

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 Gráfica de Variables Cualitativas

Se está realizando un estudio sobre la **satisfacción** de los ciudadanos con un nuevo programa de reciclaje implementado en un municipio.

Para ello, se han encuestado a **40 residentes** y se les ha preguntado si están satisfechos o no con el programa.

La variable dicotómica que estás analizando es:

- **1:** "Satisfecho con el programa"
- **0:** "No satisfecho con el programa"



Datos

1	1	1	0	0	0	1	1	0	1	1	1	1
1	1	1	1	1	1	1	1	1	0	1	1	1
0	0	1	0	1	0	0	0	1	1	0	1	0

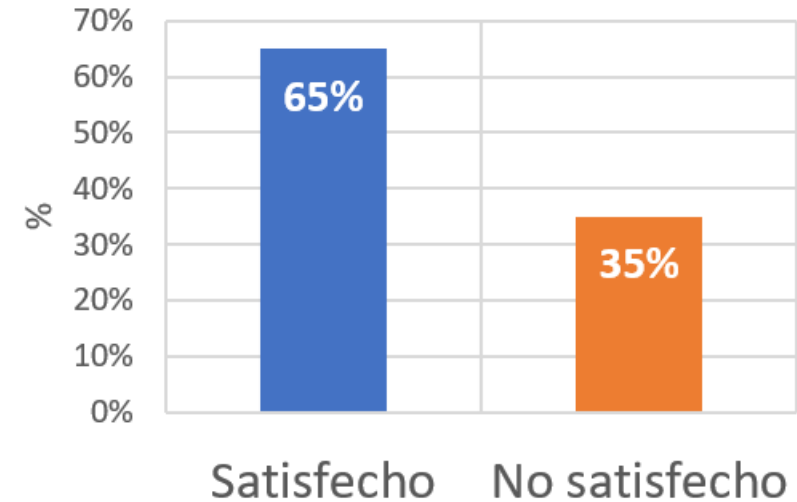
¿Qué puede decir usted acerca de los resultados obtenidos?

Variables cualitativas o cuantitativas discretas

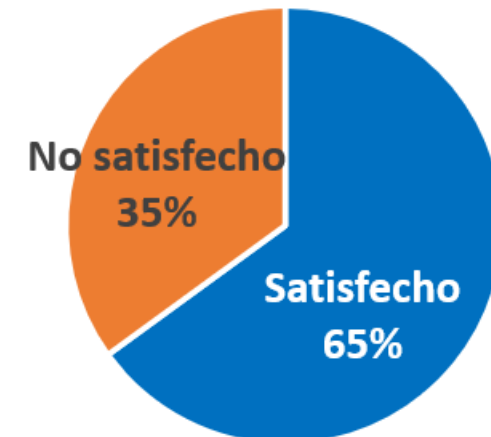
Tablas de Frecuencia

Satisfacción con el programa de reciclaje	Frecuencia	%
Satisfecho	26	65%
No satisfecho	14	35%
Total	40	100%

Diagramas de Barra



Diagramas de sectores



Variables cualitativas o cuantitativas discretas

Tablas de Frecuencia

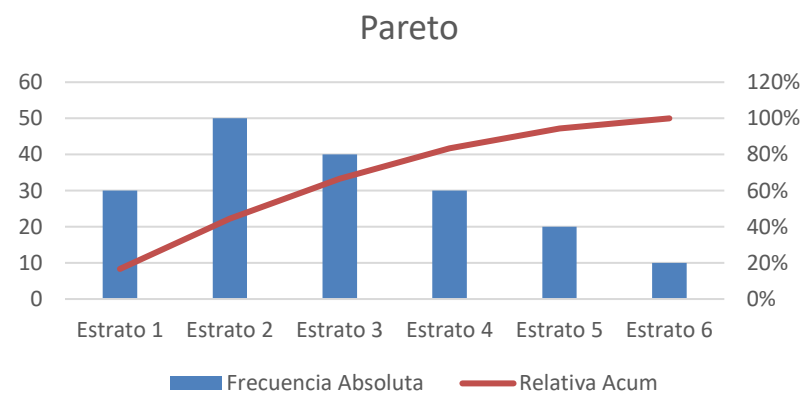
Categorías o Valores	Frecuencia Absoluta	Frecuencia Relativa	Frecuencia Acumulada	
			Absoluta	Relativa
X_1	n_1	$n_1/n*100$	n_1	$n_1/n*100$
X_2	n_2	$n_2/n*100$	n_1+n_2	$(n_1+n_2)/n*100$
.
.
.
X_j	n_j	$n_j/n*100$	n	100%
Total	n	100%		

Variables cualitativas o cuantitativas discretas

Variable Ordinal

Se pueden calcular las frecuencias acumuladas

Categorías o Valores	Frecuencia Absoluta	Frecuencia Relativa	Frecuencia Acumulada	
			Absoluta	Relativa Acum
Estrato 1	30	17%	30	17%
Estrato 2	50	28%	80	44%
Estrato 3	40	22%	120	67%
Estrato 4	30	17%	150	83%
Estrato 5	20	11%	170	94%
Estrato 6	10	6%	180	100%
Total	180	100%		



En variables ordinales no es recomendable el grafico de torta, pero se puede realizar según el objetivo

¿Barras o Pastel?

