Liceo Parroquial San Antonio

APUNTE DE ESTADISTICA. CUARTOS E. M Montoya.-

QUE ESTUDIA LA ESTADISTICA.

ANTES DE RESPONDER A ESTA PREGUNTA VEREMOS ALGUNAS CONSIDERACIONES.

EL CUARTO AÑO DEL Liceo Parroquial San Antonio REALIZARÁ UNA CONVIVENCIA PARA FINALIZAR EL AÑO ESCOLAR Y SE LE PIDE A UD. QUE COMPRE EL QUESO QUE CONSUMIRÀN. UD. CONCURRE A UN SUPERMERCADO Y EL VENDEDOR LE OFRECE TRES MARCAS DIFERENTES DE ESTE PRODUCTO. COMO NO SABE POR CUAL DECIDIRSE LE PIDE AL VENDEDOR QUE LE DE UNA "PROBADITA" DE CADA UNA DE LAS MARCAS, PARA DECIDIR CUAL COMPRAR. EN OTRAS PALABRAS UD. TOMA UNA "MUESTRA" DE CADA UNA DE LAS MARCAS Y LUEGO JUZGA EL COMPORTAMIENTO DEL TROZO COMPLETO DE QUESO, BAJO EL RAZONABLE SUPUESTO QUE LA "PROBADITA" TIENE EL MISMO SABOR QUE EL TROZO COMPLETO.

ESTE PRINCIPIO OPERA EN ESTADÍSTICA. LA ESTADÍSTICA EMPLEA DATOS OBTENIDOS DE MUESTRAS PARA SACAR CONCLUSIONES DE GRUPOS TOTALES O POBLACIONES.

POBLACIÓN: COLECCIÒN COMPLETA DE GENTE, objetos o, ítems en los cuales se tiene un interés de estudio.

Una vez elegida una muestra representativa" de una población se pueden sacar conclusiones acerca de las características de toda la población en función de datos muestrales, lo que se denomina "inferencia estadística".

En estadística, es necesario ajustar los datos obtenidos de una muestra a modelos teóricos matemáticos; en este sentido la teoría de las probabilidades nos entrega las herramientas necesarias.

Las muestras se toman entonces con el fin de descubrir " algunas" características de la población de la cual se tomó la muestra

También podemos tomar una muestra con el propósito de "descubrir" características de la muestra, sin hacer inferencias de la población representada

Por ejemplo si consideramos a todos los alumnos del cuarto medio del Liceo Parroquial San Antonio, podemos tomar una muestra y describir solamente cuantos de ellos fuman, o la proporción de estos cuyos padres son fumadores, etc. Tales rasgos son referidos como características descriptivas y los métodos empleados para obtenerlos se conoce con el nombre de "estadística descriptiva".

Objetivos de la estadística descriptiva.

*obtener una visión amplia de la distribución de los datos obtenidos de la muestra, identificando cualquier característica que pudiera estar presente. Usualmente los datos se organizan en tablas, gráficos o diagramas, Valores numéricos que resumen información

También pueden ser resumidos en proporción de porcentajes

*determinar una medida numérica que resuma de alguna manera el "centro" de los datos (ejemplo promedio). A menudo hay más de una medida de tendencia central, que permitirá una mejor interpretación de los datos.

Importante: si utilizamos como medida de tendencia central solo el promedio hacemos una drástica reducción de los datos, y nucas veces contiene escasa información sobre el verdadero comportamiento de estos.

*determinar una medida numérica que resuma el grado de dispersión entre los valores de la muestra.

Por ejemplo, el 4ª B y C del liceo parroquial san Antonio obtuvieron las siguientes calificaciones parciales cuyos promedios por curso se indican en la siguiente tabla:

4°.C.	5.5	4.5	5.2	4.8	$\overline{X} = 5.0$
4° B.	2.2	6.8	4.0	7.0	$\overline{X} = 5.0$

VEMOS QUE AMBOS CURSOS TIENEN EL MISMO PROMEDIO 5.0, PERO LAS NOTAS DE 4ª B, están más disparejas "que las del 4a.C. las medidas que resumen este tipo de variabilidad se conocen con el nombre de "medidas de dispersión"

Entonces, con esto hemos sentado los conceptos previos de lo que se conoce como estadística descriptiva, que en lenguaje más moderno es conocido como análisis exploratorio de datos.

Si además quisiéramos llegar mas allá del solo propósito de describir las características de la muestra, podríamos usar el análisis descriptivo para buscar un modelo o patrón que represente a la población y luego obtener conclusiones acerca de esta.

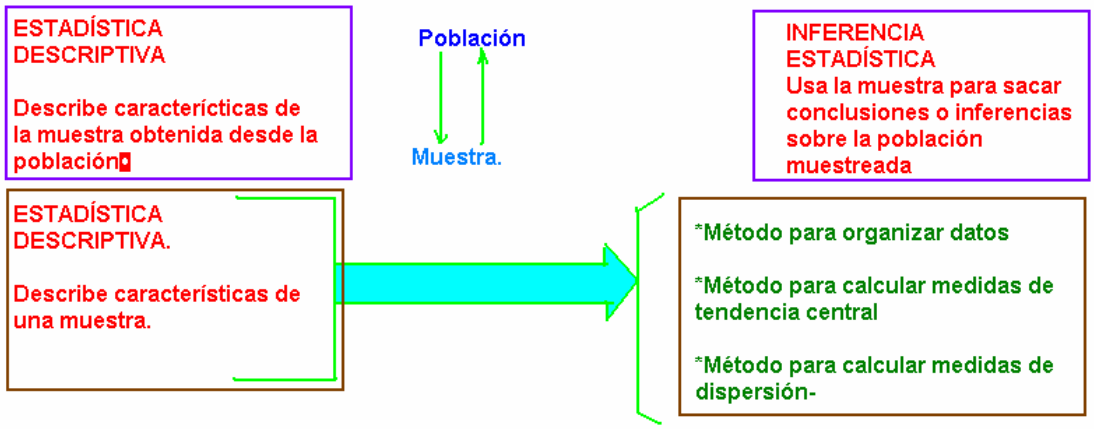
Por ejemplo , si el promedio de edad de los fumadores en una muestra es de 27 años , podríamos concluir que el promedio de edad de todos los fumadores en chile es también alrededor de 27 años , con mas o menos 2 años de diferencia, o sea 27 ± 2 años.

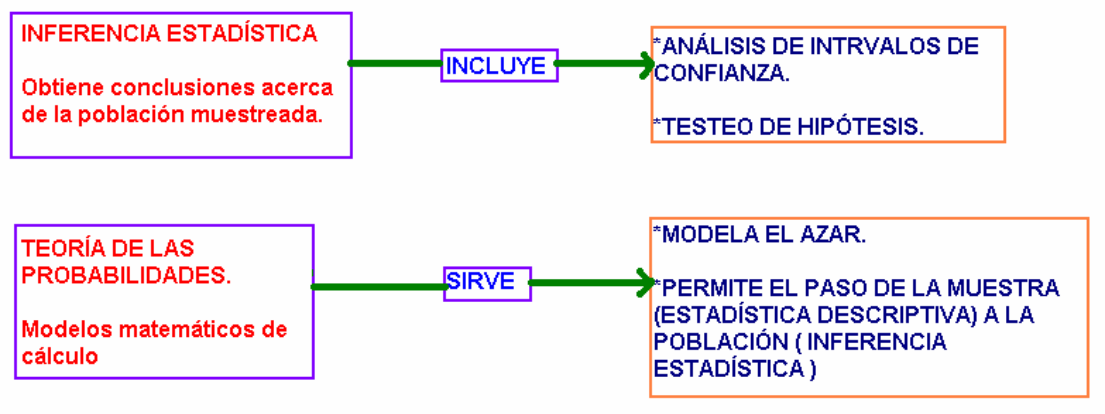
En la inferencia estadística el proceso que se usa para lograr aproximaciones sobre la probable preedición que nos entregan los datos de la muestra, como en el ejemplo recién mencionado, es llamado "estimación por medios de intervalos de confianza"

Alternativamente podemos usar los datos para testear hipótesis, es decir hacer algunas inferencias sobre la población desde la cual fue tomada la muestra. Los modelos matemáticos que permiten el estudio de patrones de comportamiento y que en particular sirven para hacer inferencias sobre una población (ejemplo, construir intervalos de confianza y testeo de hipótesis), es la teoría de las probabilidades.

En particular el estudio formal de la inferencia estadística, en el caso de intervalos de confianza y de testeo de hipótesis escapa a los objetivos de este curso, pero trataremos algunos aspectos básicos como herramientas que nos servirán a futuro.

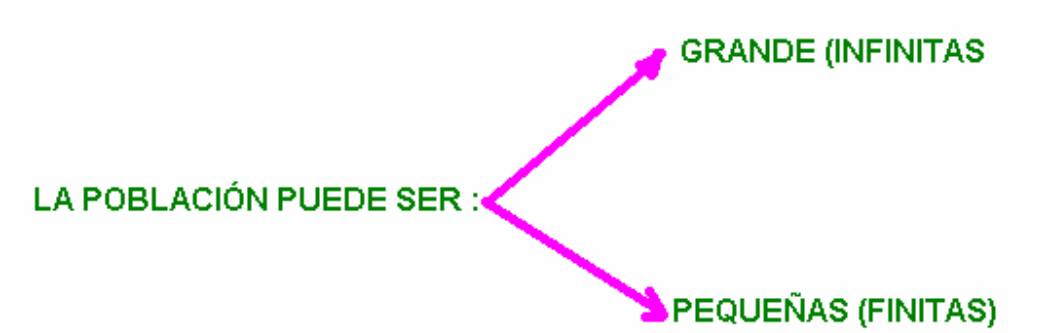
Resumen de lo expuesto.





TEMA 1: ANÁLISIS EXPLORATORIO DE DATOS.

MUESTRAS:



Suponga que queremos estudiar cuantos ratones infectados con Hanta virus tenemos en chile. ¿Es posible?
Lo razonable es estudiar solo una muestra en lugar de toda la población completa de ratones (¡no me gustaría verlo a la caza de todos los ratones, con el propósito de estudiarlos estadísticamente!)
Así, entonces podemos establecer:

RAZONES PARA ESTUDIAR MUESTRAS EN LUGAR DE POBLACIONES.

1.- si la población es muy grande. (Incluso de tamaño finito). Por cuestiones de tiempo, recursos etc..., no es posible estudiar a cada individuo que compone la población.

2.- razones éticas o de seguridad. Los medicamentos se usan primero en ratones de laboratorio, e incluso con personas voluntarias primero, antes de masificar su uso.

3.-a veces los miembros de una población son difíciles de identificar. Ejemplo: estudio de un virus que afecte a la humanidad, el cual no produce síntomas, sino hasta muy avanzada la enfermedad.

4.-estudio de tipo destructivo. Ejem.: estudio de durabilidad de un artículo, el que es sometido a uso permanente y a presiones, para medir resistividad (productos enlatados).

Es importante tener en cuenta que malas elecciones de una muestra corren el riesgo de ser sesgadas.

Ejemplo: Si estuviésemos investigando los rasgos físicos de los alumnos de 4° E.M, y eligiéramos para tal efecto una muestra de alumnos solo del liceo san Antonio, las conclusiones que saquemos de ello (físicas, intelectuales, socioculturales, etc.), no serán necesariamente validas para todos los alumnos de cuarto año medio de todo el país.

Al elegir una muestra de una población debemos esperar que sea lo más representativa de la población, es decir que la muestra refleje todas las características de la población, mientras ello ocurra mayor será la confiabilidad de cualquier conclusión que se haga al respecto de la población, basada en la muestra.

En los estudios de opinión, política, de mercado, etc, si esta se hace en función de un determinado sector socioeconómico, no tiene un valor estadístico real y objetivo. Lo mismo ocurre cuando se hace a través de un solo medio, por ejemplo el telefónico o Internet, pues la población que no cuenta con ellos no tiene representación en la muestra.

¿Cómo SE ELIGE UNA MUESTRA REPRESENTATIVA DE UN POBLACIÓN?
¿CUÁL ES EL TAMAÑO QUE DEBE TENER LA MUESTRA?

La primera pregunta resulta más fácil de responder. Se sugiere un método que incluya el azar, de modo que todos los individuos (elementos de la población), tengan igual chanca de ser elegido, este método se denomina: **MUESTREO ALEATORIO.**

En general, se sugiere aplicar los siguientes modelos matemáticos para la determinación de la muestra.

1.- elegir proporcionalmente la muestra de acuerdo a como está constituida la población.

Ejemplo: si el total de alumnos del liceo san Antonio es de 580 alumnos, de los cuales 380 son hombres. Y si se quiere elegir una muestra de 40 alumnos para indeterminado estudio estadístico, la proporción de mujeres y hombres en la muestra se debe mantener.

En este caso:

Si H representa el número de hombres de la muestra, y M, el número de mujeres de la misma:

$$\frac{H}{M}=\frac{380}{200} \quad ; \frac{H+M}{H}=\frac{58}{20} \quad ; \quad \frac{40}{M}=\frac{58}{20} \quad ; \text{ M=14 y H= 26}$$

En segundo lugar la muestra debe ser absolutamente aleatoria y para ellos se pueden usar diversos métodos. RULETAS, TÓMBOLAS CON BOLAS NUMERADAS,

EL MAS RECOMENDADO ES LA ELECCION POR EL MÉTODO DE MONTECARLO (TABLA DE NUMEROS ALEATORIOS DE LA RULETA DECIMAL DE LAPLACE)

Observación: Si hacemos un muestreo al azar, no tendremos garantía de que la muestra contenga la misma proporción de mujeres y hombres que hay en la población .Si queremos asegurar que los estratos en que hemos subdividido a la población estén presentes en la muestra en la misma proporción como lo están en la población, entonces tomamos una muestra aleatoria de cada estrato, con tamaño proporcional al que tiene el estrato en la población (Muestreo aleatorio Estratificado).

En el ejemplo anterior, la proporción del estrato hombres en la población es del 65,5% y la del estrato mujeres de 34,5% (alumnos del liceo san Antonio)

VARIABLES. (Tipo de datos)

Ya hemos establecido los criterios para analizar datos muéstrales (datos obtenidos de una muestra), que se pueden resumir en:

Estadística descriptiva	Inferencia Estadística
Describe las características principales	Utiliza la interpretación de los da-
De una muestra.	Tos para investigar la población.

Otras consideraciones.

Tipos de escala.

Toda variable estadística puede ser clasificada en uno de los niveles o escalas que se indican a continuación.

***escala nominal**: etiquetas simples, solo permite identificar el objeto en estudio.
Ejemplo, el código de barras, Rut, patente.

***ordinal**: mediciones en el que existe un orden implícito. Admite grados de calidad.

***Intervala**r: considera no solo la información, permite el orden, permite también cuantificar las diferencias entres los individuos que pertenecen a clases o categorías distintas.

En esta escala no existe el cero absoluto"

***de razón**: considera todas las cualidades de la escala anterior, pero si existe el cero absoluto. Ejem. Escalas de medición (Km., Pts.

Algunos ejemplos de los tipos de escala:

*sexo ------nominal
Edad---------razón
Carrera--------nominal
Estrato social------ordinal
Puntaje de ingreso ------razón
Año ingreso ----------nominal
Numero de aprobados -------razón
Región de procedencia--------nominal.

Técnicas de muestreo.

Las razonas que nos obligan a tomar una muestra son: tiempo, costos, pruebas destructivas, poblaciones muy grandes, etc.

Existen diferentes métodos de muestreo, asociadas a las características que presenta la población, lo cual permitirá tener las herramientas necesarias par lograr una muestra que sea imagen de la población en estudio.

Muestra aleatoria simple: En este caso cada uno de los elementos de la población tiene la posibilidad (probabilidad) de ser seleccionado: Para aplicar esta técnica de muestreo, se sugiere que los elementos de la población presenten cierto grado de homogeneidad, es decir que no existan diferencias notables entre cada uno de los elementos.

En la practica al utilizar una muestra aleatoria simple debemos enumerar las unidades de la población del 1 al N y a continuación debemos seleccionar "n" números aleatorios entre 1 y N, utilizando una tabla de números aleatorios, la calculadora, el computador, etc.

Ejemplo: Cuantas muestras de tamaño 6 se pueden obtener en un curso de 30 alumnos.

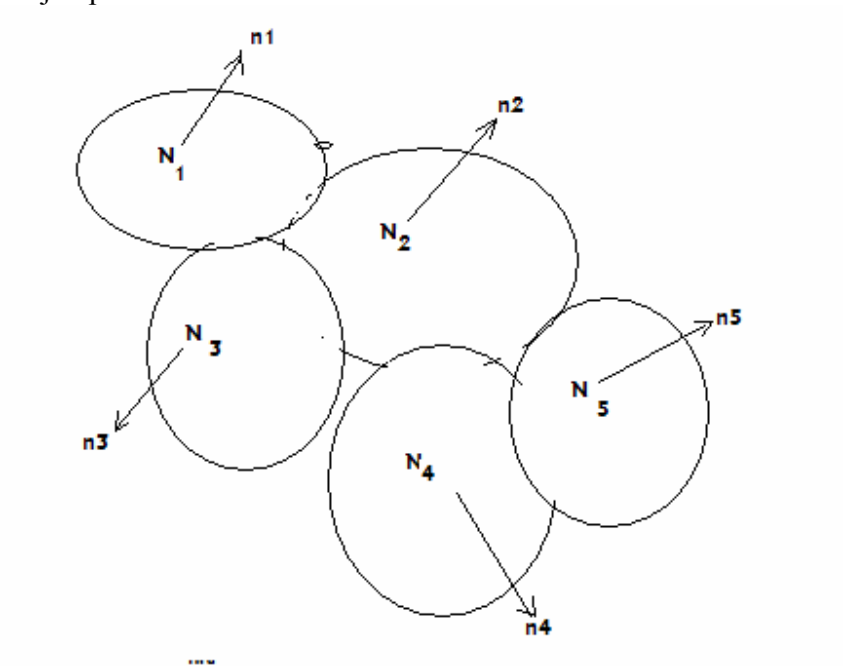
$$C_6^{30} = \frac{30!}{6!*24!}.$$

Muestra estratificada: (referido a grupos o estratos).

Esta técnica de muestreo se utiliza cuando las unidades de la población se encuentran agrupados en estratos (preferentemente pocos), cada uno de ellos con muchos elementos.

Una característica importante de este tipo de muestra es que la población presenta cierto grado de heterogeneidad.

Ejemplos: clases sociales

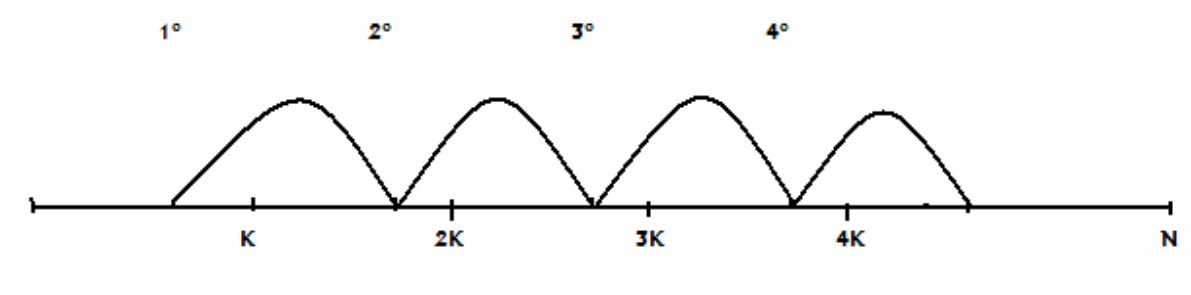


$N = \sum_{i=1}^{K} Ni$ $n = \sum_{i=1}^{k} ni$ ni debe ser proporcional al tamaño del estrato.

En este tipo de muestreo se debe seleccionar una muestra aleatoria de cada uno de los estratos.

Muestra sistemática:

Esta técnica de muestreo se utiliza cuando las unidades de la población están de algún modo totalmente ordenadas. Para seleccionar una muestra se utiliza tal ordenación dividiendo la población (de tamaño N) en "n" subpoblaciones de tamaño K=N/n, a continuación se debe seleccionar al azar una unidad de las "K" primeras (denominada no) y en adelante se selecciona cada K esima unidad.



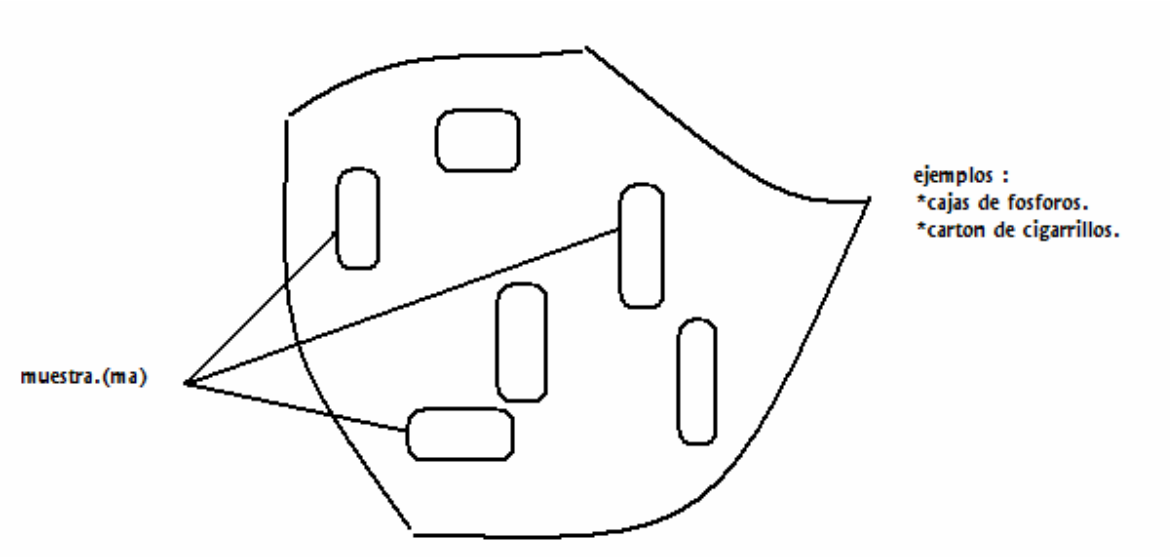
$\therefore m_a = \{n_{oo}, n_o + k, n_o + 2k \ldots\ldots\ldots\ldots\ldots\ldots, n_o + (n-1)k\},$

4.- Muestras por conglomerados.

Esta técnica de muestreo se utiliza cuando los elementos de la población se encuentran en grupos o conglomerados (generalmente muchos) cada uno de ellos con pocos elementos.

Los elementos de cada grupo presentan cierto grado de homogeneidad.

Esta técnica de selección consiste en elegir aleatoriamente grupos y posteriormente censar cada uno de los grupos seleccionados.



Ejercicios: determine el tipo de muestreo adecuado en los siguientes casos:

1.- Chilquita, para efectos de nuevas instalaciones se interesa en conocer el consumo de energía eléctrica de las familias de la quinta región.

2.- Un importador de circuitos se interesa en determinar la proporción estimada, de los defectuosos en una partida de 15000 de 20 unidades cada una.

