Integrando.

CURSOS ACADÉMICOS

Integrando.

CURSOS ACADÉMICOS

Capítulo 1

Estadística descriptiva

Temario del capítulo 1

Integrando.
CURSOS ACADÉMICOS

1.1 Variables y recopilación de datos

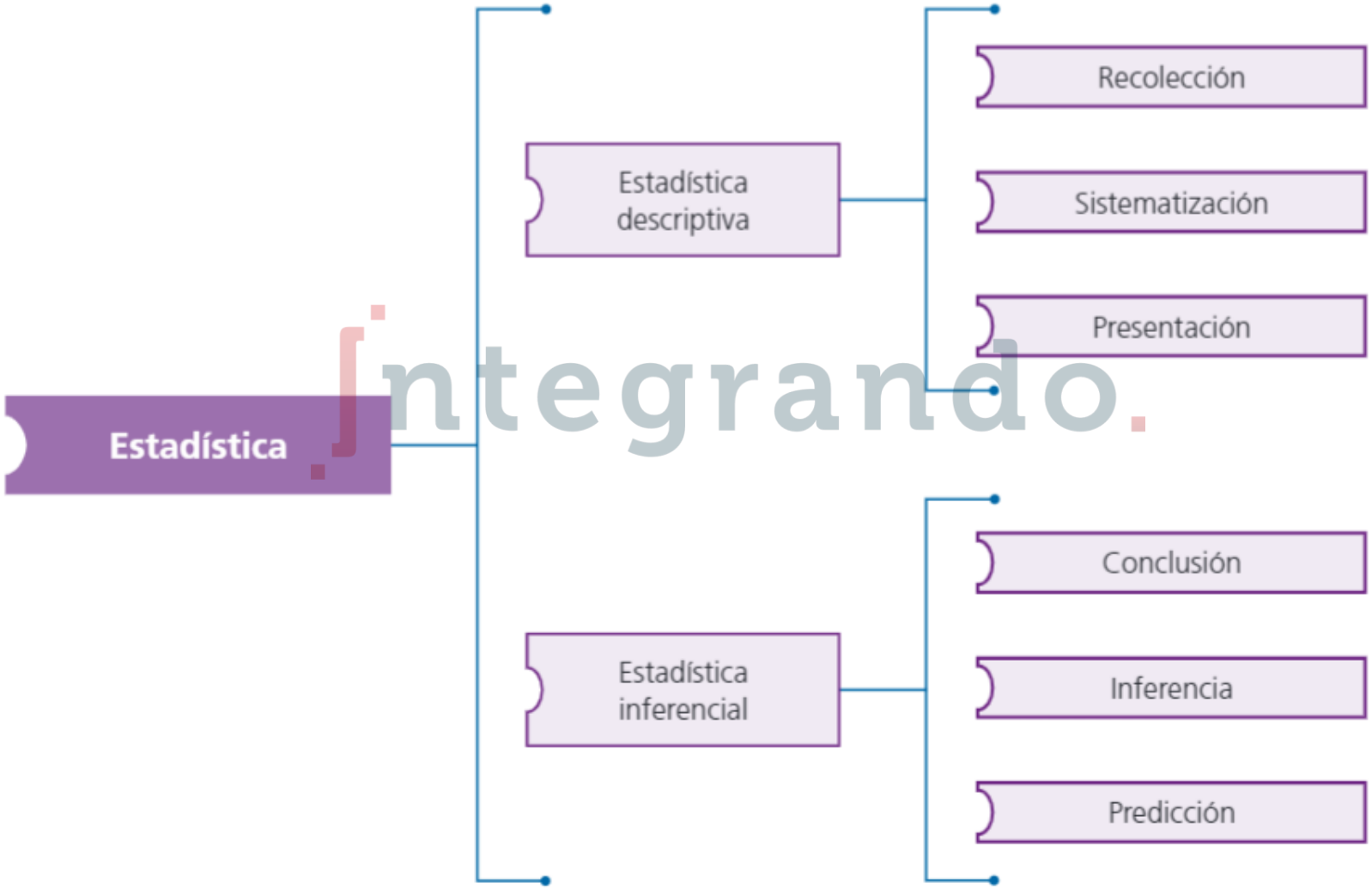
Integrando.
CURSOS ACADÉMICOS

El estudio descriptivo de la estadística consiste en la recolección, sistematización y presentación de datos obtenidos al analizar una población.

Los datos son valores que pueden ser observados o medidos, mientras que la variable es un característica que puede tener distintos valores.

