Capítulo 1

Estadística Descriptiva

1.1. Conceptos Generales

1.1.1. Introducción.

Este capitulo tiene como propósito establecer el marco de referencia para el estudio de la estadística. En el mismo se destacará la importancia y campo de acción de esta. Se introducen algunas definiciones básicas que permiten comprender en forma intuitiva y real lo que es Estadística Descriptiva.

1.1.2. Origen.

Durante mucho tiempo se considero que el campo propio del estudio científico era exclusivo de fenómenos que bajo las mismas condiciones producen los mismos resultados, es decir, de fenómenos deterministicos. Sin embargo, aquellos fenómenos o situaciones donde está presente la incertidumbre en cuanto a lo que va a ocurrir, es decir, fenómenos aleatorios, son de gran importancia y su estudio corresponde a la Estadística. Algunos

ejemplos de fenómenos aleatorios son:

1. Lanzamiento de un dado.

2. Cantidades vendidas en un Supermercado en días sucesivos.

3. La duración de los equipos eléctricos en un lote producido por determinada empresa

La ciencia Estadística tiene su origen en las siguientes corrientes históricas:

1. Recopilación de datos en forma de Censo.

2. Juegos de azar.

3. Conocimiento inductivo. Paso de lo particular a lo general.

Estadística

Cuando hablamos de Estadística, tradicionalmente nos referimos a números presentados ordenada y sistemáticamente. Esta idea es consecuencia del concepto popular que existe sobre esta ciencia y que cada vez se extiende más debido a la influencia de nuestro entorno. Sin embargo cuando profundizamos en el campo de la investigación podemos entender que la estadística no solo son números, sino que representa la única herramienta que permite dar luz y obtener información en cualquier tipo de investigación, cuyo comportamiento no puede ser abordado desde el punto de vista deterministico. Podríamos decir entonces, que la estadística es la ciencia que permite determinar como usar la información referente a una investigación y como actuar en situaciones practicas donde esta presenta la incertidumbre.

Definición 1.1 (Estadística) *Es la ciencia de coleccionar, ordenar, presentar y describir la información relativa a un fenómeno en el cual esta presente la incertidumbre para su estudio, con el objeto de deducir la ley que rige dicho fenómeno y así poder tomar decisiones y obtener conclusiones. Para el estudio de un fenómeno, necesitamos contar con información relacionada con el mismo.*

Esta información obtenida bien sea experimentalmente o, mediante la observación, esta dada por datos. Estos datos son el resultado de medir en un conjunto de elementos o individuos, una o varias características a ser analizadas en una investigación.

Definición 1.2 (Elemento) *Es un ser vivo, objeto o cosa que posee características que se desean investigar.*

En sentido estadístico un elemento puede ser algo con existencia real, como un automóvil o una casa, o algo más abstracto como la temperatura, un voto, o un intervalo de tiempo.

Definición 1.3 (Universo Estadístico) *Se denomina universo estadístico a un conjunto finito o infinito de seres vivos o cosas, sobre las cuales están definidas las características que interesan analizar.*

Ejemplo 1.4 *1. Los Habitantes de la ciudad de Mérida.*

2. Los estudiantes de la Facultad de Ciencias Económicas y Sociales.

3. Los trabajadores de una empresa.

4. Los animales en un bosque.

5. Los carros que entran en un estacionamiento al día.

Cada elemento del universo tiene una serie de características que pueden ser objeto del estudio estadístico. Así por ejemplo si consideramos como elemento a una persona, podemos distinguir en ella los siguientes caracteres: Sexo, Edad, Nivel de estudios, Profesión, Peso, Altura, Color del cabello, etc.

Por lo tanto, de cada elemento del universo podremos estudiar uno o más aspectos cualidades o caracteres.

El universo puede ser según su tamaño de dos tipos:

- Universo finito: cuando el número de elementos que la forman es finito, por ejemplo el número de alumnos de un centro de enseñanza, o grupo clase.
- Universo infinito: cuando el número de elementos que la forman es infinito, o tan grande que pudiesen considerarse infinitos. Por ejemplo si se realizase un estudio sobre los productos que hay en el mercado. Hay tantos y de tantas calidades que este universo podría considerarse infinito.

Definición 1.5 (Población) *Es el conjunto de todas las posibles mediciones que pueden hacerse de una característica en estudio de los elementos del universo. Por lo tanto, la población está constituida por valores o datos bien sea numéricos o no.*

Ejemplo 1.6 *:*

1. *El sexo de los habitantes de la ciudad de Mérida*
2. *La edad de los estudiantes de la Facultad de Ciencias Económicas y Sociales.*
3. *El sueldo de los trabajadores de una empresa.*
4. *El color de ojos de los animales en un bosque.*

5. *La marca de los carros que entran en un estacionamiento al día.*

Se puede notar que un Universo puede estar constituido por una o varias poblaciones. Además, al igual que el universo, la población puede ser finita o infinita, dependiendo del número de valores que la constituyen. En el caso de que la población sea finita, se dice que esta tiene tamaño N.

Definición 1.7 (Muestra) *Es una parte de una población.*

Ejemplo 1.8 *:*

1. *El sexo de los habitantes de la ciudad de Mérida mayores a 60 años.*
2. *La edad de los estudiantes de la Facultad de Ciencias Económicas y Sociales que tienen un promedio mayor a 15 puntos.*
3. *El sueldo de los trabajadores de una empresa que son mujeres.*
4. *El color de ojos de los animales en un bosque que se encontraron en un día.*
5. *La marca de los carros tipo sedan que entran en un estacionamiento al día.*

Definición 1.9 (Parámetro) *Es una función de los valores de la población que sirve para sintetizar alguna característica relevante de la misma. Ejemplos de parámetros son: La media poblacional, La proporción poblacional, la varianza poblacional, entre otros.*

Definición 1.10 (Estadístico) *Es una función de los valores de la muestra que sirve para sintetizar alguna característica relevante de la misma. Ejemplos de parámetros son: La media muestral, La proporción muestral, la varianza muestral, entre otros.*

Como se ha dicho anteriormente, la estadística se encarga del estudio de un fenomeno a través del manejo de la información que se tiene sobre una o más características del mismo. En el lenguaje estadístico al igual que en el matemático a las características se les conocen como variables y a las distintas formas en que pueden presentarse, modalidades o valores de las variables.