3 De las 95 declaraciones de impuesto a la renta numerados de 1 al 95, se requiere una muestra aleatoria de tamaño 10 para efectos de investigación.

4.-se requiere un estudio que incluye el gasto en transporte diario de los estudiantes del Liceo Parroquial San Antonio.

5.- se requiere un estudio que entregue información sobre la opinión de los estudiantes del Liceo Parroquial San Antonio sobre la estructura del colegio.

6.- en el año 1997 en cierta región del país se instalo una central termoeléctrica, la cual posee dos turbinas A y B, que han tenido que ser reparadas con repuestos de origen japonés. Se requiere un estudio sobre costos (en dólares) de los repuestos en los dos últimos años.

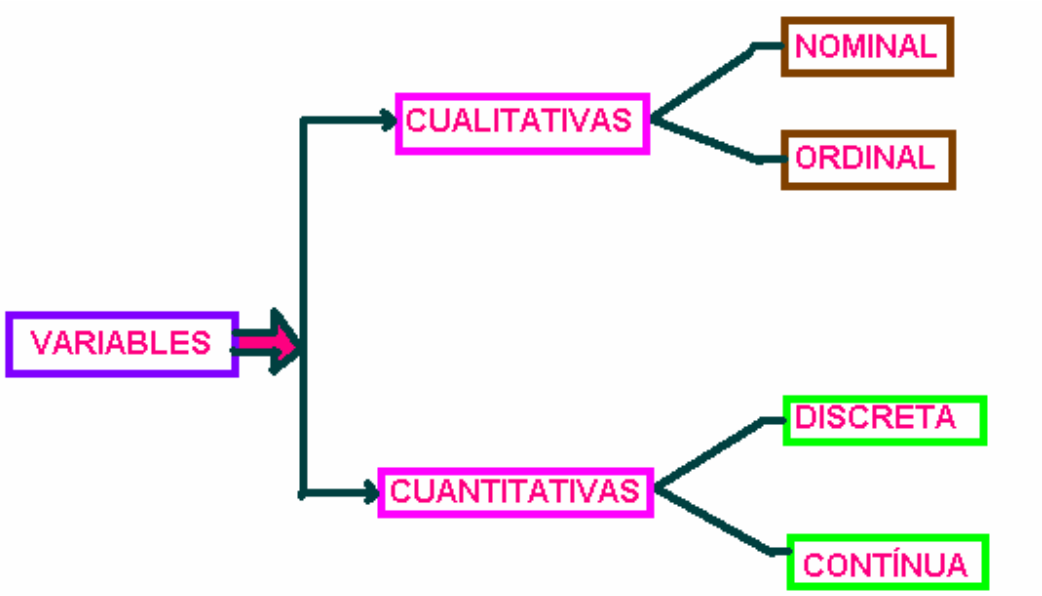
Respuestas: 1.- Estratificado. 2.- Conglomerado. 3.- Aleatoria simple o sistemática 4.- Estratificada.
5.- Estratificada o aleatoria simple.) 6.- Estratificada.

CONCEPTOS PREVIOS.

VARIABLE.
DEFINICION INFORMAL: Parámetro cuyo valor puede variar de persona a persona, de ítems a ítems, de momento a momento.Ejemplos:
*la estatura varia de persona a persona.
*el numero de camas disponibles en un hospital puede asumir los valores: 1, 2, 3 ,4…….
*el grupo sanguíneo: A, AB, B o O
*La calificación de conducta: muy bueno, bueno, satisfactorio. Deficiente
LAS VARIABLES ESTADISTICAS SE REPRESENTAN CON LAS ULTIMAS LETRAS DE ALFABETO: x, y, z

DEFINICION FORMAL DE VARIABLE ESTADISTICA: Nombre que se le asigna a cualquier parámetro que puede asumir diferentes valores o atributos. Se presentan de dos tipos:
CUALITATIVAS
CUANTITATIVAS

Es muy importante establecer con que tipo de variable tratamos, por que los métodos que podemos utilizar en la interpretación o en cualquier análisis descriptivo o inferencial dependen del tipo de variable en estudio.
Diferencias entre el tipo de variables, significa en el fondo diferencias entre el tipo de datos.



VARIABLE CUALITATIVA NOMINAL (DATOS NOMINALES).

- Consisten en etiquetas simples o categorías tales como: masculino/femenino. , si /no, casado/ soltero, sangre del tipo A, B AB.

Las categorías o etiquetas no tienen un orden inherente, da lo mismo escribir por ejemplo masculino/femenino que femenino /masculino.

En otras palabras, el orden de las categorías es completamente arbitrario

Las categorías que se usan con este tipo de variables son invariablemente ALFABETICAS O NO NUMÉRICAS, en consecuencia en ellas no podemos usar las reglas de la aritmética. (Con ellas no podemos hacer operaciones o cálculos)
Piense UD. Si tiene sentido calcular el promedio entre hombres y mujeres!

Ejemplo hipotético: Se encuesto a una muestra de alumnos del Liceo Parroquial San Antonio, cuyos datos se recogen en la siguiente tabla:

Razones para estudiar en el Liceo	Variable Cualitativa Nominal					
	Hombres		Mujeres		Total	
	SI	NO	SI	NO	SI	NO
¿Le agrada el Liceo?	15	2	12	3	27	5
Está de acuerdo con las normas del Liceo.	8	7	12	10	20	17
Encuentra el Liceo muy exigente	10	22	18	24	28	46
Valora Ud. el aspecto formativo del Liceo.	18	4	15	1	108	73
TOTALES	86		95		181	

VARIABLE CUALITATIVA ORDINAL (DATOS ORDINALES)

Constan de categorías tales como: satisfecho, insatisfecho, muy de acuerdo, de acuerdo, en desacuerdo.
Las categorías tienen un orden natural o inherente.

Los datos ordinales tienen un orden natural pero, no se pueden cuantificar (medir con precisión) las diferencias entre las categorías.

En consecuencia al igual que para las variables nominales, no se pueden aplicar las reglas de la aritmética a estos datos.
¿Qué sentido tendría calcular el promedio entre las categorías "de acuerdo y desacuerdo"?.
Ejemplo: Si de la tabla anterior tomamos la variable. ¿Le agrada el Liceo? , podemos hacerla cualitativa ordinal del siguiente modo:

	Muy poco	poco	medianamente	Mucho
¿le agrada el Liceo(Hombres)	4	6	18	60

En este caso podemos decir que lo que se mide es "el grado de satisfacción" del alumno en el liceo, y este puede asumir cualquiera de los cuatro valores que muestra la tabla.
La "lectura" de datos ordinales se debe hacer con cuidado, pues sugieren juicios subjetivos que no han sido medidos con instrumentos científicos, pues son más o menos apreciaciones, y que incluso pueden cambiar en cualquier momento.

El dato arrojado o medido dependerá de muchas consideraciones: la experiencia del encuestado, la valoración positiva o negativa que tenga del dato medido, del estado de ánimo del encuestado etc.

Si a Vd., Se le aplica esta encuesta. ¿Cuál seria su respuesta si ese día Ud. Ha tenido una experiencia negativa en el Liceo?

Tampoco podemos sacar conclusiones como "el nivel de agrado en el Liceo es el triple de los que no están a gusto en el"

¿Cómo valoro yo si el liceo me agrada poco, muy poco? ¿Que criterio aplico para establecer esta diferencia?

Las variables vistas así, tienen como valores categorías (orden), las cuales son alfabéticas o no numéricas. Pero también es común que tomen forma numérica

Ejemplo de esta escala de clasificación:

Volvamos al ejemplo anterior:
Consideremos la variable:

Encuestado n° 1	Responda calificando del 1 al 10
¿Esta de acuerdo con las normas del liceo?	1 2 3 4 5 6 7 8 9 10
	x

Volvemos a lo mismo, las escalas de clasificación, requieren la combinación de un cuestionario que usualmente involucran juicios subjetivos y/o la elección entre categorías cuyo limite es confuso.
En el ejemplo: ¿se puede estar de acuerdo entre el promedio 8 y 9?
Entonces los datos presentados en este tipo de escala, a pesar de ser numéricos, no son exactas y no podemos en consecuencia aplicar reglas numéricas o aritméticas.
Estos datos ordinales representados numéricamente no son numéricos.
Un ejemplo claro de variable cualitativa ordinal es la escala de Mercalli (modificada), aplicada a la medición de la intensidad de los terremotos. Esta corresponde a una escala descriptiva desde un grado hasta 12 grados, observando los efectos o daños producidos por el siniestro en las construcciones, objetos y, el impacto que provoca en las personas.

Ejemplo: grado 9: ¡pánico general! .las estructuras de albañilería mal proyectadas o mal construidas se desploman, etc.

VARIABLES CUANTITATIVAS. (DATOS NUMÉRICOS O MÉTRICOS).

Una variable se denomina cuantitativa si los datos numéricos (también llamados métricos), se usan para medir su valor.
Son datos numéricos si:
1.-consisten en valores específicamente definidos.Ejem: 40 cm., 8Kg ,12°
2.-si: x, z, x+h, z+h, son datos entonces: z-x = (z+h)-(x+h).

Esta última propiedad es una diferencia fundamental que existe entre datos ordinales "numéricos" y datos numéricos o métricos.

Ej: la temperatura, el peso, la talla, el número de hijos, las calificaciones, etc. Son variables numéricas o métricas
Los datos numéricos o métricos son números "reales".es decir operaciones aritméticas (suma resta, multiplicación, división) son validos en este tipo de datos.
Las variables cuantitativas tienen unidades de medida.

Ej: consideremos el peso en kg. De 5 personas.

N°personas	1°	2°	3°	4°	5°
Peso (Kg.)	40	55	70	70	80

En este caso la diferencia entre el primer y segundo dato es 10, que equivale a la diferencia entre la quinta y cuarta persona.
En otras palabras, si: x=40 y z=50 y h=30, entonces: z-x= (z+h)-(z+h).

También vemos que el peso de la quinta persona es exactamente el doble que el peso de la primera.

También podemos determinar el peso promedio: $\frac{40+50+55+70+80}{5}$ =59 kg.

¡Y YA HEMOS USADO LAS CUATRO OPERACIONES!

En el caso de las variables cuantitativas también existe una subdivisión, clasificándose en: discretas y continuas.

Discreta: si el número posible de valores que asume es un número contable (es decir finito o numerable). Frecuentemente viene de conteo de cosas.Ej:n° de camas de un hospital, numero de hijos etc.

Continuas: si el conjunto de posibles valores que asume, es un conjunto no contable.
Ej.: el peso de una persona (aunque sugiere que por el instrumento limitado utilizado para medir pareciera ser discreto)
Usualmente variables continuas provienen de mediciones de cosas, por ejemplo: el peso, altura, tiempo, volumen, temperatura, presión, etc...

ORGANIZACIÓN DE DATOS CUALITATIVOS.

El primer método consiste en ordenar los datos alfabéticamente. En todo caso esto depende del tipo de variable a estudiar.

DISTRIBUCION DE FRECUENCIAS: Es simplemente una lista de categorías o valores que una variable puede tomar, junto con el numero de valores que cada categoría puede tomar.
Usualmente se presentan en la forma de una tabla, lo que se conoce como datos tabulados, y las variables que intervienen se les denominan variables o datos intervalares.
Para hacer comparaciones entre tablas se debe hacer en relación a las frecuencias. Utilizaremos para este efecto las frecuencias relativas, la cual relaciona cada frecuencia con el total de los datos, es decir:

$F_i = \frac{N_i}{N}$, DONDE: F_i : Frecuencia relativa

N_i : Frecuencia observada o absoluta

N: Total de datos

Por otro lado si esta frecuencia se expresa como porcentaje del total, se denomina frecuencia relativa porcentual:

Montoya.-

$F_i(\%) = F_i * 100\ \%$

Sin embargo con algunas variables ordinales numéricas, el numero de posibles valores puede ser muy grande, por lo que una tabla de distribución de frecuencias requerirá una inmensa cantidad de filas (tantas como valores diferentes pueda asumir la variable en los datos observados). En estas circunstancias es usual agrupar los valores en clases o intervalos, de manera que el número de filas en la tabla de frecuencias sea de un tamaño "razonable". Este tipo de distribución de frecuencias se conoce con el nombre de distribuciones agrupadas o distribuciones intervalares.

Estos datos podríamos aguzarlos, en una tabla mediante intervalos.

¿COMO SE DEFINEN LOS INTERVALOS APROPIADOS?

En otras palabras :

$N = N_1 + N_2 + \ldots\ldots\ldots\ldots N_k$

Esto indica que : la suma de las frecuencias absolutas de los datos observados es igual al total de la muestra (columna 1)

$2.\text{-}fr_1 + fr_2 + fr_3 + \ldots\ldots + fr_k = \frac{N_1}{N} + \frac{N_2}{N} + \frac{N_3}{N} + \ldots\ldots\ldots + \frac{N_k}{N}$

$== \frac{N}{N} = 1$

LA SUMA DE LAS FRECUENCIAS RELATIVAS ES 1 (COLUMNA 2)

$3.\text{-}F_1\% + F_2\% + F_3\% + \ldots\ldots\ldots + = 100\%$

LA SUMA DE LAS FRECUENCIAS RELATIVAS PORCENTUALES EEQUIVALE AL 100% (COLUMNA 3)

MEDIDAS DE CENTRALIZACION (TENDENCIA CENTRAL).

TRES MEDIDAS SE USAN PARA DESCRIBIR EL "CENTRO" O "LOCALIZACION" DE UN CONJUNTO DE DATOS.
LA MODA
LA MEDIANA
LA MEDIA (PROMEDIO)
LA MEDIA GEOMETRICA
LA MEDIA ARMONICA

La elección de una de ellas para interpretar la idea de centro depende de:
*los aspectos de la "localización central" que se pretende.
*los tipos de datos de que se disponen (nominales ordinales o numéricos)

La tabla resume la medida de tendencia central que es posible usar dependiendo de los tipos de datos.

	Tipos de datos		
Medida de tendencia central	Numéricos	Ordinales	Nominales
Moda	si	si	si
Mediana	si	si	no
Media	si	no	no

LA MODA. Interpreta el significado de centro como el valor que ocurre con mayor frecuencia

1.- Para datos a granel: los siguientes datos representan la duración, en horas de 18 pilas eléctricas
237 242 232 242 282 244 242 254 262 234 220 225 246 232 218 228 240.

Usando un diagrama de tallo (las primeras dos cifras a la izquierda) y hoja (la última cifra) ordenamos los datos de la forma:

Tallos	hojas
21	3
22	0 5 8
23	7 2 2 4 2
24	2 2 4 2 6 0
25	4
26	2
27	
27	2

Este diagrama de tallo y hoja se ordena de menor a mayor en la siguiente forma:

Tallos	hojas
21	3
22	0 5 8
23	2 2 2 4 7
24	0 2 2 2 4 6
25	4
26	2
27	
27	2

Asì vemos que existen dos modas, las que corresponden a los valores 232 y 242 (cada una con frecuencia 3)

Ejemplo 2.- consideremos los valores diarios, en días hábiles, del, precio de dólar (transformado al peso chileno) en el mes de febrero del 2001. Estos valores se presentan en la tabla siguiente:

Valor del dólar	frecuencia
557	

Podemos observar que los datos con más frecuencia son 562 y 563, por lo que estos valores corresponden a la moda.

El modelo para calcular la moda de datos intervalares se discutirá en un capitulo aparte

Propiedades de la moda:

El cálculo de la moda no usa toda la información en la muestra, es decir, no involucra todos los valores obtenidos en la muestra, solo usa el valor que ocurre más a menudo. Este hecho podría considerarse como una seria debilidad si pensamos que toda la información disponible en la muestra debería ser usada.

*debido a lo anterior, la moda podría no ser muy descriptiva de lo que ocurre en la muestra .por ejemplo, si las notas de un alumno en la asignatura de MATEMÁTICA fuesen:

6.2 5.6 3.4 5.9 3.1 5.8 6.5 3.4 6.6 6.0

Que ordenadas de menor a mayor resultan:
3.1 3.4 3.4 5.6 5.8 5.9 6.0 6.2 6.5 6.6

Obtenemos que la moda es la nota 3.4, por lo que juzgar el rendimiento de este alumno en MATEMÁTICA, por medio de la moda, parece poco adecuado.

La moda puede no ser única (como vimos en el ejemplo anterior). Es común llamar distribución bimodal al caso en que los datos tienen dos modas.
*la modal es una medida volátil, esto significa que es sensitiva a pequeños cambios de los valores muéstrales.

La moda es particularmente afectada por valores extremos en la muestra (estos valores extremos son conocidos como puntos aislados mas comúnmente llamados OUTLIER). Son esos valores que resultan mucho más grandes o mucho más pequeños que el resto de los datos. La inclusión de un valor **outlier** puede determinar que una medida de localización central sea poco representativa de los datos

La moda es siempre igual a uno de los valores presentes en la muestra (en el caso de datos a granel)

La mediana para datos numéricos a granel:

-12-

La mediana identifica el valor central de los valores provenientes de una muestra. La mediana es entonces una medida de centralidad. **La mitad de los valores de la muestra serán mas grandes que la mediana y la otra mitad serán mas pequeños** (estrictamente) hablando, la mitad de los valores serán igual o menor que la mediana y la otra mitad igual o mayor que la mediana). Veamos un ejemplo:

Ejemplo:
Los índices diarios de calidad del aire registrado en la zonas Pudahuel / Cerro Navia / Lo Prado . Durante el mes de Abril 2000, se presentan en la tabla siguiente:

DIA	Índice
1	52
2	53
3	67
4	45
5	76
6	54
7	59
8	58
9	67
10	54
11	59
12	62
13	66
14	67
15	70
16	66
17	59
18	45
19	103
20	52
21	76
22	59
23	103
24	103
25	54
26	65
27	103
28	67
29	103
30	52

Ordenando estos datos en un modelo de tallo y hoja, se obtiene el diagrama:

tallos	Hojas
4	5 5
5	2 3 4 9 8 4 9 9 2 9 4 2
6	7 7 2 6 7 6 5 7
7	6 0 6
10	3 3 3 3 3

El diagrama de tallo y hoja ordenado de menor a mayor resulta:

-13-

tallos	**Hojas**
4	**5 5**
5	**2 2 2 3 4 4 4 8 9 9 9 9**
6	**2 5 6 6 7 7 7 7**
7	**0 6 6**
10	**3 3 3 3 3**

Otra forma de ver los datos ordenados de menor a mayor es:

Lugar de ordenamiento	Índice de calidad del aire
1	45
2	45
3	52
4	52
5	52
6	53
7	54
8	54
9	54
10	58
11	59
12	59
13	59
14	59
15	62
16	65
17	66
18	66
19	67
20	67
21	67
22	67
23	70
24	76
25	76
26	103
27	103
28	103
29	103
30	103

Como la cantidad de datos N= 30 (par), entonces existen dos lugares centrales estos son el 15 y el 16. La mediana se define como el promedio de estos valores es decir: $\frac{62+65}{2} = 63.5$

Según la clasificación de la calidad del aire dada en la ultima tabla, y usando como medida de tendencia central la mediana, el mes de abril se clasificaría como bueno. Sin embargo si se usa la moda que en este caso es el valor 103, el mes de Abril se catalogaría como regular.

Procedimiento para calcular la mediana para "n" datos a granel.

*Ordenamos los datos de menor a mayor, es decir, en forma ascendente, colocando frente a cada dato el lugar que ocupa en la ordenación: primer lugar, segundo lugar, tercer lugar, etc.

Se calcula el valor: $\frac{1}{2}$ (n+1).= $\frac{50}{100}$ (n+1). Se usa el valor $\frac{50}{100}$ porque se quiere indicar el valor que deje aproximadamente el 50% de los datos bajo el y aproximadamente el 50% de los datos sobre el.

Si $\frac{1}{2}$(n+1) es entero, lo que ocurre cuando "n" es impar, se ubica el dato que esta en el lugar: $\frac{1}{2}$(n+1). Este valor corresponde a la mediana.

En el caso que: $\frac{1}{2}$(n+1), no sea entero, lo que ocurre cuando n es par), se ubican los datos que están en los lugares mas próximos, ya sea a la izquierda o a la derecha de $\frac{1}{2}$(n+1).en este caso los lugares mas próximos son: $\frac{n}{2}$ y $\frac{n}{2}$+1.la mediana corresponde al promedio entre estos datos.

Ejemplo: consideremos nuevamente los valores diarios del precio del dólar en el mes de febrero del 2001

De los parámetros de posición o medidas de tendencia central.

Propiedades

Propiedades de la media.

*el calculo de la media involucra todos los datos o valores de la muestra, es decir, usa toda la información disponible, lo cual no se da en el caso de la moda ni en el de la mediana.
*por la propiedad anterior, la media es afectada por outliers, por ejemplo, la nota promedio entre: 5,7; 5,9; 5,5; 2,1; y 6,3. Es 4,9, mientras que el promedio de 5,7; 5,9; 5,5 y 6,3 es 5,85.
*la media no puede ser determinada gráficamente, como el caso de la mediana.
*si los datos muestrales se transforman linealmente, entonces la media de los datos transformados es igual a la media de los datos originales. Es decir, si $x_1, x_2 \ldots\ldots x_n$, son los datos originales y $\overline{X}$, entonces la media de los datos ,ax_1+b , ax_2+b,……ax_n+b , es igual a ,$a\overline{X}+b$.
*/Si se tienen dos muestras diferentes, digamos, $x_1, \ldots\ldots x_n$ e $y_1 \ldots\ldots y_n$, entonces la media de la muestra formada por ambas, es la suma de cada una de las medias.Esta propiedad no se cumple para la moda ni tampoco para la mediana.

Ejemplo. : En la sección deportes de una tienda comercial, se ha ofrecido a sus 5 empleados un incentivo económico por lograr cierta meta en las ventas mensuales. A cada uno de los vendedores se le ofrece un incentivo de $ 80.000 y para el jefe de la sección el incentivo de $400.000.

Analicemos la información de que el incentivo promedio de la sección deportes es de $114.000.

En este ejemplo, la madia de los cinco incentivo es de $114.000, pero dificilmente describe la situación que ocurre en la sección deportes, la media por otra parte es, $80.000 y es al menos representativa de 4 de los 5 empleados que recibirán el incentivo.

Este ejemplo nos muestra que siempre existe un peligro intrínseco al representar un conjunto de datos por un solo número. En todo caso el uso de una o de otra medida de centralización dependerá de las características del problema en si.

Ejemplo: un estudiante rindió, en el año 2004, la PSU, y las pruebas de conocimientos específicos de Historia y Geografía de Chile, Física, MATEMÁTICA y Biología .Los puntajes que obtuvo fueron:

PSU	Parte MATEMÁTICA	610
PSU	PARTE Verbal	590
PCE	Historia y Geografía de Chile	580
PCE.	Física	550
PCE.	Biología	640
PCE	MATEMÁTICA.	X

Si el puntaje promedio de todas las pruebas fue de 596 puntos.
¿Cuál fue el puntaje en la PCE DE MATEMÁTICA?

Se puede escribir la ecuación:

$$596 = \frac{610 + 590 + 580 + 550 + 640 + X}{6}$$

DE DONDE: X= 600.