Por ejemplo: la edad, estatura y peso de una persona son tipos de *variables.* Los valores particulares de cada variable en el caso de Rubén son: 15 años, 1.74 metros y 65 kilogramos; estas medidas constituyen sus *datos.*

1.1 Variables y recopilación de datos



1.1 Variables y recopilación de datos

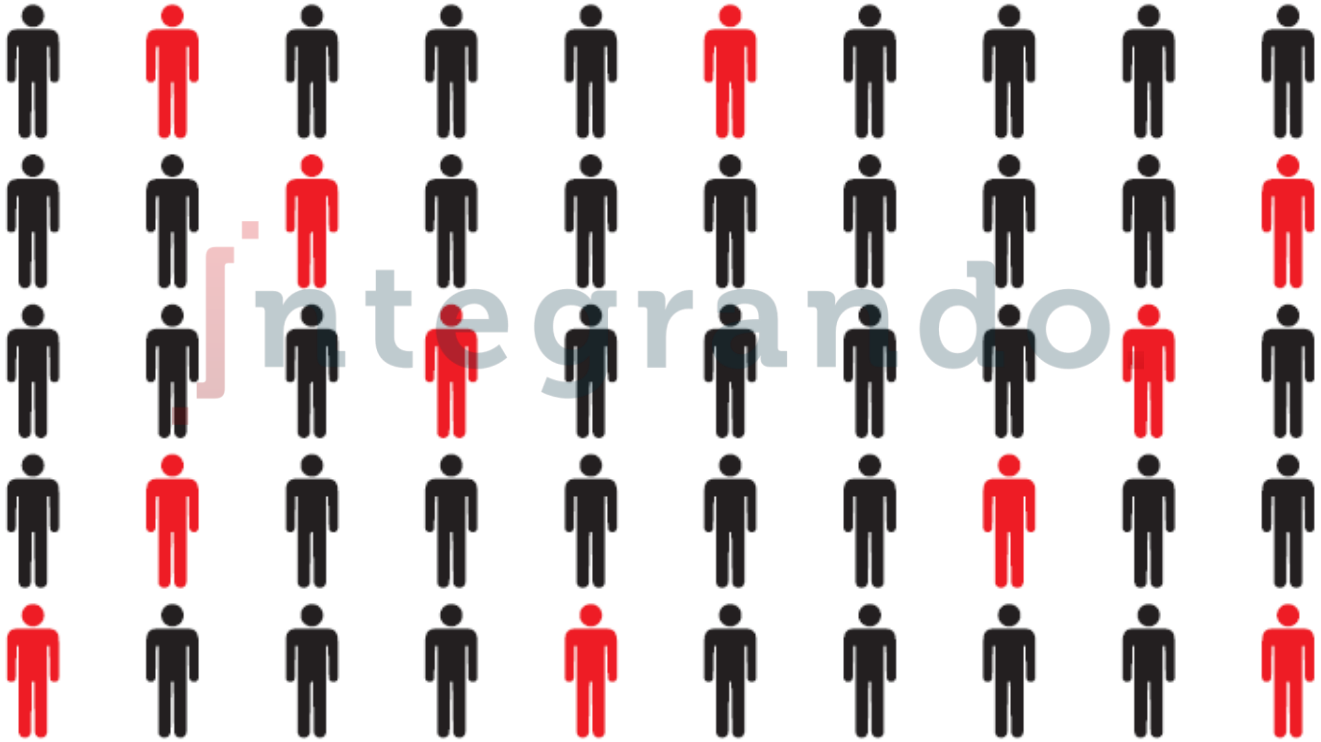
Las variables pueden denotar la parte cualitativa o cuantitativa del fenómeno; esta última puede ser del tipo discreto, como el número de habitaciones de una casa, o continuo, como la cantidad de agua que consume una persona en un día.

La población es el conjunto de objetos de estudio, los cuales suelen tener características comunes.

Una muestra es un subconjunto representativo de la población: debe tener un tamaño adecuado y considerar a la variedad de sujetos que la conforman.

1.1 Variables y recopilación de datos

Población = Conjunto de todos los elementos



Muestra = Elementos seleccionados

1.1 Variables y recopilación de datos

Integrando.
CURSOS ACADÉMICOS

Es posible recurrir a fuentes para recabar información confiable y de forma rápida; las fuentes se clasifican como:

- Primarias: cuando se observa o se experimenta directamente con el fenómeno de interés, o bien, al realizar encuestas o cuestionarios a la población involucrada

- Secundarias: con datos recabados anteriormente por alguna institución o persona

1.1 Variables y recopilación de datos

Integrando.
CURSOS ACADÉMICOS

Un experimento permite establecer los posibles resultados de un fenómeno en ciertas circunstancias.

Para recopilar la información de cada miembro de una población se realiza un censo; por otro lado, el muestreo es un proceso basado en diferentes métodos para obtener una muestra y realizar proyecciones.

El parámetro es una medida que se calcula para describir una característica de la población, mientras que el estadístico es la medición que describe una propiedad de la muestra.

1.1 Variables y recopilación de datos

Integrando.
CURSOS ACADÉMICOS

Por ejemplo, en un estudio que se realiza para conocer el

Ingreso promedio en México

- La *población* será el conjunto de los habitantes del país que perciben un salario, la población económicamente activa (PEA)
- El *parámetro* es el ingreso promedio de la PEA
- El *estadístico* podría ser el ingreso promedio de una *muestra* que incluya algunas personas del centro, del sur y del norte

1.2 Frecuencias y presentación de datos

Integrando. CURSOS ACADÉMICOS

Una vez que se han recolectado los datos es necesario organizarlos y presentarlos de modo que proporcionen información útil; una forma de lograrlo es contabilizando la frecuencia con la que ocurren ciertos valores.