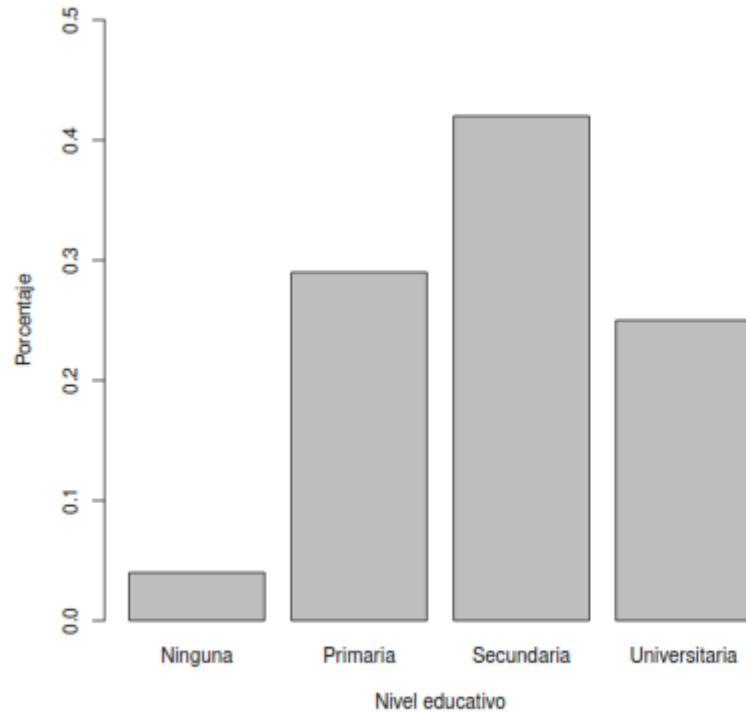


Figura: Diagrama de Barras

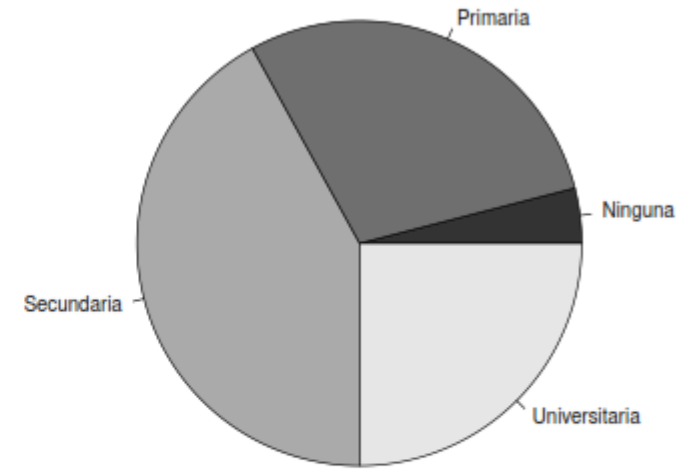


Figura: Diagrama de pastel

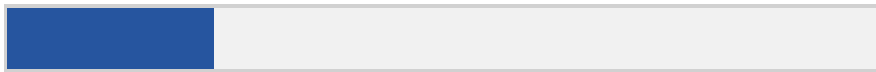
Variables cualitativas o cuantitativas discretas

¿Tiene usted ya definido por quién votará en las elecciones legislativas de este domingo?

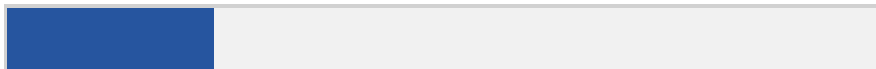
Sí (2,204 votos) 52.14%



No (1,008 votos) 23.85%



No va a votar (1,015 votos) 24%



- ¿Cuál es la variable?
- ¿Cuál es el tipo de variable y escala de medición?
- ¿Cuántas personas respondieron la encuesta?
- ¿Más de la mitad de las personas tienen definido por quién votará?
- ¿El 23.85% de las personas no va a votar?

Tablas bivariadas para variables cualitativas

Se analiza la relación entre la carrera que realizan los estudiantes y el sistema de transporte utilizado para llegar a la universidad ICESI

Frecuencia absoluta

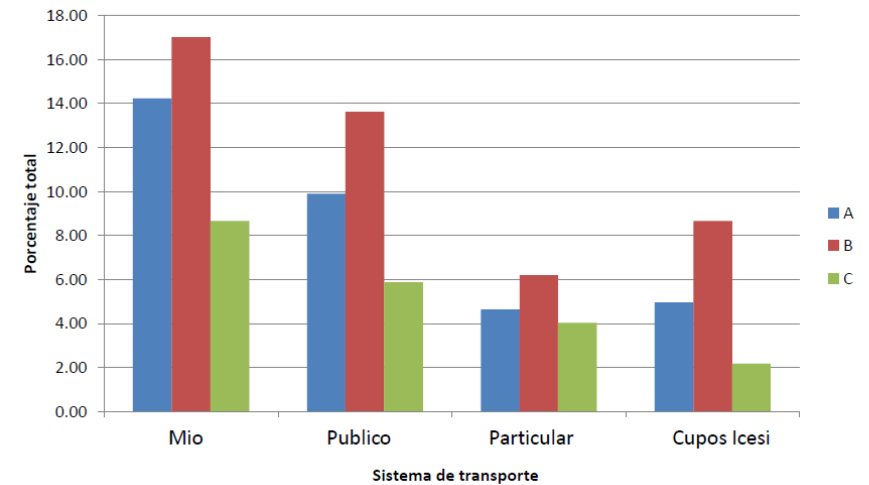
Distribución marginal de fila

Porcentaje del total

Distribución marginal de columna

Transporte	Carrera			Total
	A	B	C	
Mio	46	55	28	129
	14.2	17.0	8.7	39.9
Publico	32	44	19	95
	9.9	13.6	5.9	29.4
Particular	15	20	13	48
	4.6	6.2	4.0	14.9
Cupos Icesi	16	28	7	51
	5.0	8.7	2.2	15.8
Total	109	147	67	323
	33.7	45.5	20.7	100.0