Definición 1.11 (Variable) *Una variable es una característica que poseen los elementos del universo que pueden o no variar entre cada uno de ellos.*

Ejemplo 1.12

- *El color de ojos de las personas.*
- *La edad de las personas.*
- *El sueldo de un emplead.*
- *La raza de los perros.*
- *La nota de los alumnos de Métodos Estadísticos I.*

Al conjunto de las modalidades o valores de una variable se le denomina Escala de Medida. Las Escalas de Medida pueden clasificarse de acuerdo a las relaciones que existen entre los valores y las operaciones aritméticas que pueden realizarse entre las mismas en: Nominal, Ordinal, De Intervalos, De Razón y Absolutas.

1. Escala Nominal: Son aquellas en que la única relación que se define entre sus valores es la igualdad o diferencia, es decir solo podemos decir que dos valores de una variable son iguales o diferentes. No hay operaciones aritméticas definidas, por lo tanto, los números no tienen sentido como magnitudes.

   Ejemplo 1.13
   - *El grupo sanguíneo.*

- *El sexo.*
- *El color de ojos.*
- *El estado civil.*
- *Los números que llevan los atletas en la espalda*

2. Escala Ordinal: Son aquellas en que entre sus valores están definidas las relaciones de igualdad, diferencia, mayor que o menor que, es decir solo podemos decir que dos valores de una variable son iguales, diferentes y en el caso de que sean diferentes se puede establecer un orden entre ellos. No hay operaciones aritméticas definidas.

   Ejemplo 1.14
   - *Dureza de los minerales.*
   - *Grado de satisfacción.*
   - *Intensidad de un dolor.*
   - *Rango militar.*
   - *Nivel de educación.*

3. Escala De Intervalo: Los valores de las variables son números y entre ellos tienen sentido las relaciones de igualdad, de orden y de las distancias. La resta es la única operación aritmética definida. Esta escala posee dos propiedades de gran importancia.

   - Existe una unidad de medida cuyo significado se mantiene constante para todos los valores.
   - Posee un cero u origen relativo. El cero no significa ausencia de la característica.

Ejemplo 1.15
- *Puntuación obtenida en una evaluación.*
- *La temperatura.*
- *La distancia sobre el nivel del mar.*

4. Escala de Razón o Escala proporcional: Los valores de la variable son números y entre ellos tienen sentido las relaciones de igualdad, orden y están definidas las operaciones aritméticas de suma, diferencia y proporciones (múltiplos). Estas escalas tienen un cero absoluto, el cual representa la ausencia de la característica.

Ejemplo 1.16
- *El Sueldo de los habitantes del Estado Mérida.*
- *La edad de los alumnos de Métodos Estadísticos I.*
- *El nivel de hemoglobina.*

5. Escala Absoluta: Los valores que puede tomar la variable son el resultado de un conteo, por lo tanto, esta escala está constituida por todos los número enteros positivos y el cero.

Ejemplo 1.17
- *Número de accidentes automovilísticos el fin de semana.*
- *Número de integrantes de una familia.*
- *Numero de alumnos en un salón de clase.*

Tipos de Variables

Las variables se clasifican de acuerdo a su escala de medida en cualitativas y cuantitativas.

Definición 1.18 (Variable Cualitativa) *Son aquellas cuya escala de medida es nominal u ordinal, es decir, na variable es cualitativa si sus valores representan una cualidad o atributo del elemento en estudio. Por ejemplo:*

- *El sexo de las personas.*
- *El Tipo de sangre.*
- *La nacionalidad.*
- *El color de los ojos.*

Definición 1.19 (Variable Cuantitativa) *Hablamos de variables cuantitativas cuando los valores posibles son cantidades numéricas con las que podemos hacer operaciones aritméticas. Es decir, son aquellos cuya escala de medidas es de intervalos, proporcional o absoluta. Por ejemplo:*

- *El Sueldo de los habitantes del Estado Mérida.*
- *La edad de los alumnos de Métodos Estadísticos I.*
- *Número de integrantes de una familia.*

Las variables cuantitativas pueden dividirse en discretas y continuas.

Definición 1.20 (Variables Cuantitativas Discretas) *Son aquellas formadas por un conjunto numerable de puntos, es decir, se puede establecer correspondencia entre los valores que puede tomar la variable y el conjunto de los números reales, por lo tanto, son variables que no admiten valor alguno entre dos valores consecutivos de las mismas. Por ejemplo:*

- *La edad en años de los alumnos de Métodos Estadísticos I.*
- *Número de integrantes de una familia.*
- *Número de pares de zapatos que compran las mujeres al mes.*

Definición 1.21 (Variables Cuantitativas Continuas) *Son aquellas formadas por un conjunto numerable de puntos, es decir, se puede establecer correspondencia entre los valores que puede tomar la variable y el conjunto de los números reales, por lo tanto, son variables que no admiten valor alguno entre dos valores consecutivos de las mismas. Por ejemplo:*

- *La edad en años de los alumnos de Métodos Estadísticos I.*
- *Número de integrantes de una familia.*
- *Número de pares de zapatos que compran las mujeres al mes.*

Clasificación de la Estadística

La Estadística puede clasificarse de acuerdo a su función en el tratamiento de los datos en estadística descriptiva y estadística inferencial.

