Unidad I

Estadística Descriptiva.

1.1 Introducción, notación sumatoria

Históricamente, a través del tiempo podemos conocer y comprender la creación y el desarrollo de la estadística moderna, mediante el análisis de dos fenómenos diferentes y separados entre sí (Levine, 1992):

a) La estadística nació por la necesidad de los gobiernos de recopilar información sobre sus ciudadanos.

b) Posteriormente, la estadística se desarrolló debido a la aplicación de las, técnicas matemáticas en la teoría de la probabilidad.
A lo largo de la historia registrada de la humanidad, se observa la permanente necesidad de parte de las castas gobernantes de realizar la recopilación de información relativa a ciertos
.datos de interés con respecto a su población. En las civilizaciones egipcias, griega y romana, se obtenía información primordialmente con el propósito de cobrar impuestos y reclutar soldados.

En la edad media, era frecuente que las instituciones eclesiásticas llevaran registros acerca de los nacimientos, muertes, y matrimonios. En Estados Unidos de Norteamérica, se mantenían diversos registros durante los tiempos de la 'colonia y a principios de 1790, la constitución Federal de ese país implantó el levantamiento de un censo de población cada10 años. En la actualidad, estos datos se utilizan con diversos propósitos, incluyendo la distribución de curules en el congreso, y la asignación de fondos federales. Estas y otras necesidades en el ámbito nacional de los gobiernos, estuvieron estrechamente interrelacionados con el actual desarrollo de la moderna estadística.
En este capítulo, se presenta una definición acerca de la estadística moderna, su clasificación, los métodos gráficos y numéricos utilizados para llevar a cabo la recopilación y presentación de la información de interés, y posteriormente efectuar, el análisis correspondiente mediante la aplicación de las herramientas estadísticas más comúnmente empleadas como son: las medidas de tendencia central, las medidas de posición relativa y las medidas de dispersión.

ESTADÍSTICA DESCRIPTIVA

Definición

De estadística descriptiva consisten procedimientos usados para resumir y escribir las características importantes de un conjunto de mediciones.

Si el conjunto de mediciones esa población entera, sólo necesita sacar conclusiones con base en estadística descriptiva. Sin embargo, podría ser demasiado caro o tardado enumerar toda la población. Quizá enumerar la población la describía, como en el caso de la prueba del “tiempo para falla”. Por estas otras razones, podría tener sólo una muestra de la población. Al examinar la muestra, desea contestar las preguntas acerca de la población como en todo. La rama de la estadística que trata con este problema se llama estadística inferencias.

1.1.1 Datos no agrupados.

Datos no agrupados es el conjunto de observaciones que se presentan en su forma original tal y como fueron recolectados, para obtener informacion directamente de ellos.

Ejemplos:

5,7,2,15,2,6,12,5,5,20,10. numero de personas que ayudaron a una causa.

TRATAMIENTO PARA DATOS NO AGRUPADOS.

¿A qué se refiere esto? Cuando la muestra que se ha tomado de la población o proceso que se desea analizar, es decir, tenemos menos de 20 elementos en la muestra, entonces estos datos son analizados sin necesidad de formar clases con ellos y a esto es a lo que se le llama tratamiento de datos no agrupados.

1.1.2 Medidas de tendencia central

Medidas de tendencia central. Se les llama medidas de tendencia central a la media aritmética, la mediana, la media geométrica, la moda, etc. debido a que al observar la distribución de los datos, estas tienden a estar localizadas generalmente en su parte central. A continuación definiremos algunas medidas de tendencia central y la forma de calcular su valor.

1) Media aritmética ($\bar{x}$). También se le conoce como promedio ya que es el promedio de las lecturas o mediciones individuales que se tienen en la muestra, se determina con la fórmula siguiente:

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

donde:

$\bar{x}$ = media aritmética

x_i = dato i

n = número de datos en la muestra

Ejemplos:

1. Se han tomado como muestra las medidas de seis cables usados en un arnés para lavadora, las cuales son; 15.2 cm, 15.0, 15.1, 15.2, 15.1 y 15.0, determine su media aritmética.

Solución:

$$\bar{x} = \frac{15.2 + 15.0 + 15.1 + 15.2 + 15.1 + 15.0}{6} = 15.1cm$$

2. Se toman varias muestras de cierto tipo de queso y se determina la cantidad de proteína por cada 100 gramos de queso, encontrándose lo siguiente: 26.5 gramos, 24.8, 25.3, 30.5, 21.4, determine la cantidad promedio de proteína encontrada en la muestra por cada 100 gramos de queso que se elabora.

Solución:

$$\bar{x} = \frac{26.5 + 24.8 + 25.3 + 30.5 + 21.4}{5} = 25.7 grs$$

3. Se hacen varias lecturas de una muestra que contiene cobre, las lecturas se hacen en un espectrofotómetro de absorción atómica y son la siguientes: 12.3%, 12.28, 12.27, 12.3, 12.24, 15.01, determine la concentración promedio de Cu en la muestra.

Solución:

$$\bar{x} = \frac{12.3 + 12.28 + 12.27 + 12.3 + 12.24 + 15.01}{6} = \frac{76.4}{6} = 12.73\%Cu$$

Si observamos las lecturas del espectrofotómetro nos damos cuenta que el valor de 15.01% es un valor diferente al de las lecturas anteriores, por lo que se descarta el valor ya que se considera un valor atípico, es decir un valor que es debido a circunstancias especiales, en este caso puede ser que se deba al hecho de que se está descalibrando el aparato de absorción atómica o simplemente que se ha equivocado el operador del aparato al tomar la lectura, por lo que la media se debe calcular con las primeras cinco lecturas; como se muestra a continuación:

Solución:

$\bar{x} = \frac{12.3 + 12.28 + 12.27 + 12.3 + 12.24}{5} = \frac{61.39}{5} = 12.278\%Cu$ y esta sería la media correcta

4. Si deseamos determinar la edad promedio de los estudiantes de una escuela de nivel superior al iniciar sus estudios, suponga que se toman las

edades de algunos de los alumnos de cierta clase y estas son las que siguen: 20, 18, 18, 19, 18, 19, 35, 20, 18, 18, 19.

