}$)
b) Frecuencias Acumuladas ($\mathbf{N_i}$ y $\mathbf{H_i}$)

a) ***Frecuencias absolutas y relativas simples*** **($\mathbf{n_i}$ y $\mathbf{h_i}$)**

Se corresponde con el denominado ***gráfico de los bastones***, **idénticos,** para ambos tipos de frecuencias, sólo con un cambio de escala en el eje de las ordenadas, en el caso de las *relativas*, puesto que:

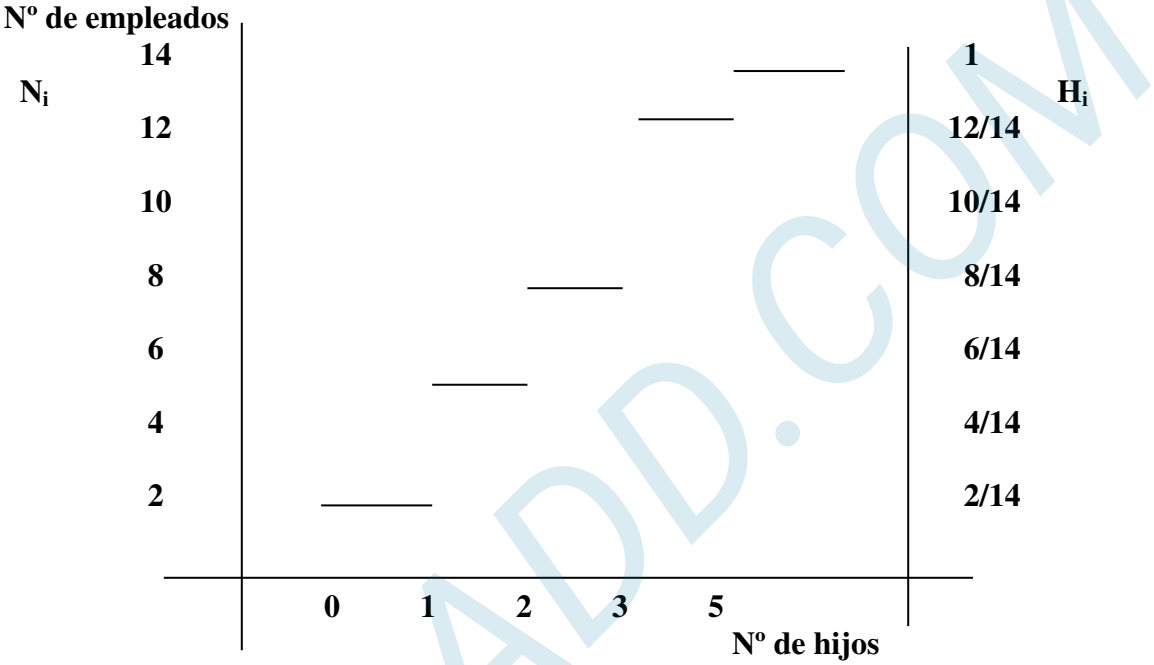
$$\mathbf{h_i = n_i / n \Rightarrow h_i = n_i . 1/n}$$

Considerando los datos del problema anterior, el gráfico responde a la forma:



b) ***Frecuencias Absolutas y Relativas Acumuladas*** **(N_i y H_i)**

Su representación gráfica está dada por el ***diagrama escalonado con trazos discontinuos***, siendo **el mismo,** para ambos tipos de frecuencias, sólo con un cambio de escala en el eje de las ordenadas, en el caso de las relativas, por lo expuesto anteriormente. Considerando los datos del problema anterior, el gráfico responde a la forma:



DISTRIBUCIÓN DE FRECUENCIA DE VARIABLE CONTINUA

Cuando los datos de una serie simple, se corresponden con una *variable discreta con* ***valores heterogéneos***, o bien, ***con una variable continua***, en forma agrupada los valores de la variable, se expresan en grupos denominados genéricamente ***"intervalos de clase"***, conteniendo valores **homogéneos** de la variable, en lo posible, **dentro** de cada intervalo, denotándose a los extremos izquierdo y derecho de los mismos, de la forma: y'_{i-1} y y'_i, respectivamente.

Los *intervalos de clase,* no necesitan ser del mismo tamaño, pero para facilitar la interpretación de la representación gráfica de la distribución de frecuencia, así como para el cálculo de las distintas medidas estadísticas, es *conveniente* que sean del *mismo tamaño*.

Cabe destacar que los intervalos se consideran ***"semiabiertos por la derecha"***, es decir "**desde el extremo izquierdo**……**hasta <u>menos</u> el extremo derecho**", con excepción del último intervalo, que es ***"cerrado"*** en ambos extremos, es decir **"<u>desde</u>…….<u>hasta</u>"**.

Los conceptos de *frecuencias absolutas y relativas simples,* se corresponden con el número de observaciones asociadas a los valores de la variable de cada intervalo, y se las denota de la misma forma que en una distribución de frecuencia de variable discreta, ($\mathbf{n_i}$ y $\mathbf{h_i}$ respectivamente). En el caso de las *Acumuladas,* son representativas del *total acumulado de frecuencias* (absolutas o relativas), hasta el *valor de la variable* correspondiente al **extremo derecho** del intervalo donde se encuentra, ***<u>sin considerar el valor</u> de dicho extremo***, ya que son *semiabiertos por la derecha*, excepto el *último intervalo*. Se las denota con $\mathbf{N_i}$ a las *Absolutas Acumuladas,* y con $\mathbf{H_i}$ a las *Relativas Acumuladas*.

Para la construcción de una Distribución de Frecuencias de Variable Continua, se procede de la siguiente forma:

1) Se calcula el ***Recorrido*** de la variable (**R**), dado por la diferencia entre el *valor mayor* y el *valor menor,* de los observados de la variable en la Serie Simple. Así por ejemplo, si el mayor valor es 93 y el menor 43, el Recorrido estará dado por:

$$\mathbf{R = 93 - 43 = 50}$$

2) **a)** Si se tiene como **<u>dato</u>** *el número de intervalos*, es necesario determinar ***<u>la amplitud</u> de los mismos (*** $\mathbf{c_i}$), de la forma:

Amplitud ($\mathbf{c_i}$) $\Rightarrow$ Recorrido / Nº de intervalos

Cabe destacar que la ***amplitud de un intervalo,*** está dada por la diferencia entre el extremo derecho e izquierdo de cada intervalo de clase, es decir:

$$\mathbf{c_i = y'_i - y'_{i-1}}$$

Así por ejemplo, si se desea una distribución de frecuencias con 5 intervalos de clase, considerando el Recorrido anterior, se tendrá:

Amplitud = 50/5 = 10

<u>Observación</u>: Es conveniente que la amplitud sea un número divisible por 2. En caso de no serlo, debe ampliarse el Recorrido.

b) Si se da como **dato** *la amplitud de los intervalos*, debe determinarse ***el número de los mismos***, de la forma:

Nº de intervalos ⇒ Recorrido / Amplitud

Así por ejemplo, considerando el Recorrido anterior y una amplitud de los intervalos de 10 se tendrá:

Nº de intervalos = 50/10 = 5

OBSERVACIÓN: En el caso de que al efectuar el cociente, se obtenga un número decimal, debe ampliarse el Recorrido, para de esta manera, lograr un número natural.

3) Una vez determinados los intervalos de clase, acorde a lo expuesto en el punto anterior, debe efectuarse el **conteo**, consistente en asignar a cada intervalo, los valores de la variable que figura en la Serie Simple, es decir las *frecuencias absolutas*, considerando el principio de que los intervalos son *semiabiertos por la derecha*, excepto el *último que es cerrado* en ambos extremos, quedando determinada una tabla que es la ***"Distribución de Frecuencias de Variable Continua".***

Ejemplo:

Se clasificó al personal de una empresa de acuerdo a sus edades, obteniéndose la siguiente ***Distribución de Frecuencia de Variable Continua***:

Edades		**Nº de empleados**	**FR**	**FAA**	**FRA**	**Marca de clase**
y'_{i-1}	y'_i	n_i	h_i	N_i	H_i	y_i
18	26	20	0,10	20	0,10	22
26	34	30	0,15	50	0,25	30
34	42	40	0,20	90	0,45	38
42	50	50	0,25	140	0,70	46
50	58	40	0,20	180	0,90	54
58	66	20	0,10	200	1,00	62
	Σ	200	1,00			

La ***"marca de clase"*** ($\mathbf{y_i}$) está dada por el punto medio de cada intervalo de clase y se calcula a los efectos de determinar un valor representativo de la variable de cada intervalo de clase, para obtener las distintas medidas estadísticas. Su valor surge del cociente:

$$y_i = \frac{y'_{i-1} + y'_i}{2}$$

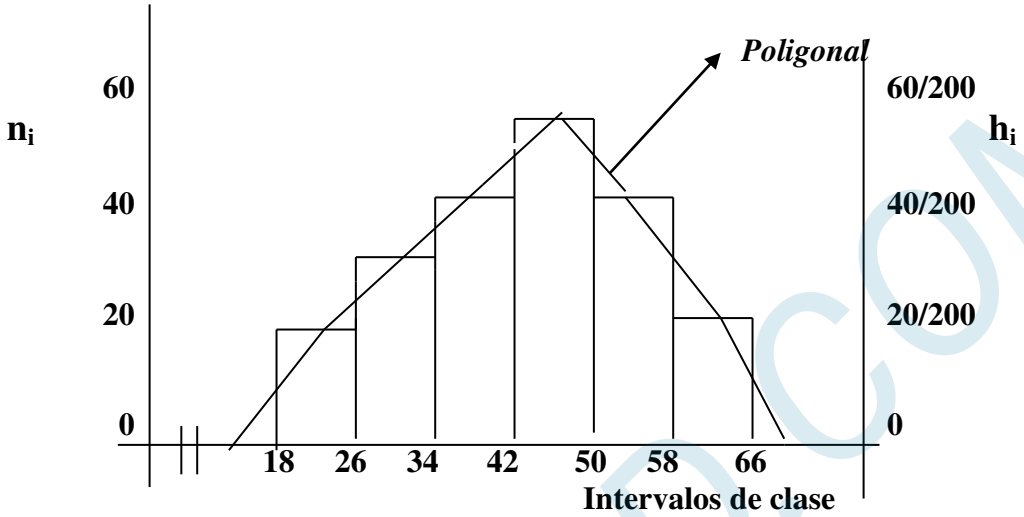
Expresión genérica de una Distribución de Frecuencias de Variable Continua

Suponiendo una distribución de frecuencia de *variable continua,* con **k** intervalos de clase, *simbólicamente,* la podemos expresar de la forma:

Inter. de clase		**Marca de clase**	**Frec. Abs.**	**Frec. Rel.**	**Frec.Abs.Ac.**	**Frec.Rel.Ac.**
$\mathbf{y'_{i-1}}$	$\mathbf{y'_i}$	$\mathbf{y_i}$	$\mathbf{n_i}$	$\mathbf{h_i}$	$\mathbf{N_i}$	$\mathbf{H_i}$
y'_0	y'_1	y_1	n_1	h_1	N_1	H_1
y'_1	y'_2	y_2	n_2	h_2	N_2	H_2
.....		.....	.....	.....	.....	.....
y'_{k-1}	y'_k	y_k	n_k	h_k	N_k	H_k
		Σ	n	1		

Representación gráfica de las frecuencias absolutas y relativas simples en una Distribución de Frecuencias de Variable Continua (n_i y h_i)

La representación gráfica de las frecuencias absolutas y relativas simples, da lugar al denominado **"*Histograma de Frecuencias",*** que adopta la forma, considerando los datos del problema dado:

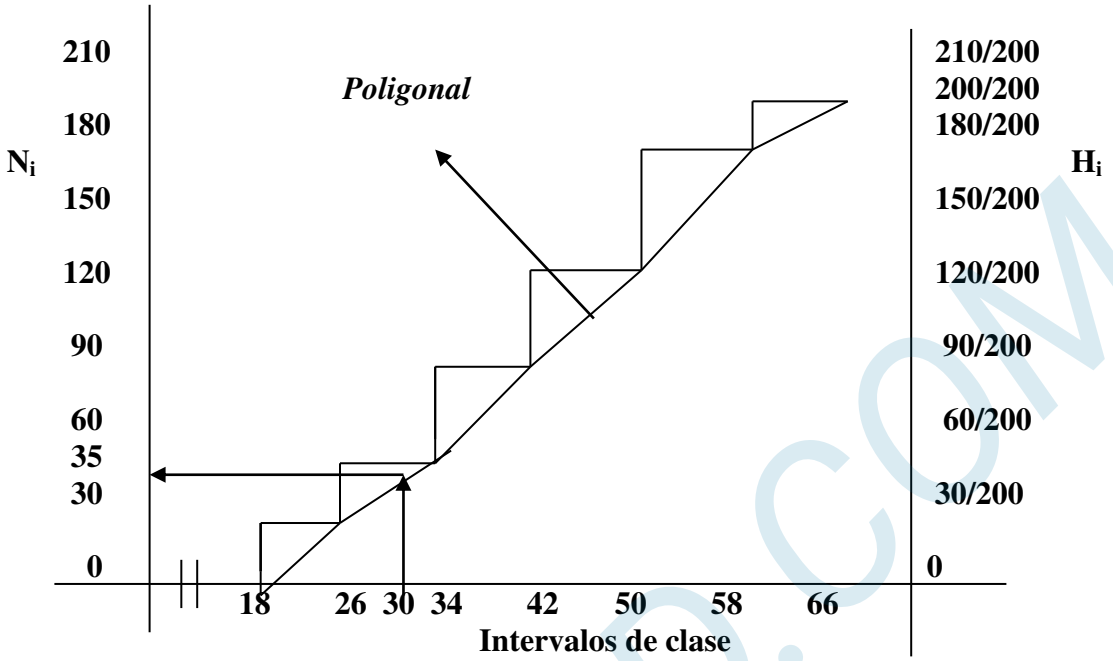


Uniendo los puntos medios de cada rectángulo del Histograma, que se corresponden con los valores de las **marcas de clase de cada intervalo**, queda determinada la ***"Poligonal de frecuencias absolutas o relativas Simples",*** *la que sirve* ***para caracterizar gráficamente*** a una *distribución de frecuencias de variable continua*, siendo el área entre la poligonal, y el eje de las abscisas, igual a la suma de las superficies de los rectángulos del histograma, que da lugar a la *"densidad de frecuencias"*. Cabe destacar que anteriormente, se planteó la conveniencia de que la amplitud de los intervalos de clase sea la misma, para facilitar la *interpretación gráfica de una distribución de frecuencias*.

Representación gráfica de las frecuencias absolutas y relativas acumuladas, en una Distribución de Frecuencias de Variable Continua (N_i y H_i)

Tanto en el caso de las Absolutas Acumuladas, como en el de las Relativas Acumuladas su representación gráfica consiste en el denominado ***"Diagrama escalonado con trazos continuos"***, siendo idéntico para ambos tipos de frecuencias, sólo con un cambio de escala en el eje de las ordenadas, en el caso de las relativas acumuladas.

Considerando los datos del problema anterior, responde a la forma:



La **poligonal de Frecuencias Acumuladas,** *permite realizar interpolaciones lineales gráficas*. Así por ejemplo si quisiéramos saber el número de empleados con edades de hasta 30 años, se tendrá: 20 + 15 (dado que en el segundo intervalo la frecuencia absoluta es 30, y al ser la interpolación lineal hasta 30 años, que es la marca de clase se tendrá 30/2 = 15).

DIAGRAMA DE TALLO Y HOJA

Otra forma de *organizar datos sin procesar,* en grupos, es a través de un ***Diagrama o Gráfico de Tallo y Hoja***, *el que se construye al separar los dígitos de cada número de los datos, en dos grupos, un* ***tallo*** *y una* ***hoja***. Los dígitos del *extremo izquierdo,* son el **tallo,** y están formados por los *dígitos,* de *más alto valor*. Los dígitos del extremo derecho, son las **hojas,** y contienen los *valores más bajos*. Si un conjunto de datos tiene *sólo dos dígitos*, el **tallo** es el *valor de la izquierda,* y la **hoja**, el de la derecha. Por ejemplo, si 56 es uno de los números, el **tallo** es 5 y la **hoja** 6. Para números con más dígitos, queda librado al juicio del investigador, la determinación del **tallo** y **hoja**.

Ejemplo:

Los siguientes valores se corresponden con las calificaciones obtenidas por 35 aspirantes a un puesto de trabajo:

86	77	91	60	55
76	92	47	88	67
23	59	72	75	83
77	68	82	97	89
81	75	74	39	67
79	83	70	78	91
68	49	56	94	81

Construir un ***Diagrama de Tallos y Hojas*:**

Solución:

Tallo	**Hoja**									
2	3									
3	9									
4	7	9								
5	5	6	9							
6	0	7	7	8	8					
7	0	2	4	5	5	6	7	7	8	9
8	1	1	2	3	3	6	8	9		
9	1	1	2	4	7					

Como se observa, podemos decir que un ***Diagrama de Tallos y Hojas*** presenta las siguientes ventajas:

1) Puede verse si las calificaciones están en el extremo superior o inferior dentro del rango de valores.

2) Los valores de los datos originales sin procesar se mantienen, permitiendo dar una estructura del histograma de frecuencia. (Recordar que en una distribución de frecuencia de variable continua se utiliza como representativo de los valores de las variables de cada intervalo a las marcas de clase).

UNIDAD Nº III

"PARÁMETROS Y ESTADÍSTICOS"

CONSIDERACIONES

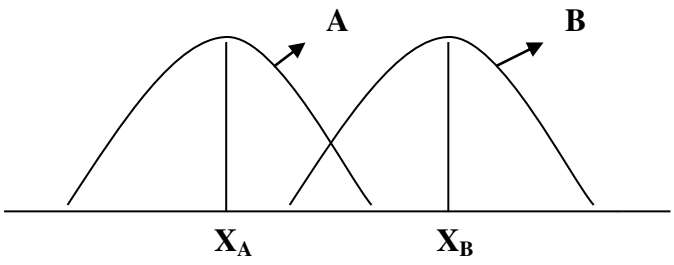
En las unidades anteriores, se desarrollaron *conceptos básicos de Estadística,* así como distintas formas de presentar en forma *organizada y resumida,* los datos relevados en una investigación estadística.

En la presente unidad, consideraremos las principales medidas estadísticas *descriptivas,* de ***resumen***, *de un conjunto de datos*. Están dadas por:

1) **Medidas de Posición.**
2) **Medidas de Dispersión.**
3) **Medidas de Forma, Asimetría o Sesgo.**
4) **Medidas de Puntiagudez o Curtosis.**

1) MEDIDAS DE POSICIÓN

Son medidas *descriptivas de resumen*, representativas del ***valor de la variable,*** en torno a la cual se concentran las observaciones. Así por ejemplo, si se tiene a las distribuciones **A** y **B**:



Se observa que ambas tienen ***la misma forma***, concentrándose los datos en torno al valor de la variable $\mathbf{X_A}$, en la distribución A, y en torno a $\mathbf{X_B}$, en la distribución B, siendo $\mathbf{X_A}$ y $\mathbf{X_B}$ valores de una ***medida de posición***.

PRINCIPALES MEDIDAS DE POSICIÓN

1) *MEDIA ARITMÉTICA*

a) Serie Simple:

Dada una Serie Simple, correspondiente a una muestra de **n** observaciones, de una variable $\mathbf{x_i} \Rightarrow x_1 , x_2 , x_3 , \ldots\ldots\ldots, x_n$, la **Media Aritmética o Promedio MUESTRAL**, simbolizada con **M(x)** o $\bar{x}$ **,** surge: *de la suma de los valores de la variable observada, dividido por el número de los mismos*, dando lugar a la denominada **"Fórmula de la Media Simple".** Es decir:

$$\mathbf{M(x) = \bar{x} = \frac{x_1 + x_2 + x_3 + ... + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}}$$

$$\Downarrow$$

Media Simple Muestral

En el caso de que la Serie Simple, se corresponda con todos los valores de la variable de la *población bajo estudio,* (**N** observaciones) **:** $x_1 , x_2 , x_3 , \ldots\ldots\ldots, x_N$, la **Media Aritmética o Promedio POBLACIONAL**, denotada con μ va a estar dada por:

$$M(x) = \mu = \frac{x_1 + x_2 + x_3 + ... + x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{N}$$

⇓

Media Simple Poblacional

b) <u>Datos Agrupados</u>:

En el caso de una Distribución de Frecuencias, tanto de variable discreta como de variable continua, la **Media Aritmética** denotada con M(y) o $\bar{y}$, surge: *de la suma de los productos entre cada valor de la variable o marca de clase,* (según sea distribución de frecuencia de variable discreta o de variable continua, respectivamente), *y las respectivas frecuencias absolutas*, dando lugar a la denominada **"Fórmula de la Media Ponderada"**, pudiendo ser **muestral** o **poblacional** según se considere los datos de una *muestra* o de *toda la población*, respectivamente. Es decir:

$$M(y) = \bar{y} = \frac{y_1.n_1 + y_2.n_2 + y_3.n_3 + ... + y_n.n_n}{n} = \frac{\sum_{i=1}^{n} y_i.n_i}{n}$$

⇓

Media Ponderada Muestral

$$M(y) = \mu = \frac{y_1.n_1 + y_2.n_2 + y_3.n_3 + ... + y_N.n_N}{N} = \frac{\sum_{i=1}^{N} y_i.n_i}{N}$$

⇓

Media Ponderada Poblacional

⇓

$y_i \Rightarrow$ **valor de la variable (DISCRETA)**

$y_i \Rightarrow$ **marca de clase (CONTINUA)**

Ejemplo 1:

El número de empleados que faltaron por día en una empresa, en los últimos 20 días laborables, está dado por:

4 ; 5 ; 2 ; 2 ; 0 ; 1 ; 1 ; 1 ; 6 ; 2 ; 2 ; 2 ; 2 ; 2 ; 5 ; 4 ; 4 ; 1 ; 1 ; 2

Calcular el *promedio diario* de empleados, que faltaron considerando:

a) Serie Simple.
b) Distribución de frecuencia de variable discreta.

Solución:

a) $\boxed{M(x) = \bar{x} = \frac{\sum_i x_i}{n}}$ $\Rightarrow$ ***Serie Simple***

$\Downarrow$

$$\boxed{M(x) = \bar{x} = \frac{49}{20} = 2{,}45 \text{ empleados por día}}$$

b) $\boxed{M(y) = \bar{y} = \frac{\sum_i y_i . n_i}{n} \Rightarrow y = \text{valor de la variable}}$ $\Rightarrow$ ***Datos Agrupados***

Construyendo la distribución de frecuencia de Variable Discreta:

Nº de empleados ausentes yi	**Nº de días ni**	**yi.ni**
0	1	0
1	5	5
2	8	16
4	3	12
5	2	10
6	1	6
$\sum$	**20**	**49**

Operando se obtiene $\Rightarrow$ $\boxed{M(y) = \bar{y} = \frac{49}{20} = \textbf{2,45 empleados por día}}$

Ejemplo 2:

La antigüedad en sus empleos (en años), de los empleados de un Hipermercado está dada por:

Antigüedad (en años)		**Nº de empleados**
y'_{i-1}	y'_i	n_i
3	5	50
5	7	70
7	9	90
9	11	40
11	13	30
13	15	20
	$\sum$	300

Calcular la **antigüedad promedio** de los empleados.

