ANÁLISIS EXPLORATORIO DE DATOS

Explorando los Datos

Ivan Mauricio Bermúdez Vera
Estadístico. M.Sc. en Ingeniería
mauricio.bermudez@correounivalle.edu.co

Análisis Exploratorio

Todo estudio basado en datos. sin importar su alcance. debe superar la fase inicial del análisis exploratorio.

"Tabular, graficar, resumir, para identificar patrones y comportamientos regulares y presencia de irregularidades en los datos"

Preguntas a resolver:

1. ¿Existen patrones de comportamiento regular en los datos?
2. ¿Se presentan datos atípicos? ¿Que hacer con ellos?
3. ¿Como se relacionan las variables de análisis?
4. ¿Existen diferencias en el comportamiento de la variable entre grupos de análisis?

Es un paso necesario, que consume tiempo y que en ocasiones es descuidado por los analistas

Análisis Exploratorio

Proporciona un conjunto de herramientas que intentan descubrir patrones de comportamiento en los datos en un ambiente de variabilidad e incertidumbre.

No siempre se requiere aplicar todas las herramientas exploratorias, cada una presenta una utilidad de acuerdo a la necesidad y al propósito de la investigación.

Hipótesis ----------> Herramientas
(Objetivo) (Plan de Exploración)

Antes de Continuar.....

El Análisis Exploratorio de datos no es una rutina, es una actividad individual en la cual el analista escoge su ruta.

Para este tipo de análisis no existe una receta, existen herramientas, cuya implementación dependerá de la tipología de variables de análisis y de la necesidad de síntesis de la información.

Definiciones

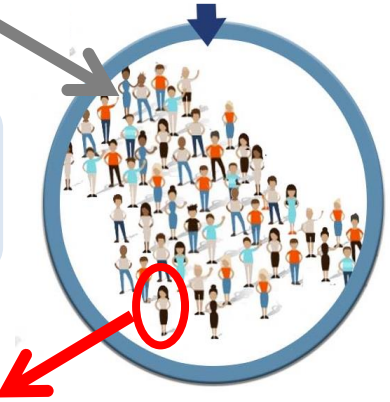
POBLACIÓN: Conjunto de Elementos de interés en una investigación.

1. El numero de elementos pueden ser finitos o infinitos

2.No debe asociarse exclusivamente con población humana

MUESTRA: Subconjuntos de elementos obtenidos desde la población de interés.

INDIVIDUO: Son los elementos que tienen información sobre el fenómeno que se estudia.



Definiciones

PARAMETRO: Término con el cual se identifica un indicador que hace referencia a la población.

Ejemplo:

- Edad promedio de los estudiantes de Estadística.
- Proporción de personas sin acceso a seguridad social.

ESTIMADOR: Indicador calculado sobre la muestra

Ejemplo:

- **Edad promedio de una muestra de estudiantes.**
- **Inversión promedio en desarrollo tecnológico en una muestra de colegios publicos de la ciudad de Cali.**

VARIABLE: Cualquier característica de interés de un individuo. Una variable puede tomar distintos valores para distintos individuos.

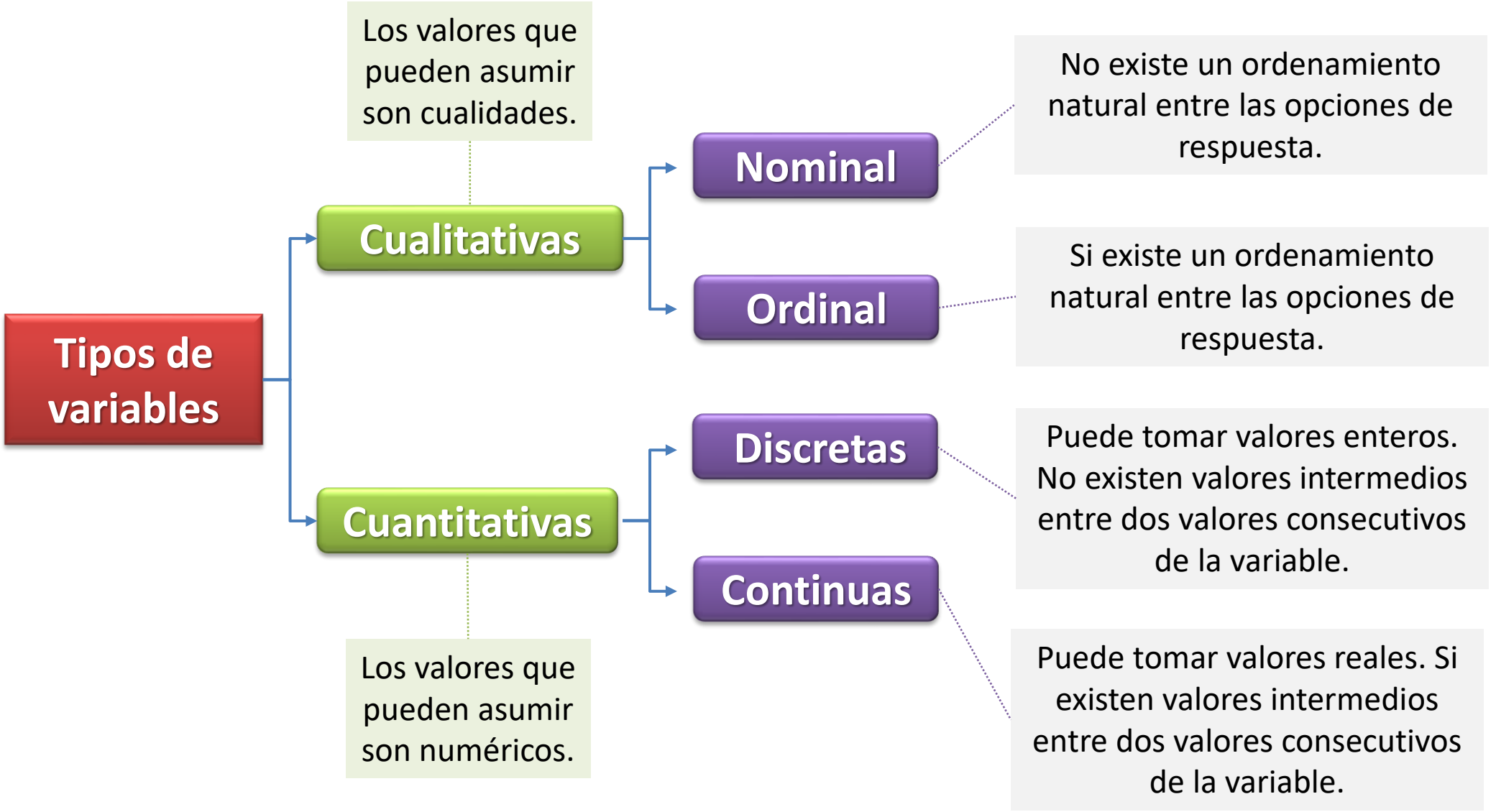
¿Que características puede identificar en el siguiente individuo?

EDAD
PESO
ESTATURA
INGRESOS
NRO. DE HIJOS

VARIABLES CUANTITATIVAS

SEXO
USA LENTES
ESTADO CIVIL
COLOR DE OJOS
NIVEL DE ESTUIO

VARIABLES CUALITATIVAS


Tipos de variables
Cualitativas
Los valores que pueden asumir son cualidades.
Nominal
No existe un ordenamiento natural entre las opciones de respuesta.
Ordinal
Si existe un ordenamiento natural entre las opciones de respuesta.
Cuantitativas
Los valores que pueden asumir son numéricos.
Discretas
Puede tomar valores enteros. No existen valores intermedios entre dos valores consecutivos de la variable.
Continuas
Puede tomar valores reales. Si existen valores intermedios entre dos valores consecutivos de la variable.

Ejemplos:

Variable	Tipo de Variable
Concentración de oxigeno disuelto en el agua (mg/l)	**Cuantitativa Continua**
El estado del clima.	**Cualitativa Nominal**
El número de hojas en una planta.	**Cuantitativa Discreta**
Nivel de riego (Bajo, medio, alto).	**Cualitativa Ordinal**
Temperatura ambiente.	**Cuantitativa Continua**

Herramientas para la descripción de datos

Indicadores Cuantitativos

- De frecuencia:
  - Conteos
  - Porcentajes
  - Tasas
- Tendencia Central:
  - Promedio
  - Mediana
  - Moda
- Dispersión:
  - Varianza
  - Desviación
  - Coeficiente de Variación
- Posición:
  - Percentiles, Deciles, Cuantiles
- Forma:
  - Asimetría, Curtosis
- De asociación:
  - Correlación

Resúmenes gráficos

- Gráficos de Barras
- Gráficos de sectores (Pastel)
- Histogramas
- Diagramas de Cajas y Alambres (Boxplot)
- Gráficos Temporales (de líneas)
- Gráficos Espaciales (Mapas)
- Diagramas de Dispersión (de correlación)

La idea es generar una combinación adecuada de gráficos. tablas e indicadores. que contribuyan a resumir la información

Tabulación y Representación Grafica de

Variables Cualitativas / Cuantitativas discretas

Tabulación y Representación Grafica de Variables Cualitativas

Un estudio quiere valorar la realidad actual respecto al consumo de cigarrillos en jóvenes con edades comprendidas entre los 15 y 20 años. Para ello ha tomado una muestra aleatoria de 40 jóvenes a los cuales les indaga acerca de su consumo de cigarrillos, los resultados son lo siguientes:

Si Si Si NO NO NO Si Si NO Si Si Si Si Si Si Si Si Si Si Si Si NO Si

Si Si NO NO NO Si NO Si NO NO NO Si Si NO Si NO

Muestra Bruta = Datos

¿Que puede decir usted acerca de los resultados obtenidos?