Solución:

Luego, la media se determinará con solo 10 de las edades ya que es necesario descartar la edad de 35 años, que es un dato atípico o un caso especial, por lo que;

$$\bar{x} = \frac{20+18+18+19+18+19+20+18+18+19}{10} = \frac{187}{10} = 18.7 años$$

1.1.3 Medidas de posición.

Las **medidas de posición** dividen un conjunto de datos en grupos con el mismo número de individuos.

Para calcular las **medidas de posición** es necesario que los **datos** estén ordenados de **menor a mayor**.

La **medidas de posición** son:

Cuartiles

Los **cuartiles** son los **tres valores** de la variable que **dividen** a un **conjunto** de **datos ordenados** en **cuatro partes iguales**.

Q_1, Q_2 y Q_3 determinan los valores correspondientes al **25%, al 50% y al 75%** de los **datos**.

Q_2 coincide con la **mediana**.

1.1.4 Medidas de dispersión.

Las **medidas de dispersión**, también llamadas medidas de variabilidad, muestran la variabilidad de una distribución, indicando por medio de un número, si las diferentes puntuaciones de una variable están muy alejadas de la media.Cuánto mayor sea ese valor, mayor será la variabilidad, cuanto menor sea, más homogénea será a la media. Así se sabe si todos los casos son parecidos o varían mucho entre ellos.

Para calcular la variabilidad que una distribución tiene respecto de su media, se calcula la media de las desviaciones de las puntuaciones respecto a la media aritmética. Pero la suma de las desviaciones es siempre cero, así que se adoptan dos clases de estrategias para salvar este problema. Una es tomando las desviaciones en valor absoluto (Desviación media) y otra es tomando las desviaciones al cuadrado (Varianza).

1.1.5 Medidas de forma

Diremos que una distribución es simétrica cuando su mediana, su moda y su media aritmética coinciden.

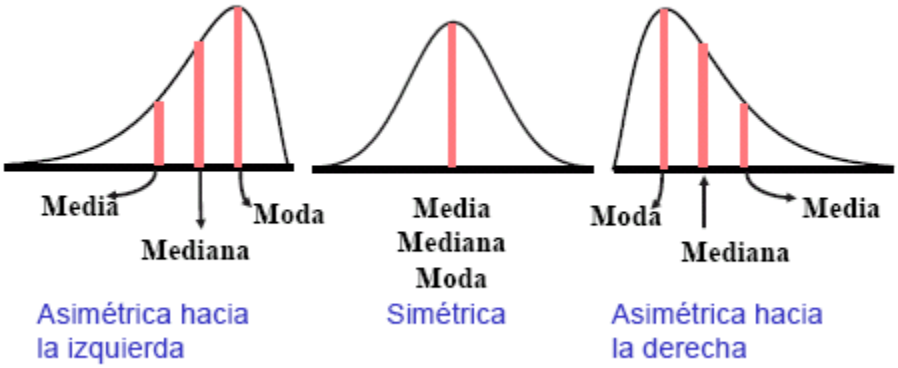
Diremos que una distribución es *asimétrica a la derecha* si las frecuencias (absolutas o relativas) descienden más lentamente por la derecha que por la izquierda.

Si las frecuencias descienden más lentamente por la izquierda que por la derecha diremos que la distribución es *asimétrica a la izquierda.*

Existen varias medidas de la asimetría de una distribución de frecuencias. Una de ellas es el **Coeficiente de Asimetría de Pearson**:

$$A_s = \frac{\bar{x} - M_o}{s}$$

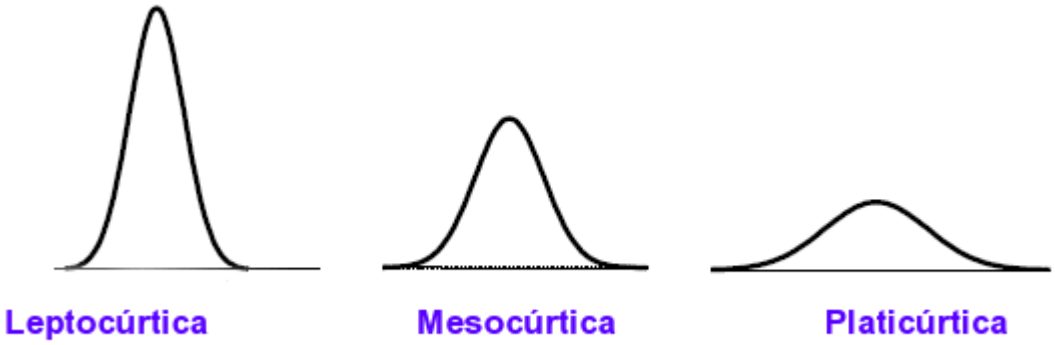
Su valor es cero cuando la distribución es simétrica, positivo cuando existe asimetría a la derecha y negativo cuando existe asimetría a la izquierda.



MEDIDA DE APUNTAMIENTO O CURTOSIS

Miden la mayor o menor cantidad de datos que se agrupan en torno a la moda. Se definen 3 tipos de distribuciones según su grado de curtosis:

Distribución mesocúrtica: presenta un grado de concentración medio alrededor de los valores centrales de la variable (el mismo que presenta una distribución normal). **Distribución leptocúrtica**: presenta un elevado grado de concentración alrededor de los valores centrales de la variable. **Distribución platicúrtica:** presenta un reducido grado de concentración alrededor de los valores centrales de la variable.



1.2. Datos agrupados

Es decir, tenemos menos de 20 elementos en la muestra, entonces estos datos son analizados sin necesidad de formar clases con ellos y a esto es a lo que se le llama tratamiento de datos no agrupados.

