



## Estadística y plataformas para el análisis de datos

---

Profesora: Katherine Barquero Mejías



## Clase 2: Fundamentos del análisis descriptivo

**¿ Qué es el análisis  
descriptivo?**

# Análisis descriptivo

## ❖ **Objetivo:**

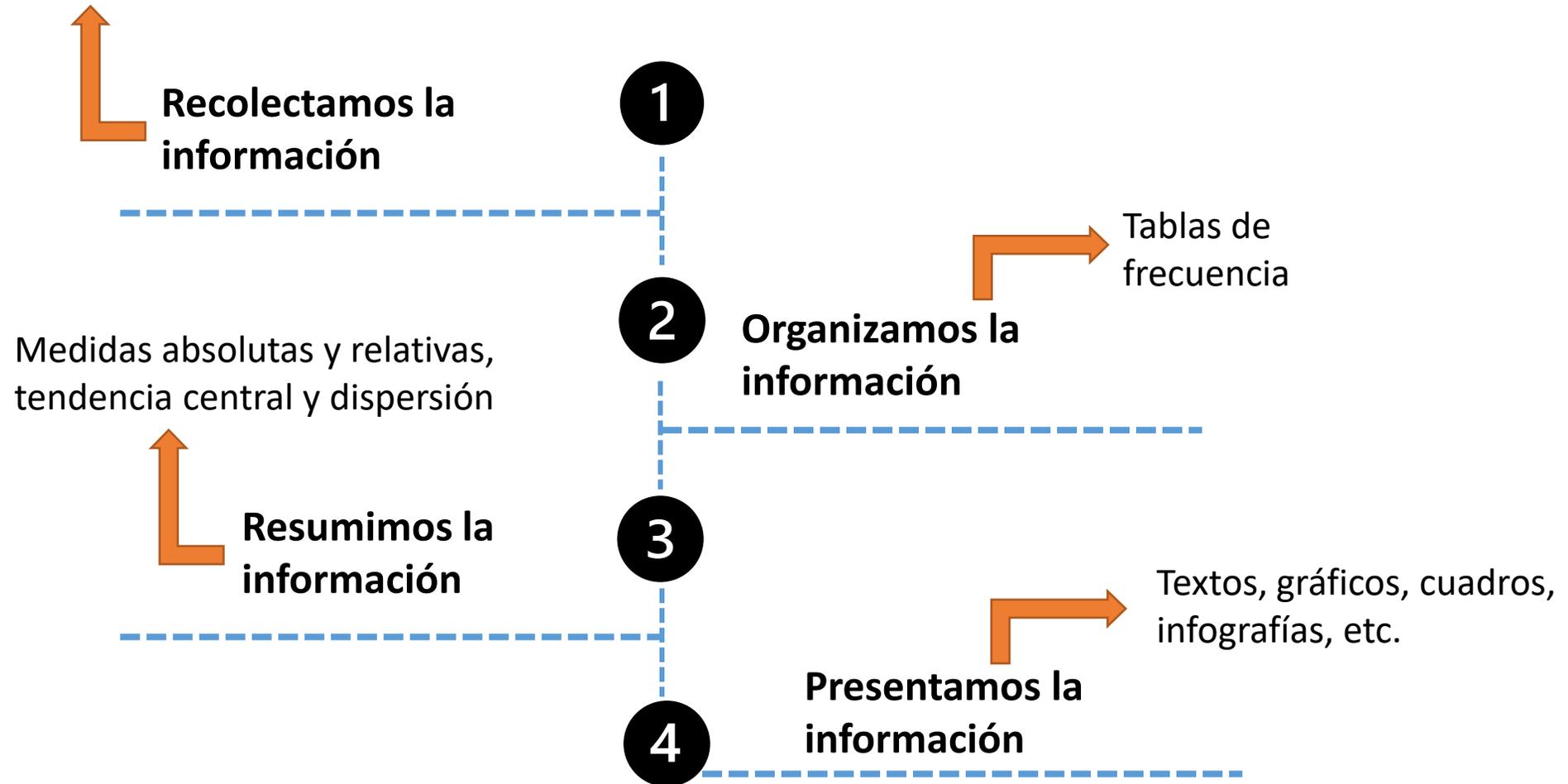
resumir y describir los datos mediante estadísticas descriptivas (tablas de frecuencias, promedios, desviaciones estándar, gráficos)

## ❖ **¿Cómo se resumen y describen los datos?**

A partir de estadísticos que resumen la información contenida en un conjunto de datos

# Proceso del análisis descriptivo

Registros administrativos,  
cuestionarios, sondeos,  
encuestas, etc.



# Técnicas para resumir datos: importancia

- Usualmente la información o datos disponibles para estudiar algún fenómeno de interés se encuentran desorganizados, en filas o registros, los cuales requieren de un procesamiento previo a su interpretación.

Edades de los  
docentes

25	30
26	25
29	46
30	53
34	53
35	55
36	55
38	56
40	56
40	57
41	62
41	62
43	64
44	64

Ventas en dólares de  
una inmobiliaria

1387	754	2230
1754	1621	2341
1817	870	3292
1040	1174	1108
1273	1412	1295
1529	1809	1344
3082	2415	1906
1951	1546	1952
2692	2148	2070
1342	2201	2454
1206	1298	1606
1342	1266	1680
443	1882	2338

# Tipos de Variables



# Tipos de variables



# Tipos de variables



Variables  
Cualitativas

- Se estudia una **característica no numérica**: género, provincia de nacimiento, estado civil, formación académica, etc.
- Comúnmente lo que interesa es el número o porcentaje de observaciones en cada categoría.
- Los datos se representan en gráficos de barras, entre otros.