Variables cualitativas o cuantitativas discretas

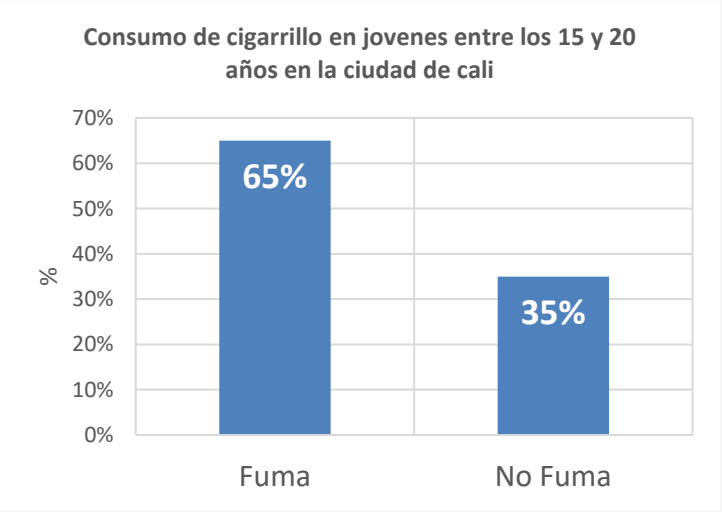
Tablas de Frecuencia

FUMA	Nro. Casos	%
SI	26	65%
NO	14	35%
Total	40	100%

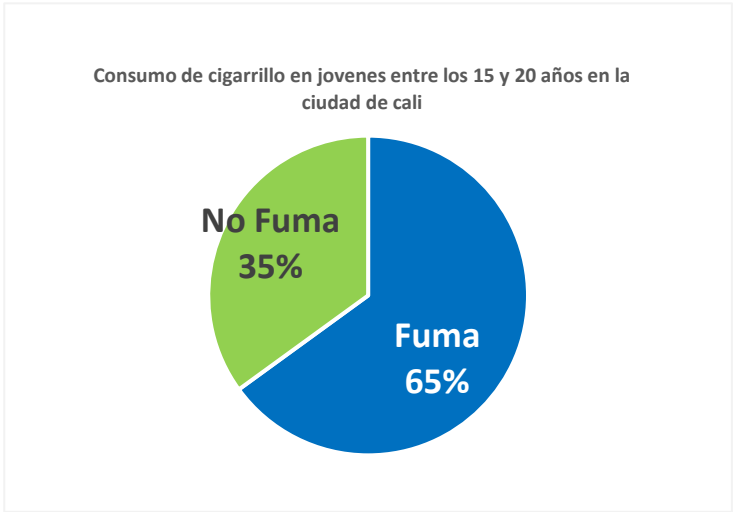
Funciones en R

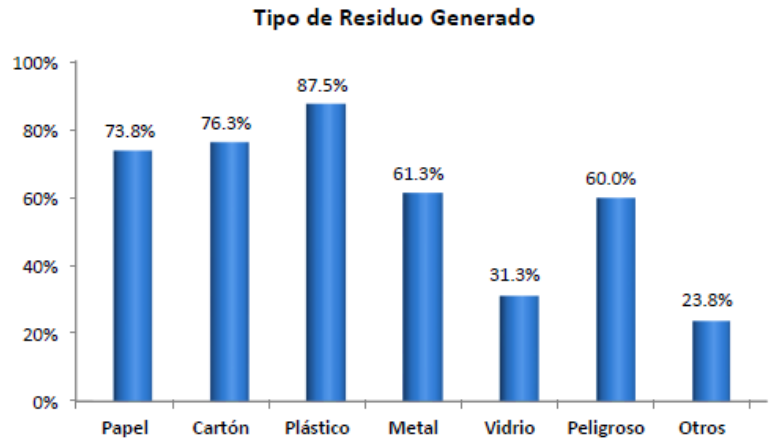
- table(x), prop.table(x, margin)
- barplot(table)
- pie(table

Diagramas de Barra



Diagramas de sectores

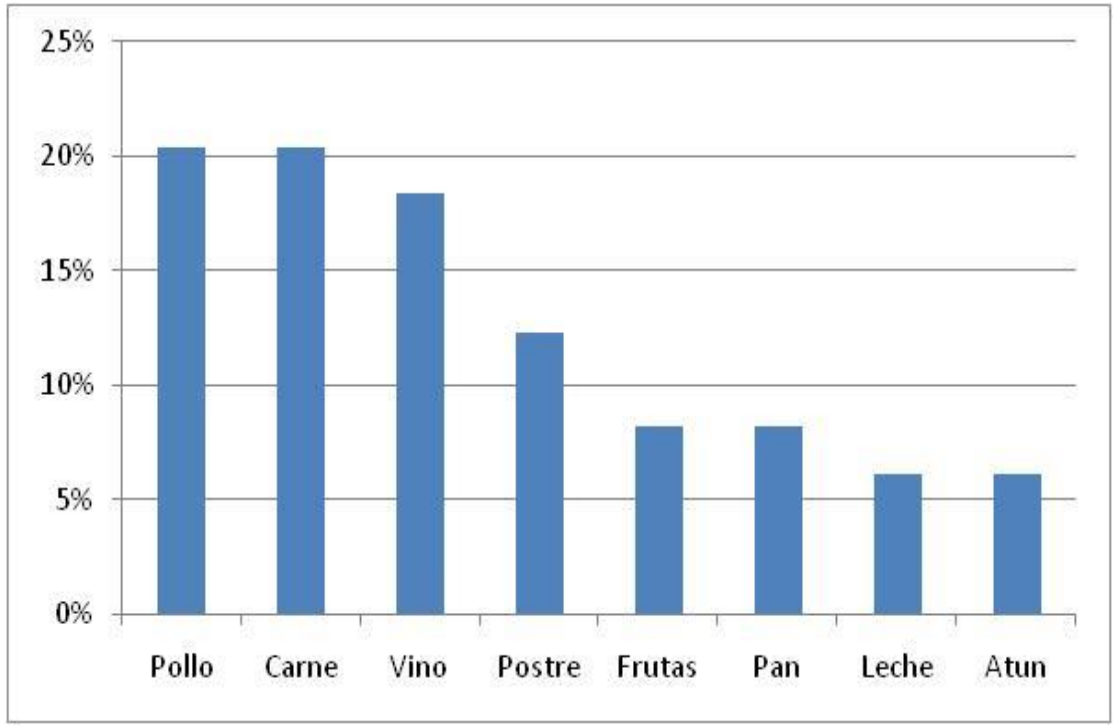




Características con muchas opciones de respuesta

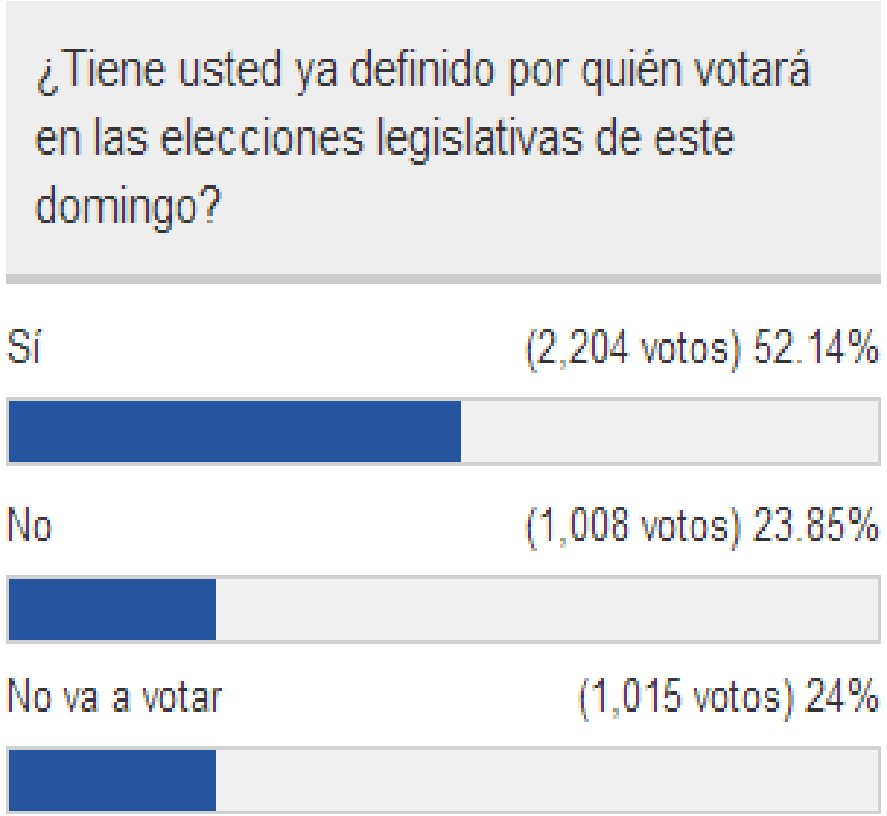
Buscando la fuente de una Intoxicación grupal:

¿Consumió alguno de los siguientes alimentos?



1. ¿Porque no es adecuado utilizar un grafico de pastel?
2. ¿Es suficiente para pensar que la culpa es del Pollo o de la Carne?

Variables cualitativas o cuantitativas discretas



a. Defina la variable.

b. Defina el tipo de variable y escala de medición.

c. Cuantas personas respondieron la encuesta?

d. Más de la mitad de las personas tienen definido por quién votará?

e. El 23.85% de las personas No va a votar?

¿Barras o Pastel?