Grafica del estudiantes por carrera y sistema de transporte



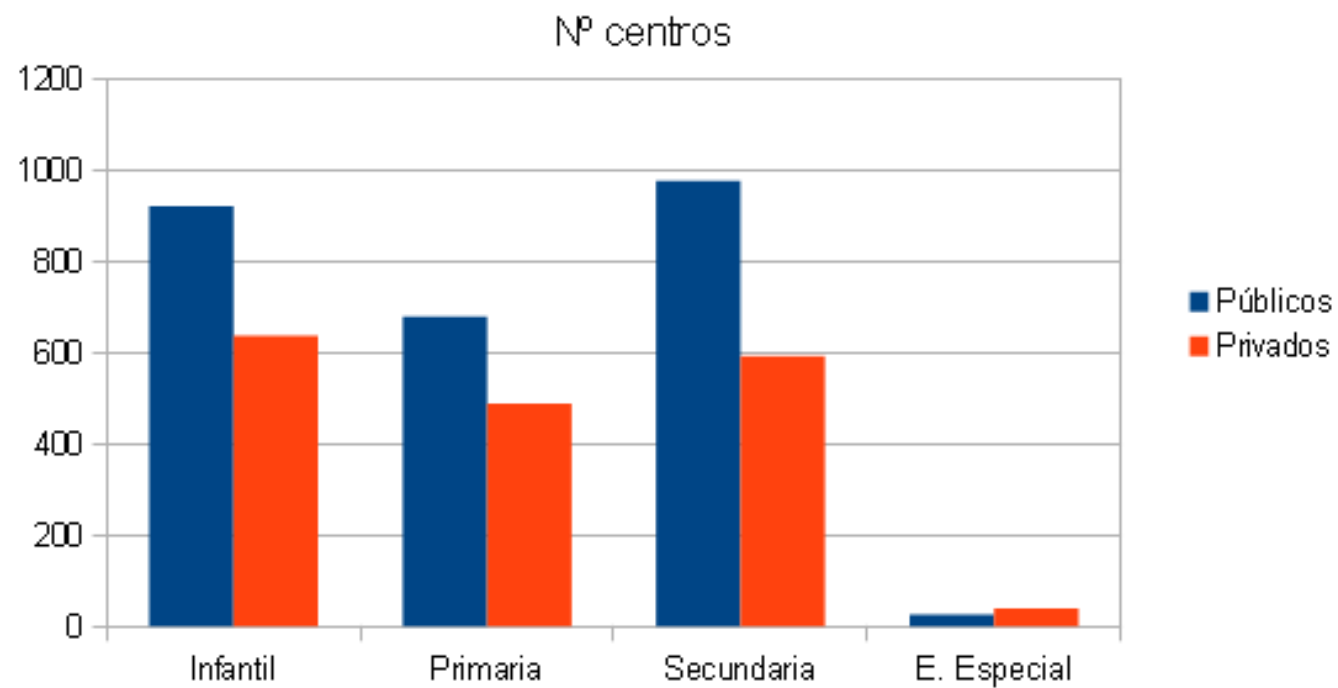
Distribución condicional
para la carrera A

Transporte	Carrera
	A
Mio	46
	42.2
Publico	32
	29.4
Particular	15
	13.8
Cupos Icesi	16
	14.7
Total	109
	100

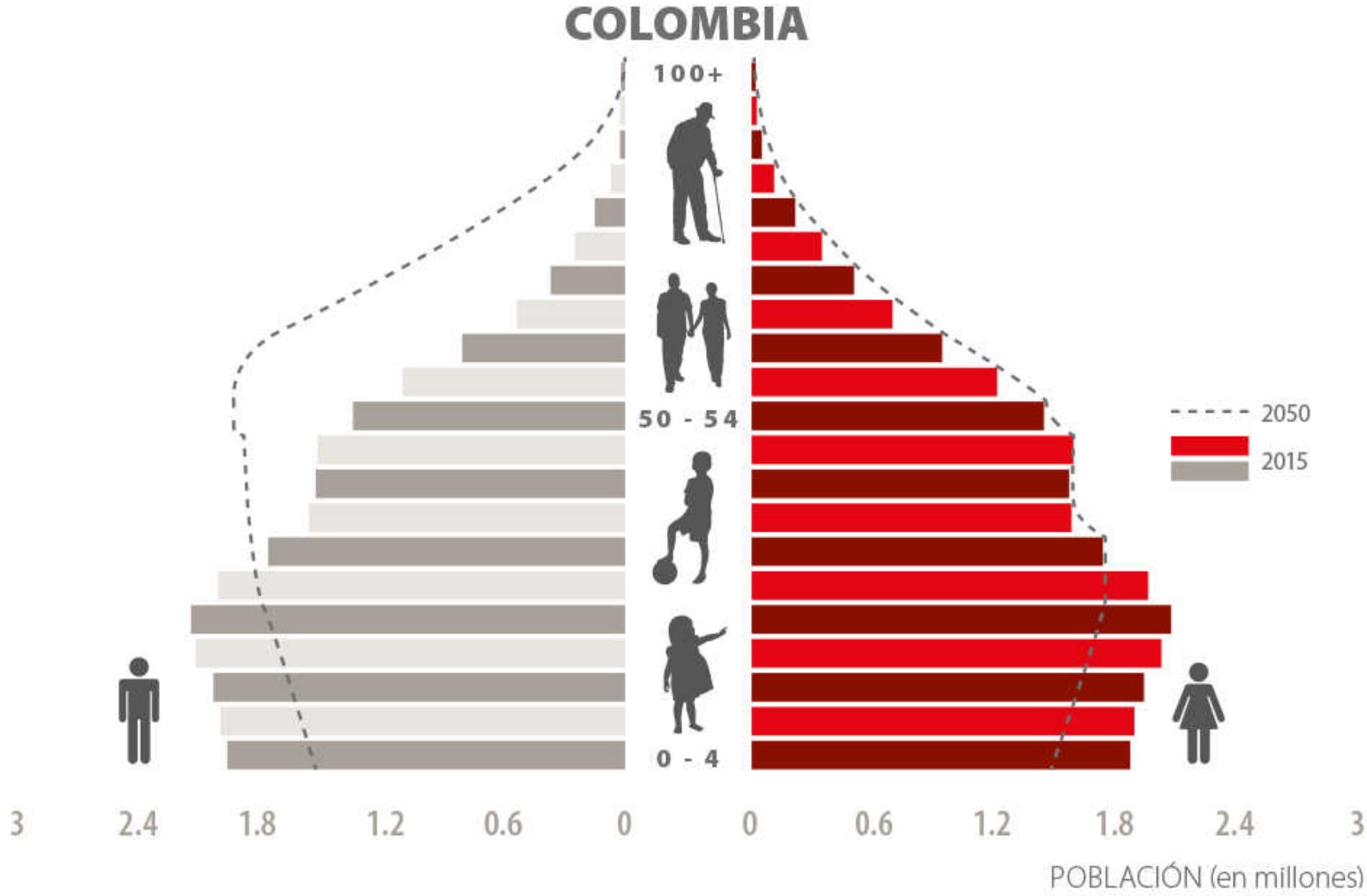
Distribución condicional para el
transporte MIO