Cuando la muestra consta de 20 o más datos, lo aconsejable es agrupar los datos en clases y a partir de estas determinar las características de la muestra y por consiguiente las de la población de donde fue tomada.

Distribución de frecuencia para datos no Agrupados (n<20): Es aquella distribución que indica las frecuencias con que aparecen los datos estadísticos, desde el menor de ellos hasta el mayor de ese conjunto sin que se haya hecho ninguna modificación al tamaño de las unidades originales. En estas distribuciones cada dato mantiene su propia identidad después que la distribución de frecuencia se ha elaborado. En estas distribuciones los valores de cada variable han sido solamente reagrupados, siguiendo un orden lógico con sus respectivas frecuencias.

Distribución de frecuencia de clase o de datos Agrupados (n>20): Es aquella distribución en la que la disposición tabular de los datos estadísticos se encuentran ordenados en clases y con la frecuencia de cada clase; es decir, los datos originales de varios valores adyacentes del conjunto se combinan para formar un intervalo de clase.

No existen normas establecidas para determinar cuándo es apropiado utilizar datos agrupados o datos no agrupados; sin embargo, se sugiere que cuando el número total de datos (*N*) es igual o superior 20, se utilizará *la distribución de frecuencia para datos agrupados*, también se utilizará este tipo de distribución cuando se requiera elaborar gráficos lineales como el histograma, el polígono de frecuencia o la ojiva. La razón fundamental para utilizar la distribución de frecuencia de clases es proporcionar mejor comunicación acerca del patrón establecido en los datos y facilitar la manipulación de los mismos. Los datos se agrupan en clases con el fin de sintetizar, resumir, condensar o hacer que la información obtenida de una investigación sea manejable con mayor facilidad.

1.2.1 Tabla de frecuencia

Una distribución de frecuencias o tabla de frecuencias es una ordenación en forma de tabla de los datos estadísticos, asignando a cada dato su frecuencia correspondiente.

Tipos de frecuencia

Frecuencia absoluta

La frecuencia absoluta es el número de veces que aparece un determinado valor en un estudio estadístico.

Se representa por f_i.

La suma de las frecuencias absolutas es igual al número total de datos, que se representa por N.

$$f_1 + f_2 + f_3 + \ldots + f_n = N$$

Para indicar resumidamente estas sumas se utiliza la letra griega Σ (sigma mayúscula) que se lee suma o sumatoria.

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

Frecuencia relativa

La frecuencia relativa es el cociente entre la frecuencia absoluta de un determinado valor y el número total de datos.

Se puede expresar en tantos por ciento y se representa por n_i.

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

La suma de las frecuencias relativas es igual a 1.

1.2.2 Medidas de tendencia central y de posición

Teniendo la siguiente Tabla.

LI LS	Frecuencia	Marca de clase	Límite real inferior	Límite real superior	Frecuencia relativa	Frecuencia Relativa acumulada
5.97 – 6.18	2	6.075	5.965	6.185	2/40 = 0.05	0.05
6.19 – 6.40	5	6.295	6.185	6.405	5/40=0.125	0.175
6.41 –	7	6.515	6.405	6.625	0.175	0.350

6.62						
6.63 – 6.84	13	6.735	6.625	6.845	0.325	0.675
6.85 – 7.06	7	6.955	6.845	7.065	0.175	0.850
7.07 – 7.28	6	7.175	7.065	7.285	0.15	1.000
Total	40				1.000	

Media, media ponderada.

Media ($\bar{x}$).

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i * f_i}{n} = \frac{(6.075)(2)+(6.295)(5)+...+(7.175)(6)}{40} = \frac{12.15+31.475+...+43.05}{40} =$$

$$= \frac{268.52}{40} = 6.713\,pulgadas$$

Donde:

k = número de clases

x_i = marca de clase i

f_i = frecuencia de la clase i

n = $\sum_{i=1}^{k} f_i =$ número de datos en la muestra

Mediana.

Mediana (X_{med}).

$$Xmed = Li + \left[\frac{n/2 - Fme - 1}{fme}\right]A = 6.625 + \left[\frac{40/2 - 14}{13}\right](0.22) = 6.7265$$

Donde:

Li = límite real inferior de la clase que contiene a la mediana

F_{me-1} = sumatoria de las frecuencias anteriores a la clase en donde se encuentra la mediana

f_{me} = frecuencia de la clase en donde se encuentra la mediana

A = amplitud real de la clase en donde se encuentra la mediana

A = LRS-LRI

LRS = límite real superior de la clase que contiene a la mediana

LRI = límite real inferior de la clase que contiene a la mediana

N = número de datos en la muestra

Moda.

Moda (X_{mod}).

$$X\,mod = Li + \left[\frac{d1}{d1 + d2}\right]A = 6.625 + \left[\frac{6}{6+6}\right](0.22) = 6.735\,pu\,lg\,adas$$

Donde:

Li = límite real inferior de la clase que contiene a la moda

$d_1 = |fmo - fmo-1| = |13-7| = 6$

$d_2 = |fmo - fmo+1| = |13-7| = 6$

fmo = frecuencia de la clase que contiene a la moda

fmo-1= frecuencia de la clase anterior a la que contiene a la moda

fmo+1= frecuencia de la clase posterior a la que contiene a la moda

A = amplitud real de la clase que contiene a la moda

A = LRS – LRI

LRS = límite real superior de la clase que contiene a la moda

LRI = límite real inferior de la clase que contiene a la moda

Relación entre media, mediana y moda.

En el caso de distribuciones unimodales, la mediana está con frecuencia comprendida entre la media y la moda (incluso más cerca de la media).

En distribuciones que presentan cierta inclinación, es más aconsejable el uso de la mediana. Sin embargo en estudios relacionados con propósitos estadísticos y de inferencia suele ser más apta la media.

Veamos un ejemplo de cálculo de estas tres magnitudes.

Ejemplo

Consideramos una tabla estadística relativa a una variable continua, de la que nos dan los intervalos, las marcas de clase c_i, y las frecuencias absolutas, n_i.