Una distribución o tabla de frecuencias permite observar dónde se localiza la mayor parte de la población, o bien, si los datos son muy dispersos entre sí.

La frecuencia absoluta (o simplemente frecuencia) denota el número de individuos que comparten el mismo valor de una variable i y se representa como f_i. La frecuencia relativa f_r indica el porcentaje de cada variable respecto al total y la frecuencia acumulada f_a es igual a la suma de las frecuencias anteriores hasta la actual.

1.2 Frecuencias y presentación de datos

Integrando. CURSOS ACADÉMICOS

Como ejemplo, analicemos los siguientes datos no agrupados de una encuesta realizada a 20 jóvenes sobre la cantidad de refrescos que beben al día:

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

No. de Refrescos	f_i	f_a	f_r (%)
0	2	2	10
1	4	6	20
2	4	10	20
3	3	13	15
4	3	16	15
5	4	20	20
	$\Sigma f_i = 20$		$\Sigma f_r = 100\%$

1.2 Frecuencias y presentación de datos

Integrando. CURSOS ACADÉMICOS

Cuando el número de datos es grande es conveniente distribuirlos a partir de clases o categorías; a esto se le como como distribución de frecuencias para datos agrupados.

Como un caso particular, consideremos la siguiente tabla que muestra el tiempo en minutos que le toma a 50 trabajadores llegar de su casa al trabajo.

Intervalo de tiempo	f_i	f_a	f_r (%)
$5 - 10$	12	12	24
$11 - 16$	14	26	28
$17 - 22$	15	41	30
$23 - 28$	9	50	18

1.2 Frecuencias y presentación de datos

Integrando. CURSOS ACADÉMICOS

Cada clase se conforma por un intervalo entre el límite inferior y el límite superior de la misma

$$5-10 \quad \begin{cases} 10 & límite\ superior \\ 5 & límite\ inferior \end{cases}$$

El promedio entre el límite superior de una clase y el límite inferior de la siguiente se llama frontera de clase o límite real; representan valores que teóricamente están contabilizados dentro del intervalo de clase

$$5-10$$

$$\frac{10+11}{2}=10.5$$

$$11-16$$

$$\frac{16+17}{2}=16.5$$

$$17-22$$

1.2 Frecuencias y presentación de datos

Integrando. CURSOS ACADÉMICOS

El ancho o amplitud de clase c es la diferencia entre las fronteras de cada clase

$$c = Frontera\ superior - Frontera\ inferior$$

Por debajo del límite inferior de la primera clase y por encima del límite superior de la última también existe una frontera, tal que el ancho sea el mismo.

Se conoce como marca de clase X al punto medio del intervalo de clase, y funciona como una parte representativa de cada categoría

$$X = \frac{Límite\ inferior + Límite\ superior}{2}$$

1.2 Frecuencias y presentación de datos

Integrando. CURSOS ACADÉMICOS

El rango R de la distribución es la diferencia entre el dato mayor y el dato menor; si los datos están agrupados, se usan los valores de la frontera superior e inferior

$$R = Dato\ mayor - Dato\ menor$$

Ahora podemos completar la tabla de frecuencias para el caso previo:

Ancho de clase

$$c = 10.5 - 4.5$$
$$c = 16.5 - 10.5$$
$$c = 22.5 - 16.5$$
$$c = 28.5 - 22.5$$
$$c = 6$$

Intervalo de tiempo	Límite inferior	Límite superior	Frontera inferior	Frontera superior	Marca de clase
5 – 10	5	10	4.5	10.5	7.5
11 – 16	11	16	10.5	16.5	13.5
17 – 22	17	22	16.5	22.5	19.5
23 – 28	23	28	22.5	28.5	25.5