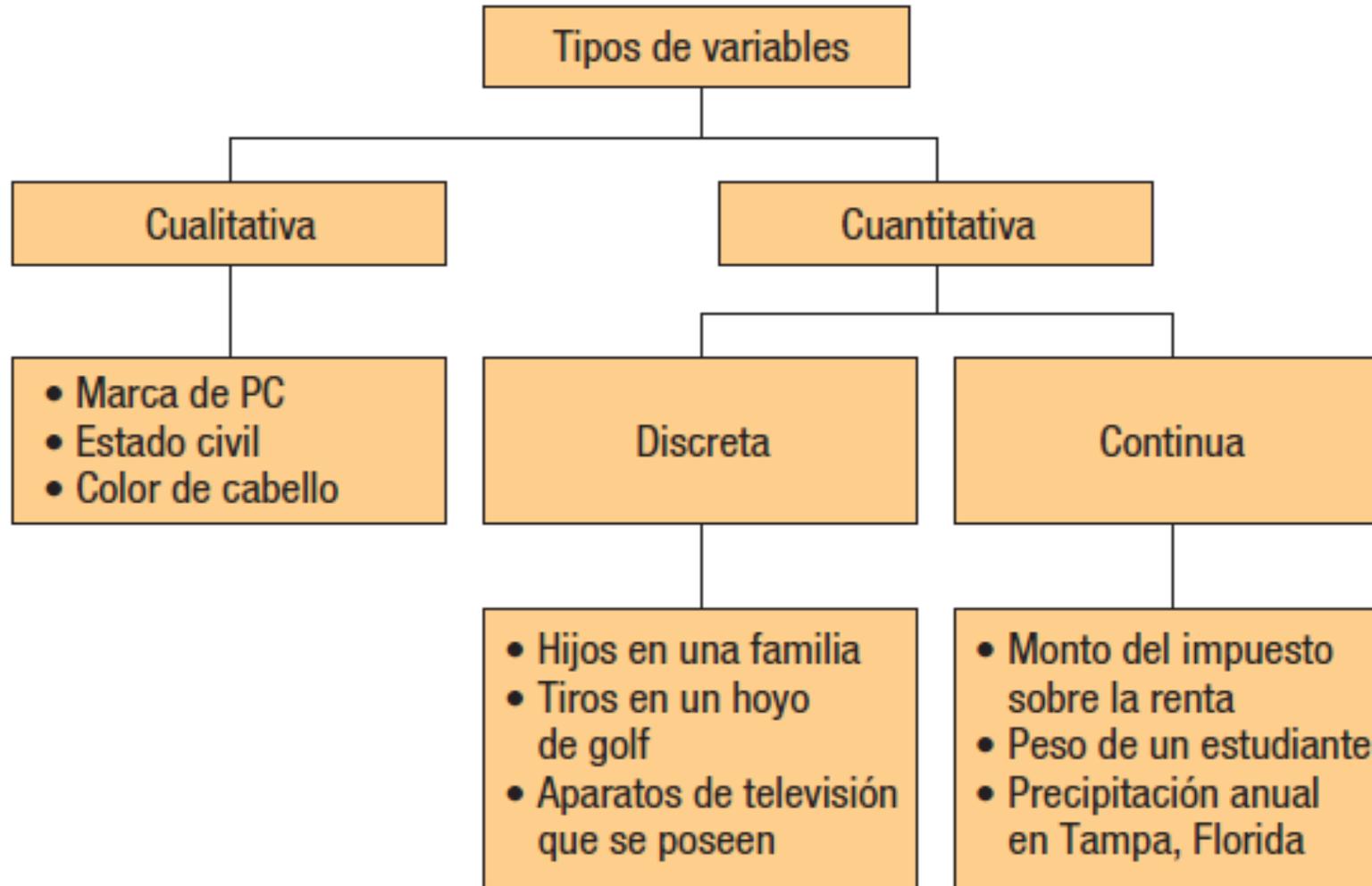
# Tipos de variables



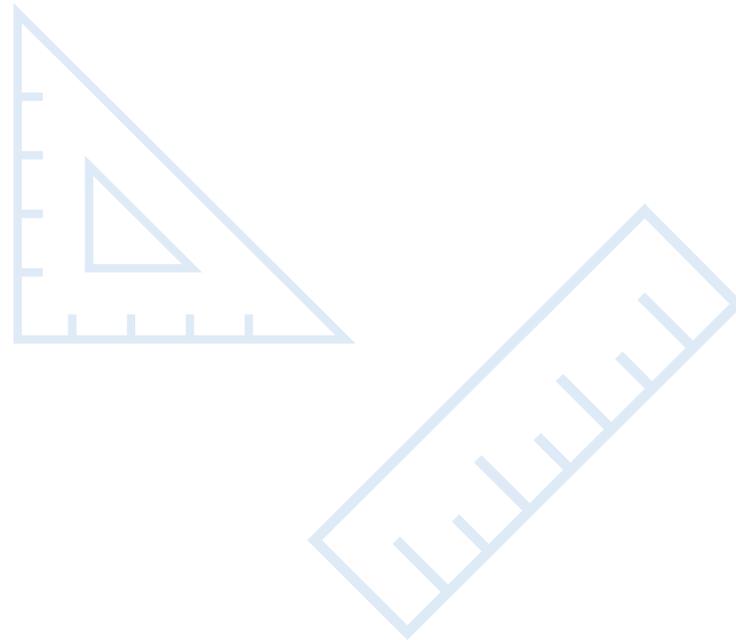
Variables  
Cuantitativas

- Variable bajo estudio aparece de forma numérica: edad, ingreso económico, cantidad de hijos, número de carros, etc.
- Se dividen en dos: Discreta y continuas:
  - **Discreta:** Adoptan solo ciertos valores.
  - **Continua:** Toman cualquier valor dentro de un intervalo específico

# Tipos de variables



# Niveles de medición



# Niveles de medición o escalas

## 1. Datos de nivel **nominal**:

- Observaciones de una **variable cualitativa** solo se clasifican y cuentan
- Variable de interés se divide en **categorías o resultados**
- No existe un orden natural de los resultados

## Ejemplo

### **Variable Beneficios de los trabajadores:**

Mejor atención médica, Más dinero, Mejor pensión, Balance entre trabajo/familia, otros

En términos de procesamientos de la información estas opciones se codifican de forma numérica por ejemplo:

- 1= Mejor atención médica
- 2= Más dinero
- 3= Mejor pensión
- 4= Balance entre trabajo/familia
- 5= Otros

# Niveles de medición o escalas

## Ejemplo

**Variable Sexo de los trabajadores:** es hombre o mujer

En términos de procesamientos de la información estas opciones se codifican de forma numérica por ejemplo:

1= Hombre  
2= Mujer

**Variable Estado civil de los trabajadores:**  
Divorciado, Soltero, Casado o en Unión libre

En términos de procesamientos de la información estas opciones se codifican de forma numérica por ejemplo:

1= Divorciado  
2= Soltero  
3= Casado  
4= Unión libre

# Niveles de medición o escalas

## 1. Datos a nivel **ordinal**:

- Clasificaciones de los datos se representan por conjunto de etiquetas o nombres con valores relativos
- Los valores relativos de los datos se pueden **clasificar u ordenar**

## Ejemplo

### Variable Satisfacción de los trabajadores:

Muy buena, Buena, Mala, Muy mala

En términos de procesamientos de la información estas opciones se codifican de forma numérica por ejemplo:

1= Muy mala

2= Mala

3= Buena

4= Muy buena

# Niveles de medición o escalas

## Ejemplo

**Variable Exposición al riesgo de los trabajadores:** Su nivel es bajo, medio o alto

En términos de procesamientos de la información estas opciones se codifican de forma numérica por ejemplo:

1= Bajo  
2= Medio  
3= Alto

**Variable Puntualidad de los trabajadores:**  
Nunca, casi nunca, casi siempre, siempre

En términos de procesamientos de la información estas opciones se codifican de forma numérica por ejemplo:

1= Nunca  
2= Casi nunca  
3= Casi siempre  
4= Siempre

# Niveles de medición o escalas

## 3. Datos a nivel de Numérico:

- Números enteros

**Variable:** Edad de los trabajadores

**Valor de 15 a 100**

- Decimales

**Variable:** Estatura de los trabajadores

**Intervalos (100cm-200cm)**

Descripción de  
datos: Resumir  
variables cualitativas



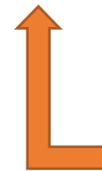
## Variables cualitativas

- Recordemos que las variables cualitativas corresponden a las **características no numéricas** de los individuos u objetos bajo estudio.
- Algunos **ejemplos**: género, lugar de nacimiento, grado académico, tipo de colegio, etc.

¿Cómo se resume la información de las variables cualitativas?

Para esto se utilizan **tablas de frecuencias**

Registros administrativos,  
cuestionarios, sondeos,  
encuestas, etc.



