CAPÍTULO 2

ESTADÍSTICA DESCRIPTIVA

2.1. Introducción.

Aun cuando en la actualidad la mayor parte del uso de la estadística esta dirigido a la Inferencia, la Estadística descriptiva tiene una utilidad importante fundamentalmente en la primera fase de una investigación.
La estadística descriptiva se refiere al proceso en el que los datos son ordenados, resumidos y clasificados con objeto de tener una visión más precisa y conjunta de las observaciones, intentando descubrir posibles relaciones entre los datos, observando similitudes y diferencias entre los mismos, destacando hechos de posible interés, entre otras cosas. Esto es, tiene como objetivo caracterizar, describir y extraer conclusiones sobre los datos de forma tal que permitan sugerir cuestiones a analizar con mayor profundidad, llegando en ocasiones a ayudar en el establecimiento de las primeras hipótesis acerca de la naturaleza del fenómeno que se estudia o investiga.
La Estadística Descriptiva además permite estudiar si pueden mantenerse algunos supuestos necesarios para procesos de inferencia, tales como la de simetría, normalidad, homocedasticidad, etc.
Este tema tiene como propósito principal, introducir técnicas que permitan tanto matemática como gráficamente describir apropiadamente un conjunto de datos y al finalizar el mismo, el estudiante debe estar en capacidad, una vez coleccionados los datos, de:

- Ordenarlos y clasificarlos,
- presentarlos a través de cuadros estadísticos y gráficos,
- calcular medidas descriptivas numéricas y
- analizar la información obtenida en los pasos anteriores.

2.2. Organización de los Datos

La organización de los datos consiste en una agrupación apropiada de los mismos. Es importante dicha agrupación, ya que por lo general la información obtenida de un estudio implica gran cantidad de datos que no es fácil interpretar directamente. Esta organización depende del tipo de variable que se maneje. Por lo tanto, se debe estudiar como realizar dicha agrupación cuando la variable es cualitativa y cuando es cuantitativa.
Los datos se organizan en una distribución de frecuencias, la cual es una tabla resumen en la que los datos se disponen en agrupamientos o categorías convenientemente establecidas de clases ordenadas numéricamente. Su estructura dependerá del tipo de variable a analizar.

2.2.1. Organización de Datos Cualitativos

Cuando los datos son cualitativos de escala nominal, la organización consiste en la construcción de una tabla de frecuencias con los siguientes columnas: la enumeración de las distintas modalidades que presenta la variable, el número de datos que corresponde a cada modalidad (frecuencia absoluta, f_i) y la proporción que cada uno de ellos representa con respecto al total (frecuencia relativa, fr_i). La tabla 2.1 muestra la estructura de una tabla de frecuencias para este caso.

Tabla 2.1: Tabla de Frecuencias para datos cualitativos de escala nominal

Modalidades	f_i	fr_i
1	f_1	fr_1
2	f_2	fr_2
$\vdots$	$\vdots$	$\vdots$
k	f_k	fr_k

donde

$\sum_{i=1}^{k} f_i = n$: representa el número total de datos.

$fr_i = \frac{f_i}{n}$ y debe cumplirse que $\sum_{i=1}^{k} fr_i = 1$

Ejemplo 2.1 *A continuación se muestran los resultados obtenidos al aplicar una encuesta a 50 estudiantes de FACES donde se les preguntó sobre la carrera que estudiaban (A: Administración, C: Contaduría, E: Economía, ES: Estadística):*
La variable en este ejemplo es la carrera que estudian las personas, la cual es cualitativa de escala nominal. Al organizar los datos en una distribución de frecuencia se obtiene la tabla 2.2

C	A	A	C	C	A	A	E	A	C
E	E	C	ES	E	A	C	C	A	C
C	A	ES	C	E	A	A	C	A	C
C	C	A	E	E	A	C	C	C	A
C	C	A	C	C	C	C	ES	A	E

Tabla 2.2: Tabla de frecuencia para los datos del ejemplo 2.1

Carrera	f_i	fr_i
Administración	16	0.32
Contaduría	23	0.46
Economía	8	0.16
Estadística	3	0.06

Si los datos son cualitativos de escala ordinal, su organización implican dos cosas: en primer lugar, las clases llevan un orden preestablecido por las modalidades de la variable; en segundo lugar se incorporan a la tabla 2.1, columnas que muestren la frecuencia absoluta acumulada, F_i, y la proporción que cada uno de ellos representa con respecto al total, frecuencia relativa acumulada, Fr_i. La tabla 2.3 muestra la estructura de una tabla de frecuencias para este caso.

Tabla 2.3: Tabla de Frecuencias para datos cualitativos de escala ordinal

Modalidades	f_i	fr_i	F_i	Fr_i
1	f_1	fr_1	F_1	Fr_1
2	f_2	fr_2	F_2	Fr_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	f_k	fr_k	F_k	Fr_k

donde

$$F_l = \sum_{i=1}^{l} f_i$$

$$Fr_l = \sum_{i=1}^{l} fr_i = \frac{F_l}{n} \text{ y debe cumplirse que } Fr_k = 1$$

Ejemplo 2.2 *Los siguientes datos corresponden a una consulta realizada a 45 pacientes acerca de su percepción sobre la atención prestada en el Instituto Autónomo Hospital Universitario de Los Andes, IAHULA (MB: Muy Bueno, B: Bueno, A: Aceptable, M: Malo, MM: Muy Malo):*

MB	B	B	A	A	M	A	MM	B	A
B	B	MM	MB	A	A	M	M	B	B
M	A	MM	MB	B	A	B	MB	A	B
B	M	M	B	B	A	B	B	M	A
MB	B	M	MM	A					

Aquí la variable es la percepción de los pacientes, la cual es cualitativa de escala ordinal. En la tabla 2.4 se muestra la organización de estos datos en una tabla de frecuencias.

Tabla 2.4: Percepción de la calidad de atención en pacientes del IAHULA

Percepción	f_i	fr_i	F_i	Fr_i
Muy buena	5	0.1111	5	0.1111
Buena	16	0.3555	21	0.4667
Aceptable	12	0.2667	33	0.7333
Mala	8	0.1778	41	0.9111
Muy mala	4	0.0889	45	1

2.2.2. Organización de Datos Cuantitativos

Si los datos son cuantitativos, los mismos pueden ser discretos o continuos. Para su organización se usa un procedimiento similar al utilizado con los datos cualitativos, considerando otros aspectos que la hacen más laboriosa.

Ejemplo 2.3 *En un curso de estadística I se observa el número de asignaturas aprobadas por cada uno de los alumnos:*

3	*6*	*1*	*2*	*3*	*7*	*5*	*5*	*4*	*5*
3	*5*	*7*	*2*	*3*	*4*	*2*	*7*	*6*	*1*
4	*3*	*2*	*4*	*6*	*3*	*7*	*6*	*1*	*1*
2	*3*	*5*	*2*	*7*	*5*	*5*	*7*	*6*	*1*
4	*5*								

Ejemplo 2.4 *Para los alumnos del ejemplo 2.3, se obtiene su estatura:*

1.55	*1.55*	*1.58*	*1.57*	*1.59*	*1.60*	*1.65*	*1.70*	*1.73*	*1.57*
1.56	*1.60*	*1.61*	*1.62*	*1.69*	*1.68*	*1.71*	*1.71*	*1.74*	*1.79*
1.77	*1.67*	*1.65*	*1.65*	*1.59*	*1.58*	*1.55*	*1.63*	*1.62*	*1.61*
1.64	*1.68*	*1.70*	*1.72*	*1.72*	*1.76*	*1.74*	*1.71*	*1.75*	*1.75*
1.58	*1.71*								

La variable número de asignaturas en el ejemplo 2.3 es discreta, mientras que en el ejemplo 2.4, la variable estatura es continua. En estos casos la tabla de frecuencias contiene los siguientes elementos:

- **Intervalos de Clase:** El intervalo total en que están repartidas las observaciones es dividido en k intervalos parciales. A estos intervalos se les denomina intervalos de clase o, simplemente clases. Deben ser excluyentes

- **Límites de Clase:** Extremos de los intervalos de clase. Al menor de estos valores se le llama límite inferior y al mayor, limite superior.

- **Marcas de Clase** (m_i)**:** Punto medio o centro de intervalo. Es una forma abreviada de representar el intervalo. De esta forma, todos los cálculos que se realizan como si en lugar de tener n_i valores en la clase i, se tiene n_i veces el mismo valor, m_i

- **Frecuencia Absoluta** (f_i)**:** Número de observaciones contenidas o incluidas en una clase. Se debe satisfacer la siguiente igualdad

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

  donde n es el número total de datos.

- **Frecuencia Relativa** (fr_i)**:** Proporción de los datos contenidos en la clase. Se obtiene al dividir la frecuencia absoluta entre el número total de observaciones. Debe cumplirse que

$$1 = \sum_{i=1}^{k} fr_i$$

- **Frecuencia Absoluta Acumulada** (F_i)**:** Suma de frecuencias absolutas hasta la clase correspondiente. De esta forma, la frecuencia acumulada para la clase k es el número total de datos, n.

- **Frecuencia Relativa Acumulada** (Fr_i)**:** Suma de las Frecuencias Relativas hasta la clase correspondiente. Se pueden obtener dividiendo la frecuencia absoluta acumulada entre el número total de observaciones. Para la clase k se cumple que $1 = Fr_k$.

Nota: En el caso discreto, cuando el número de valores diferentes que puede tomar la variable es pequeño, entonces cada uno de ellos representa una clase. De esta forma las marcas de clase coinciden con las clases. Lo mismo es valido en el caso continuo, cuando el número de datos es pequeño.
Para construir una tabla o distribución de frecuencias, en el caso de variables cuantitativas se debe seguir el siguiente procedimiento:

1. Obtener los extremos del intervalo total (V_{max} y V_{min}).

2. Obtener el rango o recorrido de la variable, R = V_{max}-V_{min}.

3. Determinar el número de clases y la amplitud de las mismas.Para determinar el número de clases no existe una regla fija. Una primera aproximación es tomar

$$K = \text{N}^\circ \text{ de clases} = \sqrt{n}$$

   Esta aproximación no siempre es conveniente, sobre todo cuando n es grande.

   Existe una fórmula para calcular el número óptimo de clases, denominada fórmula de Stugers

$$K = \text{N}^\circ \text{ de clases} = 1 + 3{,}3 \log n$$

   Cuando se particionan los datos en clases, es generalmente recomendado usar entre 5 y 15 clases. Fuera de estos extremos, la organización resulta poco eficiente. Si hay pocas clases la pérdida de información es por lo general significativa. Si hay muchas clases y adicionalmente el número de datos es pequeño, las frecuencias de clases tienden a subir y bajar de un manera desordenada evitando que se produzca una distribución ideal de los datos.

   Una vez que se toma una decisión en cuanto al número de clases, la amplitud de las clases, es simplemente

$$A = Amplitud = \frac{R}{K}$$

   Esto permite en resumen, particionar los datos en K clases, cada una con amplitud A. Es importante hacer notar que, no siempre es posible contar con clases de igual amplitud.

   Si la amplitud de los intervalos no es constante, se debe corregir entonces las frecuencias, dividiendo las mismas por la amplitud del intervalo.

4. Construir los Intervalos de Clase: Para construir la primera clase, seleccionamos como un límite inferior el valor mínimo (V_{min}). El limite superior se obtiene al sumarle al limite inferior la amplitud, A. Para la segunda clase se tiene que el limite inferior es el limite superior de la primera clase y el limite superior, resulta de sumarle a este, A. Siguiendo este procedimiento construimos las k clases. Como el limite superior de una clase representa el limite inferior de la clase siguiente, conviene considerar las clases como intervalos del tipo [L_{inf} - L_{sup}); esto es, intervalos cerrados por la izquierda y abiertas por la derecha.

5. Calcular las marcas de clase (m_i): Las marcas de clase están representadas por los puntos medios de los intervalos de clase, es decir, $m_i = ls_i - li_i$

6. Obtener las frecuencias absolutas, relativas, absolutas acumuladas y relativa acumulada. La tabla 2.5 muestra la estructura de una tabla de frecuencias para datos cuantitativos

Tabla 2.5: Tabla de Frecuencias para datos cuantitativos

Clases	m_i	f_i	fr_i	F_i	Fr_i
$[li_1 - ls_1)$	m_2	f_1	fr_1	F_1	Fr_1
$[li_2 - ls_2)$	m_2	f_2	fr_2	F_2	Fr_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[li_k - ls_k)$	m_k	f_k	fr_k	F_k	Fr_k

Ejemplo 2.5 *A continuación se muestra la información sobre el número de hijos que tienen 40 Mujeres extraídas al azar de la ciudad de Mérida.*

1	*1*	*3*	*3*	*2*	*4*	*4*	*1*
1	*2*	*1*	*3*	*3*	*2*	*1*	*3*
2	*1*	*2*	*2*	*4*	*3*	*4*	*4*
4	*0*	*3*	*0*	*4*	*1*	*5*	*2*
2	*3*	*3*	*4*	*4*	*4*	*1*	*2*

Antes de organizar los datos en una distribución de frecuencia, observemos que la variable es discreta y además posee pocos valores diferentes, pues su rango esta dado por $\{0, 1, 2, 3, 4, 5\}$. *Entonces las clases de la distribución de frecuencia están dadas por los valores individuales de la variable. En la tabla 2.6 se presenta la organización de estos datos.*

Tabla 2.6: Distribución del N° de Hijos que tienen 40 Mujeres

N° de Hijos	f_i	fr_i	F_i	Fr_i
0	2	0,050	2	0,050
1	9	0,225	11	0,275
2	9	0,225	20	0,500
3	9	0,225	29	0,725
4	10	0,250	39	0,975
5	1	0,025	40	1

En la tabla 2.6 se observa entre otras cosas que el 97.5 % de la mujeres en la muestra tienen 4 o menos hijos. Obsérvese que el 25 % de las mujeres encuestadas tienen 3 hijos, representado este el valor más frecuente. Estos porcentajes se obtienen simplemente al multiplicar los valores de fr_i *y* Fr_i *por* 100. *Es decir,* $97{,}5\,\% = 0{,}975 * 100$ *y* $25\,\% = 0{,}250 * 100$.

