Probabilidad y Estadística

Clase 2:

Distribución de frecuencias, datos agrupados, medidas de tendencia central, media aritmética

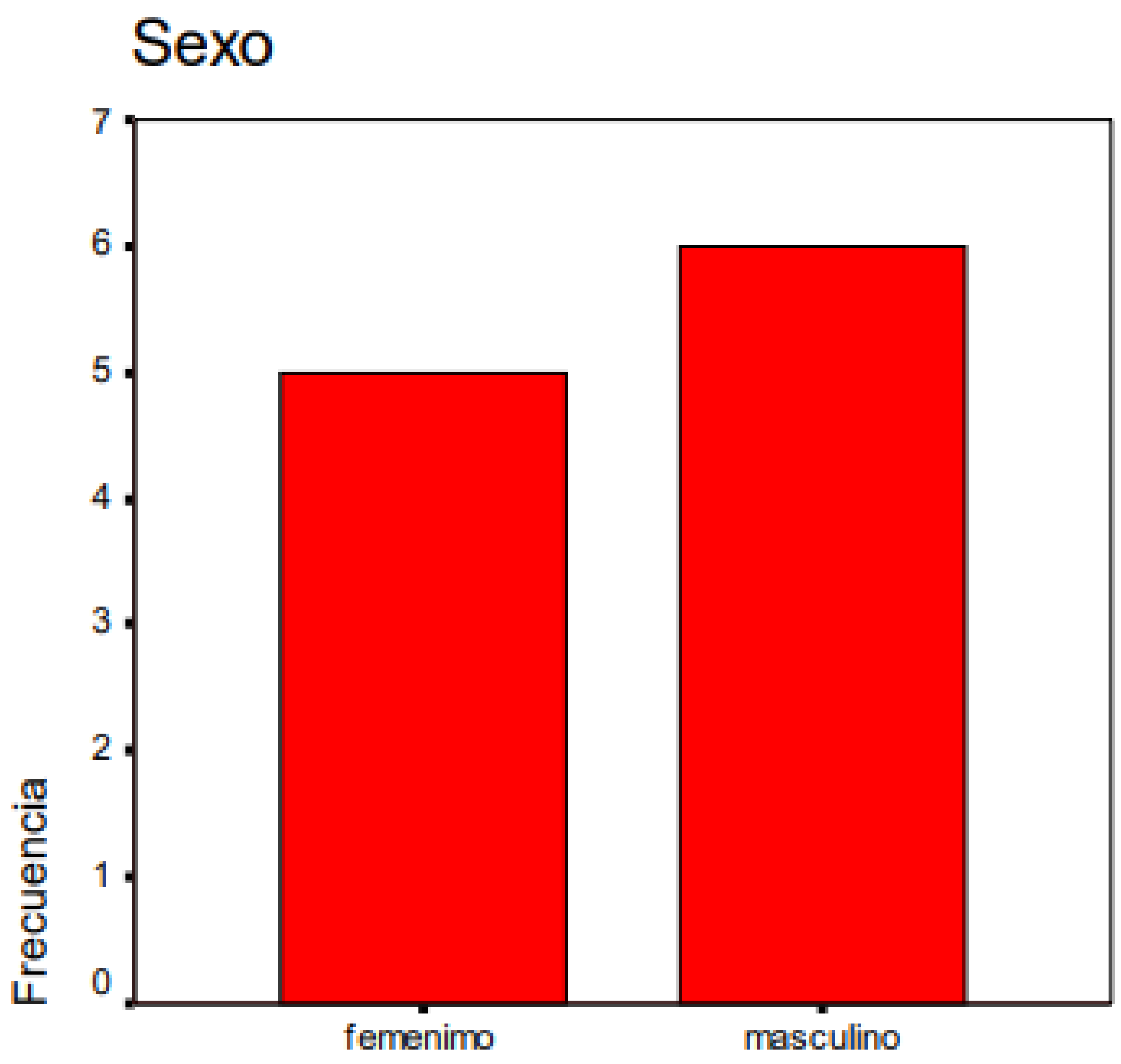
Bianca Di Biaggio

Distribución de frecuencias

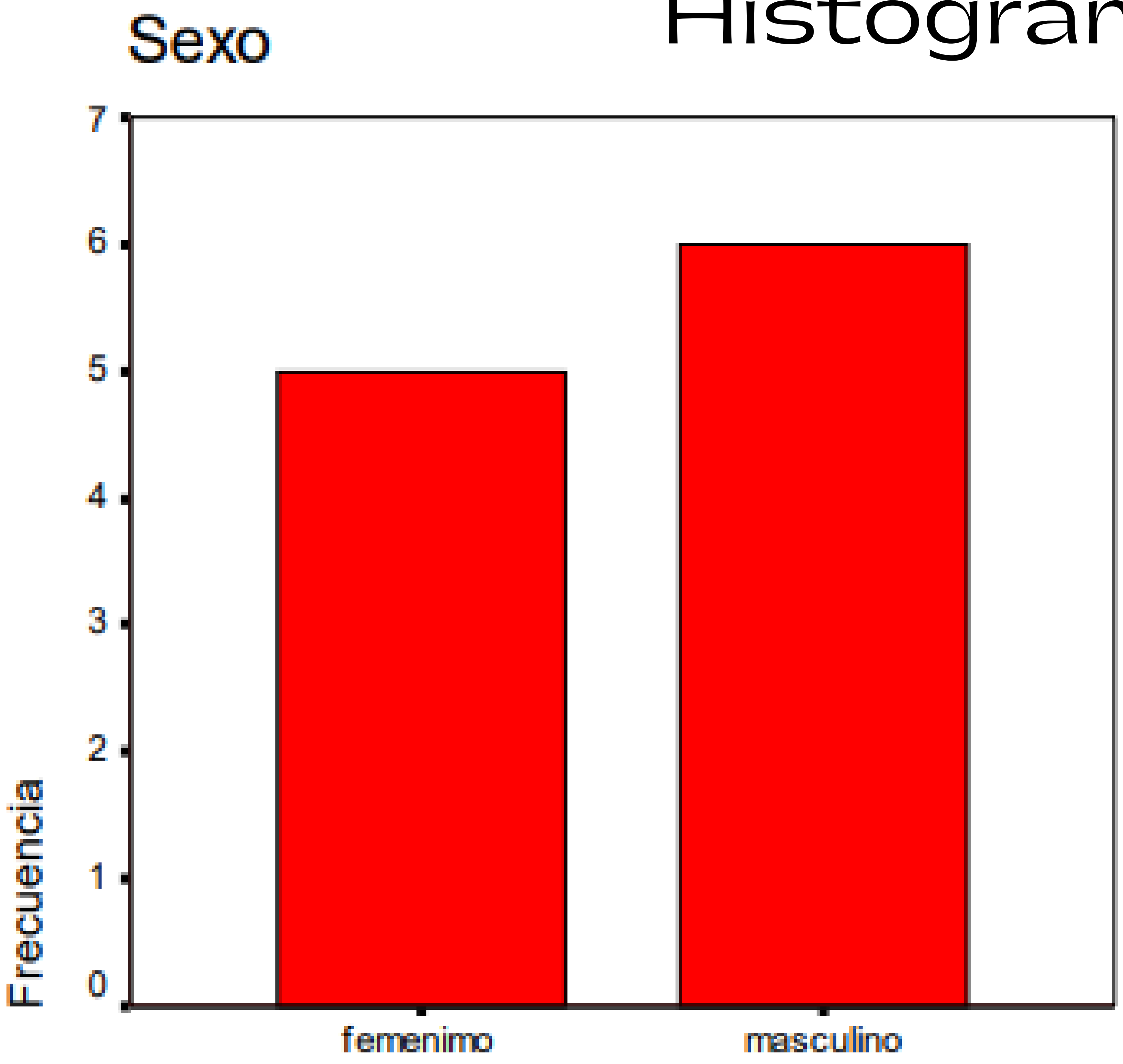
Los datos se van recopilando a medida que los hechos ocurren o se observan por lo que son registrados en una forma desordenada (no agrupados - díficil de trabajar)

Es preciso entonces agrupar los datos a medida que estos aparecen. **A la relación que hace corresponder a los valores de una variable la cantidad de veces que éstos aparecen o son observados se la denomina Distribución de frecuencias** y se la representa en una tabla o en un histograma de frecuencia.

SEXO	CÓDIGO	CANTIDAD O FRECUENCIA
Femenino	1	5.000
Masculino	2	6.000



Histograma de Frecuencia



El eje de coordenadas que generalmente representa a la variable x (abscisas) corresponde al ***sexo***.

El eje de las ordenadas (y) corresponde a la ***frecuencia***.

La escala de este eje está en miles, es decir al 5 le corresponde 5000.

Para cada valor de la variable de estudio se dibujan rectángulos adyacentes de igual ancho y con una altura igual a la frecuencia con que ocurre ese valor.

Frecuencia

En la siguiente tabla se registra la cantidad de personas, clasificada por su relación laboral, de una provincia de Argentina.