**Recolectamos la  
información**

**1**



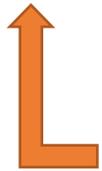
**2**

**Organizamos la  
información**



Tablas de  
frecuencia

Medidas absolutas, relativas y  
de tendencia central



**Resumimos la  
información**

**3**



**4**

**Presentamos la  
información**

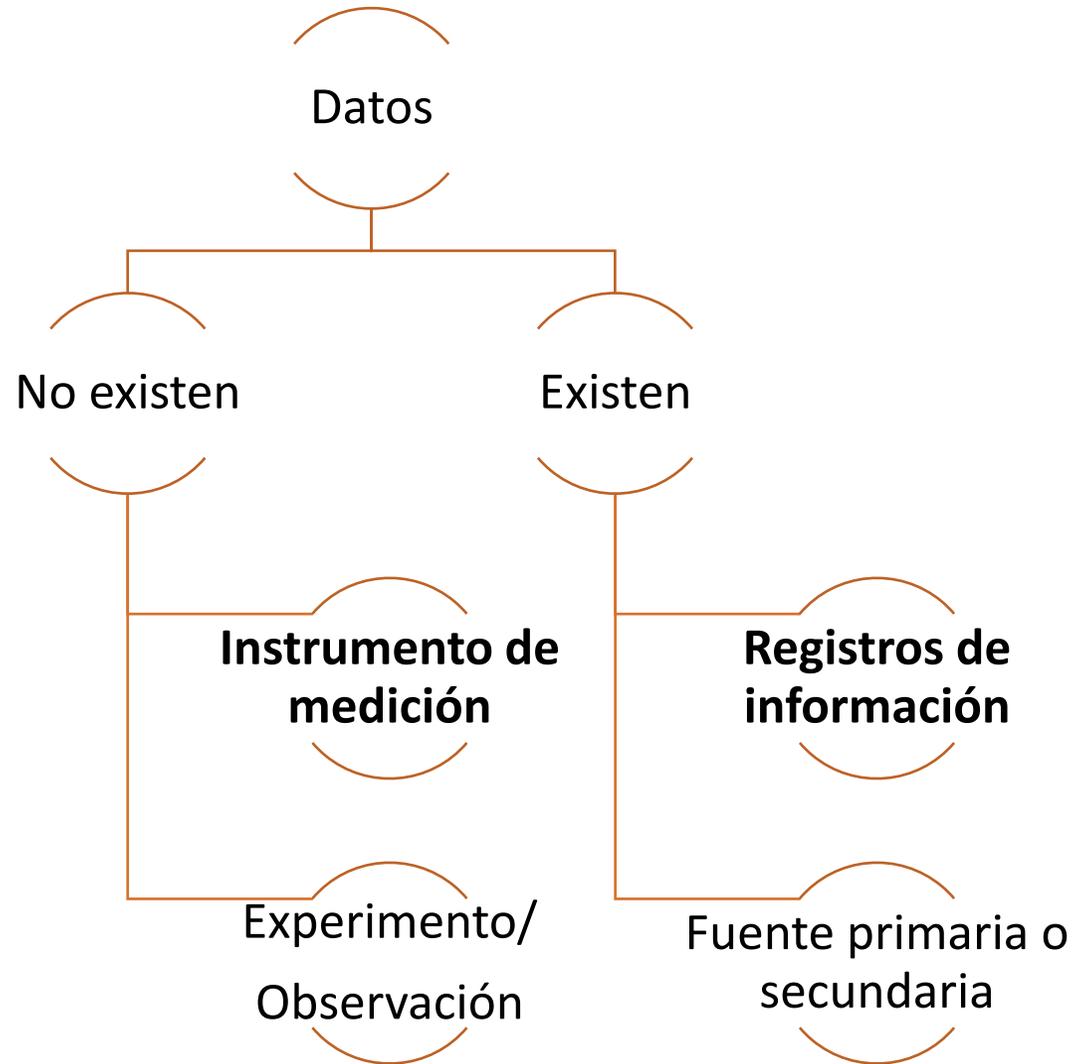


Textos, gráficos, cuadros,  
infografías, etc.



# 1

## Recolectamos la información



## 2 Organizamos la información

- Para esto empleamos **tablas de frecuencia**
- **Definición** de tabla de frecuencias:  
“Agrupación de datos cualitativos en clases mutuamente excluyentes que presenta el número de observaciones dentro de cada categoría”
- Pero, ¿cómo construimos una tabla de frecuencias?
  1. Analizamos la información e identificamos la variable de estudio
  2. Identificamos las categorías de análisis
  3. Determinamos el nivel de medición de la variable (nominal u ordinal)
  4. Agrupamos la información

## **Ejemplo 1:**

Una empresa de autos desea resumir las ventas de autos del mes anterior que se vendieron en cada una de las provincias de Costa Rica donde se cuentan con alguna sucursal. Se conoce que en Limón se vendieron 52 autos, en Guanacaste 40, en Heredia 45 y en Alajuela 43. Construya una tabla de frecuencias que permita agrupar esta información.

**1. Variable de estudio:** Ventas de autos en cada provincia de Costa Rica

**2. Categorías de Análisis:**

1. Limón
2. Guanacaste
3. Heredia
4. Alajuela

### 3. Nivel de medición: Nominal

### 4. Agrupación de los datos: Variable de análisis, frecuencia absoluta y frecuencia relativa

Tabla 1. Venta de autos en cada provincia de Costa Rica.

Provincia	Frecuencia absoluta: cantidad de autos	Frecuencia relativa:
Limón	52	0,28
Guanacaste	40	0,22
Heredia	45	0,25
Alajuela	43	0,23
<b>Total</b>	<b>180</b>	<b>1,00</b>

## **Ejemplo 2:**

El viceministerio académico del MEP giró una directriz en la que le solicitó a los directores de las escuelas públicas del país identificar la cantidad de estudiantes con dificultades en el acceso a recursos tecnológicos en los hogares y que por tanto enfrentan mayores problemas para acceder a las clases a distancia producto de la pandemia. Con esta información el MEP seleccionará las escuelas con mayor cantidad de estudiantes que carecen de estos recursos para construir laboratorios de informática educativa.

La Escuela A indagó con cinco profesores y encontró que del total de 1000 estudiantes matriculados, 300 de ellos no tenían conexión a internet, 325 no tenían computadora y 295 no poseen Tablet o celular. Con esta información, elabore una tabla de frecuencias que permita organizar la información que la Escuela A debe reportar al MEP para tomar una decisión adecuada.

**1. Variable de estudio:** Acceso a recursos tecnológicos

**2. Categorías de Análisis:**

1. Sin conexión a internet
2. Sin computadora
3. Sin Tablet o celular

## Ejemplo:

### 3. Nivel de medición: Nominal

### 4. Agrupación de los datos: Variable de análisis, frecuencia absoluta y frecuencia relativa

Tabla 1. Cantidad de estudiantes en la Escuela A con problemas de acceso a recursos tecnológicos.

Problema de acceso a recursos tecnológicos	Frecuencia absoluta: cantidad de estudiantes	Frecuencia relativa: porcentajes
Sin conexión a internet	300	33%
Sin computadora	325	35%
Sin Tablet o celular	295	32%
<b>Total</b>	<b>920</b>	<b>100%</b>

### 3 Resumir la información

- En el caso de las variables cualitativas en la práctica lo más común es utilizar los porcentajes agrupados en las distribuciones de frecuencia para describir la información:
  1. En el primer ejemplo se conoce que la provincia con mayor cantidad de ventas fue Limón quien registró el 35% del total de ventas reportadas en el mes anterior. Aunque todas las provincias vendieron un porcentaje similar de carros.
  2. En el segundo ejemplo el principal problema de acceso a recursos tecnológicos que enfrentan los estudiantes es que no cuentan con computadora, el 35% indicó que no tenían este dispositivo en sus hogares.
- También se pueden utilizar datos absolutos, pero son más complejos para realizar comparaciones. Por ejemplo, en una escuela 300 estudiantes con problemas de acceso a recursos tecnológicos puede ser muy poco si la matrícula es de 2000 estudiantes, pero para una escuela con 600 estudiantes matriculados en total este valor representaría prácticamente la mitad de ellos

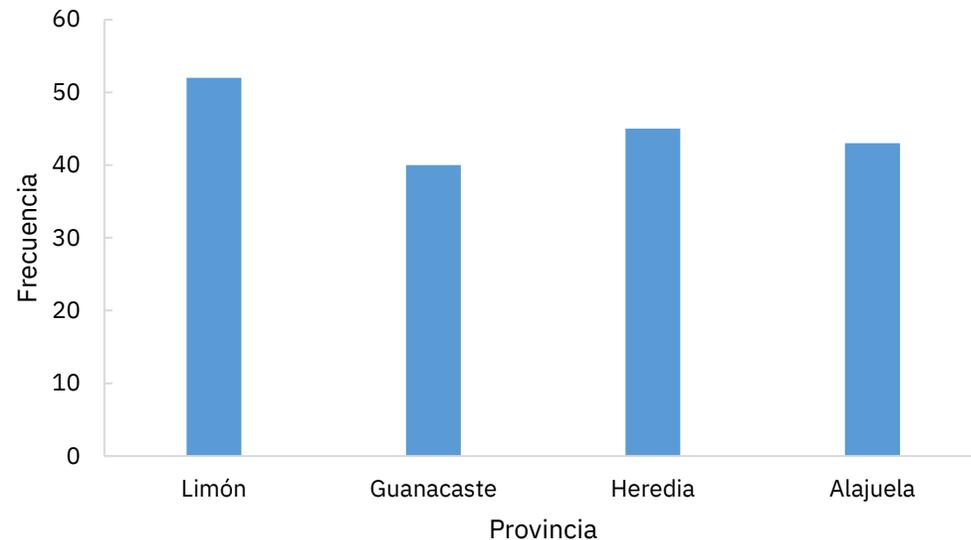
## 2 Presentamos la información

- Para presentar la información podemos utilizar varias formas: presentar un texto escrito, infografías o gráficos. Usualmente se combinan: Por ejemplo texto y gráficos
- En las variables cualitativas, lo más común es utilizar gráficos de barras (para presentar datos absolutos) y gráficos circulares para presentar datos relativos