Intervalos	c_i	n_i
0 -- 2	1	2
2 -- 4	3	1
4 -- 6	5	4
6 -- 8	7	3
8 - 10	9	2

Para calcular la media podemos añadir una columna con las cantidades $n_i\, c_i$. La suma de los términos de esa columna dividida por n=12 es la media:

Intervalos	c_i	n_i	N_i	$n_i\, c_i$
0 -- 2	1	2	2	2
2 -- 4	3	1	3	3
4 -- 6	5	4	7	20
6 -- 8	7	3	10	21
8 - 10	9	2	12	18
	12		64	

$$\overline{x} = \frac{64}{12} = 5,\widehat{3}$$

La mediana es el valor de la variable que deja por debajo de sí a la mitad de las *n* observaciones, es decir 6. Construimos la tabla de las frecuencias absolutas acumuladas, N_i, y vemos que eso ocurre en la modalidad tercera, es decir,

$$\begin{aligned} i &= 3 \qquad \text{Observación} \\ (l_{i-1}, l_i] &= (4;6] \qquad \text{Intervalo donde se encuentra la mediana} \\ M_{ed} &= l_{i-1} + \frac{\frac{n}{2} - N_{i-1}}{n_i} \cdot a_i = 4 + \frac{\frac{12}{2} - 3}{4} \cdot 2 = 5,5 \in (l_{i-1}, l_i] \end{aligned}$$

1.2.3 Medidas de dispersión

Las medias de tendencia central o posición nos indican donde se sitúa un dato dentro de una distribución de datos. Las medidas de dispersión, variabilidad o variación nos indican si esos datos están próximos entre sí o sí están dispersos, es decir, nos indican cuán esparcidos se encuentran los datos. Estas medidas de dispersión nos permiten apreciar la distancia que existe entre los datos a un cierto valor central e identificar la concentración de los mismos en un cierto sector de la distribución, es decir, permiten estimar cuán dispersas están dos o más distribuciones de datos.
Estas medidas permiten evaluar la confiabilidad del valor del dato central de un conjunto de datos, siendo la media aritmética el dato central más utilizado. Cuando existe una dispersión pequeña se dice que los datos están dispersos o acumulados cercanamente respecto a un valor central, en este caso el dato central es un valor muy representativo. En el caso que la dispersión sea grande el valor central no es muy confiable. Cuando una distribución de datos tiene poca dispersión toma el nombre de distribución homogénea y si su dispersión es alta se llama heterogénea.

Desviación media o desviación promedio

La desviación media o desviación promedio es la media aritmética de los valores absolutos de las desviaciones respecto a la media aritmética.
1.1) PROPIEDADES
Guarda las mismas dimensiones que las observaciones. La suma de valores absolutos es relativamente sencilla de calcular, pero esta simplicidad tiene un inconveniente: Desde el punto de vista geométrico, la distancia que induce la desviación media en el espacio de observaciones no es la *natural* (no permite definir ángulos entre dos conjuntos de observaciones). Esto hace que sea muy engorroso trabajar con ella a la hora de hacer inferencia a la población. Cuando mayor sea el valor de la desviación media, mayor es la dispersión de los datos. Sin embargo, no proporciona una relación matemática precisa entre su magnitud y la posición de un dato dentro de una distribución.

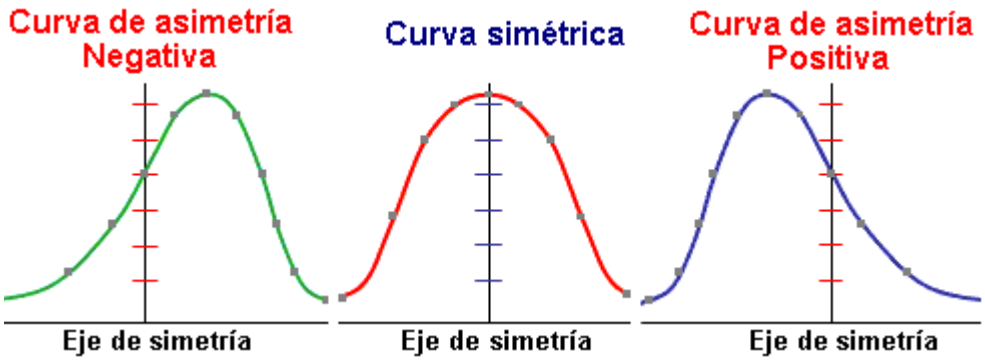
La desviación media al tomar los valores absolutos mide una observación sin mostrar si la misma está por encima o por debajo de la media aritmética.

1.2.4 Medidas de asimetría y curtosis

Las medidas de distribución nos permiten identificar la forma en que se separan o aglomeran los valores de acuerdo a su representación gráfica. Estas medidas describen la manera como los datos tienden a reunirse de acuerdo con la frecuencia con que se hallen dentro de la información. Su utilidad radica en la posibilidad de identificar las características de la distribución sin necesidad de generar el gráfico. Sus principales medidas son la *Asimetría* y la *Curtosis*.

1. ASIMETRÍA

Esta medida nos permite identificar si los datos se distribuyen de forma uniforme alrededor del punto central (Media aritmética). La asimetría presenta tres estados diferentes, cada uno de los cuales define de forma concisa como están distribuidos los datos respecto al eje de asimetría. Se dice que la *asimetría es positiva* cuando la mayoría de los datos se encuentran por encima del valor de la media aritmética, la curva es *Simétrica* cuando se distribuyen aproximadamente la misma cantidad de valores en ambos lados de la media y se conoce como *asimetría negativa* cuando la mayor cantidad de datos se aglomeran en los valores menores que la media.