Relación Laboral	**Frecuencia**
Empresarios	4648
Trabajadores Autónomos	16423
Trabajadores en relación de Dependencia	2405
Trabajadores Temporal	64981
Trabajadores pasantes	45236
Total	**133693**

Donde relación laboral es una variable cualitativa medida en una escala nominal y se relaciona con la variable frecuencia que es cuantitativa discreta. La variable frecuencia justamente mide la cantidad de cada uno de los tipos de relación laboral.

Propiedades de la Frec. Absoluta

Propiedad 1: Es un número entero mayor o igual a cero

$$f_i \geq 0$$

Propiedad 2: La suma de las frecuencias absolutas es igual al tamaño de la muestra o la cantidad de datos u observaciones.

$$\sum f_i = n$$

Frecuencia Absoluta

Relación Laboral	Frecuencia
Empresarios	4648
Trabajadores Autónomos	16423
Trabajadores en relación de Dependencia	2405
Trabajadores Temporal	64981
Trabajadores pasantes	45236
Total	**133693**

Para poder operar con los datos del cuadro 1.1 o referirnos a ellos, podemos representar la característica a observar mediante la variable X y a la modalidad número **i** de dicha variable con la notación x_i El número de individuos que presentan esa modalidad se llama ***frecuencia absoluta*** y se representará por f_i

Ejemplo**:** Supongamos que i=2, obtendríamos que x_2 = Trabajadores autónomos; entonces observamos que la frecuencia absoluta es f_2 =16.423, ¿Cómo se interpreta? "16.423 es la cantidad de Trabajadores autónomos".

<u>Ejemplo</u>: Supongamos que la variable en estudio es:
X: Cantidad de hijos por mujer.
Entonces debemos contar la cantidad de veces que se repite cada valor de la variable en estudio, frecuencia absoluta.

4 mujeres tienen 1 hijo cada una.
5 mujeres tienen 2 hijos cada una.
8 mujeres tienen 3 hijos cada una.
6 mujeres tienen 4 hijos cada una.
3 mujeres tienen 5 hijos cada una.

El total de mujeres con hijos observados es 26, por lo tanto, la suma de las frecuencias es:

$\sum f_i = n$ Como n = 26 entonces $\sum f_i = 26$

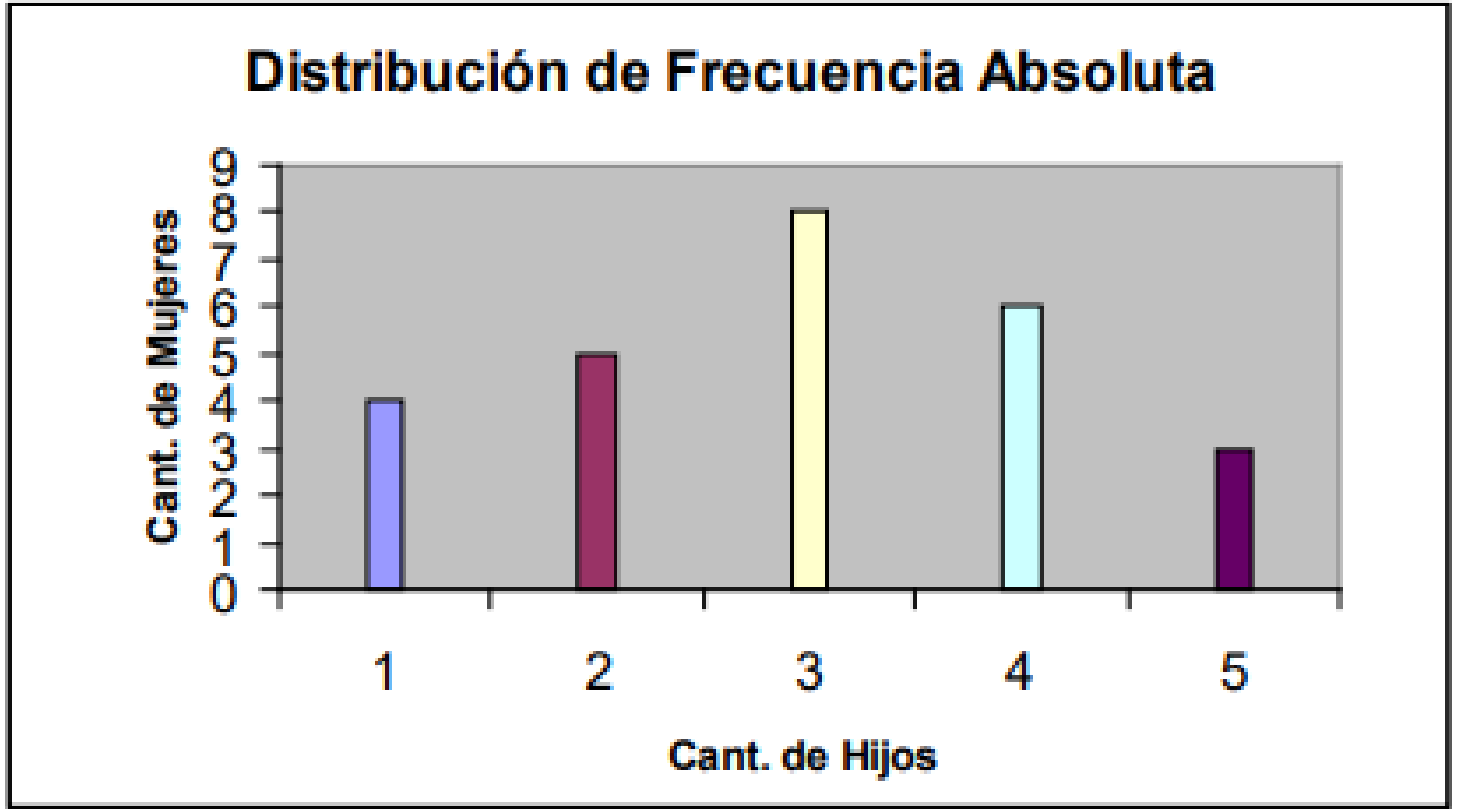
<u>Cuadro 1.2</u>: Distribución de frecuencia absoluta