Ejemplo 2.6 *Para los datos del ejemplo 2.3, obtener la tabla o distribución de frecuencia. Puede observarse en la tabla 2.7 que el 19.05 % de los alumnos han aprobado 5 materias y el*

Tabla 2.7: Distribución del Nº de asignaturas aprobadas.

N° de Asig. aprobadas	f_i	fr_i	F_i	Fr_i
1	5	0,119	5	0,119
2	6	0,1429	11	0,2619
3	7	0,1666	18	0,4285
4	5	0,119	23	0,5475
5	8	0,1905	31	0,738
6	5	0,119	36	0,857
7	6	0,1429	42	1

54.75 % han aprobado menos de 5 materias. Nótese que un 26 % de los estudiantes tienen más de 5 materias aprobadas

Ejemplo 2.7 *Los siguientes datos corresponden a la edad de 40 estudiantes de FACES.*

30	*28*	*22*	*28*	*34*	*32*	*32*	*23*	*28*	*35*
34	*28*	*20*	*29*	*21*	*30*	*30*	*19*	*27*	*19*
25	*30*	*34*	*32*	*31*	*24*	*32*	*20*	*21*	*30*
31	*19*	*18*	*27*	*19*	*26*	*26*	*27*	*29*	*34*

Si se organizan los datos en una distribución de frecuencia cuyas clases son valores individuales, como en el ejemplo anterior, el arreglo resultante es el mostrado en la tabla 2.8.

Esta agrupación de los datos es poco eficiente ya que la variable edad posee muchos valores diferentes (modalidades), lo que conlleva a un arreglo que no tiene una fácil interpretación. Para mejorar la organización de los datos, es necesario considerar a las clases como intervalos. El procedimiento para tal caso se describe a continuación.

Tabla 2.8: Distribución de frecuencia de las edades en clases individuales.

Edad	f_i	fr_i	F_i	Fr_i
18	1	0,025	1	0,025
19	4	0,100	5	0,125
20	2	0,050	7	0,175
21	2	0,050	9	0,225
22	1	0,025	10	0,250
23	1	0,025	11	0,275
24	1	0,025	12	0,300
25	1	0,025	13	0,325
26	2	0,050	15	0,375
27	3	0,075	18	0,450
28	4	0,100	22	0,550
29	2	0,050	24	0,600
30	5	0,125	29	0,725
31	2	0,050	31	0,775
32	4	0,100	35	0,875
34	4	0,100	39	0,975
35	1	0,025	40	1

a) *Identificación de los valores extremos del intervalo total.*

$$V_{max} = 35 \ y \ V_{min} = 18$$

b) *Calculo del Rango.*

$$R = V_{max} - V_{min} = 35 - 18 = 17$$

c) *Determinación del Número de Clases (K) y de la amplitud de las clases (A) Para determinar el número de clases se usa la regla de Sturges, obteniéndose:*

$$K = 1 + 3{,}3\log(n) = 1 + 3{,}3\log(40) = 6{,}28$$

Por lo tanto se deben tener aproximadamente 6 clases. La amplitud de las clases está dada por:

$$A = \frac{R}{K} = \frac{17}{6{,}28} = 2{,}7$$

lo cual se puede aproximar a 3, ya que, se ha asumido que la variable edad es discreta.

d) *Construcción de los intervalos de clases.*

- *El primer intervalo se construye utilizando como limite inferior el valor mínimo de los datos, en este caso 18, y el limite superior se obtiene al sumarle la amplitud (A) al limite inferior, es decir, 18 + 3 =21. Por lo tanto el primer intervalo es* $[18-21)$.
- *El segundo intervalo tiene como limite inferior el limite superior de la clase anterior, es decir, 21, y el limite superior se obtiene al sumarle la amplitud al limite inferior, es decir, 21+3=24. Por lo tanto el segundo intervalo es* $[21-24)$.
- *Los demás intervalos se obtienen de manera similar al segundo intervalo. El último intervalo construido debe contener al valor máximo. Si el límite superior de este intervalo coincide con el valor máximo de los datos, entonces el intervalo debe ser cerrado, es decir, de la forma [,].*

e) *Los intervalos de clases obtenidos al seguir el procedimiento anterior son:*

[18 - 21)
[21 - 24)
[24 - 27)
[27 - 30)
[30 - 33)
[33 - 36)

f) *Calculo de las marcas de clase: La marcas de clase para cada una de los intervalos de clases se muestran a continuación*

Clase	Marca de Clase
$[18-21)$	$\frac{18+21}{2} = 19{,}5$
$[21-24)$	$\frac{21+24}{2} = 22{,}5$
$[24-27)$	$\frac{24+27}{2} = 25{,}5$
$[27-30)$	$\frac{27+30}{2} = 28{,}5$
$[30-33)$	$\frac{30+33}{2} = 31{,}5$
$[33-36)$	$\frac{33+36}{2} = 34{,}5$

g) *Calculo de las frecuencias absolutas y relativas.*

- *Las frecuencias absolutas* (f_i) *representan el número de observaciones que se encuentran en el intervalo i. Para el primer intervalo de clase por ejemplo, la frecuencia absoluta* (f_1) *es 7, esto quiere decir que hay 7 estudiantes con edades mayores o iguales a 18 años pero menores a 21 años.*

- *Las frecuencias relativas* (fr_i) *se obtienen al dividir la frecuencia absoluta entre el numero de observaciones. Para el primer intervalo de clase* $fr_1 = \frac{7}{40} = 0{,}175$. *donde 40 es el número de observaciones.*
- *Las frecuencias acumuladas* (F_i) *se obtienen al sumar las frecuencias absolutas de esa clase con las anteriores. En este caso, la frecuencia acumulada del tercer intervalo de clase es* $F_3 = f_1 + f_2 + f_3 = 7 + 4 + 4 = 15$. *En general, la frecuencia acumulada para la clase* c $(1 < c < k)$ *está dada por* $F_c = \sum_{i=1}^{c} f_i$
- *Las frecuencias relativas acumuladas* (Fr_i) *se obtienen al sumar las frecuencias relativas de esa clase con las anteriores. En este caso, la frecuencia relativa acumulada del tercer intervalo de clase es* $Fr_3 = fr_1 + fr_2 + fr_3 = 0{,}175 + 0{,}100 + 0{,}100 = 0{,}375$. *Otra manera de obtener esta valor es dividir la frecuencia acumulada entre el numero de observaciones,* $Fr_3 = \frac{15}{40} = 0{,}375$

De esta forma, en la tabla 2.9 se muestra la distribución de frecuencia para los datos del ejemplo 2.7. Esta tabla presenta los datos de manera mas resumida que la tabla 2.8, lo cual

Tabla 2.9: Distribución de frecuencia de las edades de 40 estudiantes.

Edad	f_i	fr_i	F_i	Fr_i
$[18-21)$	7	0,175	7	0,175
$[21-24)$	4	0,100	11	0,275
$[24-27)$	4	0,100	15	0,375
$[27-30)$	9	0,225	24	0,600
$[30-33)$	11	0,275	35	0,875
$[33-36)$	5	0,125	40	1

la hace más fácil de interpretar. Por ejemplo, se puede decir que un 27.5 % de los estudiantes tienen edades inferiores a 33 años y mayores o iguales a 30 años. El 60 % de los estudiantes tiene edades inferiores a 30 años.

Tablas como las anteriores se utilizan cuando se está estudiando una variable. Existen situaciones en las que se registra información acerca de dos o más variables para cada individuo o elemento. Si este es el caso, la serie de datos se dice es multidimensional. Para el caso de dos variables, digamos A y B, los datos se pueden organizar mediante el uso de una tabla de doble entrada, denominada distribución conjunta o, tabla de contingencia en el caso de variables cualitativas. Esta tabla se construye enumerando en la parte superior las modalidades o valores de una variable (variable columna) y en el extremo derecho las modalidades de la otra variable (variable fila). La tabla 2.10

muestra la estructura de una distribución conjunta o tabla de contingencia. Esta es una tabla con r filas y c columnas, por tanto, tiene $r \times c$ celdas. La celda correspondiente a la fila i y la columna j, C_{ij}, contiene el número de elementos que presenta simultáneamente la categoría i de la variable fila y la categoría j de la variable columna. Por ejemplo, si sobre un conjunto de individuos se miden las variables estado civil y nivel educativo, la celda C_{ij} registrara el número de individuos que presentan la modalidad i de estado civil y la modalidad j de nivel educativo.

Tabla 2.10: Tabla de Contingencia

		Variable B				Totales
		B_1	B_2	...	B_k	
Variable A	A_1					
	A_2					
	$\vdots$					
	A_k					
Totales						

Ejemplo 2.8 *Los siguientes datos corresponden a la región de procedencia (1=Centro (C), 2=Occidente (Occ), 3=Oriente (Or) y 4=Región Zuliana (RZ)) y carrera que cursan (1=Admintración (Ad), 2=Contaduria (C), 3=Economía (Ec) y 4=Estadística (Es))* 40 *estudiantes de FACES.*

Estudiante	1	2	3	4	5	6	7	8	9	10
Procedencia	1	2	2	4	3	4	3	1	2	2
Carrera	4	1	1	4	4	1	3	3	3	2
Estudiante	11	12	13	14	15	16	17	18	19	20
Procedencia	1	1	1	2	3	3	4	1	2	3
Carrera	1	2	3	2	1	4	1	3	1	2
Estudiante	21	22	23	24	25	26	27	28	29	30
Procedencia	3	3	3	4	4	2	4	4	3	3
Carrera	3	4	2	2	3	1	2	2	2	1
Estudiante	31	32	33	34	35	36	37	38	39	40
Procedencia	1	2	3	4	3	1	4	2	2	3
Carrera	1	2	2	2	4	4	4	3	2	2

En la tabla 2.11 se presenta la organización de estos datos en una tabla de contingencia. Obsérvese por ejemplo, que la celda $C_{22} = 4$, indica que existen 4 estudiantes que proceden de la región occidental del pais y estudian contaduría.

Tabla 2.11: Tabla de Contingencia para los datos del ejemplo 2.8

		Carrera				Totales
		C	*Occ*	*Or*	*RZ*	
Procedencia	*Ad*	2	1	3	2	8
	C	4	4	2	0	10
	Ec	2	6	1	4	13
	Es	2	4	1	2	93
Totales		10	15	7	8	40

La tabla de contingencia puede construirse en términos de frecuencias relativas, simplemente dividiendo los elementos de la tabla inicial entre el total de datos. De esta forma las celdas de la nueva tabla representa proporciones. La tabla siguiente muestra la tabla de contingencia para en términos de proporciones para los datos del ejemplo 2.8

Tabla 2.12: Tabla de Contingencia para los datos del ejemplo 2.8 en términos de proporciones

		Carrera				Totales
		C	*Occ*	*Or*	*RZ*	
Procedencia	*Ad*	0,05	0,025	0,075	0,05	0,2
	C	0,1	0,1	0,05	0	0,25
	Ec	0,05	0,15	0,025	0,1	0,325
	Es	0,05	0,1	0,025	0,05	0,225
Totales		0,25	0,375	0,175	0,2	1

2.3. Presentación de los Datos

Una vez organizados los datos, corresponde ahora representarlos de tal manera que se pueda comprender en forma sencilla y apreciar así, sus particularidades. Tres formas existen para representar un conjunto de datos: escrita, tabular y gráfica. De las tres, la menos eficiente es la escrita. La principal razón es que por lo general, hay dificultad para entender tanto número dentro del texto. En esta sección se consideran las formas tabular y gráfica.

2.3.1. Representación Tabular

Anteriormente se dijo que para efectuar una organización de datos se utiliza una tabla de frecuencia, la cual tiene como finalidad presentar en forma ordenada los datos de tal manera que permitan al lector tener una visión clara del conjunto a analizar. Sin embargo, esa tabla no debe ser utilizada para representar dichos datos, pues dicha representación debe ser directa, concisa, clara, fácil de leer y tener un título que se explique por sí mismo.
La representación tabular debe tener, por lo menos, las siguientes partes: título; columna base; encabezados de las columnas, cuerpo, notas al pie y fuente.