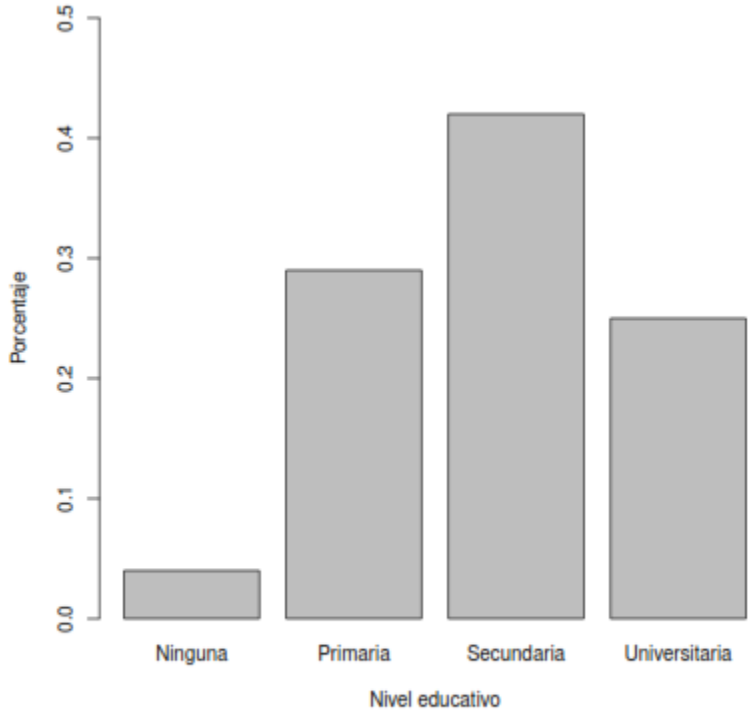


Figura: Diagrama de Barras

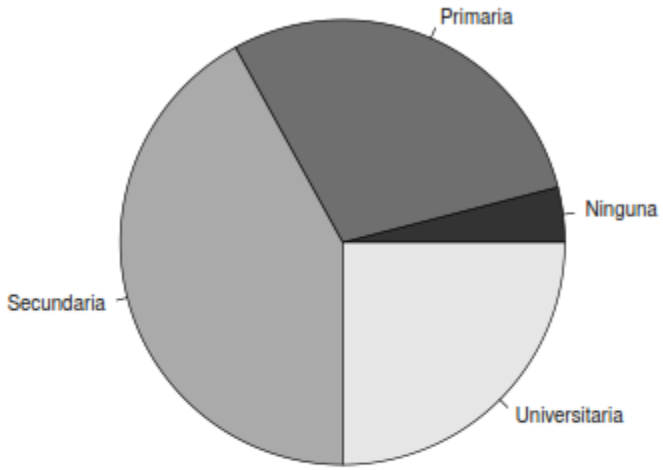
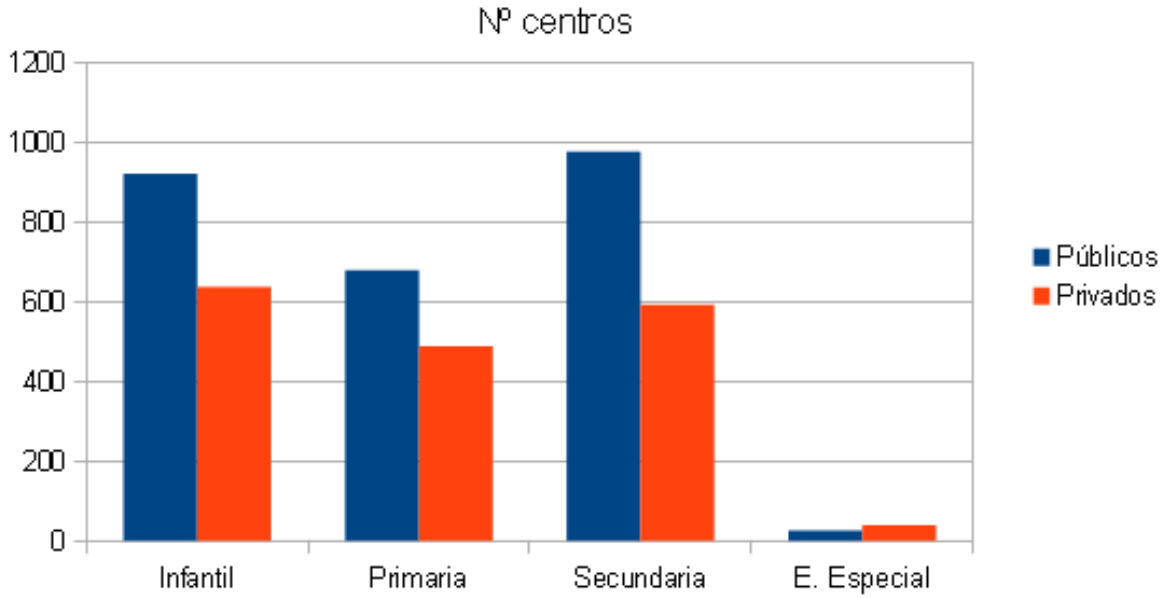
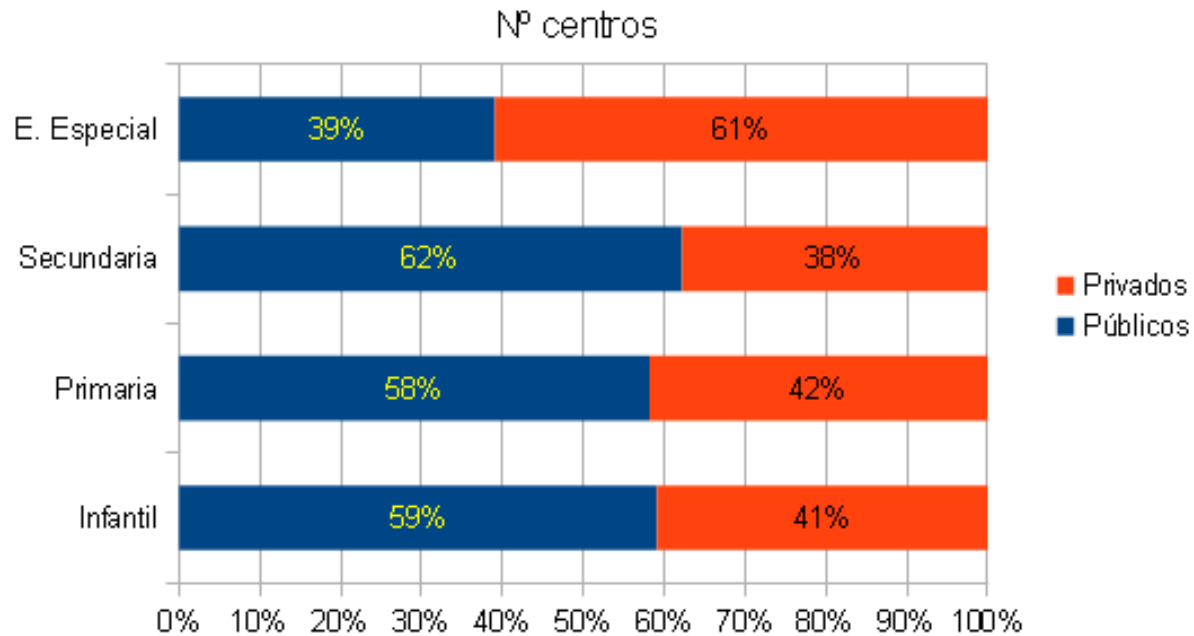