Cantidad de Hijos x_i	Cantidad de Mujeres f_i
1	4
2	5
3	8
4	6
5	3

Cuadro 1.2: Distribución de frecuencia absoluta

Cantidad de Hijos x_i	Cantidad de Mujeres f_i
1	4
2	5
3	8
4	6
5	3

Observación: Las barras del gráfico deberían ser bastones, es decir, segmentos, ya que corresponde a un número natural. Este gráfico fue construido en Excel y además es muy común ver estos gráficos con rectángulos en lugar de bastones.



Situación Problemática 1:

Supongamos que la variable de estudio es:
X: cantidad de materias que cursa una persona por cuatrimestre

2 personas cursan 1 materia por cuatrimestre.
3 personas cursan 2 materias por cuatrimestre.
6 personas cursan 3 materias por cuatrimestre.
4 personas cursan 4 materias por cuatrimestre.
1 personas cursan 5 materias por cuatrimestre.

a- ¿Cuál es el tipo de Variable que estamos trabajando?
b- Construir una distribución de frecuencia absoluta con su correspondiente gráfico.
c- Interpretar: f_3 y f_5

Frecuencia Absoluta Acumulada

Algunas veces, para una variable es interesante conocer el número de valores que son menores que uno dado. Las **frecuencias absolutas acumuladas** se obtienen: "sumando a la frecuencia absoluta de un determinado valor todas las anteriores". A esta se la simboliza con Fi

Ejemplo 2: Utilizando la distribución de frecuencia del Ejemplo 1, donde la variable es cantidad de hijos por mujer, construiremos una distribución de frecuencias absolutas acumuladas.

Frecunecia absoluta

Cantidad de Hijos x_i	Cantidad de Mujeres f_i
1	4
2	5
3	8
4	6
5	3

x_i	f_i	F_i
1	4	4
2	5	9
3	8	17
4	6	23
5	3	26

Frecunecia absoluta acumulada

Realizaremos algunas interpretaciones referidas a este cuadro:

x_i	f_i	F_i
1	4	4
2	5	9
3	8	17
4	6	23
5	3	26

Sólo 4 mujeres tienen 1 hijo o menos, cada una, luego la frecuencia absoluta acumulada hasta el primer valor de la variable, x = 1, es:

$$F(x=1)=F_1=4$$

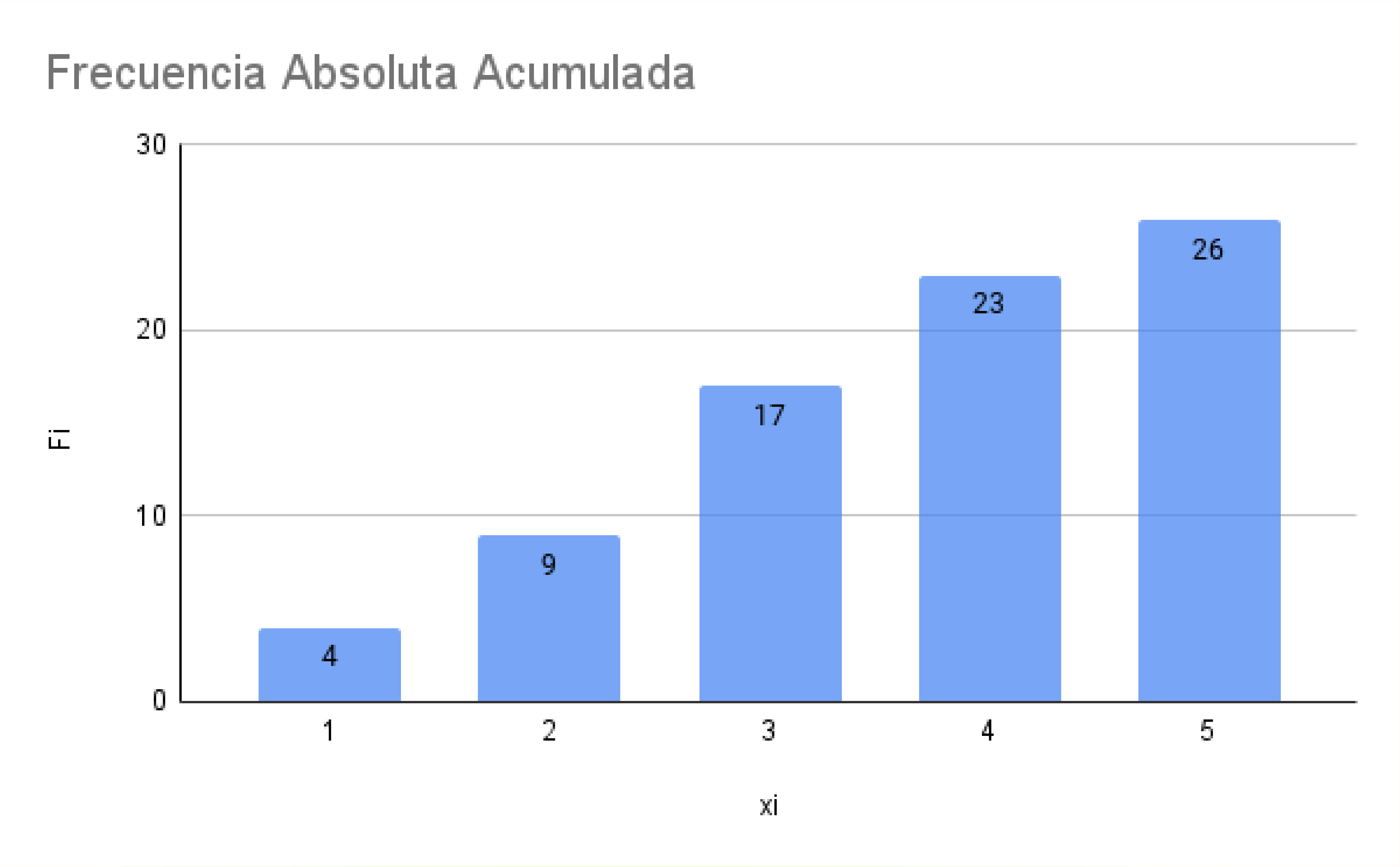
Sólo 9 mujeres tienen 2 hijos o menos cada una (4 que tienen 1 hijo, más 5 que tienen 2 hijos), luego la frecuencia absoluta acumulada hasta el segundo valor de la variable, x = 2, es:

$$F(x=2)=F_2=9$$

Sólo 23 mujeres tienen 4 hijos o menos cada una, luego la frecuencia absoluta acumulada hasta el cuarto valor de la variable, x = 4, es:

$$F(x=4)=F_4=23$$

x_i	f_i	F_i
1	4	4
2	5	9
3	8	17
4	6	23
5	3	26



Propiedades de la Frecuencia Absoluta Acumulada

Propiedad 1: Indica el número de observaciones menores o iguales a un determinado valor de la variable

Propiedad 2: Es un número entero mayor o igual a cero

Propiedad 3: Las frecuencias absolutas acumuladas forman una sucesion finita no decreciente entre 0 y n.

$$0 \leq F_1 \leq F_2 \leq \ldots\ldots\ldots\ldots \leq F_s = n$$

Intervalo de Clases

Para un adecuado análisis estadístico del comportamiento de una variable cuantitativa continua, es necesario agrupar a los valores individuales de ella en clases de equivalencia, llamadas Intervalos de Clases.

Veamoslo a partir de un **ejemplo:** En una empresa se realizaron 25 ventas en un día determinado, cuyos montos, en pesos, son los siguientes:

206,1	216,9	214,4	210,4	228,9
216,1	201,2	203,4	211,3	218,3
208,4	210,0	224,1	212,2	207,8
214,8	204,7	205,8	213,9	217,4
222,4	215,2	219,8	211,4	210,8

Intervalo de Clases

Primer Paso: Hallamos el rango de la variable que es la diferencia entre el Máximo y el Mínimo. A esa diferencia la llamamos "A".

