Universidad Nacional del Litoral
Facultad de Ingeniería y Ciencias Hídricas

ESTADÍSTICA

Ingeniería Informática

TEORÍA

Mg.Ing. Susana Vanlesberg
Profesor Titular

UNIDAD 5
ESTADÍSTICA DESCRIPTIVA Y ANÁLISIS EXPLORATORIO

Estadística Descriptiva

- Se utiliza cuando los resultados del análisis estadístico no pretende ir más allá del conjunto de datos investigados.
- Describe numéricamente, analiza y representa un conjunto de datos ordenados mediante la utilización de métodos numéricos, tablas y gráficas, simplificando y resumiendo la información.

Una vez que hemos recogido los datos, tenemos que:

1. Confeccionar tablas acompañadas de gráficos para una mejor visión de los datos.
2. Hacer una recopilación y reducción de dichos datos a unas pocas medidas representativas.
3. **Interpretar los resultados y obtener conclusiones para predecir y tomar decisiones estadísticas. Campo de la Estadística Inferencial.**

Podríamos pensar en analizar a todos los individuos de la población. Sin embargo, esto puede ser inviable por su costo o por el tiempo que requiere. Entonces nos conformamos con extraer una muestra. La muestra proporciona información sobre el objeto de estudio. Lo habitual en nuestro contexto es que en el procedimiento de extracción intervenga el azar.

Por Ejemplo: Se quiere analizar el número de horas de estudio semanal que dedican los estudiantes de Ingeniería en Informática de esta Facultad. Para ello se consulta a 50 alumnos de esta ingeniería.

Población: Todos los estudiantes de Ingeniería en Informática de esta Facultad.
Variable: Número de horas de estudio semanal.
Muestra: 50 alumnos encuestados.

Más formalmente se denomina muestra de una población original con función F(x), a la sucesión $x_1\ x_2\ \ldots\ x_n$ de los valores observables de la variable aleatoria x, que corresponden a n repeticiones independientes de un experimento aleatorio. Se define de forma análoga a la muestra en el caso que el experimento aleatorio esté relacionado con varias variables aleatorias (variable bidimensional, por ejemplo).

Uno de los tipos de muestreo más utilizado es el *muestreo aleatorio simple* (m.a.s.) en el que cada individuo de la población tiene la misma probabilidad de ser incluido en la muestra.

No siempre es necesario tomar una muestra, ya que si queremos estudiar el fracaso de un curso determinado, deberíamos analizar todos los alumnos de dicho curso, y no una muestra de ellos.

Antes de hacer un estudio estadístico, tenemos que plantearnos bien que problema vamos a estudiar, y cuáles van a ser los objetivos de nuestra investigación o trabajo, fijando los pasos a seguir, las clasificaciones que se van a realizar, las variables que debemos observar y cómo medirlas, los gráficos que vamos a realizar.

Los datos que constituyen la muestra son llamados observaciones porque representan lo que se observa

actualmente.

Caracteres estadísticos: es una propiedad que permite clasificar a los individuos de una población. Se distinguen dos tipos:

a) Cualitativos. Son aquellos cuya variación se recoge por la presentación de distintas cualidades, es decir, los que no se pueden medir. Ejemplo: estado civil, color de ojos, sexo, profesión de una persona, carrera que piensa elegir un alumno.
Las modalidades son las diferentes situaciones de un carácter, por ejemplo, las modalidades del carácter profesión podrían ser: ingeniero, economista, psicólogo, informático, periodista …
b) Cuantitativos. Son aquellos que se pueden medir o contar y están formadas por cantidades numéricas

Una observación puede ser numérica o no numérica; las primeras se denominan cuantitativas y las segundas cualitativas. Las cuantitativas se refieren en general a medidas, por ejemplo precios, cantidad de PC en una oficina, etc.; las cualitativas se refieren a clasificaciones; por ejemplo, mediciones buenas o malas, datos confiables o anómalos etc.

También se hará la distinción entre datos de tipo continuo (por ejemplo ganancias, tiempo de ejecución de un programa, km recorridos) y datos de tipo discreto (número de días que falta un empleado de una empresa, número de sectores en un hipermercado).

Las observaciones que se obtienen de forma aleatoria (uno de los métodos de obtención de muestras) y que no se han ordenado de ninguna forma constituyen los datos crudos. Es necesario por lo tanto ordenar, presentar, agrupar y resumir los datos para cumplir con el objetivo de la investigación, estudio, trabajo etc. Desde que se dispone de computadoras, y cada vez más avanzadas, es posible manejar una gran cantidad de datos, pero de todas maneras la organización es siempre necesaria.

Cuando los datos se ordenan de acuerdo a su magnitud lo que se obtiene es una distribución de frecuencias; si se tiene en cuenta el tiempo en que ocurrió ese dato, lo que se obtiene es una serie cronológica, y si lo que se toma en consideración es la ubicación geográfica se obtiene una distribución espacial.

En el ordenamiento de los datos se deberá hacer la distinción entre datos o variables de tipo continuo y discreto.

Si los valores de la muestra se han presentado sólo una vez, se los ordena de acuerdo a su magnitud y la serie se denomina variacional o serie simple. Ahora supóngase que un elemento se encuentra más de una vez en la muestra, ese número de veces que se repite ese valor se denomina frecuencia f_i del elemento x_i. Se denomina serie estadística o serie de datos agrupados a la sucesión de pares $x_i \sim f_i$.

La forma de la distribución de los datos (de una variable) se denomina *distribución de frecuencias*.

El estudio de las distribuciones de frecuencias tiene por objeto la construcción de tablas de frecuencias que podrán utilizarse para una mejor presentación e interpretación de la información contenida en los datos observados en la muestra. En este apartado, nos referimos a las distribuciones unidimensionales de frecuencias, que son aquellas utilizadas para describir una variable individual sin tener en cuenta la información de otras variables que pudieran haberse incluido en el estudio.

Para poder obtener la forma general de una distribución de frecuencias unidimensional, es necesario introducir algunos conceptos previos.

Consideremos una población estadística de N individuos, descrita según una variable o carácter X, cuyas modalidades han sido agrupadas en un número n de clases, para cada una de esas clases i=1, . . ., n, vamos a definir:

Frecuencia absoluta de la clase: Es el número de observaciones que existen en dicha clase o sea es el número de veces que se repite dicho valor (f_i).

Frecuencia absoluta acumulada de la clase: Es el número de elementos de la muestra cuya modalidad es inferior o equivalente a las de la clase considerada (F_i).

Además se cumple que:

$$F_i = f_1 + f_2 + ... + f_i = \sum_{j=1}^{i} f_j$$

Frecuencia relativa de la clase: Es el cociente entre las frecuencias absolutas de dicha clase y el número total de observaciones o datos que denotamos por N:

$$h_i = \frac{f_i}{N}$$

Si estamos interesados en trabajar con porcentajes, sólo tenemos que multiplicar la frecuencia relativa por 100 y así representamos el porcentaje (%) de la muestra que comprende a esa clase.

Frecuencia relativa acumulada de la clase: es el número de elementos de la población que están en alguna de las clases inferior o igual a la clase.

$$H_i = \frac{F_i}{N}$$

Como normalmente el conjunto de datos que se recolecta suele ser muy grande, es necesario disponer de alguna herramienta mediante la cual podamos visualizarlos. Para ello, una vez ordenados, hacemos un recuento de dichos datos y realizamos tablas estadísticas. En estas tablas, deberán figurar los valores de la variable en estudio, y sus frecuencias correspondientes. Si bien este ordenamiento puede evitarse al trabajar con programas específicos o alguno que posea este tipo de análisis, es útil para la realización de algunos gráficos.

La principal dificultad para la obtención de una distribución de frecuencias, reside en la construcción de las modalidades, ya que ésta variará de acuerdo con el tipo de variable que se pretende describir: si la variable es cualitativa, se tomarán como modalidades las distintas respuestas observadas de la muestra; si la variable es discreta (que tome pocos valores distintos), las modalidades coincidirán con los distintos valores medidos en la muestra; si la variable es continua (o bien discreta, pero toma muchos valores distintos), se tomarán como modalidades intervalos de clase. Son los intervalos donde se encuentran los datos agrupados, se simbolizan por $[L_{i-1}, L_i)$.