Título: Describe el contenido de la tabla e indica su número de orden.

Columna base: Describe las modalidades o categorías de la variable.

Encabezados de las columnas: Identifica el tipo de datos y descripciones alineados verticalmente.

Cuerpo: Espacio que contiene los datos numéricos y los términos o frases descriptivos. Constituye el mensaje de la tabla.

Notas al pie: Explican detalles del contenido de la tabla.

Fuente: Se usan para indicar de donde vienen los datos en el caso de que no sean propios.

Existen varios tipos de tablas:

1. Generales o de referencia. Ubicados por lo general en un apéndice, son muy extensas y dan información muy detallada
2. De texto o de resumen. Tiene un tamaño reducido y trata de resaltar con la mayor intensidad posible, un dato, o varios datos estrechamente relacionados. Su construcción es relativamente sencilla.

La construcción de cuadros para la representación tabular de datos, generalmente considera los siguientes elementos:

1. Redacción del título e identificación. Debe redactarse con claridad y debe expresar en forma concisa los datos que se presentan en el cuadro. Se ubica por lo general en la parte superior del cuadro. El título expresa en el siguiente orden: qué, dónde, cómo y cuándo se ha clasificado.

2. Nota de introducción y al pie. Pueden a añadirse al cuadro una nota de introducción (abajo del título), una o más notas al pie (abajo del cuerpo del cuadro) y una nota indicando la fuente (al final del cuadro).

3. Observaciones generales. Si se usan porcentajes debe estar claro que representa el 100 %.

Ejemplo 2.9 *La tabla 2.9 muestra la organización en una tabla de frecuencia para los datos del ejemplo 2.7. La tabla muestra la representación tabular de dichos datos.*

Tabla 2.13: Distribución de las edades de 40 estudiantes de FACES.

Edades	N° de estudiantes	% de estudiantes
$[18-21)$	7	0,175
$[21-24)$	4	0,100
$[24-27)$	4	0,100
$[27-30)$	9	0,225
$[30-33)$	11	0,275
$[33-36)$	5	0,125

Obsérvese que la tabla 2.13 representa sólo una parte de la 2.9. Su titulo describe con claridad el contenido de dicha tabla. El encabezado de la columnas identifican en forma precisa los datos que allí se muestran. Es más claro decir número de estudiantes o porcentajes de estudiantes, que f_i o fr_i. Cualesquiera otra información que se requiera de dichos datos, se pueden obtener haciendo uso de las columnas que se muestran en esta representación tabular.

2.3.2. Representación Gráfica

En la sección anterior se discutió como resumir un conjunto de datos procedentes de una determinada población. Este método tiene como objetivo fundamental facilitar la comprensión y análisis de ese conjunto de datos. En los análisis estadísticos, es frecuente utilizar representaciones visuales complementarias de las tablas que resumen los datos de estudio, lo que permite esclarecer aun más las características asociadas con la población. La representación gráfica de datos presenta como su principal ventaja, la capacidad de ofrecer de forma rápida un panorama general de los resultados y permite captar las características fundamentales de los datos. Entre las funciones que cumple la representación gráfica de datos se pueden señalar las siguientes:

- Hacen más visibles los datos.
- Poner de manifiesto sus variaciones y su evolución en el tiempo o espacio.
- Evidenciar las relaciones entre los diversos elementos de un sistema o de un proceso y representar la correlación entre dos o más variables.

- Sistematizar y sintetizar los datos.
- Aclarar y complementar las tablas y las exposiciones teóricas o cuantitativas.
- Pueden sugerir hipótesis nuevas.

Existe una gran variedad de gráficos y la selección apropiada de algunos de ellos para la representación de la información dependerá, entre otras cosas, del tipo de datos y la preferencia e interés del investigador. La tabla 2.14 muestra los gráficos más apropiados de acuerdo al tipo de variable.

Tabla 2.14: Tipos de Gráficos de acuerdo al tipo de variable

Variable	**Escala**	**Gráfico**
Cualitativa	Nominal	Barra, pictograma, sectores
	Ordinal	Curvas, Barras, sectores
Cuantitativa		Curvas, histograma, diagrama de línea, polígono de frecuencias, ojiva

La representación gráfica, al igual que la tabular, debe contener una identificación, un título, el gráfico y sus leyendas. Estos elementos, salvo el gráfico como tal, son equivalentes a los de una representación tabular.
Aún cuando parece tarea sencilla, en la construcción de una representación gráfica se puede cometer error; errores de forma y errores de contenido. El uso de gráficos no acordes con el tipo de datos y la omisión de leyendas para identificar claves o símbolos, representan los errores de contenido más comunes.
Los principales errores de forma son:

1. omite la identificación.
2. no se usa un título, título extremadamente extenso
3. gráficos muy cargados

1. **Gráficos para Variables Cualitativas**

   - **Diagrama de Barras:** Gráfica que representa en el eje de las abscisas (X), las distintas categorías de la variable y en eje de las ordenadas (Y), la frecuencia absoluta o la frecuencia relativa asociada con cada categoría. A cada categoría se le asocia una barra vertical cuya longitud es proporcional a la frecuencia (bien sea absoluta o relativa), como se muestra en la figura 6.5. Puede ser usado para comparar poblaciones.

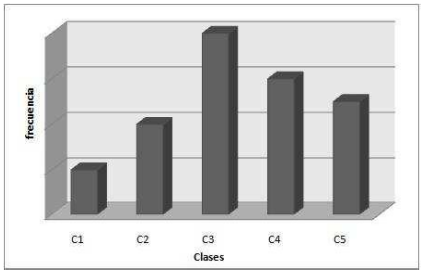

Figura 2.1: Gráfico de Barras

Ejemplo 2.10 *La figura 2.2 presenta el diagrama de barras para los datos del ejemplo 2.1.*

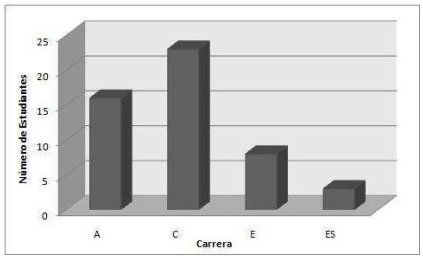

Figura 2.2: Distribución de las carreras de FACES

- **Pictogramas:** se usan para hacer mas llamativas la representación. En lugar de barras, para graficar las frecuencias, se usan dibujos alusivos al tema de estudio. Cada dibujo representa un número determinado de unidades, por lo que debe repetirse tantas veces como sea necesario para reflejar una magnitud determinada. Otra forma es representando en diferentes escalas un mismo dibujo donde las áreas son proporcionales a la frecuencia. En la figura 2.3 se presenta un pictograma.

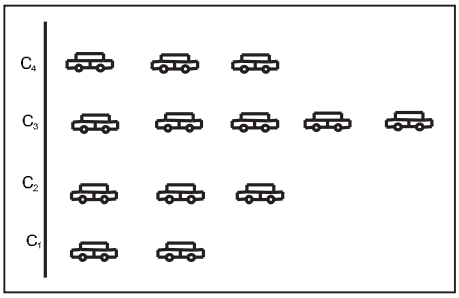

Figura 2.3: Pictograma

- **Diagrama de Sectores:** llamado también gráfico de torta o pastel. Consiste en dividir el circulo en tantos sectores como categorías tenga la variable y donde a cada sector se le corresponde un área proporcional a la frecuencia absoluta o relativa asociada con la modalidad que representa, como se muestra en la figura 2.4.

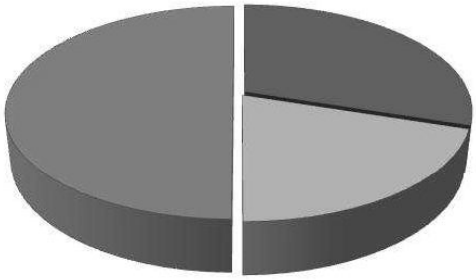

Figura 2.4: Fig.1.

Ejemplo 2.11 *En la figura 2.5 se muestra el diagrama de sectores asociado con los datos del ejemplo 2.1.*

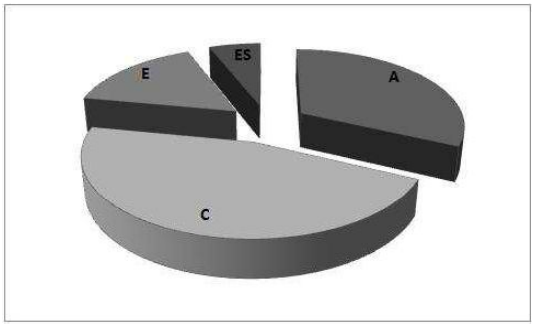

Figura 2.5: Distribución de las carreras de FACES

2. **Gráficos para Variables Cuantitativas:**

a) Gráficos a utilizar cuando las clases son valores individuales:

- **Diagrama de Líneas:** Para representar gráficamente una variable de tipo cuantitativo y cuyas clases son valores individuales, se usa el diagrama de líneas el cual se construye colocando en el eje de las abscisas los valores de la variable y en el eje de las ordenadas, la frecuencia absoluta o relativa. Para cada valor se traza una línea recta vertical cuya altura es igual a la frecuencia absoluta o relativa asociada con ese valor.

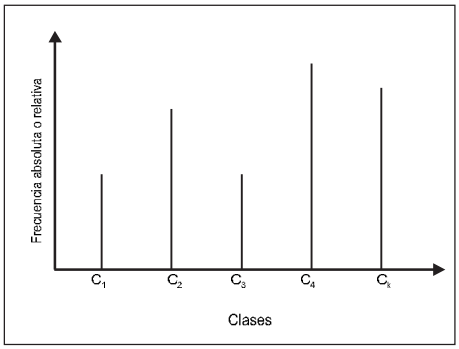

Figura 2.6: Diagrama de Líneas

Ejemplo 2.12 *La figura 2.7 presenta el diagrama de línea para los datos del ejemplo 2.5.*

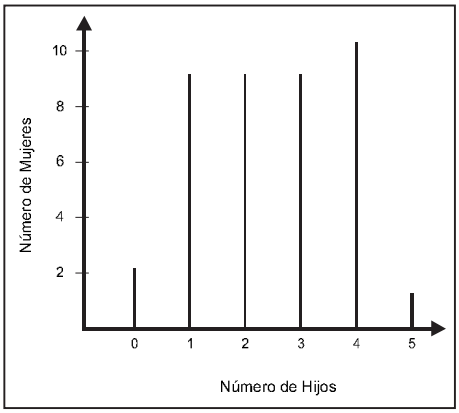

Figura 2.7: Distribución del número de hijos por familia

- **Diagrama Escalonado o de Frecuencias Acumuladas:** Por la naturaleza de la variable, tiene forma de escalera. Cada escalón corresponde al paso de un valor

de la variable a otro (al siguiente). Para su construcción se colocan en el eje de las X los valores de las variables y en el eje de las Y, las frecuencias acumuladas. La frecuencia acumulada de cada valor se representa con una línea horizontal que va desde ese valor hasta donde se señala el siguiente.

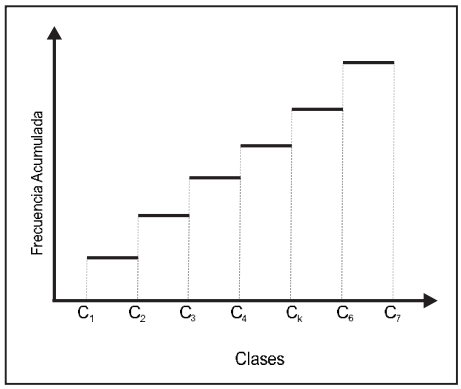

Figura 2.8: Fig.1.

Ejemplo 2.13 *En la figura 2.9 se muestra el diagrama escalonado para los datos del ejemplo 2.5.*

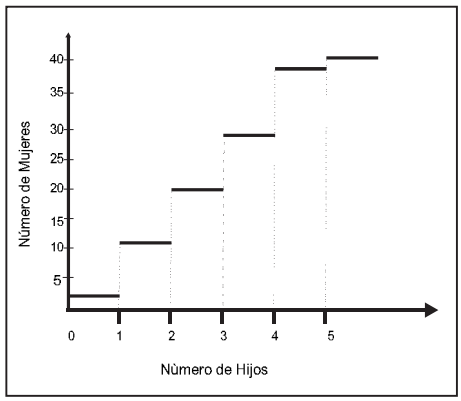

Figura 2.9: Distribución del número de hijos por familia

b) Gráficos a utilizar cuando las clases son intervalos:

Los gráficos que a continuación se discuten son usados exclusivamente con datos cuantitativos agrupados en distribuciones de frecuencias cuyas clases son intervalos.