En este caso, el menor valor observado es $201,2 y el mayor es $ 228,9. Entonces: A = $230 - $200 = $30

206,1	216,9	214,4	210,4	228,9
216,1	201,2	203,4	211,3	218,3
208,4	210,0	224,1	212,2	207,8
214,8	204,7	205,8	213,9	217,4
222,4	215,2	219,8	211,4	210,8

Intervalo de Clases

Segundo paso: Se determina la cantidad total de intervalos a utilizar con la siguiente fórmula:

$$h = [1 + \frac{\log(n)}{\log 2}]$$

donde n es el tamaño de la muestra.

En este caso la muestra tiene tamaño 25, entonces:

$$[1 + \frac{\log(25)}{\log 2}] = [5{,}6438...] = 6$$

206,1	216,9	214,4	210,4	228,9
216,1	201,2	203,4	211,3	218,3
208,4	210,0	224,1	212,2	207,8
214,8	204,7	205,8	213,9	217,4
222,4	215,2	219,8	211,4	210,8

Intervalo de Clases

Tercer paso: Se determina la amplitud **a** de cada intervalo utilizando la siguiente fórmula:

$$a = h/A$$

Se simboliza con Li a los límites inferiores de los intervalos y Ls a los límites superiores. A los intervalos se los considera semi-abiertos a la derecha, [Li; Ls).

206,1	216,9	214,4	210,4	228,9
216,1	201,2	203,4	211,3	218,3
208,4	210,0	224,1	212,2	207,8
214,8	204,7	205,8	213,9	217,4
222,4	215,2	219,8	211,4	210,8

En este caso, la amplitud es:

$$a = \frac{A}{h} = \frac{30}{6} = 5$$

El límite inferior del primer intervalo es, entonces, 200. A partir de este número se forman los seis intervalos de amplitud 5 y se procede al conteo; como muestra el siguiente cuadro

Intervalo de Clases

Cuadro 1.4: Distribución de frecuencias absolutas

Intervalos de Clases	Números de Ventas (f_i)
[200;205)	3
[205;210)	4
[210;215)	9
[215;220)	6
[220;225)	2
[225;230)	1

El límite inferior del primer intervalo es, entonces, 200. A partir de este número se forman los seis intervalos de amplitud 5 y se procede al conteo; como muestra el siguiente cuadro

Observar que el cuadro 1.3 y 1.4 tienen ambos el mismo titulo, lo que varia es el tipo de variable.

Cuadro 1.4: Distribución de frecuencias absolutas

Intervalos de Clases	Números de Ventas (f_i)
[200;205)	3
[205;210)	4
[210;215)	9
[215;220)	6
[220;225)	2
[225;230)	1

x_i	f_i	F_i
1	4	4
2	5	9
3	8	17
4	6	23
5	3	26

En general se definen algunas reglas para la formulación de intervalos:

- No establecer ni menos de 5 intervalos ni más de 16
- Los intervalos deben ser entre sí exhaustivos y excluyentes, es decir deben estar ***todos*** los datos incluidos en algún intervalo y cada dato debe pertenecer a un solo intervalo.
- Se deben evitar intervalos cuya frecuencia sea cero.
- No se debe perder información relevante, por ejemplo, máximos, mínimos, etc.
- Si algún valor de la variable coincide con los límites de clase de algún intervalo, se considerará en la clase inmediata superior.

Situación Problemática 5:

En una empresa se realizaron 20 compras en un día determinado, cuyos montos, en pesos, son los siguientes:

105,2	135,8	113,3	108,3	139,9
115,2	100,1	102,3	110,2	117,2
107,3	109,0	123,0	111,1	106,7
112,6	103,6	104,7	118,8	116,3

a- Definir la variable y la frecuencia

b- Clasificar la variable

c- Determinar la cantidad de intervalos y la amplitud de cada uno de ellos.

d- Interpretar f_2, f_3, F_3, F_5

e- ¿Cuál es la frecuencia acumulada que coincide con el tamaño de la muestra?

f- Graficar

Frecuencia Relativa

Para valorar la representatividad de cada categoría respecto al total de datos se calcula la frecuencia relativa (hi), dividiendo la frecuencia absoluta Fi por el número total de observaciones (n).

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

Usualmente, también se utiliza la frecuencia relativa porcentual que es la frecuencia relativa multiplicada por cien:

$$h_i\% = \frac{f_i}{n} \cdot 100$$

Ejemplo 5:

Con las distribuciones de frecuencias del Ejemplo 1, donde la variable es la cantidad de hijos por mujeres, vamos a construir las correspondientes distribuciones de frecuencia relativa y relativa porcentual.