Rango

$$R = 28.5 - 4.5 = 24$$

1.2 Frecuencias y presentación de datos

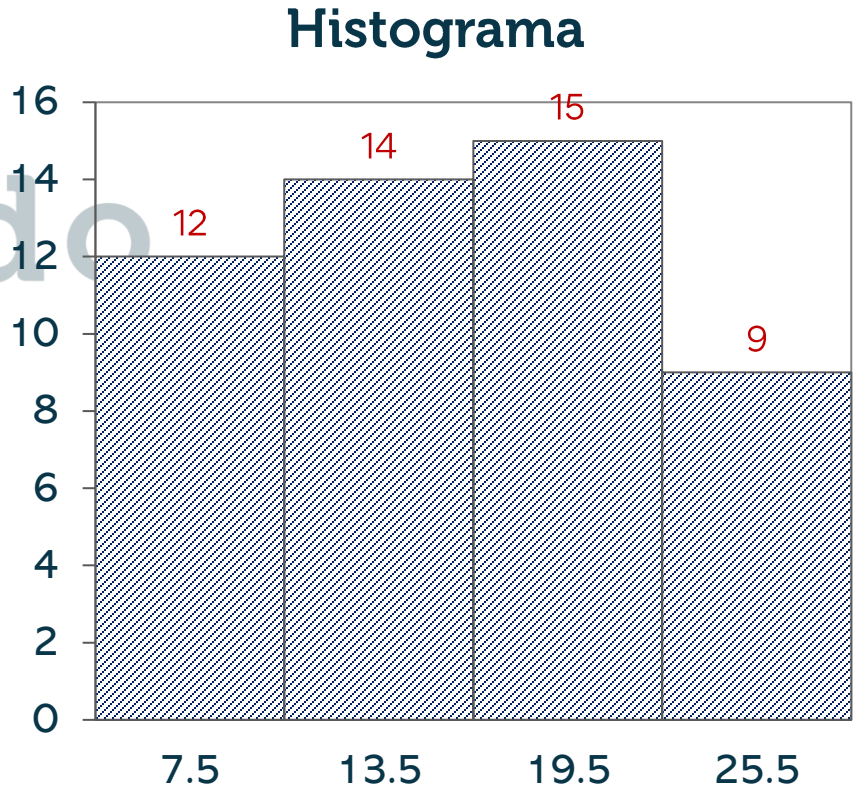
Una vez que se tiene la tabla de frecuencias, se busca una forma conveniente de visualizar los resultados. Las formas usuales para presentar datos son:

i. Histograma

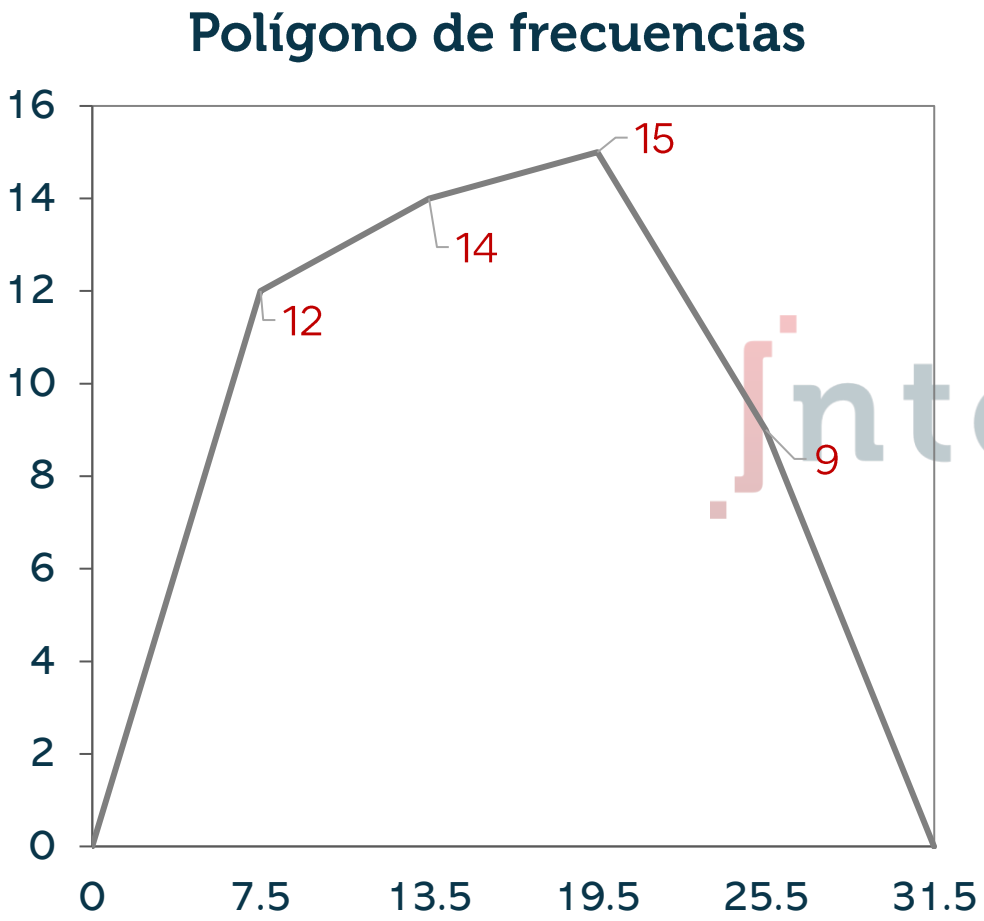
Gráfica de barras rectangulares, sin espacios, para mostrar la forma de la distribución.

El eje horizontal contiene las variables, representadas por la marca de clase, y el eje vertical las frecuencias.

Cada columna abarca un intervalo de clase.



1.2 Frecuencias y presentación de datos



ii. Polígono de frecuencias

Gráfica de líneas continuas para observar cómo es la distribución de datos.

Para cada clase se dibuja un punto, la coordenada x es la marca de clase y la y la frecuencia.