Transporte	Carrera			Total
	A	B	C	
Mio	46	55	28	129
	35.7	42.6	21.7	100

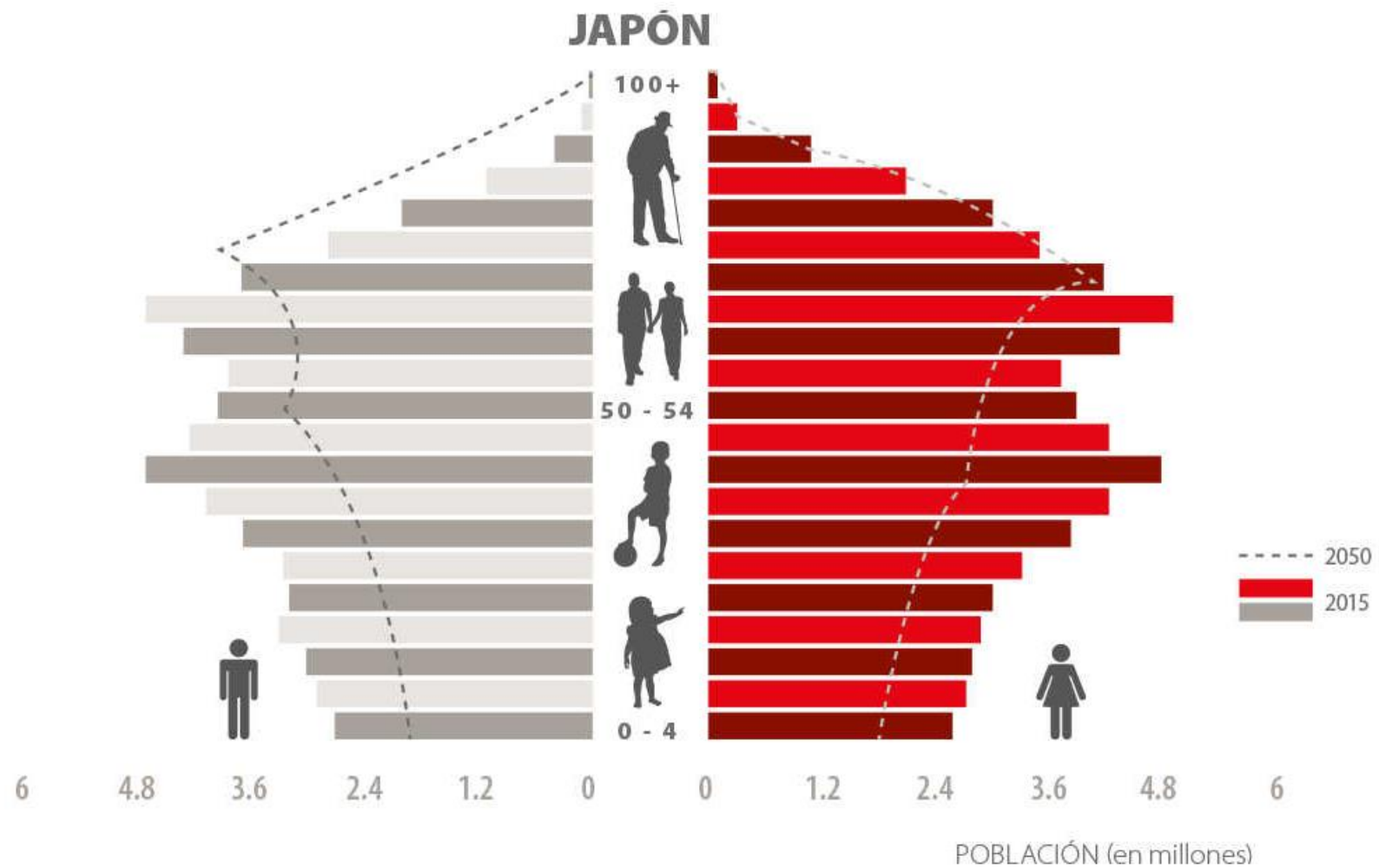
Diagramas de Barras



Pirámides Poblacionales



Pirámides Poblacionales



Un gráfico vale más que mil palabras!