- **Histograma de Frecuencias:** es un diagrama de barras con la característica que las barras están juntas unas de otras. Se obtiene construyendo sobre cada intervalo de

clase de la variable, un rectángulo cuya área es proporcional a la frecuencia absoluta o relativa correspondiente al intervalo, como se muestra en la figura 2.10

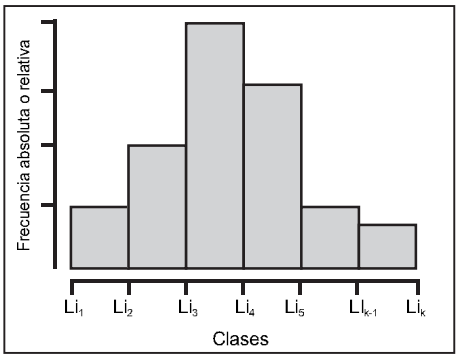

Figura 2.10: Fig.1.

Si deseamos comparar histogramas, la forma apropiada de construirlas es utilizando las frecuencias relativas y haciendo la altura de cada barra igual a $h_i = \frac{fr_i}{A_i}$ donde A_i es la amplitud de la clase i, cuando $A_1 = A_2 = \ldots = A_k$ entonces h_i coincide con f_i o fr_i.

Ejemplo 2.14 *La figura 2.11 representa el histograma de frecuencias asociado con los datos del ejemplo 2.6.*

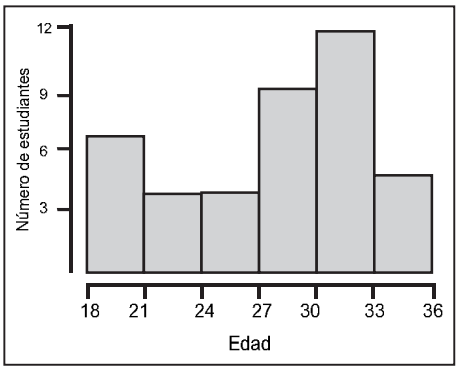

Figura 2.11: Distribución de las Edades de los estudiantes de FACES

- **Polígono de Frecuencia:** Consiste en unir mediante líneas rectas los puntos del histograma que corresponden a los puntos medios o marcas de cases. Para representarlo en el primer y ultimo intervalo, suponemos que adyacentes a ellos existen otros intervalos de la misma amplitud y frecuencia cero y se unen por una línea recta los puntos del histograma que corresponden a sus puntos medios.

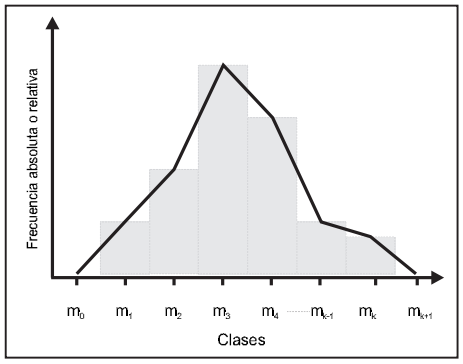

Figura 2.12: Polígono de Frecuencia

- **Ojiva o Polígono de frecuencias acumuladas:** para su construcción se usan los limites superiores de la clase y las frecuencias acumuladas (relativas o absolutas) de la clase. Para cada límite superior de la clase se indica con un punto su correspondiente frecuencia acumulada, luego estos puntos se unen mediante segmentos de recta obteniéndose así, una curva no decreciente. Los limites superiores se ubican en el eje de abscisas y las frecuencias acumuladas en el eje de las ordenadas. También se ubica el limite inferior de la primera clase, al cual se le asigna frecuencia acumulada igual a cero. Cuando el gráfico es construido usando las frecuencias relativas acumuladas, se le denomina Ojiva Porcentual.

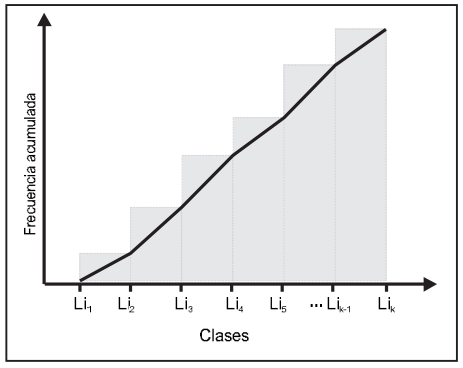

Figura 2.13: Ojiva

Ejemplo 2.15 *La ojiva para los datos del ejemplo 2.6 se muestra en la figura 2.14.*

La Ojiva puede ser usada para calcular gráficamente el número o porcentaje aproximado de datos que son menores o, mayores e igual que un valor determinado. Si queremos conocer el número de datos que es inferior a X_0 , simplemente ubicamos

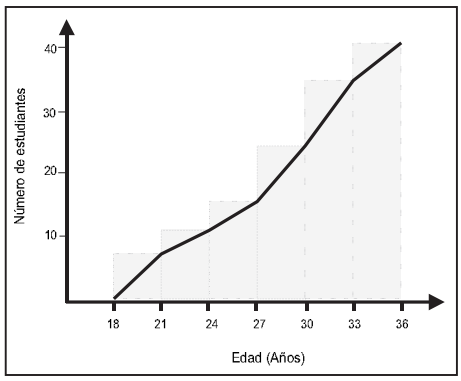

Figura 2.14: Distribución de las Edades de los Estudiantes de FACES

en el eje de las abscisas a X_0 y luego proyectamos una línea perpendicular hasta la Ojiva. Desde allí se traza una línea paralela al eje de las abscisas y el punto, digamos F_0, donde esta línea corta al eje de las ordenadas representa el número a calcular.

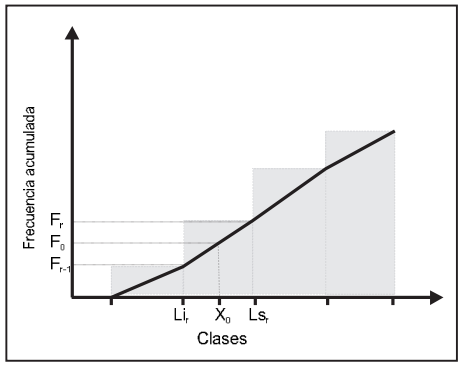

Figura 2.15: Fig.1.

El valor F_0 puede ser calculado algebraicamente mediante interpolación. Supongamos que se desea calcular el número de valores que son menores a X_0. Supongamos además que X_0 esta incluido en la clase $[LI_r - LS_r)$, la cual tiene frecuencia absoluta acumulada igual a F_r . Entonces F_0 se obtiene al resolver la ecuación:

$$\frac{X_0 - LI_r}{LS_r - LI_r} = \frac{F_0 - F_{r-1}}{F_r - F_{r-1}}$$

donde F_{r-1} representa la frecuencia absoluta acumulada de la clase anterior a la que contiene a X_0 .
De igual manera, podemos calcular mediante la ojiva aquel valor X_0, tal que un

número o porcentaje de datos dado, sea menor o mayor que el. Esto se logra simplemente realizando el procedimiento anterior en sentido opuesto.

3. **Gráficos Especiales**

   Hay gráficos o diagramas que se utilizan con gran frecuencia que no hemos considerado hasta ahora por no encontrarse enmarcados en la calificación anterior.

   - **Diagrama de Dispersion:** Gráfico de especial utilidad para analizar la relación entre dos variables. Se construye ubicando en el eje de las abscisas los valores de la variable X y en el eje de las ordenadas los valores de la variable Y.

   - **Diagrama de Causa - Efecto:** Son representaciones gráficas que permiten identificar las posibles causas asociadas a un problema (efecto) estructuradas según una serie de factores genéricos. Reciben también el nombre de "Diagrama de espina de pescado", "Diagrama de río" o "Diagrama de Ishikawa".

   - **Gráfico de Pareto:** Son diagramas de barras, donde estas se representan en orden descendente en altura. De esta forma, la barra mas alta corresponde a la modalidad de mayor frecuencia. Esta representación permite ubicar las modalidades mas relevantes por su frecuencia.

   - **Diagrama de Tallo y Hoja de Tukey:** Técnica que permite clasificar los datos sin perder precisión, cuando el número de datos no es muy grande.

   - **Diagrama de Caja:** Gráfico que describe la distribución de un conjunto de datos mediante el uso de los cuartiles como medida de posición y el rango intercuartílico como medida de dispersión. Representa una de las principales alternativas en el Análisis Exploratorio de Datos. Son especialmente útiles si se desea comparar la distribución de dos o más grupos de datos.

2.4. Medidas Descriptivas Numéricas

En las secciones anteriores examinamos algunas técnicas que permiten describir visualmente un conjunto de datos, es decir, procedimientos que ofrecen una idea cualitativa de las características del mismo. Usualmente, esa descripción gráfica o cualitativa, es acompañada por algunas medidas numéricas sencillas de calcular e interpretar, denominadas medidas descriptivas numéricas. El propósito de esta sección es el de introducir técnicas que permitan la descripción de un conjunto de datos desde el punto de vista matemático.

Al concluir esta sección el alumno debe estar en la capacidad de definir y usar las principales medidas de tendencia central, de posición , de dispersión y de forma (Asimetría y Curtosis) de un conjunto de datos, así como las técnicas para manipular distribuciones de frecuencias.

Definición 2.1 (Medidas Descriptivas) *Son cantidades que de manera resumida proveen información acerca de características importantes de un conjunto de datos. Es decir, son coeficientes que resumen una serie de datos y que contienen la mayor parte de la información relevante, permitiendo así descubrir aspectos importantes de dicha serie.*

Las medidas descriptivas las podemos clasificar de acuerdo a lo que se mide en los siguientes tres grupos: Medidas de localización, medidas de dispersión y medidas de forma.

2.4.1. Medidas de Localización

Son coeficientes de tipo promedio que tratan de representar una determinada distribución, pueden ser de dos tipos; centrales (o de tendencia central) y no centrales. Las medidas centrales son parámetros alrededor de los cuales se distribuyen los datos de la distribución y se toman como el centro de la misma. Tratan de identificar el valor más representativo de un conjunto de datos. Las medidas no centrales permiten ubicar partes de la distribución. Algunas medidas de localización son la media, la mediana, la moda y los cuantiles.

1. **La Media.** Es la medida de tendencia central más popular para datos cuantitativos, entre otras cosas por poseer propiedades matemáticas deseables. Representa el centro de gravedad o punto de equilibrio de un conjunto de datos. Existen distintos tipos de medias:

   - **Media Aritmética.** La media aritmética de una variable es simplemente el promedio de los datos. Su cálculo depende si los datos están o no agrupados en una distribución de frecuencia.
     - Para datos no agrupados, la media aritmética está dada por:

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

donde

x_i representa la i-ésima observación.

n el número de observaciones

Ejemplo 2.16 *Los siguientes datos corresponden al número de hijos de 18 matrimonios en el Estado Mérida:*

1	*2*	*2*	*3*	*2*	*2*
2	*1*	*3*	*3*	*3*	*1*
2	*2*	*3*	*1*	*2*	*1*

La media aritmética de estos datos es

$$\overline{x} = \frac{36}{18} = 2$$

- **Media Aritmética Ponderada:** Cuando se calcula la media aritmética se considera que todas las observaciones tienen igual importancia. Hay situaciones en las que los valores de una variable están afectadas por un factor que las modifica. Es decir,hay situaciones en las que la importancia (peso o ponderación) de las observaciones no es siempre la misma. Sea w_i la importancia de la observación i. Entonces, para un conjunto de n datos, la media, denominada en este caso media aritmética ponderada, se define de la siguiente manera:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i * w_i}{\sum_{i=1}^{n} w_i}$$

Ejemplo 2.17 *El profesor de estadística II informa a sus alumnos que ha asignado los siguientes pesos a sus 4 evaluaciones; 35 %, 20 % 30 % y 15 %. Para un estudiante que ha obtenido una nota de 15, 12, 17 y 14 puntos en los cuatro exámenes, su nota promedio debe calcularse usando un procedimiento que tome en cuenta la importancia relativa de cada observación. Luego, la nota media o promedio de este alumno es*

$$\overline{x} = \frac{15 * 0{,}35 + 12 * 0{,}20 + 17 * 0{,}30 + 14 * 0{,}15}{1} = 14{,}85 puntos.$$

Ejemplo 2.18 *Un supermercado vende pollos procedentes de 4 empresas. El gerente desea calcular la utilidad promedio del supermercado por la venta de pollos. La siguiente tabla muestra la empresa, la utilidad por pollo y la cantidad vendida.*

Empresa	***Utilidad en BsF.***	***Cantidad vendida***
A	5	300
B	7	150
C	4	600
D	2	700

Por tanto, la media aritmética ponderada es

$$\overline{x} = \frac{5*300 + 7*150 + 4*600 + 2*700}{1750} = 3{,}63 BsF.$$

Se puede pensar erróneamente en calcular la utilidad promedio como $\overline{x} = \frac{18}{4} = 4{,}5 BsF$*, pues se vende más pollo de algunas empresas que de otras. Aquellos tipos de pollos que se venden más, tienen una mayor incidencia sobre la utilidad.*

- Para datos agrupados en tablas de frecuencias. A menudo, se quiere calcular la media a través de una tabla de frecuencias previamente hecha. Su fórmula de cálculo depende de si los clases están conformadas por valores individuales o por intervalos.