Es decir el estudiante obtuvo 600 puntos en la prueba específica de MATEMÁTICA.

Ejemplo; los datos siguientes corresponden a las notas trimestrales de 4 alumnos de la cátedra de MATEMÁTICA en el Liceo Parroquial San Antonio dictada por el profesor Montoya.

	NOTAS	MEDIANA(Md)	$\overline{X}$
ALUMNO 1	**2.1 5.0 6.1**	**5.0**	4.4
ALUMNO 2	**4.2 5.0 5.8**	**5.0**	5.0
ALUMNO 3	**3.5 5.0 6.5**	**5.0**	5.0
ALUMNO 4	2.3 5.0 5.0	5.0	4.1

OBSERVAMOS QUE LA MEDIA NO CAMBIO PARA ESTOS 4 CONJUNTOS DE OBSERVACIONES, PERO SI LO HIZO LA MEDIA. COMO DIJIMOS ANTERIORMENTE, A DIFERENCIA DE LA MEDIA, LA MEDIANA NO SE VE AFECTADA CUANDO MODIFICAMOS LOS VALORES EXTREMOS .ESTA

PROPIEDAD DE LA MEDIANA PUEDE SER DESEABLE EN ALGUNOS CASOS Y EN OTROS NO. POR EJEMPLO, ¿USARÌA UD. LA MEDIANA COMO MEDIDA DE TENDENCIA CENTRSL PARA DECIDIR SI UN ALUMNO PASA O NO E CURSO? ESTO ES, ¿LE PARACE LA MEDIANA UNA BUENA MEDIDA PARA CUANTIFICAR EL RENDIMIENTO DEL AÑO DE UN ALUMNO?

Observamos también que la medida en cada uno de los cuatro conjuntos de datos es un punto de equilibrio del conjunto de observaciones, ya que compensa los desvíos negativos de las observaciones respecto de la media con los desvíos positivos.

Nota: desvío = diferencia entre el valor observado y la media de los datos.

ALUMNO 1

notas	**2.1**	**5.0**	**6.1**
desvíos	2.1-4.4	5.0-4.4	6.1-4.4

SUMA DE LOS DESVÌOS ES IGUAL A 0 (CERO).

ALUMNO 2

NOTAS	**4.2**	**5.0**	**5.8**
DESVÌOS	4.2-5.0	5.0-5.0	5.8-5.0

SUMA DE DESVÌOS = 0

ALUMNO 3

NOTAS	**3.5**	**5.0**	**6.5**
DESVÌOS	3.5-5.0	5.0-5.0	6.5-5.0

ALUMNO 4

NOTAS	**2.3**	**5.0**	**5.0**
DESVÌOS	2.3-4.1	5.0-4.1	5.0-4.1

SUMA DE DESVÌOS = 0

¿SERÀ SIEMPRE CERO LA SUMA DE LOS DESVÌOS?

GENERALIZANDO LA SITUACION:

Sean: X_1,……..X_n , un conjunto de datos, y sea $\overline{X}$ SU MEDIA .Entonces los desvíos serán:

$X_1 - \overline{X}$; $X_2 - \overline{X}$,……..,$X_n - \overline{X}$,

POR LO QUE:

$$(X_1 - \overline{X}) + (X_2 - \overline{X}) + \ldots\ldots\ldots\ldots + (X_n - \overline{X}) = (X_1 + X_2 + \ldots\ldots + X_n) - n\overline{X}$$

$$= (X_1 + X_2 + \ldots\ldots\ldots + X_n) - n\frac{X_1 + X_2 + \ldots\ldots\ldots X_n}{n}$$

$$= 0$$

Ejemplo: los siguientes datos representan el tiempo (medido en semanas) desde que se detecto SIDA a un grupo de 24 personas hasta que murieron.

45 69 73 51 84 70 68 57
52 49 80 95 76 65 56 50
66 58 62 54 58 158 104 60

Un diagrama de Tello y hojas para estos datos es:

Tallos	Hojas
4	5 9
6	1 7 2 6 0 8 4 8
5	9 8 5 6 2 0
7	3 0 6
8	4 0
9	5
10	4
11	
12	
13	
14	
15	8

El diagrama de tallos y hojas con los datos ordenados de menor a mayor es el siguiente:

Tallos	Hojas
4	5 9
5	0 0 1 2 4 6 8 8
6	0 0 2 5 6 8
7	0 3 6
8	0 4
9	5
10	4
11	
12	
13	
14	
15	8

Desde este diagrama podemos observar que la distribución de los datos muestra que la mayoría de los valores están entre 50 y 80 semanas.Tambien, hay un outlier en el valor de 158 semanas, que representa un enfermo de la muestra que sobrevivió mucho mas que el resto de los enfermos considerados en ella.

Determinación de la moda.

El càlculo de la moda **no usa toda la informa de la muestra** , es decir , no involucra todos los valores obtenidos en la muestra , solo usa el valor que ocurre mas a menudo .este hecho podría considerarse como una seria debilidad si pensamos que toda la información disponible en la muestra debería estar representado en estos parámetros.

Debido a lo anterior, la moda podría no ser muy descriptiva de lo que ocurre en la muestra.Por ejemplo, si las notas de un alumno en la asignatura de matemática fuesen

6.2 5.6 3.4 5.9 3.1 5.8 6.5 3.4 6.6 6.0

Que ordenadas de mayor a menor resultan:
3.1 3.4 3.4 5.6 5.8 5.9 6.0 6.2 6.5 6.6

Obtenemos que la moda es 3.4, por lo que juzgar el rendimiento de este alumno en matemática, por medio de la moda, parece poco adecuado.

La moda puede no ser única (si hay dos o mas datos u observacio0nes que se repiten el mismo numero de veces o tienen la misma frecuencia) .es común llamar distribución bimodal al caso en que os datos tienen dos modas.

La moda es un medida volátil, esto significa que es sensitiva a pequeños cambios de los valores muestrales.

La moda no es particularmente afectada por los valores extremos en la muestra (estos valores extremos son conocidos como puntos aislados, mas comúnmente llamados outliers).son estos valores que resultan mucho mas grandes o mucho mas pequeños que el resto de los datos. La inclusión de un valor outlier puede determinar que una medida de localización central sea poco representativa de los datos.

Resumiendo :

SI :$X_1, X_2 \ldots\ldots\ldots\ldots\ldots\ldots X_n$, n Datos. Entonces

$\overline{X} = \frac{X_1 + X_2 + \ldots\ldots\ldots + X_n}{n}$ que equivale a :

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

Ejemplo : la media para los datos a granel : 8,3,5,12,10 es

$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{8+3+5+12+10}{5} = 7.6$

Ahora bien , si los números: , $X_1, X_2 \ldots\ldots\ldots\ldots\ldots\ldots X_n$ o datos se presentan con la frecuencia $f_1, f_{2,\ldots\ldots\ldots\ldots,f_n}$,entonces la media estará dada por : $\overline{X} = \frac{X_1 * f_1 + X_2 * f_2 + \ldots\ldots\ldots\ldots X_n * f_n}{f_1 + f_2 + \ldots\ldots\ldots\ldots\ldots f_n}$

De donde: $\sum f = N$, DE OTRO MODO:

$\overline{X} = \frac{\sum X_i * f_i}{N}$

EJEMPLO : calcular la media si los números : 5,8,6,2 se presentan con frecuencia 3,2,4,1 respectivamente.

Aplicamos: $\overline{X} = \frac{X_1 * f_1 + X_2 * f_2 + \ldots\ldots\ldots\ldots X_n * f_n}{f_1 + f_2 + \ldots\ldots\ldots\ldots\ldots f_n}$

$$\overline{X} = \frac{5*3+8*2+6*4+2*1}{3+2+4+1} = \frac{15+16+24+2}{10} = \frac{57}{10} = 5.7$$

MEDIA ARITMETICA PONDERADA: es una herramienta estadística muy útil, sobre todo en la construcción de polígonos de costo, producción etc., También tiene especial aplicación en el campo de la educación.

Cuando se asocia a los números X_1, X_2 ……X_n ,ciertos factores ,w_1, w_2 ……..w_k ,que dependen de la significación o importancia de cada uno de los números . En este caso la media ponderada se define por el modelo:

$$\overline{X} = \frac{W_1 * X_1 + W_2 * X_2 + ... + W_3 * X_3}{W_1 + W_2 + ... + W_N}$$, DE OTRO MODO:

$$\overline{X} = \frac{\sum X_i * W_i}{\sum W}$$

EJEMPLO: si cada examen final de curso se valora como tres veces los exámenes parciales un estudiante tiene una nota final de 85 y sus notas parciales son 70 y 90, su promedio ponderado será:

$$X = \frac{1*70+1*90+3*85}{1+1+3} = 83$$

Propiedades de la media aritmética:

Primera propiedad: La suma algebraica de las desviaciones de un conjunto de números respecto a su media aritmética es cero:

En efecto, consideremos los números o datos a granel: 8, 3, 5, 12,10, cuya media es $\overline{x} = 7.6$ .calculando las desviaciones $x_i - \overline{x}$, que para mayor claridad se han dispuesto en una tabla:

Dato	$x_i - \overline{x}$
8	8-7.6=0.4
3	3-7.6=-4.6
5	5-7.6=-2.6
12	12-7.6=4.4
10	10-7.6=2.4

$$\sum (x_i - \overline{x}) = 0$$

Segunda propiedad: la suma de los cuadrados de las desviaciones de un conjunto de números x_i de cualquier numero “ajes mínima solamente si “a”= $\overline{x}$

Tercera propiedad: si f_i números tienen media m_1 , f_2 números tiene media m_2 ,…….f_k números tiene media m_k .entonces la media de todos los números es:

$$\overline{X} = \frac{f_1 * m_1 + f_2 * m_2 + ... + f_k * m_k}{f_1 + f_2 + ... + f_k}$$, es decir una media ponderada de todas las medias.

Ejemplo: para los números a) 2, 3, 4,5 la media es 3.5 n=4

Para los números b) 2, 4, 6,8.La media es 5 n=4

Para los números c) 3, 5, 6, 7, 4,8.La media es: 5.5 n=6

La media ponderada de las tres medias corresponde a:

$$\overline{X} = \frac{3.5*4+5*4+5.5*6}{4+4+6} = \frac{14+20+33}{14} = \frac{67}{14} = 4.79$$

QUE EVIDENTEMENTE EQUIVALE A CALCULAR LA MEDIA COMO:

$$\overline{X} = \frac{2+3+4+5+2+4+6+8+3+5+6+7+4+8}{14} = 4.79$$

Cuarta propiedad: si $\overline{X_S}$ Es cualquier supuesta media aritmética (puede ser cualquier valor, no necesariamente uno de los datos observados) y si $d_j = x_j - \overline{X_S}$, son las desviaciones de x_i de $\overline{X_S}$, la ecuación:

Montoya.

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

SE CONVIERTE EN:

$$\overline{X} = \overline{X_S} + \frac{\sum_{i=1}^{n} d_j}{n}$$ QUE CORRESPONDE A:

$$\overline{X} = \overline{X_S} + \frac{\sum d}{n}$$ O sea

$$\overline{X} = \overline{X_S} + \frac{\sum f*d}{n}$$

$$\overline{X} = \overline{X_S} + \overline{d}$$

EJEMPLO: Consideremos los números: 2, 3, 4, 5, 6, 4,8 cuya media es $\overline{x}$=4.57

Si hubiésemos aplicado la propiedad cuatro se obtiene.

Supongamos que la media supuesta es arbitrariamente 5, calculamos entonces las desviaciones $x_i - \overline{x_s}$, que para efectos de mejor claridad se han ordenado en una tabla:

numerous	*Desviaciones*
2	2-5=-3
3	3-5=-2
4	4-5=-1

5	5-5=0
6	6-5=1
4	4-5=-1
8	8-5=3

$\sum d$ =-3

$\bar{x} = 5 + \frac{-3}{7} = 5 - 0.43 = 4.57$, valor que efectivamente corresponde a la media.

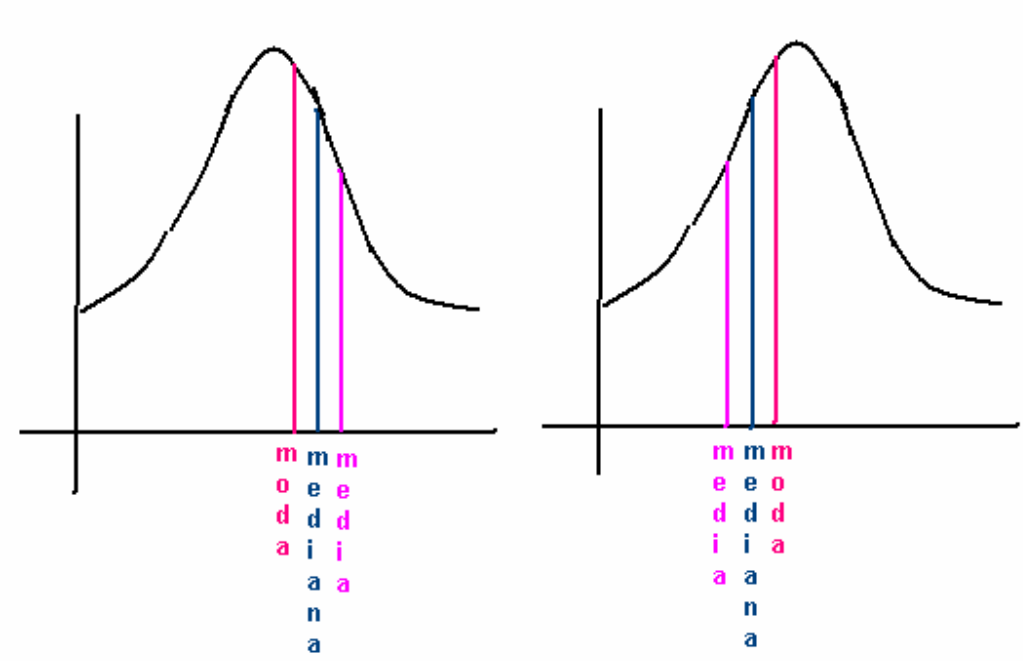
Media aritmética calculada a partir de datos intervalares.Este tema se tratara con detalles en la sección especial de datos intervalares.

Relación empírica entre mediana, media y moda

Media –moda = 3(media-mediana)

$\bar{X}$ -Mod(X) = 3($\bar{X}$ -Med)

Para curvas de frecuencias unimodales que sean moderadamente sesgadas (asimétricas), se verificara la relación anterior.



MEDIA GEOMETRICA :(G)

SI :X_1,X_2 …….X_n , n Datos. Entonces “G”

G=$\sqrt[n]{X_1 * X_2 * .....X_n}$

Ejemplo: para los datos a granel: 2, 4, 6,8. La media armónica queda expresada por:

G=$\sqrt[4]{2*4*6*8} = \sqrt[4]{1920}$

En la práctica se calcula aplicando logaritmos:

Log G = $\frac{1}{4} Log 1920$

LogG=$\frac{1}{4} * 3.28330$

LogG=0.820825

G=6.62

LA MEDIA ARMONICA: (H)

SI :X_1,X_2 …….X_n , n Datos. Entonces

H=$\frac{1}{\frac{1}{X_1}+\frac{1}{X_2}+.....+\frac{1}{X_n}}$

Ejemplo: la media armónica para los datos a granel: 2, 4, 5, 6,8 corresponde a:

$\frac{1}{H}=\frac{1}{2}+\frac{1}{4}+\frac{1}{5}+\frac{1}{6}+\frac{1}{8}=0.5+0.25+0.2+0.16+0.125=1.235$, ENTONCES

H=0.809

RELACION ENTRE LAS MEDIAS: ARITMETICA, GEOMETRICA Y ARMONICA

$$\mathbf{H} \leq \mathbf{G} \leq \overline{X}$$

LA MEDIA CUADRATICA (RMS).

SI: X_1, X_2 …….X_n , n Datos. Entonces

RMS=$\sqrt{X_1^2 + X_2^2 + ... + X_n^2}$

Ejemplo: la media cuadrática para los datos a granel: 2, 3,4 corresponde a:

RMS=$\sqrt{2^2+3^2+4^2} = \sqrt{4+9+16} = \sqrt{29} = 3.109$

SE APLICA FUNDAMENTALMENTE A LA FISICA.
ES DECIR:

MEDIDAS DE DISPERSION.

Estas medidas describen la variabilidad existente entre los datos , es decir , si estos están "estrechamente " o " ampliamente" dispersos. Por ejemplo , dos balanzas diferentes ,pesan 100 bolsas de azúcar con un peso nominal de 1 Kg. cada una (1000gr).Los registros se observan en la tabla siguiente :

Balanza 1

Valores	**995**	**996**	**997**	**998**	**999**	**1000**	**1001**	**1002**	**1003**	**1004**	**1005**
Frecuencia	3	5	8	12	15	16	14	12	8	4	3

Balanza 2

Valores (gr.)	**998**	**999**	**1000**	**1001**	**1002**
frecuencia	17	20	27	19	17

Los siguientes gráficos de frecuencias asociadas a las balanzas 1 y 2 ilustran la idea de dispersión.

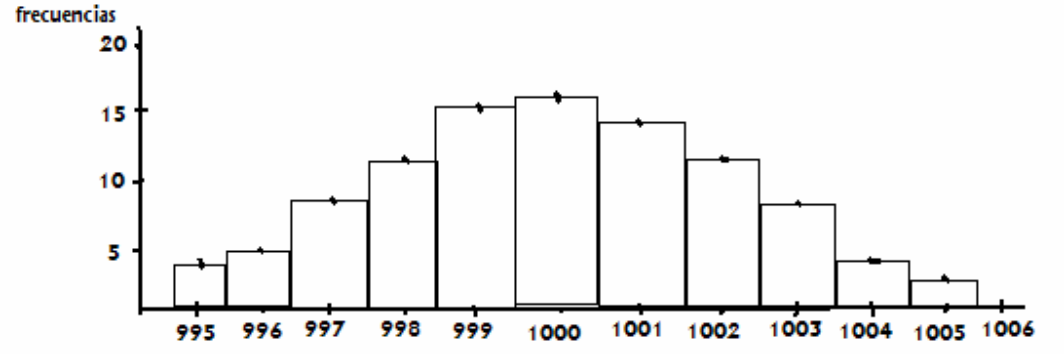


Grafico de frecuencias de la balanza 1

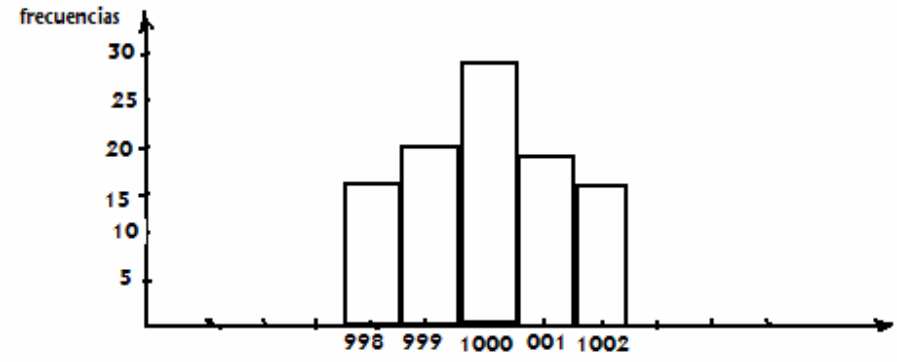


Grafico de frecuencias de la balanza 2

Notemos que la media para los registros que entrega la balanza 1 es de 999.95 gr. Y la media para los valores obtenidos para la balanza 2 es 999.99 gr. , es decir son prácticamente iguales.Sin embargo , los valores obtenidos con la balanza 1 están bastante mas dispersos que los que entrega la balanza 2,en un caso varían ente 995 gr. y 1005 gr. , y en el otro varían ente 998 gr. y 1002 gr.

También se observa que os registros obtenidos con la balanza 2 están mas cerca de la media que en el caso de la balanza 1.

Como consecuencia de lo anterior, la mayor dispersión de los datos se han obtenidos con la balanza 1, significa que algunos de estos no estarán "muy cerca" del valor medio 999.95 gr.

En general la "media" de una distribución con dispersión reducida estará mas cerca de cada valor individual de la muestra que en el caso en que la distribución tenga una dispersión mas grande.

En resumen, medidas de dispersión indican, hablando de un modo general, la distancia promedio de los valores muestrales al valor centro de la distribución.

En esta sección se tratan tres tipos diferentes de dispersión, basadas en las frecuencias, en el rango y en la desviación.

LAS MEDIDAS DE DISPERSION BASADAS EN LAS FRECUENCIAS, USAN LA FORMA EN QUE LOS VALORES DE LA MUESTRA ESTAN DISPERSOS EN LAS DIFERENTES CATEGORIAS O CLASE.ESTAS SON USADAS PRINCIPALMENTE EN DATOS NOMINALES Y ORDINALES (NO NUMERICOS) E INCLUYE EL RADIO O TAZA DE VARIACIO (RV), EL INDICE DE DIVERSIDAD /ID) Y EL INDICE DE VARIACION CUALITATIVA (IVQ), ESTOS DOS ULTIMOS SOLO LOS MENCIONAREMOS COMO REFERENCIA GENERAL.

Cabe recordar que datos ordinales numéricos, significa que pueden tomar forma numérica, pero que no cumplen las reglas algebraicas básicas, de los números reales.

*las medidas de dispersión basadas en el rango usan la diferencia (o rango) entre el valor mayor y el menor.Estas **son usadas principalmente en datos ordinales (numéricos) y datos numéricos, e incluye el rango y el rango intercuartil.**

***las medidas de dispersión basadas en las desviaciones, usan la diferencia entre cada valor de la muestra y la media.Estas son usadas con datos numéricos e incluye la desviación estándar (SD) y la desviación media (DM).**

Como en el caso de las medidas de tendencia central, la elección de una apropiada medida de dispersión esta determinada por el tipo de datos que contenga la muestra.

LAS ELECCIONES SON LAS SIGUIENTES:

1.- Con datos nominales se debería normalmente restringir a las medidas basadas en las frecuencias.

2.- Con los datos ordinales (no numéricos), también se usan las medidas basadas en las frecuencias.

3.- Con datos ordinales "numéricos" normalmente se debería elegir entre las medidas basadas en el rango, no obstante, si el número de valores posibles es suficientemente pequeño para no requerir una distribución de frecuencias agrupadas, podría usarse una medida basada en las frecuencias.

4.- Con los datos numéricos debería usualmente trabajarse con medidas basadas en las desviaciones, no obstante cuando la distribución de los datos es asimétrica, medidas basadas en el rango pueden ser apropiadas.

LA TABLA SIGUIENTE RESUME LO EXPUESTO ANTERIORMENTE.