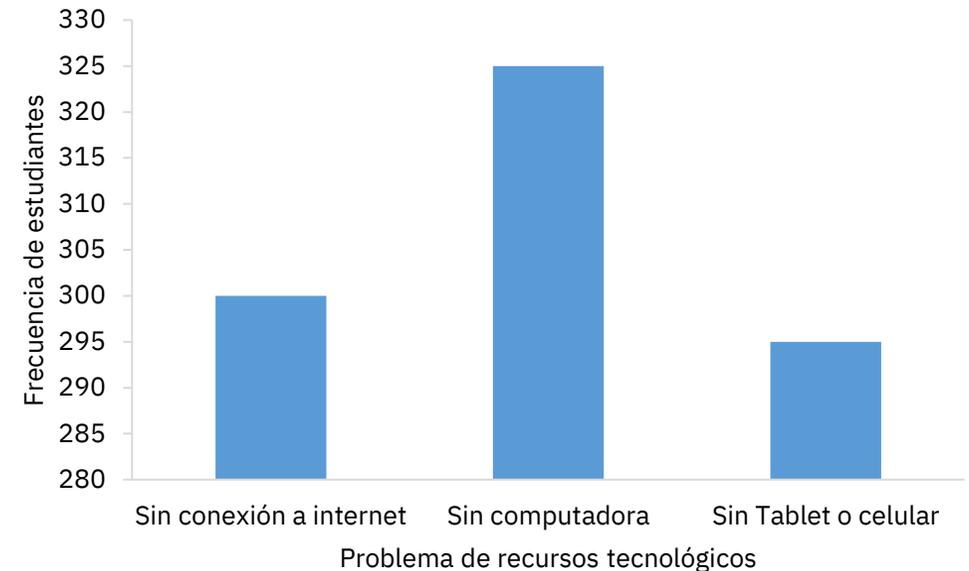
## 2 Presentamos la información

- **Gráfico de barras:** Las categorías o clases cualitativas se representan en el eje horizontal y la frecuencia en el eje vertical. Las barras son proporcionales a la altura de las barras

Ejemplo 1



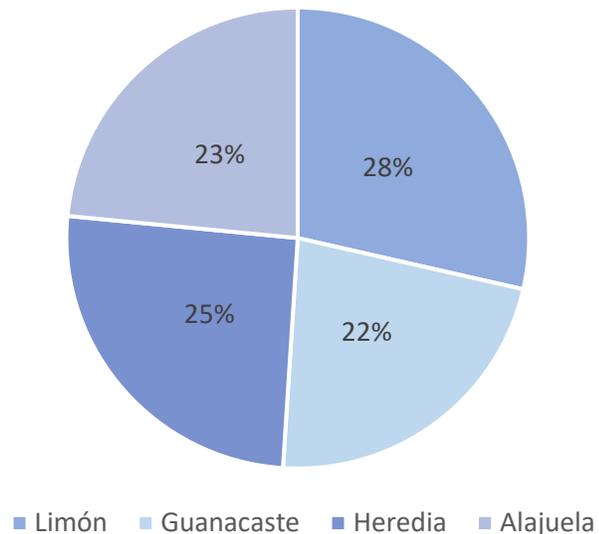
Ejemplo 2



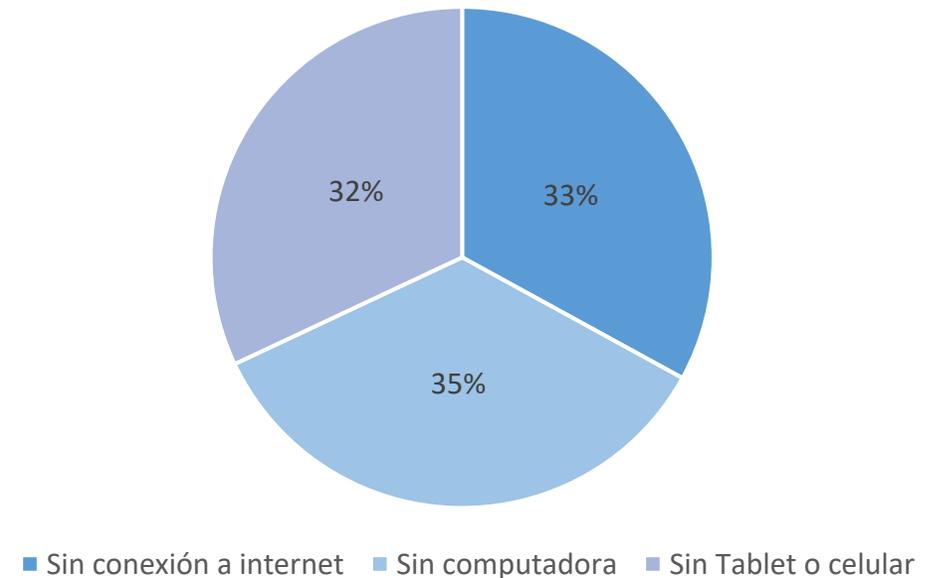
## 2 Presentamos la información

- **Gráfico circular o pastel:** Muestra la parte o el porcentaje que representa cada clase del total de números de frecuencia

Ejemplo 1



Ejemplo 2

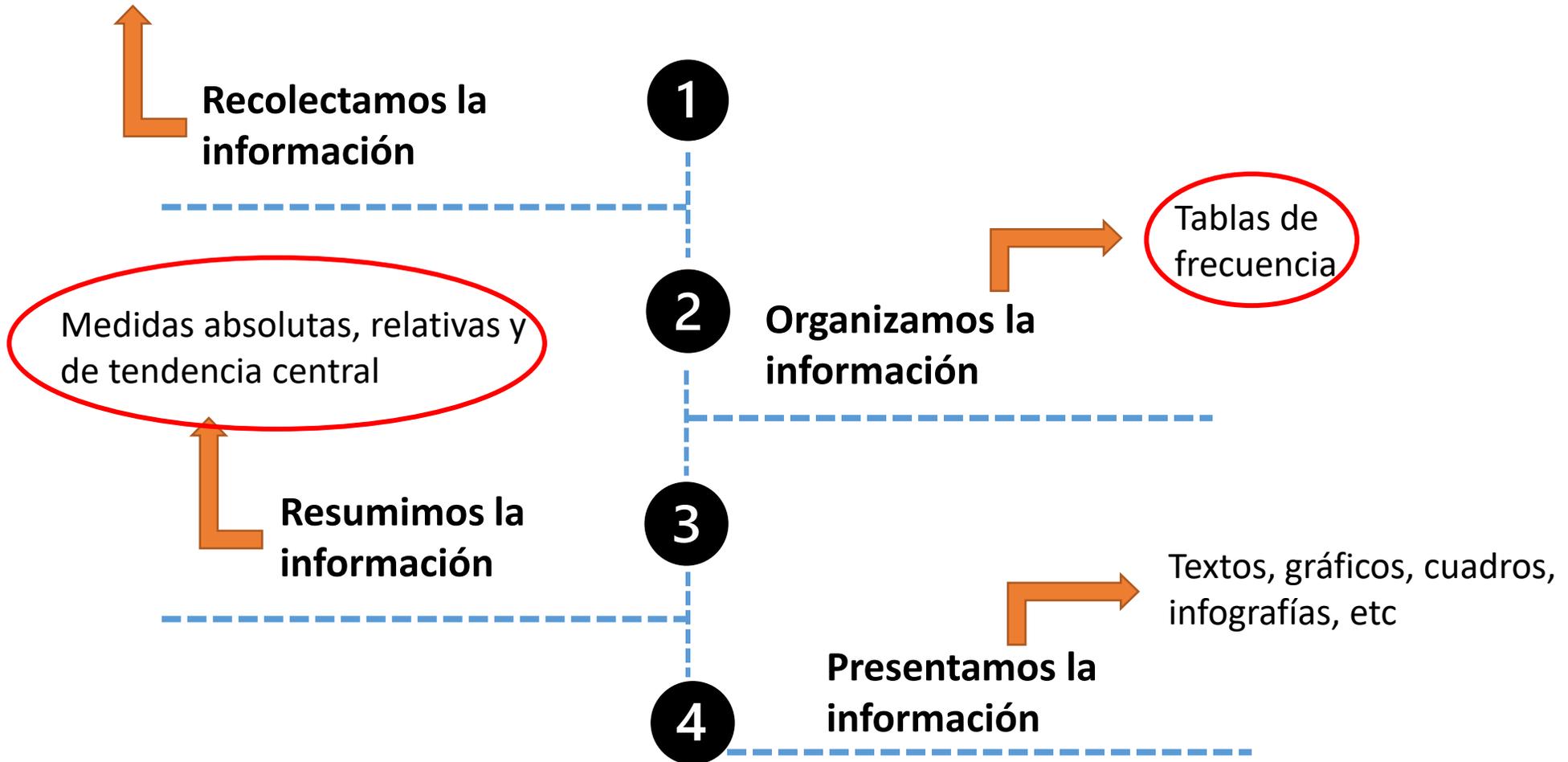




Descripción de  
datos: variables  
cuantitativas

# Procesamiento descriptivo de variables cuantitativas

Registros administrativos, cuestionarios, sondeos, encuestas, etc



## 2 Organizamos la información

- Para esto empleamos **tablas de frecuencia**

- **Definición** de tabla de frecuencias:

“Agrupación de datos cualitativos en clases **mutuamente excluyentes** que presenta el número de observaciones dentro de cada categoría”

- Pero, ¿cómo construimos una tabla de frecuencias?