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

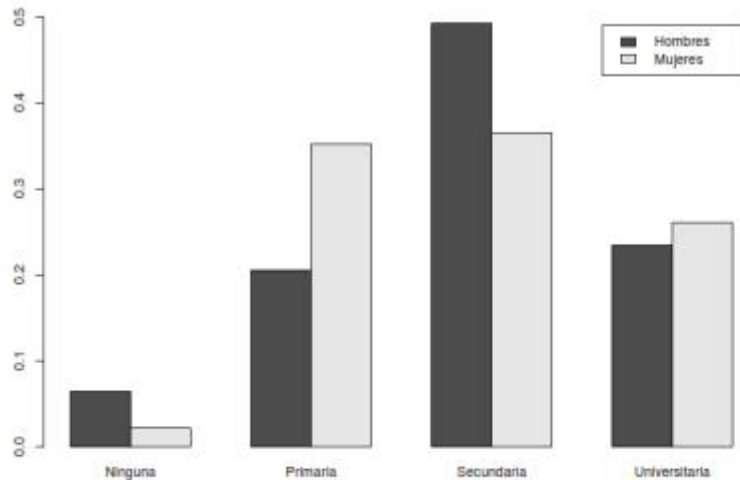


Figura: Buena representación

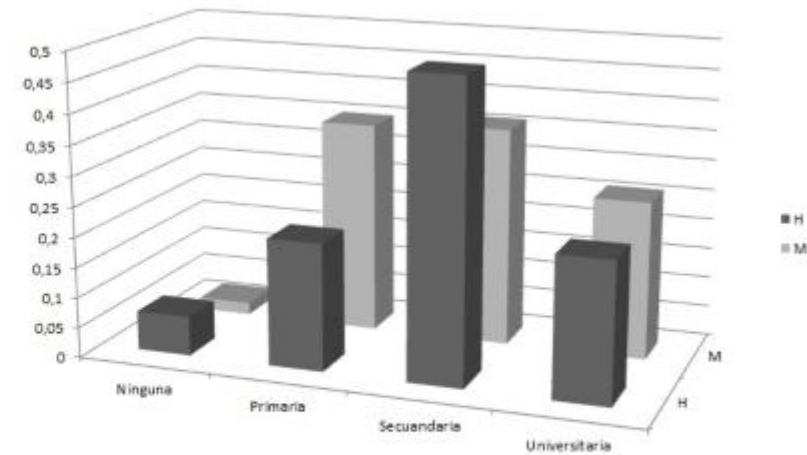
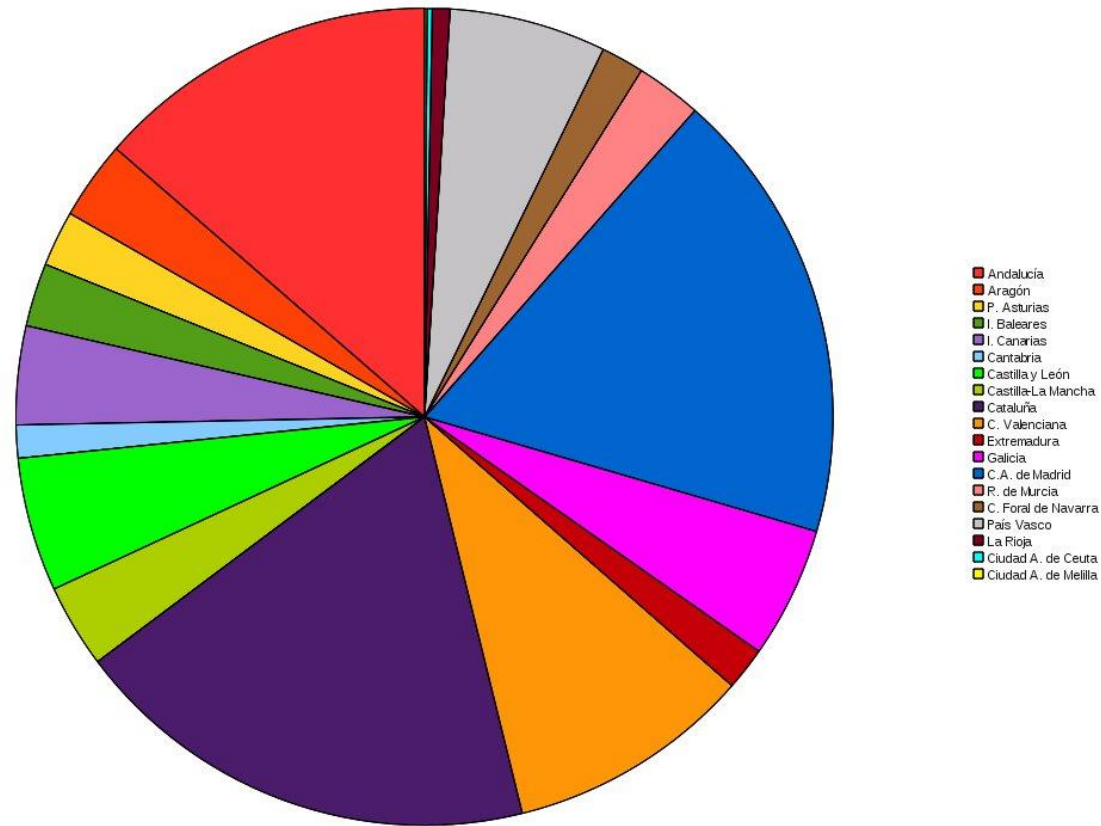