Definición 1.22 (Estadística Descriptiva) *Denominada también Estadística Deductiva. Es la encargada de describir, analizar y representar un conjunto de datos, utilizando métodos numéricos, tablas y gráficos que resumen y presentan la información contenida en ellos. Puede llevarse a cabo sobre una muestra o sobre toda una población.*

Definición 1.23 (Estadística Inferencial) *Denominada también Inferencia Estadística o Estadística Inductiva. Es la que apoyándose en la Teoría de Probabilidades y la*

Teoría del Muestreo, se encarga de efectuar estimaciones, permitir la toma decisiones, predicciones u otras generalizaciones sobre una población partiendo del estudio de una muestra.

La estadística descriptiva e inductiva pueden ser usadas separadas o conjuntamente. Lo usual es que en una investigación participen las dos.

1.2. Estadística Descriptiva

1.2.1. Introducción.

Esta sección tiene como propósito principal, introducir técnicas que permitan tanto matemática como gráficamente describir apropiadamente un conjunto de datos.

Al finalizar el tema, 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.

1.2.2. Organización de los Datos

La organización de los datos consiste en una agrupación apropiada de los mismos. Es importante dicha agrupacuión, 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, como dijimos en la sección anterior, del tipo de variable que se maneje. Por lo tanto, vamos a estudiar como se realiza la agruapación cuando la variable es cualitativa y cuando es cuantitativa.

Organización de Datos Cualitativos

Cuando los datos son cualitativos, la organización consiste en la construcción de una tabla, la cual contendrá la enumeración de las distintas modalidades que presenta la variable, el número de datos que corresponde a cada modalidad (frecuencia absoluta, fi) y la proporción que cada uno de ellos representa con respecto al total (frecuencia relativa, fri). Esta tabla recibe el nombre de Tabla de Frecuencia. La tabla 1 muestra la estructura de una tabla de frecuencias para datos cualitativos.

Tabla 1. Tabla de Frecuencias para datos cualitativos

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 1.24 *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:*

$C \quad A \quad A \quad C \quad C \quad A \quad A \quad E \quad A \quad C$

$E \quad E \quad C \quad ES \quad E \quad A \quad C \quad C \quad A \quad C$

$C \quad A \quad ES \quad C \quad E \quad A \quad A \quad C \quad A \quad C$

$C \quad C \quad A \quad E \quad E \quad A \quad C \quad C \quad C \quad A$

$C \quad C \quad A \quad C \quad C \quad C \quad C \quad ES \quad A \quad E$

donde

A: Administración

C: Contaduría

E: Economía

ES: Estadística

La variable en este ejemplo es la carrera que estian las personas, la cual es cualitativa de escala nominal, dicha variable presenta cuatro modalidades representadas por A,E,C y ES. Por lo tanto, al organizar los datos en una distribución de frecuencia se tiene que:

Tabla 2. Distribución de frecuencia de las carreras que se estudian en FACES

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

La tabla anterior es utilizada cuando se está estudiando una variable. Para el caso de dos variables, se usan comunmente las llamadas tablas de doble entrada o tablas de contingencia, pues las mismas permiten agrupar el numero de observaciones que cumplen con con las dos modalidades.

Tabla 3. Tabla de Contingencia

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

Organización de Datos Cuantitativos

Si los datos son cuantitativos, usamos un procedimiento similar al utilizado con los datos cualitativos, excepto, que es más laborioso. En este caso la tabla de frecuencias contiene los siguientes elementos:

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

- **Frecuencia Absoluta** (f_i)**:** Número de observaciones contenidas o incluidas en una clase.
- **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.
- **Frecuencia Absoluta Acumulada** (F_i)**:** Suma de frecuencias absolutas hasta la clase correspondiente.
- **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.

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 debemos seguir el siguiente procedimiento:

1. Obtener los extremos del intervalo total (Vmax y Vmin).
2. Obtener el rango o recorrido de la variable, R = Vmax - Vmin.
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

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

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

$$K = 1 + 3{,}3 \log n$$

Cuando particionamos los datos en clases, es generalmente recomendado usar entre 5 y 15 clases. Fuera de estos extremos, la organización resulta poco eficiente.

Una vez que hemos decidido en cuanto al número de clases, la amplitud de las clases, es simplemente

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

Esto nos 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, debemos 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 (Vmin). 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 [Linf - Lsup); esto es, intervalos cerrados por la izquierda y abiertas por la derecha.

5. Calcular las marcas de clase (m_i): Las marcas de clase estan 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 3 muestra la estructura de una tabla de frecuencias para datos cuantitativos

Tabla 4. 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_i	Fr_i	Fr_i
$[li_2 - ls_2)$	m_2	f_2	fr_i	Fr_i	Fr_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[li_k - ls_k)$	m_k	f_k	fr_i	Fr_i	Fr_i

Ejemplo 1.25 *A continuación se muestra la información sobre el número de hijos que tienen 40 Mujeres extraidas al azar de la ciudad de Mérida.*

Tabla 5. Número de Hijos

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 disreta 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 estan dadas por los valores individuales de la variable. A continuación se muestra se muestra dicha tabla:*

Tabla 6. Distribución del N de Hijos que tienen 40 Mujeres extraidas al azar de la ciudad de Mérida.

$NdeHijos$	f_i	F_i	fr_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

Donde se observa que gran parte de las mujeres estudiadas tiene de 1 a 4 hijos de manera bastante uniforme.