Solución:

$$\boxed{M(y) = \bar{y} = \frac{\sum_i y_i . n_i}{n} \Rightarrow y = \text{marca de clase}}$$

Antigüedad (en años)		**Nº de empleados**	**Marca de clase**	
y'_{i-1}	y'_i	n_i	y_i	$y_i\, n_i$
3	5	50	4	200
5	7	70	6	420
7	9	90	8	720
9	11	40	10	400
11	13	30	12	360
13	15	20	14	280
	$\sum$	300		2.380

Aplicando la fórmula, resulta:

$$M(y) = \bar{y} = \frac{\sum y_i . n_i}{n} = \frac{2.380}{300} = \textbf{7,93 años de antiguedad}$$

PROPIEDADES DE LA MEDIA ARITMÉTICA

I. ***"La media aritmética de una constante, es la constante misma".***

$$\Downarrow$$

$$\boxed{M(k) = k \;;\; k \rightarrow \text{constante}}$$

II. ***"La media aritmética del producto entre una constante y una variable es igual al producto de la constante por la media aritmética de la variable".***

$$\Downarrow$$

$$\boxed{M(k.x) = k.M(x) \;;\; k \rightarrow \text{constante}}$$

III. ***"La media aritmética de la suma de una constante y una variable es igual a la suma de la constante y la media aritmética de la variable".***

$$\Downarrow$$

$$\boxed{M(k + x) = k + M(x) \;;\; k \rightarrow \text{constante}}$$

IV. ***"La media aritmética de una suma de variables <u>expresadas en igual unidad de medida</u> es igual a la suma de las medias aritméticas de cada una de las variables consideradas".***

$$\Downarrow$$

$$\boxed{M(x + y) = M(x) + M(y) \quad ; x, y \rightarrow \text{variables}}$$

V. ***"La suma de los desvíos o diferencias entre cada valor de la variable y su media aritmética, ponderadas en el caso de datos agrupados, <u>es igual a cero</u>".***

$$\Downarrow$$

a) ***Serie Simple*** $\Rightarrow \sum (x_i - \bar{x}) = 0$

b) ***Datos Agrupados*** $\Rightarrow \sum (y_i - \bar{y})n_i = 0$

VI. ***"La suma de los desvíos o diferencias entre cada valor de la variable y su media aritmética elevadas al cuadrado y ponderadas en el caso de datos agrupados, <u>es un mínimo</u>".***

$$\Downarrow$$

a) ***Serie Simple*** $\Rightarrow \sum (x_i - \bar{x})^2 < \sum (x_i - k)^2 \; ; k \neq \bar{x}$

b) ***Datos Agrupados*** $\Rightarrow \sum (y_i - \bar{y})^2 . n_i < \sum (y_i - k)^2 . n_i \; ; \; k \neq \bar{y}$

<u>MEDIA GENERAL O TOTAL</u>

Dado un conjunto de **<u>n</u>** observaciones, las que subdividen en **<u>k</u>** grupos distintos, y se calcula en cada uno de ellos la media aritmética, la "***MEDIA TOTAL O GENERAL",*** se obtiene de la: *suma de los productos entre la media de cada grupo, y su respectivo tamaño, dividido todo por el total de observaciones.*

<u>Grupo</u>	$\bar{y}_i$	n_i
1	$\bar{y}_1$	n_1
2	$\bar{y}_2$	n_2

3	$\overline{y}_3$	n_3	$\Rightarrow$ **Media General o Total** $= M(\overline{y}) = \overline{\overline{y}} = \dfrac{\sum_{i=1}^{k} y_i . n_i}{n}$
.	.	.	.
.	.	.	.
k	$\overline{y}_k$	n_k	

Ejemplo:

Se clasificó al personal de una empresa en dos grupos y se calculó el sueldo medio mensual (en dólares) en cada grupo, obteniéndose los siguientes valores:

Empleados (A) $\overline{y}_A = 700$; $n_A = 450$

Directivos (B) $\overline{y}_B = 8.200$; $n_B = 15$

Calcular **el** ***promedio mensual de los sueldos pagados*** a **todo el personal** .

Solución:

$$\mathbf{M(\overline{y}) = \overline{\overline{y}} = \frac{\sum_{i=1}^{k} \overline{y}_i n_i}{n}} \Rightarrow M(\overline{y}) = \overline{\overline{y}} = \frac{\overline{y}_A . n_A + \overline{y}_B . n_B}{n_A + n_B} = \frac{700.450 + 8200.15}{450 + 15} = \frac{438.000}{465} = \mathbf{941,94\ U\$S}$$

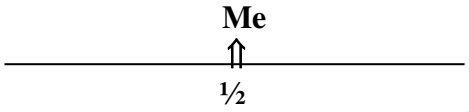
Promedio Erróneo $\Rightarrow \mathbf{M(\overline{y}) = \overline{\overline{y}} = \frac{\overline{y}_A + \overline{y}_B}{2} = \frac{700 + 8.200}{2} = \frac{8900}{2} = 4.450\ U\$S}$

$\Downarrow$

No tomar en consideración a las Ponderaciones

2) ***MEDIANA***

La Media Aritmética como medida de posición resulta **útil,** *siempre que los valores de la variable observada,* ***sean homogéneos***. En caso contrario, resulta más adecuada como Medida de Posición la ***MEDIANA***, dada por ***el valor de la variable que supera a no más de la mitad de las observaciones y es superada por no más de la mitad de las mismas, siendo menos sensible que la media aritmética ante la presencia de valores extremos***. Es decir, la ***MEDIANA,*** es el valor de la variable ubicado ***en el medio,*** de un conjunto de valores ordenados de la variable, o sea el ***valor central.*** Para que *un valor de la mediana tenga sentido*, el nivel de medida de los datos, debe ser por lo menos ordinal, para poder efectuar el ordenamiento. Se la denota con M_e o $X_{.5}$.



CÁLCULO DE LA MEDIANA

a) Serie Simple:

1) Se ordenan los datos en sentido creciente o decreciente.

2) Se determina la **Mediana,** que se corresponde con el valor de la variable ubicado en el *valor central*. Es decir:

$$\mathbf{Me} = x_{.5} = x_{\left[\frac{n+1}{2}\right]^{o}}$$

Observaciones:

- Si el número de datos (**n**) es **par,** el valor de la *Mediana* se obtiene de la ***media aritmética de los valores centrales de la*** **variable,** siendo *su valor coincidente con uno observado, sólo si son iguales los valores centrales.*

- Si el número de datos (**n**) es **impar,** el valor de la *Mediana* será igual a un *valor realmente observado de la variable.*

Ejemplo 1:

Las temperaturas observadas en un cierto lugar en grados centígrados (ºC), están dadas por:

-15 ; 2 ; 1 ; -2 ; -1 ; -3 ; -4 ; 1 ; -1 ; 0 ; 3

Calcular:

1) Temperatura media.
2) Temperatura mediana.

Solución:

1) ***Media*** $\Rightarrow M(x) = \bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n} \Rightarrow M(x) = \dfrac{-15+2+1+.....+3}{11} = \dfrac{-19}{11} =$ **-1,73 ºC**

2) ***Mediana:***

1º) Ordenar los datos

-15
- 4
- 3
- 2
- 1

$x_{6^o} \Rightarrow$ - 1 $\Rightarrow$ **Me = -1 ºC**

0
1
1
2
3

1. $Me = x_{.5} = x_{\left[\frac{n+1^o}{2}\right]} \Rightarrow n = 11 \Rightarrow Me = x_{\left[\frac{11+1^o}{2}\right]} = x_{6^o} =$ **-1 ºC**

Ejemplo 2:

Las temperaturas observadas en un cierto lugar en grados centígrados (ºC), están dadas por:

-15 ; 2 ; 1 ; -2 ; -1 ; -3 ; -4 ; 1 ; -1 ; 0 ; 3 ; 45

Calcular:

1) Temperatura media.

2) Temperatura mediana.

Solución:

1) ***Media*** $\Rightarrow M(x) = \bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n} \Rightarrow M(x) = \dfrac{-15+2+1+.....+3+45}{12} = \dfrac{26}{12} = \underline{\mathbf{2{,}17\ ^{o}C}}$

2) ***Mediana:***

1. **Ordenar los datos**

-15
- 4
- 3
- 2
- 1

$x_{6^o} \Rightarrow$ **-1** $\Big\} \Rightarrow$ **Me = - 0,5 ºC**

$x_{7^o} \Rightarrow$ **0**

1
1
2
3
45

2. $\mathbf{n = 12} \Rightarrow \mathbf{Me} = \mathbf{x}\left(\frac{\mathbf{n+1}}{\mathbf{2}}\right)^{o} = \mathbf{x}\left(\frac{12+1}{2}\right)^{} = x(6{,}5)^{o} = x\left(\frac{x6^{o}+x7^{o}}{2}\right) = \frac{-1+0}{2} = \underline{-\mathbf{0{,}5\ ^{o}C}}$

Observación:

Analizando los Ejemplos **1** y **2**, se observa que en el **2**, sólo se agregó a la serie simple del **1,** el valor **45**, lo que generó que la ***media aritmética*** pase de (-1,73 ºC) a (2,17 ºC), mientras que la ***mediana,*** operó modificación poco significativa. Es por ello, que se dice que la ***media aritmética*** **es muy sensible,** a la presencia de **valores extremos**, siendo **conveniente** en estos casos, calcular como **medida de posición** a la ***mediana***.

b) Datos Agrupados de Variable Discreta:

En el caso de una distribución de frecuencias o datos agrupados de variable discreta, el valor de la Mediana surge aplicando los siguientes pasos:

1. Calcular la ***mitad de las observaciones***, es decir: **n/2**, siendo **n** el total de datos.

2. Determinar $\mathbf{N_j}$, que es la ***menor de las frecuencias absolutas acumuladas, que supera a la mitad de las observaciones (n/2).***

3. Relacionar $\mathbf{N_{j-1}}$, que es la ***frecuencia absoluta acumulada inmediata anterior a*** $\mathbf{N_j}$, con **n/2**, pudiéndose presentar dos situaciones:

a) $\boxed{N_{j-1} = n/2}$

$\Downarrow$

$\boxed{Me = \frac{y'_{j-1} + y'_j}{2}} \Rightarrow$ Promedio aritmético de los valores de las variables relacionadas con N_{j-1} y N_j respectivamente.

b) $\boxed{N_{j-1} < n/2} \Rightarrow \boxed{Me = y_j} \Rightarrow y_j$ es el valor de la variable asociada a $\mathbf{N_j}$.

Ejemplo 1:

Calcular la Mediana de la siguiente distribución de frecuencia de variable discreta:

	y_i	n_i	N_i	
	1	9	9	
	3	12	**21**	$\rightarrow N_{j-1}$
Me = $y_j \rightarrow$	**7**	15	**36**	$\rightarrow N_j$
	9	10	46	
	20	5	51	
	100	3	54	
	$\sum$	54		

Solución:

Procediendo acorde a lo antes expuesto:

1. **$n/2 = ? \Rightarrow n/2 =$** $54/2 =$ **27**
2. **$N_j = ? \Rightarrow N_j = 36$**
3. **$N_{j-1} = 21 \Rightarrow N_{j-1} < n/2 \Rightarrow Me = y_j \Rightarrow Me = 7$**

Ejemplo 2:

Calcular la Mediana de la siguiente distribución de frecuencia de variable discreta:

	y_i	n_i	N_i	
	1	8	8	
	3	12	20	
$y_{j-1} \rightarrow$	**7**	7	**27**	$\rightarrow N_{j-1}$
$y_j \rightarrow$	**9**	13	**40**	$\rightarrow N_j$
	20	8	48	
	100	6	54	
	$\sum$	54		

Solución:

Procediendo en forma análoga al problema anterior:

1. **n/2 = ? ⇒ n/2 =** 54/2 = **27**

2. **N_j = ? ⇒ N_j = 40**

3. **N_{j-1} = 27 ⇒ N_{j-1} = n/2 ⇒** $\mathbf{Me} = \frac{y_{j-1} + y_j}{2} \Rightarrow \boxed{\mathbf{Me} = \frac{7+9}{2} = 8}$

c) <u>Datos Agrupados de Variable Continua</u>:

En el caso de una distribución de frecuencias o datos agrupados de variable continua, el valor de la Mediana surge de la aplicación de los siguientes pasos:

1. Calcular la ***<u>mitad de las observaciones</u>***, es decir: **n/2,** siendo **n** el total de observaciones.

2. Determinar **N_j,** que es la ***<u>menor</u> de las frecuencias absolutas acumuladas, <u>que supera</u> a la mitad de las observaciones (n/2).***

3. Relacionar **N_{j-1}** que es la ***frecuencia absoluta acumulada inmediata anterior a*** **N_j**, con **n/2.**

Se ***<u>demuestra</u>*** que, la ***<u>fórmula de cálculo de la mediana,</u> en una distribución de frecuencia de variable continua***, está dada por:

⇓

$$\mathbf{Me} = \mathbf{y'_{j-1}} + \mathbf{c_j}\,\frac{\mathbf{n/2} - \mathbf{N_{j-1}}}{\mathbf{n_j}}$$

Siendo:

- **y'_{j-1}** ⇒ extremo izquierdo del intervalo correspondiente a **N_j,** denominado ***intervalo mediano***.