El *Coeficiente de asimetría*, se representa mediante la ecuación matemática,

$$g_1 = \frac{\frac{1}{n}\sum (X_i - \overline{X})^3 * n_i}{\left(\frac{1}{n}\sum (X_i - \overline{X})^2 * n_i\right)^{\frac{3}{2}}}$$

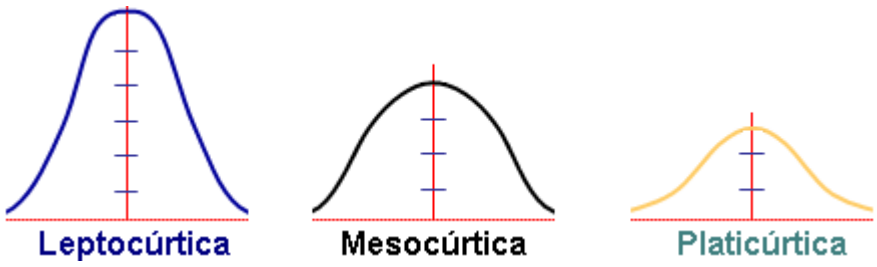
Donde (g1) representa el coeficiente de asimetría de Fisher, (Xi) cada uno de los valores, ($\overline{X}$) la media de la muestra y (ni) la frecuencia de cada valor. Los resultados de esta ecuación se interpretan:

- (g1 = 0): Se acepta que la distribución es Simétrica, es decir, existe aproximadamente la misma cantidad de valores a los dos lados de la media. Este valor es difícil de conseguir por lo que se tiende a tomar los valores que son cercanos ya sean positivos o negativos (± 0.5).
- (g1 > 0): La curva es asimétricamente positiva por lo que los valores se tienden a reunir más en la parte izquierda que en la derecha de la media.
- (g1 < 0): La curva es asimétricamente negativa por lo que los valores se tienden a reunir más en la parte derecha de la media.

Desde luego entre mayor sea el número (Positivo o Negativo), mayor será la distancia que separa la aglomeración de los valores con respecto a la media.

2. CURTOSIS

Esta medida determina el grado de concentración que presentan los valores en la región central de la distribución. Por medio del *Coeficiente de Curtosis*, podemos identificar si existe una gran concentración de valores (*Leptocúrtica*), una concentración normal (*Mesocúrtica*) ó una baja concentración (*Platicúrtica*).



Para calcular el coeficiente de Curtosis se utiliza la ecuación:

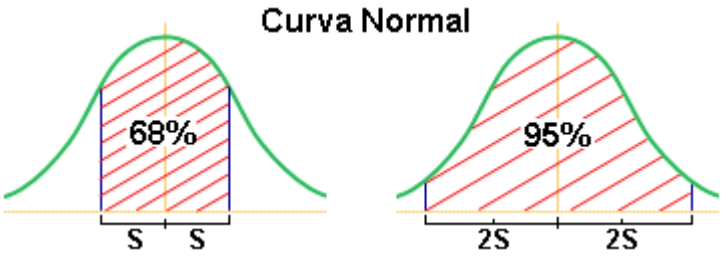
$$g_2 = \frac{\frac{1}{n}\sum (X_i - \overline{X})^4 * n_i}{\left(\frac{1}{n}\sum (X_i - \overline{X})^2 * n_i\right)^2} - 3$$

Donde (g2) representa el coeficiente de Curtosis, (Xi) cada uno de los valores, ($\overline{X}$) la media de la muestra y (ni) la frecuencia de cada valor. Los resultados de esta fórmula se interpretan:

- **(**g2 = 0) *la distribución es Mesocúrtica***:** Al igual que en la asimetría es bastante difícil encontrar un coeficiente de Curtosis de cero (0), por lo que se suelen aceptar los valores cercanos (± 0.5 aprox.).
- (g2 > 0) *la distribución es Leptocúrtica*
- (g2 < 0) *la distribución es Platicúrtica*

Cuando la distribución de los datos cuenta con un coeficiente de asimetría (g1 = ±0.5) y un coeficiente de Curtosis de (g2 = ±0.5), se le denomina Curva Normal. Este criterio es de suma importancia ya que para la mayoría de los procedimientos de la estadística de inferencia se requiere que los datos se distribuyan normalmente.

La principal ventaja de la distribución normal radica en el supuesto que el 95% de los valores se encuentra dentro de una distancia de dos desviaciones estándar de la media aritmética; es decir, si tomamos la media y le sumamos dos veces la desviación y después le restamos a la media dos desviaciones, el 95% de los casos se encontraría dentro del rango que compongan estos valores.



Desde luego, los conceptos vistos hasta aquí, son sólo una pequeña introducción a las principales medidas de Estadística Descriptiva; es de gran importancia que

los lectores profundicen en estos temas ya que la principal dificultad del paquete SPSS radica en el desconocimiento de los conceptos estadísticos.

Las definiciones plasmadas en este capítulo han sido extraídas de los libros *Estadística para administradores* escrito por *Alan Wester* de la editorial *McGraw-Hill* y el libro *Estadística y Muestreo* escrito por *Ciro Martínez* editorial *Ecoe editores* (*Octava edición*). No necesariamente tienes que guiarte por estos libros ya que en las librerías encontraras una gran variedad de textos que pueden ser de bastante utilidad en la introducción a esta ciencia.

1.3. Representaciones gráficas

Los gráficos más usuales para representar variables de tipo nominal son los siguientes:

Diagramas de barras:

Se representa en el eje de ordenadas las modalidades y en abscisas las frecuenciasabsolutas o las frecuencias relativas. Si se intentan comparar varias poblaciones entre sí,usando el diagrama, existen otras modalidades, como las mostradas. Cuand los tamaños de las dos poblaciones son diferentes, es conveniente utilizar las frecuencias relativas, ya que en otro caso podrían resultar engañosas.

1.3.1 Diagrama de Dispersión

Un **diagrama de dispersión** es un tipo de diagrama matemático que utiliza las coordenadas cartesianas para mostrar los valores de dos variables para un conjunto de datos.

Los datos se muestran como un conjunto de puntos, cada uno con el valor de una variable que determina la posición en el eje horizontal y el valor de la otra variable determinado por la posición en el eje vertical.[1] Un diagrama de dispersión se llama también *gráfico de dispersión*.