TIPOS DE MEDIDAS	TIPOS DE DATOS			
	NOMINALES	ORDINALES NO NUMERICAS	ORDINALES NUMERICAS	NUMERICOS
BASADAS EN LAS FRECUENCIAS (RV,ID,IVQ)	SI	SI	NO	NO
BASADAS EN EL RANGO (R,RIQ)	NO	NO	SI	SI
BASADAS EN LAS DESVIACIONES (SD,DM)	NO	NO	NO	SI

MEDIDAS DE DISPERSION PARA DATOS NOMINALES:

Como lo indica la tabla anterior, las medidas de dispersión para datos nominales están basadas en las frecuencias, estas miden el grado de heterogeneidad de los valores de la muestra, en otras palabras, el grado en el cual, los valores de la muestra están divididos entre las categorías.

-25-

En un extremo, si todos los datos están en una sola categoría, diremos que la dispersión es homogénea, es decir la dispersión es cero .En otro extremo, si hay un numero igual de datos en cada categoría, describiremos la dispersión como heterogénea, es decir, los valores de la muestra están dispersos tanto como es posible.

Ejemplo DISTRIBUCIN DE FRECUENCIAS HOMOGENEAS.

SEXO DE UN GRUPO CURSO	**Nº DE ALUMNOS**
MUJER	**0**
HOMBRE	**42**

DISTRIBUCION DE FRECUENCIAS HETEROGENEAS.

SEXO DE UN GRUPO CURSO	Nº DE ALUMNO
MUJER	21
HOMBRE	21

EL RADIO TASA) DE VARIACION (RV)

El radio de variación mide la proporción de observación de la muestra que no están en la clase modal (clase con mayor frecuencia) .El valor de RV, cuando no hay dispersión, es decir, cuando todos los datos están en una sola categoría, es cero.

El valor máximo teórico es 1, pero este puede ser alcanzado solamente si hay un número infinito de categorías.

En la practica, el valor máximo (el cual será menor que 1), depende del tamaño de la muestra, y del numero de categorías, pero resulta cercano a 1 cuando el numero de valores de la muestra es el mismo en cada categoria.Esta dependencia, del máximo valor de RV, sobre el numero de categorías, hace difícil comparar distribuciones con diferente numero de categorías, pero el RV es una buena medida de dispersión para distribuciones similares.

El cálculo de RV se realiza de la siguiente forma:

*CONSTRUIR LA DISTRIBUCION DE FRECUENCIA PARA OS VALORES DE LA MUESTRA.
*ENCONTRAR LA MODA, ES DECIR, LA CATEGORIA CON MAS ALTA FRECUENCIA, LA FRECUENCIA ASOCIADA A ESTA CATEGORIA LA DENOMINAREMOS FRECUENCIA MODAL (FM)
*DIVIDIR fm POR EL TAMAÑO DE LA MUESTRA (DIGAMOS n)
*EL VALOR DE RV SE DEFINE COMO;

$$RV=\frac{n-fm}{n}$$

$$RV=1-\frac{fm}{n}$$

Para la primera de las tablas anteriores, la categoría con más alta frecuencia es "hombre" y su frecuencia es 42.Así fm= =42, de donde:

$$RV=1-\frac{42}{42}=0$$

Para los datos de la segunda TABLA, fm=21 (CUALQUIERA DE LAS CATEGORIAS SIRVE,) notar que en este caso hay dos modas, ambas con igual frecuencia, en consecuencia: $RV=1-\frac{21}{42}=1-\frac{1}{2}=\frac{1}{2}$

Esto significa que el 50% de los valores esta fuera de la categoría modal.

Ejemplo: La siguiente tabla muestra el número de alumnos del Liceo Parroquial San Antonio, inscritos para rendir la PSU en los años 2004 y 2005, según la categoría cursos.

	PROMOCION	
CURSOS	2004	2005
4ºA	40	36
4ºB	42	34
4ºC	43	40
TOTALES	125	110

Para los inscritos de la promoción 2004 se tiene que: $fm=1-\frac{43}{125}=0.656$

O sea el 65,6% de los alumnos inscritos en el 2004 de la clase modal son del 4ºC

Para los inscritos en el 2005 $RV=1-\frac{40}{110}=1-0.636$

O sea el 0.656 es más cercano a 1 que 0,636, entonces la dispersión sobre las categorías (cursos) es mayor en el año 2004 que en el año 2005

Montoya.

OBSERVACION: Lo que se puede decir sobre el RV es que valores bajos (es decir valores tendientes a cero), indican bajo monto de dispersión (la mayoría de los valores de la muestra en solo unas pocas categorías, algunas categorías tal vez vacías), mientras valores altos (es decir, tendientes a 1) indican alto monto de dispersión (similar numero de valores en cada categoría)

No es fácil decir que significa en una muestra que el valor de RV sea, digamos 0.4 o 0.7. Por esta razón, el RV a menudo es más usado para comparar dispersión de dos o más distribuciones muestrales.

En términos de propiedades deseables, el RV no usa toda la información de la muestra (esta basado solo en la moda) y es sensitivo a pequeño cambios en los datos.Sin embargo, es fácil de calcular y es resistente a los outliers.

MEDIDAS DE DISPERSION PARA DATOS ORDINALES.

Para este tipo de datos también se usa como medida de dispersión (basada en frecuencias) el radio de variación RV.

Ejemplo: Los siguientes datos muestran los resultados de estudios realizados por dos compañías telefónicas acerca de la calidad de atención que ellos entregan a sus clientes.

	GRADOS DE	SATISFACCION	DEL CLIENTE	
Compañía	MALA	REGULAR	BUENA	**MUY BUENA**
VTR	13.4%	47.5%	34.2%	**4.9%**
C.M.E.T	**32.9%**	**53.8%**	**10.6%**	**2.7%**

En el caso de VTR. El radio de variación resultante es $RV=1-\frac{47.5\%}{100\%}=0.525$

Mientras que el radio de variación de C.M.E.T es: $RV=1-\frac{53.8\%}{100\%}=0.462$

OTRAS MEDIDAS DE DISPEERSION USADA EN DATOS ORDINALES, PERO "NUMERICOS", SON EL "RANGO" Y EL "RANGO INTERCUARTIL"

-27-

EL RANGO.

Es la más simple de las medidas de dispersión para datos ordinales "numéricos", y es fácilmente calculable, ya que es la diferencia entre el valor máximo y el valor mínimo de la muestra.

Ejemplo: Se aplican dos ensayos de PSU, en el taller del Liceo Parroquial San Antonio, una en Marzo y la otra en Noviembre del 2005.Los resultados se presentan en la tabla siguiente.

alumno	Puntaje ensayo Marzo	**Puntaje ensayo Noviembre**
1	425	**604**
2	556	**508**
3	325	**616**
4	466	**704**
5	643	**690**
6	512	**610**
7	506	**594**
8	570	**450**
9	565	**580**
10	**540**	**624**

En el ensayo de Marzo, el rango es: R= 643-325=318, mientras que en el ensayo de Noviembre el rango corresponde a R=704-450=254

Lo anterior indica que la dispersión (medida con el rango) es mayor en el ensayo de marzo que en el que se aplica en Noviembre.

Notar que si se omiten los puntajes más bajos obtenidos por los alumnos en ambos ensayos, los rangos serán:
Marzo: R= 643-425=218
Noviembre: 704-508=196

Las diferencias disminuyen, entonces la dispersión asociada en cada ensayo son mas cercanas .Esto muestra lo sensible que es esta medida de dispersión frente a los outliers.

EL RANGO INTERCUARTIL. (RIQ)

Una manera de resolver el problema de la sensibilidad del rango frente a los outlier, es eliminar algún número predeterminado, o un porcentaje predeterminado, de observaciones desde la parte inferior y superior de la muestra.Lo anterior elimina la influencia de valores extremos (outliers)

La medida de dispersión es ahora el valor máximo menos el valor mínimo de los datos que no fueron eliminados.

Si el porcentaje que eliminamos, tanto en la parte inferior como en la superior es el 25% de los datos muestreados, entonces la diferencia entre el valor máximo y el valor mínimo de los datos que no fueron eliminados, es la medida de dispersión conocida con el nombre de RANGO INTERCUARTIL (RIQ).

Recordemos que el valor que deja por debajo de un 25% de la muestra se llama PRIMER CUARTIL (Q_1) y el valor que deja por arriba del 25% de la muestra se llama TERCER CUARTIL (Q_3). En consecuencia:

$$RIQ = Q_3 - Q_1$$

Ejemplo: Los siguientes son los puntajes en la PSU, en las categorías Matemática y verbal de un grupo de 18 alumnos del Liceo Parroquial San Antonio .en el año 2004.

	Verbal	**Matemática**
Nº de alumno	puntaje	Puntaje
1	556	425
2	428	704
3	612	696
4	643	589
5	596	576
6	502	504
7	645	768
8	704	743
9	496	694
10	508	683
11	546	654
12	612	564
13	495	552
14	501	492
15	510	604
16	696	563
17	770	710
18	772	740

Ordenando estos datos en forma ascendente, tenemos:

Lugar	verbal	**Matemática**
1	428	**425**
2	495	**492**
3	496	**504**
4	501	**552**
5	502	**563**
6	508	**564**
7	510	**576**
8	546	**589**
9	556	**604**

10	596	**654**
11	612	**683**
12	612	**694**
13	643	**696**
14	645	**704**
15	696	**743**
16	704	**768**
17	710	**770**
18	**740**	**772**

Como n=18. Entonces

$Q_1 = \frac{25}{100} * 18 = 4.7$ $\qquad$ $Q_3 = \frac{75}{100} * 18 = 13.5$

Entonces Q_1 es el promedio de los datos ubicados en los lugares 4 y 5

Q_3 es el promedio de los datos ubicados en los lugares 13 y 14.

Así para los alumnos en la categoría Matemática: RIQ = Q_3 - Q_1

$$\text{RIQ} = \frac{696+704}{2} - \frac{552+563}{2} = \frac{1400}{2} - \frac{1115}{2} = 700 - 555.75 = 144.25$$

Ahora para los alumnos de la categoría Verbal

$$\text{RIQ} = \frac{643+645}{2} - \frac{501+502}{2} = \frac{1288}{2} - \frac{1003}{2} = 644 - 501.5 = 142.5$$

Esto muestra que la dispersión en ambas categorías, tienen prácticamente el mismo valor, es decir, en términos de movilidad, el grupo de 18 alumnos, tanto en Matemáticas como en Verbal, tienen puntajes homogéneos.

Cabe señalar que el parámetro estadístico RIQ es resistente (no sensible) a los OUTLIERS, y que el hecho de eliminar el 25% inferior y el 25% superior de los datos, es solo una convención, eventualmente podrían eliminarse el 10% inferior y el 15% superior u otros porcentajes.

MEDIDAS DE DISPERSION PARA DATOS NUMERICOS.(CUANTITATIVOS)

Para datos numéricos, medidas de dispersión basadas en las desviaciones son generalmente usadas, aunque en los casos en que los datos tengan outliers o sean marcadamente sesgados, la mediana y el rango intercuartil pueden ser mas representativos del conjunto de los datos.Las diferentes desviaciones entre cada valor de la muestra y su valor medio son conocidas como “ERRORES DE REPRESENTACION”.

LA DESVIACION MEDIA.

Una idea que surge para medir la dispersión en base a las desviaciones (errores de representación), seria sumar todas las desviaciones que se obtienen en la muestra.Un primer problema es que si dos muestras tienen tamaños diferentes, el numero de desviaciones para cada muestra seria distinto, por lo que comparar la suma de sus respectivos desvíos, no seria apropiado.Una solución a este problema seria comparar los promedios de desvíos de cada muestra.

Sin embargo la dificultad fundamental permanece, puesto que, en cualquier muestra, siempre, los desvíos negativos se cancelan con los desvíos positivos, obteniéndose que la suma de los desvíos es invariablemente cero...Es decir esta forma de medir la dispersión carece absolutamente de sentido.

Una manera de resolver la dificultad es ignorar el signo de cada desvío (que en adelante llamaremos “errores”).Tales valores son conocidos como “errores absolutos”.Si calculamos la suma de los errores absolutos y

la dividimos por el tamaño de la muestra, es decir, calculamos el promedio de los errores absolutos, obtenemos una medida de dispersión conocida con el nombre de "DESVIACION MEDIA" (D.M)

EJEMPLOS:

Consideremos los datos de la tabla anterior, cuyas medias correspondientes son:

Media Verbal: Xv=588 media Matemática: Xm= 626.

Lugar	verbal	**Matemática**
1	428	**425**
2	495	**492**
3	496	**504**
4	501	**552**
5	502	**563**
6	508	**564**
7	510	**576**
8	546	**589**
9	556	**604**
10	596	**654**
11	612	**683**
12	612	**694**
13	643	**696**
14	645	**704**
15	696	**743**
16	704	**768**
17	710	**770**
18	**740**	**772**

AHORA LISTAMOS LOS ERRORES ABSOLUTOS PARA LA CATEGORIA MATEMATICA, TENEMOS:

Dato (Xi)	425	704	696	589	576	504	768	743	694	683	654	564	552	492	604	563	710	**740**
E.A	201	78	70	37	50	122	142	117	68	57	28	62	74	134	22	63	84	144

$$D.M=\frac{\sum_{n=1}^{18} E:A}{18}=\frac{1553}{18}=86.27$$

Y para la categoría Verbal, cuya media es 588

Dato (Xi)	***428***	***495***	***496***	***501***	***502***	***508***	***510***	***546***	***556***	***596***	***612***	***612***	***643***	***645***	***696***	***704***	***710***	***740***
E.A	160	93	92	87	86	80	78	42	32	8	24	24	55	57	108	116	122	152

$$D.M=\frac{\sum_{n=1}^{18} E.A}{18}\frac{1416}{18}=78.6=79$$

Montoya.

De esta manera, se puede concluir, que: los datos de la categoría matemática están algo más dispersos que los de la categoría Verbal.

Para el cálculo de los errores absolutos no es necesario que los datos estén ordenados.

Decir que la desviación del dato 501 en la categoría Verbal con respecto a la media de estos datos (588) es 88, significa que este dato esta a una distancia de 87 puntos con respecto a esta medida de centralización.

Ejemplo 2:
Para los siguientes datos que se muestran en la tabla de distribución de frecuencias, calculemos su desviación media (D.M)

Dato (Xi)	**Frecuencias absolutas (fi)**
0	4
1	5
2	7
3	3
5	1
total	20

Primeramente observemos que el valor medio resultante es:

$$\overline{X} = \frac{0*4+1*5+2*7+3*3+5*1}{20} = 1.65$$

Los errores o desvíos son: 0-1.65=-1.65

1-1.65= -0.65

2-1.65= 0.35

3-1.65 = 1.35

5-1.65 = 3.35.

Luego, la suma de los desvíos es: 4*(-1.65) +5*(-0.65)+7*(0.35+3*1.35+1*3.35=0,

Y la suma de los errores absolutos resulta: 19.7.

En consecuencia, la dispersión medida por los desvíos medios es D.M = 19.7:20 = 0.985

En general si disponemos del conjunto de valores x_1 ,x_2 …….x_n , la desviación media para estos valores se define por:

$$\textbf{D.M=}\frac{\left|x_1-\overline{x}\right|+\left|x_2-\overline{x}\right|+.....+\left|x_n-\overline{x}\right|}{n}$$

Donde $\overline{x}$, representa la media de los valores x_1 ,x_2 …….x_n ,

Note que si transformamos el conjunto de valores linealmente, es decir el valor x, se transforma en :$y_i = ax_i + b$, entonces:

$$\overline{y} = \frac{(ax_1+b)+(ax_2+b)+.....+(ax_n+b)}{n}$$

$$=\frac{a(x_1+x_2+.....x_n)}{n}+\frac{nb}{n}$$

$$=a\overline{x}+b,$$

Y la desviación media de los valores transformados es:

$$\text{D.M}=\frac{\left|y_1-\overline{y}\right|+\left|y_2-\overline{y}\right|+.....+\left|y_n-\overline{y}\right|}{n}$$

$$=\frac{\left|ax_1+b-(a\overline{x}+b)\right|+\left|ax_2+b-(a\overline{x}+b)\right|+.....+\left|ax_n+b-(a\overline{x}+b)\right|}{n}$$

$$DM=\frac{|a||x_1-\bar{x}|+|a||x_2-\bar{x}|+.....+|a||x_n-\bar{x}|}{n}$$

$$DM=|a|\frac{|x_1-\bar{x}|+|x_2-\bar{x}|+.....+|x_n-\bar{x}|}{n}$$

D.M = $|a|$ *D.M (valores originales).

Es decir:

D.M (ax_1+b) = $|a|$ D.M (x_1)

Cabe señalar que D.M es muy poco usada como medida de dispersión, principalmente por las dificultades algebraicas que produce la suma de valores absolutos, lo cual no permite obtener buenas propiedades algebraicas para esta medida.

LA DESVIACION ESTÁNDAR COMO MEDIDA DE DISPERSION.

La desviación estándar resulta ser la distancia que se obtiene entre un dato observado o medido y la media de los mismos,

Es decir: $$S_x=\sqrt{\frac{(x_1)^2+(x_2)^2+.....(x_n)^2}{n}-\frac{(x_1+x_2+.....x_n)^2}{b}}$$

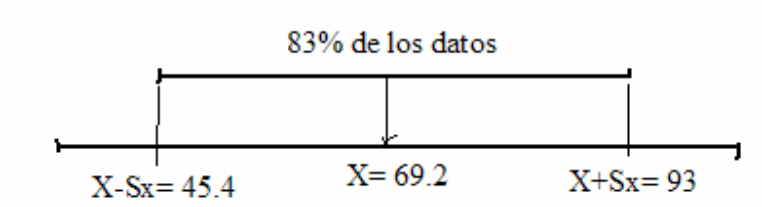
Ejemplo: consideremos los datos de la tabla que se indica:

Dato (Xi)	Error : Xi- $\bar{X}$	**Error cuadrático(Xi- $\bar{X}$) ^{2}**
45	-24.2	585.64
69	-0.2	0.04
73	3.8	14.44
51	-18.2	331.24
84	14.8	219.04
70	0.8	0.64
68	-1.2	1.44
57	-12.2	148.84
52	-17.2	295.84
49	-20.2	408.04
80	10.8	116.64
95	25.8	665.64
76	6.8	46.24
65	-4.2	17.64
56	-13.2	174.24
50	-19.2	368.64
66	-3.2	10.24
58	-11.2	125.44
62	-7.2	518.40
54	-15.2	231.04
58	-11.2	125.44
158	88.8	7885.44
104	34.8	1211.04

60	-9.2	84.64
Total	-0.8	13585.92

Por lo tanto: $S_x = \sqrt{\frac{13585.92}{24}} = 23.8$

Notemos que, aproximadamente el 83% de los datos se encuentran entre 69.2-23.8 = 45.4 y 69.2+23.8 =93. Se puede establecer entonces que:



OBSERVACION: En muchos casos la desviación estándar se define como el valor:

$$\sigma_{n-1} = \sqrt{\frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + ...(x_n - \overline{x})^2}{n-1}}$$

La razón de dividir por "n-1", en lugar de n, tiene relación con el uso de la desviación estándar como "estimador" de la desviación estándar de toda la población.

Es común encontrar en las calculadoras los símbolos σ_{n-1} y σ_n

σ_{n-1} : Indica la desviación estándar de la muestra

σ_n : Indica la desviación estándar poblacional

Ejemplo: Supongamos que disponemos de los datos: 2 4 6 8 10.

La media corresponde a: $\overline{X} = 6$, y la desviación estándar es $Sx = \sqrt{\frac{2^2+4^2+6^2+8^2+10^2}{5} - 6^2} = \sqrt{8}$

Supongamos ahora que a cada dato le sumamos una constante, digamos 7, se tendrán entonces los valores:
9 11 13 15 17.

El nuevo promedio9 que resulta es: $\overline{X}' = 13 = \overline{X} + 7$, y la desviación estándar es:

$Sx' = \sqrt{\frac{9^2+11^2+13^2+15^2+17^2}{5} - 13^2} = \sqrt{\frac{885}{5} - 169} = \sqrt{8}$

Si en lugar de sumarle una constante a cada dato, multiplicamos cada dato por una constante, digamos 4, tendríamos: 8 16 24 32 40.

Entonces el promedio que resulta es: $\overline{X}'' = 24 = 4 * \overline{X}$. Y la desviación estándar:

$Sx'' = \sqrt{\frac{8^2+16^2+24^2+32^2+40^2}{5} - 24^2} = \sqrt{128} = 4\sqrt{8}$.

Este ejemplo es un caso particular de la siguiente propiedad que posee la desviación estándar cuando los datos se transforman linealmente.

SEAN: $X_1, X_2, ...X_n$, un conjunto de n datos, los cuales se transforman linealmente en $W_1 .....W_n$ con $W_1 = AX_1 + B$ .Entonces, la media de los datos transformados es:

$$\overline{W} = \frac{(AX_1 + B) + (AX_2 + B) + \ldots\ldots\ldots\ldots + (AX_n + B)}{n}$$

$$\overline{W} = A\frac{(X_1 + X_2 + \ldots\ldots\ldots\ldots + X_{n)}}{n} + \frac{nB}{n}$$

$$\overline{W} = A\overline{X} + B$$

Y la desviación estándar:

$$Sw = \sqrt{\frac{(W_1 - \overline{W})^2 + (W_2 - \overline{W})^2 + \ldots\ldots\ldots\ldots + (W_n - \overline{W})^2}{n}}$$

$$Sw = \sqrt{\frac{(AX_1 + B - (A\overline{X} + B))^2 + (AX_2 + B - (A\overline{X} + B))^2 + \ldots\ldots\ldots\ldots + (AX_n + B - (A\overline{X} + B))^2}{n}}$$

$$Sw = \sqrt{\frac{A^2(X_1 - \overline{X})^2 + A^2(X_2 - \overline{X})^2 + \ldots\ldots\ldots\ldots + A^2(X_n - \overline{X})^2}{n}}$$

$$Sw = \sqrt{\frac{A^2\left[(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \ldots\ldots\ldots\ldots + (X_n - \overline{X})^2\right]}{n}}$$

$$Sw = A\sqrt{\frac{\left[(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \ldots\ldots\ldots\ldots + (X_n - \overline{X})^2\right]}{n}}$$

Sw = A*Sx

Es decir, si se tienen los datos: $X_1, X_2, \ldots\ldots\ldots\ldots X_n$, y cada dato se multiplica por una constante A, y luego se le suma la constante B, entonces la nueva media es la media de los datos iniciales por la constante A mas la constante B.

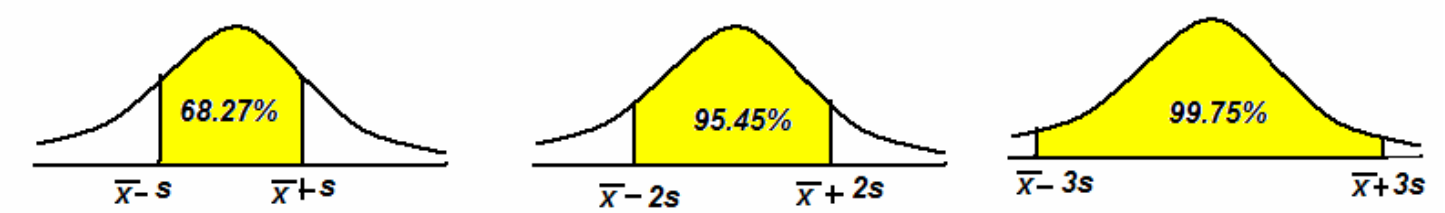
Se dice que la nueva media es una combinación lineal de la media de los datos iniciales.