Se unen los puntos por medio de rectas, incluyendo el cero tanto al inicio como al final

1.2 Frecuencias y presentación de datos

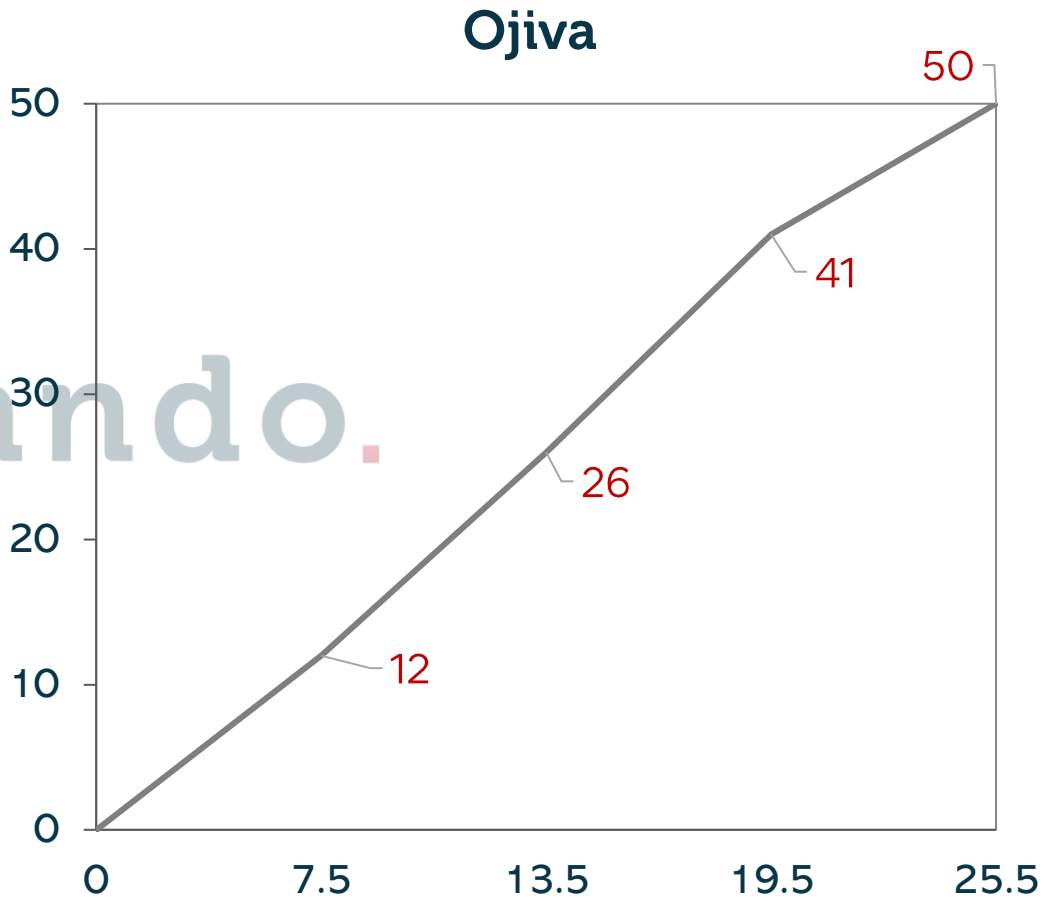
Integrando.
CURSOS ACADÉMICOS

iii. Ojiva

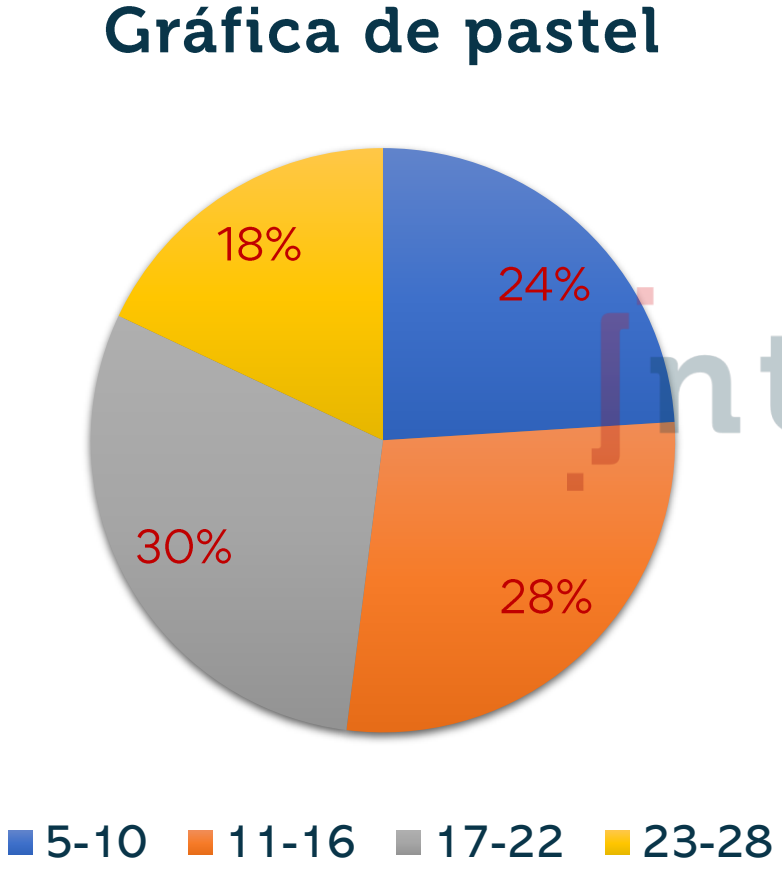
Gráfica de líneas continuas que muestra la distribución de frecuencias acumuladas (absoluta o relativa)

Para cada clase se dibuja un punto, la coordenada x es la marca de clase y la y la frecuencia acumulada.