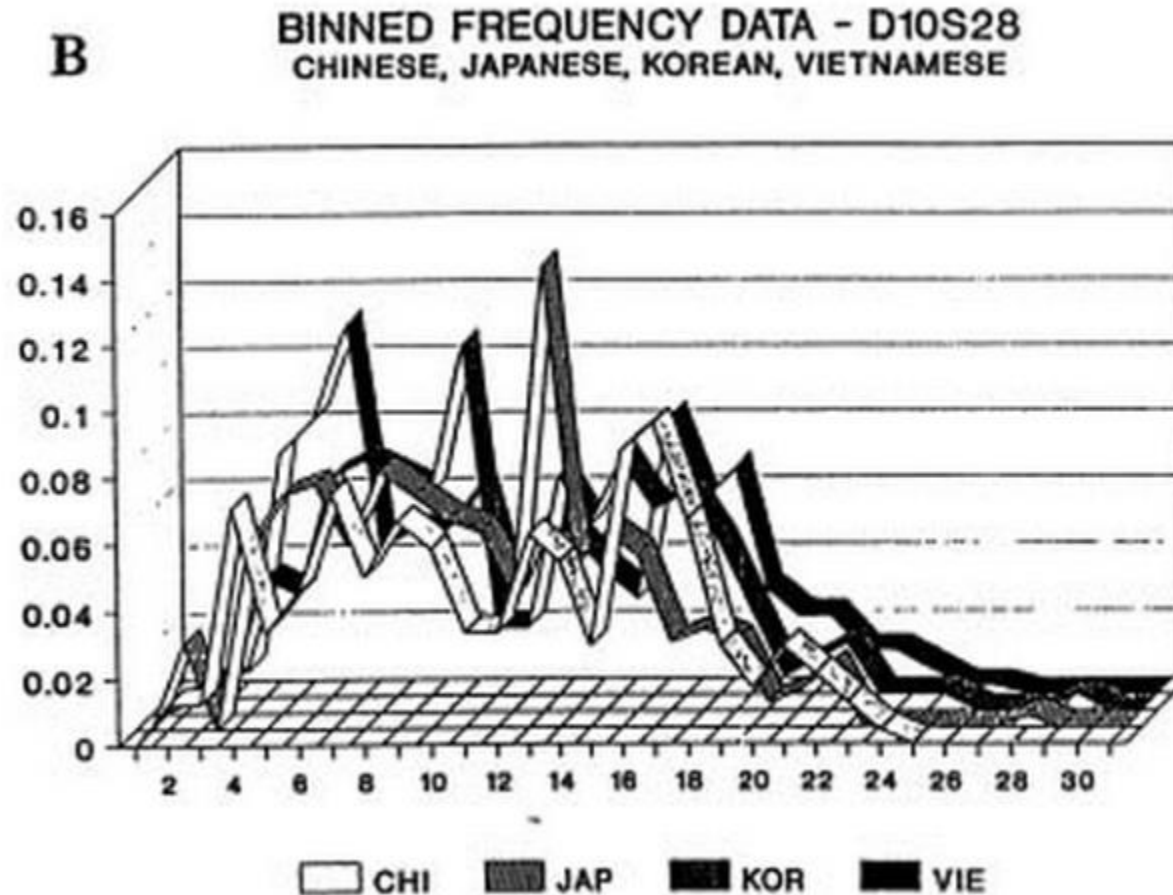
Figura: Mala representación

Un gráfico vale más que mil palabras!

Aportación autonómica al PIB(%)

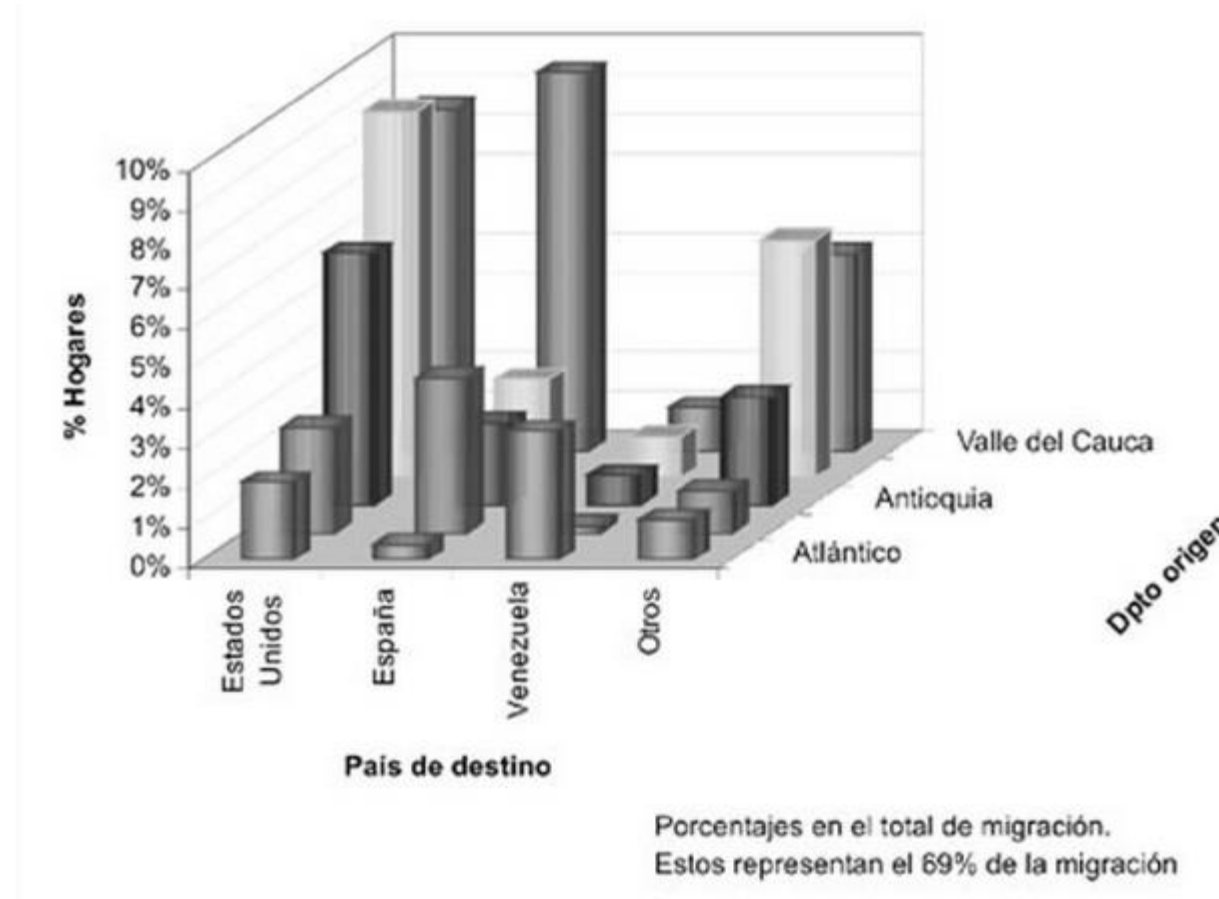


Un gráfico vale más que mil palabras!



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

Un gráfico vale más que mil palabras!



“La perspectiva hace difícil la comparación de la altura de los cubos”

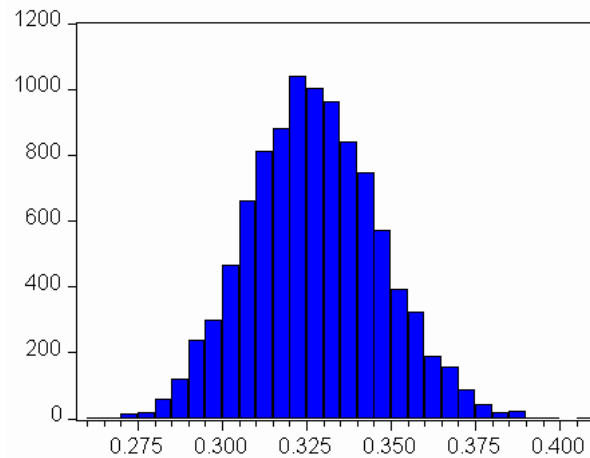
Para la próxima clase

Realizar la Tarea 1. Agrupar Datos cualitativos

Revisar la ruta de aprendizaje 2.

¿Ya empezaron alguno de los cursos de Python? Es hora

Tabulación y Representación grafica de datos cuantitativos



Datos Cuantitativos Discretos

Una planta industrial desea analizar el número de fallos en su sistema eléctrico durante un mes.