1. Analizamos la información e identificamos la variable de estudio
2. **Identificamos y construimos las clases de análisis**
3. Agrupamos la información

- En las variables continuas usualmente se tienen listados o registros de datos que **no están agrupados**

- **Ejemplo 3:**

El MEP solicitó a 6 colegios de la Dirección Regional de Alajuela registrar la edad de los docentes para analizar cuántos se podrían pensionar y estimar las posibles plazas vacantes que se tendrían en los próximos años.

Edades de los  
docentes



Docente 1	35
Docente 2	40
Docente 3	25
Docente 4	64
Docente 5	53
Docente 6	56
Docente 7	62
Docente 8	36
Docente 9	41
Docente 10	25
Docente 11	56
Docente 12	43
Docente 13	38
Docente 14	55

Docente 15	26
Docente 16	64
Docente 17	53
Docente 18	40
Docente 19	30
Docente 20	34
Docente 21	46
Docente 22	55
Docente 23	29
Docente 24	57
Docente 25	62
Docente 26	41
Docente 27	44

# Identificar y construir las clases de análisis

- A diferencia de las variables cualitativas, en las que las clases o categorías están definidas, en las variables continuas debemos crearlas
- Para crear las clases debemos realizar los siguientes pasos:
  1. Estimar la cantidad de clases (categorías)
  2. Estimar el ancho de cada intervalo de las clases
  3. Establecer los límites de las clases
  4. Realizar el conteo o agrupación de los datos en las clases definidas

## Paso 1: Estimar la cantidad de clases (categorías)

En nuestro ejemplo tenemos **27 observaciones**

En este ejemplo:

- $n = 27$
- $k = ?$

¿Cuántas clases recomendaría?

**Opción 1:  $2^4 = 16$**

**Opción 2:  $2^5 = 32$**

**Opción 1:  $K=4$   $16 < 27$**

**Opción 2:  $K=5$   $32 > 27$  ✓**

Docente 1	35
Docente 2	40
Docente 3	25
Docente 4	64
Docente 5	53
Docente 6	56
Docente 7	62
Docente 8	36
Docente 9	41
Docente 10	25
Docente 11	56
Docente 12	43
Docente 13	38
Docente 14	55
Docente 15	26
Docente 16	64
Docente 17	53
Docente 18	40
Docente 19	30
Docente 20	34
Docente 21	46
Docente 22	55
Docente 23	29
Docente 24	57
Docente 25	62
Docente 26	41
Docente 27	44

Encontramos que **debemos crear 5 clases**

## Paso 2: Determinar el ancho del intervalo de la clase

$$i = \frac{H-L}{k}$$

- $i$ : ancho del intervalo
- $H$ : valor máximo de los datos
- $L$ : valor mínimo de los datos

- Para aplicar la fórmula primero ordenamos de menor a mayor nuestros datos
- Identificamos el valor máximo de los datos (H): 64
- Identificamos el valor mínimo de los datos (L): 25
- Aplicamos la fórmula

$$i = \frac{64 - 25}{5} = 7,8 = 8$$



Docente 3	25
Docente 10	25
Docente 15	26
Docente 23	29
Docente 19	30
Docente 20	34
Docente 1	35
Docente 8	36
Docente 13	38
Docente 2	40
Docente 18	40
Docente 9	41
Docente 26	41
Docente 12	43
Docente 27	44
Docente 21	46
Docente 5	53
Docente 17	53
Docente 14	55
Docente 22	55
Docente 6	56
Docente 11	56
Docente 24	57
Docente 7	62
Docente 25	62
Docente 4	64
Docente 16	64

L: Valor mínimo

H: Valor máximo

## Recapitulando...

Hasta el momento tenemos:

- 5 clases o categorías
- Con un ancho de intervalo de 8 unidades

## Paso 3: Establecer los límites de cada clase

- Iniciamos con el valor mínimo de los datos

Clases	Ancho del intervalo
25-33	8
34-42	8
43-51	8
52-60	8
61-69	8

- Evitar la sobreposición de los límites

25-33

33-42

Docente 3	25
Docente 10	25
Docente 15	26
Docente 23	29
Docente 19	30
Docente 20	34
Docente 1	35
Docente 8	36
Docente 13	38
Docente 2	40
Docente 18	40
Docente 9	41
Docente 26	41
Docente 12	43
Docente 27	44
Docente 21	46
Docente 5	53
Docente 17	53
Docente 14	55
Docente 22	55
Docente 6	56
Docente 11	56
Docente 24	57
Docente 7	62
Docente 25	62
Docente 4	64
Docente 16	64

L: Valor mínimo

H: Valor máximo

# Identificar y construir las clases de análisis

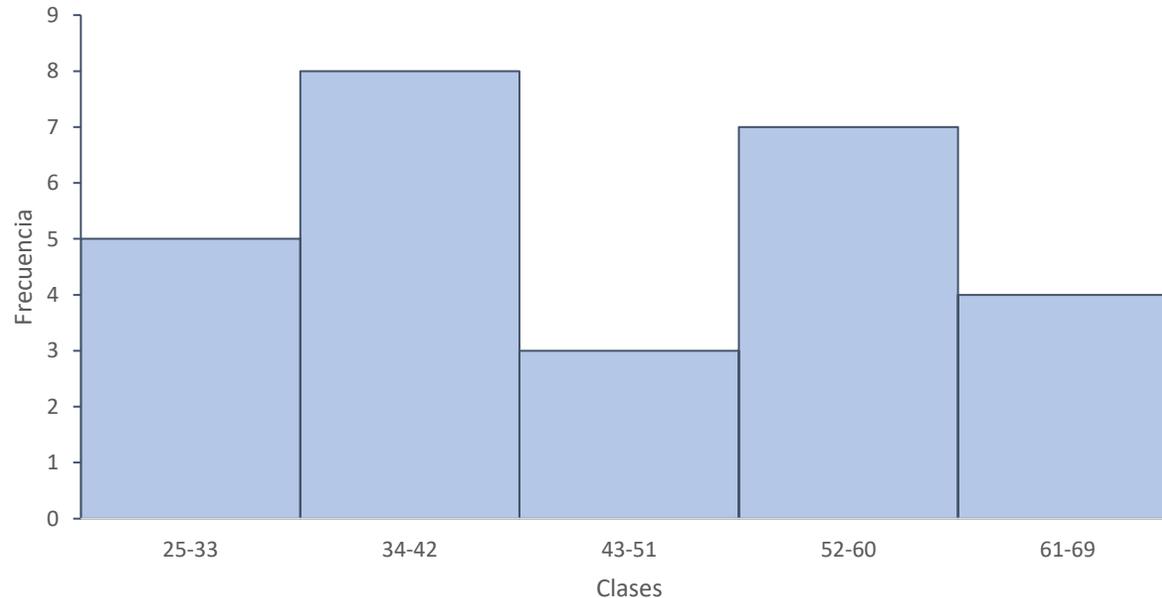
## Paso 4: Realizar el conteo y agrupación de los datos

Clases	Cantidad de docentes	Frecuencia relativa
25-33	5	19%
34-42	8	30%
43-51	3	11%
52-60	7	26%
61-69	4	15%
<b>Total</b>	<b>27</b>	<b>100%</b>