Ejemplo 1.26 *Se tienen los siguientes datos correspondientes 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

Tabla 7. Edad de 40 estudiantes de FACES

Si organizamos los datos en una distribución de frecuencia cuyas clases son valores individuales obtenemos lo siguiente:

Tabla 8. Distribución de fracuencia de las edades en clases individuales.

$NdeHijos$	f_i	F_i	fr_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

Esta agrupación de los datos es poco eficiente ya que la variable edad posee muchos valores diferentes (modalidades), lo que conlleva a no ser de fácil interpretación.

Para mejorar la organización de los datos, es necesario considerar a las clases como intervalos, tal como se describe a continuación:

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

$$Vmax = 35 \ y \ Vmin = 18$$

b) *Calculo del Rango.*

$$R = Vmax - Vmin = 35 - 18 = 17$$

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, obteniendose:

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

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 sumrle 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.*

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 intervales 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 numero de observaciones que se encuentran en el intervalo, para el primer intervalo de clase la frecuencia absoluta* (f_1) *es 7, esto quiere decir que hay 7 estudiantes con edades mayores o iguales a 18 años pero con edad menor 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* $F_3 = f_1 + f_2 + f_3 = 7 + 4 + 4 = 15$
- *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 obener esta valor es dividir la frecuencia acumulada entre el numero de observaciones,* $Fr_3 = \frac{15}{40} = 0{,}375$

La distribución de frecuencia está dada en la siguiente tabla:

Tabla 9. Distribución de frecuencia de las edades de 40 estudiantes.

Esta tabla presenta los datos de manera mas resumida que la tabla 8, lo cual la hace más fácil de interpretar. Por ejemplo, se puede decir que el 50 % de los estudiantes tienen edades entre 27 y 30 años.

$Edad$	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

1.2.3. Presentación 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 y el resumen puede ser representado gráficamente, lo que permite esclarecer aun más las características asociadas con la población. El uso de gráficos permite captar rápidamente las características fundamentales de los datos.

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, la preferencia e interés del investigador. La tabla 10 muestra los gráficos más apropiados de acuerdo al tipo de variable.

Tabla 10. Tipos de Gráficos de acuerdo al tipo de variable

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

1. **Gráficos para Variables Cualitativas**

- **Diagrama de Barras:** Grafica que representa en el eje de las abcisas (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). Puede ser usado para comparar poblaciones.

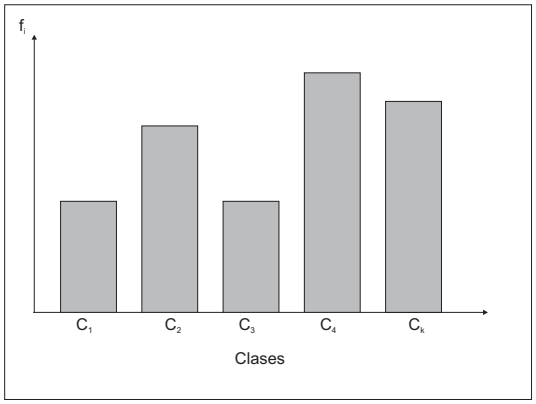

Figura 1.1: Gráfico de Barras

Ejemplo 1.27 *El diagrama de barras para el ejemplo 1.24 es:*

- **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 tanto, 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.
- **Diagrama de Sectores:** llamado también gráfico de torta. Consiste en dividir el circulo en tantos sectores como categorías tenga la variable y donde

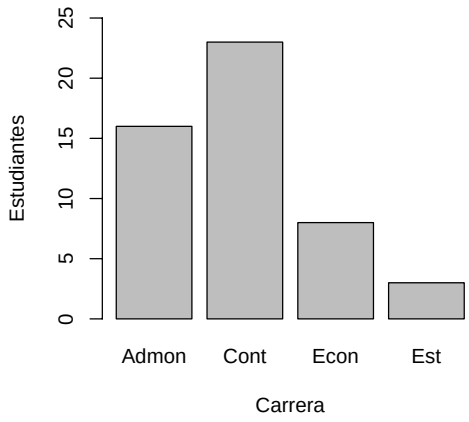

Figura 1.2: Distribución de las carreras de FACES

a cada sector se le corresponde una área proporcional a la frecuencia absoluta o relativa asociada con la modalidad que representa.

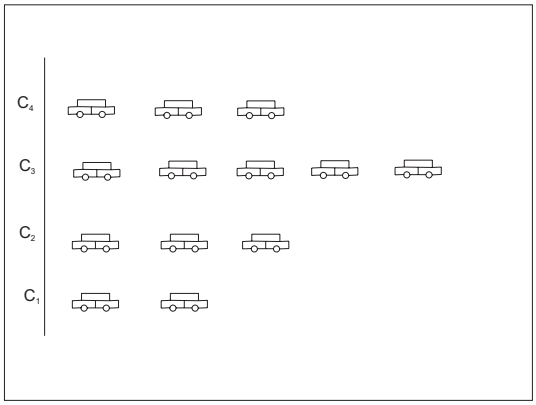

Figura 1.3: Pictograma

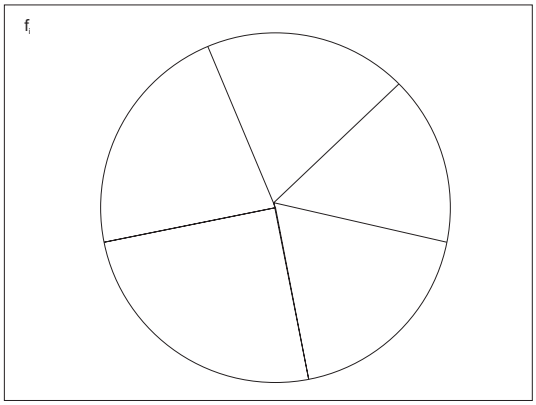

Figura 1.4: Fig.1.