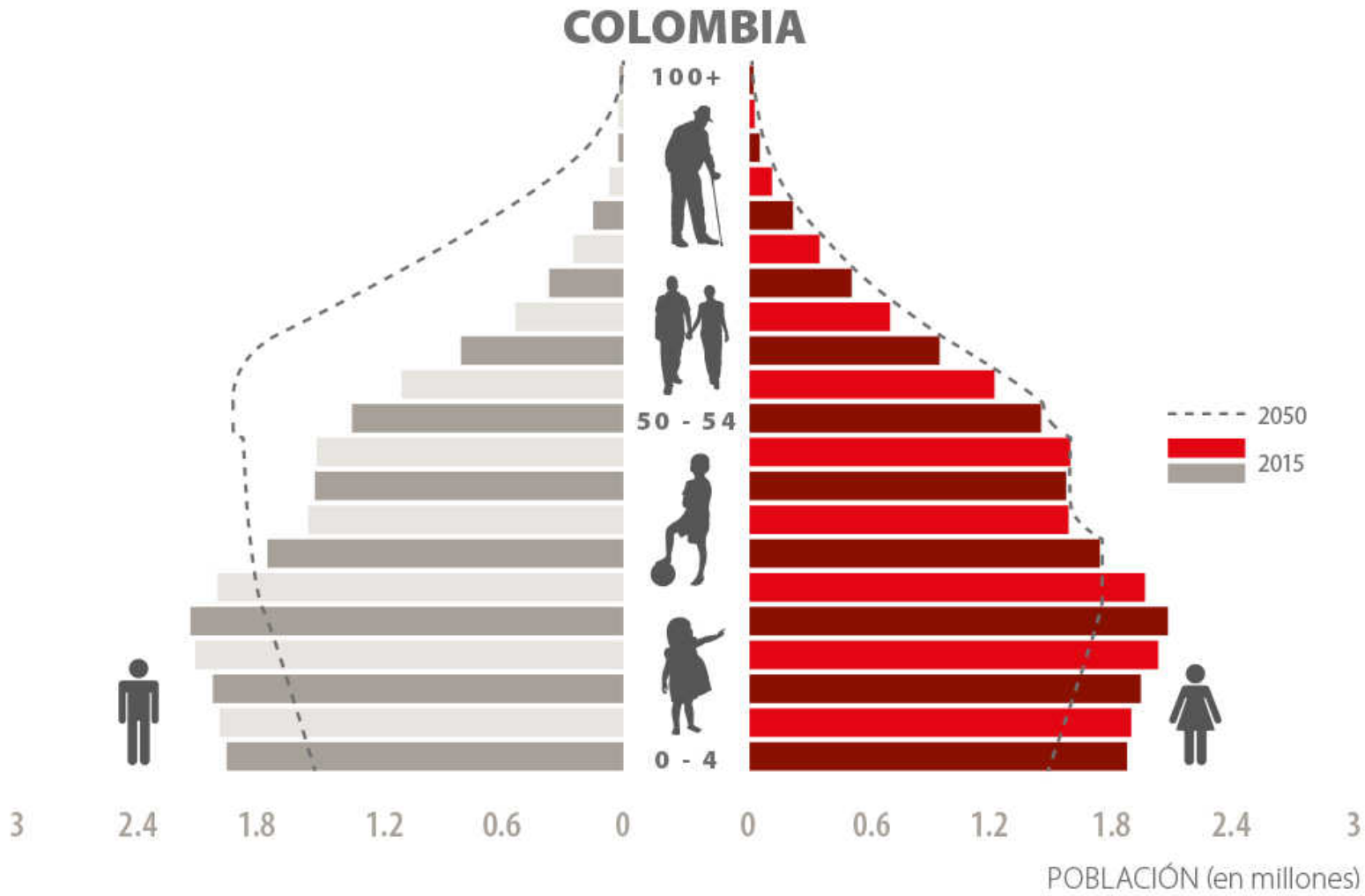
Figura: Diagrama de pastel

Diagramas de Barras

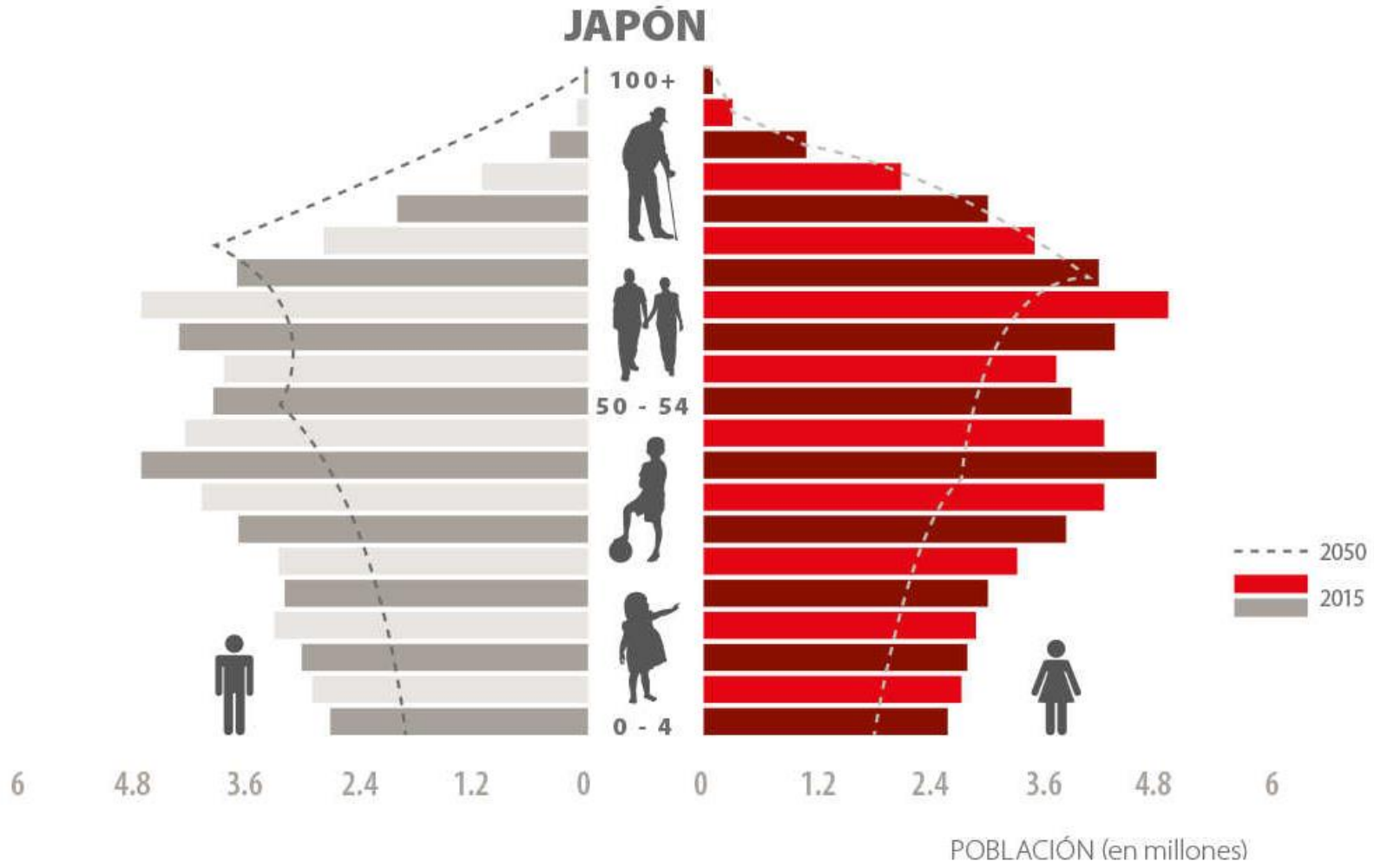




Pirámides Poblacionales



Pirámides Poblacionales



Un grafico vale más que mil palabras!

¿La distribución del nivel educativo máximo alcanzado es la misma para hombres y mujeres?

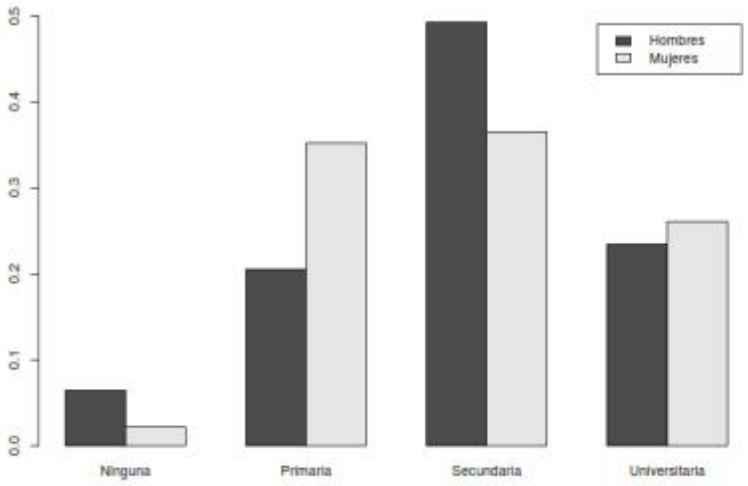


Figura: Buena representación

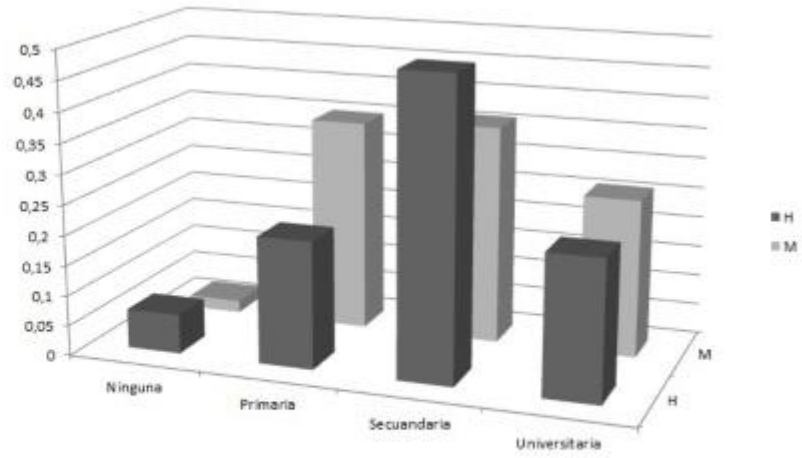
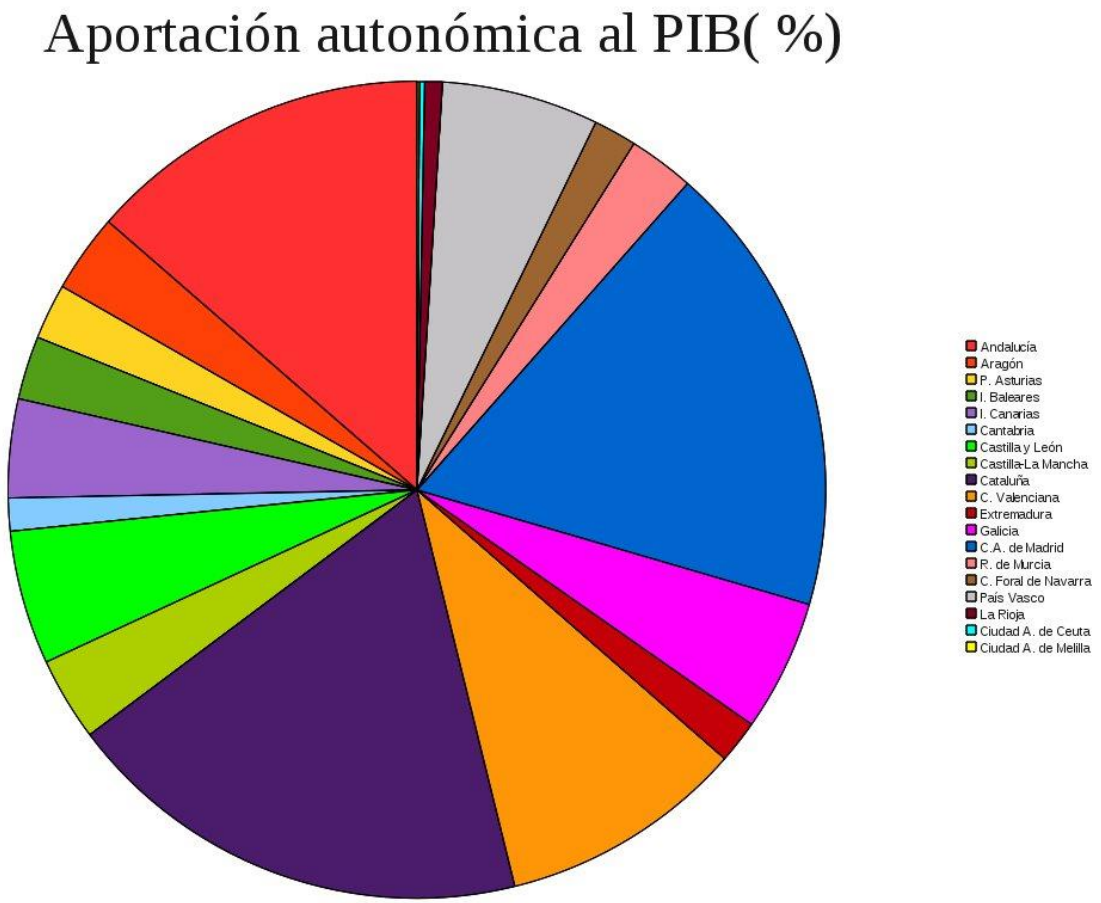
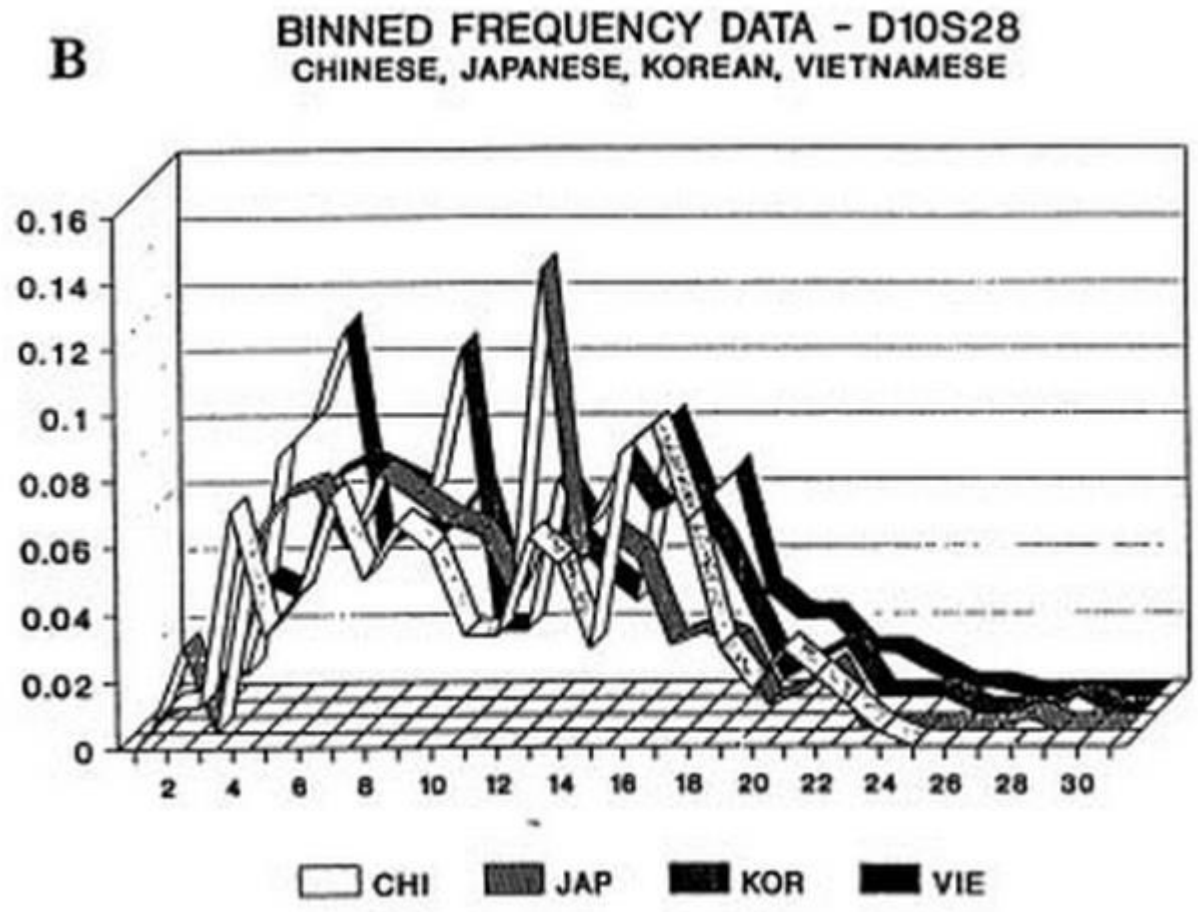