$$\overline{x} = \begin{cases} \dfrac{\sum_{i=1}^{k} x_i * f_i}{n} = \sum_{i=1}^{k} x_i * fr_i, & \text{clases individuales;} \\ \dfrac{\sum_{i=1}^{k} m_i * f_i}{n} = \sum_{i=1}^{k} m_i * fr_i, & \text{clases en intervalos.} \end{cases}$$

Ejemplo 2.19 *El gerente de una sucursal de MRW en Mérida quiere conocer el número medio de enmiendas recibidas diariamente. La siguiente tabla de frecuencias muestra la distribución del número de enmiendas* (x_i) *registrados durante 30 días.*

x_i	f_i	fr_i
0	2	0,07
1	3	0,1
2	10	0,33
3	12	0,4
4	3	0,1

La cantidad media de enmiendas diarias recibidas es

$$\begin{aligned} \overline{x} &= \frac{0*2 + 1*3 + 2*10 + 3*12 + 4*3}{30} \\ &= 0*0{,}07 + 1*0{,}1 + 2*0{,}33 + 3*0{,}4 + 4*0{,}1 = 2{,}37. \end{aligned}$$

Ejemplo 2.20 *La siguiente tabla corresponde a la distribución de los ingresos (en miles de euros) en 38 partidos del equipo Real Madrid en la liga española de fútbol.*

Empresa	***Utilidad en BsF.***	***Cantidad vendida***
$[80-100)$	5	$0,1316$
$[100-120)$	7	$0,1842$
$[120-140)$	10	$0,2632$
$[140-160)$	9	$0,2368$
$[160-180)$	5	$0,1316$
$[180-200)$	2	$0,0526$

$$\overline{x} = \frac{5100}{38} = 90 * 0{,}1316 + \ldots + 190 * 0{,}0526 = 134{,}21$$

Cuando las clases son valores individuales, el valor de la media es exacto, mientras que cuando son intervalos existe una perdida de precisión ya que se supone que todos los valores dentro de una clase son iguales al punto medio de la misma. Esta pérdida de precisión es sin embargo despreciable.

Obsérvese que si los datos están agrupados en una tabla de frecuencia, su media aritmética es un caso particular de la media aritmética ponderada con $f_1, f_2, ..., f_k$ o $fr_1, fr_2, ..., fr_k$ como ponderaciones.

Propiedades de la Media Aritmética:

- La suma de los desvíos de los datos con respecto a su media es nula:

$$\sum_{i=1}^{n}(x_i - \overline{x}) = 0$$

  En general

$$\sum_{i=1}^{n}(x_i - \overline{x}) * f_i = 0$$

- Para cualquier valor $k \neq \overline{x}$ que se considere:

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 < \sum_{i=1}^{n}(x_i - k)^2$$

  es decir $\sum_{i=1}^{n}(x_i - \overline{x})^2$ es un mínimo.

- Si todos los datos son iguales a un valor constante c, entonces:

$$\overline{x} = c$$

- Si a cada uno de los datos $x_1, x_2, ..., x_k$ cuya media es $\overline{x}$ se le suma un valor constante c, la media $\overline{x}$ aumenta en el mismo valor. Es decir, si $y = x + c$, entonces

$$\overline{y} = \overline{x} + c$$

- Si a cada uno de los datos $x_1, x_2, ..., x_k$ cuya media es $\overline{x}$ se multiplica por un valor constante c, la media $\overline{x}$ queda multiplicada por dicho valor. Es decir, si $y = c * x$, entonces

$$\overline{y} = c * \overline{x}$$

- Si $y = a + bx \Rightarrow \overline{y} = a + b\overline{x}$ para $a, b \in \mathbb{R}$;
- Dados r diferentes grupos de datos de tamaño $n_1, n_2, ..., n_r$, con medias $\overline{x}_1, \overline{x}_2, ..., \overline{x}_r$, entonces la media de los $n = n_1 + n_2 + ... + n_r$ datos es:

$$\overline{x} = \frac{n_1\overline{x}_1 + n_2\overline{x}_2 + ... + n_r\overline{x}_r}{n}$$

Ventajas de la media aritmética

Las principales Ventajas de esta medida de localización son:

- Toma en cuenta todos los datos.
- Fácil de calcular y de operar algebraicamente.
- A medida que la distribución sea más simétrica mayor será la aproximación entre el valor medio de los datos no agrupados y el valor medio de los datos agrupados.

Desventajas de la media aritmética

Sus principales desventajas son:

- Es sensible a valores extremos o atípicos. Por ejemplo, Supongamos que tenemos una muestra de 49 datos iguales a 1 y un dato extremo igual a 51. La media de esta muestra es 2, valor que no es una buena medida de localización de la mayoría de los datos.
- No ofrece siempre una buena aproximación cuando las distribuciones son asimétricas.
- No se puede calcular para tablas de frecuencias con intervalos de clases abiertas.

- **Media Geométrica.** Es una medida poco utilizada, que se emplea fundamentalmente para calcular un promedio cuando se combinan el valor de la variable y el tiempo. Resulta la más apropiada cuando se quiere promediar porcentajes, índices y cifras relativas. También es útil para determinar incrementos porcentuales de un periodo a otro. Tiende a reducir la influencia de valores grandes y destacar la de los valores pequeños. Es de

amplia aplicación en los negocios y la economía. Se define como:

$$\overline{x}_G = \sqrt[n]{x_1 x_2 ... x_n} = \sqrt[n]{\prod_{i=1}^{I} x_i^{n_i}}$$

Ejemplo 2.21 *La siguiente tabla muestra los ingresos (en millones de euros) del equipo Real Madrid durante las últimas 5 campañas del fútbol español. Se desea determinar la tasa de crecimiento promedio de dichos ingresos*

Año	***Ingreso***	***Porcentaje respecto del año anterior***
2004	500	$----$
2005	550	1,1
2006	670	1,2182
2007	660	0,9851
2008	780	1,1818

La tasa de crecimiento promedio (cambio promedio) en los ingresos del Real Madrid está dada por

$$\overline{x}_G = \sqrt[4]{1{,}1 * 1{,}2182 * 0{,}9852 * 1{,}1818} = 1{,}1176(11{,}76\,\%)$$

Ejemplo 2.22 *Las siguientes temperaturas (en grados centígrados) han sido tomadas durante un proceso químico;* 13,4; 12,8; 11,9 *y* 13,6. *La temperatura promedio del proceso se obtiene a través de la media geométrica*

$$\overline{x}_G = \sqrt[4]{13{,}4 * 12{,}8 * 11{,}9 * 13{,}6} = 12{,}91\,\%$$

Ejemplo 2.23 *Se desea conocer cual fue el interés promedio generado por cierto tipo de certificado. Para ello se cuenta con los datos para el año 2008: 16.72, 17.73, 17.47, 16.17, 15.5, 15.04, 13.85, 13.68, 13.71, 13.13, 14.38 y 11.78. El interés promedio generado por dicho certificado es*

$$\overline{x}_G = \sqrt[12]{1{,}1279^{14}} = 14{,}83\,\%$$

- **Media Armónica.** Su uso es poco frecuente. Se utiliza para el calculo de variaciones con respecto al tiempo y se presenta un marcado sesgo hacia la derecha en el comportamiento

de los datos. Igual que la media geométrica, tiende a reducir la influencia de valores grandes y destacar la de los valores pequeños. Su desventaja es que cuando algún valor de la variable es 0 o próximo a cero, no se puede calcular. Su fórmula de calculo está dada por

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

Al transformar los datos a $\frac{1}{x_i}$, el comportamiento de los datos tiende corregirse el sesgo hacia la derecha.

Ejemplo 2.24 *La media armónica para los datos 4, 6 y 9 es*

$$\overline{x}_a = \frac{3}{\frac{1}{4} + \frac{1}{6} + \frac{1}{9}} = 5{,}68$$

Ejemplo 2.25 *Supóngase que una familia realiza un viaje en automóvil a un ciudad y cubre los primeros 100 km a 60 km/h, los siguientes 100 km a 70 km/h y los últimos 100 km a 80 km/h. La velocidad media utilizada es*

$$\overline{x}_a = \frac{3}{\frac{1}{60} + \frac{1}{70} + \frac{1}{80}} = 69{,}04$$

- **Media Cuadrática.** Cuando se observa un conjunto de datos, hay situaciones en las que es de interés tener en cuenta la influencia del signo de los mismos. Un caso típico es el de los errores, pues el error, bien sea negativo o positivo, es error. En este tipo de problemas se usa la media cuadrática como medida de localización. Se define como

$$\overline{x}_c = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}} = \sqrt{\frac{\sum_{i=1}^{k} x_1^2 f_i}{n}}$$

2. **La Mediana:** La mediana de un conjunto de datos es el valor del centro de los datos, una vez que los mismos sean ordenados de menor a mayor. Es decir, la mediana es aquel valor que deja el mismo número de datos antes y después que el, una vez que son ordenados.

   Su aplicación se ve restringida por el hecho de que solo considera el orden jerárquico de los datos y no alguna propiedad propia de los datos, como en el caso de la media. Igual que en el caso de la media, existen dos procedimientos para calcular la mediana, dependiendo de si los datos se consideran tal cual, o si están agrupados en intervalos de clase. Veamos cada uno de ellos.

- Para datos no agrupados se distinguen dos casos de acuerdo al número de datos n. Si n es impar, la mediana es el valor central del conjunto ordenado, mientras que si el número de datos es par, la medina es el promedio de los valores centrales del conjunto ordenado. Esto es, si denotamos por M_d a la mediana, se tiene que:

$$M_d = \begin{cases} \frac{x_{n/2}+x_{(n/2)+1}}{2}, & \text{si } n \text{ es par;} \\ x_{(n+1)/2}, & \text{si } n \text{ es impar.} \end{cases}$$

Ejemplo 2.26 *Se tienen los 5 siguientes datos correspondientes a notas; 15, 12, 17, 10 y 14. Una vez ordenados, la mediana de estos datos es*

$$M_d = x_{(n+1)/2} = x_{(5+1)/2} = x_3 = 14$$

Ejemplo 2.27 *Los siguientes datos corresponden a los puntos marcados por Trotamundos de Carabobo en 6 partido de la liga nacional de Basketball; 88, 102, 98, 94, 111 y 106. Se ordenan y se obtiene su mediana*

$$M_d = \frac{x_{n/2} + x_{(n/2)+1}}{2} = \frac{x_{6/2} + x_{(6/2)+1}}{2} = \frac{x_3 + x_4}{2} = \frac{98 + 102}{2} = 100$$

Esto significa que en la mitad de los juegos se marcaron menos de 100 puntos, y en la mitad de los juegos el numero de puntos marcados excedieron dicha cantidad.

- Para datos agrupados en tablas de frecuencias.
  - Si las clases son valores individuales, el procedimiento es el siguiente:
    a) Se calcula $n/2$.
    b) Se ubica $n/2$ en la columna de la frecuencia acumulada, F_i Si $n/2$ coincide con F_i, la mediana es el promedio de ese valor de la variable y el siguiente, es decir

$$M_d = \frac{x_i + x_{(i+1)}}{2}$$

    c) Si $n/2$ no coincide con F_i, ubicamos aquella frecuencia acumulada inmediatamente superior a $n/2$ y la mediana es el correspondiente valor de variable.

Ejemplo 2.28 *Calcular la mediana para los datos del ejemplo 2.5. La tabla 2.5 muestra la distribución de frecuencias para estos datos.*
En este caso $n/2 = 20$, coincide con la frecuencia acumulada asociada con el tercer valor de la variable. Por lo tanto, la mediana será el promedio entre el tercer y cuarto

valor de la variable, es decir

$$M_d = \frac{x_3 + x_4}{2} = \frac{2+3}{2} = 2{,}5$$

Ejemplo 2.29 *La siguiente tabla de frecuencia corresponde a la distribución del número de hijos en 80 familias seleccionadas en el Estado Mérida.*

$N°deHijos$	f_i	fr_i	F_i	Fr_i
0	15	0,1875	15	0,1875
1	21	0,2625	36	0,45
2	19	0,2375	55	0,6875
3	18	0,225	73	0,9125
4	7	0,0875	80	1

Aquí $n/2 = 40$ *está contenido en la frecuencia acumulada asociada con el tercer valor de la variable, por lo que la mediana es igual a dicho valor. Esto es,* $M_d = 2$.

- Si los datos están agrupados en tablas de frecuencias y las clases son intervalos, suponiendo que los mismos están igualmente espaciados, la mediana se calcula mediante el siguiente procedimiento:
  a) Calcular $n/2$.
  b) Ubicar la clase cuya frecuencia acumulada es igual o superior a $n/2$. A esta clase se le llama clase medianal.
  c) Obtener la mediana mediante la fórmula

$$M_d = li_m + \left(\frac{n/2 - F_{am}}{f_m}\right) * A_m$$

donde

F_{am}=Frecuencia Acumulada de la clase anterior a la medianal.
A_m=Amplitud de la clase medianal.
li_m=Limite inferior de la clase medianal.
f_m=Frecuencia absoluta de la clase medianal.