1.3.2 Diagramas de Tallo y Hojas

El diagrama "tallo y hojas" (Stem-and-Leaf Diagram) permite obtener simultáneamente una distribución de frecuencias de la variable y su representación gráfica. Para construirlo basta separar en cada dato el último dígito de la derecha (que constituye la hoja) del bloque de cifras restantes (que formará el tallo).

Esta representación de los datos es semejante a la de un histograma pero además de ser fáciles de elaborar, presentan más información que estos.

1.3.3 Histogramas

Un **histograma** es una **representación gráfica** de una **variable** en forma de **barras**.

Se utilizan para **variables continuas** o para **variables discretas**, con un gran número de datos, y que se han agrupado en **clases**.

En el **eje abscisas** se construyen unos **rectángulos** que tienen por **base la amplitud del intervalo**, y por **altura**, la **frecuencia absoluta** de cada **intervalo**.

La **superficie** de cada **barra** es **proporcional** a la **frecuencia** de los **valores** representados.

Polígono de frecuencia

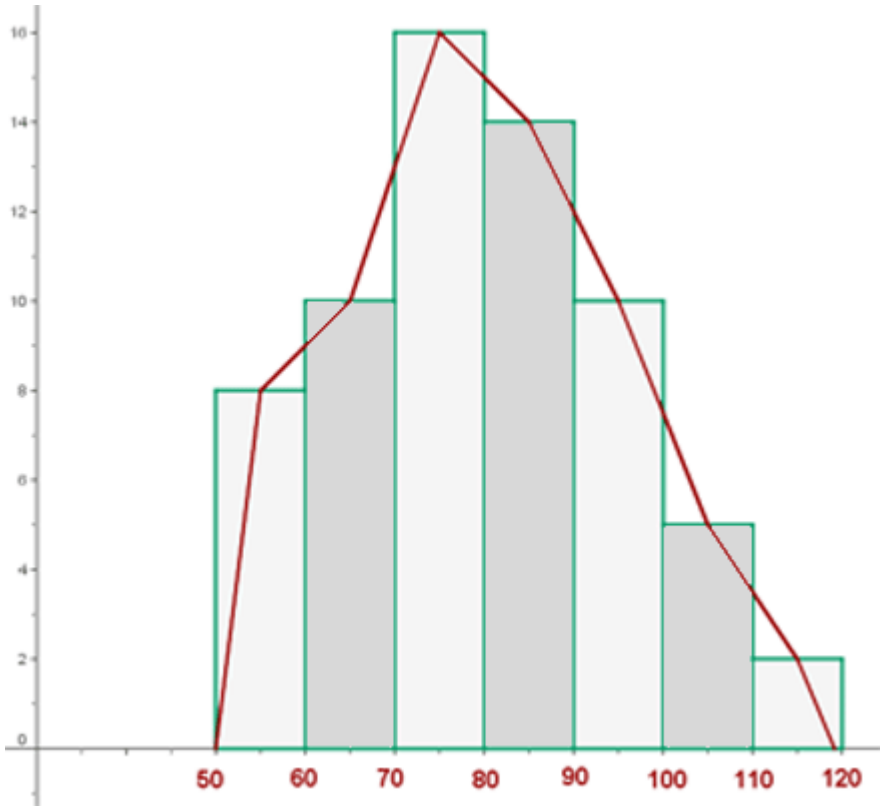
Para construir el **polígono de frecuencia** se toma la **marca de clase** que coincide con el **punto medio** de cada **rectángulo**.

Ejemplo

El peso de 65 personas adultas viene dado por la siguiente tabla:

	c_i	f_i	F_i
[50, 60)	55	8	8
[60, 70)	65	10	18
[70, 80)	75	16	34
[80, 90)	85	14	48
[90, 100)	95	10	58
[100, 110)	110	5	63
[110, 120)	115	2	65

		65	



Histograma y polígono de frecuencias acumuladas

Si se representan las **frecuencias acumuladas** de una **tabla de datos agrupados** se obtiene el **histograma de frecuencias acumuladas** o su correspondiente **polígono**.

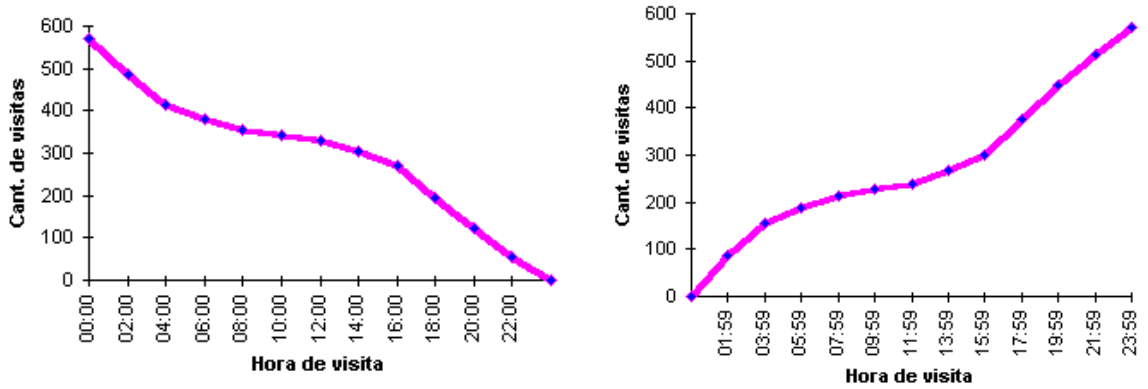
1.3.4 Ojivas

Una gráfica similar al polígono de frecuencias es la **ojiva**, pero ésta se obtiene de aplicar parcialmente la misma técnica a una distribución acumulativa y de igual manera que éstas, existen las **ojivas mayor que** y las **ojivas menor que**.