Ejemplo 1.28 *El diagrama de sectores para el ejemplo 1.24 es:*

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 abcisas los valores de la variable y en el eje de las ordenadas, la frecuencia absoluta

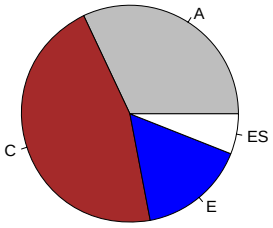

Figura 1.5: Distribución de las carreras de FACES

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 1.6: Diagrama de Líneas

Ejemplo 1.29 *El diagrama de línea para el ejemplo 1.25 es:*

- **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

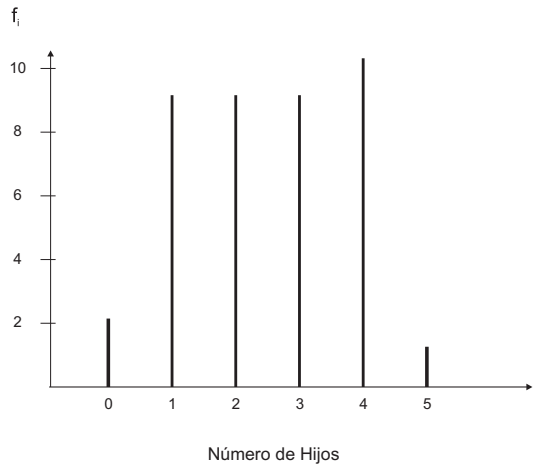

Figura 1.7: Distribución del número de hijos por familia

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.

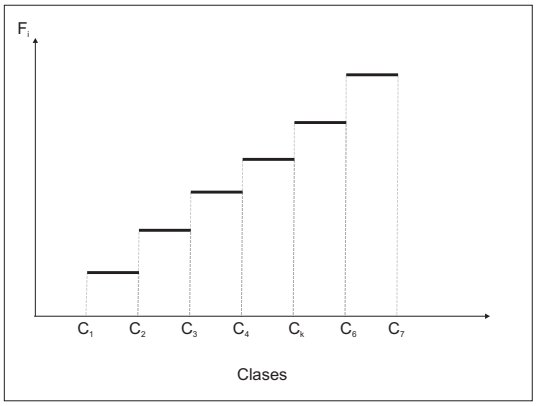

Figura 1.8: Fig.1.

Ejemplo 1.30 *El diagrama escalonado para el ejemplo 1.25 es:*

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

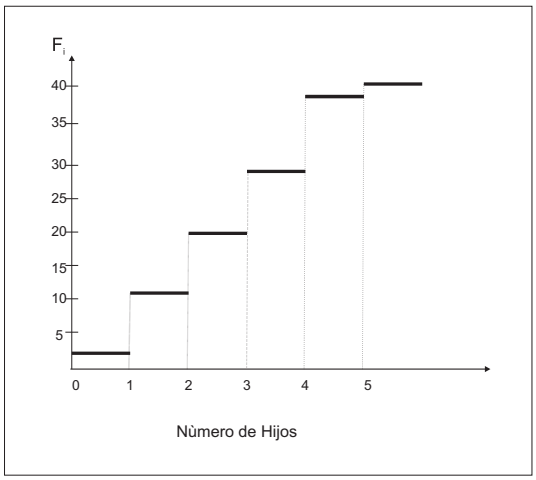

Figura 1.9: Distribución del número de hijos por familia

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 correspondiente al intervalo, como se muestra en la figura

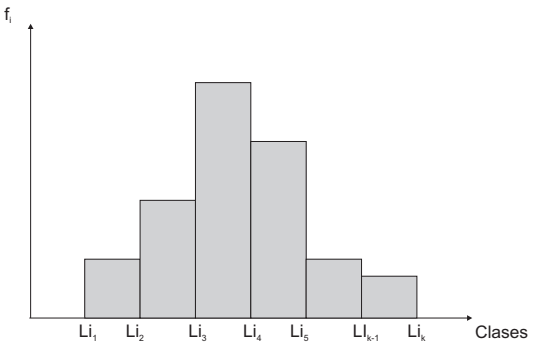

Figura 1.10: Fig.1.

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 1.31 *El histograma para el ejemplo 1.26 es:*

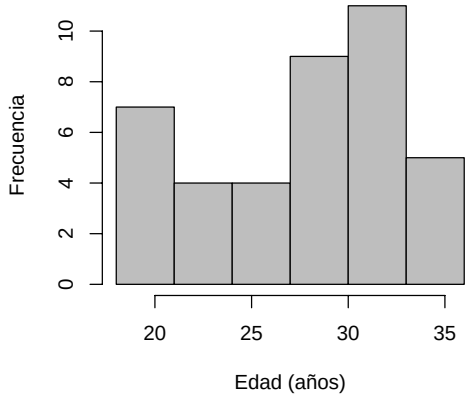

Figura 1.11: Distribución de las Edades de los estudiantes de FACES

- **Polígono de Frecuencia:** Consiste en unir mediante líneas rectas los puntos del histograma que corresponden a los puntos medios. 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.
- **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, lue-

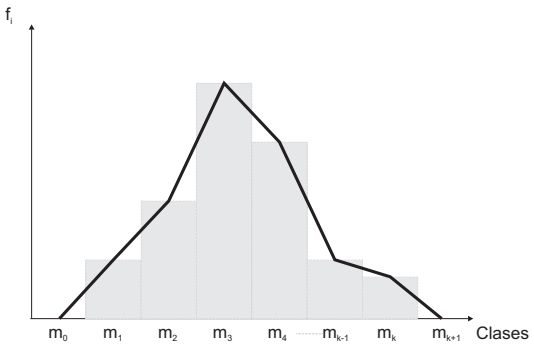

Figura 1.12: Poligono de Frecuencia

go estos puntos se unen mediante segmentos de recta obteniéndose así, una curva no decreciente. Los limites superiores se ubican en el eje de abcisas y las frecuencias acumuladas en 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.