Para ello se registra el **número de fallos en cada transformador durante el mes.**

0, 1, 0, 2, 3, 1, 1, 2, 0, 1
3, 2, 1, 0, 2, 1, 0, 3, 2, 1

X_i (Valor observado)	Conteo	n_i (Frecuencia absoluta)
0		5
1	-	7
2		5
3		3
Total		20

TABLA DE FRECUENCIA
Número de fallos en cada transformador durante el mes

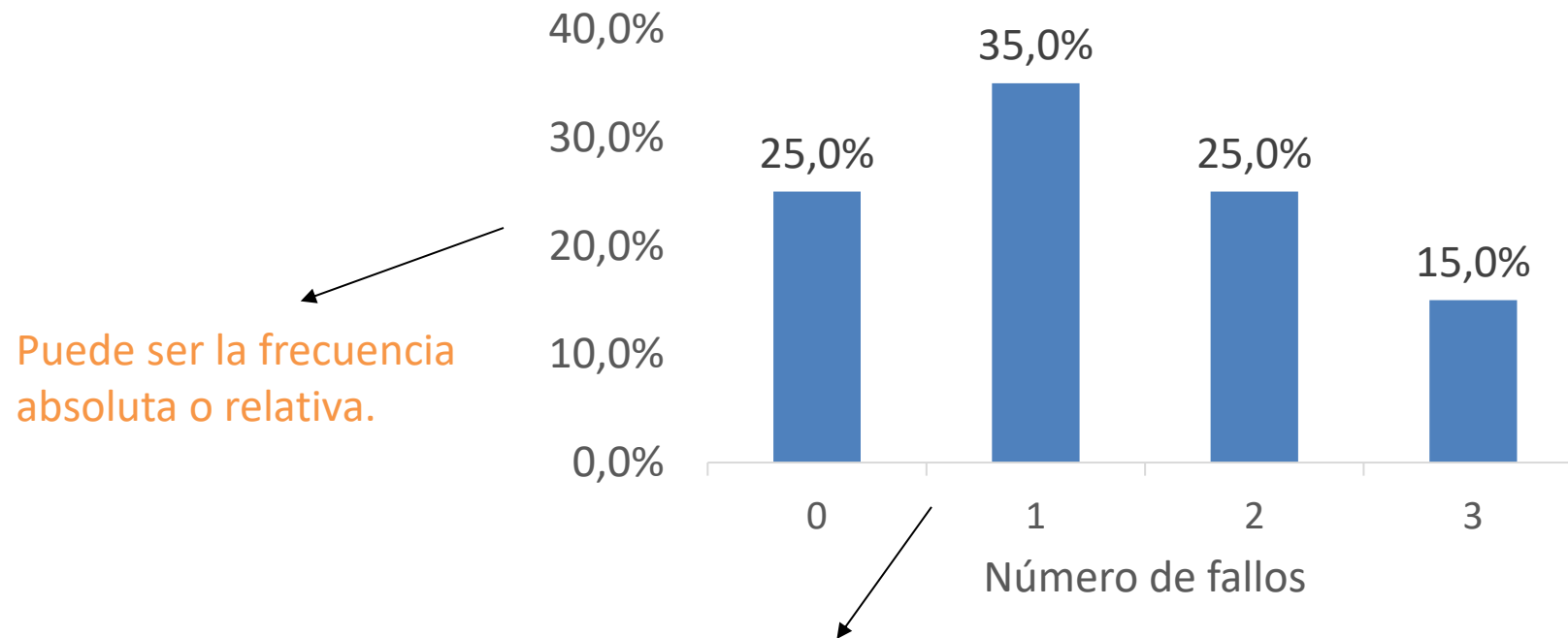
x_i Valor observado	n_i Frecuencia Absoluta	f_i Frecuencia Relativa	N_i Frecuencia Absoluta Acumulada	F_i Frecuencia Relativa Acumulada
0	5	0.25	5	0.25
1	7	0.35	12	0.60
2	5	0.25	17	0.85
3	3	0.15	20	1.0
Total	20	1.0		

REPRESENTACIÓN GRAFICA DE UNA DISTRIBUCIÓN DE FRECUENCIAS

Caso Discreto

Diagrama de Barras

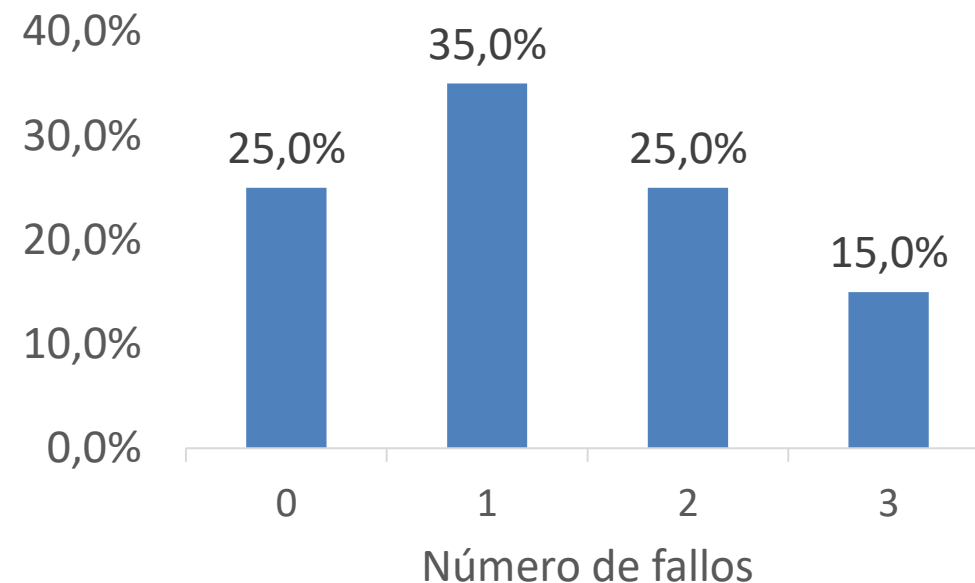
En el Eje horizontal se representan los valores que asume la variable y en el eje horizontal su frecuencia absoluta o relativa



Por ser una variable discreta las barras no deben juntarse

x_i	n_i	f_i
0	5	0.25
1	7	0.35
2	5	0.25
3	3	0.15
Total	30	1.0

x_i Valor observado	n_i Frecuencia Absoluta	f_i Frecuencia Relativa	N_i Frecuencia Absoluta Acumulada	F_i Frecuencia Relativa Acumulada
0	5	0.25	5	0.25
1	7	0.35	12	0.60
2	5	0.25	17	0.85
3	3	0.15	20	1.0
Total	20	1.0		



Preguntas

- ¿Cuál es el número de fallos más frecuente en los transformadores durante el mes?