Figura: Mala representación

Un grafico vale más que mil palabras!

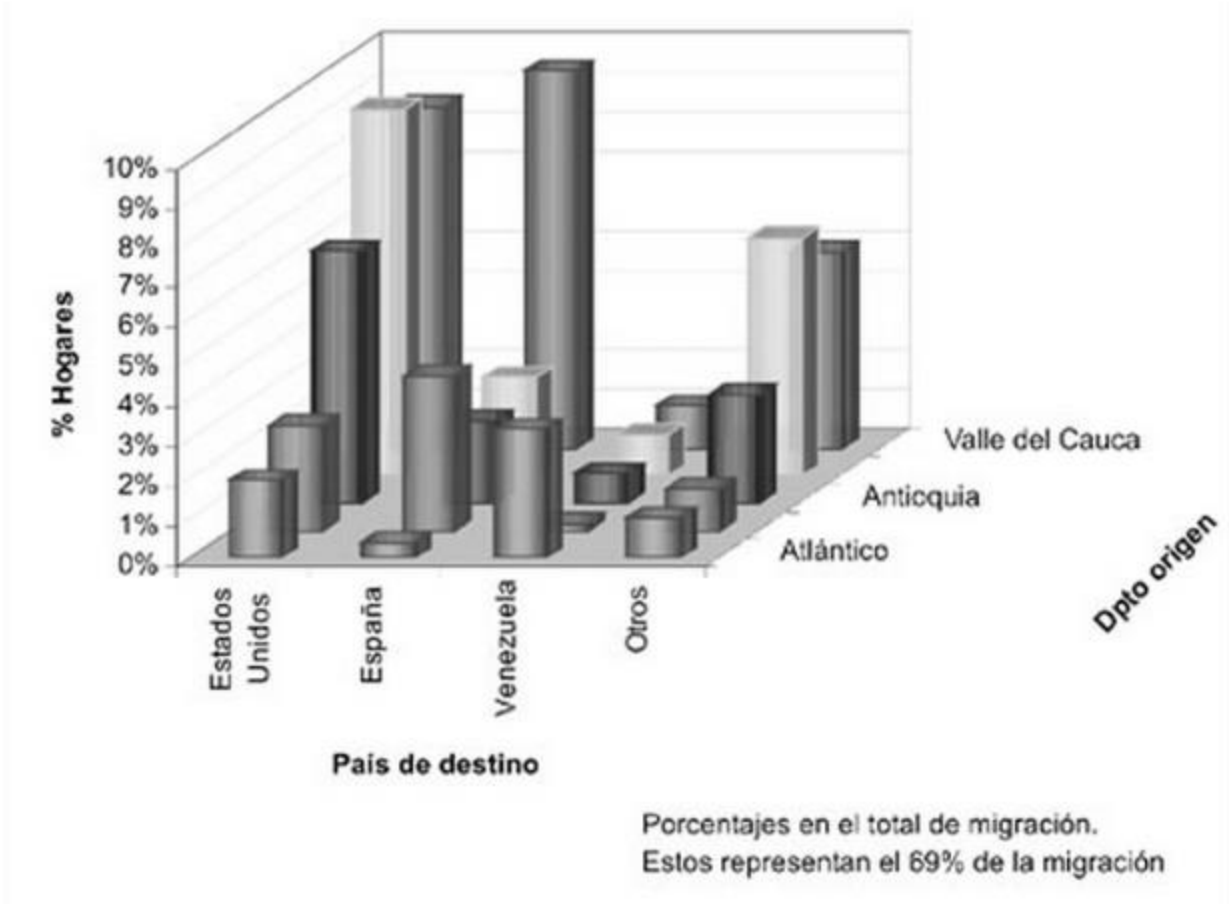


Un grafico vale más que mil palabras!



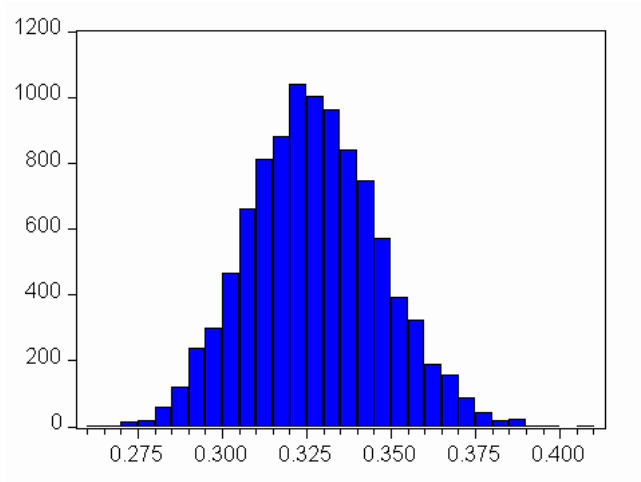
"Los gráficos no deben ser más complejos que los datos que describe" (evite efectos 3D).

Un grafico vale más que mil palabras!



"La perspectiva hace difícil la comparación de la altura de los cubos"

Tabulación y Representación grafica de datos cuantitativos



Datos Cuantitativos Discretos

Ejemplo: número de chocolatinas defectuosas que contiene cada caja de un lote de producción.

Muestra bruta:

3, 2, 0, 2, 3, 3, 1, 1, 0, 1, 3, 3, 4, 4, 3,
2, 4, 2, 4, 2, 0, 2, 4, 3, 1, 2, 4, 3, 0, 2.

x_i (Valor observado)	Conteo	n_i (Frecuencia absoluta)
0	||||	4
1	||||	4
2	𝍸 |||	8
3	𝍸 |||	8
4	𝍸 |	6
Total		30

TABLA DE FRECUENCIA
NUMERO DE PIEZAS DEFECTUOSAS QUE CONTIENEN LAS CAJAS.