Por otro lado una combinación lineal de Sx (datos iniciales), corresponde a la desviación inicial Sx multiplicada por el factor de linealizacion, eliminándose la constante aditiva.

OBSERVACIONES : Para distribuciones normales

- el 68.27 % de los casos están comprendidos entre $\overline{X} - S_X y \overline{X} + S_X$
- el 95.45% de los casos están comprendidos entre : $\overline{X} - 2S_x y \overline{X} + S_x$
- el 99.75% de los casos esta comprendido entre : $\overline{X} - 3S_X y \overline{X} + 3S_X$

Gráficamente:



Empleando cálculos bastante laboriosos, se puede demostrar que el modelo de la función densidad que corresponde a la tabla de distribución, conocida también como CAMPANA DE GAUSS, viene dada por la formula:

$$F(x) = \frac{1}{S_x\sqrt{2\pi}} * e^{\frac{-(x-\overline{X})^2}{S_x^2}}$$

Sx: desviación típica , Sx^2 : varianza.

Domf : R

Máximo: $\left(\overline{X}, \frac{1}{Sx\sqrt{2\pi}}\right)$

Puntos de inflexión : $\overline{X}$ +Sx , $\overline{X}$ -Sx ,

Asintotas: eje OX

Simetría: respecto a la recta x = $\overline{X}$

Signo: siempre positivo.

Intersección con el eje OY = $\left(0, \frac{1}{Sx\sqrt{2\pi}} e^{\frac{-\bar{x}}{2Sx}}\right)$

DIAGRAMAS DE CAJA.

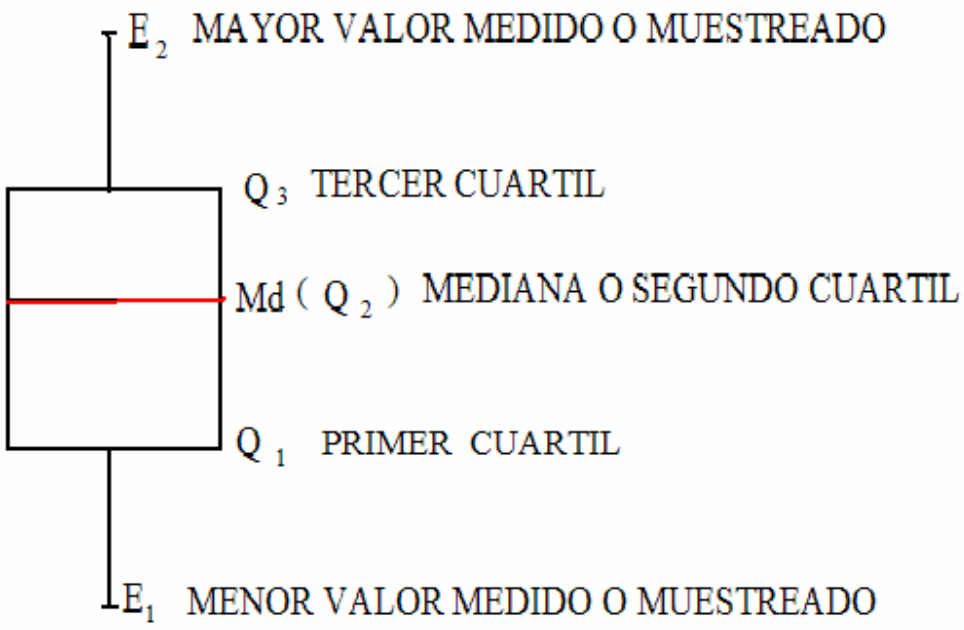
Es un modelo que permite analizar datos visualmente.

-35-

Es adecuado tanto para la $\overline{X}$ como para Sx

Son útiles para representar un conjunto de datos a granel como para datos intervalares.

Se representan de acuerdo al siguiente modelo.



EJEMPLO:

Los siguientes datos corresponden a las notas de una prueba de estadística, aplicada al cuarto año B del Liceo Parroquial San Antonio:

5.8	4.0	6.5	4.1	5.7	6.3	4.2	5.8	5.9	5.7	4.0	5.9	5.9	6.6
4.0	4.2	4.1	6.4	4.5	4.3	6.2	4.5	7.0	3.2	2.3	6.7	3.9	6.6

Aplicamos la técnica de conteo tallo y hoja .tenemos:

TALLOS	HOJAS
1	
2	3
3	2 9
4	0 0 0 1 1 2 2 3 5 5
5	7 7 8 8 9 9 9
6	2 3 4 5 6 6 7
7	0

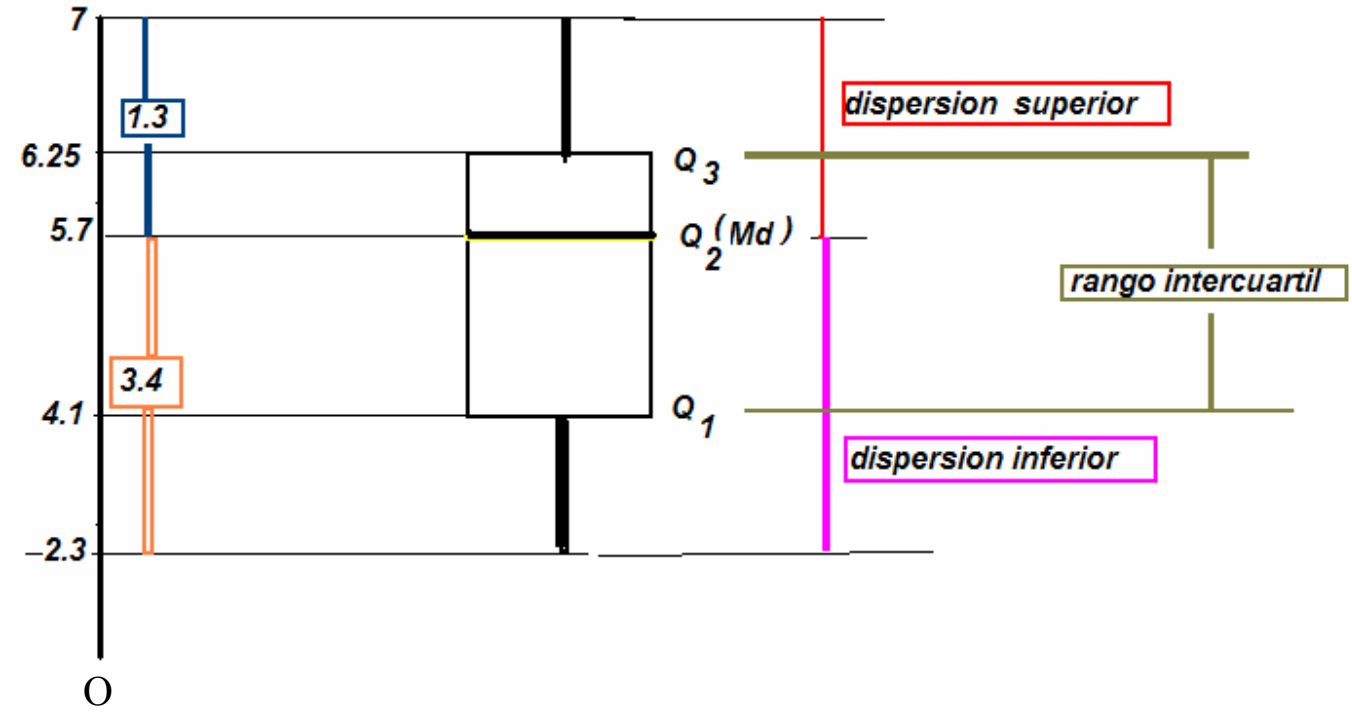
De aquí, entonces:

$E_1 = 2.3$ $E_2 = 7.0$

$\frac{25}{100} * 28 = 7$, entonces. $Q_1 \frac{41+41}{2} = 4.1$

$\frac{50}{100} * 28 = 14$, entonces $Q_2 \frac{57+57}{2} = 5.7$ (Md)

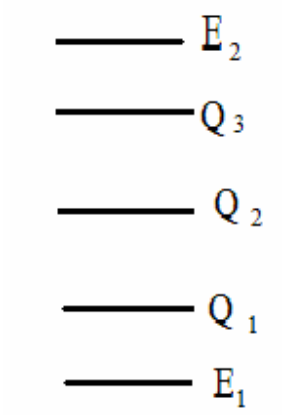
$\frac{75}{100} * 28 = 21$, entonces $Q_3 = \frac{62+63}{2} = 6.25$



La distancia entre E_1 y Q_2, corresponde a la diferencia entre $Q_2 - E_1$=5.7-2.3 = 3.4, que se denomina dispersión inferior .La distancia entre Q_2 y E_2, es decir la diferencia $E_2 - Q_2$ = 7.0-5.7 =1.3, se denomina dispersión superior

Recordemos que la distancia entre Q_3 y Q_1 , esto es: $Q_3 - Q_1$= 6.25-4.1 =2.15, se denomina RANGO **INTERCUARTIL.**

La comparación de estas distancias entrega información sobre la dispersión de los datos y también sobre la forma de su distribución.Por ejemplo, si la distribución fuese relativamente simétrica, entonces la dispersión inferior debería ser aproximadamente igual a la dispersión superior.Tambien; $Q_2 - Q_1$ debería ser similar a $Q_3 - Q_2$ y $Q_1 - E_1$ similar a $E_2 - Q_3$.



En consecuencia, las notas del cuarto B, en ele ejemplo, parecieran no ser simétricas (en torno a su media).

Los diagramas de cajas son de especial eficacia para retratar comparaciones entre conjuntos de datos.

Por ejemplo: En el Liceo Parroquial San Antonio, las notas finales de la asignatura de Matemáticas del 4º año Medio A y del 4º año Medio C, son las siguientes:

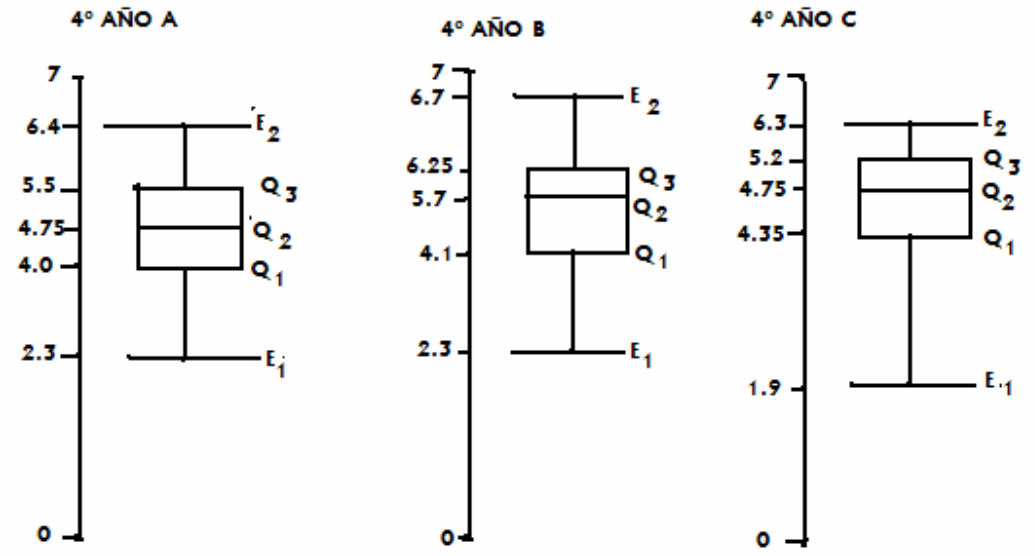
4º año Medio A:

4.8	5.2	57	4.0	4.1	5.4	6.2	3.7	4.9	5.0	4.1	3.5	6.4	3.3
5.6	4.7	6.2	4.5	4.0	2.3	4.1	3.6	4.8	5.7	5.9			

Cuarto año C:

4.5	4.9	4.2	4.4	4.1	5.2	1.9	4.7	5.4	5.0	5.2	5.1	4.4	4.5
4.3	4.7	4.5	5.2	5.9	4.1	4.4	5.1	5.5	6.1	5.2	3.7	4.8	4.9
4.3	4.0	5.1	4.5	6.3	5.1	5.3	4.9	5.4	3.3	4.4	3.9		

A CONTINUACION SE MUESTRAN LOS DIAGRAMAS DE CAJA DE LOS TRES CUARTOS AÑOS MEDIOS:



A partir de estos diagramas , podríamos concluir que el 4º año A y el 4º año C tienen igual mediana , pero las notas del 4a.C. tienen menor dispersión , es decir son mas homogéneas que las del 4º A.

Del mismo modo, podemos concluir que las notas del 4º B están mas dispersas.

OTRAS MEDIDAS DE DISPEERSION

DISPERSION ABSOLUTA: SE REFIERE A CUALQUIER VARIACION REAL DETERMINADA POR LA DESVIACION TIPICA.

DISPERSION RELATIVA= $\frac{DISPERSION.ABSOLUTA}{MEDIA}$

COEFICIENTE DE VARIACION: V= $\frac{S}{\overline{X}}$ **, S: desviación típica.**

REFERENCIA TIPIFICADAS.

Z= $\frac{X_i - \overline{X}}{S}$

X_i : Valor observado.

$\overline{X}$ **: MEDIA**

S: desviación típica

Desviación media (MD)

MD= $\frac{\sum(X - \overline{X})}{N}$

DESVIACION MEDIA PARA DATOS ORDENADOS POR FRECUENCIAS:

MD= $\frac{\sum f*(X_i-\overline{X})}{N}$

X_i : Marca de clase

$\overline{X}$ **: MEDIA**

f:frecuencias de marcas de clase.

N: $\sum f$

Ejemplo:

	X_i	$(x_i-\bar{x})$	F	F*$(x_i-\bar{x})$
60-62	**61**	**6.45**	**5**	**32.25**
63-65	**64**	**3.45**	**18**	**62.10**
66-68	**67**	**0.45**	**42**	**18.90**
69-71	**70**	**2.55**	**27**	**68.85**
72-74	**73**	**5.55**	**8**	**44.40**
			100	$\sum$226.50

MD= $\frac{\sum f*X_i-\overline{X}}{N}$

MD= $\frac{226.50}{100}=2.26$

PARA ESTE EJEMPLO LA DESVIACION TIPICA ES:

S= $\sqrt{\frac{\sum(X_i-\overline{X})^2}{N}}$

ES DECIR:

S= $\sqrt{\frac{(226.36)^2}{100}}=\sqrt{\frac{11965.36}{100}}=10.93$

Métricas estadísticas para datos a granel (EJEMPLO)

39.-

Consideremos los siguientes datos, que corresponden a las calificaciones de un alumno en la asignatura de algebra en una escala de 1 a 10.

NOTAS: 4 6 7 6 9 10

MEDIA DE LOS DATOS: $\overline{X}=\frac{4+6+7+6+9+10}{6}=\frac{42}{6}=7$

También se puede calcular la media por el criterio de la media supuesta.Esto, suponemos que la media es uno de los datos, ejemplo: $\overline{X}$ s =9.Calculamos luego los desvíos de cada dato respecto a la media supuesta.

Xi	***4***	***6***	***7***	***6***	***9***	***10***
Xi- $\overline{X}$ **s**	-5	-3	-2	-3	0	1

Suma de desvíos: $\sum_{n=1}^{n}(Xi-\overline{Xs})$ = -12. Por lo tanto $\overline{X}=9+\frac{-12}{6}=7$

Suma de errores:

Xi	4	6	7	6	9	10	
Xi- $\overline{X}$	-3	-1	0	-1	2	3	$\sum_{n=1}^{6}(Xi-\overline{X})=0$

Errores absolutos:

Xi	4	6	7	6	9	10	
$\lvert Xi-\overline{X}\rvert$	3	1	0	1	2	3	$\sum_{n=1}^{6}\lvert\ Xi-\overline{X}\ \rvert=10$

Radio de variación: RV = 1- $\frac{fm}{n}$ = 1- $\frac{2}{6}=1-0.33=0.67$

Desviación media: DM= $\frac{\sum_{n=1}^{6}\lvert Xi-\overline{X}\rvert}{6}=\frac{10}{6}=1.66$

Desviación estándar: Sx= $\sqrt{\frac{\sum_{n=1}^{6}(Xi)^2}{n}-\overline{X}^2}$

Sx= $\sqrt{\frac{4^2+6^2+7^2+6^2+9^2+10^2}{6}-7^2}=\sqrt{\frac{16+36+49+36+81+100}{6}-49}=\sqrt{\frac{318}{6}-49}=\sqrt{53-49}=2$

De otro modo:

Xi	4	6	7	6	9	10	
(Xi- $\overline{X}$) ^{2}	9	1	0	1	4	9	$\sum_{n=1}^{6}(Xi-\overline{X})^2=24$

Sx= $\sqrt{\frac{\sum_{n=1}^{6}(Xi-\overline{X})^2}{6}}$, Sx= $\sqrt{\frac{24}{6}}=\sqrt{4}=2$

Intercuartiles: Se ordenan los datos en forma creciente .4 6 6 7 9 10

Q_1: Primer cuartil: Corresponde al 25% de los datos.

Para el ejemplo: $\frac{25}{100}*6=1.5$ .Se toma el promedio entre el primer y segundo dato, esto es $Q_1=\frac{4+6}{2}=5$

Q_2 :Segundo cuartil . Corresponde al 50% de los datos y equivale a la mediana.

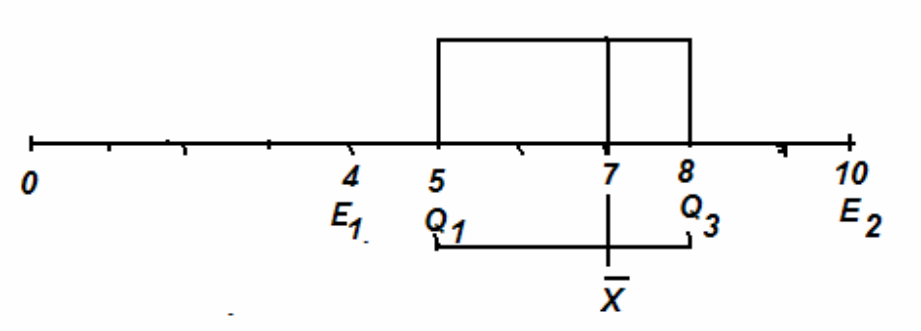
Para el ejemplo: $\frac{50}{100}*6=3$, lo que indica que se debe tomar el tercer dato, esto es : 6

Q_3 :Tercer cuartil . Corresponde al 75% de los datos

Para el ejemplo : $\frac{75}{100}*6=4.5$ .Se toma el promedio entre el cuarto y quinto dato . Esto es : $\frac{7+9}{2}=8$

Rango intercuartìlico: Corresponde a la diferencia entre el tercer y el primer cuartil . Esto es Q_3 - Q_1

Para el ejemplo : $Q_3 - Q_1 = 8-5=3$



MODA DE LOS DATOS A GRANEL (Mo) =6 . Corresponde al dato que mas se repite.

MEDIANA DE LOS DATOS A GRANEL: (Md) $= \frac{50}{100} * 6 = 3$ .Que corresponde a la posición que ocupa la mediana entre los datos previamente ordenados en forma creciente, es decir la tercera posición .Como en en este caso hay dos valores centrales, la mediana corresponderá al promedio entre los dos datos centrales. Esto es Md= $\frac{7+6}{2} = 6.5$

VARIABLES CUALITATIVAS

DATOS NOMINALES: Son simples categorías

De una muestra elegida aleatoriamente entre los apoderados del Liceo San Antonio, se recogen los siguientes datos:

Estado civil	Frecuencia (fi)	Frecuencia relativa (fr)	**Angulo central (α º)**
Casados	40	40/60	**240º**
Separados	12	12/60	**72º**
Solteros	4	4/60	**24º**
Viudos	4	4/60	**24º**
total	60	1	**360º**

De estos datos cualitativos, se puede establecer:

1**.- La media de los datos NO SE PUEDE CALCULAR**... ¿Qué sentido tiene determinar el promedio entre las categorías casados, separados, solteros y viudos?

2**.-La moda .Se puede determinar** y para el ejemplo resulta ser CASADOS (corresponde a la categoría que tiene la mayor frecuencia absoluta o relativa)

3.-**La mediana: NO SE PUEDE CALCULAR**, ¿Qué orden creciente seguiría UD? Para ordenar las categoría? Y luego ¿Cómo determina matemáticamente el valor central?

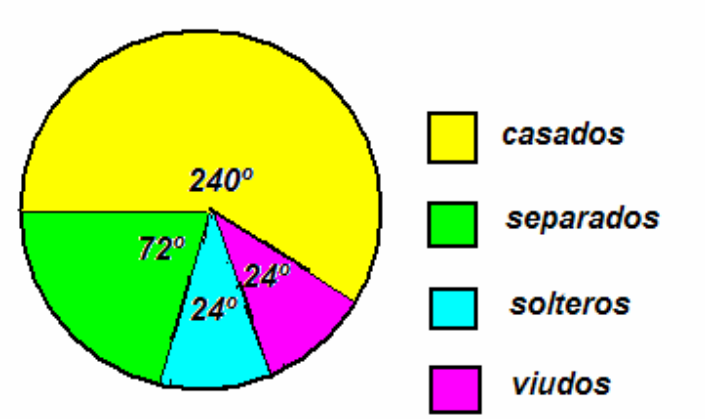
Cuando las categorías son cualitativas nominativas (de numero reducido) se recomienda representarlas en un grafico de torta (circular)

DEFINICION DE LOS ANGULOS CENTRALES PARA UN GRAFICO DE TORTAS.

$\alpha^{o} = 360^{o} * fr.$

Para la tabla: La categoría CASADOS: $\alpha^{o} = \frac{40}{60} * 360^{o} = 240^{o}$ .El resto de los ángulos centrales para las categorías restantes, están indicadas en la tabla.

El grafico circular o de torta quedaría:



DATOS CUALITATIVOS ORDINALES:

Recordemos que estos datos no son numéricos, pero se representan por números, se gradúan de acuerdo a un criterio de valoración.

Ejemplo: Las calificaciones finales de la asignatura de Religión del 4º A, del Liceo Parroquial San Antonio, se expresan en la tabla:

concepto	*Frecuencias absolutas*
MB	20
B	4
S	2
I	34
TOTAL	34

MEDIA: NO SE PUEDE CALCULAR
MEDIANA: NO SE PUEDE CALCULAR
MODA. En este caso corresponde al concepto MB, (mayor frecuencia)

También se recomienda un grafico de torta , un grafico de barras o un pictograma

Obs. : Elegir un grafico para representar datos muestrales es bastante relativo y depende fundamentalmente de los datos, de las características de los mismos y del criterio de quien estudia los datos.

El principio básico es que un grafico sea de fácil lectura, que muestre la información en forma apropiada, que se pueda VER fácilmente lo que se quiere informar.

Existen variados tipos de gráficos.