Gráficos

Una de las herramientas más populares y utilizada dentro de la estadística descriptiva es, sin lugar a dudas, el análisis gráfico de los datos. Las tablas estadísticas, resumen los datos que disponemos sobre una muestra y dan toda la información necesaria, pero como se suele decir, *"Una imagen vale más que mil palabras"*, es conveniente expresar la información que disponemos mediante un gráfico o diagrama, con el fin de hacerla más clara y captar de un solo vistazo las características de los datos.

Gracias a la informática y los programas que se han desarrollado se pueden realizar fácilmente todo tipo de representaciones gráficas y de gran calidad.

Gráficos para variables cualitativas o atributos

Diagrama de barras o bastones

En este tipo de gráficos se representan en el eje de abcisas (X) las diferentes modalidades de la variable y en el eje de ordenadas (Y) la frecuencia relativa o absoluta.

Este tipo de gráficos también se puede hacer en el espacio, incorporando una nueva variable (Z) y realizando un dibujo tridimensional.

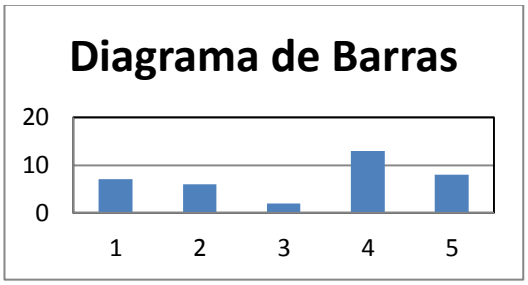


Figura Nº 1 – Diagrama de Barras

Diagramas de sectores

Se utilizan para hacer comparaciones de las distintas modalidades de un carácter mediante sectores circulares. Para construirlos se divide un círculo en tantas porciones como modalidades existan de manera que el ángulo central de cada sector ha de ser proporcional a la frecuencia absoluta o relativa correspondiente.

Este tipo de diagramas recibe también el nombre de *tartas* o *tortas*, por la forma que tiene su representación.

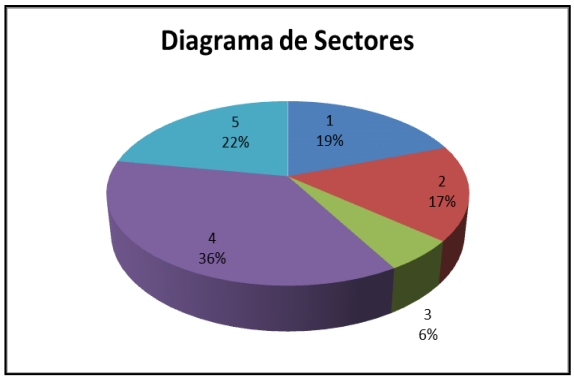


Figura Nº 2 – Diagrama de Sectores

Pictogramas

Para realizarlos se representan a diferentes escalas un mismo dibujo teniendo en cuenta que el perímetro del dibujo tiene que ser proporcional a la frecuencia, pero esto puede ocasionar un efecto visual engañoso ya que a frecuencia doble corresponde un dibujo de área cuádruple, con lo cual tiene un inconveniente debido a la falta de precisión.

A pesar de este inconveniente este tipo de dibujos son muy utilizados por los medios de comunicación a la hora de hacer que el público no especializado comprenda temas complejos sin necesidad de dar una explicación complicada.

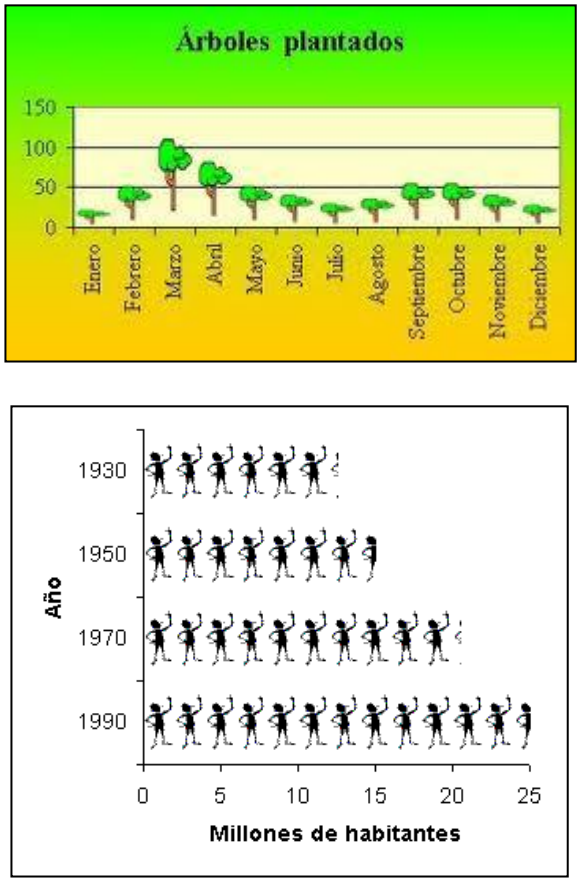


Figura Nº 3 – Pictogramas

Gráficos para Variables Cuantitativas

Para este tipo de variables, tenemos diferentes gráficos según el tipo de frecuencia que usemos y además tenemos que tener en cuenta si la variable es discreta o continua.

Gráficos para variables cuantitativas discretas

Diagrama de barras

Su representación es idéntica a la explicada para variables cualitativas, las barras deben de ser estrechas para mostrar que los valores que toma la variable son discretos.

Gráficos para variables cuantitativas continuas

Histograma

Es una manera sencilla de representar una gran masa de datos y debe ser el comienzo de cualquier estudio más sofisticado y en el que pueden observarse tres propiedades esenciales de una distribución: forma, tendencia central o acumulación y dispersión o variabilidad.

Este se obtiene por graficar en el eje x las clases y en el eje y las frecuencias. La altura de las barras del histograma tienen distinta significación según el ancho de clase sea constante o no. En el primer caso se representan frecuencias, o sea la cantidad de valores en cada clase; en el segundo caso densidad de frecuencias, o sea es el promedio, en cada clase, de cuántos valores hay por unidad de ancho de clase: $f_i / c_i = h$ **.** En este caso el área de cada rectángulo es proporcional a la frecuencia.

A diferencia del diagrama de barras, los rectángulos verticales, se representan contiguos para reflejar la idea de que la variable es continua. La forma del histograma refleja propiedades importantes de la variable estadística a la que se refiere.

El número de clases o intervalos y la longitud que se consideran, depende de cada problema y de la utilización que se quiera dar a las tablas estadísticas. Lo normal es que todos los intervalos sean de la misma amplitud ($L_i - L_{i-1}$), aunque pueden existir múltiples razones donde se aconseje tomar intervalos de amplitud variable, como puede ser el caso en el que existan uno o dos intervalos donde se concentren la mayoría de los datos.

La construcción de los intervalos de clase, introduce algunas cuestiones subjetivas, como son:
1) *¿Cuántos intervalos construir?*
Aunque no existe una regla general para usar, es evidente que el número de intervalos debe ser mayor al aumentar el tamaño muestral, lo ideal entre 5 y 20.
2) *¿Qué valor se elige como extremo inferior del primer intervalo L_0 ?*
Se toma como L_0 un valor "un poco menor" que el mínimo de la muestra (o el mínimo).

Es muy importante hacer una buena elección de la cantidad de clases a utilizar. Para este fin se utilizan distintas reglas, una de ellas consiste en tomar el número de clases igual al entero más próximo a la raíz cuadrada del número de observaciones que se estudian, $\sqrt{N}$ y no ser inferior a 5 ni superior a 20, ya que en el primer caso se produciría una concentración de datos que no sería representativa de la muestra, y en el segundo caso podrían quedar intervalos vacíos, en los cuales no habría ningún valor.

Consejos:

1. Usar intervalos de la misma longitud
2. Los intervalos no pueden solaparse
3. Cada observación sólo puede pertenecer a un intervalo
4. Todos los datos deben pertenecer a algún intervalo
5. La forma del histograma depende de la amplitud del intervalo que se elija.