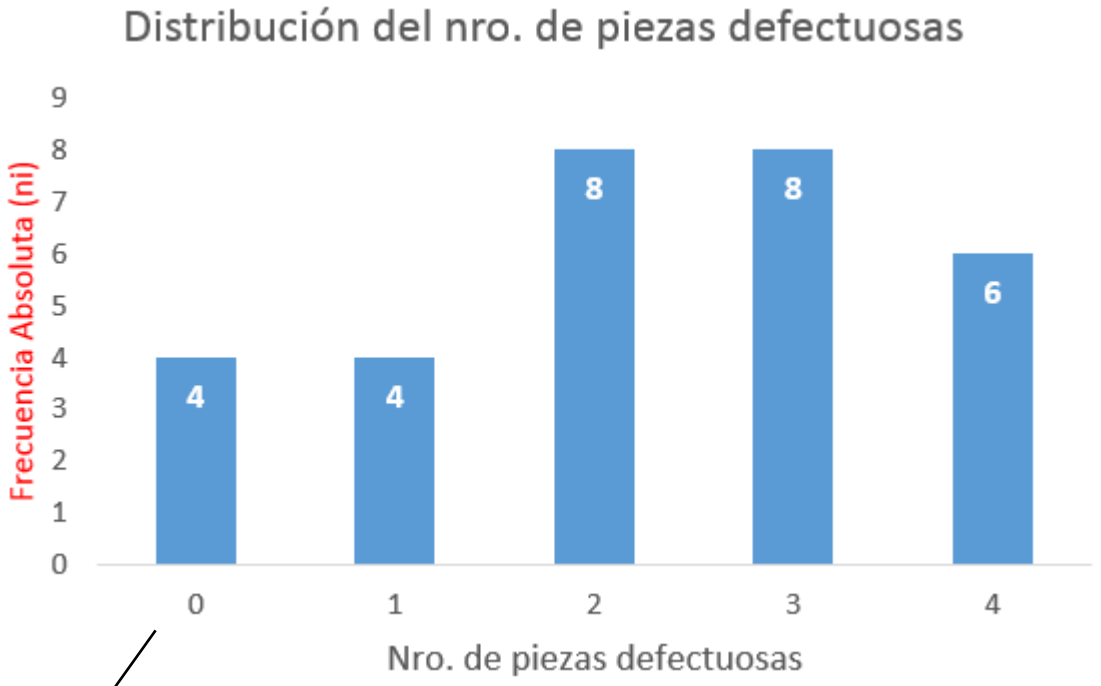
x_i Valor observado	n_i Frecuencia Absoluta	f_i Frecuencia Relativa	N_i Frecuencia Absoluta Acumulada	F_i Frecuencia Relativa Acumulada
0	4	0.133	4	0.133
1	4	0.133	8	0.267
2	8	0.267	16	0.533
3	8	0.267	24	0.800
4	6	0.200	30	1.0
Total	30	1.0		

REPRESENTACIÓN GRAFICA DE UNA DISTRIBUCIÓN DE FRECUENCIAS

Caso Discreto

Diagrama de Barras

En el Eje horizontal se representan los valores que asume la variable y en el eje horizontal su frecuencia absoluta o relativa



Puede ser la frecuencia absoluta o relativa (%)

x_i	n_i	f_i
0	4	0.133
1	4	0.133
2	8	0.267
3	8	0.267
4	6	0.200
	30	1.0

Por ser una variable discreta las barras no deben juntarse

Ejercicio

Numero de clientes que llegan por hora a un cajero automático:

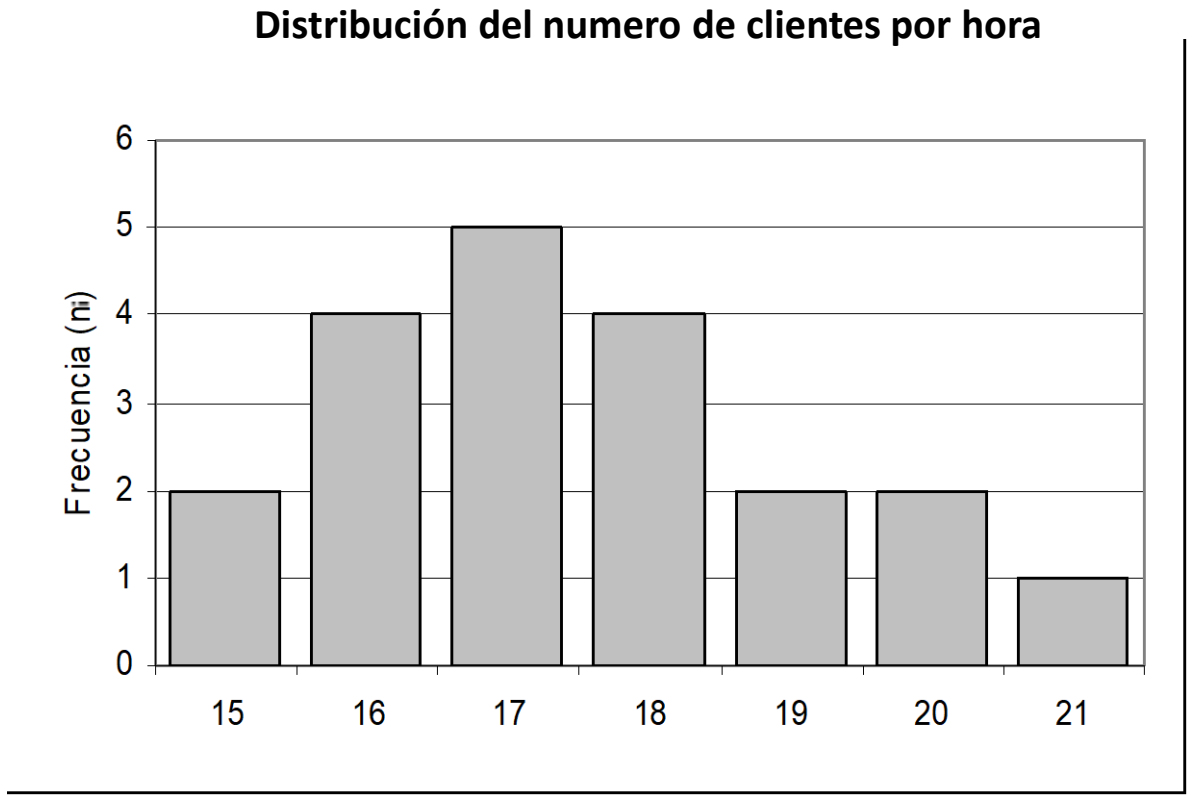
15, 16, 19, 18, 16, 17, 15, 18, 18, 17,
20, 16, 17, 18, 17, 19, 20, 21, 16, 17

a. Identifique el tipo de variable.
b. Construya su respectiva distribución de frecuencias.
c. Realice el grafico para la frecuencia absoluta simple.
d. En que porcentaje de las ocasiones se presentan al servicio menos de 19 clientes?
e. En que porcentaje 19 o mas clientes?

Distribución de Frecuencias

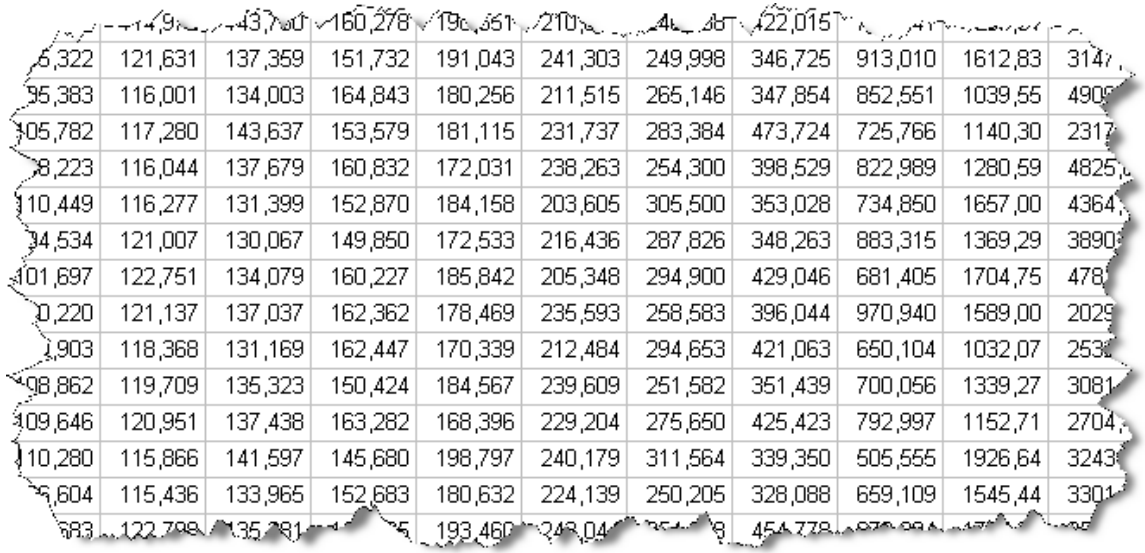
TABLA DE FRECUENCIA DEL NUMERO DE CLIENTES POR HORA

X_i Valor observado	n_i (Frecuencia Absoluta)	f_i (Frecuencia Relativa)	N_i (Frecuencia Absoluta Acumulada)	F_i (Frecuencia Relativa Acumulada)
15	2	0.1	2	0.1
16	4	0.2	6	0.3
17	5	0.25	11	0.55
18	4	0.2	15	0.75
19	2	0.1	17	0.85
20	2	0.1	19	0.95
21	1	0.05	**20**	**1.00**
Total	**20**	**1.0**		



Datos Cuantitativos Continuos

Suponga que se tiene la siguiente información de la duración en horas de cierto dispositivo electrónico.