Realizamos entonces el cociente entre la frecuencia absoluta correspondiente a cada valor de la variable y el número de observaciones.

Frecuencia relativa: $h_i = \dfrac{f_i}{n}$

$$h_1 = \frac{4}{26} = 0,15$$

$$h_2 = \frac{5}{26} = 0,19$$

$$h_3 = \frac{8}{26} = 0,31$$

$$h_4 = \frac{6}{26} = 0,23$$

$$h_5 = \frac{3}{26} = 0,12$$

Luego hallamos la frecuencia relativa porcentual, donde hallamos el mismo cociente, pero multiplicado por cien. Entonces:

Frecuencia relativa porcentual:

$$h_i\% = \frac{f_i}{n} \cdot 100$$

$$h_1\% = \frac{4}{26} \cdot 100 = \%15$$

$$h_2\% = \frac{5}{26} \cdot 100 = \%19$$

$$h_3\% = \frac{8}{26} \cdot 100 = \%31$$

$$h_4\% = \frac{6}{26} \cdot 100 = \%23$$

$$h_5\% = \frac{3}{26} \cdot 100 = \%12$$

Propiedades de la Frecuencia Relativa

Propiedad 1: Es un número comprendido entre 0 y 1.

En símbolos: $0 < h_i < 1$

Propiedad 2: La suma de las frecuencias relativas es igual a 1.

En símbolos: $\sum h_i = \sum \frac{f_i}{n} = 1$

Propiedades de la Frecuencia Relativa Porcentual

Propiedad 1: Es un número comprendido entre 0 y 100.

En símbolos: $0 < h_i\% < 100$

Propiedad 2: La suma de las frecuencias relativas porcentuales es igual a 100%.

En símbolos: $\sum h_i\% = 100\%$

Frecuencia Relativa y Relativa porcentual Acumulada

Dado el siguiente cuadro 2.1 ya visto anteriormente recordando el concepto de frecuencia acumulada y razonando por analogía la frecuencia relativa acumulada será:

x_i	h_i	$h_i\%$	H_i	$H_i\%$
1	0,15	15	0,15	15
2	0,19	19	0,34	34
3	0,31	31	0,65	65
4	0,23	23	0,88	88
5	0,12	12	1,00	100

Frecuencia relativa acumulada: $H_i = \frac{F_i}{n}$

$H_1 = \frac{4}{26} = 0{,}15$

$H_2 = \frac{9}{26} = 0{,}34$ y así sucesivamente.

Frecuencia relativa porcentual: $H_i\% = \frac{F_i}{n} \cdot 100$

$H_1\% = \frac{4}{26} \cdot 100 = 15\%$

$H_2\% = \frac{9}{26} \cdot 100 = 34\%$ y así sucesivamente.

Situación Problemática 7:

Utilizando la distribución de frecuencia del Ejemplo 4, donde la variable es el monto de las ventas, construir las correspondientes distribuciones de frecuencia relativa y relativa acumulada.

Contestar:

a- ¿Qué porcentajes de ventas tuvieron un monto entre $210 y $215?
b- ¿Qué porcentajes de ventas tuvieron un monto inferior a $215?
c- ¿Qué porcentajes de ventas tuvieron un monto superior a $220?
d- Interpretar $h_2\%$ y H_5
e- Interpretar $H_4\%$

Medidas de Tendencia Central

Para realizar un análisis estadístico es preciso repetir el experimento varias veces bajo condiciones similares, entonces para una variable en particular se cuenta con varios valores observados.

Estos valores tienden a agruparse alrededor de algunos puntos centrales (de ahí su nombre) que fijan una posición.

Estas medidas entonces resumen los datos en un solo valor que los representa.

Medidas de Tendencia Central

Promedios Simples: Los promedios pueden ser calculados sólo cuando las variables son cuantitativas, porque las variables cualitativas no admiten operaciones algebraicas.

1) Promedio o Media Aritmética:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

El ejemplo más sencillo sería el cálculo del promedio de notas obtenidas en un examen por un grupo de alumnos

Medidas de Tendencia Central

Si los datos **están organizados en una distribución de frecuencia**, es decir que tengo una tabla de frecuencias entonces la fórmula será la misma, pero multiplicando cada valor de la variable por la frecuencia correspondiente. O sea:

$$\bar{x} = \frac{x_1 . f_1 + x_2 . f_2 + \dots\dots + x_n . f_n}{n} = \sum_{i=1}^{n} \frac{x_i . f_i}{n}$$

Si los datos están agrupados en intervalos, se toma el punto medio del intervalo (como el punto medio de un segmento) que se lo nota: ' i x Luego la fórmula resulta:

$$\bar{x} = \frac{x'_1 . f_1 + x'_2 . f_2 + \dots\dots + x'_n . f_n}{n} = \sum_{i=1}^{n} \frac{x'_i . f_i}{n}$$

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

Situación Problemática 8:

Supongamos que la variable de estudio es:

X = edad del alumno (podríamos decir cantidad de años del alumno, pero en lenguaje coloquial no utilizamos esta forma).