Ejemplo 2.30 *Calcular e interpretar la mediana para los datos del ejemplo 2.7. La tabla de frecuencia es*

Inicialmente se ubica la clase medianal. Para ello se calcula $n/2 = 40/2 = 20$, *luego se localiza la primera frecuencia acumulada que contiene a 20. La clase medianal es entonces* $[27 - 30)$.

$Edad$	f_i	F_i	fr_i	Fr_i
$[18-21)$	7	0,175	7	0,175
$[21-24)$	4	0,100	11	0,275
$[24-27)$	4	0,100	15	0,375
$[27-30)$	9	0,225	24	0,600
$[30-33)$	11	0,275	35	0,875
$[33-36)$	5	0,125	40	1

De esta forma se tiene que

$$M_d = 27 + \left(\frac{20-15}{9}\right) * 3 = 28{,}67$$

De esta manera, $M_d = 28{,}67$ *años representa el valor central de las edades. Es decir, aproximadamente la mitad de los estudiantes tiene una edad inferior a 28.67 años y aproximadamente la otra mitad tiene más de 28.67 años.*

Método gráfico para el cálculo de la Mediana

De acuerdo a lo visto en el capitulo 2, La mediana puede ser calculada gráficamente mediante el uso de la Ojiva. El procedimiento es:

a) Se localiza $n/2$ (o 0.50 si se usa la frecuencia relativa acumulada) en el eje de las ordenadas.

b) Desde este punto se traza una línea paralela al eje de las abscisas hasta cortar la ojiva.

c) Desde este punto de intersección se traza una línea paralela al eje de las ordenadas hasta cortar el eje de las abscisas. Este punto de corte es la mediana.

Propiedades de la Mediana

- En su cálculo no se incluyen todos los valores de la variable.
- No se ve afectada por valores atípicos, lo que hace recomendable su uso en el caso de distribuciones asimétricas.
- Es de cálculo rápido y de interpretación sencilla.
- Es función de los intervalos escogidos.
- Puede calcularse en el caso de las clases abiertas.
- Solamente toma en cuenta la posición que ocupan las observaciones y no el valor de las mismas, por lo que no es susceptible de operaciones algebraicas, característica que limita su utilidad.

- Para cualquier conjunto de datos, la mediana es el valor mas cercano o próximo a todos ellos. Esto es, $\sum_{i=1}^{n} |x_i - M_d|$ es un mínimo.

3. **La Moda:** Se denota por M_o y es el valor más común entre los datos, el valor de la variable que se presenta mayor número de veces, es decir, el valor de mayor frecuencia. Para un conjunto de datos puede no existir la moda o, existir una o más (polimodal).

Ejemplo 2.31 *Calcular la moda para el siguiente conjunto de datos: 5, 3, 3, 3, 2, 6, 5, 3, 2, 2. En este caso la moda es* $M_o = 3$*, pues es el valor que más se repite.*

Ejemplo 2.32 *Calcular la moda para el siguiente conjunto de datos: 5, 3, 3, 3, 2, 6, 2, 3, 2, 2. Aquí existen dos modas,* $M_{o1} = 2$ *y* $M_{o2} = 3$

Ejemplo 2.33 *Calcular la moda para el siguiente conjunto de datos: 5, 3, 4, 1, 2, 6, 7, 9, 8. Este conjunto de datos no posee moda, ningún valor se repite más que los otros.*

Cuando los datos están agrupados en tablas de frecuencias, se presentan dos situaciones:

- Si las clases son valores individuales entonces la moda es el valor (modalidad) o los valores (modalidades) que posee(n) la(s) mayor(es) frecuencia(s) absoluta(s).

  Ejemplo 2.34 *Se seleccionan 50 hombres en la ciudad de Mérida y se registra su estado civil. Los resultado se muestran en la siguiente tabla:*

Estado Civil	f_i
Soltero	16
Casado	12
Divorciado	8
Viudo	4
Otro	10

  La modalidad que presenta la mayor frecuencia es Casado,$(f_i = 16)$*, representando así la moda para este conjunto de datos*

  Ejemplo 2.35 *Obtener la moda para la siguiente distribución de frecuencias:*
  La frecuencia absoluta más alta es 19 *y corresponde a la cuarta clase, es decir, la moda en este caso es* $M_o = 35$*.*

Clase	f_i
20	7
25	10
30	12
35	19
40	13
45	11
50	8

- Si los datos están agrupados en tablas de frecuencias y las clases son intervalos, se ubica la clase con la mayor frecuencia absoluta, clase modal, luego se obtiene la moda mediante la fórmula

$$M_o = li_o + \frac{\Delta_1}{\Delta_1 + \Delta_2} * A_o$$

donde:

li_o = Limite inferior de la clase con mayor frecuencia absoluta (clase modal).
$\Delta_1 = f_o - f_{o-1}$ Frecuencia absoluta de la clase modal - Frecuencia absoluta de la clase Pre - modal.
$\Delta_2 = f_o - f_{o+1}$ Frecuencia absoluta de la clase modal - Frecuencia absoluta de la clase Post - modal.
A_o = Amplitud de la clase modal.

Ejemplo 2.36 *Calcular la moda para el ejemplo 2.7. La clase modal es* $[30 - 33)$*. De esta forma,* $f_o = 11$*,* $f_{o-1} = 9$*,* $f_{o+1} = 5$ *y la moda es*

$$M_o = 30 + \frac{2}{2+6} * 3 = 30{,}75$$

Propiedades de la Moda:

- Es muy fácil de calcular.
- No es susceptible de operaciones algebraicas, por lo tanto, su uso es limitado.
- Es la única medida que puede ser usada para datos cualitativos.
- Es una medida muy imprecisa e inestable.
- Puede no ser única.
- No siempre es una medida de tendencia central. En realidad indica punto(s) de concentración de datos

Cuál Medida es Mejor

La moda tiene como principal ventaja sobre el resto de medidas de tendencia central su aplicabilidad en todas las escalas de medida. Si el tamaño muestral no es suficientemente grande, la moda no es una medida confiable. La mediana por su lado, es una medida excelente para representar el nivel característico o representativo de los datos. Es una medida más confiable que la moda. La media aritmética tiene un error de muestreo menor que las medidas anteriores, por lo tanto es la más confiable de las tres.

Para fines descriptivos, la mediana es la medida de tendencia central preferida mientras que para fines inferenciales, la media es la de mayor uso.

En la tabla 2.15 se muestran las distintas medidas de tendencia central clasificadas de acuerdo al tipo de datos.

Tabla 2.15: Medidas de Tendencia Central según en tipo de datos.

Variable	**Escala**	**Medida de tendencia central**
Cualitativa	Nominal	Moda
	Ordinal	Mediana, Moda
Cuantitativa		Media, Mediana y Moda

Además del tipo de escala de medida, existen otros factores que deben considerarse en la selección de la medida a utilizar en cada caso. La naturaleza de la distribución de los datos, aspecto que interesa reflejar, presencia de valores extremos y alcance del estudio, son algunos de estos aspectos.

4. **Cuantiles:** Son medidas de localización similares a las anteriores. Denotados por Q_h, tienen como objetivo fundamental identificar el valor de la variable por debajo del cual queda la h-ésima parte, en tanto por ciento, de todos los valores de la colección ordenada. Se puede decir que los cuantiles son unas medidas que dividen a la distribución en Q partes de manera que en cada una de ellas hay el mismo porcentaje de valores de la variable. Los más importantes son:

   - Cuartiles. Dividen a la distribución en cuatro partes porcentualmente iguales (3 divisiones). Se denotan por C_1, C_2, C_3, y corresponden al 25 %, 50 %,75 %.
   - Deciles. Dividen a la distribución en 10 partes iguales (9 divisiones). Se denotan por D_1, ..., D_9, y corresponden al 10 %,...,90 %.
   - Percentiles. Dividen a la distribución en 100 partes (99 divisiones). P_1, ..., P_{99}, y corresponden al 1 %,...,99 %.

- Para datos no agrupados. Para la obtención del cuantil h, se deben seguir los siguientes pasos:
  a) Ordenar los datos de menor a mayor
  b) Calcular el valor t

$$t = \frac{h}{q} * n$$

  donde h es el cuantil deseado, q es iguala 4, 10 y 100, para cuartiles (C_h), deciles (D_h) y percentiles (P_h), respectivamente.
  c) Si t es entero, el cuantil h, Q_h, es el promedio de los valores en las posiciones t y $t+1$, es decir

$$Q_h = \frac{x_t + x_{t+1}}{2}$$

  en caso contrario, t debe ser redondeado y Q_h será el valor en la posición asociada con el entero inmediatamente mayor que t.

Ejemplo 2.37 *Determinar el cuartil 3, el decil 7 y el percentil 85 para el siguiente conjunto de datos: 33 34 38 31 36 30 35 35 37 29 32 39.*

Esta serie ordenada es: 29 30 31 32 33 34 35 35 36 37 38 39. Para el cálculo del cuartil 3, $t = \frac{3}{4} * 12 = 9$*, y su valor está dado por el promedio de los valores en las posiciones 9 y 10 de la serie ordenada, es decir,*

$$C_3 = \frac{36 + 37}{2} = 36{,}5$$

Esto significa que el 75 % de los datos se encuentran por debajo de 36.5.

Para hallar el decil 7, $t = \frac{7}{10} * 12 = 8{,}4$*. Como t no es entero, el decil 7 es el valor asociado con la posición 9, es decir,* $D_7 = 36$*. El 70 % de los datos está por debajo de 36. Igualmente, para obtener el percentil 85, se obtiene t. En este caso,* $t = \frac{85}{100} * 12 = 10{,}2$*, y su valor es aquel que ocupa la posición 11,* $P_{85} = 38$*. Por lo tanto, el 85 % de los datos están por debajo de 38.*

- Para datos agrupados en tablas de frecuencia.
  - Si las clases son valores individuales, el procedimiento es el siguiente:
    a) Se calcula $\frac{h*n}{q}$.
    b) Se ubica $\frac{h*n}{q}$ en la columna de la frecuencia acumulada, F_i Si $\frac{h*n}{q}$ coincide con F_i, el cuantil es el promedio de ese valor de la variable y el siguiente, es decir

$$Q_h = \frac{x_i + x_{(i+1)}}{2}$$

c) Si $\frac{h*n}{q}$ no coincide con F_i, ubicamos aquella frecuencia acumulada que contiene a $\frac{h*n}{q}$ y el cuantil es el correspondiente valor de variable.

- Si los datos están agrupados en tablas de frecuencias y las clases son intervalos, suponiendo que los mismos están igualmente espaciados, el cuantil Q_h se calcula mediante el siguiente procedimiento:
  a) Calcular $\frac{h*n}{q}$.
  b) Ubicar la clase cuya frecuencia acumulada es igual o superior a $\frac{h*n}{q}$. A esta clase se le llama clase cuantil.
  c) Obtener el $h-$ésimo cuantil mediante la fórmula

$$Q_h = li_c + \left(\frac{\frac{h*n}{q} - F_{ac}}{f_c} \right) * A_c$$

donde

F_{ac}=Frecuencia Acumulada de la clase anterior a la clase cuantil.
A_c=Amplitud de la clase cuantil.
li_c=Limite inferior de la clase cuantil.
f_c=Frecuencia absoluta de la clase cuantil.

Ejemplo 2.38 *Para los datos del ejemplo 2.7, calcular el cuartil 1, el decil 6 y el percentil 90.*

Para cada caso, se debe inicialmente obtener el valor de $\frac{h*n}{q}$. *Estos valores son:*

$$\frac{h*n}{q} = \begin{cases} \frac{1*40}{4} = 10, & \textit{para el cuartil 1;} \\ \frac{6*40}{10} = 24, & \textit{para el decil 6;} \\ \frac{90*40}{100} = 36, & \textit{para el percentil 90.} \end{cases}$$

Luego, las clase 2, 4 y 6 están asociadas con el cuartil 1, decil 6 y percentil 90, respectivamente. El valor de estos cuantiles son por lo tanto:

$$C_1 = 21 + \left(\frac{10-7}{4} \right) * 3 = 23{,}25$$

$$D_6 = 27 + \left(\frac{24-15}{9} \right) * 3 = 30$$

y

$$P_{90} = 33 + \left(\frac{36-35}{5} \right) * 3 = 33{,}6$$

El 25 % de los estudiantes tiene edad inferior a 23.25 años, un 60 % inferior a 30 años y el 90 % tiene una edad inferior a 33.6 años.

2.4.2. Medidas de Dispersión.

Para variables cuantitativas o numéricas, en las que por lo general se observa un gran número de valores distintos, el análisis debe ser tal que de respuesta al siguiente conjunto de interrogantes:

- ¿Alrededor de qué valor se agrupan los datos?
- ¿Que valor es el más frecuente?
- Como se agrupan los datos, ¿muy concentrados? ¿muy dispersos?