- $\mathbf{c_j} \Rightarrow$ amplitud del ***intervalo mediano***.
- $\mathbf{n/2} \Rightarrow$ mitad de las observaciones.
- $\mathbf{N_{j-1}} \Rightarrow$ frecuencia absoluta acumulada, inmediata anterior a $\mathbf{N_j}$.
- $\mathbf{n_j} \Rightarrow$ frecuencia absoluta, correspondiente al ***intervalo mediano***.

Observación:

Si: $\boxed{\mathbf{N_{j-1} = n/2}} \Rightarrow \boxed{\mathbf{Me = y'_{j-1}}} \Rightarrow$ extremo izquierdo del ***intervalo de clase mediano***.

Ejemplo 1:

Calcular la ***mediana,*** de la siguiente distribución de frecuencia de variable continua:

	y'_{i-1}	y'_i	n_i	N_i
	4	8	5	5
	8	12	9	**14** $\rightarrow \mathbf{N_{j-1}}$
$y'_{j-1} \rightarrow$	**12**	16	20	**34** $\rightarrow \mathbf{N_j}$
	16	20	6	40
	20	24	10	50
		$\sum$	50	

Solución:

1. $\mathbf{n/2 = ?} \Rightarrow \mathbf{n/2} = 50/2 = \mathbf{25}$

2. $\mathbf{N_j = ?} \Rightarrow \mathbf{N_j = 34} \Rightarrow \mathbf{N_{j-1} = 14}$

3. Relacionar $\mathbf{N_{j-1}}$ con $\mathbf{n/2} \Rightarrow 14 < 25 \Rightarrow \boxed{\mathbf{Me} = y'_{j-1} + c_j \dfrac{\mathbf{n/2} - \mathbf{N_{j-1}}}{\mathbf{n_j}}}$

Reemplazando en la fórmula por sus iguales, resulta:

⇓

$$\mathbf{Me = 12 + 4\frac{25 - 14}{20} = 14{,}2}$$

Ejemplo 2:

Calcular la mediana de la siguiente distribución de frecuencia de variable continua:

	y'_{i-1}	y'_i	n_i	N_i
	4	8	5	5
	8	12	9	14
	12	16	20	**25** → N_{j-1}
Me = y'_{j-1} →	**16**	20	16	**41** → N_j
	20	24	10	50
		$\sum$	50	

Solución:

1. **$n/2 = ? \Rightarrow n/2 = 50/2 = 25$**
2. **$N_j = ? \Rightarrow N_j = 41 \Rightarrow N_{j-1} = 25$**
3. Relacionar **N_{j-1}** con **$n/2 \Rightarrow 25 = 25 \Rightarrow Me = y'_{j-1} \Rightarrow$ Me = 16**

PROPIEDAD DE LA MEDIANA

Se demuestra que: *"La suma de los desvíos **tomados en valores absolutos**, entre **cada valor de la variable y la mediana**, ponderados en el caso de una distribución de frecuencia, **es un mínimo**".*

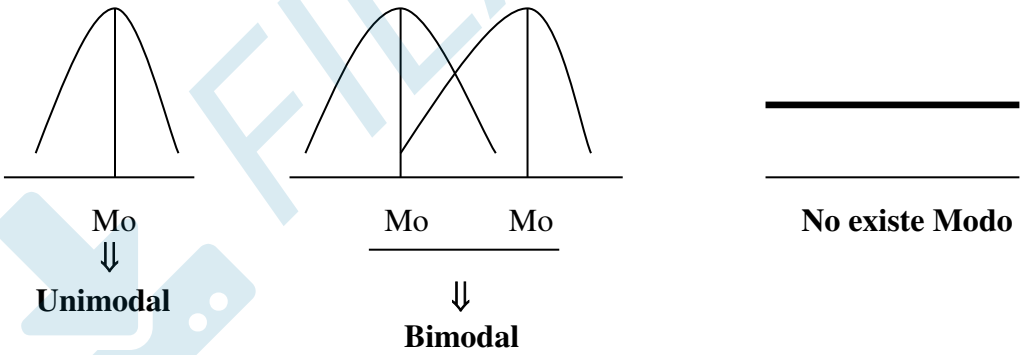
⇓

$$\sum | y_i - Me | n_i < \sum | y_i - k | n_i \text{ ; siendo } k \neq Me$$

3) MODO, MODA O VALOR MODAL

Es otra ***medida de posición***, definida ***por el valor de la variable con mayor frecuencia de presentación***. También se dice, que es el ***valor típico*** de un conjunto de datos, y resulta aplicable tanto a variable *cuantitativa*, como a variable *cualitativa*.

Si de un conjunto de datos, *un valor de la variable* tiene mayor frecuencia que los demás, se dice que la *distribución es* ***unimodal***, mientras que si *son dos los valores* que tienen mayor frecuencia, se dice que es ***bimodal***. Cabe destacar que, si un conjunto de datos no es exactamente ***bimodal*** pero contiene dos valores que son más dominantes que otros, algunos investigadores denominan al conjunto de datos como bimodal incluso sin un empate exacto para la moda. Los *conjuntos de datos con más de dos modas*, se conocen como ***multimodales***, dándose situaciones, en las que una distribución ***carezca de modo***. Se denota al modo con **Mo**.



En el mundo de los negocios, el concepto de moda, se usa con frecuencia al determinar medidas. Por ejemplo, la industria del vestido produce camisas, vestidos y trajes y muchos otros productos de vestido en *talles modales*. Al reducir el número de talles a unas pocas medidas modales, las compañías reducen sus costos incrementando de esta forma sus utilidades.

CÁLCULO DEL MODO

a) Serie Simple: El ***Modo***, si existe, será el valor de la variable que se presenta con más frecuencia en la serie de datos.

Ejemplos:

Calcular en cada caso, si existe, el ***Modo*:**

1. El Nº de artículos fallados por día en una línea está dado por:

   6 ; 8 ; 10 ; 8 ; 8 ; 6 ; 5 ; 3.

2. El Nº de empleados que faltaron a su trabajo, por día, está dado por:

   3 ; 1 ; 1 ; 1 ; 2 ; 2 ; 2 ; 2 ; 1 ; 4 ; 5 ; 3.

3. El estado civil de un conjunto de empleados es:

   Casado; Soltero; Casado; Casado; Soltero; Soltero; Soltero.

4. El medio de movilidad escogido por un grupo de empleados para ir a su lugar de trabajo, está dado por:

   Moto; Auto; Colectivo; Caminando; Moto; Colectivo; Caminando; Auto.

5. El color de los autos vendidos en una semana en una concesionaria es:

   Blanco ; Rojo ; Gris ; Azul ; Negro.

Soluciones:

1. **Mo = 8 ⇒ *Unimodal***
2. **Mo = 1 y Mo = 2 ⇒ *Bimodal***
3. **Mo = soltero ⇒ *Unimodal***
4. **Mo ⇒ *No existe***
5. **Mo ⇒ *No existe***

4) PERCENTILES o CENTILES

Son medidas de *posición o tendencia central, que* ***dividen a un grupo de datos*** *en* ***100 partes***. Hay 99 **Percentiles**, porque se necesita 99 divisores para separar un grupo de datos en 100 partes. El ***n-ésimo*** **Percentil** es el valor de la variable tal que ***"al menos n por ciento de los datos están bajo de ese valor, y a lo sumo (100 – n) por ciento superan a ese valor".*** Así por ejemplo, el percentil 67 es un valor de la variable tal que, al menos el 67% de los datos están por debajo de ese valor, y no más del 33% están sobre ese valor.

Observación: *Si consideramos la definición de* ***Mediana*** *y de* ***Percentiles****, podemos concluir que* ***siempre*** *se va a verificar que:* $\mathbf{P_{50} = Me}$ (*valor de la variable que supera a no más de la mitad de las observaciones y es superado por no más de la mitad de las mismas).*

5) CUARTILES

Los ***cuartiles*** son *medidas de posición o tendencia central,* que dividen a un conjunto de datos en cuatro partes iguales, teniendo por lo tanto 3 cuartiles:

- $\mathbf{Q_1 \Rightarrow}$ **Primer Cuartil**: representativo del valor de la variable que supera a no más del 25% de las observaciones y es superado por no más del 75% de las mismas. Es decir, separa el cuarto más bajo de los datos de los tres cuartos más altos, siendo por lo tanto igual al **Percentil 25**.

$$\Downarrow$$

$$\boxed{\mathbf{Q_1 = P_{25}}}$$

- $\mathbf{Q_2 \Rightarrow}$ **Segundo Cuartil:** representativo del valor de la variable que supera a no más del 50% de las observaciones y es superado por no más del 50% de las mismas. Es decir, separa la mitad más baja de los datos de la mitad más alta, siendo igual al **Percentil 50,** e igual a la **Mediana.**

$$\Downarrow$$

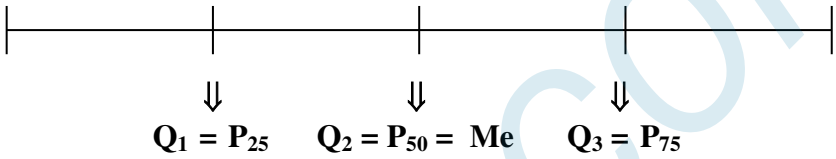
$$\boxed{\mathbf{Q_2 = P_{50} = Me}}$$

- **$Q_3 \Rightarrow$ Tercer Cuartil:** representativo del valor de la variable que supera a no más del 75% de las observaciones y es superado por no más del 25% de las mismas. Es decir, separa los tres cuartos más bajos del cuarto más alto, siendo igual al **Percentil 75**.

$$\Downarrow$$

$$\boxed{Q_3 = P_{75}}$$

***<u>Esquemáticamente</u>*:**



6) <u>DECILES</u>

Son medidas de *posición o tendencia central, que* ***dividen a un grupo de datos*** *en* ***10 partes iguales***. Hay 9 **Deciles**, porque se necesita 9 divisores, para separar un grupo de datos en 10 partes iguales. El ***n-ésimo* Decil,** es el valor de la variable tal que ***"al menos <u>n</u> por ciento de los datos están bajo de ese valor, y a lo sumo <u>(100 – n)</u> por ciento, superan a ese valor".*** Así por ejemplo, el **decil 6,** es un valor de la variable tal que, *al menos el 60%* de los datos están por debajo de ese valor, y *no más del 40% están sobre ese valor*.

Si denotamos al **Decil** que ocupa el lugar **i** de la forma: $\mathbf{D_i}$ podemos establecer una relación entre ***Deciles y Percentiles*** de la forma:

$$D_1 = P_{10}$$

$$D_2 = P_{20}$$

$$\vdots$$

$$\vdots$$

$$D_5 = P_{50} \Rightarrow \text{Mediana}$$

$$\vdots$$

$$\vdots$$

$$D_9 = P_{90}$$

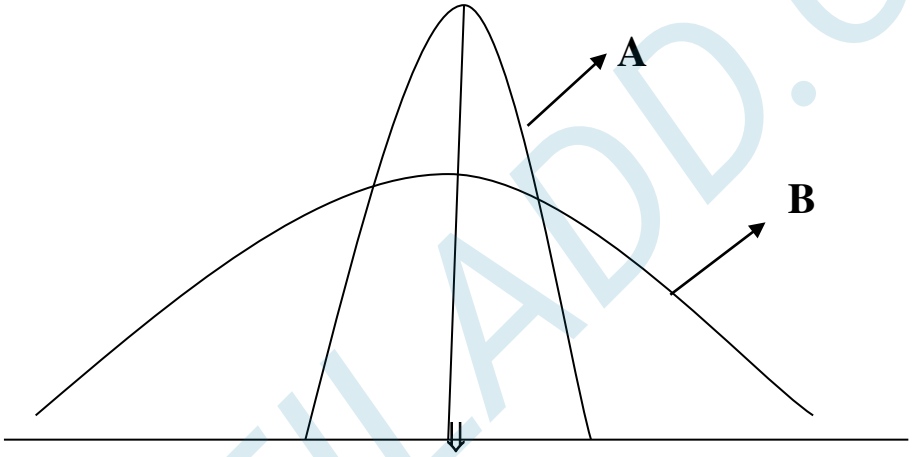
<u>OBSERVACIÓN</u>: En forma genérica a los ***Percentiles, Cuartiles y Deciles,*** se los denomina **FRACTILES,** dado que dividen a un conjunto de datos, en fracciones o partes. Como antes se comprobó, se verifica que:

$$P_{50} = Q_2 = D_5 = M_e$$

2) MEDIDAS DE DISPERSIÓN

Las medidas de posición o tendencia central, dan información acerca del valor de la variable en torno al cual *se concentran los datos*, pero *nada dicen* de *qué forma, los mismos se distribuyen.* Es por ello, que es necesario para describir a un conjunto de valores de una variable, de otras herramientas analíticas que son las ***medidas de dispersión o variabilidad***, representativas de la ***dispersión de un conjunto de datos***, de forma tal que, conjuntamente con *las medidas de posición*, es posible obtener una descripción numérica más completa.