Existen dos diferencias fundamentales entre las ojivas y los polígonos de frecuencias (y por ésto la aplicación de la técnica es parcial):

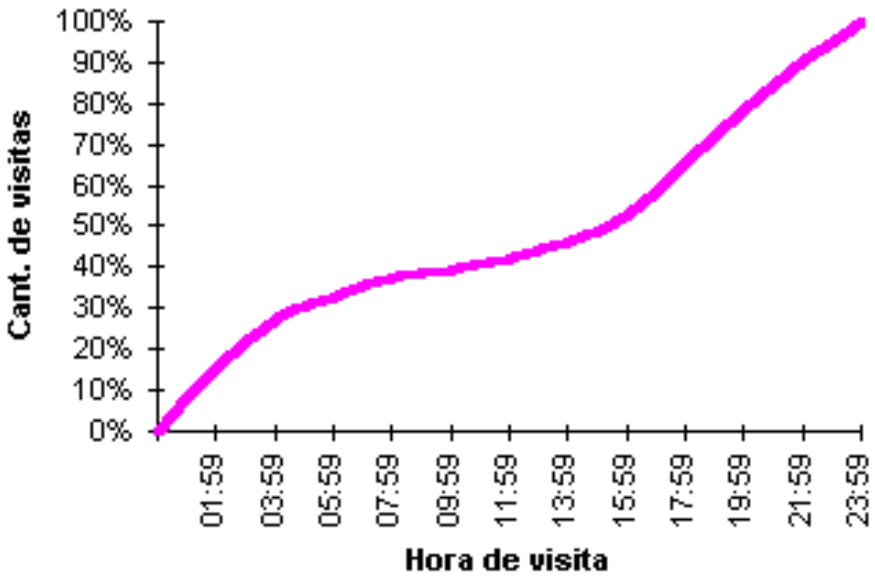
1. Un extremo de la ojiva no se "amarra" al eje horizontal, para la *ojiva mayor que* sucede con el extremo izquierdo; para la *ojiva menor que*, con el derecho.
2. En el eje horizontal en lugar de colocar las marcas de clase se colocan las fronteras de clase. Para el caso de la *ojiva mayor que* es la frontera menor; para la *ojiva menor que*, la mayor.

Las siguientes son ejemplos de ojivas, a la izquierda la *mayor que*, a la derecha la *menor que*, utilizando los datos que se usaron para ejemplificar el histograma:



La ojiva mayor que (izquierda) se le denomina de esta manera porque viendo el punto que está sobre la frontera de clase "4:00" se ven las visitas que se realizaron en una hora *mayor que* las 4:00 horas (en cuestiones temporales se diría: *después de las 4:00 horas*). De forma análoga, en la ojiva menor que la frecuencia que se representa en cada frontera de clase son el número de observaciones *menores que* la frontera señalada (en caso de tiempos sería el número de observaciones *antes* de la hora que señala la frontera).

Si se utiliza una distribución porcentual acumulativa entonces se obtiene una ojiva (*mayor que* o *menor que* según sea el caso) cuyo eje vertical tiene una escala que va del 0% al 100%. El siguiente ejemplo es la misma *ojiva menor que* que se acaba de usar, pero con una distribución porcentual:



1.3.5 Polígono de Frecuencias

Variables discretas

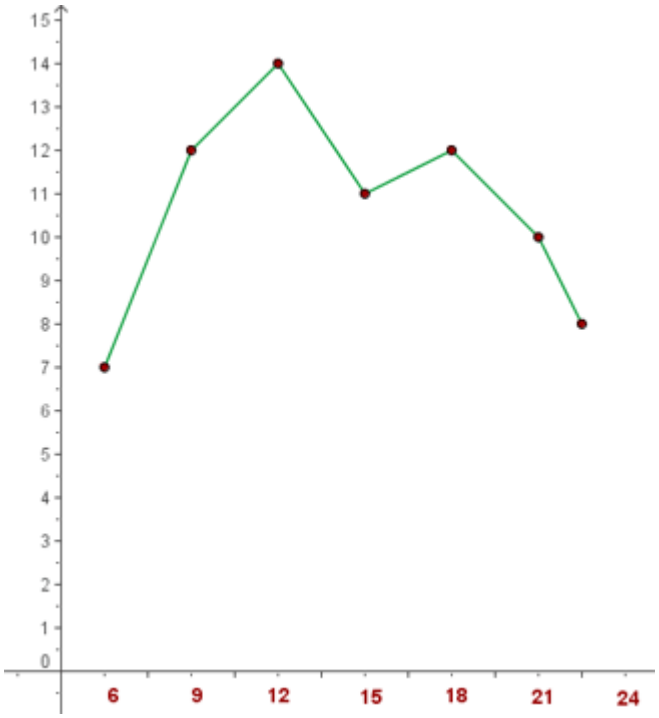
Los polígonos de frecuencias se realizan trazando los **puntos** que

representan las **frecuencias** y uniéndolos mediante **segmentos**.

Ejemplo

Las temperaturas en un día de otoño de una ciudad han sufrido las siguientes variaciones:

Hora	Temperatura
6	7º
9	12°
12	14°
15	11°
18	12°
21	10°
24	8°



Variables continuas o datos agrupados

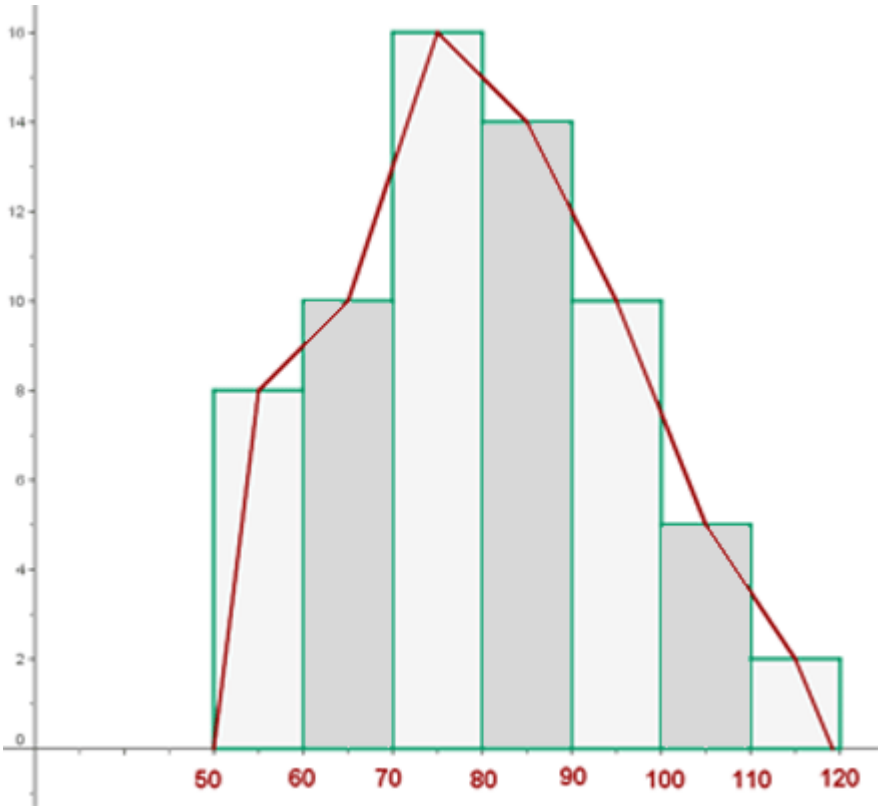
Los polígonos de frecuencias se realizan trazando los **puntos** formados las **marcas de clase** y las **frecuencias,** y uniéndolos mediante **segmentos**.

También se puede construir el **polígono de frecuencia** uniendo los **puntos medios** de cada **rectángulo** de un **histograma**.

Ejemplo

El peso de 65 personas adultas viene dado por la siguiente tabla:

	c_i	f_i	F_i
[50, 60)	**55**	**8**	**8**
[60, 70)	**65**	**10**	**18**
[70, 80)	**75**	**16**	**34**
[80, 90)	**85**	**14**	**48**
[90, 100)	**95**	**10**	**58**
[100, 110)	**110**	**5**	**63**
[110, 120)	**115**	**2**	**65**
		65	



1.3.6 Diagrama de Caja y Ejes

Un **Diagrama de caja** es un gráfico, basado en cuartiles, mediante el cual se visualiza un conjunto de datos. Está compuesto por un rectángulo, la "caja", y dos brazos, los "bigotes".

Es un gráfico que suministra información sobre los valores mínimo y máximo, los cuartiles Q1, Q2 o mediana y Q3, y sobre la existencia de valores atípicos y la simetría de la distribución. Primero es necesario encontrar la mediana para luego encontrar los 2 cuartiles restantes

Cómo expresarlo gráficamente[editar · editar código]

```
                  +-----+-+
 *        o     |-------|     | |---|
                  +-----+-+

+---+---+---+---+---+---+---+---+---+---+---+---+---+
```

```
0          5          10   12
```

- Ordenar los datos y obtener el valor mínimo, el máximo, los cuartiles Q1, Q2 y Q3 y el Rango Inter Cuartilico (RIC)

  En el ejemplo:

  - Valor 7: es el Q1 (25% de los datos)
  - Valor 8.5: es el Q2 o mediana (el 50% de los datos)
  - Valor 9: es el Q3 (75% de los datos)
  - Rango Inter Cuartilico RIC (Q3-Q1)=2

- Para dibujar los bigotes, las líneas que se extienden desde la caja, hay que calcular los límites superior e inferior, Li y Ls, que identifiquen a los valores atípicos.

  Para ello se calcula cuándo se consideran atípicos los valores. Son aquellos inferiores a Q1-1.5*RIC o superiores a Q3+1.5*RIC.

  En el ejemplo:

  - inferior: 7-1.5*2=4
  - superior: 9+1.5*2=12

  Ahora se buscan los últimos valores que **NO** son atípicos, que serán los extremos de los bigotes.

  - En el ejemplo: 5 y 10
    - Marcar como atípicos todos los datos que están fuera del intervalo (Li, Ls).

  En el ejemplo: 0.5 y 3.5

    - Además, se pueden considerar valores extremadamente atípicos aquellos que exceden Q1-3*RIC o Q3+3*RIC.

  De modo que, en el ejemplo:

  - inferior: 7-3*2=1
  - superior: 9+3*2=15

1.3.7 Diagrama de Sectores

También conocido como gráfico de torta o gráfico circular.
Se representan los datos en un círculo, de modo que la frecuencia de cada valor viene dada por un trozo de área del círculo. Así, el círculo queda dividido en sectores cuya amplitud es proporcional a las frecuencias de los valores. Sirve para representar cualquier tipo de variable.

EJEMPLO:

En la clase se realizó la siguiente encuesta: ¿Qué libros prefieres leer?

Los resultados se ordenaron en esta tabla

Tipos de libros	De Misterio	De Aventuras	Historietas	Total
N° de alumnos	15	9	6	30

Si queremos representar esta información en un gráfico de torta tenemos que realizar unos cálculos previamente.

Como la medida de la superficie de cada sector es directamente proporcional a la medida del ángulo central, se hace una proporcionalidad directa entre la cantidad de alumnos que hay en el sector con respecto al total de alumnos y el ángulo del sector (la incógnita) con respecto al ángulo central de todo el círculo, o sea 360°.

Para el sector de libros de misterio tenemos:

30 alumnos ----------360° (todo el círculo)

15 alumnos ---------- incógnita (sector del círculo correspondiente a libros de misterio)

Ángulo sector Misterio = 360° X 15 alumnos / 30 alumnos = 180° (la mitad del círculo)

Ángulo sector Aventuras = 360° X 9 alumnos / 30 alumnos = 108°

Ángulo sector Historietas = 360° X 6 alumnos / 30 alumnos = 72°

Si sumamos la amplitud de los tres sectores nos tiene que dar el círculo completo:

180° + 108° + 72° = 360°