Se unen los puntos por medio de rectas, incluyendo el cero al inicio.



1.2 Frecuencias y presentación de datos



iv. Diagrama de pastel

Gráfica circular para representar las frecuencias relativas.

El círculo se divide en sectores, cada uno de tamaño correspondiente al porcentaje del total.

A cada porcentaje se le asocia un color y se indican las clases.

1.3 Medidas de tendencia central

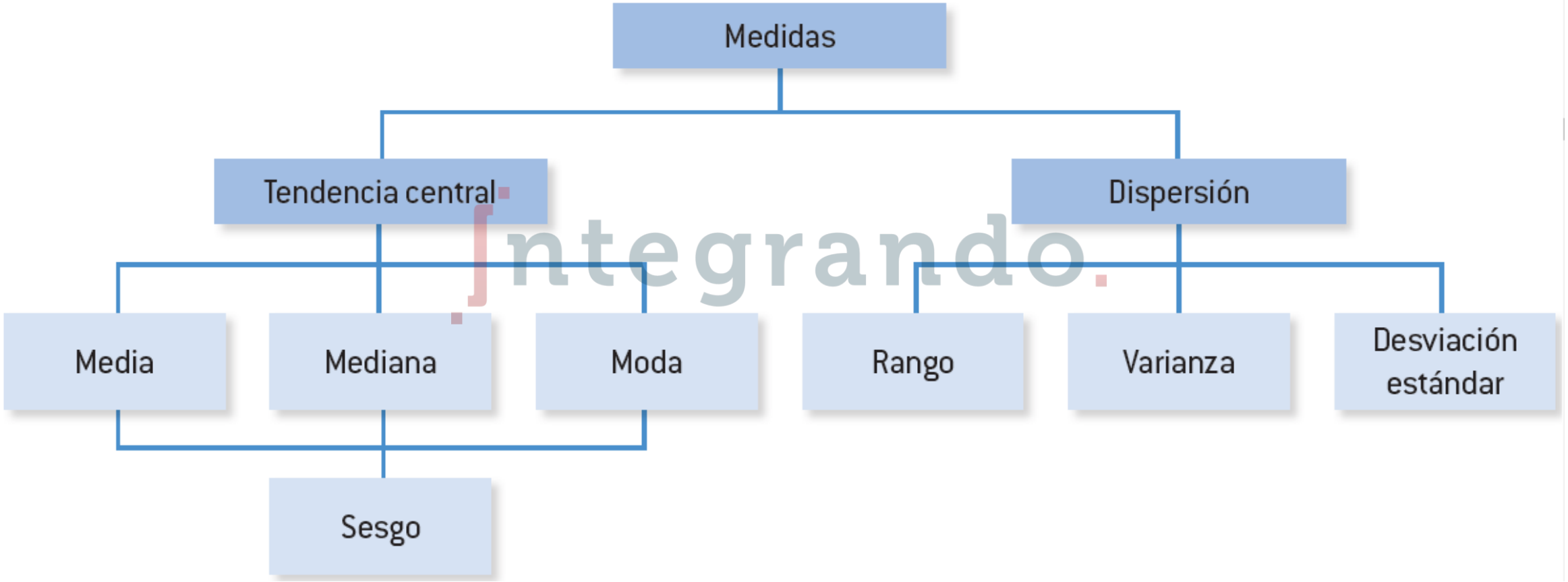
Para comprender con mayor profundidad el comportamiento de las distribuciones se realiza un análisis numérico alrededor de un valor central basado en:

1. Acumulación de datos
2. Dispersión de datos

Las medidas de tendencia central permiten identificar el valor más representativo del conjunto de datos; esto es, el valor central.

Las medidas de dispersión o de variabilidad describen cuánto se acercan o alejan los datos con respecto al valor central.