Docente 3	25
Docente 10	25
Docente 15	26
Docente 23	29
Docente 19	30
Docente 20	34
Docente 1	35
Docente 8	36
Docente 13	38
Docente 2	40
Docente 18	40
Docente 9	41
Docente 26	41
Docente 12	43
Docente 27	44
Docente 21	46
Docente 5	53
Docente 17	53
Docente 14	55
Docente 22	55
Docente 6	56
Docente 11	56
Docente 24	57
Docente 7	62
Docente 25	62
Docente 4	64
Docente 16	64

## 2 Presentamos la información

- **Histograma:** Gráfico donde las clases se señalan en el eje horizontal y las frecuencias de la clase en el eje vertical. Las frecuencias se representan por medio de las alturas de las barras que se dibujan de manera adyacente



### 3 Resumir la información

- Cuando estamos ante variables continuas podemos resumir la información a través de dos formas:
  1. AGRUPANDO LOS DATOS EN DISTRIBUCIONES DE FRECUENCIAS
  2. UTILIZANDO **MEDIDAS DE UBICACIÓN EN DATOS NO AGRUPADOS**
- El primer punto lo abordamos anteriormente
- Ahora procederemos a estudiar las **medidas de ubicación o tendencia central**
- No perdamos de vista que el interés es siempre **resumir la información**

# Medidas de tendencia central

- Utilidad: sirven para describir datos cuantitativos
- **Propósito:** Señalar el centro de un conjunto de valores o el punto central de la distribución.
- Medidas posición / ubicación : a menudo llamadas **promedios**
- No se pueden resumir los datos tomando en cuenta únicamente las medidas de posición, se debe completar con **las de variabilidad o dispersión**

# Medidas de posición

- El centro de una distribución puede ser definida de distintas formas por lo tanto existen diferentes medidas de posición central

1. Moda

2. Media aritmética (promedio)

3. Mediana

# Moda

- Es el **valor más frecuente**, el que más se repite
- Indica el número donde la variable ocurre con mayor densidad
- Se obtiene la frecuencia de valores y se identifica el que exhibe la mayor frecuencia
- **Ventaja:** No se ve afectado por valores extremos

14, 15, 17, 17, 21, 21, 21, 21, 33, 36, 40

**Nota: los datos tienen que estar ordenados para facilitar su cálculo**

## Ejemplo:

A través de una encuesta realizada en setiembre del 2009 a una muestra de 350 adultos en CR se preguntó cuál era del centro comercial que más les gustaba visitar. La distribución de las respuestas se muestra a continuación:

<b>Centro comercial</b>	<b>Porcentaje de mención</b>
Multiplaza de Escazú	25,7%
Mall San Pedro	23,4%
Paseo de las Flores	15,2%
Terramall	14,9%
Real Cariari	5,6%
Multiplaza del Este	5,0%
Mall Internacional	3,2%
Otro	7,0%
<b>Total</b>	<b>100,0%</b>

**¿Cuál medida de posición se implementó para resumir las frecuencias?**

# Media de una muestra (Media aritmética)

- Suma de todos los valores observados en la población dividido entre el número de valores de la población

$$\text{Media poblacional} = \frac{\text{Suma de todos los valores observados en la población}}{\text{Número de valores en la población}}$$

$$\bar{X} = \frac{1}{n} \sum X_i = \frac{\sum_{i=1}^n X_i}{n}$$

20, 20, 22, 20, 30, 25, 25, 18, 20, 18, 22, 36

$$\bar{X} = \frac{20+20+22+20+30+25+25+18+20+18+22+36}{12} = 23 \text{ años}$$

# Media de una muestra (Media aritmética)

- Esta medida se ve muy afectada cuando hay presencia de valores extremos
- Mismo caso pero con **valores extremos**

20, 20, 22, 20, 30, 25, 25, 18, 20, 18, 22, 36, **90**

$$\bar{X} = \frac{20+20+22+20+30+25+25+18+20+18+22+36+\mathbf{90}}{13} = 28 \text{ años}$$

# Mediana

- Valor central de una serie de **datos ordenados**: valor en el que no más de la mitad de las observaciones son menores que ella y no más de la mitad son mayores
- En datos sin agrupar se deben ordenar primero de acuerdo con su magnitud
- Luego se determina el valor central de la serie  $(n+1)/2$ 
  - **Si  $n$  es par, existirán dos valores centrales. Se obtiene el promedio de ellos**

**6, 8, 8, 10, 12, 19, 23**

# Mediana

- Medida de posición
- Valor central de una serie de **datos ordenados**: valor en el que no más de la mitad de las observaciones son menores que ella y no más de la mitad son mayores
- En datos sin agrupar se deben ordenar primero de acuerdo con su magnitud
- Luego se determina el valor central de la serie  $(n+1)/2$ 
  - **Si n es par, existirán dos valores centrales. Se obtiene el promedio de ellos**

**6, 8, 8, 10, 12, 19, 23**

Posición de la mediana: ?

Valor de la mediana = 10 (4º término)

# Mediana

- Si  $n$  es par, existirán dos valores centrales. Se obtiene el promedio de ellos

**3,4,4,5,16,19,25,30**

Posición de la mediana: ?

$$\frac{n+1}{2} = \frac{8+1}{2} = 4,5$$

**Valor de la mediana = 10,5 (4,5° término)**

**Promedio entre 5 y 16 = 10,5**

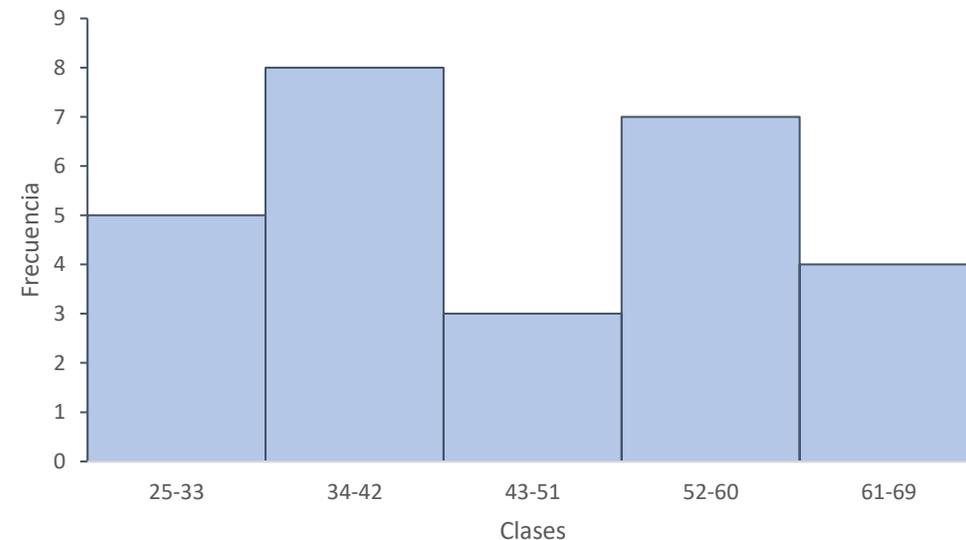
### 3 Presentamos la información

- Utilizamos la información que proporciona la **distribución de frecuencias** y las **medidas de posición** para caracterizar los datos
- Complementamos el análisis con el histograma presentado