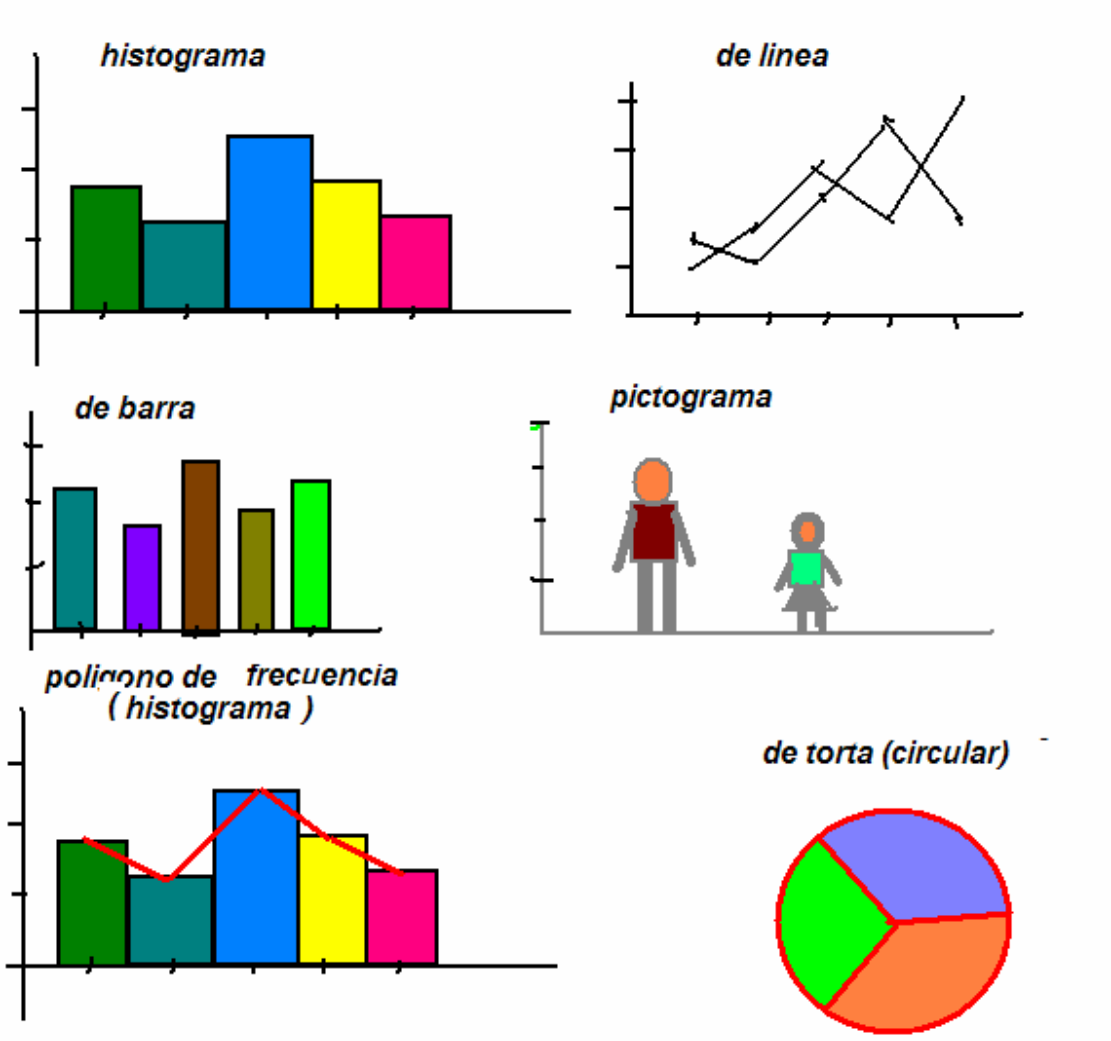
DE TORTA.: Circulares

DA BARRAS .En el eje OX, se ubican las categorías y en el eje OY las frecuencias (absolutas, relativas, porcentuales, acumuladas).Cada una de las categorías se representan por rectángulos del mismo ancho, con centro en las categorías del eje OX, y de alto las frecuencias respectivas en el eje OY.

DE LINEAS.El criterio de construcción es similar al de barra, pero se representan solamente los puntos y la alinea quebrada que los une.

PICTOGRAMAS.Utiliza figura para representar una categoría .Una figura suele representar una cierta cantidad prefijada .Son muy usados en los mapas de economía.

Algunos ejemplos de estos gráficos:



REGLAS GENERALES PARA FORMAR LAS DISTRIBUCIONES DE FRECUENCIAS DE DATOS CUANTITATIVOS TABULADOS (INTERVALARES) Y CALCULOS DE PARAMETROS ESTADISTICOS PARA ESTOS MODELOS.(EJEMPLO)

En general se aplican métodos aritméticos.

Los datos a granel se pueden representar mediante tablas, lo que se denomina representación tabular o datos intercalares.

En primer lugar y es lo más importante es establecer el número de clases o de intervalos que se van a usar en la representación tabularen este caso aplicaremos un modelo matemático ampliamente probado y refrendado por la experiencia empírica estadística: el modelo de STURGES, QUE ESTABLECE UN CRITERIO MATEMATICO PARA LA DETERMINACION DEL NÚMERO DE CLASES "K"
Cuya formula o modelo matemático es:
K= 1+3.3log N, donde:
K:N° de intervalos
N: N° de datos observados, o datos a granel

Por ejemplo: los siguientes datos a granel corresponden al estudio del crecimiento al cabo de una semana en cm. de algún tipo de alga marina a las cuales se le aplico un determinado tratamiento,

0.3	0.5	0.5	0.5	0.9	0.6	5.8	0.9	1.0	1.2	3.3	1.3	1.4	1.4	6.3	6.4	9.6
1.5	1.8	1.6	1.7	1.9	2.0	2.4	2.5	2.5	2.6	2.8	2.9	2.9	8.0	6.9	7.3	9.8
3.2	3.4	3.4	3.5	3.8	3.9	5.5	4.5	4.7	4.8	6.3	0.5	5.7	5.8	7.7	7.9	7.

PRIMER PASO:En este caso la cantidad de intervalos será: k= 1+ Log 51, pues el total de datos a granel son precisamente 51.

Luego k= 1+ 3,3*1.7075

K= 6,63 = 7 intervalos

SEGUNDO PASO: se determina la desviación máxima (Rg) de la muestra: que corresponde a la diferencia entre el mayor de los datos medidos y el menor de los mismos.

Para el ejemplo: valor mayor observado: 9.8

Valor menor observado: 0.3

Entonces la desviación máxima corresponde a: 9.5

TERCER PASO: se determina la unidad de medida (a.m.) de la muestra: corresponde al numero de decimales (d) en que se han medidos datos, se expresa por el modelo: $(\frac{1}{10})^d$ .para el ejemplo:

UM= $(\frac{1}{10})^1 = 0.1$

CUARTO PASO: se determina el recorrido de la muestra (RM), que se expresa por el modelo:

RM = Rg +UM, que para el caso:

RM = 9.5 + 0.1

RM = 9.6

QUINTO PASO: se determina la amplitud de los intervalos (representa el "ancho de cada clase"), se expresa por:

a = $\frac{RM}{K}$, que para el ejemplo:

a = $\frac{9.6}{7} = 1.4$

SEXTO PASO: se determina el recorrido de la tabla (Rt): que corresponde a: Rt= a*K, que para el ejemplo es:

Rt = 1.4*7 = 9.8

SEPTIMO PASO: se determina la diferencia de recorrido (Dr.): que se expresa por:

Dr= Rt – RM, que para el ejemplo corresponde a:

Dr = 9.8-9.6 = 0.2

OCTAVO PASO: se definen los limites superior e inferior como:

Limite.inferior = menor valor medido - $\frac{Dr}{2}$

Limites superior = mayor valor medido + $\frac{Dr}{2}$, entonces para el ejemplo.

Limite inferior = 0.3-0.1 =0.2

Limite superior = 9.8+0.1 = 9.9

NOVENO PASO: determinación de cada intervalo o clase: partiendo de la base que los intervalos son cerrados y que el limite inferior esta definido y que la amplitud de cada uno de los intervalos también lo esta, es fácil deducir cada uno de ellos:

El primero será: $[\lim ite * \inf rior - \lim ite * \inf erior + a]$, para el ejemplo;
Primera clase: $[0.2 - 0.2 + 1.4] = [0.2 - 1.6]$
Segunda clase: $[\lim.\inf + a + UM - (\lim.\inf + UM + 2a]$, para el ejemplo
Segunda clase: $[0.2 + 1.4 + 0.1 - 0.2 + 0.1 + 2 * 1.4]$
$[1.7 - 3.1]$ Y, axial sucesivamente hasta completar los siete intervalos
Para el ejemplo: $[3.2 - 4.6]$; $[4.7 - 6.1]$; $[6.2 - 7.6]$; $[7.7 - 9.1]$; $[9.2 - 106]$

Los que ordenados en una tabla, en las cuales incluiremos además de las clases, otros datos como las frecuencias, las marcas de clase que definiremos luego:

Clases	Marcas de Clases	fi	fr	Fr%
$[0.2 - 1.6]$	0.9	15	15/51	1500/51
$[1.7 - 3.1]$	2.4	11	11/51	1100/51
$[3.2 - 4.6]$	3.9	8	8/51	800/51
$[4.7 - 6.1]$	5.4	6	6/51	600/51
$[6.2 - 7.6]$	6.9	5	5/51	500/51
$[7.7 - 9.1]$	8.4	4	4/51	400/51
$[9.2 - 10.6]$	9.9	2	2/51	200/51
total		51	1	100

A partir de la tabla definiremos otros conceptos que involucra la estructura intervalar de datos:

Clase: cualquiera de los intervalos que definen la tabla: corresponden por definición a intervalos cerrados, son de la forma entonces $[a - b]$.
En la tabla uno de los intervalos puede ser: $[6.2 - 7.6]$

Limites de clase: corresponden a las cotas extremas de cada uno de los intervalos; en el intervalo: $[a - b]$, los limites de clase son: inferior: a; superior b

En la tabla y para la clase: $[6.2 - 7.6]$, limites: inferior: 6.2 y superior: 7.6

Limites reales: se definen

Límite real inferior: limite inferior - $\frac{UM}{2}$

Límite real superior: limite superior + $\frac{UM}{2}$

Para el ejemplo, considerando la clase: $[6.2 - 7.6]$

Limite real inferior: Lim.real inf.: 6.2 - $\frac{0.1}{2}$ = 6.2-0.05 =6.15

Limite real superior: Lim.real .sup. : $7.6+\frac{0.1}{2} = 7.6 + 0.05 = 7.65$

Un método práctico para determinar los límites reales es operar del siguiente modo:

Supongamos la clase hipotética: – $[2.135 - 3.458]$, entonces los valores que quedan, fuera del intervalo son:

$2.134\,[2.135 - 3.458]\,3.459$; calculando los promedios correspondiente en cada extremo:

Límite real inferior: $\frac{2.134 + 2.135}{2} = 2.1345$

Límite real superior: $\frac{3.458 + 3.459}{2} = 3.4585$

Amplitud de clase: corresponde al "ancho de la clase, y esta definido por :b-a
Para la tabla en cuestión y para la clase tomada como ejemplo, la amplitud será :7.6-6.2 =1.4

Marca de clase (Xi): corresponde al punto medio de la marca de clase.El modelo matemático equivale a:
Xi = $\frac{\sum \lim ites - de - clase}{2}$; en otras palabras corresponde al promedio de los limites de clase.Para tal efecto se pueden considerar para el calculo, indistintamente los limites de clase o limites reales de clase.
En general los límites de clase supone que todos los datos observados o medidos en la clase correspondiente se consideran como datos que tienen un valor equivalente precisamente a la marca de clase.

Para el ejemplo: consideremos la clase: $[6.2 - 7.6]$.la marca de clase será:

$Xi = \frac{6.2 + 7.6}{2} = \frac{13.8}{2} = 6.9$

O bien, si tomamos para el calculo los limites reales tendremos: $Xi = \frac{6.15 + 7.65}{2} = \frac{13.8}{2} = 6.9$

Es importante hacer notar el siguiente concepto : si observamos la frecuencia absoluta de esta clase en la tabla , vemos que es 5,significa que para efectos de cálculos o interpretación estadística los cinco valores del intervalo los consideramos de valores 6.9 ,aunque sabemos que estos valores son en realidad : 6.3 ; 6.4 ; 6.9 ; 7.3 ; 6.3

Frecuencia relativa en una distribución intervalar : corresponde a la frecuencia absoluta de cada clase en relación al total de los datos
En el ejemplo y para la clase tomada como ejemplo. $[6.2 - 7.6]$, la frecuencia relativa corresponde a 5/51

Frecuencia relativa porcentual: corresponde a porcentaje correspondiente al total de la frecuencia absoluta de cada clase.
En el ejemplo y para la clase tomada como ejemplo $[6.2 - 7.6]$, la frecuencia relativa es 500/51 =9.80%. Lo que significa que el 9.80% de los datos de la muestra se encuentran en esta clase.

Ahora ampliaremos la tabla anterior a la que agregaremos otras dos columnas que estudiaremos su significado:

clase	Fi	fr	Fr %	Fa	Fa%
$[0.2 - 1.6]$	15	15/51	29.41	15	29.41
$[1.7 - 3.1]$	11	11/51	21.56	26	50.97
$[3.2 - 4.6]$	8	8/51	15.68	34	66.65

$[4.7-6.1]$	6	6/51	11.76	40	78.41
$[6.2-7.6]$	5	5/51	9.80	45	88.21
$[7.7-9.1]$	4	4/51	7.84	49	96.05
$[9.2-10.6]$	2	2/51	3.95	51	100.00
total		1	100 %		

La cuarta columna (Fa) se define como frecuencia acumulada en una distribución intervalar.Se obtiene sumando la frecuencia de la clase inmediatamente anterior, por ejemplo para obtener la frecuencia acumulada de la 5° clase: $[6.2-7.6]$, que según la tabla es 45, debemos sumar las frecuencias absolutas de todas las las clases que antecede incluyendo la de la clase mencionada, esto es: (15+11+8+6+5 = 45). También se puede obtener sumando la frecuencia acumulada de la clase inmediatamente anterior con la frecuencia absoluta de la clase en cuestión, esto es: 40 + 5 = 45

El dato estadístico de la frecuencia acumulada es un parámetro muy importante en términos estadístico, ya con el podemos hacer una lectura rápida de información .supongamos que debemos responder a la interrogante ¿Cuántas mediciones están por debajo o son inferiores a 7.4? .podemos estimar rápidamente que el valor es 45 según lo indica la frecuencia acumulada de la clase donde se sitúa dicho valor

La última columna sugiere el mismo análisis, salvo que esta vez el parámetro se presenta en porcentaje respecto del total.

Observación: podemos agregar también la frecuencia acumulada relativa, que corresponde a la frecuencia acumulada respecto del total, dato que no se ha representado en la tabla, pero que sugiero que Ud. Puede construir sin contratiempo

REPRESENTACION GRAFICA DE DATOS INTERVALARES.

Histograma: consideremos los datos a granel, que corresponden a los puntajes de un ensayo PSU de MATEMÁTICA, EN EL TALLER del liceo parroquial san Antonio, en una muestra de 40 alumnos

485.56	785.24	556.23	589.26	564.23	654.63	645.25	598.78	605.23	712.56
689.24	648.23	706.49	546.68	485.62	543.62	609.23	546.89	685.69	694.21
586.36	708.62	679.56	754.23	654.23	684.56	684.85	692.64	602.35	634.82
602.03	608.96	564.25	524.65	555.56	498.23	568.74	589.39	625.64	689.30

Aplicando el modelo tabular tenemos:

1.- numero de clases: K = 1 + Log 40 = 6
2.-desviación máxima de la muestra: Rg = 785.24-485.56 = 299.68
3.-unidad de medida = 0.01
4.- recorrido de la muestra: 299.68+0.01 = 299.69

5.- amplitud de clase = $\frac{299.69}{6} = 49.95$
6.- recorrido de la tabla: 49.95*6 =299.70

7.-diferencia de recorrido: 299.70-299.69 = 0.01
8.- limite inferior de la primera clase: 485.56 - $\frac{0.01}{2} = 485.55$
9.-limite superior de la primera clase: 485.55+48.95 = 534.50

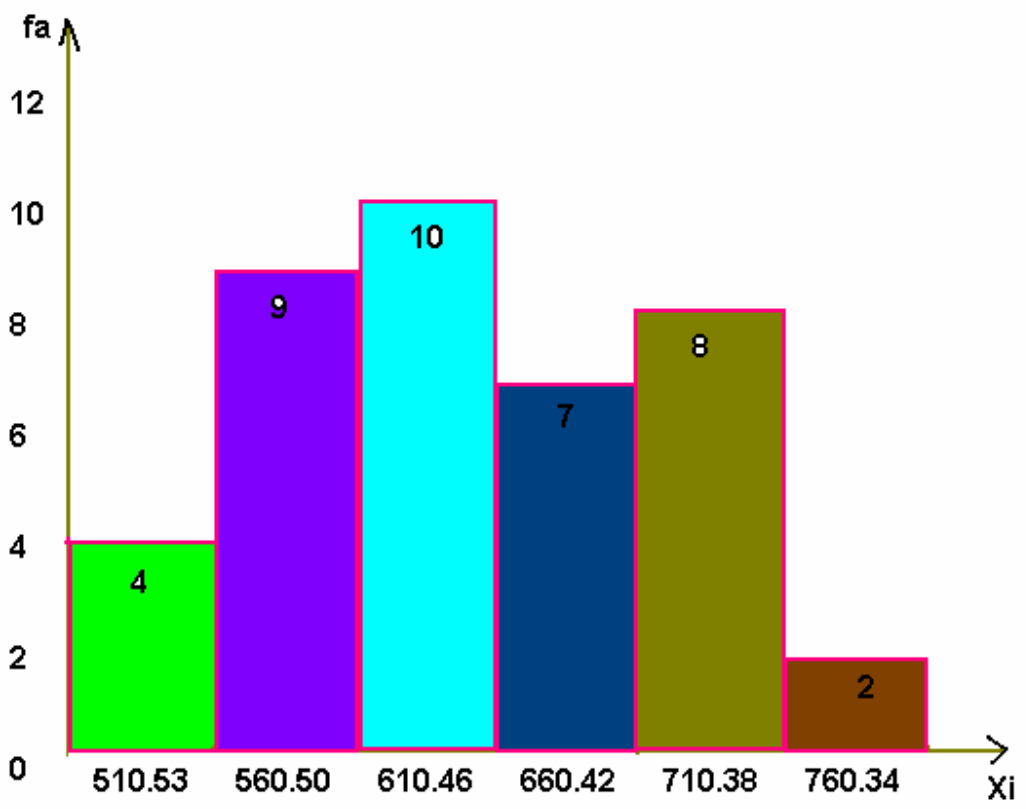
El resto se resume en la tabla:

Clases	Marca de clas	FREC absoluta.	Fi %	fa	Fa%
[485.55 – 535.51]	510.53	4	10.00	4	10.00
[535.52 – 585.47]	560.50	9	22.50	13	32.50
[585.48 – 635.43]	610.46	10	25.00	23	57.50
[635.44 – 685.39]	660.42	7	17.50	30	75.00
[685.40 – 735.35]	710.38	8	20.00	38	95.00
[735.36 – 785.31]	760.34	2	5.00	40	5.00
		40	100.00		

El histograma se obtiene haciendo un grafico Xi v/s fa. Es decir se ubican en el eje horizontal (ox), las marcas de clase, y en el eje vertical las frecuencias absolutas.
Se levantan luego las barras teniendo como centro la marca de clase correspondiente , de ancho equivalente a la amplitu de la clase y de alto la frecuencia absoluta definida en el eje vertical (oy).

Es importante que las barras tengan el mismo ancho, de lo contrario una lectura "visual" del grafico puede llevar a una interpretación equivocada sugerida por las superficies de estos rectángulos (dos rectángulos de distinta base e igual altura tienen naturalmente superficies distintas)

Para este ejemplo el **HISTOGRAMA** será:



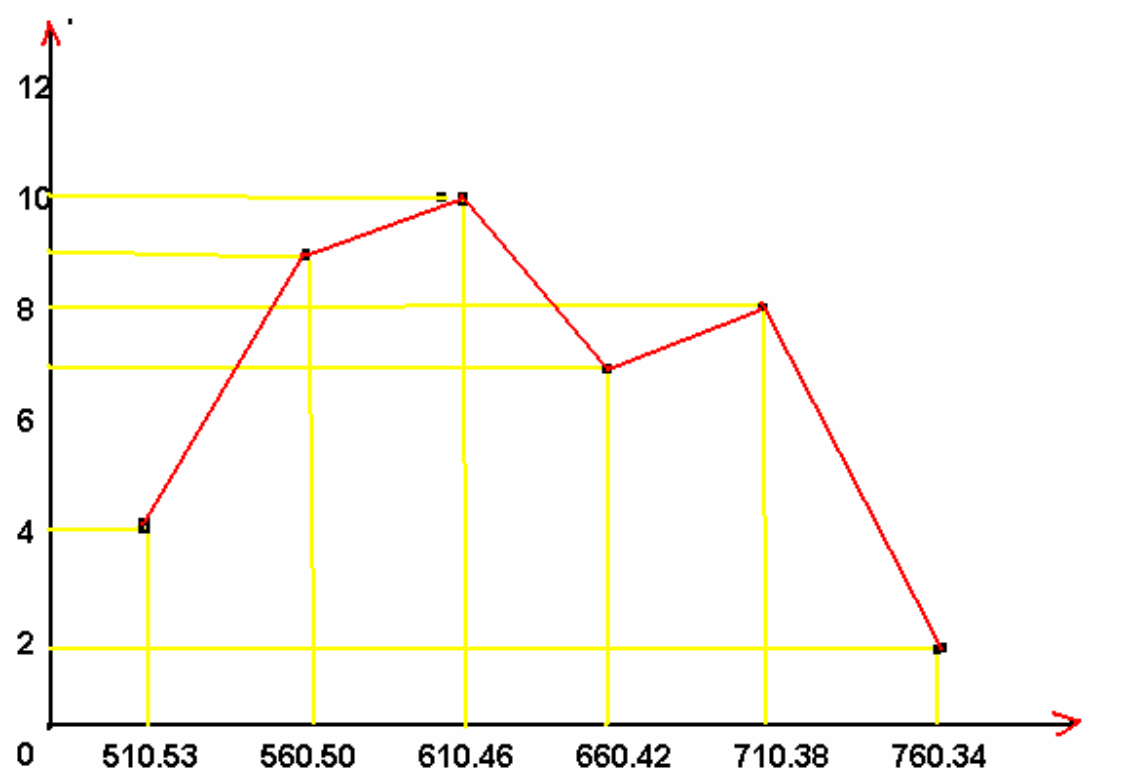
ES IMPORTANTE HACER NOTAR QUE EL HISTOGRAMA DIFIERE DEL GRAFICO DE BARRA, POR QUE EL PRIMERO ESTA RELACIONADO CON DATOS CUANTITATIVOS INTERVALARES Y EL SEGUNDO CON DATOS CUANTITATIVOS DE CARÁCTER ORDINAL

POR OTRO LADO EN UN GRAFICO DE BARA (CUALITATIVA – ORDINAL), PUEDEN ESTAR REPRESENTADAS EN UN MISMO GRAFICO MAS DE UNA VARIABLE, MIENTRAS QUE EN EL HISTOGRAMA (CUANTITATIVA – INTERVALAR) UNA SOLA VARIABLE...