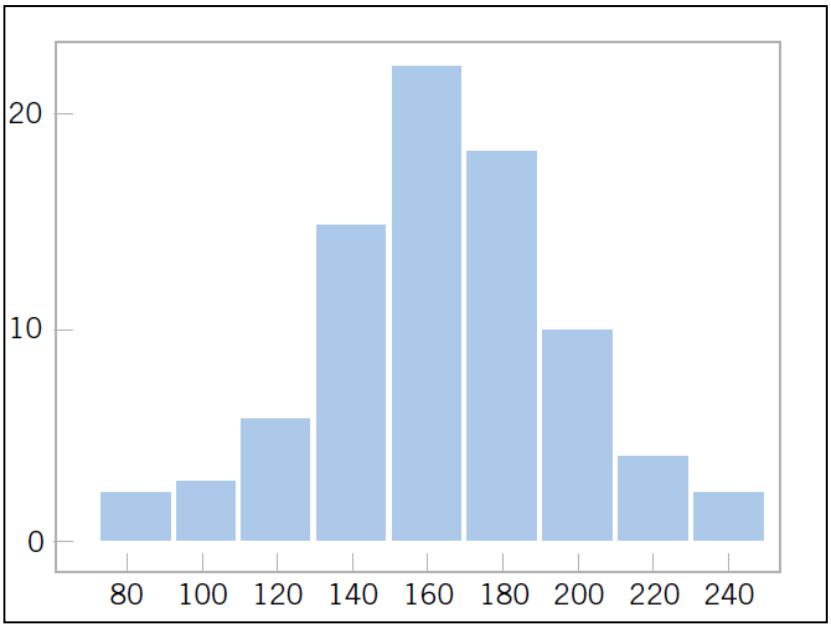


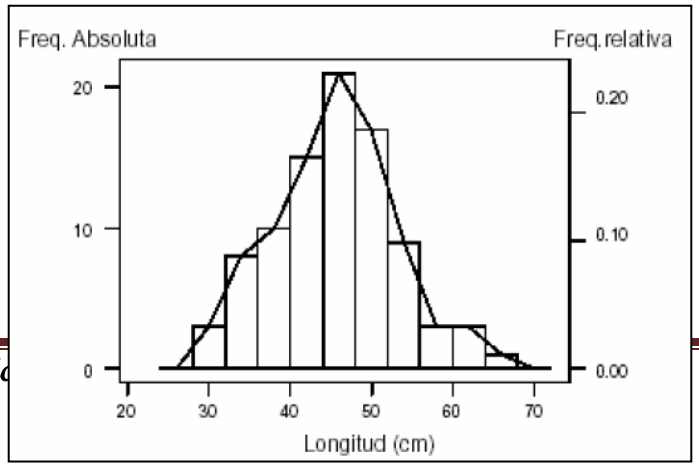
Figura Nº 4 - Histograma

Polígono de frecuencias

Se construye fácilmente una vez representado el histograma, y consiste en unir los puntos del histograma que corresponden a las marcas de clase de cada intervalo mediante una recta.

El diagrama, para variables continuas, se denomina *polígono de frecuencias acumulado u ojiva*. En estos polígonos obtenidos se aprecian con claridad propiedades importantes, da idea aproximada de qué curva teórica le correspondería a la población de la cual se obtuvo la muestra.

Si las frecuencias se expresan como proporciones - es decir, divididas por el total de observaciones en la muestra lo que se obtiene es una distribución de frecuencias relativas. Cuando se realice el histograma en este caso**,** el área total de las barras será igual a 1.



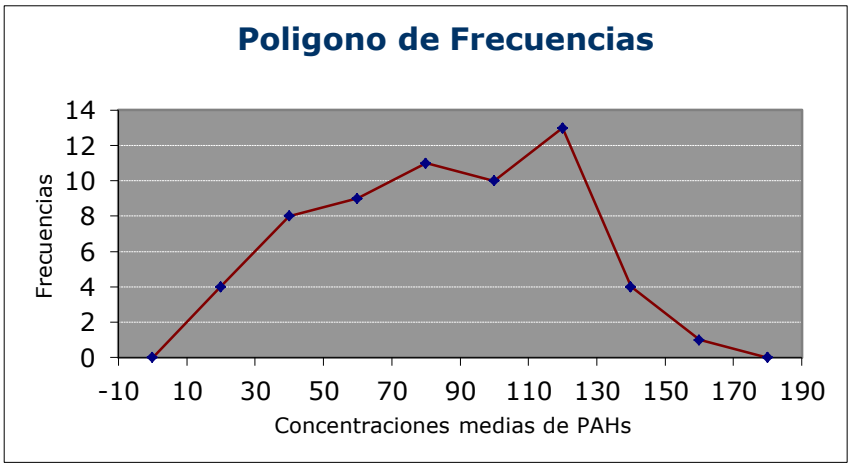
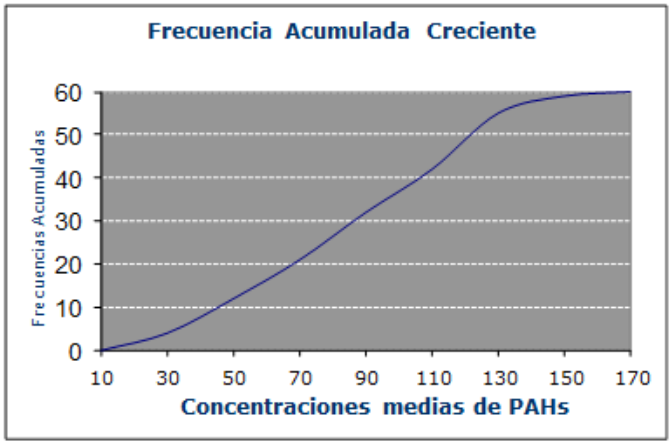
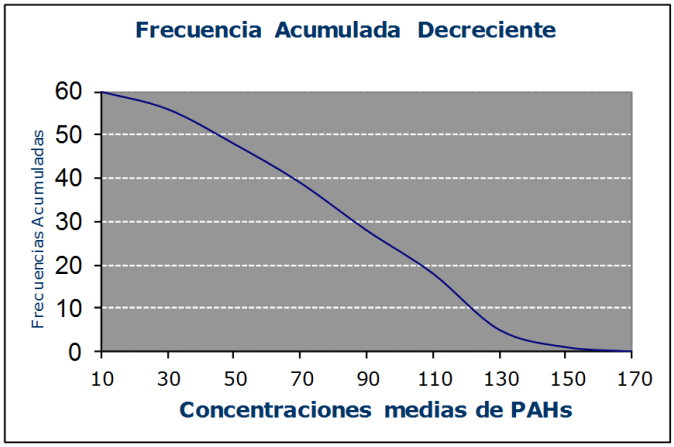


Figura Nº 5 – Polígono de Frecuencias

Polígono de frecuencias acumuladas: se utilizan en variables continuas. El eje de abscisas se construye igual que en los histogramas, pero en el de ordenadas se incluyen las frecuencias acumuladas, ya sean absolutas o relativas. Sobre cada límite se levanta una perpendicular cuya longitud sea idéntica a la frecuencia acumulada y se unen los extremos superiores de dichas perpendiculares.





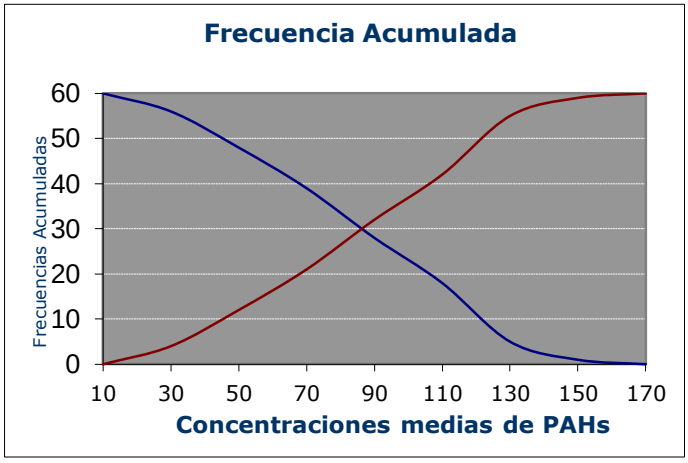


Figura Nº 6 – Curvas de Frecuencias Acumuladas

Todos estos gráficos y tablas se pueden realizar de una manera rápida y sencilla con ayuda de los softwares disponibles tanto planillas de cálculo como específicos de estadística.