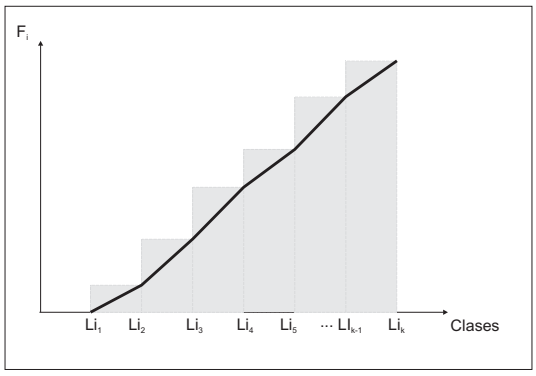

Figura 1.13: Ojiva

Ejemplo 1.32 *La ojiva para el ejemplo 1.26 se muestra en la siguiente figura.*

La Ojiva puede ser usada para calcular gráficamente el número o por-

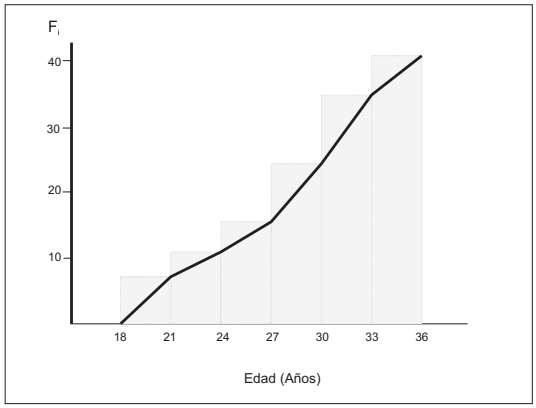

Figura 1.14: Distribución de las Edades de los Estudiantes de FACES

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

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 an-

terior a la que contiene a X_0 .

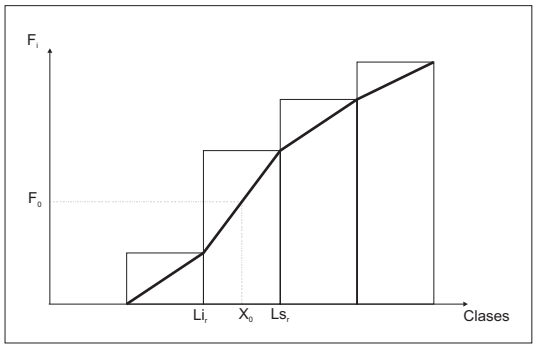

Figura 1.15: Fig.1.

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 Dispersión:** Gráfico de especial utilidad para analizar la relación entre dos variables. Se construye ubicando en el eje de las abcisas 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 graficas 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".

- **Grafico 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.

1.2.4. Medidas Descriptivas Númericas

En la sección anterior examinamos algunas técnicas que permiten describir visualmente un conjunto de datos, es decir, procedimientos que ofrecen una idea cualitativa de las características de un conjunto de datos. El propósito de esta sección es el de introducir técnicas que permitan la descripción desde el punto de vista matemático.

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

Definición 1.33 (Medidas Descriptivas) *Son cantidades que de manera resumida proveen información acerca de características importantes de un conjunto de datos.*

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.

1. **Medidas de Localización**

   También conocidas como medidas de tendencia central, son parámetros alrededor de los cuales se distribuyen los datos de la distribución y se toman como el centro de la misma. Algunas medidas de tendencia central son la media, la mediana y la moda.

   a) **La Media.** Es la medida de tendencia central más popular. 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-esima observación.

    el número de observaciones

  - Para datos agrupados en tablas de frecuencias, su fórmula de cálculo es:

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

  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.

La media de una serie de datos representa el centro de gravedad o punto de equilibrio de esos datos.

La media aritmética es fácil de obtener y explicar y tiene varias propiedades matemáticas que hacen mas ventajosos su uso que el de las otras medidas de tendencia central.

Propiedades:

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

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

- Para cualquier valor k que consideremos:

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

  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 $y = a + bx \Rightarrow \overline{y} = a + b\overline{x}$ 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:

$$\overline{x} = \frac{n_1\overline{x_1} + n_2\overline{x_2} + ... + n_r\overline{x_r}}{n}$$

- Si a cada uno de los datos $x_1, x_2, ..., x_k$ cuya media es $\overline{x}$ se le suma una constante k, entonces se obtiene una nueva colección de datos: $x_1 + k, x_2 + k, ..., x_n + k$ y la media de esta nueva colección sería: $\overline{x} + k$
- Si cada uno de los datos $x_1, x_2, ..., x_k$ cuya media es $\overline{x}$ se multiplica por una constante k, entonces se obtiene una nueva colección de datos: $x_1k, x_2k, ..., x_nk$ y la media de esta nueva colección sería: $\overline{x}k$

Ventajas

Las principales Ventajas son:

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

Desventajas

Sus principales desventajas son:

- Es sensible a valores extremos.
- 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 Aritmética Ponderada:** Existen situaciones en las que a los valores de la variable se le asigna un peso, ponderación o importancia. Es decir, existen situaciones en las que los valores de una variable están afectadas por un factor que las modifica. A este factor se le conoce con el nombre de ponderación y, debe ser considerada al momento de calcular la media aritmética de esos valores. La media aritmética calculada considerando esa ponderación recibe el nombre de Media Aritmética Ponderada y se define de la siguiente manera:

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

donde

p_i representa la ponderación de la i-esima observación.