1.3 Medidas de tendencia central



1.3 Medidas de tendencia central

Integrando. CURSOS ACADÉMICOS

La primera de las medidas de tendencia central es el promedio o media aritmética $\bar{x}$; se define como la suma de todos los valores x_i dividida por la cantidad de datos N

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{\sum_{i=1}^{N} f_i}$$

Para una distribución de datos agrupados, se obtiene al sumar todos los productos de la frecuencia f_i por la marca de clase X_i y dividir por el total de datos N

$$\bar{x} = \frac{f_1 \cdot X_1 + f_2 \cdot X_2 + \cdots + f_N \cdot X_N}{N} = \frac{\sum_{i=1}^{N} f_i \cdot X_i}{\sum_{i=1}^{N} f_i}$$

1.3 Ejemplos

1. Determina la media para la siguiente distribución de frecuencias:

x_i	f_i
1	3
2	4
3	2
4	6
5	5

a) $\bar{x} = 3.3$

1.3 Ejemplos

2. Completa la tabla de la siguiente distribución de frecuencias para determinar el promedio.

x_i	f_i	X_i	$f_i \cdot X_i$
$0 - 99$	16		
$100 - 199$	12		
$200 - 299$	14		
$300 - 399$	2		
$400 - 499$	1		

a) $\bar{x} = 160.6$

1.3 Medidas de tendencia central

Integrando. CURSOS ACADÉMICOS

La siguiente medida central es la mediana Me; para obtenerla es necesario ordenar los datos de menor a mayor.

Si los datos no están agrupados, se obtiene de la siguiente forma:

i. Si el número de datos es impar, la mediana es el dato ubicado en el centro, es decir, la posición

$$\frac{n+1}{2}$$

ii. Si el número de datos es par, la mediana es el promedio de los datos centrales, esto es, en las posiciones

$$\frac{n}{2}, \frac{n}{2}+1$$

1.3 Medidas de tendencia central

Integrando. CURSOS ACADÉMICOS

Si los datos están agrupados, la mediana se obtiene de la manera siguiente:

i. Localizar la posición de la mediana como en el caso no agrupado
ii. Identificar el límite inferior L_M y la frecuencia f_M de la clase donde se ubica la mediana
iii. Calcular la frecuencia acumulada f_a anterior a la clase de la mediana
iv. Utilizar la fórmula

$$Me = L_M + \left[\frac{\frac{N}{2} - f_{a,M-1}}{f_M}\right] c$$

donde N es el número de datos y c el ancho de clase

1.3 Ejemplos

3. Una compañía registró los siguientes ingresos por ventas mensuales en miles de dólares, durante siete meses. Determina la mediana.

$$58, 56, 67, 54, 48, 50, 63$$

4. Encontrar la mediana del siguiente conjunto de datos:

$$6, 5, 9, 7, 18, 5, 12, 12, 11, 15$$

a) $Me = 56$ b) $Me = 10$

1.3 Ejemplos

5. Completa la siguiente tabla con las frecuencias acumuladas para calcular la mediana de la distribución

x_i	f_i	$f_{a,i}$
$20-29$	10	
$30-39$	16	
$40-49$	27	
$50-59$	32	
$60-69$	15	

a) $Me = 48.9$

1.3 Medidas de tendencia central

Integrando.
CURSOS ACADÉMICOS

El dato que se repite con mayor frecuencia recibe el nombre de moda Mo. Existen 3 caso posibles:

a) Existe un dato con mayor frecuencia que los demás, por lo tanto, existe la moda

b) Más de una cantidad aparece con la frecuencia más grande, entonces hay más de una moda. Esto recibe el nombre de distribución multimodal

c) Si no hay algún dato con frecuencia mayor a la de los demás, la moda no existe

1.3 Medidas de tendencia central

Integrando. CURSOS ACADÉMICOS

Cuando se tiene una distribución de datos agrupados, se llevan a cabo los siguientes pasos para encontrar la moda:

i. Identificar la clase con mayor frecuencia, llamada clase modal
ii. Ubicar el límite inferior L_M y frecuencia f_M de la clase modal
iii. Localizar las frecuencias anterior f_{M-1} y posterior f_{M+1} a la de la clase modal
iv. Utilizar la fórmula

$$Mo = L_M + \left[\frac{f_M - f_{M-1}}{(f_M - f_{M-1}) + (f_M - f_{M+1})}\right] c$$

donde c es el ancho de clase

1.3 Ejemplos

Integrando.
CURSOS ACADÉMICOS

6. Determina la moda de los siguientes conjuntos de datos:

a) $4, 7, 4, 5, 7, 4, 8, 5$

b) $52, 48, 50, 49, 57, 46, 51, 45$

c) $9, 8, 9, 7, 5, 6, 3, 4, 8$

a) $Mo = 4$ b) No existe c) $Mo = 8, 9$

1.3 Ejemplos

Integrando. CURSOS ACADÉMICOS

7. Encuentra la moda para la siguiente distribución de frecuencias:

x_i	f_i
$50 - 59$	3
$60 - 69$	7
$70 - 79$	18
$80 - 89$	8
$100 - 109$	2

a) $Mo = 76.5$

1.4 Medidas de dispersión

Integrando.
CURSOS ACADÉMICOS

Para analizar cuán cercanos o lejanos están los datos, ya sea entre los extremos, o bien con respecto al promedio, usaremos las medidas de dispersión.

El rango R, también llamado recorrido, es igual a la diferencia entre los valores máximo y mínimo de los datos, una vez que han sido ordenados

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

Esta medida solo involucra los extremos, por lo que no describe el tipo de distribución ni depende de la media.