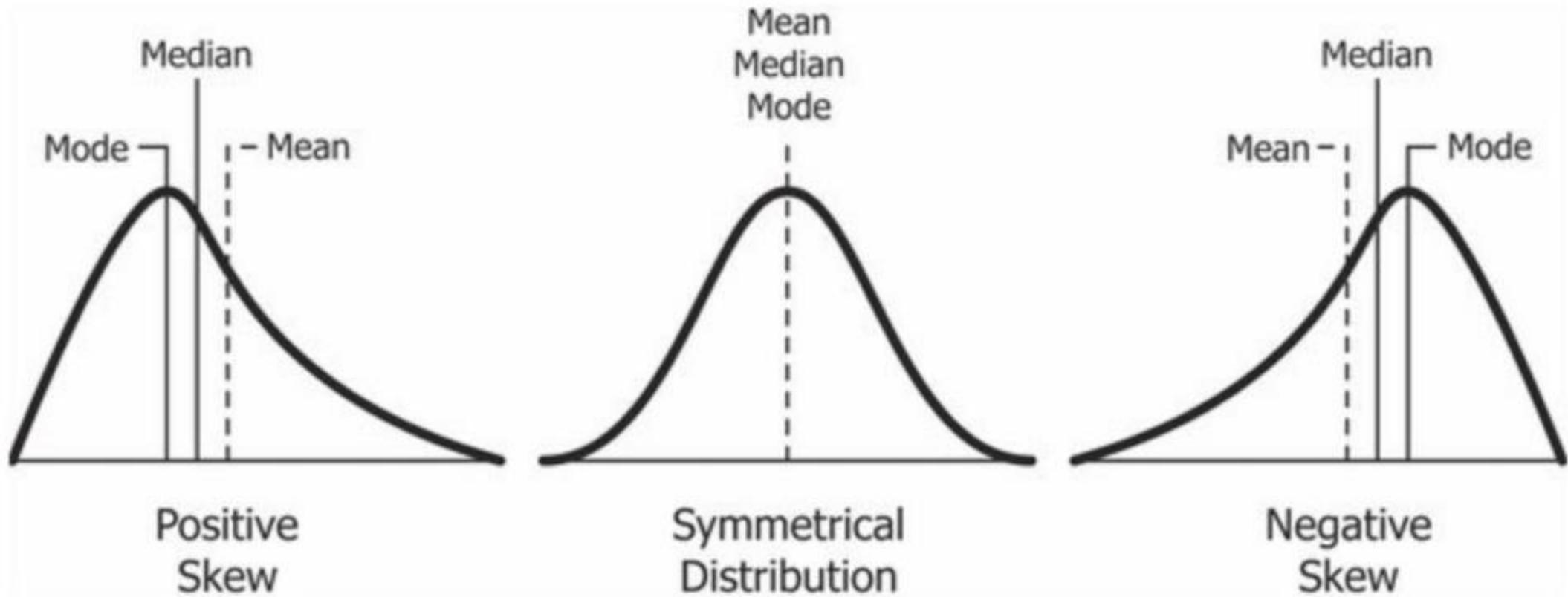
La Dirección Regional de Alajuela informa al MEP que el 41% de los docentes tienen edades mayores a 52 años. El 26% del profesorado tiene edades entre los 52 y 60 años. Mientras que un 49% tiene edades inferiores a los 42 años. Esta distribución se presenta con mayor detalle en el gráfico 1.

La **edad promedio** de los docentes es de **27 años**, mientras que las edades **de 62 y 64 años** son las **más frecuentes**. Además, se encontró que el 50% de los datos son edades inferiores a los **43 años** y el otro 50% son edades superiores a este valor.

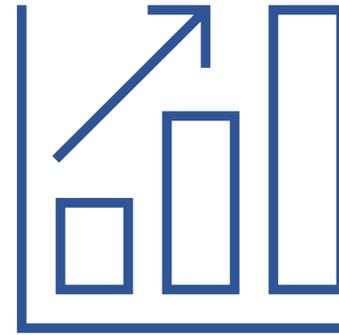
Gráfico 1: Distribución de los docentes de la Dirección Regional de Alajuela, según el grupo de edad



# Interpretando el histograma



# Descripción de datos: Medidas de dispersión



1

2

3

# Medidas de dispersión

Analizar y describir un conjunto de datos implica al menos tres objetivos:

1. Determinar cómo se distribuyen los datos
2. Identificar los valores centrales
3. Establecer la medida en la que los datos se agrupan o dispersan alrededor de la media

Medidas de posición solo abordan un aspecto de los datos: **su ubicación central**, pero no indica si los datos se agrupan o dispersan alrededor de los valores centrales

Por esto, deben calcularse **medidas de dispersión para completar el panorama descriptivo de la distribución**

## Medidas de dispersión

Es tan importante conocer el promedio como la variabilidad de los datos a su alrededor, la validez de un promedio para resumir o representar el conjunto de datos depende de si los datos se dispersan o se concentran a su alrededor.

Cuanto más concentrados estén los datos alrededor del promedio aritmético, más confianza se tendrá en este valor para caracterizarlos.

El interés por la variabilidad de los datos se origina en la certeza de que los datos constantemente varían, pero tienen una tendencia natural a agruparse alrededor de los valores centrales.

**Entre más concentrados estén los datos alrededor de ellos más representativos son del conjunto y más confiables las conclusiones que se logren usando valores centrales o promedios**

# ¿Cómo se mide la variabilidad?

Disponemos de las siguientes opciones



# El rango

- También es conocido como recorrido o amplitud
- Se origina al considerar los valores extremos del grupo de datos: es decir la **amplitud de los datos**
- El rango es la diferencia entre el valor mayor y el menor de un conjunto de datos

**Datos: 3,10,2,8,7**

**Recorrido:  $10-2=8$**

- A pesar de su simplicidad y facilidad de calculo presenta limitaciones que hace que se use poco:
  - No considera todas las observaciones del grupo
  - Depende sensiblemente del número de datos, puede crecer fácilmente
- En la práctica se usa cuando no se pueden estimar métodos complejos (por tiempo y personal) y se tienen muestras pequeñas de datos

- Pasos prácticos para calcular el rango
  - ✓ **Paso 1:** Ordene los datos
  - ✓ **Paso 2:** Identifique el valor mínimo de los datos
  - ✓ **Paso 3:** Identifique el valor máximo de los datos
  - ✓ **Paso 4:** Reste el valor máximo y el valor mínimo para obtener la amplitud de los datos.
  - ✓ **Paso 5:** Interprete los resultados

Apliquemos estos pasos en el siguiente ejercicio

## Ejercicio:

El MEP solicitó a 6 colegios de la Dirección Regional de Alajuela registrar la edad de los docentes para analizar cuántos se podrían pensionar y estimar las posibles plazas vacantes que se tendrían en los próximos años. **Estime el rango o amplitud de los datos.**

Edades de los  
docentes



Docente 1	35
Docente 2	40
Docente 3	25
Docente 4	64
Docente 5	53
Docente 6	56
Docente 7	62
Docente 8	36
Docente 9	41
Docente 10	25
Docente 11	56
Docente 12	43
Docente 13	38
Docente 14	55

Docente 15	26
Docente 16	64
Docente 17	53
Docente 18	40
Docente 19	30
Docente 20	34
Docente 21	46
Docente 22	55
Docente 23	29
Docente 24	57
Docente 25	62
Docente 26	41
Docente 27	44

- Aplicación de los pasos:
  - ✓ **Paso 1:** Ordenar los datos

Docente 3	25
Docente 10	25
Docente 15	26
Docente 23	29
Docente 19	30
Docente 20	34
Docente 1	35
Docente 8	36
Docente 13	38
Docente 2	40
Docente 18	40
Docente 9	41
Docente 26	41
Docente 12	43
Docente 27	44
Docente 21	46
Docente 5	53
Docente 17	53
Docente 14	55
Docente 22	55
Docente 6	56
Docente 11	56
Docente 24	57
Docente 7	62
Docente 25	62
Docente 4	64
Docente 16	64

- Aplicación de los pasos:
- ✓ **Paso 2:** Identifique el valor mínimo

Docente 3	25
Docente 10	25
Docente 15	26
Docente 23	29
Docente 19	30
Docente 20	34
Docente 1	35
Docente 8	36
Docente 13	38
Docente 2	40
Docente 18	40
Docente 9	41
Docente 26	41
Docente 12	43
Docente 27	44
Docente 21	46
Docente 5	53
Docente 17	53
Docente 14	55
Docente 22	55
Docente 6	56
Docente 11	56
Docente 24	57
Docente 7	62
Docente 25	62
Docente 4	64
Docente 16	64

- Aplicación de los pasos:
- ✓ **Paso 3:** Identifique el valor máximo

Docente 3	25
Docente 10	25
Docente 15	26
Docente 23	29
Docente 19	30
Docente 20	34
Docente 1	35
Docente 8	36
Docente 13	38
Docente 2	40
Docente 18	40
Docente 9	41
Docente 26	41
Docente 12	43
Docente 27	44
Docente 21	46
Docente 5	53
Docente 17	53
Docente 14	55
Docente 22	55
Docente 6	56
Docente 11	56
Docente 24	57
Docente 7	62
Docente 25	62
Docente 4	64
Docente 16	64

- Aplicación de los pasos:

- ✓ **Paso 4:** Reste el valor máximo y el valor mínimo para obtener la amplitud de los datos.

$$\text{Rango} = \text{Valor máximo} - \text{Valor mínimo}$$

$$\text{Rango} = 64 - 25$$

$$\text{Rango} = 39$$

- ✓ **Paso 5:** Interprete los resultados

La amplitud o recorrido de los datos es de 39 años. Las edades de los docentes se encuentran entre los 25 y 54 años de edad

# Desviación estándar

- Surge de la necesidad de contar con una medida de dispersión que considere todos los datos y no esté estrechamente ligada al número de ellos
- Por tanto, debe basarse en las desviaciones o diferencias de los datos individuales respecto a un valor central o típico. Es decir, considerar la suma de las desviaciones de los datos con respecto a su media aritmética

$$Desviaciones = \sum Xi - \bar{X}$$

- $Xi$ : el valor de los datos
  - $\bar{X}$ : promedio o media aritmética
- Indica cuánto se alejan, en promedio, las observaciones de la media aritmética del conjunto.