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 $f_1, f_2, ..., f_k$ como ponderaciones.

- **Media Geométrica**

$$\overline{x_G} = \sqrt[n]{x_1 x_2 ... x_n}$$

- **Media Armónica**

$$\overline{x_a} = \sqrt{\frac{x_1^2 + x_2^2 + ... + x_n^2}{n}}$$

b) **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. Esto es, la mediana es aquel valor por debajo (encima) del cual se encuentra el 50 % de los datos.

Al igual que la media el calculo de la mediana depende de si los datos estan o no agrupados en una distribuón de frecuencias.

- Para datos no agrupados, la mediana es el valor central del conjunto ordenado , mientras que cuando el número de datos es par, la medina es el promedio de los valores centrales del conjunto ordenado, es decir:

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

- Para datos agrupados en tablas de frecuencias.
  - Si los datos están agrupados en tablas de frecuencias y las clases son valores individuales, el procedimiento es el siguiente:
    1) Se calcula $n/2$.
    2) Si $n/2$ coincide con F_a, la mediana es el promedio de ese valor de la variable y el siguiente.
    3) Si $n/2$ no coincide con F_a, ubicamos aquella frecuencia acumulada

que contiene a $n/2$ y la mediana es su correspondiente valor de variable.

- Si los datos están agrupados en tablas de frecuencias y las clases son intervalos, la mediana viene dada por:

$$m_d = LI_m + \frac{n/2 - F_{am}}{f_m} * a_m$$

El procedimiento para su calculo es:

1) Calcular $n/2$.
2) Ubicar la clase cuya frecuencia acumulada es igual o superior a $n/2$. A esta clase se le llama clase medianal.
3) Identificar los elementos de la fórmula anterior:

   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.

Calculo de la Mediana graficamente

La mediana puede ser calculada gráficamente mediante el uso de la Ojiva. El procedimiento es:

1) Localizamos 50 % en el eje de las ordenadas.
2) Desde este punto trazamos una línea paralela al eje de lasa abcisas hasta cortar la ojiva.

3) Desde este punto de intersección trazamos una línea paralela al eje de las ordenadas hasta cortar el eje de las abcisas. Este punto de corte es la mediana.

Propiedades de la Mediana

- No se ve afectada por observaciones extremas.
- Es de cálculo rápido y de interpretación sencilla.
- Es función de los intervalos escogidos.
- Puede calcularse en el caso de las clases abiertas.
- Su mayor defecto es las propiedades matemáticas que posee.
- 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.

c) **La Moda:** Es el valor más común entre los datos.

- Si las clases son valores individuales entonces la moda es el valor o los valores que posee(n) la(s) mayor(es) frecuencia(s) absoluta(s).
- Si los datos están agrupados en tablas de frecuencias y las clases son intervalos, la moda viene dada por:

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

donde:

LI_o = Limite inferior de la clase con mayor frecuencia absoluta (clase modal).

Δ_1 = Frecuencia absoluta de la clase modal - Frecuencia absoluta de la clase Pre - modal.

Δ_2 = Frecuencia absoluta de la clase modal - Frecuencia absoluta de la clase Post - modal.

A_o = Amplitud modal.

Propiedades:

- Es muy fácil de calcular.
- No es susceptible de operaciones algebraicas.
- 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.

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 bastante 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 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 11 se muestran las distintas medidas de posición y tendencia central clasificadas de acuerdo al tipo de datos.

Tabla 11. Medidas de Posición y Tendencia Central

Variable	**Escala**	**Medida de Localización**
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.

2. **Medidas de Dispersión.**

Son medidas que permiten medir el grado de agrupación o disgregación en un conjunto de datos. Esto es, permiten determinar si los valores están cercanos o separados entre si. Se pueden clasificar en absolutas y relativas. Las absolutas pueden o no, estar referidas a un valor central. En la tabla 5 se muestran las distintas medidas de dispersión.

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

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 intercuartílico 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 intercuartílico relativo.

a) **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

$$R = Vmax - Vmin$$

Es decir, el rango es la diferencia entre el valor máximo y el valor mínimo

del conjunto de datos.

- **Recorrido Intercuartílico:** Es una medida de la dispersión en la zona intermedia de los datos. Viene dada por la diferencia entre los cuartiles 3 y 4. Esto es,

$$RIC = Q_3 - Q_1$$

Su principal ventaja es que no se ve influenciada por los valores extremos.

- **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:

$$DM = \begin{cases} \dfrac{\sum_{i=1}^{k} |x_i - \overline{x}| f_i}{n}, & \text{Individuales;} \\ \dfrac{\sum_{i=1}^{k} |m_i - \overline{x}| f_i}{n}, & \text{Intervalos.} \end{cases}$$

- **Varianza:** Se define como la media de las diferencias al cuadrado de los datos respecto de su media, es decir,

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

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

$$DM = \begin{cases} \dfrac{\sum_{i=1}^{k}(x_i - \overline{x})^2 f_i}{n-1}, & \text{Individuales;} \\ \dfrac{\sum_{i=1}^{k}(m_i - \overline{x})^2 f_i}{n}, & \text{Intervalos.} \end{cases}$$

Las siguientes fórmulas son usadas comunmente por su facilidad de cálculo

$$DM = \begin{cases} \dfrac{\sum_{i=1}^{k} {x_i}^2 - n\overline{x}^2}{n-1}, & \text{No agrupados;} \\ \dfrac{\sum_{i=1}^{k} x_i^2 f_i - n\overline{x}^2}{n-1}, & \text{Individuales;} \\ \dfrac{\sum_{i=1}^{k} m_i^2 f_i - n\overline{x}^2}{n-1}, & \text{Intervalos.} \end{cases}$$