Como se mostró en la sesión anterior, las medidas de tendencia central dan respuesta a la primera interrogante. Estas medidas sirven para describir sólo un aspecto de los datos, no dicen nada acerca de la dispersión de los valores observados. Para esto es necesario el uso de otro conjunto de medidas, las medidas de dispersión o variabilidad. Si el valor de estas medidas de dispersión es pequeño, indica que los datos están concentrados. Si es una medida de dispersión referida a un valor central, por ejemplo la media, para un valor pequeño de dicha medida se dice que los datos están concentrados alrededor de la media. En este caso, la media se considera representativa de los datos, es decir, es un promedio confiable. En caso contrario, la media no es confiable, no es representativa de los datos.

Las medidas de dispersión permiten medir el grado de agrupación o disgregación en un conjunto de datos, es decir, permiten determinar que tan cercanos o separados entre si están los valores. Esto es, las medidas de dispersión cuantifican la separación, la dispersión, la variabilidad de los valores de la distribución. Se pueden clasificar en absolutas y relativas. Las absolutas pueden o no, estar referidas a un valor central y no son comparables entre diferentes muestras. Las medidas relativas permiten comparar varias muestras. El siguiente cuadro muestra las distintas medidas de dispersión.

$$\text{Medidas de Dispersión} = \begin{cases} \text{Absolutas} = \begin{cases} \text{Rango;} \\ \text{Recorrido Intercuartilico;} \\ \text{Desviación Media;} \\ \text{Varianza y} \\ \text{Desviación Estándar.} \end{cases} \\ \text{Relativas} = \begin{cases} \text{Recorrido Intercuartilico Relativo y} \\ \text{Coeficiente de Variación.} \end{cases} \end{cases}$$

Al igual que en el caso de las medidas de tendencia central, la selección de la medida de dispersión a utilizar, dependerá, entre otras cosas, del objetivo a cumplir en el estudio. Si se quiere tener una visión general de la variabilidad de los datos, el rango y el recorrido intercuartilico son apropiadas. Si el objetivo es medir la variabilidad de los datos respecto de su media, entonces deben usarse medidas como la varianza, desviación media o desviación estándar.

Para comparar grupos de datos con valores promedios diferentes y unidades de medida diferentes, las mejores opciones resultan ser el coeficiente de variación y el rango intercuartilico relativo.

1. **Medidas de Dispersión Absolutas**

   - **Rango o Recorrido:** Medida de poco utilidad ya que puede llevar a conclusiones erróneas acerca del verdadero comportamiento de los datos. Viene dada por

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

     Es decir, el rango es la diferencia entre el valor máximo y el valor mínimo del conjunto de datos. Dos aspectos se deben resaltar:

     - Cuanto menor es su valor, es más representativo de las medidas de tendencia central.
     - Sólo depende de los valores extremos. Valores muy alejados afectan dicha medida.
     - No es aconsejable usarlo para muestras grandes, pues puede conducirnos a errores. Se le utiliza en muestras pequeñas de 4 a 5 observaciones, básicamente en el control estadístico de la calidad.

   - **Recorrido Intercuartilico:** Es una medida de la dispersión definida en la zona intermedia de los datos. Viene dada por la diferencia entre los cuartiles 3 y 1. Esto es,

     $$RIC = C_3 - C_1$$

     Esta medida indica en cuántas unidades de los valores que toma la variable se concentra el cincuenta por ciento central de los casos. Su principal ventaja es que es una medida resistente a los datos atípicos. Si su valor es muy pequeño, implica que la mayoría de los datos están en el centro, existe poca o baja dispersión. En caso contrario, los datos se distribuyen ampliamente, existe una alta dispersión.

   - **Desviación Media:** Está dada por el promedio de los valores absolutos de las diferencias entre cada valor del conjunto de datos y su media. Mide la diferencia que hay en cualquier sentido, positivo o negativo, entre los valores de una variable y su media. Su fórmula de calculo es,

     $$DM = \frac{\sum_{i=1}^{n} |x_i - \overline{x}|}{n}$$

     Si los datos están agrupados en una tabla de frecuencias, entonces su fórmula de cálculo

es:

$$DM = \begin{cases} \dfrac{\sum_{i=1}^{k} |x_i - \overline{x}| f_i}{n}, & \text{Si las clases son valores individuales;} \\ \dfrac{\sum_{i=1}^{k} |m_i - \overline{x}| f_i}{n}, & \text{Si las clases son intervalos.} \end{cases}$$

- **Varianza:** La desviación media presenta el inconveniente de no destacar cuando un valor está separado significativamente de la media y destaca excesivamente pequeñas diferencias respecto de la media. Para evitar tal situación se propone en su lugar la varianza, definida como la media de las diferencias al cuadrado de los datos respecto de su media, es decir,

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

Si los datos están agrupados en una tabla de frecuencias, entonces su fórmula de cálculo es:

$$S^2 = \begin{cases} \dfrac{\sum_{i=1}^{k} (x_i - \overline{x})^2 f_i}{n}, & \text{Si las clases son valores individuales;} \\ \dfrac{\sum_{i=1}^{k} (m_i - \overline{x})^2 f_i}{n}, & \text{Si las clases son intervalos.} \end{cases}$$

Las siguientes fórmulas son usadas comúnmente por su facilidad de cálculo

$$S^2 = \begin{cases} \dfrac{\sum_{i=1}^{k} {x_i}^2 - n\overline{x}^2}{n} = \dfrac{\sum_{i=1}^{k} {x_i}^2}{n} - \overline{x}^2, & \text{Para datos no agrupados;} \\ \dfrac{\sum_{i=1}^{k} x_i^2 f_i - n\overline{x}^2}{n} = \dfrac{\sum_{i=1}^{k} x_i^2 f_i}{n} - \overline{x}^2, & \text{Si las clases son valores individuales;} \\ \dfrac{\sum_{i=1}^{k} m_i^2 f_i - n\overline{x}^2}{n} = \dfrac{\sum_{i=1}^{k} m_i^2 f_i}{n} - \overline{x}^2, & \text{Si las clases son intervalos.} \end{cases}$$

Dado que esta medida viene expresada en unidades de los datos al cuadrado, por ejemplo, si las observaciones se miden en metros, la varianza lo hace en metros al cuadrado. De esta forma su interpretación se dificulta, siendo esta su principal desventaja.

- **Desviación Estándar:** Dada la dificultad presentada con la interpretación de la

varianza, surge una medida de dispersión función de ella y que viene expresada en las mismas unidades que los datos, desviación estándar o típica. Representa la medida de dispersión más utilizada en estadística y está dada por,

$$S = \sqrt{S^2}$$

Propiedades de la Varianza y Desviación Estándar:

a) La varianza y la desviación estándar no pueden ser negativas.

b) Al aumentar el tamaño de la muestra, disminuye la varianza y la desviación estándar.

c) Si todos los datos son iguales a una constante c, entonces $S^2 = 0$ y $S = 0$.

d) Si a cada dato original se le suma una constante k, la varianza y la desviación estándar no se ven afectadas.

e) Si cada dato original se multiplica por una constante k, la varianza y la desviación estándar del nuevo conjunto de datos están dadas por k^2S^2 y kS .

f) Supongamos que se tiene un conjunto de datos digamos,$x_1, x_2, ..., x_n$, cuya varianza es S^2, entonces la varianza y la desviación estándar de $a + bx_1, a + bx_2, ..., a + bx_n$, están dadas por, b^2S^2 y $|b|S$

Cuando se desea medir la dispersión o variabilidad de una variable, por lo general, esta se mide con respecto a un valor central, es decir, se usan medidas absolutas referidas a un valor central. Son las que tiene mayor sentido cuando los datos son simétricos o tienden a una distribución simétrica.

Todas las medidas de dispersión consideran que a mayor valor de la medida de dispersión, mayor es la variabilidad.

2. **Medidas de Dispersión Relativas**

Por lo general están dadas por el cociente entre una medida de dispersión y una medida de tendencia central y sirven para comparar la variabilidad de dos conjuntos de valores.

- **Rango Intercuartílico Relativo:** Resulta del cociente entre el rango intercuartílico y la mediana, es decir,

$$IQ = \frac{Q_3 - Q_1}{Md}$$

Indica que tamaño tiene el rango intercuartílico con respecto a la mediana. Es una medida independiente de las unidades de medida y resulta interesante para comparar la variabilidad de diferentes variables.

- **Coeficiente de Variación:** Igual que el rango intercuartílico relativo, su utilidad estriba en que permite comparar la dispersión o variabilidad de dos o más grupos. Indica el

tamaño relativo de la desviación estándar respecto a la media. Es la medida de dispersión relativa de mayor uso y su fórmula de calculo es

$$CV = \frac{S}{\overline{x}} * 100$$

El coeficiente de variación se utiliza para comparar la homogeneidad de dos series de datos, aún cuando estén expresados en distintas unidades de medida. A medida que el Coeficiente de variación disminuye, se observa una mayor homogeneidad en los datos, es decir, los datos están más concentrados alrededor del promedio.

Propiedades:

a) Si x tiene coeficiente de variación $CV_x = \frac{S}{\overline{x}} * 100$, entonces $y = a + x$ tiene coeficiente de variación dado por $CV_y = \frac{S}{a+\overline{x}} * 100$. Esto es, el coeficiente de variación no es invariante ante cambios de origen.

b) Si x tiene coeficiente de variación $CV_x = \frac{S}{\overline{x}} * 100$, entonces $y = bx$ tiene coeficiente de variación dado por $CV_y = \frac{bS}{b\overline{x}} * 100 = \frac{S}{\overline{x}} * 100 = CV_x$. Esto es, el coeficiente de variación es invariante ante cambios de escala.

2.4.3. Medidas de Forma

Hasta ahora, se han analizado y estudiado la tendencia así como la dispersión de una distribución, pero, parece evidente que es necesario conocer más sobre el comportamiento de una distribución. En esta parte, se analizaran las medidas de forma.
Las medidas de forma permiten comprobar si una distribución de frecuencia tiene características especiales como simetría, asimetría, nivel de concentración de datos o nivel de apuntamiento que la clasifiquen en un tipo particular de distribución. Son medidas necesarias para determinar el comportamiento de los datos y así, poder adaptar herramientas para el análisis probabilístico.
Las medidas de forma de una distribución se pueden clasificar en dos grandes grupos: **medidas de asimetría y medidas de curtosis**. Estas medidas permiten evaluar la situación de los datos desde los ejes vertical (simetría) y horizontal (curtosis).

1. **Medidas de Asimetría** La asimetría resulta conveniente en muchas situaciones. Muchos modelos asumen una distribución normal, esto es, simétrica alrededor de la media. La distribución normal tiene una asimetría cero. En el mundo real, los valores no son nunca perfectamente simétricos y la asimetría de la distribución proporciona una idea sobre si las desviaciones de la media son positivas o negativas. Una asimetría positiva implica que hay más valores distintos a la derecha de la media. Las medidas de asimetría, junto a las medidas de curtosis se utilizan para verificar si se puede aceptar que un conjunto de datos sigue la distribución normal, lo que es necesario para realizar inferencia estadística.

   Cuando el diagrama de líneas o histograma de frecuencias de una variable presenta una forma acampanada, diremos que los datos tienen una distribución simétrica. En caso contrario, dicha

distribución será asimétrica o diremos que presenta asimetría.

Ahora bien, comparando las medidas de tendencia central, podemos establecer relaciones que permitan determinar la presencia o no, de asimetría en un conjunto de datos. De esta forma podemos indicar que:

Si $\overline{x} = Md = Mo$ la Distribución es simétrica.

Si $\overline{x} < Md < Mo$ la Distribución es asimétrica negativa.

Si $\overline{x} > Md > Mo$ la Distribución es asimétrica positiva.

Otra manera de evaluar la simetría de un conjunto de datos es calculando ciertos coeficientes de asimetría. Las medidas de asimetría son indicadores que permiten establecer el grado de simetría (o asimetría) que presenta una distribución de una variable aleatoria sin tener que hacer su representación gráfica.

Como base de simetría consideramos una recta paralela al eje de ordenadas que pasa por la media de la distribución. Si existe el mismo número de valores a la derecha que a la izquierda de la media y por lo tanto, el mismo número de desviaciones con signo positivo que con signo negativo, se tiene una distribución es simétrica. Se dice que hay asimetría positiva (o a la derecha) si la "cola" a la derecha de la media es más larga que la de la izquierda, es decir, si hay valores más separados de la media a la derecha. En caso contrario, hay asimetría negativa (o a la izquierda).

- **Coeficiente de Asimetría de Fisher:** Para determinar el grado de asimetría de un conjunto de datos una posibilidad es el coeficiente de Fisher, cuya fórmula de cálculo es

$$A_F = \begin{cases} \dfrac{\sum_{i=1}^{n}(x_i - \overline{x})^3}{nS^3}, & \text{Datos no agrupados;} \\ \dfrac{\sum_{i=1}^{k}(m_i - \overline{x})^3 f_i}{nS^3}, & \text{Datos agrupados en intervalos.} \end{cases}$$

  Si $A_F = 0$ la Distribución es simétrica.

  Si $A_F < 0$ la Distribución es asimétrica negativa.

  Si $A_F > 0$ la Distribución es asimétrica positiva.