6,322	121,631	137,359	151,732	191,043	241,303	249,998	346,725	913,010	1612,83	314
95,383	116,001	134,003	164,843	180,256	211,515	265,146	347,854	852,551	1039,55	490
105,782	117,280	143,637	153,579	181,115	231,737	283,384	473,724	725,766	1140,30	231
8,223	116,044	137,679	160,832	172,031	238,263	254,300	398,529	822,989	1280,59	4825
110,449	116,277	131,399	152,870	184,158	203,605	305,500	353,028	734,850	1657,00	4364
94,534	121,007	130,067	149,850	172,533	216,436	287,826	348,263	883,315	1369,29	3890
101,697	122,751	134,079	160,227	185,842	205,348	294,900	429,046	681,405	1704,75	478
0,220	121,137	137,037	162,362	178,469	235,593	258,583	396,044	970,940	1589,00	2029
1,903	118,368	131,169	162,447	170,339	212,484	294,653	421,063	650,104	1032,07	253
98,862	119,709	135,323	150,424	184,567	239,609	251,582	351,439	700,056	1339,27	3081
109,646	120,951	137,438	163,282	168,396	229,204	275,650	425,423	792,997	1152,71	2704
110,280	115,866	141,597	145,680	198,797	240,179	311,564	339,350	505,555	1926,64	3243
6,604	115,436	133,965	152,683	180,632	224,139	250,205	328,088	659,109	1545,44	3301

"Seguramente todos los datos sean distintos y la tabla de frecuencia no resumiría en nada la información"

Solución ⟶ **Realizar Agrupaciones**

Intervalos de Clase

Para variables continuas es preferible agrupar la información en **intervalos de clase**. pero,

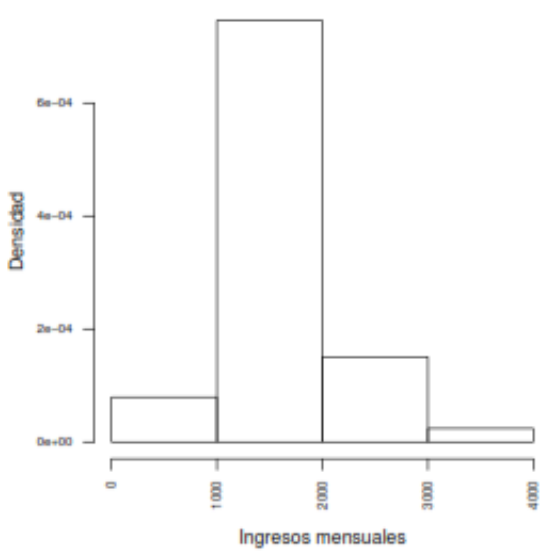
¿Cuántos intervalos?

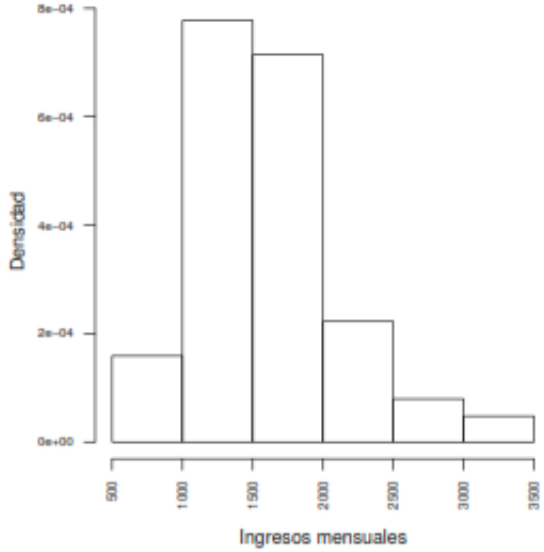
Siempre que se agrupan los datos en intervalos de clase se produce pérdida de información.

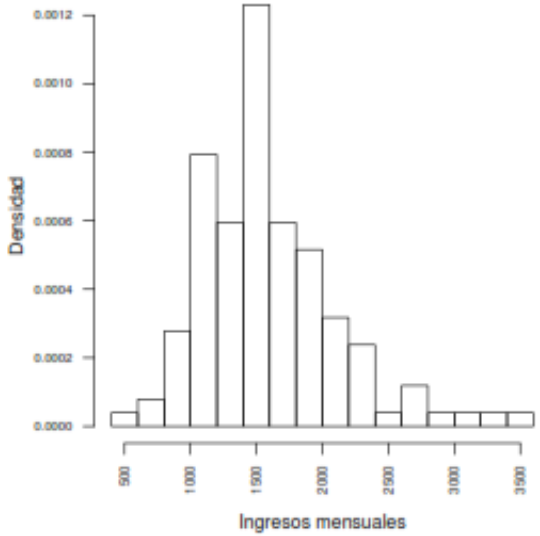
- Si se usan pocos intervalos se globalizan más los datos y se pierde más información.
- Si se usan muchos intervalos la manipulación de los datos se hace compleja y su presentación poco visible.
- Se recomienda utilizar entre 5 y 10 intervalos de clase.
- Una posible solución (aunque la selección puede ser arbitraria) es: $2^k > n$

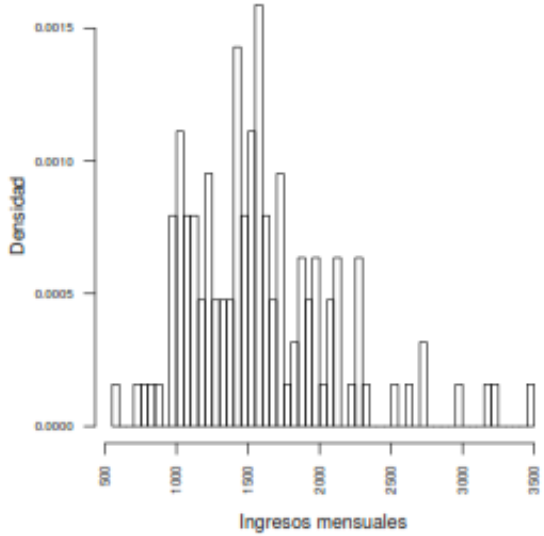
$1+3.3\text{Log}_{10}(n)$

¿Cuantas Clases usar?









Datos Cuantitativos Continuos

Ejemplo:

Una entidad encargada del control de contaminación de cierto río, lleva registros sobre el oxígeno disuelto (x), expresado en mg/l; éstos se presentan a continuación:

2.6	4.0	2.8	1.9	3.5
3.6	3.2	1.8	4.5	1.6
3.1	2.5	4.2	1.2	3.2
2.6	1.7	3.5	2.2	4.4
2.7	0.3	2.4	2.2	1.4
3.9	3.1	2.2	3.0	0.7
2.4	2.6	3.4	2.1	2.8
2.7	1.3	3.7	1.8	3.3
2.5	4.3	0.8	2.9	0.5
2.3	1.5	2.3	3.8	2.3

Distribución de Frecuencia

PASOS PARA CONSTRUIR UNA DISTRIBUCIÓN DE FRECUENCIA EN DATOS AGRUPADOS

1. Determinar el numero de intervalos **(k)** que deseamos construir:

$$2^k > n \quad \longrightarrow \quad 2^6 = 64 > 50 \quad \longrightarrow \quad \boldsymbol{k = 6}$$

2. Determinar el rango de variación **(R)**:

$$Rango = Max(x_i) - Min(x_i) \quad \longrightarrow \quad R = 4.5 - 0.3 = \mathbf{4.2}$$

3. Fijar el ancho de clases **(C)**:

$$C = \frac{R}{k} \quad \longrightarrow \quad C = 4.2 / 6 = \mathbf{0.7}$$

Distribución de Frecuencia

PASOS PARA CONSTRUIR UNA DISTRIBUCIÓN DE FRECUENCIA EN DATOS AGRUPADOS

4. Determinar los limites **(L_0, L_1, L_2, …, L_k):**

L_o = Min

$L_1 = L_0 + C$

$L_2 = L_1 + C$

$L_3 = L_2 + C$

$L_i = L_{i-1} + C$

$L_o = 0.3$

$L_1 = 0.3 + 0.7 = 1.0$

$L_2 = 1.0 + 0.7 = 1.7$

$L_3 = 1.7 + 0.7 = 2.4$

$L_4 = 2.4 + 0.7 = 3.1$

$L_5 = 3.1 + 0.7 = 3.8$

$L_6 = 3.8 + 0.7 = 4.5$

Intervalos de Clase	x'_i Marca de clase
[0.3 , 1.0]	0,65
(1.0 , 1.7]	1,35
(1.7 , 2.4]	2,05
(2.4 , 3.1]	2,75
(3.1 , 3.8]	3,45
(3.8 , 4.5]	4,15

5. Calcular la marca de clase **(x'_i):**

$$x'_i = \frac{L_{i-1} + L_i}{2}$$

Distribución de Frecuencia

TABLA DE FRECUENCIA DEL REGISTRO DE OXIGENO DISUELTO DE CIERTO RÍO (mg/l)

Intervalos de Clase	x'_i Marca de clase	n_i	f_i	N_i	F_i
[0.3 , 1.0]	0,65	4	0,08	4	0,08
(1.0 , 1.7]	1,35	6	0,12	10	0,20
(1.7 , 2.4]	2,05	12	0,24	22	0,44
(2.4 , 3.1]	2,75	13	0,26	35	0,70
(3.1 , 3.8]	3,45	9	0,18	44	0,88
(3.8 , 4.5]	4,15	6	0,12	50	1,00
	Total	**50**	**1.0**		

El 18% de las mediciones presentaron un registro de oxigeno disuelto entre 3.1 y 3.8 mg/l.

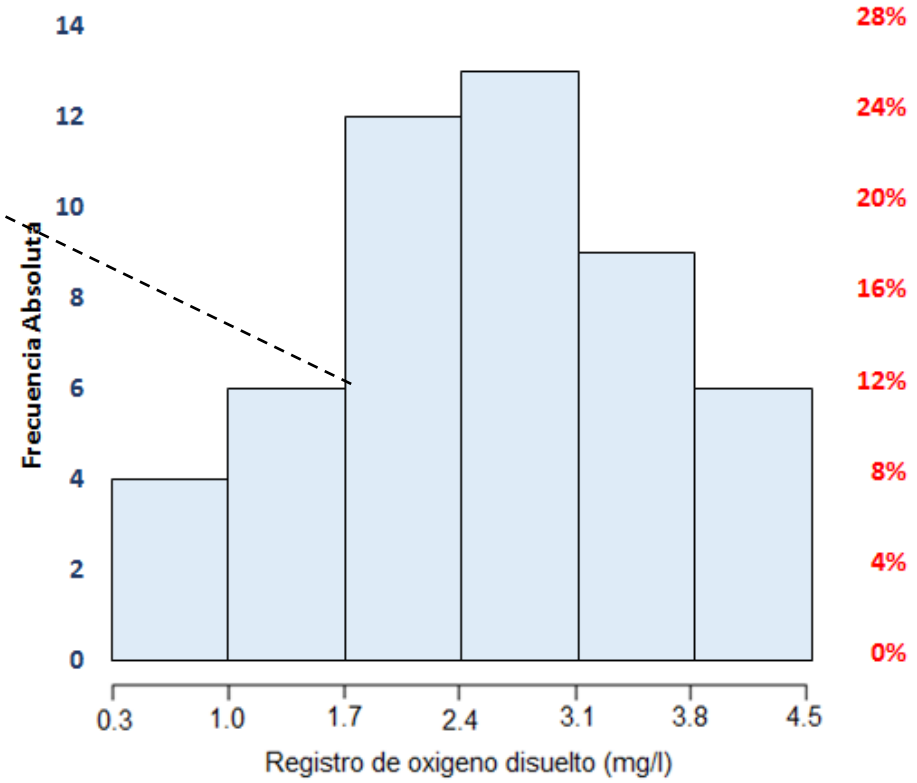
El 88% de las mediciones presentaron un registro de oxigeno disuelto menor o igual a 3.8 mg/l.