EN UN GRAFICO DE BARRA (CUALITATIVA ORDINAL), SE PUEDE HACER SOLAMENTE UNA LECTURA DE LOS DATOS, MIENTRAS QUE EN EL HISTOGRAMA (CUANTITATIVA – INTERVALAR), SE PUEDEN HACER ITERPOLACIONES Y EXTRAPOLACIONES MATEMATICAS Y/O ALGEBRAICAS. CUALQUIER GRAFICO DEBE PERMITIR UN LECTURA VISUAL RAPIDA DE LOS DATOS, EN ESE SENTIDO LOS GRAFICOS TIENEN UN CONTENIDO HOLISTICO.

<u>POLIGONO DE FRECUENCIAS ABSOLUTAS</u>: Si se determinan los puntos medios de las bases superiores de los rectángulos del histograma y unimos estos puntos se obtiene el polígono de frecuencia.

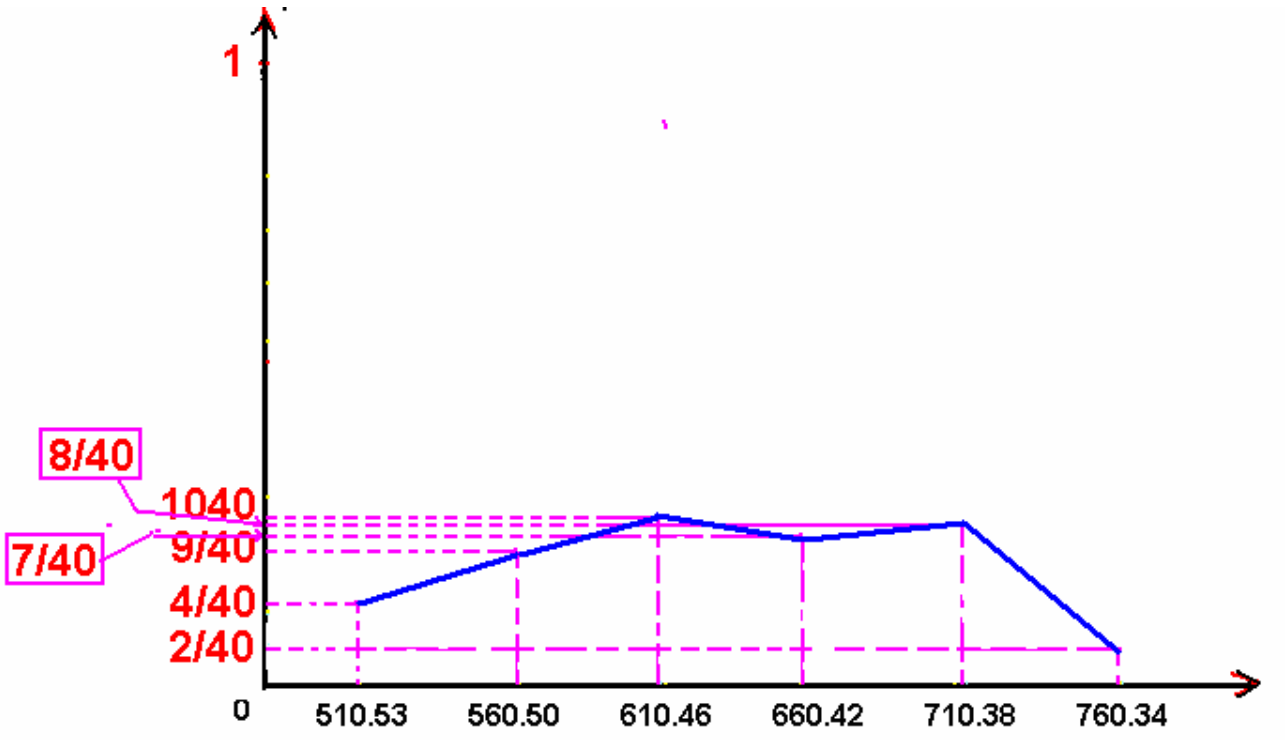
Ejemplo: para los datos del problema anterior:



Nota: el polígono de frecuencias representado mas arriba corresponde al polígono de frecuencias absolutas.
ES FACIL OBSERVAR EN UN POLIGONO DE FRECUENCIS ABSOLUTAS, QUE LOS PUNTOS O VERTICES MÁS ALTOS EN RELACION AL EJE HORIZONTAL CORRESPONDEN A AQUELAS MARCAS DE CLASE QU REPRESENTAN EL CONJUNTO DE DATOS QUE SE REPITEN CON MAYOR FRECUENCIA.

Polígono de frecuencias relativas: si en el eje horizontal disponemos las marcas de clase correspondientes a cada clase, y en el eje vertical los porcentajes de las frecuencias relativas con respecto al total, lo que se obtiene es el polígono de frecuencias relativas.

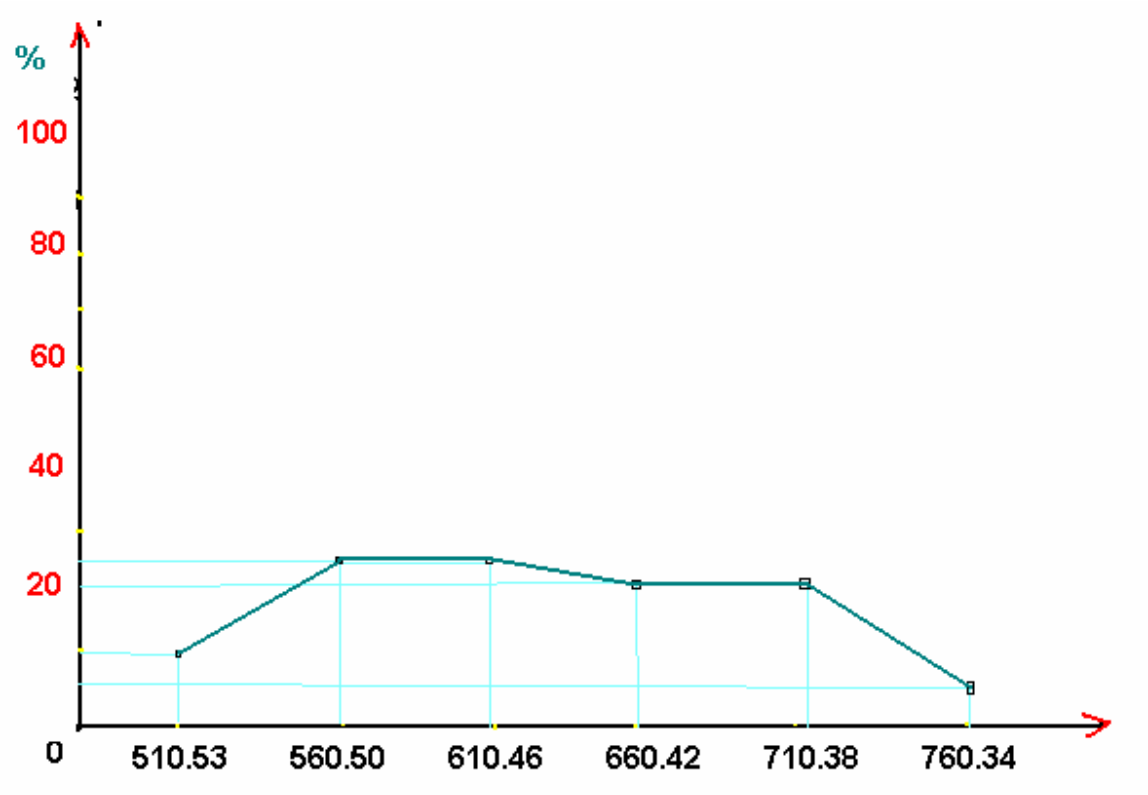
Montoya.-



EJEMPLO: El grafico corresponde a l polígono de frecuencias relativas de los datos intervalares del ensayo de la PSU, mencionado anteriormente.

LA LECTURA INMEDIATA QUE SE PUEDE HACER EN ESTE CASO ES QUE, LA MARCA DE CLASE EQUIVALENTE A 610.50 ES LA QUE CONTIENE UN MAYOR NUMERO DE DATOS RESPECTO DEL TOTAL

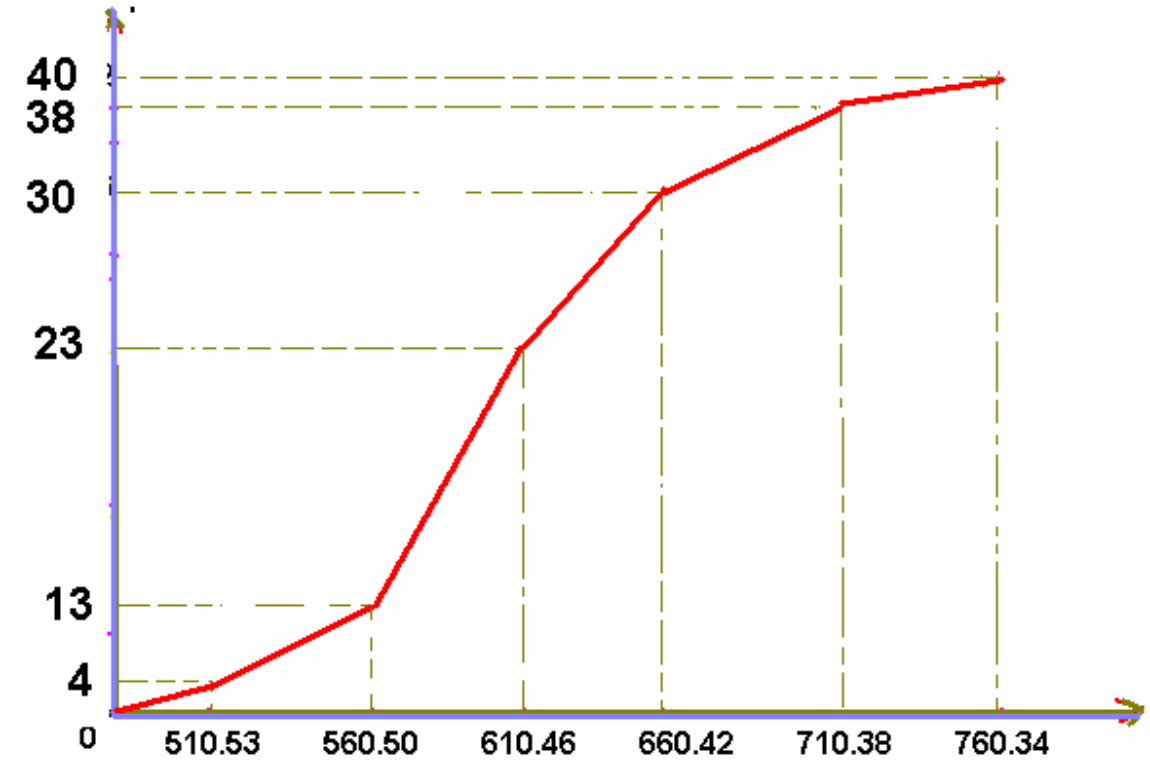
POLIGONOS DE FRECUENCIAS RELATIVAS PORCENTUALES: SI LAS FRECUENCIAS RELATIVAS SE EXPRESAN EN PORCENTAJES, SE OBTIENE EL GRAFICO QUE SE MUESTRA A CONTINUACION, QUE SE DENOMINA “POLIGONO DE FRECUENCIAS RELATIVAS PORCENTUALES.



AHORA LA LECTURA SE PUEDE HACER EN TERMINOS DE PORCENTUALES.

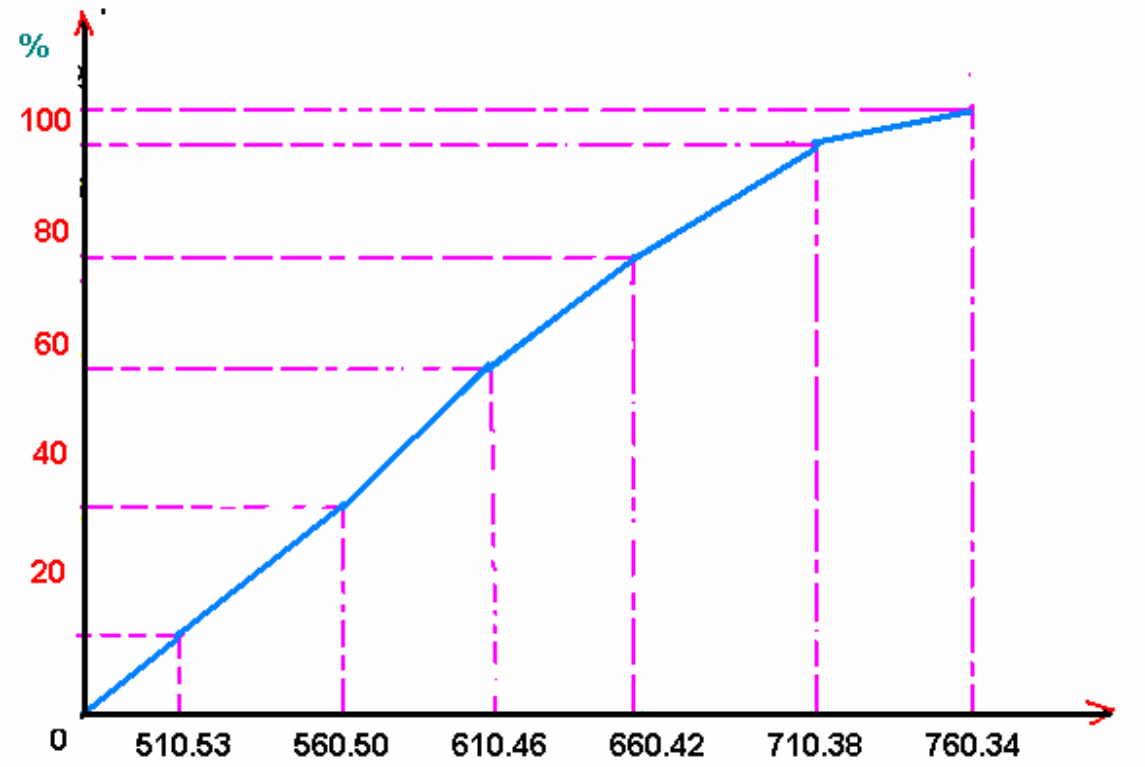
OJIVA. Corresponde al grafico de las frecuencias relativas acumuladas, Se obtiene disponiendo en el eje horizontal las marcas de clase y, en el eje vertical las frecuencias acumuladas a la clase correspondiente.

Para el ejemplo en estudio, la ojiva de frecuencias relativas acumuladas corresponde a:



LA OJIVA ES UN GRAFICO MUY ÚTIL EN TÉRMINOS ESTADÍSTICOS Y DE LECTURA RÁPIDA DE INFORMACION:
***PERMITE CALCULAR PARAMETROS ESTADISTICOS, QUE ESTUDIAREMOS MAS ADELANTE.**
***PERMITE ESTIMAR CON PRESICION QUE DATOS ESTAN POR DEBAJO O POR ENCIMA DE UN VALOR DETERMINADO.**
EJEMPLO: EN EL GRAFICO SE PUEDE "LEER", QUE 23 ALUMNOS OBTUVIERON UN PUNTAJE INFERIOR A 610.46

OJIVA PORCENTUAL PARA DATOS CUANTITATIVOS INTERVALARES: LA UNICA DIFERENCIA CON LA ANTERIOR ES QUE AHORA LAAS FRECUENCIAS ACUMULADAS SE EXPRESAN EN PORCENTAJES.
PARA EL EJEMPLO ANTERIOR QUEDARA.



AHORA LAS LECTURAS DE LOS DATOS SE PUEDEN HACER EN PORCENTAJES.
EJEMPLO: ALREDEDOR DEL 62% DE LOS DATOS ESTAN SOBRE LOS 660.42 PUNTOS.

CALCULO DE MEDIDAS DE POSICION (CENTRALIZACION) PARA DATOS CUANTITATIVOS AGRUPADOS O INTERVALARES.

MEDIA ARITMETICA (PROMEDIO)

CUANDO LOS DATOS SE PRESENTAN MEDIANTE UNA DISTRIBUCION DE FRECUENCIAS, TODOS LOS VALORES CAEN EN UN INTERALO DE CLASE COINCIDENTES CON LAS MARCAS DE CLASE.
RETOMEMOS LA TABLA DE MÁS ARRIBA:

	Marca de clas	FREC absoluta.	Fi %	fa	Fa%
Clases	510.53	4	10.00	4	10.00
[535.52 – 585.47]	560.50	9	22.50	13	32.50
[585.48 – 635.43]	610.46	10	25.00	23	57.50
[635.44 – 685.39]	660.42	7	17.50	30	75.00
[685.40 – 735.35]	710.38	8	20.00	38	95.00
[735.36 – 785.31]	760.34	2	5.00	40	5.00
		40	100.00		

LOS DATOS de la clase $[535.52-585.47]$ es decir los COMPRENDIDOS ENTRE: 535.52 Y 585.47, podemos considerarlos todos de valor único equivalente a la marca de clase 510.53. Entonces tendríamos 4 datos cuya frecuencia es 510.53

Luego entonces la media aritmética se puede determinar por el siguiente modelo matemático:

$$\overline{X} = \frac{\sum f * Xi}{N}$$, DONDE: f: frecuencias absolutas de clase.

s.f.: marcas de clase

N: numero total de datos.

Para el ejemplo, la tabla se puede disponer como se indica: (se agrega la columna f*Xi)

	Xi	f	f*Xi
clases	Marca de clase	FREC absol	
[485.55-535.51]	510.53	4	2042.12
[535.52 – 585.47]	560.50	9	5044.50
[585.48 – 635.43]	610.46	10	6104.60
[635.44 – 685.39]	660.42	7	4622.94
[685.40 – 735.35]	710.38	8	5683.04
[735.36 – 785.31]	760.34	2	1520.68
		40	
		25017.88	

$$\Rightarrow \bar{x} = \frac{25018.16}{40} = 625.45$$

De donde 625.45 corresponde a la media aritmética de los puntajes obtenidos por los 40 alumnos en el ensayo de matemáticas.

El cálculo se puede hacer también por el criterio de la media supuesta. En efecto

Procedimiento:
*se elige una marca de clase arbitraria, digamos A (QUE SERA LA MEDIA SUPUESTA, $\overline{X_s}$)
*SE DETERMINAN LAS DESVIACIONES, X_i - $\overline{X_s}$, DE CADA UNA DE LAS MARCAS DE CLASE CON ESTA MEDIA SUPUESTA
*LUEGO LA MEDIA SERA:

$$\overline{X} = \overline{X_s} + \frac{\sum f * (X_I - \overline{X_s})}{N}$$

Para el ejemplo:
Supongamos como media supuesta la marca de clase de la segunda clase, esto es: 560.50, luego la tabla con los cálculos correspondientes se pueden ordenar en forma simplificada como sigue:

MARCA DE CLASE	DESVIACION: $X_i - X_s$	F*(Xi-Xs)
510.53	510.53-560.50=-49.97	**4*49.97 = -199.98**
560.50	560.50-560.50= 0	**9*0 = 0**
610.46	610.46-560.50= 49.96	**10*49.96 =499.60**
660.42	660.42-560.50= 99.92	**7*99.92 = 699.44**
710.38	710.38-560.50=149.88	**8*149.88 = 1199.04**
760.34	760.34-560.50=199.84	**2*199.84 = 399.68**
		$\sum$ **f*(Xi-Xs)=2597.78**

Entonces la media quedara expresada por:

$$\overline{X} = 560.50 + \frac{2597.78}{40}$$

$$\overline{X} = 625.44.$$

VALOR QUE COINCIDE CON EL CALCULADO ANTERIORMENTE.

Observación: El cálculo de la media se puede hacer considerando como media supuesta cualquiera de las marcas de clases, obteniéndose el mismo valor para la media de los datos intervalares.

Un caso especial es cuando cada una de las desviaciones son múltiplos del intervalo de clase "C" (siempre y cuando todas las clases tengan la misma amplitud), entonces se puede dividir esta diferencia por este múltiplo "C".

EJEMPLO: Consideremos la siguiente tabla:

Clases	Xi	fi	Xi-Xs	(Xi-Xs)/C	Fi*(Xi-Xs)/C
60-62	61	5	-6	-2	-10
63-65	64	18	-3	-1	-18
66-68	(67)	42	0	0	0
69-71	70	27	3	1	27
72-74	73	8	6	2	16

$\sum$ =15

ENTONCES LA MEDIA PARA ESTE CASO SERA: $\overline{X} = \overline{X_s} + \frac{\sum fi*(Xi - Xs)/C}{N}$ *C

$$\overline{X} = 67 + \frac{15}{100} * 3$$

$$\overline{X} = 67.45$$

Valor de la media que coincide, si calculamos del modo que sigue:

$$\overline{X} = \frac{61*5 + 64*18 + 67*42 + 70*27 + 73*8}{100} = 67.45$$

La mediana: la mediana identifica el valor central de los valores provenientes de una muestra. La Mediana es entonces una medida de centralidad

clases	Marcas de Clases	fi	fr	Fr%
$[0.2-1.6]$	0.9	15	15/51	1500/51
$[1.7-3.1]$	2.4	11	11/51	1100/51
$[3.2-4.6]$	3.9	8	8/51	800/51
	5.4	6	6/51	600/51
$[6.2-7.6]$	6.9	5	5/51	500/51
$[7.7-9.1]$	8.4	4	4/51	400/51
$[9.2-10.6]$	9.9	2	2/51	200/51
total		51	1	100

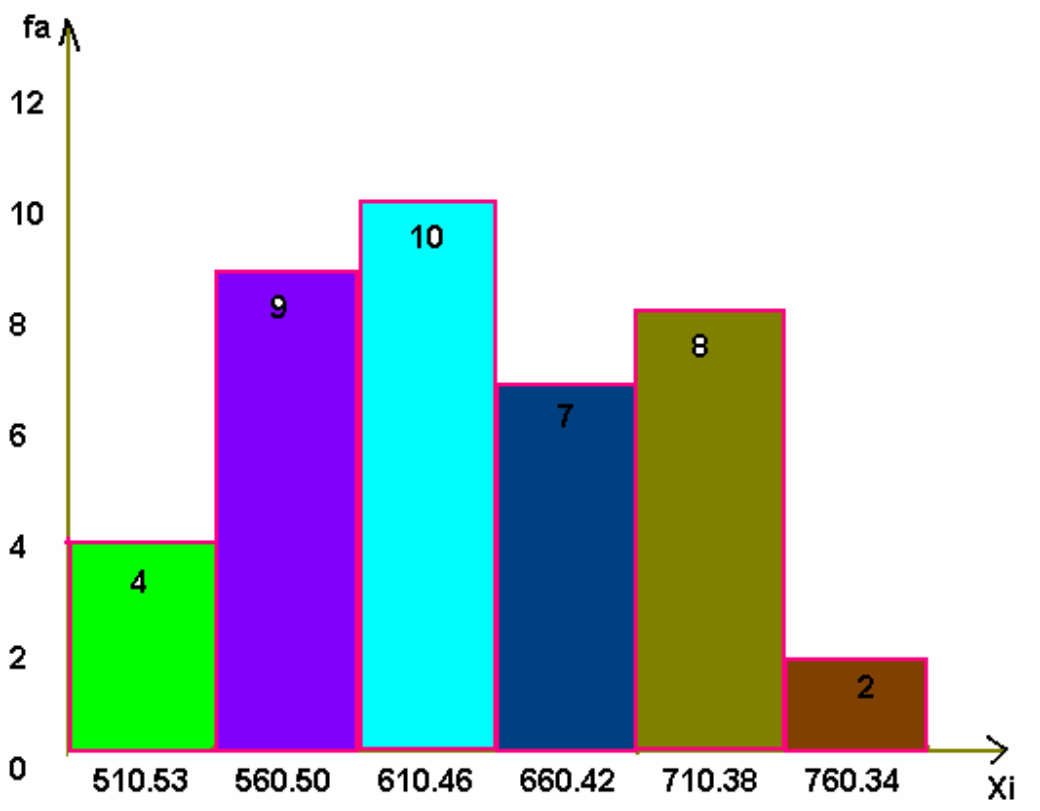
Construiremos en primer lugar un histograma, que corresponde a un grafico de barras, en el cual las marcas de clase se sitúan en el eje $\xrightarrow[ox]{}$ y las frecuencias de las marcas de clase correspondientes en el eje $\xrightarrow[oy]{}$

Para este efecto consideraremos los 40 datos tabulados de los puntajes obtenidos por los alumnos del liceo san Antonio en el ensayo de matemáticas:

Cuyo HISTOGRAMA es:

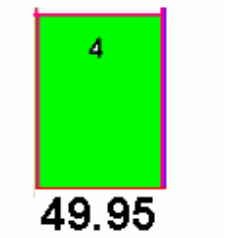
El histograma se obtiene haciendo un grafico Xi v/s fa . Es decir se ubican en el eje horizontal (ox), las marcas de clase, y en el eje vertical las frecuencias absolutas.
Se levantan luego las barras teniendo como centro la marca de clase correspondiente , de ancho equivalente a la amplitu de la clase y de alto la frecuencia absoluta definida en el eje vertical (oy).