ANÁLISIS EXPLORATORIO DE DATOS

Un análisis reciente que ha impulsado la estadística se debe al esfuerzo de John Tukey, quién ha producido una gran cantidad de métodos innovadores para el análisis de datos. En su libro "Exploratory Data Analysis" de 1977 y en otras publicaciones recientes (Hoaglin, Mosteller y Tukey, 1983) , Tukey ha expuesto una filosofía práctica para el análisis de datos. La escuela de Tukey se ha extendido en los últimos años (Breckenridge, 1983) haciendo énfasis en la exploración de los datos por métodos gráficos previos al clásico análisis estadístico tradicional. La visualización de los datos permite al investigador penetrar en su estructura, minimizando los supuestos probabilísticos que tradicionalmente se asumen con

respecto a su comportamiento y distribución. Lo anterior equivale a proporcionarle al investigador "una lente" de aumento que le permite:

- Exhibir características o patrones ocultos dentro de los datos.
- Resaltar con claridad la tendencia que conforman los datos.
- Proporcionar hipótesis o modelos acerca del comportamiento de los datos

La finalidad del Análisis Exploratorio de Datos (AED) es examinar los datos previamente a la aplicación de cualquier técnica estadística. De esta forma el analista consigue un entendimiento básico de sus datos y de las relaciones existentes entre las variables analizadas.

El AED proporciona métodos sencillos para organizar y preparar los datos, detectar fallos en el diseño y recogida de datos, tratamiento y evaluación de datos ausentes, identificación de casos atípicos y comprobación de los supuestos subyacentes en la mayor parte de las técnicas

El examen previo de los datos es un paso necesario, que lleva tiempo, y que habitualmente se descuida por parte de los analistas de datos. Las tareas implícitas en dicho examen pueden parecer insignificantes y sin consecuencias a primera vista, pero son una parte esencial de cualquier análisis estadístico.

Cabe mencionar que esta parte de la Estadística se ha robustecido con la reciente aparición de diversos programas como por ejemplo Statgraphics, Statistica, SPLUS, etc

Algunas de las herramientas más importantes son:

-El diagrama de tallo y hoja.

-El diagrama de caja.

- Las profundidades.

-El diagrama de letras.

-Las transformaciones matemáticas. -Las suavizaciones.

-Las series de tiempo.

Diagrama de Tallo y Hoja

El objetivo del **diagrama de tallo y hojas** es mostrar la frecuencia con la que ocurren los valores dentro de un conjunto de datos lo cual es muy parecido a lo que hace un histograma pero la diferencia es que en el diagrama de tallo y hojas no se observan barras sólidas sino que son los mismos números los que dan forma al diagrama.

Puede definirse como un híbrido que combina los aspectos visuales del histograma con la información numérica que proporciona una tabla de distribución de frecuencias.

Este diagrama se construye colocando en una columna todos los números que conforman los datos eliminando la última cifra, es decir las unidades. Esta columna debe ordenarse de menor a mayor.

A la derecha de cada número se escribe la última cifra o unidad de cada dato que comienza con ese número. Luego se ordenan de menor a mayor los números de cada fila.

Cada valor se subdivide en tres componentes el más significativo, o sea el situado más a la derecha, se usa para formar el tallo, el segundo en significación forma la hoja, que servirá para generar un histograma y con ello proporcionar una idea de la forma de la variable, y el tercero, si existe, que es el menos significativo, se puede despreciar.

Cuando existen valores muy separados del conjunto, se puede simplificar el gráfico eliminando las filas sin hojas e indicando los valores altos o bajos completos precedidos de esas palabras, ALTOS, si son muy elevados, o BAJOS si se da la circunstancia contraria.

La elaboración de un gráfico de tallo y hojas es muy sencilla, y se puede considerar como la técnica de representación gráfica recomendable para variables cuantitativas, por encima de otra forma muy usual como el histograma.

Se construye de la siguiente manera:

1. Ordenar el lote de datos en magnitud creciente.

2. Seleccionar un par conveniente de dígitos que permita fraccionar en dos partes el lote de datos según la característica de los datos o lo que se quiere mostrar.

3. Formar el tallo y las hojas con las fracciones respectivas.

4. Construir el tallo escribiendo verticalmente los dígitos enteros entre el 22 y 31, asociando a cada uno su hoja respectiva. Los dígitos del tallo están separados de los dígitos de la hoja por medio de una línea vertical.

En términos generales un diagrama de esta naturaleza hace visibles las siguientes características:

1. Muestra el rango de valores que los datos cubren.

2. Determina donde se concentran la mayoría de los datos

3. Describen la simetría del conjunto de datos.

4. Identifica si existen huecos en la distribución de los datos.

5. Señala aquellos valores que claramente se desvían del conjunto de datos.

Otra opción que presenta el diagrama de tallo y hoja es la comparación entre dos lotes de datos, aspecto que no considera el histograma. A esta derivación se le llama diagrama de tallo y hoja en espejo.

La observación de cualquiera de estos gráficos, el histograma o el diagrama de tallo y hoja, permite extraer ideas de las características generales de la variable representada.

```
  1   0|5
  7   0|666777
 26   0|8889999999999999999
(20)  1|00000000001111111111
 43   1|2233333
 36   1|44444444444555555
 19   1|6666666677
  9   1|889
  6   2|000
  3   2|22
```

7	6	1
8	7	1
9	7	1
10	5 1	2
11	5 8 0	3
12	1 0 3	3
13	4 1 3 5 3 5	6
14	2 9 5 8 3 1 6 9	8
15	4 7 1 3 4 0 8 8 6 8 0 8	12
16	3 0 7 3 0 5 0 8 7 9	10
17	8 5 4 4 1 6 2 1 0 6	10
18	0 3 6 1 4 1 0	7
19	9 6 0 9 3 4	6
20	7 1 0 8	4
21	8	1
22	1 8 9	3
23	7	1
24	5	1

Figura Nº 7 – Diagrama de Tallo y Hojas

Gráfico de Caja y Bigote - Box Plot

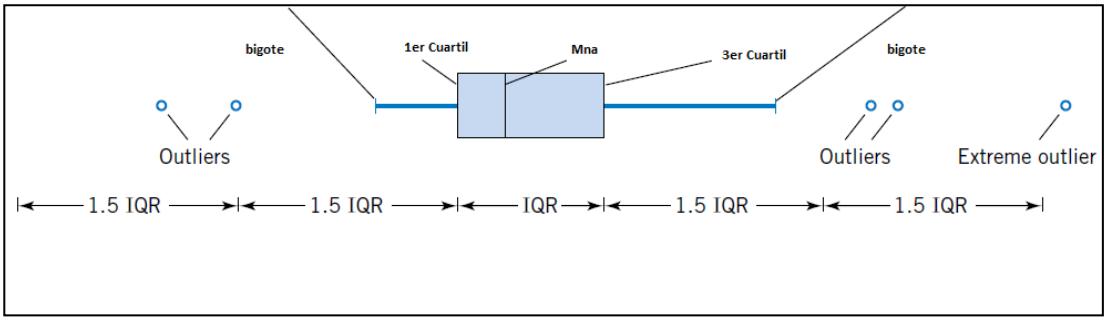
Un diagrama de caja y bigotes es una gráfica basada en cuartiles, que ayuda a visualizar un conjunto de datos.

Para construir un diagrama de caja se necesita el valor mínimo, el primer cuartil, la mediana, el tercer cuartil, y el valor máximo.

El cuadro encierra el rango intercuartil con el borde izquierdo (o más bajo) en el primer cuartil, Q_1, y el borde derecho (o superior) en el tercer cuartil, Q_3.

Se traza una línea a través de la caja en el segundo cuartil (que es el porcentil 50 o la mediana); una línea, o el bigote, se extiende desde cada extremo de la caja. El bigote inferior es una línea desde el primer cuartil hasta el punto de datos más pequeño dentro de 1,5 rango intercuartil. El bigote superior es una línea desde el tercer cuartil hacia los valores más grandes a 1,5 rango intercuartil. Datos que estén más lejos de estos bigotes se representan como puntos individuales. Un punto más allá de un bigote, pero menos de 3 rango intercuartil desde el borde de la caja, es llamado un caso atípico. Un punto más allá de

3 rangos intercuartil del borde de la caja se llama un extremo atípico. Diferentes símbolos, tales como círculos abiertos y llenos, se utilizan en ocasiones para identificar los dos tipos de valores atípicos.