Así por ejemplo, si se tiene a las distribuciones **A** y **B**:



Se observa que ambas distribuciones tienen la misma medida de posición, pero en la distribución **A** los valores de la variable están ***más concentrados o menos dispersos*** que en la distribución **B.**

PRINCIPALES MEDIDAS DE DISPERSIÓN

1) RECORRIDO, RANGO o AMPLITUD

Está dada por ***"la diferencia entre el mayor valor observado de la variable y el valor menor de la misma".***

a) En el caso de una **Serie Simple**, el ***Recorrido*** (**R**) será el valor resultante del cálculo:

$$R = x_{máx} - x_{mín}$$

b) En el caso de **Datos Agrupados**, el ***Recorrido*** (**R**) será el valor resultante del cálculo:

Variable Discreta $\Rightarrow$ $\mathbf{R} = y_m - y_0 \Rightarrow$

- $y_m \Rightarrow$ valor mayor
- $y_0 \Rightarrow$ valor menor

Variable Continua $\Rightarrow \mathbf{R} = y'_m - y'_0 \Rightarrow$

- $y'_m \Rightarrow$ extremo der.último. int.
- $y'_0 \Rightarrow$ extremo izq.primer int.

Observaciones:

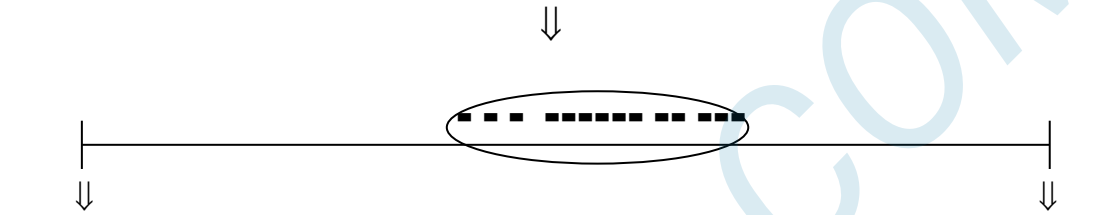
1) Cuando se desea comparar dos conjuntos de datos, expresados **en la misma unidad de medida**, la ***aplicación práctica*** *del* ***Recorrido*** como *medida de dispersión*, *está condicionada* a que, los conjuntos ***tengan el mismo valor de la Media Aritmétic****a*.

2) Si bien el ***Recorrido*** es de fácil cálculo, ya que toma en consideración sólo el valor mayor y el menor de la variable, en los casos en que éstos sean ***dispares*** y los datos *concentrados*, por ejemplo en la *zona central* o en *uno de los extremos*, su valor

como medida de dispersión *carecería de sentido*, por no tomar en consideración las situaciones mencionadas.

3) Un uso importante del ***Recorrido*** como *medida de variabilidad,* es en Control de Calidad, donde se lo utiliza para elaborar los "Gráficos de Control", que permiten determinar si un proceso opera "*bajo control*".

RECORRIDO POCO ÚTIL



o también:

Ejemplo:

Dadas las Series Simples:

A ⇒ 0 ; 25 ; 75 ; 100

B ⇒ 48 ; 49 ; 51 ; 52

Considerando el Recorrido como medida de dispersión, ¿ En **qué serie,** los datos son más heterogéneos o más dispersos?.

Solución:

$$\left.\begin{array}{l} R_A = 100 - 0 = 100 \\ R_B = 52 - 48 = 4 \end{array}\right\} \Rightarrow R_A > R_B \Rightarrow \text{“}\textit{datos más dispersos en Serie A”}$$

$$\Downarrow$$

$$M_A = M_B = 50$$

$$\Downarrow$$

“***El Recorrido como medida de dispersión aporta decisión”***

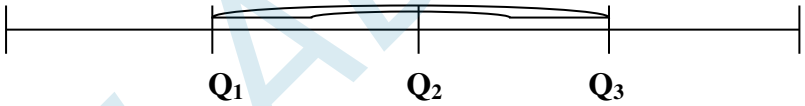
2) RECORRIDO O RANGO INTERCUARTÍLICO

Está dado por ***“la diferencia entre el tercer y primer cuartil”***. Es decir, se corresponde con *la amplitud existente entre el 50% de los valores cen*trales:

$$\Downarrow$$

$$\boxed{RIQ = Q_3 - Q_1}$$

$$\Downarrow$$



Observaciones:

1) El ***Recorrido Intercuartílico*** como medida de dispersión, tiene una aplicación práctica limitada, dado que no considera los datos ubicados a la izquierda del primer cuartil o a la derecha del tercero.

2) Al ***Recorrido Intercuartílico*** y al ***Recorrido, Rango o Amplitud***, al tomar en consideración sólo dos valores de la variable, se las denomina *“medidas posicionales de dispersión”.*

Ejemplo:

Si de un conjunto de datos se obtiene que:

$$Q_3 = 36{,}4$$

$$Q_1 = 15{,}1$$

Calcular el ***Recorrido Intercuartílico***, e interpretar.

Solución:

$$\boxed{\mathbf{RIQ = Q_3 - Q_1}} \Rightarrow \boxed{\mathbf{RIQ} = 36{,}4 - 15{,}1 = \mathbf{21{,}3}}$$

$\Downarrow$

"Existe una amplitud de 21,3 entre los extremos de los valores de la variable ubicados en el 50% de los valores centrales".

3) DESVIACIÓN MEDIA

Es otra medida de dispersión, y está dada por la ***"Media aritmética de los valores absolutos, de los desvíos existentes entre cada valor de la variable, y la media aritmética".***

a) Serie Simple:

$\Downarrow$

1. **Poblacional** $\Rightarrow$ $$DM(x) = \frac{\sum_{i=1}^{N} |x_i - \mu|}{N}$$

2. **Muestral** $\Rightarrow$ $$DM(x) = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

b) Datos Agrupados:

$\Downarrow$

1. **Poblacional** $\Rightarrow$ $$DM(y) = \frac{\sum_{i=1}^{N} |y_i - \mu| n_i}{N} \quad ; \quad y_i = \begin{cases} \nearrow \text{Valor de variable } (\mathbf{D}) \\ \searrow \text{Marca de clase } (\mathbf{C}) \end{cases}$$

Valor de variable (**D**)

2. Muestral $\Rightarrow$ $DM(y) = \dfrac{\sum_{i=1}^{n} |y_i - \bar{y}| \, n_i}{n}$; $y_i =$

Marca de clase (**C**)

Observación:

Debido a que la **Desviación Media**, considera el *promedio de los desvíos tomados en **valores absolutos***, *no permite determinar el **signo de las desviaciones***, razón por la cual como
medida de dispersión, es menos útil que otras. En el campo de pronósticos, se usa ocasionalmente como medida de error.

Ejemplo:

Un taller mecánico inició sus actividades la semana pasada, siendo el número diario de vehículos que ingresaron a reparación

Día	**Nº de vehículos**
Lunes	5
Martes	9
Miércoles	16
Jueves	17
Viernes	18

Calcular la ***Desviación Media***.

Solución:

Nº de vehículos

	x_i	$x_i - \mu$	$\lvert x_i - \mu \rvert$
	5	-8	8
	9	-4	4
	16	3	3
	17	4	4
	18	5	5
Σ	65	0	24

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} \Rightarrow \mu = \frac{65}{5} = \underline{13} \Rightarrow DM(x) = \frac{\sum_{i=1}^{N} |x_i - \mu|}{N} = \frac{24}{5} = \underline{4{,}8}$$

4) VARIANZA

Es una de la más importante medida de dispersión, y está dada por **"La media aritmética del cuadrado de los desvíos entre, cada valor de la variable y la respectiva media aritmética".** Suponiendo una variable **x**, se la denota de la forma: **V(x) o $\sigma^2(x)$.**

▪ *FÓRMULA DE LA VARIANZA SEGÚN LA DEFINICIÓN*

a) Serie Simple:

$$\Downarrow$$

$$V(x) = \sigma^2{}_x = M\left(x - \bar{x}\right)^2 = \frac{\sum_{i=1}^{n} (x - \bar{x})^2}{n}$$

b) Datos Agrupados:

$$\Downarrow$$

$$V(y) = \sigma_y = M\left(y - \bar{y}\right)^2 = \frac{\sum_{i=1}^{n} (y - \bar{y})^2}{n}$$

$y_i =$ *valor de la variable* $\Rightarrow$ *Variable Discreta*

$y_i =$ *marca de clase* $\Rightarrow$ *Variable Continua*

FÓRMULA DE CÁLCULO DE LA VARIANZA

Partiendo de la fórmula de Varianza según la definición, y suponiendo una Serie Simple, sabemos que:

$$\mathbf{V(x) = M\,(\,x - \bar{x}\,)^2}$$

Operando en el segundo miembro, desarrollando el cuadrado del binomio y aplicando propiedades de la Media Aritmética, resulta:

$$\mathbf{V(x) = M}\,(x - \bar{x}\,)^2 = M\,(x^2 - 2\,x\,\bar{x} + x^2) = M\,[x^2 - M(2x\,\bar{x}\,) + x^2] \;;\; M(x) = \bar{x} = \text{const.}$$

$$= M(x^2) - 2\,\bar{x}\,M(x) + M(\bar{x}^2) = M(x^2) - 2\,[M(x)]^2 + [M(x)]^2$$

$$\Downarrow$$

$$\mathbf{V(x) = M(x^2) - [M(x)]^2} \Rightarrow \mathbf{V(x)} = \frac{\sum x_i^2}{N} - \left[\frac{\sum x_i}{N}\right]^2$$

$$\Downarrow$$

En forma análoga se demuestra que en el caso de datos agrupados, la **fórmula de cálculo de la Varianza** va a estar dada por :

$$\mathbf{V(y) = M(y)^2 - [M(y)]^2} \Rightarrow \mathbf{V(y)} = \frac{\sum y_i^2 n_i}{N} - \left[\frac{\sum y_i n_i}{N}\right]^2$$

$$\Downarrow$$

$\mathbf{y_i}$ = *valor de la variable* $\Rightarrow$ *Variable Discreta*

$\mathbf{y_i}$ = *marca de clase* $\Rightarrow$ *Variable Continua*

IMPORTANTE: un valor particular de la Varianza, NADA INDICA

PROPIEDADES DE LA VARIANZA

1) *"La Varianza es una magnitud no negativa".*

$$\Downarrow$$

$$\boxed{V(x) \geq 0}$$

2) *"La Varianza de una constante es igual a cero".*

$$\Downarrow$$

$$\boxed{V(k) = 0 \quad ; \quad k = \text{constante}}$$

3) *"La Varianza del producto entre,* ***una constante y una variable,*** *es igual al* ***producto*** *entre,* ***la constante elevada al cuadrado,*** *por la Varianza de la variable".*

$$\Downarrow$$

$$\boxed{V(kx) = k^2\, V(x) \quad ; \quad k = \text{constante}}$$

4) *"La Varianza de la suma entre una constante y una variable es igual a la varianza de la variable".*

$$\Downarrow$$

$$\boxed{V(k + x) = V(x) \quad ; \quad k = \text{constante}}$$

5) *"La Varianza de una suma de variables, en general,* ***no es igual a la suma de las varianzas de cada una de las variables***".