REPRESENTACIÓN GRAFICA DE UNA DISTRIBUCIÓN DE FRECUENCIAS – Caso Continuo

Histograma de Frecuencias (Variable agrupada)

Las clases se indican en el eje horizontal y su frecuencias (relativas o absolutas) sobre el eje vertical

La barras se juntan por continuidad de la variable

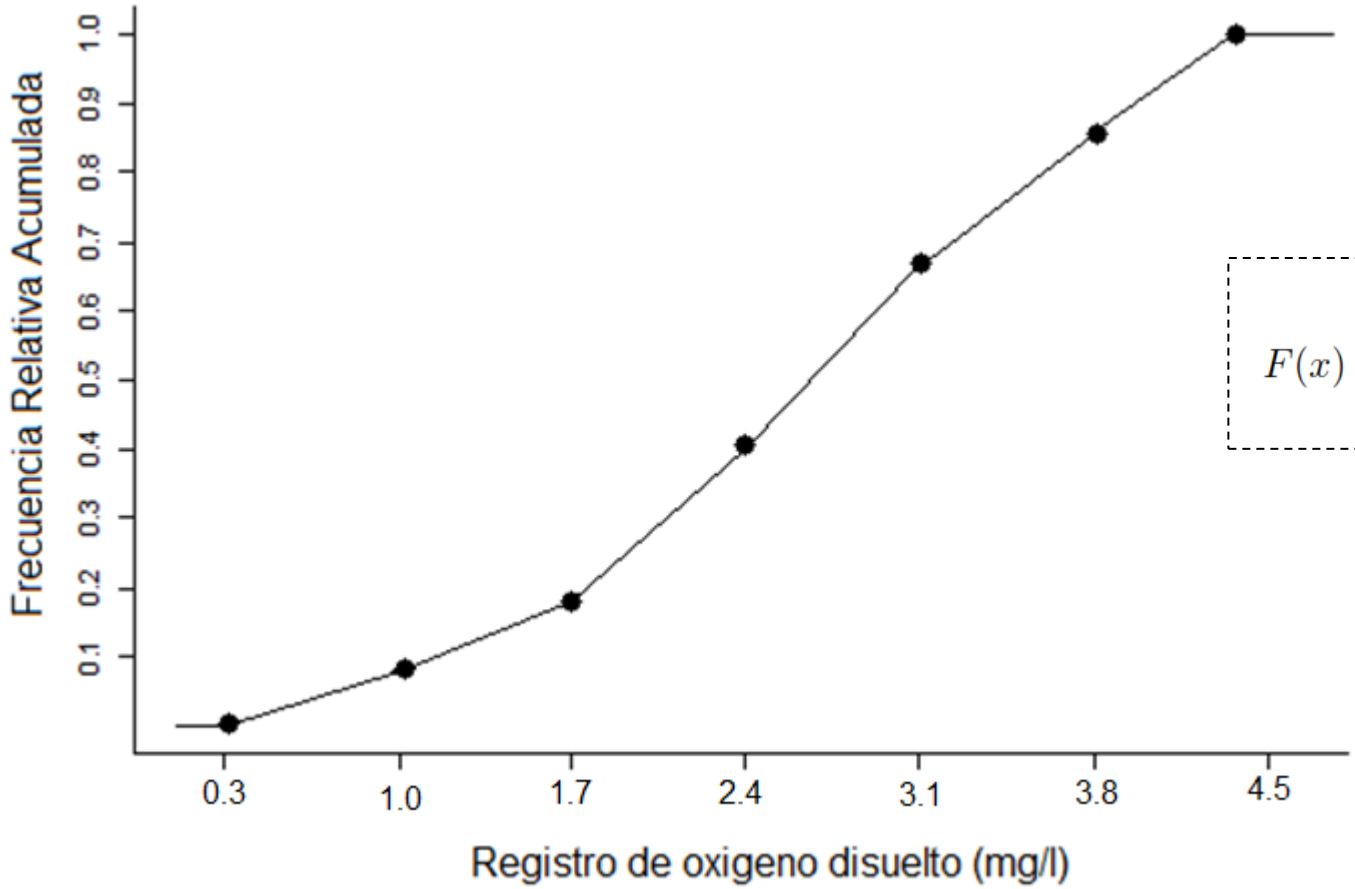


Intervalos	n_i	f_i
[0.3 , 1.0]	4	0,08
(1.0 , 1.7]	6	0,12
(1.7 , 2.4]	12	0,24
(2.4 , 3.1]	13	0,26
(3.1 , 3.8]	9	0,18
(3.8 , 4.5]	6	0,12
	50	1.0

Función empírica de distribución acumulada

Cada intervalo de F(x), representa un segmento de recta, cuya pendiente es la densidad del intervalo respectivo. Esto da origen al gráfico que lleva el nombre de **ojiva**.

Ojiva del Oxigeno Disuelto en un Río (mg/l).



$$F(x) = \begin{cases} 0, & \text{para } x < L_0, \\ F(L_{i-1}) + \frac{f_i}{C_i}(x - L_{i-1}) & \text{para } L_{i-1} < x \leq L_i \\ 1, & \text{para } x > L_m, \end{cases}$$

Distribución de Frecuencia

La entidad encargada del estudio sabe que si el nivel de oxigeno disuelto en el rio es inferior a 1.5 mg/l se pueden presentar consecuencias negativas para la calidad del agua y por lo tanto deberán de intervenir.

Intervalos	x'_i	n_i	f_i	N_i	F_i
[0.3 , 1.0]	0,65	4	0,08	4	0,08
(1.0 , 1.7]	1,35	6	0,12	10	0,20
(1.7 , 2.4]	2,05	12	0,24	22	0,44
(2.4 , 3.1]	2,75	13	0,26	35	0,70
(3.1 , 3.8]	3,45	9	0,18	44	0,88
(3.8 , 4.5]	4,15	6	0,12	50	1,00
	Total	50	1.0		

¿Que porcentaje de las mediciones presentan registros menores o iguales a 1.5 mg/l ?

Ejercicio

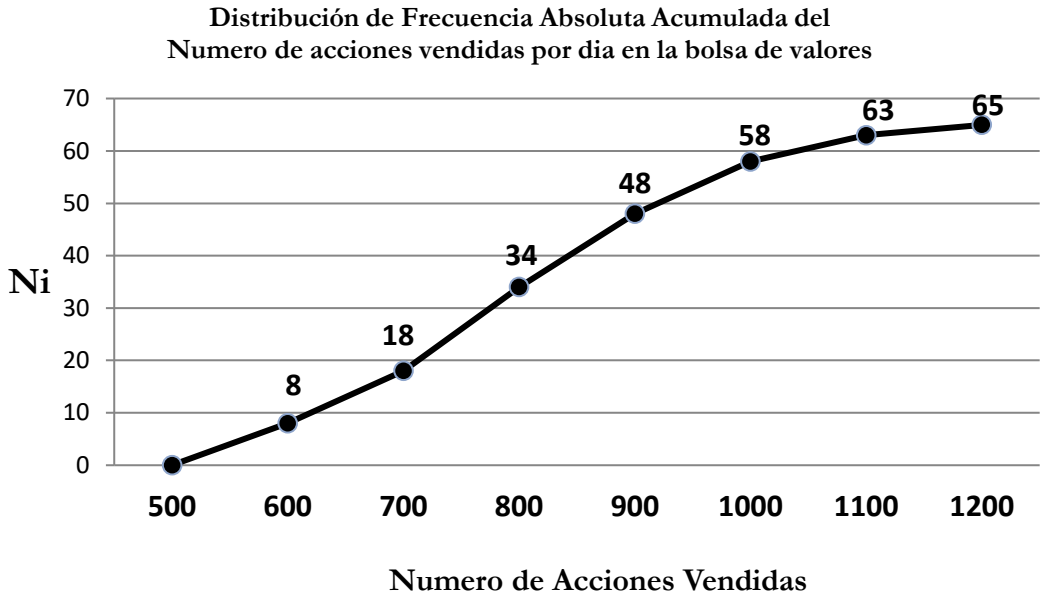
Intervalos	x'_i	n_i	f_i	N_i	F_i
[0.3 , 1.0]	0,65	4	0,08	4	0,08
(1.0 , 1.7]	1,35	6	0,12	10	0,20
(1.7 , 2.4]	2,05	12	0,24	22	0,44
(2.4 , 3.1]	2,75	13	0,26	35	0,70
(3.1 , 3.8]	3,45	9	0,18	44	0,88
(3.8 , 4.5]	4,15	6	0,12	50	1,00
	Total	50	1.0		

A partir de la tabla de frecuencias:

1. ¿Qué porcentaje de registros presentan niveles de OD superiores a 2.0 mg/l.
2. ¿Qué porcentaje de registros presentan niveles de OD entre 2.5 y 3.5 mg/l.
3. ¿A partir de que valor de OD se encuentra acumulado el 90% de los datos?

Ejercicio

La siguiente grafica de Ojiva presenta la frecuencia absoluta acumulada del número de acciones de Ecopetrol vendidas por día, para un total de 65 días.



i. De acuerdo con la gráfica construya la tabla de frecuencias respectiva (trabaje con 2 decimales).
ii. Interprete los valores de n_3 , N_4 , f_5 y F_6
iii. Presente gráficamente la frecuencia relativa del número de acciones de Ecopetrol vendidas por día.
iv. Construya su respectivo diagrama de cajas. Que puede decir de los datos?
v. ¿Qué porcentaje de días presentan entre 750 y 950 acciones vendidas.