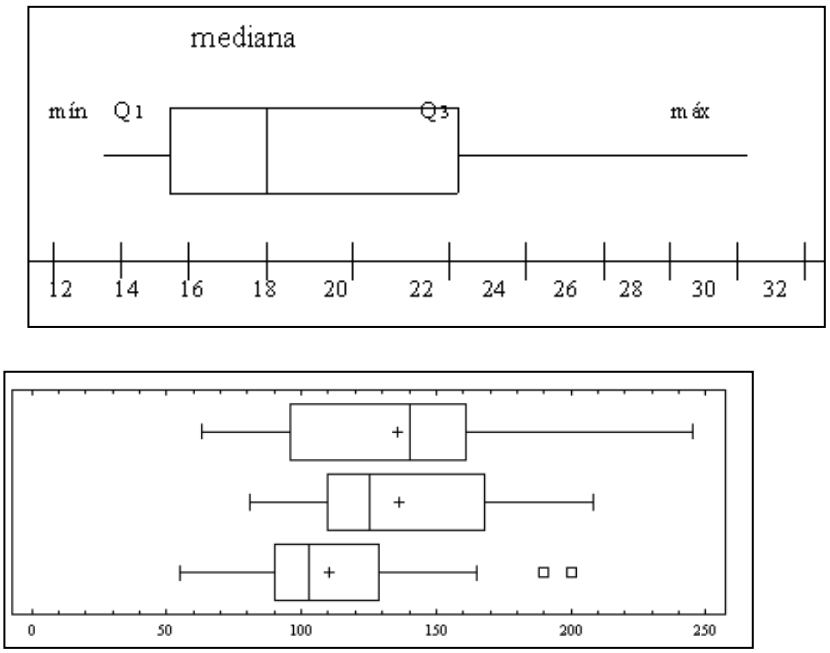
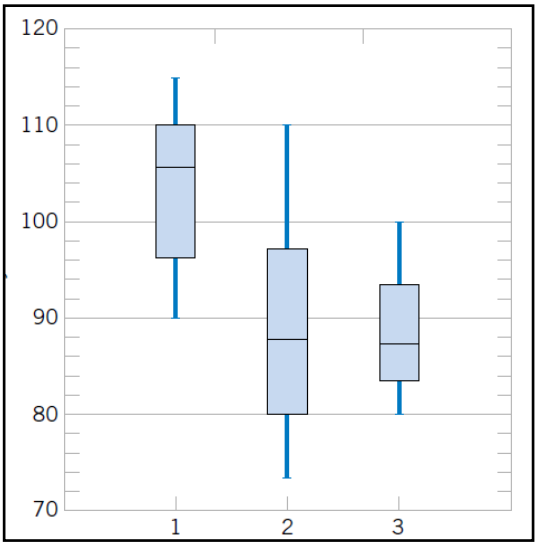


Figura Nº 8 – Diagrama de Caja y Bigote

Es posible introducir algunas variaciones en la construcción de estos diagramas, dependiendo del tipo de estudio y de la información disponible. La caja o rectángulo contiene un porcentaje de la muestra y puede construirse con diferentes rangos de variación.

Los diagramas de caja son muy útiles para realizar comparaciones gráficas entre los conjuntos de datos, debido a que tienen alto impacto visual y son fáciles de entender como se muestra en la figura siguiente.



Si la mediana está ubicada como sigue

entonces la distribución es asimétrica negativa. (izquierda)

Si la mediana está de esta manera

Entonces la distribución es asimétrica positiva. (derecha)

Todo este análisis se puede realizar también con la ayuda de softwares muchos de los cuales tiene incorporado sino todo parte de este análisis.

CARACTERÍSTICAS de muestra

Además de organizar los datos y mostrarlos en gráficos, se necesita de ciertas medidas representativas que puedan resumir una gran cantidad de ellos.

Estos números que sirven para caracterizar las **distribuciones de frecuencias** de datos univariados pueden resumirse en aquellos que tienen en cuenta las cuatro propiedades básicas:

1-Ubicación del centro de la distribución, comúnmente llamadas **MEDIDAS DE IA**

TENDENCIA CENTRAL

2-Variación de las observaciones alrededor del punto central. Estas son conocidas como **MEDIDAS DE DISPERSIÓN**

3-Grado de asimetría conocida como **MEDIDAS DE ASIMETRIA**

4-Grado de variación en altura de una distribución respecto de un modelo o patrón **MEDIDAS DE CURTOSIS**

MEDIDAS DE LA TENDENCIA CENTRAL

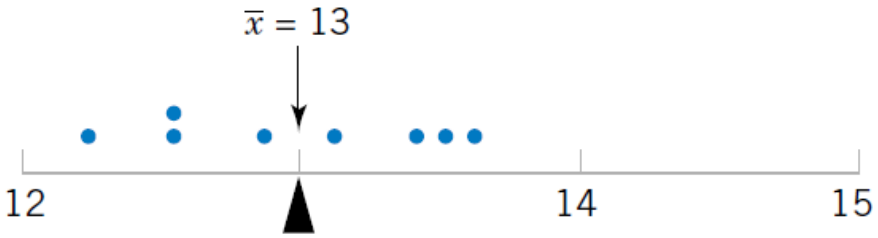
Entre estas medidas se diferencian los llamados *promedios* y las *medidas de ubicación.* El más comúnmente usado es la **media aritmética**; otros menos usados y útiles en algunas circunstancias especiales son la media geométrica y la media armónica.

Entre las medidas de ubicación se consideran la **mediana,** el **modo**, los **cuantiles**: **deciles, cuartiles, porcentiles.**

Promedios

La **media aritmética** se define como el promedio de todos los valores de la muestra:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (1)$$



Es uno de los valores de la tendencia central más usados y el más simple de calcular.

La media aritmética tiene algunas propiedades matemáticas:

Es el *centro de gravedad, un punto de equilibrio.* De su expresión se puede ver que:

$$n.\bar{x} = \sum x_i$$

Esto significa que concentra en su valor toda la información que hay en la muestra.

La suma de las desviaciones respecto a este valor es igual a cero.

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

$$\sum x_i - n\bar{x} = 0$$

La suma de las desviaciones cuadradas de los datos con respecto a la media es menor que si estas desviaciones se toman con respecto a cualquier otro valor. Se demuestra que:

$$\sum (x_i - \bar{x})^2 = mínimo$$

Debido a esta propiedad la media aritmética se emplea como base de las medidas de dispersión.

La media queda fuertemente afectada por los valores extremos, y por esto puede que en algunos casos no sea representativa.

La media puede tratarse algebraicamente, esto es: si se tiene la media de subgrupos puede obtenerse la media general promediando estas medias; si el número de elementos de cada subgrupo no es el mismo se efectuará un promedio ponderado por la cantidad de elementos en cada grupo:

$$\bar{X} = \frac{\bar{x}_1 N_1 + \bar{x}_2 N_2 + \bar{x}_n N_n}{N} \qquad (2)$$

$\bar{x}_1, \bar{x}_2 \ldots\ldots \bar{x}_n$ s on las medias de cada grupo,
N_i, **...,** N_n es la cantidad de elementos de cada grupo; $N = N_i$ + **...** + N_n es la longitud total de la muestra.

Media Geométrica: se define como la raíz n-ésima del producto de los elementos que conforman la muestra.

$$Gm = \sqrt[n]{\Pi\, x_i} \qquad (3)$$

Es apropiada para promediar razones, porcentajes o velocidades de cambio. Para facilitar su cálculo se pueden aplicar logaritmos a la expresión anterior.

$$\log G_m = \frac{1}{n}(\log x_1 + \log x_2 + ... + \log x_n)$$

Puede verse de esta expresión, que el logaritmo de la media geométrica es igual al promedio del logaritmo de los valores de las observaciones.

Si la variable presenta frecuencia:

$$G_m = \sqrt[n]{\prod x_i^{fi}}$$

$$\log G_m = \frac{1}{n}(\log x_1 \ f_1 + \log x_2 \ f_2 + ... + \log x_n \ f_n)$$

Tiene las siguientes propiedades:

-está menos afectada por valores extremos.

-para cualquier serie es siempre menor que la media aritmética.

-es muy útil en el cálculo de *números índice.*

-se puede manipular algebraicamente.

-no es muy conocida y no puede evaluarse cuando hay datos negativos o ceros.