Dado que esta medida viene expresada en unidades al cuadrado, 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 la variable. Esta

medida recibe el nombre de desviación estándar o típica y esta dada por,

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

Propiedades de la Varianza y Desviación Estándar:

1) La varianza y la desviación estándar no pueden ser negativas.
2) Si todos los datos son iguales a una constante c, entonces $S^2 = 0$ y $S = 0$.
3) Si a cada dato original se le suma una constante k, la varianza y la desviación estándar no se ven afectadas.
4) 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 .
5) 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.

b) **Medidas de Dispersión Relativas**

Por lo general están dados 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,

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

- **Coeficiente de Variación:** Indica el tamaño relativo de la desviación estándar respecto a la media y debe ser calculado para variables cuyos valores son todos positivos. Es la medida de dispersión relativa de mayor uso y su fórmula de calculo es

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

Propiedades:

1) 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.

2) 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.

3. **Medidas de Forma**

Hasta ahora, hemos estado analizando y estudiando la dispersión de una distribución, pero parece evidente que necesitamos conocer más sobre el comportamiento de una distribución. En esta parte, analizaremos las medidas de forma.

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).

a) **Medidas de Asimetría** Las medidas de asimetría permiten saber si los datos se distribuyen en forma simétrica con respecto a su valor central.

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 e datos es calculando ciertos coeficientes de asimetría, a continuación veamos los dos más usados:

- **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.

b) **Medidas de Curtosis.**

Las medidas de apuntamiento (curtosis), miden el grado de apuntamiento o achatamiento de la distribución en su parte central, 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. Una manera de evaluar la curtosis de un conjunto de datos es a través del Coeficiente de Curtosis de Fisher.

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.

1.3. Ejercicios

1.3.1. Introducción

1. Se realiza un estudio en el municipio Libertador del Estado Mérida sobre el tipo de transporte utilizado por sus residentes, para lo cual se encuesta a un grupo de ellos, obteniéndose

Tipo de Transporte	N **de Residentes**
Particular	45
Taxi	25
Trolebús	50
Bus	60
Otros	10

   Identifique:

   a) Universo

   b) Población

   c) Muestra

   d) Variable y tipo de variable.

   e) Tipo de escala.

2. Un fabricante produce tornillos para los cuales existen estrechos márgenes de tolerancia en sus diámetros. El departamento de Control de Calidad selecciona la producción de un día y la somete a proceso de control. Identifique:

   a) Universo

b) Población

c) Muestra

d) Variable y tipo de variable.

e) Tipo de escala.

3. De un lote de 1000 piezas defectuosas se toman al azar 150 de ellas encontrándose con 1,2,3 ó 4 y más defectos, 15, 52, 46 y 37 piezas respectivamente.

   Identifique:

   a) Universo

   b) Población

   c) Muestra

   d) Variable y tipo de variable.

   e) Tipo de escala.

4. Identifique el tipo de variable en cada uno de los siguientes casos:

   a) La resistencia a la ruptura de un determinado tipo de cuerda.

   b) El color del cabello de los niños que estén viendo por televisión una película.

   c) El número de señales de tránsito en poblados con menos de 500 habitantes.

   d) Si una llave de lavamanos esta defectuosa o no.

   e) El número de preguntas contestadas correctamente en un examen.

   f) El tiempo que se necesita para contestar una llamada telefónica en un a oficina de bienes raíces.

g) El resultado de la encuesta hecha a un grupo de votantes posibles acerca del candidato de su preferencia.

h) El gasto en que incurre una empresa al mes en el pago de la nomina.

i) El número de empleados del sexo femenino que hay en una empresa.

j) El precio de un producto en el mercado.

5. Para cada uno de los ítem del ejercicio 5, identifique el tipo de escala más adecuada para realizar la medición.

1.3.2. Organización y Presentación

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 de 40 funcionarios. Agrupar la información en 9 clases.

Sueldo mensual en Miles de BsF.

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 de pers**
$[52-56)$	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 pers**
$[100 - 150)$	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?

1.3.3. Medidas Descriptivas Numéricas

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.

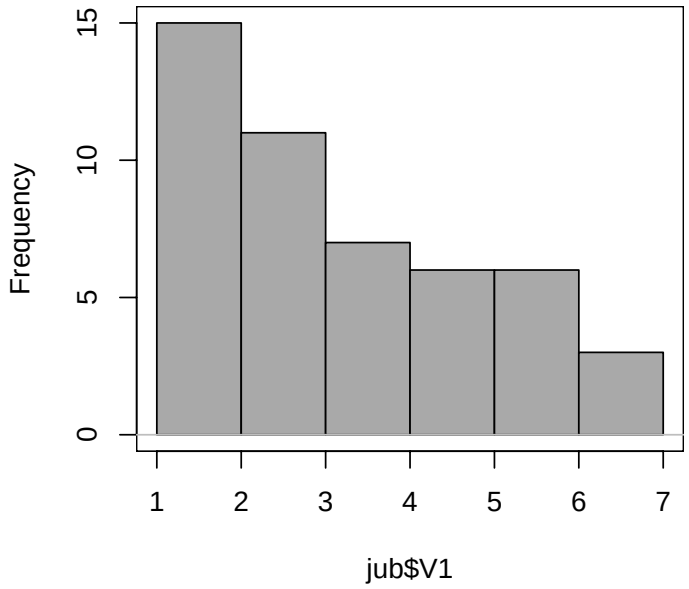

Figura 1.16: Fig.1.