Es importante que las barras tengan el mismo ancho, de lo contrario una lectura “visual” del grafico puede llevar a una interpretación equivocada sugerida por las superficies de estos rectángulos (dos rectángulos de distinta base e igual altura tienen naturalmente superficies distintas)



Ahora razonamos del siguiente modo: el 50 % de los datos corresponde a $\frac{50}{100}$ *40 = 20.

Si consideramos cada rectángulo cuya medida de la base será la amplitud de clase y cuya altura será la frecuencia absoluta de clase, esto es, para el primer rectángulo será:



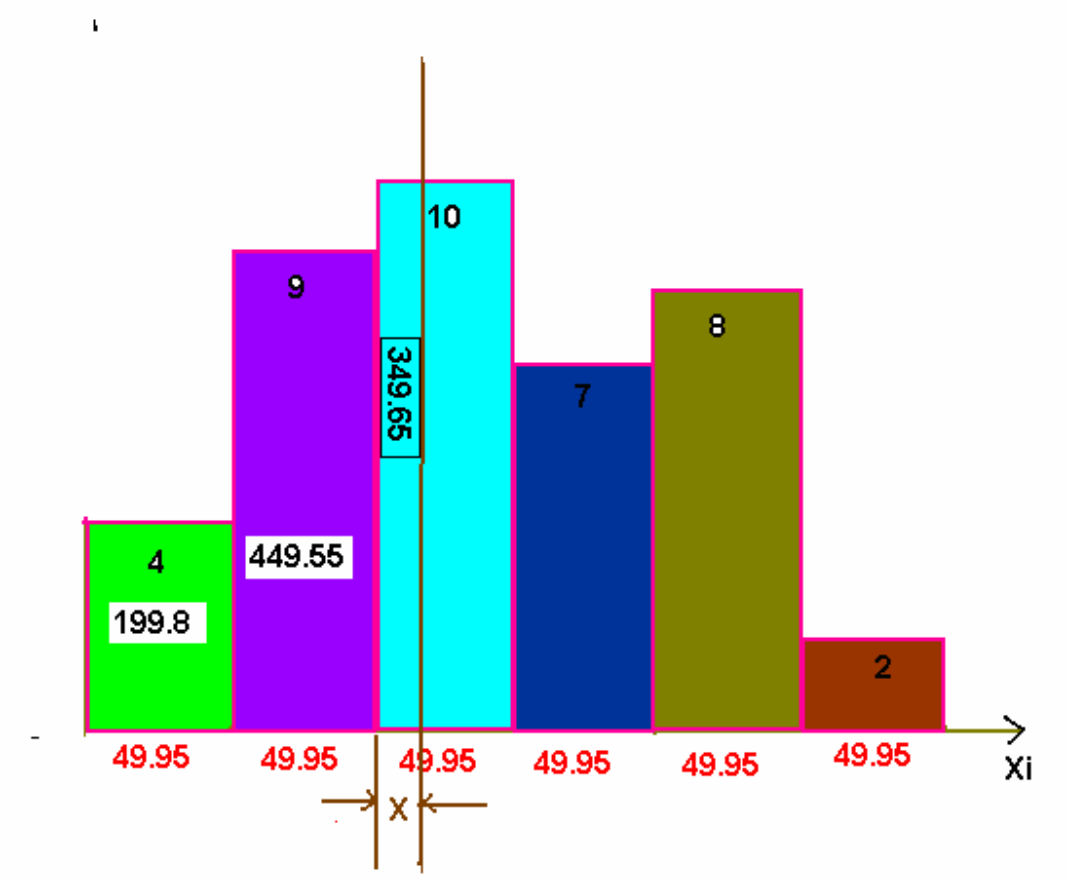
Cuya área corresponde a: 199.8 (unidades)^2 .

Ahora bien, por el momento no nos ocuparemos de las unidades, sino del hecho de que la mediana corresponde al valor de la abscisa que deja la mitad de la superficie de los rectángulos de cada lado.

Calculando la superficie total del polígono de frecuencia tenemos:

St = 49.95 (4+9+10+7+8+2) = 1998, cuya mitad corresponde a: 999.

Ahora bien, para los dos primeros rectángulos la suma de las superficies es: S (1y2) = 49.95 (4+9) = 649.35 COMO EL 50% DE LOS DATOS ES 20 Y YA HEMOS CONSIDERADO 13 DE ELLOS, LA DIFERENCIA 349.65 DE SUPERFICIE CORRESPONDIENTE AL 50% DEL TOTAL QUE ES 999, LO DEDUCIMOS DEL TERCER RECTANGULO:



Es decir X*10 = 349.65, DE DONDE X= 34.965,

Como el límite real inferior de la clase mediana es:

585.47 [585.48 – 635.43] **635.44**

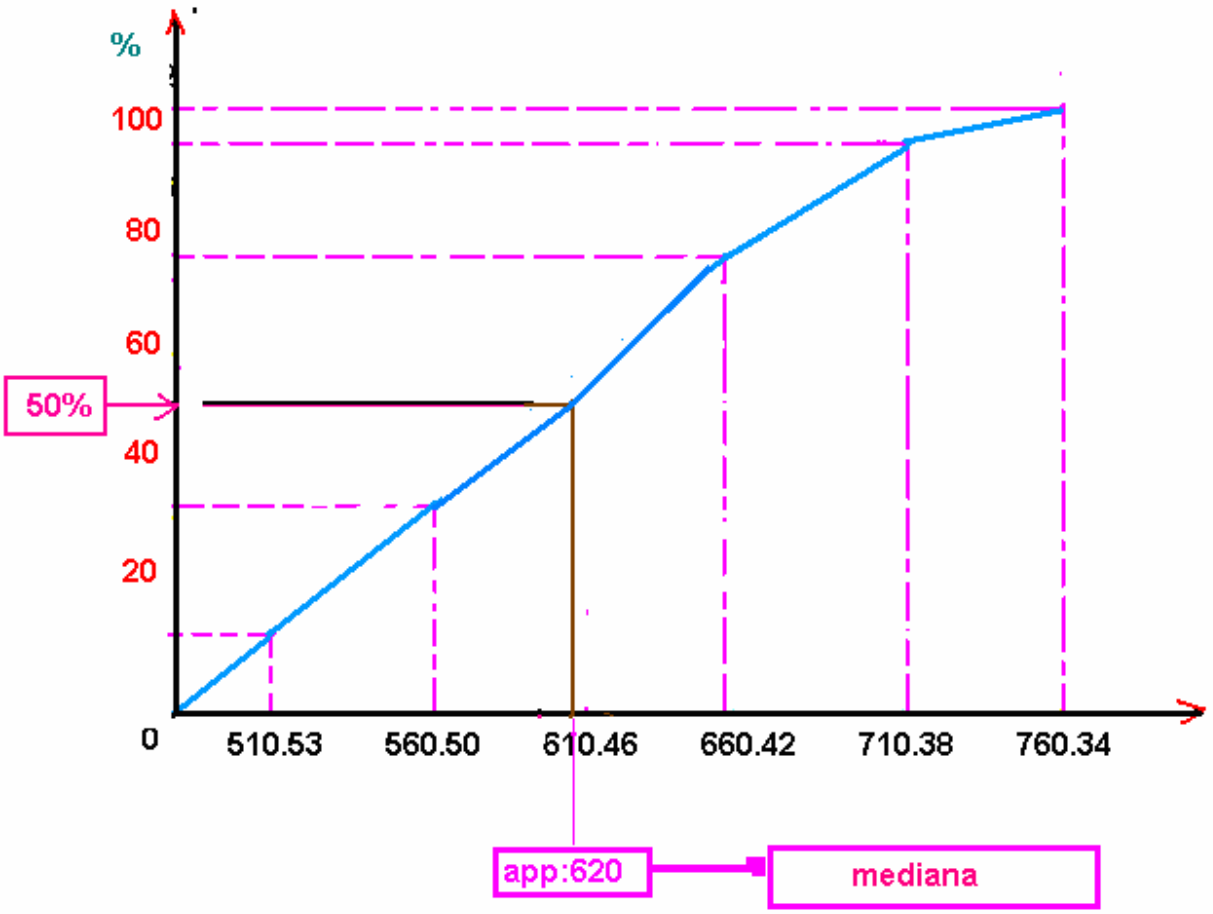
,

Lrs= $\frac{585.47+585.48}{2}$ =585.475

Luego la abscisa correspondiente al punto que deja la mitad de la superficie total a cada lado es: 585.475+34.965 =620.44 este valor entonces corresponde a la mediana.

Una forma grafica de determinar la mediana es a través de la ojiva porcentual: en efecto trazamos la recta perpendicular correspondiente al 50 % (línea negra), que interfecta a la curva ojiva en un punto "M", bajamos la recta perpendicular al eje de las marcas de clase y hacemos una interpolación aproximada.

Veámoslo en forma práctica:



Vera ud. Que este procedimiento nos da un valor estimado bastante próximo al valor real, por lo que puede considerarse un método relativamente confiable.
Ahora si nos armamos de paciencia y ordenamos los 40 datos de menor a mayor y determinamos de este modo el valor central tenemos obtendríamos exactamente el valor de este parámetro.Aqui le presento los datos y le desafío a que considere el calculo en forma personal, porque yo ya no estoy para esos jueguito….

485.56	785.24	556.23	589.26	564.23	654.63	645.25	598.78	605.23	712.56
689.24	648.23	706.49	546.68	485.62	543.62	609.23	546.89	685.69	694.21
586.36	708.62	679.56	754.23	654.23	684.56	684.85	692.64	602.35	634.82
602.03	608.96	564.25	524.65	555.56	498.23	568.74	589.39	625.64	689.30

Ahora estableceremos un criterio tabular (a partir de la tabla) para determinar la mediana directamente de datos intervalares.
Para el ejemplo anterior, cuya tabla es.

[485.55 – 535.51]	510.53	4	10.00	4	10.00
[535.52 – 585.47]	560.50	9	22.50	13	32.50
[585.48 – 635.43]	610.46	10	25.00	23	57.50
[635.44 – 685.39]	660.42	7	17.50	30	75.00
[685.40 – 735.35]	710.38	8	20.00	38	95.00
[735.36 – 785.31]	760.34	2	5.00	40	5.00
		40	100.00		

Determinamos la frecuencia total, es decir el número de datos medidos, que en este caso es 40.
Determinamos luego el 50% es decir la mitad, en este caso 20.
La suma de las frecuencias de las clases más cercanas a este valor (50%), en este caso es: 4+9 = 13.
Calculamos la diferencia entre la mitad y esta suma, en este caso: 20-13 = 7.
Determinamos luego los límites reales donde estaría por deducción la mediana, lo que se llama "la clase mediana", en este caso: la clase mediana es

585.47	[585.48 – 635.43]	635.44	610.46	10	25.00	23	57.50

Limite real inferior = $\frac{585.47 + 585.48}{2} = 585.475$

Limite real superior= $\frac{635.43 + 635.44}{2} = 635.435$

Luego se deduce que la mediana estará comprendida entre estos valores extremos o limites reales.
Por lo tanto la mediana corresponderá al límite real inferior más 7 de las 10 de la siguiente clase, esto es:

Md =limite real inferior + $\frac{7}{10} * [635.435 - 585.475]$

Md=585.475+34.972
Md=620.447
Md=620.45

Por el método de las áreas de los rectángulos del polígono de frecuencia el valor obtenido es de 620.44, valor que coincide casi exactamente con el determinado por este nuevo modelo...
En general para datos intervalares la mediana se determina por la formula:

$$Md = L_i + \frac{(\frac{N}{2} - \sum f_{\inf}) * C}{f_{clase-mediana}}, \text{ donde:}$$

L_i : límite real inferior de la clase mediana.

N: NUMERO DE DATOS

$\sum f_{\inf}$: suma de las frecuencias de todas las clases inferiores a la clase mediana.

C: tamaño del intervalo de clase.

$f_{clase-mediana}$: frecuencia de la clase mediana.

LA MODA PARA DATOS INTERVALARES:

Como ya sabemos la moda corresponde al valor que mas se repite, es decir aque que tiene mayor frecuencia. El modelo intervalar para determinar la moda es un modelo de aproximación , pues en una clase modal , es decir aquella de mayor frecuencia , no es posible determinar por observación de la tabla cual es ese valor , pero si es valido estadísticamente suponer que el dato esta ahí , aunque esto no ocurre con absoluta certeza.Para tener la seguridad que la determinación de la moda es el valor de mayor frecuencia de la muestra , tendríamos que tener la información de todos los datos y aplicar un rudimentario método de conteo.

De igual forma se define un modelo matemático intercalar adecuado:

EJEMPLO :Consideremos los datos de la tabla :

Clases	Xi	FREC absoluta.	Fi %	fa	Fa%
[485.55 – 535.51]	510.53	4	10.00	4	10.00
[535.52 – 585.47]	560.50	9	22.50	13	32.50
[585.48 – 635.43]	610.46	10	25.00	23	57.50
[635.44 – 685.39]	660.42	7	17.50	30	75.00
[685.40 – 735.35]	710.38	8	20.00	38	95.00
[735.36 – 785.31]	760.34	2	5.00	40	5.00
		40	100.00		

Aquí vemos que lo mas probable es que la moda se encuentre en la 3 º clase, esto es porque tiene mayor frecuencia absoluta.Esta clase se denomina clase modal

El modelo matemático se establece según:

$$Mo = \text{Limite real inferior de la clase modal} + \frac{\Delta_s}{\Delta_i + \Delta_s} * C$$

Donde: L_{ri} : Límite real inferior de la clase modal

Δ_s : Exceso de la clase modal sobre la frecuencia de la clase contigua superior.

Δ_i : Exceso de la clase modal sobre la frecuencia de la clase contigua inferior.

C: Amplitud de clase.

Simbólicamente.

$$Mo = Lri + \frac{\Delta_s}{\Delta_i + \Delta_s} * C$$

Para l emplo: Clase modal :3º clase

Frecuencia modal: 10

Frecuencia clase superior: 7, exceso superior: 3

Frecuencia clase inferior: 9, exceso inferior : 1.

$$Lri = \frac{585.48 + 585.47}{2} = 585.475$$

Amplitud de clase: 635.435-585.475 = 49.96

Remplazando los valores y calculando, se tiene:

$$Mo = 585.475 + \left(\frac{3}{3+1}\right) * 49.96 = 585.475 + 37.470 = 622.945$$

Mo = 622.945 , que es un estimativo del valor que mas se repite en los datos. Montoya.-

FORMULARIO DE ESTADISTICA:

DATOS A GRANEL:

MEDIA: $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$

MEDIA SUPUESTA : $\overline{X} = \overline{X_S} + \frac{\sum_{i=1}^{n} d_j}{n}$, d: DESVIOS.

MEDIA PONDERADA : $\overline{X} = \frac{\sum_{n=1}^{n} Xi * Wi}{\sum_{n=1}^{n} Wi}$

MEDIA DE MEDIAS : $\overline{X} = \frac{\sum_{n=1}^{n} \overline{X}_i * f_i}{\sum_{n=1}^{n} f_i}$

ERRORES (E): $\overline{X} - X_i$, $\sum_{n=1}^{N} \left(\overline{X} - X_i\right) = 0$

ERRORES ABSOLUTOS : $Ea = \left|\overline{X} - X_i\right|$, $\sum_{n=1}^{n} Ea = \sum_{n=1}^{n} \left|\overline{X} - X_i\right| = 0$

RANGO DE MEDICION :R = Es - Ei , Es Y Ei :VALORES EXTREMOS SUPERIOR E INFERIOR.

MEDIANA : UNA VEZ ORDENADOS : SI n ES PAR : $Md = \frac{\left(\frac{n}{2}+1\right)+\left(\frac{n}{2}-1\right)}{2}$

SI n ES IMPAR $Md = \frac{n}{2}$

MODA : CORREPONDE AL DATO DE MAYOR FRECUENCIA

MEDIA GEOMETRICA : SI :X_1, X_2 ……X_n , n Datos. Entonces “G”

$G = \sqrt[n]{X_1 * X_2 * .....X_n}$

MEDIA ARMONICA **(H):**

SI :X_1, X_2 ……X_n , n Datos. Entonces

$H = \frac{1}{\frac{1}{X_1} + \frac{1}{X_2} + ... + \frac{1}{X_n}}$

Relación empírica entre mediana, media y moda

Media –moda = 3(media-mediana)

$\overline{X}$ -Mod(X) = 3($\overline{X}$ -Med)

RAIZ CUADRADA DEL CUADRADO DE LA MEDIA : R.M.S = $\sqrt{\frac{\sum_{n=1}^{n} X_i^2}{}}$

Radio de variación: RV = 1- $\frac{fm}{n}$ =

Desviación media: DM= $\frac{\sum_{n=1}^{6} \left| Xi - \overline{X} \right|}{6}$

Desviación estándar: Sx= $\sqrt{\frac{\sum_{n=1}^{6} (Xi)^2}{n} - \overline{X}^2}$

De otro modo: Sx= $\sqrt{\frac{\sum_{n=1}^{6} (Xi - \overline{X})^2}{6}}$

VARIANZA : $Sx^2 = \frac{\sum_{n=1}^{n} \left(X_i - \overline{X} \right)^2}{n}$

VARIANZA COMBINADA PARA UN CONJUNTO DE DATOS CON VARIANZAS : $S_1^2 y S_2^2$ Y CON UNA MEDIA COMUN

$\overline{X}$ Y CON FRECUENCIAS $N_1 y N_2$ $S_{\frac{2}{C}} = \frac{N_1 * S_1^2 + N_2 * S_2^2}{N_1 + N_2} *$

RELACION EMPIRICA ENTRE LAS MEDIDAS DE DISPERSION:

D.M = $\frac{4}{5}$ *DESVIACION TIPICA.

RANGO SEMI INTERCUARTILICO : $\frac{2}{3}$ *DESVIACION TIPICA

RANGO INTERCUARTILICO : $Q_3 - Q_1$

RANGO SEMI-INTERCUARTILICO :($Q_3 - Q_1$)/2

RANGO PERCENTILICO :P(90) – P(10)

DISPERSION RELATIVA = $\frac{DISPERSION, RELATIVA, ABSOLUTA}{\overline{X}}$

PARA DATOS TABULADOS (DATOS INTERVALARES)

MEDIA : $\overline{X} = \frac{\sum f * Xi}{N}$, DONDE: f: frecuencias absolutas de clase.

Xi : MARCA DE CLASE

N: NUMERO TOTAL DE DATOS.

MEDIA SUPUESTA : $\overline{X} = \overline{X_s} + \frac{\sum f * (X_I - \overline{X_s})}{N}$

MEDIANA :Md= $L_i + \frac{(\frac{N}{2} - \sum f_{\inf}) * C}{f_{clase-mediana}}$, donde:

L_i : límite real inferior de la clase mediana.

N: NUMERO DE DATOS

$\sum f_{inf}$: suma de las frecuencias de todas las clases inferiores a la clase mediana.

C: tamaño del intervalo de clase.

$f_{clase-mediana}$: frecuencia de la clase mediana.

MODA : Mo = Mo=Lri + $\frac{\Delta_s}{\Delta_i + \Delta_s} * C$

Limite real inferior de la clase modal+ $\frac{\Delta_s}{\Delta_i + \Delta_s} * C$

Donde: L_{ri} : Límite real inferior de la clase modal

Δ_s : Exceso de la clase modal sobre la frecuencia de la clase contigua superior.

Δ_i : Exceso de la clase modal sobre la frecuencia de la clase contigua inferior.

C: Amplitud de clase.

Simbólicamente.

CRITERIO DE ORGANIZACIÓN DE DATOS TABULADOS :

PRIMER PASO: k= 1+ Log n , n NUMERO DE DATOS .

SEGUNDO PASO: (Rg) = Es - Ei

TERCER PASO: $(\frac{1}{10})^d$. d : NUMERO DE DECIMALES DE LOS DATOS

CUARTO PASO: (RM) = RM = Rg +UM

QUINTO PASO: amplitud de los intervalos (representa el "ancho de cada clase") a = $\frac{RM}{K}$

SEXTO PASO: se determina el recorrido de la tabla (Rt): que corresponde a: Rt= a*K

SEPTIMO PASO: diferencia de recorrido (Dr.)=Dr= Rt – RM

OCTAVO PASO: se definen los límites superior e inferior como:

Limite.inferior = menor valor medido - $\frac{Dr}{2}$

Limites superior = mayor valor medido + $\frac{Dr}{2}$

NOVENO PASO: determinación de cada intervalo o clase: partiendo de la base que los intervalos son cerrados y que el limite inferior esta definido y que la amplitud de cada uno de los intervalos también lo esta, es fácil deducir cada uno de ellos:

El primero será: $[\lim ite * \inf rior - \lim ite * \inf erior + a]$

Segunda clase: $[\lim.\inf + a + UM - (\lim.\inf + UM + 2a]$