$$\Downarrow$$

$$\boxed{V(x + y) \neq V(x) + V(y)}$$

5) DESVIACIÓN ESTÁNDAR

El **valor** de la ***Varianza,*** viene expresado en la *unidad de medida* de la variable *elevada al cuadrado*, mientras que las *medidas de posición* vienen expresadas en la *unidad de medida original de la variable*, lo que **no permite una comparación directa**. Es por ello que, a los fines prácticos, se utiliza como medida de dispersión a la ***Desviación Estándar o Típica,*** dada por **"La raíz cuadrada positiva de la varianza".** De esta forma, el valor de la variable asociado a la *tendencia central*, así como el *correspondiente a la dispersión*, se corresponderán con la *unidad de medida original de la variable*, permitiendo con ello, su comparación. Se la denota con **σ,** de modo tal que:

Desviación Estándar $\Rightarrow \sigma = +\sqrt{\sigma_y^2} = +\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$

IMPORTANTE: un valor particular de la Desviación Estándar, NADA INDICA

Ejemplo 1:

De las10 piezas producidas en una hora en una fábrica, se verificó si alguna de ellas presentaba algún defecto, obteniéndose los siguientes valores:

x_i = N° de defectos por pieza producida $\Rightarrow$ 0 ; 1 ; 3 ; 0 ; 1 ; 0 ; 1 ; 1 ; 1 ; 0

Obtener:

a) **Varianza**
b) **Desviación Estándar**

Considerando:

I. ***Serie Simple.***
II. ***Datos Agrupados de Variable Discreta.***

Solución:

I. ***Serie Simple:***

	x	x^2
	0	0
	1	1
	3	9
	0	0
	1	1
	0	0
	1	1
	1	1
	1	1
	0	0
Σ	8	14

a) $V(x) = \sigma^2 = M(x^2) - [M(x)]^2 \Rightarrow V(x) = \dfrac{\sum x_i^{\,2}}{N} - \left[\dfrac{\sum x_i}{N}\right]^2$

$$\Downarrow$$

$$V(x) = \frac{\sum_{i=1}^{10} xi^2}{10} - \left[\frac{\sum_{i=1}^{10} xi}{10}\right]^2 = \frac{14}{10} - \left[\frac{8}{10}\right]^{[2]} = 1{,}4 - 0{,}64 = \underline{\underline{0{,}76\ (\text{defectos}^2)}}$$

b) $\sigma = +\sqrt{V(x)} \Rightarrow \sigma = +\sqrt{0{,}76\ (\text{defectos})^2} = \underline{\underline{0{,}87\ \text{defectos}}}$

II. ***Datos agrupados de variable discreta***

Defectos y_i	Nº de piezas n_i	$y_i\, n_i$	$y_i^2\, n_i$
0	4	0	0
1	5	5	5
3	1	3	9
Σ	10	8	14

a) $$V(y) = M(y)^2 - [M(y)]^2 \Rightarrow V(y) = \frac{\sum yi^2 ni}{n} - \left[\frac{\sum yi\, ni}{n}\right]^2$$

$$\Downarrow$$

$$V(y) = \frac{14}{10} - \left[\frac{8}{10}\right]^2 = 1{,}4 - 0{,}8^2 = 1{,}4 - 0{,}64 = \underline{\underline{0{,}76\ (\text{defectos})^2}}$$

b) $$\sigma = +\sqrt{0{,}76\ (\text{defectos})^2} = \underline{\underline{0{,}87\ \text{defectos}}}$$

Ejemplo 2:

El precio de venta al público (en dólares), de un cierto artículo en distintos comercios de una ciudad, está dado por:

Precios		**Nº de comercios**	**Marca de clase**		
y'_{i-1}	y'_i	n_i	y_i	$y_i\, n_i$	$y_i^2\, n_i$
43	53	5	48	240	11.520
53	63	7	58	406	23.548
63	73	13	68	884	60.112
73	83	9	78	702	54.756
83	93	6	88	528	46.464
	Σ	**40**		**2.760**	**196.400**

Obtener:

a) Media aritmética.
b) Varianza.
c) Desviación estándar
d) En qué intervalo se encuentra el "valor modal"?. Cómo interpreta al mismo?.

Solución:

a) $$M(y) = \frac{\sum_{i=1}^{n} y_i n_i}{n} = \frac{2.760}{40} = \underline{\underline{69 \text{ dólares}}}$$

b) $$V(y) = M\left(y^2\right) - [M(y)]^2 \Rightarrow V(y) = \frac{\sum y_i^2 n_i}{n} - \left[\frac{\sum y_i n_i}{n}\right]^2$$

$$\Downarrow$$

$$V(y) = \frac{196.400}{40} - \left[\frac{2.760}{40}\right]^2 = \underline{\underline{149 \left(\text{dólares}\right)^2}}$$

c) $$\sigma_{(y)} = +\sqrt{\sigma^2} = +\sqrt{149 \text{ dólares}^2} = \underline{\underline{12{,}21 \text{ dólares}}}$$

d) **Mo** $\subseteq$ **[63 73[** $\Rightarrow$ *"El **precio más frecuente** a que **se vende el artículo**, se encuentra comprendido entre 63 y **menos** de 73 dólares".*

6) COEFICIENTE DE VARIACIÓN

Es una medida de *dispersión relativa,* dada por ***"el cociente entre la desviación estándar y la media aritmética"***, siendo una magnitud ***a-dimensional***, es decir ***sin unidad de medida***, razón por la cual puede ser aplicado para comparar distribuciones, *con igual* o *distinta unidad de medida*, pudiéndoselo expresar en porcentajes.

$$\Downarrow$$

$$\mathbf{CV(x) = \frac{Desviación estándar}{Media aritmética} = \frac{\sigma}{\mu} \quad \Rightarrow \quad CV(x)\,\% = \frac{\sigma}{\mu}.100}$$

Observación:

"Un menor valor del Coeficiente de Variación, significa una Media aritmética más representativa como medida de posición, de un conjunto de datos".

$$\Downarrow$$

Si se tiene a dos distribuciones **I** y **II**, y sus respectivos Coeficientes de Variación: $\mathbf{CV_I}$ y $\mathbf{CV_{II}} \Rightarrow$ Si: $\mathbf{CV_I < CV_{II}} \Rightarrow$ "$\mathbf{M_I}$ **es** ***más representativa*** que $\mathbf{M_{II}}$".

$$\Downarrow$$

"datos ***más uniformes*** en distribución **I"**

$$\Downarrow$$

"datos ***más parejos*** en distribución **I"**

$$\Downarrow$$

"datos ***más uniformes*** en distribución **I"**

$$\Downarrow$$

"datos ***más homogéneos*** en distribución **I"**

$$\Downarrow$$

"datos ***menos dispersos*** en distribución **I"**

$$\Downarrow$$

"datos ***con menor variabilidad*** en distribución **I"**

Ejemplo 1:

Se tiene dos acciones **A** y **B**, cuyas cotizaciones en un mercado bursátil en una semana, permitieron el cálculo de los siguientes valores, en dólares:

Acción A $\Rightarrow$ Precio Medio = 64,40 ; Desviación Estándar = 4,84

; **Acción B** $\Rightarrow$ Precio Medio = 13 ; Desv. Estándar = 3,03

¿Qué Acción es ***más riesgosa?.*** Fundamentar.

Solución:

Un indicador del ***riesgo*** de una Acción, es la ***variabilidad en los precios de su cotización,*** es decir, de la ***dispersión*** en los ***precios***.

Si consideramos los valores de la **Desviación Estándar** en **forma independiente** de cada una de las acciones, se observa que **el** de la **Acción A es mayor** que **el** de **la Acción B**, con lo que concluiríamos ***ERRÓNEAMENTE*** en decir que: *"la Acción **A es más riesgosa** que la **B,** dado que sus **precios de cotización,** presentan **mayor dispersión al ser mayor el valor de la desviación estándar**".*

La decisión es **INCORRECTA,** dado que los valores de los **Precios Medios** son **distintos**, siendo por lo tanto **necesario**, el cálculo de una ***medida de dispersión relativa,*** que considere *no sólo la desviación estándar en forma aislada*, sino que **también,** el respectivo *valor de la media aritmética.*

En consecuencia, es **necesario** el cálculo del ***Coeficiente de Variación***:

$$CV_A = \frac{\sigma_A}{\mu_A} \Rightarrow CV_A = \frac{4{,}84\ u\$s}{64{,}40\ u\$s} = \frac{4{,}84}{64{,}40} = 0{,}075 \Rightarrow 7{,}5\%$$

$$\Downarrow$$

"***La desviación estándar es de 7,5% a partir de la media***"

Procediendo en forma análoga con la Acción B, se tiene:

$$CV_B = \frac{\sigma_B}{\mu_B} \Rightarrow CV_B = \frac{3{,}03\ u\$s}{13\ u\$s} = \frac{3{,}03}{13} = 0{,}233 \Rightarrow 23{,}3\%$$

$$\Downarrow$$

"***La desviación estándar es de 23,3% a partir de la media"***

Dado que: $CV_B > CV_A$ **podemos CONCLUIR ADECUADAMENTE,** diciendo que: "Los precios de cotización de las **Acciones B** presentan ***mayor variabilidad relativa***, y por ende, **son más riesgosas que las Acciones A",** lo que es equivalente a decir que "*El **precio medio** de las Acciones **A, es más representativo,** como medida de posición, que el de la Acción **B**".*

Ejemplo 2:

La gerencia de Recursos Humanos de una empresa, desea realizar una investigación sobre sus empleados, para lo cual se consideraron dos variables: edad y peso, deseando

realizar el estudio tomando aquella variable, que presente ***valores más homogéneos***. Relevando los datos se obtuvo que:

Edad = (**x**) → Media aritmética = 47 años; Varianza = 225 $(años)^2$; Peso= (**y**) → Media aritmética = 75 kg; Varianza = 289 $(kg)^2$

Solución:

Para poder determinar *qué variable* presenta *valores* ***más homogéneos***, debemos calcular el **Coeficiente de Variación:**

$$\textbf{Edad} \Rightarrow CVx = \frac{\sigma_x}{\mu_x} \Rightarrow CV_x = \frac{\sqrt{225\ \text{años}^2}}{47\ \text{años}} = \frac{15\ \text{años}}{47\ \text{años}} = \frac{15}{47} = 0{,}319 \Rightarrow 31{,}9\%$$

⇓

"La desviación estándar es de 31,9%, a partir de la media"

$$\textbf{Peso} \Rightarrow CVy = \frac{\sigma_y}{\mu_y} \Rightarrow CVy = \frac{\sqrt{289\ \text{kg}^2}}{75\ \text{kg}} = \frac{17\ \text{kg}}{75\ \text{kg}} = \frac{17}{75} = 0{,}227 \Rightarrow 22{,}7\%$$

⇓

"La desviación estándar es de 22,7% a partir de la media"

$CV_y < CV_x \Rightarrow$ M(y) *"es más representativa"* ⇒ *valores "más HOMOGÉNEOS".*

⇓

Se trabajará con la variable ***PESO***

SIGNIFICADO PRÁCTICO DE LA DESVIACIÓN ESTÁNDAR

Si quisiéramos saber ¿***Cómo se interpreta el valor de la Desviación Estándar***?. Podemos hacerlo, además de la aplicación del *Coeficiente de Variación*, utilizando dos herramientas estadísticas, dadas por:

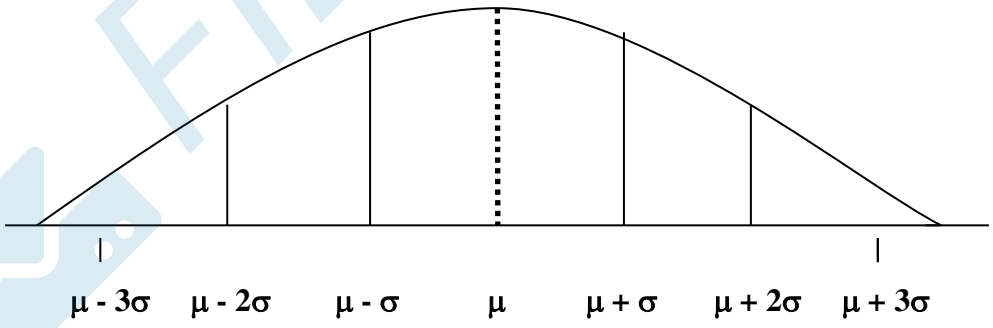
I. REGLA EMPÍRICA.