# Desviación estándar

- El cálculo de la desviación estándar utiliza los **cuadrados** de las desviaciones

$$Desviación\ estándar = \sqrt{\frac{\sum(Desviaciones)^2}{número\ de\ datos}}$$

- Indica cuánto se alejan, en promedio, las observaciones de la media aritmética del conjunto.
- Es la medida de dispersión más utilizada en la estadística
- Entre más cercano a cero sea el valor obtenido, menor es la variabilidad de los datos

# Desviación estándar

- Pasos prácticos para calcular la desviación estándar
  - ✓ **Paso 1:** Escriba en una columna los datos de los que dispone
  - ✓ **Paso 2:** Cálculo el promedio o media aritmética de los datos y la cantidad de los datos
  - ✓ **Paso 3:** Estime las desviaciones medias y cree una columna con los valores calculados
  - ✓ **Paso 4:** Cree otra columna y eleve al cuadrado los resultados obtenidos en el paso 3
  - ✓ **Paso 5:** Sume el total de los valores obtenidos en el paso 4:  $\sum(Desviaciones)^2$
  - ✓ **Paso 6:** Divida la suma anterior entre la cantidad de datos indicada en el paso 2
  - ✓ **Paso 7:** Interprete los resultados

Apliquemos estos pasos en el siguiente ejercicio

# Desviación estándar

## Ejercicio:

El MEP solicitó a 6 colegios de la Dirección Regional de Alajuela registrar la edad de los docentes para analizar cuántos se podrían pensionar y estimar las posibles plazas vacantes que se tendrían en los próximos años. **Estime el rango o amplitud de los datos.**

Edades de los  
docentes



Docente 1	35
Docente 2	40
Docente 3	25
Docente 4	64
Docente 5	53
Docente 6	56
Docente 7	62
Docente 8	36
Docente 9	41
Docente 10	25
Docente 11	56
Docente 12	43
Docente 13	38
Docente 14	55

# Desviación estándar

- ✓ **Paso 1:** Escriba en una columna los datos de los que dispone

**Datos (Xi)**

35
40
25
64
53
56
62
36
41
25
56
43
38
55

# Desviación estándar

## ✓ Paso 2:

Calcule el promedio o media aritmética de los datos ( $\bar{X}$ ) y la cantidad de los datos ( $n$ )

Datos ( $X_i$ )
35
40
25
64
53
56
62
36
41
25
56
43
38
55

$$\bar{X} = \frac{35+40+25+64+53+56+62+36+41+25+56+43+38+55}{14} = 44,9$$

- $\bar{X} = 44,9$
- $n = 14$

# Desviación estándar

## ✓ Paso 3:

Estime las desviaciones medias y cree una columna con los valores calculados

Datos: Xi	Desviaciones medias: $xi - \bar{X}$	Desviaciones medias: $xi - \bar{X}$
35	(35-44,9)	-9,9
40	(40-44,9)	-4,9
25	(25-44,9)	-19,9
64	(64-44,9)	19,1
53	(53-44,9)	8,1
56	(56-44,9)	11,1
62	(62-44,9)	17,1
36	(36-44,9)	-8,9
41	(41-44,9)	-3,9
25	(25-44,9)	-19,9
56	(56-44,9)	11,1
43	(43-44,9)	-1,9
38	(38-44,9)	-6,9
55	(55-44,9)	10,1

# Desviación estándar

## ✓ Paso 4:

Cree otra columna y eleve al cuadrado los resultados obtenidos en el paso 3

Datos: Xi	Desviaciones medias: $xi - \bar{X}$	Desviaciones medias: $(xi - \bar{X})^2$	Desviaciones medias al cuadrado: $(xi - \bar{X})^2$
35	-9,9	$(35-44,9)^2$	98,01
40	-4,9	$(40-44,9)^2$	24,01
25	-19,9	$(25-44,9)^2$	396,01
64	19,1	$(64-44,9)^2$	364,81
53	8,1	$(53-44,9)^2$	65,61
56	11,1	$(56-44,9)^2$	123,21
62	17,1	$(62-44,9)^2$	292,41
36	-8,9	$(36-44,9)^2$	79,21
41	-3,9	$(41-44,9)^2$	15,21
25	-19,9	$(25-44,9)^2$	396,01
56	11,1	$(56-44,9)^2$	123,21
43	-1,9	$(43-44,9)^2$	3,61
38	-6,9	$(38-44,9)^2$	47,61
55		$(55-44,9)^2$	102,01

**Datos:** **Desviaciones medias:** **Desviaciones medias:** **Desviaciones medias al cuadrado:**  
 **$X_i$**  **medias:  $x_i - \bar{x}$**   **$(x_i - \bar{X})^2$**  **cuadrado:  $(x_i - \bar{X})^2$**

35	-9,9	$(35-44,9)^2$	98,01
40	-4,9	$(40-44,9)^2$	24,01
25	-19,9	$(25-44,9)^2$	396,01
64	19,1	$(64-44,9)^2$	364,81
53	8,1	$(53-44,9)^2$	65,61
56	11,1	$(56-44,9)^2$	123,21
62	17,1	$(62-44,9)^2$	292,41
36	-8,9	$(36-44,9)^2$	79,21
41	-3,9	$(41-44,9)^2$	15,21
25	-19,9	$(25-44,9)^2$	396,01
56	11,1	$(56-44,9)^2$	123,21
43	-1,9	$(43-44,9)^2$	3,61
38	-6,9	$(38-44,9)^2$	47,61
55	10,1	$(55-44,9)^2$	102,01

$\Sigma(Desviaciones)^2$

**2 130,94**

# Desviación estándar

## ✓ Paso 6:

Divida la suma anterior entre y la cantidad de datos indicada en el paso 2. Luego saqué la raíz de los resultados

$$\text{Desviación estándar} = \sqrt{\frac{\sum(\text{Desviaciones})^2}{\text{número de datos}}}$$

$$\text{Desviación estándar} = \sqrt{\frac{2\,130,94}{14}}$$

$$\text{Desviación estándar} = 12,337$$

# Desviación estándar

## ✓ Paso 7:

Interprete los resultados

La desviación estándar obtenida fue de 12,34. Esto quiere decir que los datos se alejan de la media en cerca de 12 desviaciones estándar

Los datos no se alejan en gran magnitud del promedio, por lo que se puede confiar en esa medida para representar los datos

El MEP puede reportar con confianza que el promedio obtenido es de 45 años con poca variabilidad en los datos presentados

# Varianza

- Es una de las principales medidas de variabilidad, la desviación estándar surge de ella
- Es el valor cuadrado de la desviación estándar

$$\text{Variancia} = \frac{\sum(\text{Desviaciones})^2}{\text{número de datos}}$$

- Notar que la fórmula es la misma que la desviación estándar, pero no incluye la raíz.

# Varianza

- ✓ **Paso 6: Se obtiene en el paso 6 de los definidos en la desviación estándar**

Divida la suma de las desviaciones al cuadrado entre la cantidad de datos.

$$\text{Varianza} = \frac{\sum(\text{Desviaciones})^2}{\text{número de datos}}$$

$$\text{Varianza} = \frac{2130,94}{14}$$

$$\text{Varianza} = 152,21$$

Muchas gracias