1 persona tiene 22 años.
5 personas tienen 23 años.
2 personas tienen 24 años.
2 personas tienen 25 años.
6 personas tienen 26 años.
3 personas tienen 27 años.

1 persona tiene 29 años.
5 personas tienen 30 años.
2 personas tienen 31 años.
1 persona tiene 32 años.
1 persona tiene 33 años.
1 persona tiene 36 años.

a- Encontrar el Promedio o la Media Aritmética.

Propiedades de la Media Aritmética

Prop 1: La media aritmética pertenece al campo de variación de la variable.

$$x_1 \leq \bar{x} \leq x_n$$

Prop 2: La media aritmética de una constante es una constante.

$$\text{Si } x_1 = x_2 = x_3 = \ldots\ldots\ldots = x_n = c \quad \Rightarrow \quad \bar{x} = c$$

En símbolo: $\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} = \sum_{i=1}^{n} \frac{c}{n} = n.\frac{c}{n} = c$

Prop 3: La media aritmética del producto de una constante por una variable es igual a la constante por la media aritmética de la variable.

En símbolo: $\sum_{i=1}^{n} \frac{c.x_i}{n} = c.\sum_{i=1}^{n} \frac{x_i}{n} = c.\bar{x}$

Esta propiedad indica que al realizar un cambio de unidad de medida a los datos (por ejemplo, si pasamos de gr. a Kg.), la medida estará afectada por dicho cambio de escala.

Propiedades de la Media Aritmética

Prop 4: La media aritmética de la suma de una constante más una variable es igual a la constante más la media aritmética de la variable.

$$(\bar{c} + \bar{x}) = c + \bar{x}$$

En símbolos: $$\sum_{i=1}^{n} \frac{c. + x_i}{n} = \sum_{i=1}^{n} \frac{c}{n} + \sum_{i=1}^{n} \frac{x_i}{n} = c + \bar{x}$$

También es cierto que la media aritmética de la suma de dos variables x e y es igual a la suma de las medias aritméticas de cada una de dichas variables.

Prop 5: La suma de desvíos dados por la diferencia entre cada valor observado de la variable y la media aritmética es igual a cero.

En símbolos: $$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

Propiedades de la Media Aritmética

En una distribución de frecuencia esta propiedad se escribe:

En símbolos: $$\sum_{i=1}^{n}(x_i - \bar{x})f_i = 0$$

Prop 6: La suma de los cuadrados de los desvíos con respecto a la media aritmética es un valor mínimo.

En símbolos: $$\sum_{i=1}^{n}(x_i - \bar{x})^2 = mínimo$$

Media Ponderada

No siempre los valores de una variable tienen la misma importancia o el mismo peso por este motivo es que recurrimos a “ponderar” es decir a darle un peso relativo a determinados valores de dicha variable.

El promedio ponderado surge de la suma del producto entre la ponderación y el valor de la variable, dividido por la suma de las ponderaciones.

$$\bar{x}_p = \frac{\sum x_i . w_i}{\sum w_i}$$

Ejemplo 8:

Preparamos un envase con café de 3 calidades diferentes:
- 2,5kg de calidad A
- 6,0kg de calidad B
- 1,5kg de calidad C

Los precios respectivos son: $9, $7, y $12. Calcular el precio promedio por kilo del envase.

Calidad	Precio (x_i)	Kg. (w_i)	$x_i \cdot w_i$
A	9	2,5	22,5
B	7	6,0	42
C	12	1,5	18
		$n = \sum f_i = 10$	82,5

$$\bar{x}_p = \frac{\sum x_i . w_i}{\sum w_i} = \frac{82,5}{10} = 8,25$$

Situación Problemática 12:

Hay 10 personas en un ascensor, 4 mujeres y 6 hombres. El peso medio de las mujeres es 60 kilos y el peso medio de los hombres es de 80 kilos
¿Cuál es el peso medio de las 10 personas del ascensor?

Situación Problemática 13:

Supongamos que el sueldo promedio de 47 empleados hombres de una empresa es de $970, mientras que el sueldo promedio de las mujeres es de $860, donde el total de empleados es de 82 personas, determinar:

a- el sueldo promedio de todos los empleados de la empresa.
b- el nuevo sueldo promedio ante un aumento general del 5%.