II. TEOREMA O DESIGUALDAD DE CHEBYSHEV.

I. REGLA EMPÍRICA

Es una importante *regla práctica*, que se usa para expresar el *porcentaje aproximado de datos*, que están *dentro* de un número dado de desviaciones estándar desde *la media aritmética*: *"En toda* ***distribución aproximadamente simétrica****, (forma de campana), con media aritmética igual a* μ *y desviación estándar igual a* σ*, se verifica que:*

Distancia desde la media		**Valores aproximados dentro de la distancia**
a) $\mu \pm \sigma \Rightarrow [\mu \pm \sigma]$	$\Rightarrow$	**68% de las observaciones**
b) $\mu \pm 2\sigma \Rightarrow [\mu \pm 2\sigma]$	$\Rightarrow$	**95% de las observaciones**
c) $\mu \pm 3\sigma \Rightarrow [\mu \pm 3\sigma]$	$\Rightarrow$	**99% de las observaciones**



Cabe destacar que *habitualmente,* los datos de numerosos fenómenos, *están distribuidos en forma de campana*, como ocurre con la mayoría de las características humanas, tales como la estatura y el peso. Por tal motivo, la ***REGLA EMPÍRICA,*** se aplica en muchas situaciones y se usa ampliamente.

Ejemplo:

Las longitudes de las piezas derivadas de un cierto proceso productivo **A,** se distribuyen *en forma simétrica*, con una **media** de 12 cm, y una **desviación estándar** de 0,4 cm.

1) Entre qué valores de longitud se encuentran aproximadamente el 68% de las piezas?.
2) Entre qué valores de longitud se encuentran aproximadamente el 95% de las piezas?.
3) Una pieza escogida al azar tiene una longitud de 12,6 cm: ¿Pertenece al proceso productivo **A**?.
4) Una pieza escogida al azar tiene una longitud de 10,7 cm: ¿Pertenece al proceso productivo **A**?.
5) Si el peso de las piezas, es en promedio de 230 gramos con una varianza 36 gramos^2, considerando longitud y peso: ¿En qué aspecto es el proceso más estable?. **FUNDAMENTAR.**

Solución:

X = longitud de las piezas **/ X ~** simétrica $\Rightarrow$ $\mu =$ 12 cm; $\sigma = 0{,}4$ cm

$\Downarrow$

1) **68%** $\Rightarrow$ **[$\mu \pm \sigma$]** $\Rightarrow$ [12 **±** 0,4] $\Rightarrow$ **[11,6 ; 12,4]**

$\Downarrow$

"Entre 11,6 cm y 12,4 cm, se encuentran aproximadamente el 68% de las piezas".

2) **95%** $\Rightarrow$ **[$\mu \pm 2\sigma$]** $\Rightarrow$ [12 **±** 2**.** 0,4] $\Rightarrow$ **[11,2 ; 12,8]**

$\Downarrow$

*"**Entre 11,2 cm y 12,8 cm, se encuentran aproximadamente el 95% de las piezas".***

3) X = 12,6 cm ⇒ *"PUEDE pertenecer al proceso A, siendo ALTAMENTE PROBABLE".*

⇓

"Se encuentra comprendido dentro de los valores habituales de las mediciones".

4) X = 10,7 cm ⇒ ?

⇓

99% ⇒ [μ ± 3σ] ⇒ [12 ± 3. 0,4] ⇒ [10,8 ; 13,2]

⇓

*"Entre **10,8 cm y 13,2 cm** se encuentran aproximadamente el **99%** de las piezas".*

⇓

"Un 1% de las piezas tienen longitudes menores a 10,8 cm o mayores a 13,2 cm".

⇓

X = 10,7 cm

⇓

"PUEDE** pertenecer al proceso A, siendo **POCO PROBABLE".

5) Y = peso de las piezas ⇒ μ = 230 gramos; σ^2 = 36 gramos^2

⇓

*¿En qué aspecto es el proceso **más estable**?....⇒***COEFICIENTE DE VARIACIÓN***

⇓

- **Longitud ⇒** $CV(X) = \frac{\sigma_X}{\mu_X} = \frac{0{,}4\text{ cm}}{12\text{ cm}} = \frac{0{,}4}{12} = \mathbf{0{,}033}$

- **Peso ⇒** $CV(Y) = \frac{\sigma_Y}{\mu_Y} = \frac{\sqrt{36\text{ gramos}^2}}{230\text{ gramos}} = \frac{6\text{ gramos}}{230\text{ gramos}} = \frac{6}{230} = \mathbf{0{,}0261}$

⇓

CV(Y) < CV(X) ⇒ *Proceso más estable, en el Peso de las piezas*

II. TEOREMA O DESIGUALDAD DE CHEBYSHEV (Tchebyccheff)

La ***REGLA EMPÍRICA,*** es una regla práctica, que resulta ***aplicable*** cuando los datos, están distribuidos en forma ***aproximadamente simétrica***. En los casos en los que los datos ***no estén*** distribuidos ***en forma simétrica***, o bien la ***forma de la distribución es desconocida,*** resulta aplicable el ***"Teorema o Desigualdad de Chebyshev (o Tchebyccheff)".***

La ***Desigualdad de Chebyshev*** establece que: *"Si un conjunto de datos tiene media aritmética igual a* μ *y desviación estándar igual a* σ*,* ***independientemente de como sea su distribución***, ***por lo menos*** $1 - 1/k^2$ *valores, caerán dentro de:* ***± k desviaciones estándar*** *de la* ***media aritmética****".*

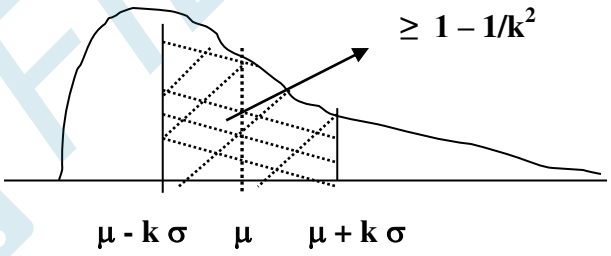
$$\Downarrow$$

"Dentro de k desviaciones estándar de la media aritmética, $\mu \pm k\sigma$, existe **por lo menos** $\left(1 - \frac{1}{k^2}\right)$ *proporción de valores*".

$$\Downarrow$$

$$\{ |x - \mu| \leq k\sigma \} \geq 1 - \frac{1}{k^2} \; ; \quad k > 1$$

$$\Downarrow$$



1) $k = 2 \Rightarrow 1 - \frac{1}{k^2} = 0{,}75 \Rightarrow$ *"Dentro de* $\pm 2\sigma$ *de la media, se encuentran* ***por lo menos*** *el* ***75% de los datos****".*

2) $k = 3 \Rightarrow 1 - \frac{1}{k^2} = 0{,}8889 \Rightarrow$ *"Dentro de* $\pm 3\sigma$ *de la media, se encuentran* ***por lo menos*** *el* ***88,89% de los datos****".*

3) $k = 4 \Rightarrow 1 - \frac{1}{k^2} = 0{,}9375 \Rightarrow$ *"Dentro de* $\pm 4\sigma$ *de la media, se encuentra* ***por lo menos*** *el* ***93,75% de los datos****".*

4) $k = 2{,}5 \Rightarrow 1 - \frac{1}{k^2} = 0{,}84 \Rightarrow$ *"Dentro de* $\pm 2{,}5\sigma$ *de la media, se encuentran* ***por lo menos*** *el* ***84 % de los datos****".*

Observación:

Se afirma que la **"*Desigualdad de Chebyshev*",** al establecer una **cota mínima** de la **proporción de datos**, que quedan a una determinada cantidad de desviaciones estándar de la media aritmética, se trata de un **"enfoque conservador".**

Ejemplo:

En una cierta rama de la industria dedicada a la electrónica, la edad promedio de los empleados profesionales tiende a ser menor que en otras actividades. Un estudio realizado determinó que el promedio de las edades de los empleados profesionales es de 36 años con una desviación estándar de 5 años, no existiendo un comportamiento definido en la distribución de las edades. Aplicar la ***Desigualdad de Chebyshev*** para determinar dentro de qué rango de edades se encuentran al menos el 85% de las edades de los trabajadores.

Solución:

x = edades de los empleados / $\mathbf{x} \sim ? \Rightarrow$ $\mu = 36$ años; $\sigma = 5$ años

Dado que se desconoce cómo es la distribución de las edades, no podemos aplicar la ***"Regla Empírica"***, debiendo aplicar la ***"Desigualdad de Chebyshev"***.

Sabemos que ***Chebyshev*** establece que:

$$\{ |x - \mu| \leq k\sigma \} \geq 1 - \frac{1}{k^2} \quad ; \quad k > 1$$

$$\Downarrow$$

Proporción de valores

Como el problema es ***determinar el rango,*** dentro del cual se encuentran ***"por lo menos el 85% de las edades":***

$$\Downarrow$$

$$1 - \frac{1}{k^2} = 0{,}85 \Rightarrow k = 2{,}58$$

$$\Downarrow$$

"Al <u>menos</u>, el 85% de las edades están comprendidas dentro de ± 2,58σ de la media aritmética"

$$\Downarrow$$

Siendo: $\mu = 36$, $\sigma = 5$ $\Rightarrow$ $\{ |x - 36| \leq 2{,}58 \cdot 5 \} \geq 0{,}85$

$$\Downarrow$$

$$x \Rightarrow 36 \pm 12{,}9$$

$$\Downarrow$$

$$[23{,}1 \; ; \; 48{,}9]$$

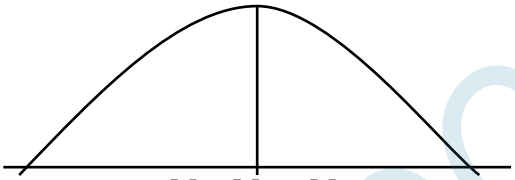
$$\Downarrow$$

"<u>Al menos el 85%</u> de las edades de los empleados, <u>se encuentran</u> entre 23,1 y 48,9 años".

3) <u>MEDIDAS DE FORMA, ASIMETRÍA O SESGO</u>

Las ***medidas de asimetría*** *tienen como objetivo "Elaborar un* ***indicador*** *que permita establecer el grado de simetría o asimetría que presenta una distribución, sin necesidad de llevar a cabo su representación gráfica".* Son medidas que permiten ***describir la forma de una distribución de datos.*** Se presentan tres situaciones:

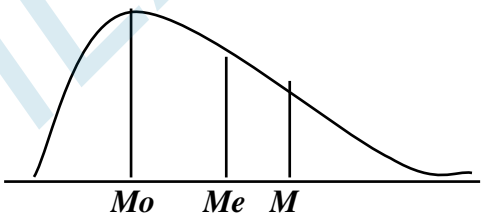
1) *Distribución simétrica:*



En toda ***distribución simétrica,*** la mitad de los datos se encuentra localizada a la izquierda de la media aritmética y la otra mitad a la derecha, verificándose que:

Distribución simétrica $\Rightarrow M = Me = Mo$

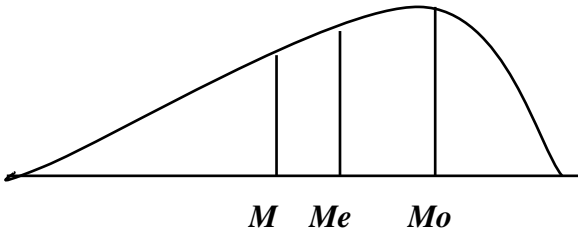
2) *Distribución asimétrica positiva, lateral derecha o sesgada positiva:*



En toda ***distribución asimétrica positiva o sesgada hacia la derecha,*** la parte alargada de la curva se encuentra ubicada hacia la derecha, indicando ello que, para **valores altos** de la variable, el número de observaciones es **bajo,** verificándose que:

Distribución asimétrica positiva $\Rightarrow Mo < Me < M$

3) ***Distribución asimétrica negativa, lateral izquierda o sesgada negativa:***



En toda distribución ***asimétrica negativa o sesgada hacia la izquierda***, la parte alargada de la curva se encuentra ubicada hacia la izquierda, indicando ello que para **valores bajos** de la variable, el número de observaciones es **bajo**, verificándose que:

Distribución asimétrica negativa $\Rightarrow$ $M < Me < Mo$

Determinación del sesgo o asimetría de una distribución

Para determinar la forma de una distribución de frecuencias pueden aplicarse los indicadores:

Coeficiente de Asimetría de Karl Pearson (S_k):

$$\Downarrow$$

$$S_k = \frac{3(M - Me)}{\sigma} \Rightarrow S_k \begin{cases} < 0 \Rightarrow \textit{Asimetría o Sesgo Negativo} \\ = 0 \Rightarrow \textit{Simétrica} \\ > 0 \Rightarrow \textit{Asimetría o Sesgo Positivo} \end{cases}$$

Ejemplo:

Una distribución, tiene una ***Media Aritmética*** igual a 40, una ***Mediana*** igual a 36 y una desviación estándar igual a 17. ¿Cómo es la **distribución,** en cuanto a su forma?.