Media armónica: Se define como la inversa de la media aritmética de las inversas de los valores muestrales. Es apropiada para el procesamiento de datos de razones que tienen dimensiones fisicas como *km/I, producción/hora,* etc. Sus expresiones para los diferentes casos son:

$$\frac{1}{Hm} = \frac{\frac{1}{x_1} + \frac{1}{x_2} + ... + \frac{1}{x_n}}{N} \qquad (4)$$

Observaciones sobre la media Geométrica y la media Armónica

El empleo de la media geométrica o de la armónica equivale a una transformación de la variable en log x ó 1/x , respectivamente, y el cálculo de la media aritmética de la nueva variable; por ejemplo, si la variable abarca un campo de variación muy grande, tal como el porcentaje de impureza de un producto químico, por lo general alrededor del 0.1%, pero que en ocasiones llega incluso al 1% o más, puede ser ventajoso el empleo de log x en lugar de x para obtener una distribución más simétrica.

Relación entre las medias:

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

Medidas de ubicación

Modo: Es un valor muy útil en la descripción de la muestra. Se lo define como el valor de la variable que aparece más veces que otro; se puede decir que es el valor con mayor frecuencia.

Si la variable analizada es continua se puede obtener con alguna de las siguientes expresiones:

$$Mo = L_{iMo} + \frac{d_1}{d_1 + d_2} c \qquad (5)$$

Li_{Mo} = límite inferior del intervalo que contiene al modo.
d_1:diferencia (sin tener en cuenta el signo) entre las frecuencias del intervalo que contiene el modo y el intervalo anterior.
d_2:diferencia entre las frecuencias del intervalo que contiene al modo y a del intervalo posterior.
c = ancho de clase.

$$Mo = L_{iMo} + \frac{f_1}{f_1 + f_2} c \qquad (6)$$

f_1=frecuencia del intervalo de clase anterior al modal.
f_2=frecuencia del intervalo de clase posterior al modal.

Es un valor muy inestable, ya que puede cambiar con el método de redondeo de los datos. Puede determinárselo gráficamente a partir del histograma interpolando en la barra más alta.

Mediana: Es el valor de la serie para el cual el 50% de los valores son menores y el 50% mayores o iguales. Se la denomina valor medio de la serie.

Tenemos que diferenciar entre:

Variables discretas:

Serie sin frecuencias

- Si el número de datos es impar Mna. es el valor central.
- Si el número de datos es par Mna. es la semisuma de los valores centrales.

Si hay frecuencias:

- Se calcula N/2 y se obtiene Ni (frecuencias acumuladas)
- Se observa cual es la primera Ni que supera o iguala a N/2, distinguiéndose dos casos:
  - Si existe un valor de x_i tal que $N_{i-1} < N/2 < N_i$, entonces se toma como $Mna. = x_i$ o sea la que se corresponde con la frecuencia Ni
  - Si existe un valor i tal que $N_i = N/2$ entonces la mediana será $Mna. = \frac{x_i + x_{i+1}}{2}$

Variables continuas

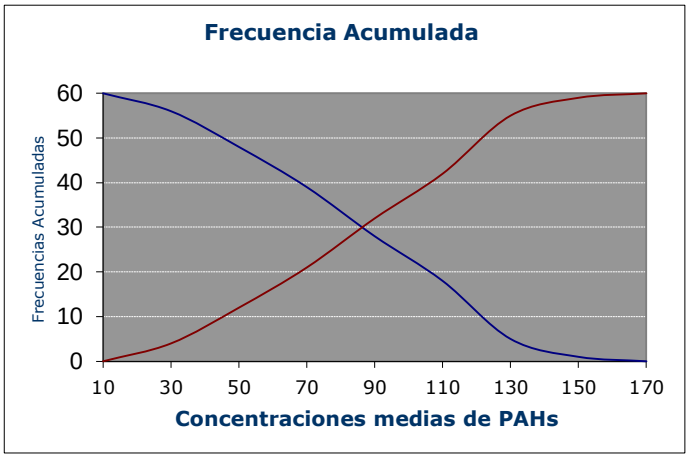
$$Mediana = L_i + \frac{N/2 - FL_i}{f_i} c \qquad (7)$$

siendo:

Li=límite inferior del intervalo que contendrá a la mediana.
FLi=frecuencia acumulada hasta el intervalo anterior a la clase mediana
fi=frecuencia del intervalo mediano.

También se la puede obtener a través de las ojivas, en su intersección o en el valor correspondiente al 50%.

Gráficamente en la intersección de las ojivas.



Propiedades

-No está influenciada por valores extremos. Por lo tanto, es una medida conveniente de la ubicación central.

-Un valor seleccionado a azar se ubicará por arriba o por debajo de ella con igual probabilidad; por esto suele llamársela valor probable.

-Su cálculo es fácil.

Algunas desventajas son:

-los datos deben ordenarse para su cálculo.

-No se la puede manipular algebraicamente.

-No es tan usada como la media aritmética, y tiene mayor error que ella.

Cuantiles (cuartiles, deciles, percentiles): son también medidas de ubicación. Como la mediana divide a la distribución de datos en dos partes, los cuartiles la dividen en cuatro, los deciles en diez y los percentiles en cien. Estas medidas posibilitan un análisis más minucioso de la distribución.

Los cuartiles dividen a la distribución en cuatro partes; por lo tanto hay tres cuartiles. El caso de los percentiles se usa cuando existen muchas observaciones.

Se calculan de la misma forma que la mediana, sólo que cambia come se determina el orden del cuantil. Por ejemplo para ubicar el primer cuartil se hace N/4, para el segundo 2N/4 y así para el resto, el cálculo es luego similar al realizado para la mediana. De la misma forma se precede para calcular los deciles y percentiles.

Se define el cuantil p como el número que deja a su izquierda una frecuencia relativa **p**. Lo que es lo mismo, la frecuencia relativa acumulada hasta el cuantil **p** es p. Claro está que los cuantiles sólo se podrán calcular con variables ordinales. Nótese que la mediana es el cuantil 005. Para calcular los cuantiles seguiremos las siguientes indicaciones.

Si la variable es discreta, o si es continua y disponemos de todos los datos:

Ordenamos la muestra. Tomamos el menor dato de la muestra (primero de la muestra ordenada) cuya frecuencia relativa acumulada es mayor o igual que **p**. Si se supera p estrictamente, este dato ya es el cuantil p; mientras que si se alcanza con igualdad, el cuantil p es la media de este dato con el siguiente.

Si la variable es continua y se encuentra agrupada en intervalos de clase:
Buscamos en la tabla de frecuencias el primer intervalo cuya frecuencia relativa acumulada es mayor o igual que **p o q** (pensemos que es el intervalo [Li ; Li+1) y, dentro de ese intervalo, calculamos el cuantil **p** por interpolación lineal, esto es:

$$C_{r/q} = L_{i-1} + \frac{\frac{rN}{q} - N_{i-1}}{n_i}.c_i \qquad r = 1,2.....q-1$$

con p o q=4 cuartil, p=5 quintil, p=10 decil, p=100 percentil o porcentil.

Primero se calcula r.(N/q), y se mira a ver en qué intervalo cae el número que salga, y es con este intervalo en el que nos tenemos que fijar para usar F_{i-1}, f_i,… c_i es la longitud de ese intervalo).

Algunos órdenes de los cuantiles tienen nombres específicos. Así los cuartiles son los cuantiles de orden (0.25, 0.5, 0.75) y se representan por Q1, Q2, Q3. Los deciles son los cuantiles de orden (0.1, 0.2,..., 0.9). Los percentiles son los cuantiles de orden j/100 donde j=1,2,...,99.

MEDIDAS DE DISPERSION

Permiten *calcular la representatividad de una medida de posición*, para lo cual es preciso cuantificar la distancia entre los diferentes valores de la distribución respecto a dicha medida. A esta distancia es a lo que se denomina ***variabilidad o dispersión de la distribución***.

La finalidad de estas medidas es estudiar *hasta qué punto para una determinada distribución de frecuencias, las medidas de tendencia central o de posición son representativas* como síntesis de toda la información de la distribución.

La medida más simple de variación o dispersión es la amplitud que se considera como la diferencia entre los valores mínimo y máximo de la serie, también conocida come *rango.* También puede obtenerse la amplitud media como el promedio de los valores extremos.