- **Coeficiente de Asimetría de Pearson:** Mide el grado de asimetría en términos de la distancia entre la media y la moda. Este coeficiente divide esta diferencia entre la desviación estándar para eliminar la dimensionalidad. Su fórmula de cálculo es

$$A_p = \frac{\overline{x} - Mo}{S}$$

Si $A_P = 0$ la Distribución es simétrica.
Si $A_P < 0$ la Distribución es asimétrica negativa.
Si $A_P > 0$ la Distribución es asimétrica positiva.

- **Coeficiente de Asimetría de Bowley:** Está basado en la posición de los cuartiles y la mediana. Su fórmula de cálculo es

$$A_B = \frac{C_3 + C_1 - Md}{C_3 - C_1}$$

Si $A_B = 0$ la Distribución es simétrica.
Si $A_B < 0$ la Distribución es asimétrica negativa.
Si $A_B > 0$ la Distribución es asimétrica positiva.

2. **Medidas de Apuntamiento o Curtosis.**

Las medidas de apuntamiento o curtosis, miden el grado de apuntamiento o achatamiento de la distribución en su parte central con respecto a la distribución normal, es decir, miden el grado de concentración de datos en la región central.

La distribución de probabilidad normal tiene gran importancia al querer estudiar el apuntamiento o curtosis de la distribución de los datos. Se dice que una distribución tiene un apuntamiento u otro, siempre en función de esta distribución normal. La distribución normal, corresponde a fenómenos muy corrientes en la naturaleza y cuya representación gráfica es una campana de Gauss. Esta campana responde a una función matemática, que es la función de densidad de la distribución.

- **Coeficiente de Curtosis de Fisher:** Permite medir el grado de apuntamiento de la distribución de un conjunto de datos. Está dada por

$$C_f = \begin{cases} \dfrac{\sum_{i=1}^{n}(x_i - \overline{x})^4}{nS^4} - 3, & \text{Datos no agrupados;} \\ \dfrac{\sum_{i=1}^{k}(m_i - \overline{x})^4 f_i}{nS^4} - 3, & \text{Datos agrupados en intervalos.} \end{cases}$$

Al comparar con la distribución normal, se tiene la siguiente interpretación:
Si $C_f > 0$ la Distribución es leptocúrtica. Más apuntada que la normal
Si $C_f < 0$ la Distribución es platicúrtica. Menos apuntada que la normal
Si $C_f = 0$ la Distribución es mesocúrtica. Similar a la normal.

- **Coeficiente de Curtosis percentílico:** Se define en función de los percentiles. Está dada por

$$C_p = \frac{\frac{1}{2}(C_3 - C_1)}{P_{90} - P_{10}} - 0{,}263$$

Al comparar con la distribución normal, se tiene la siguiente interpretación:

Si $C_p > 0$ la Distribución es leptocúrtica. Más apuntada que la normal
Si $C_p < 0$ la Distribución es platicúrtica. Menos apuntada que la normal
Si $C_p = 0$ la Distribución es mesocúrtica. Similar a la normal.

2.5. Ejercicios

1. Se registro el estado civil de 50 estudiantes de FACES seleccionados aleatoriamente y los resultados obtenidos fueron

c	s	s	s	d	c	s	s	d	c
s	s	s	s	c	d	s	s	s	s
c	s	c	c	v	s	s	c	c	s
d	v	c	c	s	s	s	s	s	c
c	s	s	s	s	s	s	s	s	s

   Organize los datos en una distribución de frecuencia y comente los resultados.

2. Los siguientes datos recogen la información del sexo de una persona, la ocupación y su opinión referente a como ha visto la participación de Venezuela en la Copa América 2007.

Sexo	**Ocupación**	**Opinión**
F	Estudiante	Buena
F	Docente	Regular
M	Estudiante	Buena
F	Estudiante	Buena
M	Empleado	Mala
F	Docente	Regular
M	Estudiante	Mala
M	Obrero	Buena
F	Empleado	Buena
F	Docente	Buena
F	Estudiante	Regular
M	Estudiante	Mala
M	Docente	Mala
F	Estudiante	Buena
M	Estudiante	Mala

   a) Organize los datos en una distribución de frecuencia para cada variable por separado.

   b) Construya todas las posibles tablas cruzadas.

   Comente los resultados.

3. Se ha realizado una encuesta a 30 personas en la que se les pregunta el número de personas que conviven en el domicilio habitualmente. Las respuestas obtenidas han sido las siguientes: 1, 4, 4, 1, 3, 5, 3, 2, 4, 1, 6, 2, 3, 4, 5, 5, 6, 2, 3, 3, 2, 2, 1, 8, 3, 5, 3, 4, 7, 2, 3.

a) Calcule la distribución de frecuencias de la variable obteniendo las frecuencias absolutas, relativas y sus correspondientes acumuladas.

b) ¿Qué proporción de hogares está compuesta por tres o menos personas? ¿Qué proporción de individuos vive en hogares con tres o menos miembros?

c) Dibuje el diagrama de barras de frecuencias y el diagrama en escalones.

d) Agrupe por intervalos de amplitud 2 los valores de la variable, calcule su distribución de frecuencias y represente el histograma correspondiente.

4. Como control de la ética publicitaria se requiere que el rendimiento, en millas por galón de gasolina, que los fabricantes de automóviles usan con fines publicitarios, este basado en un buen número de pruebas efectuadas en diversas condiciones. Al tomar una muestra de 50 automóviles se registran las siguientes observaciones en millas por galón:

27.9	29.3	31.8	22.5	34.2	34.2	32.7	26.5	26.4	31.6
35.6	31.0	28.0	33.7	32.0	28.5	27.5	29.8	31.2	28.7
30.0	28.7	33.2	30.5	27.9	31.2	29.5	28.7	23.0	30.1
30.5	31.3	24.9	26.8	29.9	28.7	30.4	31.3	32.7	30.3
33.5	30.5	31.3	32.7	30.3	30.1	30.3	29.6	31.4	32.4

Construya una distribución de frecuencia.

5. Construir una distribución de frecuencias con los datos dados a continuación que corresponden a los sueldos mensuales en miles de BsF. de 40 funcionarios. Agrupar la información en 9 clases.

1.45	1.49	1.43	1.64	1.64	1.47	1.53	1.22	1.72	1.50
1.46	1.41	1.39	1.39	1.45	1.57	1.18	1.71	1.62	1.48
1.38	1.49	1.27	1.25	1.34	1.56	1.36	1.30	1.21	1.44
1.80	1.29	1.55	1.36	1.61	1.43	1.70	1.50	1.51	1.52

6. La siguiente distribución se refiere a los pesos de un grupo de 80 personas.

Pesos (Kg)	**N^o de personas**
$[52-56)$	4
$[56-60)$	12
$[60-64)$	17
$[64-68)$	20
$[68-72)$	15
$[72-76)$	9
$[76-80)$	3

Calcule:

a) El porcentaje de personas con pesos inferiores a 62 kgs.

b) ¿Cuántas personas pesan entre 65 y 74 kgs?.

c) El número de personas con pesos superiores a 62 Kgs.

d) ¿Cuál es el peso por debajo del cual están el 75 % de las personas?

7. La distribución del ahorro mensual de 150 personas es:

Ahorro (miles/mes)	**N° de personas**
$[100-150)$	12
$[150-200)$	18
$[200-250)$	21
$[250-300)$	48
$[300-350)$	24
$[350-400)$	15
$[400-450)$	12

Calcule:

a) El porcentaje de personas con ahorro menor de 200000 Bs mensuales.

b) ¿Cuántas personas ahorran mas de 320000 Bs mensuales?.

c) ¿Cuál es el ahorro por encima del cual están el 50 % de las personas?

1. Se ha realizado un estudio entre 100 mujeres mayores de 15 años y el número de hijos de las mismas. El resultado ha sido:

N° de Hijos	**N° de mujeres**
0	13
1	20
2	25
3	20
4	11
5	7
6	4

Se pide:

a) Calcular el número medio de hijos, la mediana y la moda.

b) Analizar la dispersión de la distribución.

c) Analizar la forma de la distribución calculando los coeficientes correspondientes.

2. La siguiente distribución expresa el número de autos vendidos durante una semana por cada uno de los 50 concesionarios que una determinada firma tiene en Venezuela:

N° de autos vendidos	**N° de concesionarios**
1	3
4	6
10	5
12	20
8	5

Se pide:

a) El promedio de autos vendidos, mediana y moda.

b) Analizar la dispersión de la distribución.

c) Analizar la forma de la distribución calculando los coeficientes correspondientes.

3. Un estudio sobre remuneraciones realizado tomando como muestra 100 profesionales de una determinada especialidad, arrojo el siguiente resultado:

Remuneración (BsF/mes)	**N° de prof**
$[3000 - 3600)$	6
$[3600 - 4200)$	10
$[4200 - 4800)$	20
$[4800 - 5400)$	22
$[5400 - 6000)$	18
$[6000 - 6600)$	14
$[6600 - 7200)$	10

Se pide:

a) La media, mediana y moda.

b) Analizar la dispersión de la distribución.

c) Analizar la forma de la distribución calculando los coeficientes correspondientes.

4. Calcular las medidas descriptivas para los ejercicios de la sección 1.3.2.

CAPÍTULO 3

PROBABILIDADES

3.1. Introducción.

Cuando los resultados de un fenómeno se conocen, existe certeza completa, la única razón de que se cometa un error en la toma de decisiones sobre los mismos, es que exista un error en el análisis. Sin embargo, en la realidad por lo general se presentan situaciones que no son totalmente predecibles y aún cuando se haga un análisis correcto, hay factores que no se pueden controlar y que influyen de forma tal que los resultados no pueden determinarse con certeza absoluta, es decir, existe incertidumbre. Bajo estas condiciones, se habla de posibilidades de ocurrencia. Una medida numérica de estas posibilidades es la probabilidad, representada por un número que va desde cero (ninguna posibilidad de ocurrencia) a uno (certeza completa de ocurrencia). Por tanto, las probabilidades se utilizan para cuantificar que tan probable es un determinado evento.

Las probabilidades son muy útiles, ya que pueden servir para desarrollar estrategias o tomar decisiones. Por ejemplo, un inversionista desea invertir su dinero si existen altas posibilidades de ganar; un apostador hípico decidirá hacer una apuesta grande si existe un riesgo pequeño de perder. Situaciones como en las siguientes, resulta necesario trabajar con el concepto de probabilidades:

1. Se elige un alumno en la Facultad de Economía y se le consulta acerca del número de libros que ha solicitado a préstamo en la biblioteca durante el último mes.

2. En una prueba de control de calidad se examina un componente electrónico haciéndolo funcionar de manera ininterrumpida hasta que falla, y entonces se registra el tiempo transcurrido desde el inicio de la prueba.

No se puede afirmar con certeza que respuesta dará el estudiante ni cual sera exactamente el tiempo de duración del componente electrónico.

A continuación se dan a conocer algunos conceptos básicos, necesarios para la comprensión y manejo de la definición de probabilidades. Esta sección comienza con un tratamiento rápido sobre la teoría de conjuntos, luego se define experimento aleatorio para posteriormente, tratar las distintas definiciones de probabilidad.

3.2. Conceptos Básicos

3.2.1. Teoría de Conjuntos y Técnicas de Conteo

La teoría de conjuntos es de mucha utilidad en el desarrollo de las probabilidades, y es por ello que se debe revisar los conocimientos sobre las operaciones de conjuntos como lo son: la unión, la intersección, el complemento de un conjunto, etc. Para resolver algunos problemas de probabilidades es necesario conocer el numero de elementos que posee cierto conjunto y el conjunto universal, denominado, en probabilidades, espacio muestral. Cuando el conjunto es pequeño no hay problema, pero cuando contiene muchos elementos, esta tarea puede resultar algo complicada. Es necesario, por tanto, acudir a técnicas de conteo especiales que permitan calcular el número de elementos de cualquier conjunto.

Definición 3.1 (Conjunto) *Un conjunto es una colección de objetos, denominados miembros o elementos. En general, el conjunto se denota por una letra mayúscula A, B, C, mientras que sus elementos por una letra minúscula a, b, c.*

Ejemplo 3.1 *Son ejemplos de conjuntos:*

1. El conjunto de los números enteros.
2. El conjunto de las vocales en el alfabeto.
3. El conjunto de las edades de los estudiantes de la Escuela de Estadística.
4. El conjunto de los posibles resultados al lanzar un dado.
5. El conjunto de estaturas de todos los habitantes de una ciudad.

Dependiendo de la cantidad de elementos que contenga un conjunto, los mismos se pueden clasificar en conjuntos finitos e infinitos. Si el conjunto tiene un número conocido de elementos, se dice que es finito, en caso contrario, es decir, no se puede determinar su longitud, se dice que es infinito. En el ejemplo 3.1, los numerales 2 y 4 corresponden a conjuntos finitos, mientras que los restantes son conjuntos infinitos.
Un conjunto puede expresarse especificando todos sus elementos o, describiéndolos mediante las propiedades que deben tener sus elementos. En el primer caso, se dice que el conjunto se ha expresado por extensión, y en el segundo por comprensión.
Se definen ahora, los distintos tipos de conjuntos y las operaciones que se pueden dar entre los mismos.