Solución:

Aplicando: $\mathbf{S_k = \frac{3(M - Me)}{\sigma} \Rightarrow S_k = \frac{3(40 - 36)}{17} = +\underline{\underline{0,71}} > 0 \Rightarrow}$ ***Asimetría Positiva***

Observación:

Se **demuestra** que en toda *distribución levemente asimétrica* se verifica que:

$$\mathbf{Mo = M - 3\,(M - Me)}$$

Que es lo que se conoce con el nombre de ***"Método Empírico para el cálculo del Modo.***

4) MEDIDAS DE PUNTIAGUDEZ O CURTOSIS

Las ***medidas de Curtosis, apuntamiento o concentración central,*** estudian la distribución de frecuencias en la zona central de la misma, o también del valor modal, de forma tal que, una **mayor** o **menor** concentración de observaciones, dará lugar a una distribución **más** o **menos** puntiaguda.

Las ***medidas de Curtosis,*** se aplican a *distribuciones en forma de campana*, o sea unimodales y simétricas o con leve asimetría. Es por ello que, para su estudio, es necesario previamente definir una ***distribución tipo,*** que se toma como modelo de referencia, siendo ésta la ***distribución normal,*** que se corresponde con los datos de fenómenos muy corrientes en la naturaleza, siendo su representación gráfica la denominada ***Curva Normal o Campana de Gauss.***

Tomando la ***distribución normal como referencia***, pueden presentarse en lo que a ***Curtosis*** se refiere, tres situaciones:

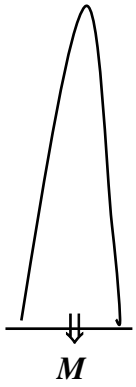
1) ***Leptocúrtica*** $\Rightarrow$ *alta concentración* $\Rightarrow$ *más puntiaguda que la normal.*

2) ***Mesocúrtica*** $\Rightarrow$ *concentración normal* $\Rightarrow$ *distribución normal.*

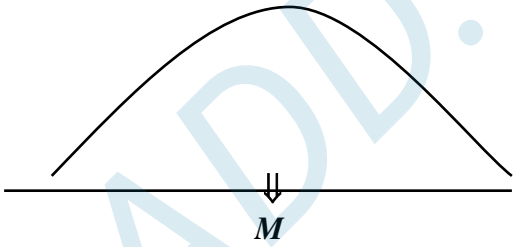
3) ***Platicúrtica*** $\Rightarrow$ *baja concentración* $\Rightarrow$ *más aplanada que la normal.*

Gráficamente:

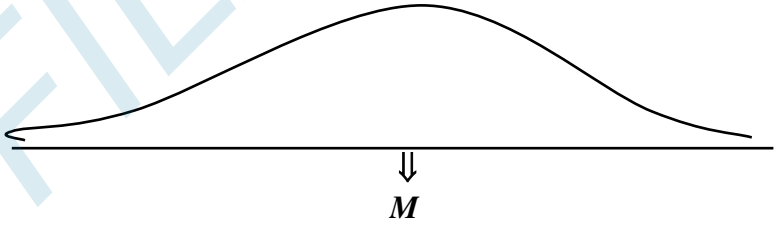
1) ***Leptocúrtica:***



2) ***Mesocúrtica o Normal:***



3) ***Platicúrtica:***



DISTRIBUCIONES BIDIMENSIONALES

Se denomina de esta forma, *"Al conjunto de **pares de valores,** resultantes de estudiar **simultáneamente**, de cada una de las unidades estadísticas de una determinada población estadística, **dos variables**".*

SERIE SIMPLE

De la misma forma que en el caso de ***distribuciones unidimensionales,*** en el caso de una ***distribución bidimensional,*** una ***serie simple*** va a estar dada por *el conjunto de pares de valores de las variables estudiadas,* obtenidos en *bruto.*

Se denota, en *forma genérica* a cada variable estudiada, de la forma:

$$\underline{x_{1i}} \text{ y } \underline{x_{2i}} \text{ ; } i = 1, 2, \ldots .. n$$

Ejemplo:

Dados los siguientes datos referidos a horas trabajadas y número de artículos producidos en 10 máquinas de una empresa:

Máquinas	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Horas trabajadas (x_1)	6	6	6	7	7	7	8	8	8	8
Artículos producidos (x_2)	70	70	70	80	80	90	90	90	100	100

La representación gráfica de los datos de una *Serie Simple Bidimensional,* está dada por el ***"diagrama de dispersión o nube de puntos"***, considerando en cada uno de los ejes de un sistema de coordenadas cartesianas ortogonales, a las variables dadas y representando en el plano a los puntos asociados a los pares de valores observados.

COVARIANZA

Es una **medida de dispersión**, representativa del ***grado de variación conjunta entre dos variables***, y se define como ***"La media aritmética del producto de las variables desvíos respecto a la media aritmética, de cada una de las variables dadas".***

Dadas dos variables **x** e **y**, la **Covarianza,** por definición, va a estar dada por:

$$\mathbf{Cov(x\ ,\ y) = M(Z_x\ .\ Z_y) = M[(x - \bar{x})\ (y - \bar{y})] = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{n}}$$

SIGNO DE LA COVARIANZA:

La **Covarianza** asume **valores,** en el intervalo: $\mathbf{-\infty < cov\ (x\ ,\ y) < \infty}$, con el siguiente significado:

Cov(x,y) →
- **< 0 ⇒ relación inversa o negativa → Δ x ⇒ ∇ y**
- **= 0 ⇒ variables independientes o incorrelacionadas.**
- **> 0 ⇒ relación directa o positiva → Δ x ⇒ Δ y**

FÓRMULA DE CÁLCULO DE LA COVARIANZA:

Por *definición*, sabemos que la **Covarianza** es igual a:

Cov(x,y) = $M[Z_x\ .\ Z_y] = M[(x - \bar{x}).(y - \bar{y})]$. Efectuando el producto de los binomios:

Cov(x,y) = $M[\ x\ y\ -\ \bar{x}\ y\ -\ x\ \bar{y}\ +\ \bar{x}\ \bar{y}\ ]$. Aplicando propiedad de media aritmética

Cov(x,y) = $M\ (x\ y\ ) -\ \bar{x}\ M(y) -\ \bar{y}\ M(x) +\ \bar{x}\ \bar{y}$; $\bar{x} = M(x)$ e $\bar{y} = M(y)$

Cov(x,y) = $M\ (x\ y\ ) -\ \bar{x}\ \bar{y} - \cancel{\bar{y}\ \bar{x}} + \cancel{\bar{x}\ \bar{y}}$

$$\mathbf{Cov(x,y) = M\ (x\ y\ ) - M(x)\ .\ M(y)\ \Rightarrow\ Cov\ (x,y) = \frac{\sum xy}{n} - \frac{\sum x}{n}\frac{\sum y}{n}}$$

VARIANZA DE UNA SUMA Y DIFERENCIA DE DOS VARIABLES

Sabemos que en general, *"La varianza de una suma o diferencia de dos variables* ***NO ES IGUAL*** *a la suma o diferencia de las varianzas de cada una de ellas".* Por lo tanto, se tratará de encontrar:

1) **V(x + y) = ?**
2) **V(x - y) = ?**

1) V(x + y) = ?

Aplicando la **definición de Varianza** y considerando a la **variable (x + y)**, resulta:

V(x + y) = M $[(x + y) - M(x + y)]^2$. Aplicando propiedad de media aritmética:

V(x + y) = M $[x + y - M(x) - M(y)]^2$; $\bar{x} = M(x)$ e $\bar{y} = M(y)$

V(x + y) = M $[x + y - \bar{x} - \bar{y}]^2$; asociando convenientemente en la base de la potencia:

V(x + y) = M$[(x - \bar{x}) + (y - \bar{y})]^2$; desarrollando el cuadrado del binomio:

V(x + y) = M$[(x - \bar{x})^2 + 2(x - \bar{x})(y - \bar{y}) + (y - \bar{y})^2]$. Tomando media aritmética:

V(x + y) = $M(x - \bar{x})^2 + M(y - \bar{y})^2 + 2\,M[(x - \bar{x})(y - \bar{y})]$

$\Downarrow$ $\Downarrow$ $\Downarrow$

V(x) **V(y)** **Cov(x,y)**

Por lo tanto:

V(x + y) = V(x) + V(y) + 2 Cov(x,y) ⇔ x e y variables cualesquiera

$\Downarrow$

"La varianza de una suma de dos variables es igual a la <u>suma</u> de las varianzas de cada una de las variables dadas <u>más</u> dos veces la Covarianza entre las mismas"

Si ***x e y son variables independientes***, sabemos que **Cov(x,y) = 0**, resultando:

V(x + y) = V(x) + V(y) ⇔ x e y variables independientes (I)

⇓

"La varianza de una <u>suma</u> de dos <u>variable independientes</u> es igual a la <u>suma</u> de las <u>varianzas</u> de cada una de las variables dadas".

2) V(x - y) = ?

Aplicando la **definición de Varianza** y considerando a la **variable (x - y)**, resulta:

V(x - y) = M $[(x - y) - M(x - y)]^2$. Aplicando propiedad de media aritmética:

V(x - y) = M $[x - y - M(x) + M(y)]^2$; $\bar{x} = M(x)$ e $\bar{y} = M(y)$

V(x - y) = M $[x - y - \bar{x} + \bar{y}]^2$; asociando convenientemente en la base de la potencia:

V(x - y) = M$[(x - \bar{x}) - (y - \bar{y})]^2$; desarrollando el cuadrado del binomio:

V(x - y) = M$[(x - \bar{x})^2 - 2(x - \bar{x})(y - \bar{y}) + (y - \bar{y})^2]$. Tomando media aritmética:

V(x - y) = $M(x - \bar{x})^2 + M(y - \bar{y})^2 - 2\,M[(x - \bar{x})(y - \bar{y})]$

⇓ ⇓ ⇓

V(x) **V(y)** **Cov(x,y)**

Por lo tanto:

V(x - y) = V(x) + V(y) - 2 Cov(x,y) ⇔ x e y variables cualesquiera

⇓

"La varianza de una diferencia de dos variables es igual a la <u>suma</u> de las varianzas de cada una de las variables dadas <u>menos</u> dos veces la Covarianza entre las mismas"

Si ***x e y son <u>variables independientes</u>***, sabemos que **Cov(x,y) = 0**, resultando:

V(x - y) = V(x) + V(y) ⇔ x e y variables independientes (II)

"La varianza de una <u>diferencia</u> de dos <u>variable independientes</u> es igual a la <u>suma</u> de <u>las varianzas</u> de cada una de las variables dadas"

OBSERVACIÓN

Relacionando lo obtenido en **(I)** y **(II)**, se concluye en que: "**La Varianza de una Suma o Diferencia de variables independientes**, es igual a la **Suma** de las varianzas, de cada una de las variables dadas.

$$\Downarrow$$

Si x e y son variables independientes $\Rightarrow$ V(x ± y) = V(x) + V(y)