1.4 Ejemplos

1. Los siguientes datos indican el tiempo en segundos que tomó a 10 automóviles acelerar desde el reposo hasta alcanzar una rapidez de 100 km/h. ¿Cuál es el rango?

$$15, 12, 18, 17, 17, 11, 19, 16, 15, 30$$

a) $R = 19$

1.4 Medidas de dispersión

Integrando.
CURSOS ACADÉMICOS

Para encontrar la variación de los datos con respecto al promedio se emplea la desviación media, la varianza y la desviación estándar.

La desviación media D_m permite determinar el valor absoluto de la separación entre cada dato x_i y la media aritmética $\bar{x}$

$$D_m = \frac{\sum_{i=1}^{N} |x_i - \bar{x}|}{N}$$

Se usa el valor absoluto ya que la resta es negativa para los datos menores que el promedio, pero la única información relevante es la distancia entre datos, mas no el signo.

1.4 Ejemplos

2. Completa la tabla y encuentra la desviación media del siguiente conjunto de datos:

x_i	$\lvert x_i - \bar{x} \rvert$
1	
2	
4	
6	
8	
9	

a) $D_m = 2.67$

1.4 Medidas de dispersión

Integrando. CURSOS ACADÉMICOS

La varianza σ^2 permite determinar el cuadrado de la diferencia entre un dato x_i y el promedio $\bar{x}$

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}$$

La desviación estándar es la raíz cuadrada de la varianza y cuenta con las mismas unidades que los datos originales (a diferencia de la varianza)

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}}$$

1.4 Medidas de dispersión

Si los datos están presentados en una tabla de frecuencias sin agrupar, es posible usar la siguiente fórmula

$$\sigma^2 = \frac{\sum_{i=1}^{N} f_i \cdot x_i^2}{N} - \bar{x}^2$$

Si los datos están agrupados, entonces

$$\sigma^2 = \frac{\sum_{i=1}^{N} f_i \cdot X_i^2}{N} - \bar{x}^2$$

donde X_i es la marca de clase respectiva

1.4 Ejemplos

Integrando. CURSOS ACADÉMICOS

3. Completa la tabla y halla la varianza de la siguiente distribución de frecuencias:

x_i	f_i	$f_i \cdot x_i$	$f_i \cdot x_i^2$
1	1		
2	2		
3	2		
4	5		
5	3		
	$\Sigma f_i =$	$\Sigma f_i \cdot x_i =$	$\Sigma f_i \cdot x_i^2 =$

a) $\sigma^2 = 1.47$

1.4 Ejemplos

4. Completa la tabla y determina la desviación estándar de la siguiente distribución de frecuencias:

x_i	f_i	X_i	$f_i \cdot X_i$	$f_i \cdot X_i^2$
$2-4$	2			
$5-7$	4			
$8-10$	5			
$11-13$	3			
$14-16$	2			
$17-19$	1			
	$\Sigma f_i =$		$\Sigma f_i \cdot X_i =$	$\Sigma f_i \cdot X_i^2 =$

a) $\sigma = 4.1$

1.5 Medidas de posición

Así como la mediana divide una distribución en dos partes iguales, es posible agrupar los datos en diferentes grupos a través de las medidas de posición.

a) Los percentiles dividen en 100 partes iguales los datos y se denotan por $P_1, P_2, \dots P_{99}$

b) Los deciles permiten dividir en 10 partes iguales la distribución, desde D_1 hasta D_9

c) Los cuartiles dividir los datos en 4 partes iguales y se escriben como Q_1, Q_2, Q_3

Previamente se deben ordenar los datos en forma creciente; las medidas de posición pueden ocupar el valor de uno de los datos, o bien, el punto intermedio entre dos.

1.5 Medidas de posición

Integrando.
CURSOS ACADÉMICOS

El primer cuartil es igual al percentil veinticinco; a la izquierda de este punto existe el 25% de los datos

$$Q_1 = P_{25}$$

La mediana, el cuartil dos, el decil cinco y el percentil cincuenta son el mismo valor; a la izquierda se ubica el 50% de la distribución

$$Me = P_{50} = D_5 = Q_2$$

El tercer cuartil ocupa el mismo lugar que el percentil setenta y cinco; el 75% de los datos se encuentra antes de esta posición

$$Q_3 = P_{75}$$

1.5 Medidas de posición

Integrando. CURSOS ACADÉMICOS

Para calcular la posición que ocupan los cuartiles:

$$P(Q_1) = \frac{n+1}{4}, \qquad P(Q_2) = \frac{2(n+1)}{4}, \qquad P(Q_3) = \frac{3(n+1)}{4}$$

Para los deciles

$$P(D_i) = \frac{i(n+1)}{10}$$

Y para los percentiles

$$P(P_i) = \frac{i(n+1)}{100}$$

1.5 Ejemplos

1. En un salón de 20 estudiantes se realizó un examen de matemáticas, y se obtuvieron los siguientes resultados; calcular Q_1, D_5, P_{75}

$$5, 5, 8, 7, 9, 10, 7, 6, 8, 7, 8, 9, 10, 10, 8, 7, 6, 5, 9, 6$$

a) $Q_1 = 6, D_5 = 7.5, P_{75} = 9$