Puede obtenerse además la *amplitud intercuantílica,* es considerar las distancias entre cuartiles, deciles y percentiles. Por ejemplo para los cuartiles se calculan el primero y el tercero y se restan sus valores, de la misma forma se precede con los deciles y percentiles.

Interesa fundamentalmente poder determinar una medida de variabilidad que involucre a todos los valores contenidos en la muestra, para esto se puede considerar el promedio de las desviaciones de los valores respecto a la media aritmética

$$\frac{\sum(x_i - \bar{x})}{n}$$

pero como se vió:

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

para evitar esto es que se toman las desviaciones sin considerar su signo y se obtiene la desviación media:

$$\frac{\sum|(x_i - \bar{x})|}{n}$$

En vez de usar valores absolutos, se pueden tomar cuadrados, ya que es más sencillo trabajar con ellos. El promedio de estos desvíos respecto a la media se denomina Varianza de la muestra y se simboliza S^2

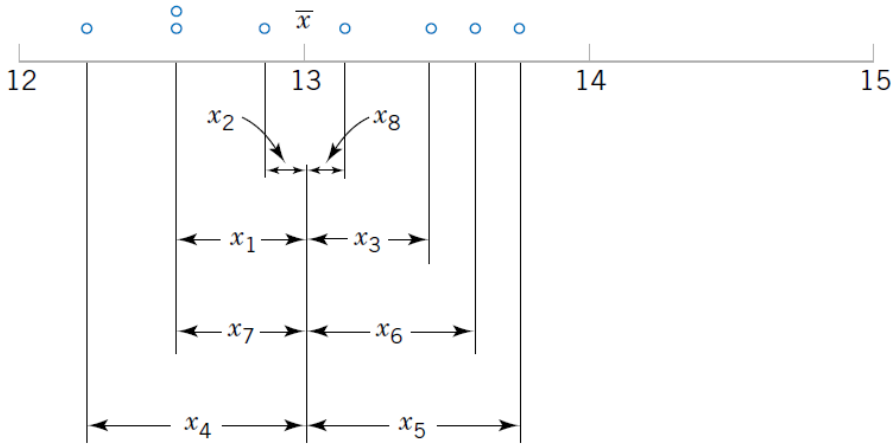
$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n} \qquad (8)$$

también se la suele expresar como:

$$S'^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \qquad (9)$$

Cuando la muestra es grande la diferencia entre ambas expresiones es despreciable y se utiliza la primera.

El promedio de las diferencias al cuadrado respecto a la media se denomina momento centrado de orden 2: $\mathbf{m_2}$



La raíz cuadrada de la varianza se denomina desvío estándar de la muestra y se lo suele preferir a la varianza ya que tiene las mismas unidades que la variable. Para simplificar su cálculo se desarrolla el cuadrado de su expresión y se consigue la expresión en función de los momentos muestrales respecto al origen (recuérdese población):

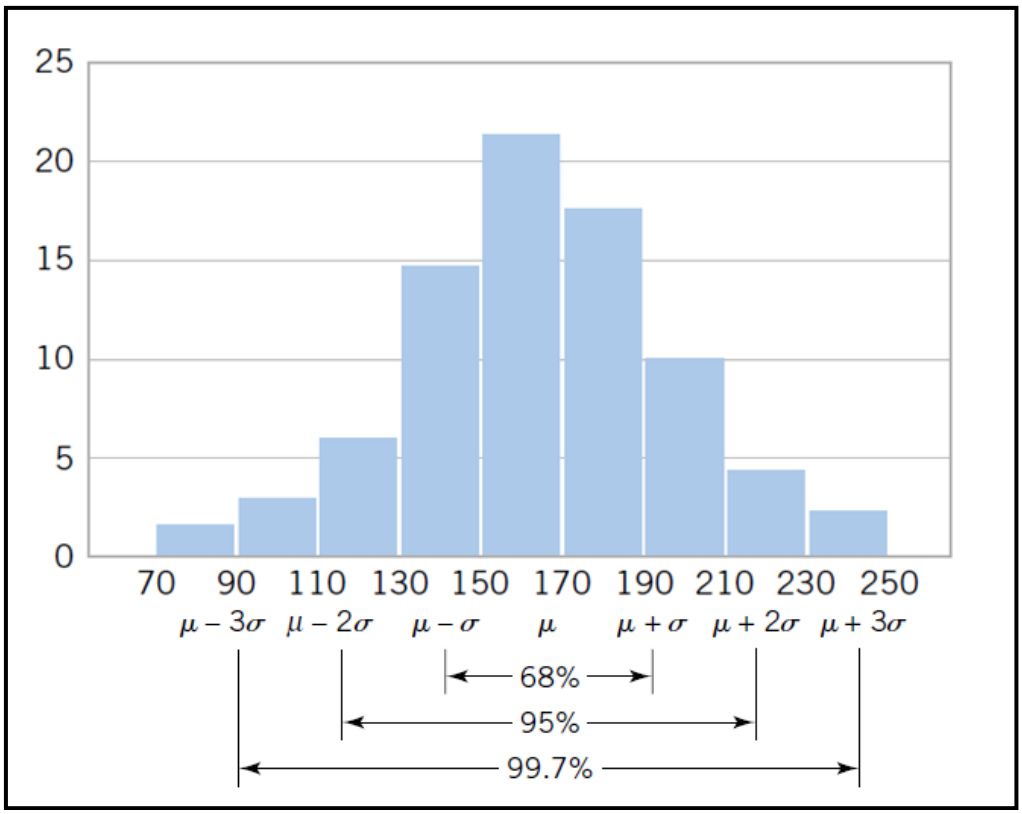
$$S^2 = \frac{\sum x_i^2}{n} - \bar{x}^2$$

$$S^2 = a_2 - a_1^2 \qquad (10)$$

Estas medidas (varianza y desvío) son las medidas de variabilidad más empleadas. Pero si se halla sólo el desvío o varianza de un conjunto de valores, no se puede asegurar si representa un valor alto, medio o bajo de variabilidad.

Se utiliza por esto una regla empírica para interpretar los valores de la varianza o desvío, se usará cuando

la muestra sea grande y la forma de la muestra sea aproximadamente de campana; esta regla considera que si se miden en el eje x y hacia ambos lados de la media una distancia igual al desvío, en ese intervalo quedarán comprendidos el 68% de las observaciones. Si se traza dos veces el desvío hacia ambos lados de la media quedarán comprendidos el 95% de las observaciones en ese intervalo, y si se trazan tres veces el desvío quedarán comprendidos el 99% de las observaciones entre esos límites.



Estas medidas descriptas son en valor absoluto. Si se comparan las variabilidades de dos conjuntos de datos en base a estas medidas anteriores, puede darse una respuesta cierta sólo si las medias (respecto a las cuales se hallan los desvíos) son aproximadamente iguales. Cuando esto no sucede, o bien tienen unidades diferentes, se recurre a medidas de variabilidad relativas para independizarlas de las unidades de medida. Una muy usada es el coeficiente de variabilidad o coeficiente de Pearson, que para la muestra es:

$$Cv = \frac{S}{\bar{x}} * 100 \qquad (11)$$

normalmente expresado en porcentaje. Por ejemplo si se quiere comparar la variabilidad de los caudales del río Paraná y los del arroyo Saladillo, la desviación o varianza no sirven ya que los valores medios son muy diferentes; en este caso es conveniente calcular este coeficiente y comparar los porcentajes obtenidos.

MEDIDAS DE FORMA

ASIMETRÍA

Cuando la distribución es simétrica la media, mediana y el modo coinciden. Cuando es asimétrica esos valores difieren. Come se ha visto la media aritmética es el valor de la tendencia central más afectado por los valores extremos, es por esto que cuanto mayor sea la distancia entre la media y el modo mayor será el grado de asimetría. Esta diferencia entre media y modo se suele usar como medida de asimetría ya que cuanto mayor sea esta distancia mayor será la asimetría. A los fines de comparar la asimetría entre dos distribuciones y salvar la diferencia de unidades y la diferencia en las dispersiones es que se divide por el desvío. Debido a que el modo se encuentra aproximadamente algunas veces, se prefiere trabajar con la mediana que se encuentra mejor y teniendo en cuenta la relación vista entre la media, mediana y modo, se pueden obtener las siguientes expresiones para la asimetría:

$$As = \frac{(\bar{x} - Modo)}{S} \qquad As = \frac{3(\bar{x} - Mediana)}{S} \qquad (12)$$

Este valor seria aproximadamente igual a 0 para una distribución simétrica, positivo para una distribución asimétrica hacia la izquierda y negativo para una distribución asimétrica hacia la derecha.

El valor exacto de la Asimetría está dado por el momento centrado de tercer orden adimensionalizado:

$$m_3 = \frac{\sum (x_i - \bar{x})^3}{n} \qquad As = \frac{m_3}{S^3} \qquad (13)$$

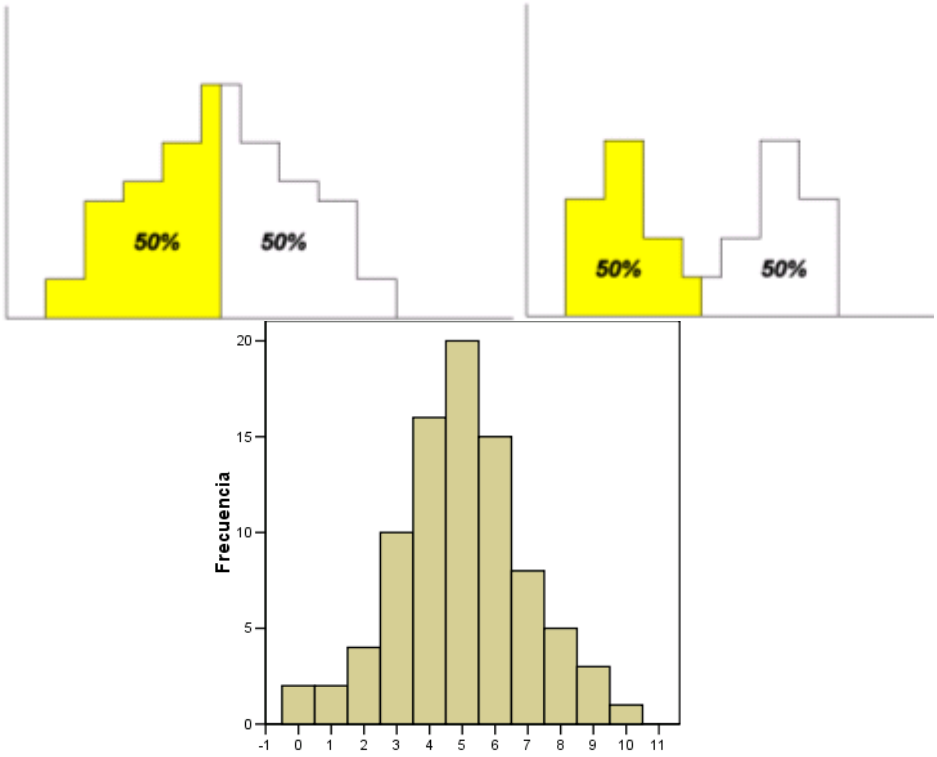


Figura Nº 9 – Distribución simétrica

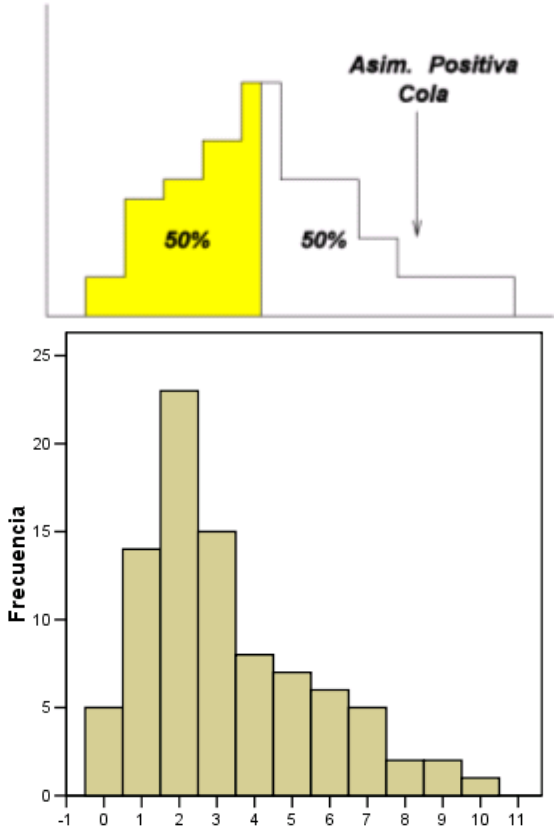
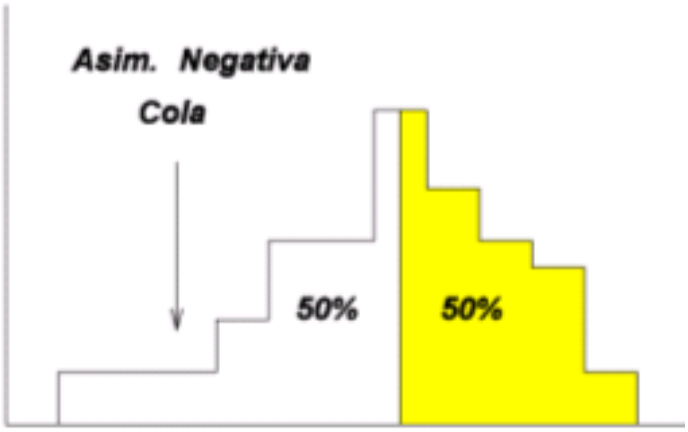


Figura Nº 10 – Distribución asimétrica positiva o a la derecha



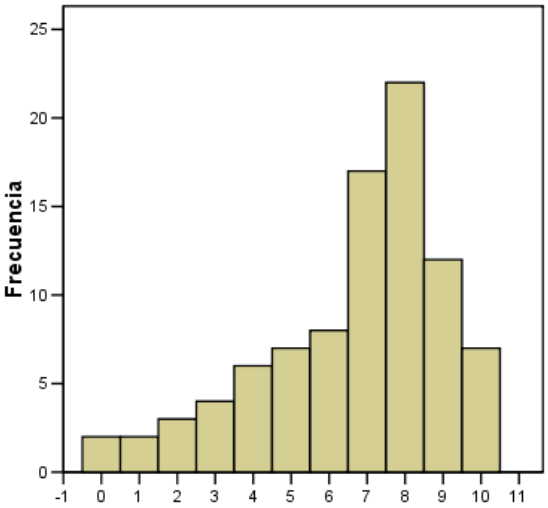


Figura Nº 11 – Distribución asimétrica negativa o a la izquierda

CURTOSIS

Esta medida dada por el coeficiente de Curtosis mide la diferencia en elevación respecto a una curva tomada come patrón e modelo que es la curva normal.

Se la puede definir aproximadamente como una razón de la amplitud semiintercuartil y la amplitud 90-10 porcentil:

$$K = \frac{1}{2}\frac{(Q_3 - Q_1)}{(P_{90} - P_{10})} \qquad (14)$$

este coeficiente clasifica distintos tipos de agudeza:

K>0	leptocúrtica o más empinada que la curva normal
K=0	mesocúrtica igual a la curva normal
K <0	platicúrtica o menos empinada que la curva normal

La forma exacta de calcular es a través del momento centrado de orden 4:

$$m_4 = \frac{\sum (x_i - \bar{x})^4}{n} \qquad K = \frac{m_4}{S^4} \qquad (15)$$

RESUMEN:

DIAGRAMAS SEGÚN EL TIPO DE VARIABLES

Tipo de variable	Diagrama o gráfico
Cualitativa	Barras, sectores, pictogramas
Cuantitativa (discreta)	Barras Escalera
Cuantitativa (continua)	Histograma, polígono de frecuencias diagramas acumulativos

Medidas Descriptivas Numéricas y Representaciones Graficas aconsejadas en función de la escala de medida de la variable

Escala de medida	Representaciones gráficas	Medidas de tendencia central	Medidas de dispersión
Nominal	Diagrama de barras Diagrama de líneas Diagrama de sectores	Moda	
Ordinal	Boxplot	Mediana	Rango Intercuartílico
Intervalo	Histogramas Polígono de frecuencias	Media	Desviación Típica
Razón		Media Geométrica	Coeficiente de Variación