El número de fallos más frecuente es **1**.

- ¿Qué porcentaje de transformadores no presentó ningún fallo durante el mes?

El **25%** de los transformadores no presentó ningún fallo.

- ¿Cuál es la frecuencia relativa de los transformadores que tuvieron 2 fallos?

La frecuencia relativa de los transformadores con 2 fallos es del **25%**.

- ¿Cuántos transformadores tuvieron al menos 2 fallos?

8 transformadores tuvieron al menos 2 fallos (sumando los que tuvieron 2 o 3 fallos)

Ejercicio

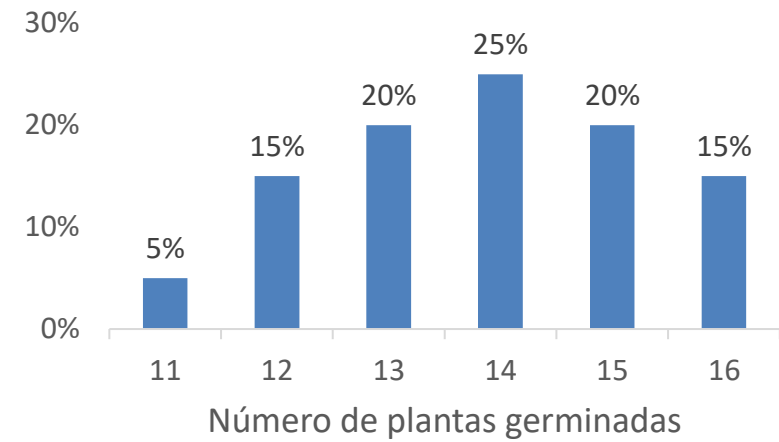
Se desea analizar la distribución el **número de plantas germinadas** por metro cuadrado en 20 parcelas diferentes después de un proceso de siembra.

11, 12, 12, 12, 13, 13, 13, 13, 14, 14,
14, 14, 14, 15, 15, 15, 15, 16, 16, 16

1. Construya su respectiva distribución de frecuencias. (simples y acumuladas).
2. Realice el grafico para la frecuencia relativa simple.
3. ¿Cuál es el número de plantas germinadas más frecuente en las parcelas?
4. ¿Qué porcentaje de parcelas tienen entre 13 y 15 plantas germinadas por metro cuadrado?
5. ¿Cuántas parcelas tienen más de 14 plantas germinadas por metro cuadrado?

- **Variable:** Número de plantas germinadas.

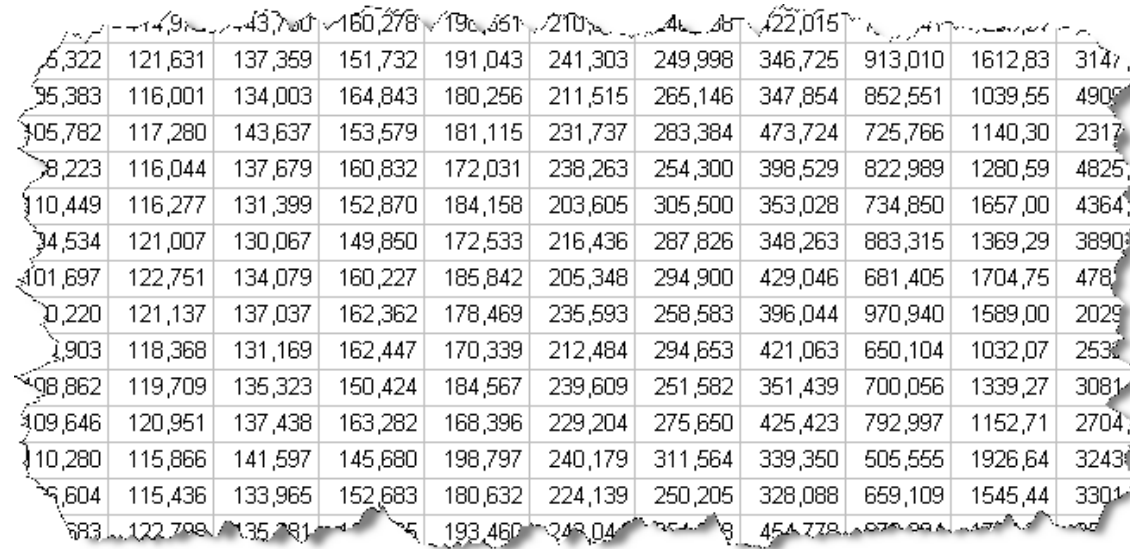
X	ni	fi	Ni	Fi
11	1	5%	1	5%
12	3	15%	4	20%
13	4	20%	8	40%
14	5	25%	13	65%
15	4	20%	17	85%
16	3	15%	20	100%
Total	20	100.0%		



- ¿Cuál es el número de plantas germinadas más frecuente en las parcelas? **14**
- ¿Qué porcentaje de parcelas tienen entre 13 y 15 plantas germinadas por metro cuadrado? **65%**
- ¿Cuántas parcelas tienen más de 14 plantas germinadas por metro cuadrado? **7**

Datos Cuantitativos Continuos

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



14,9	43,7	160,278	190,361	210,1	4,4	46,2	422,015				
6,322	121,631	137,359	151,732	191,043	241,303	249,998	346,725	913,010	1612,83	3147,1	
95,383	116,001	134,003	164,843	180,256	211,515	265,146	347,854	852,551	1039,55	4909,1	
105,782	117,280	143,637	153,579	181,115	231,737	283,384	473,724	725,766	1140,30	2317,1	
8,223	116,044	137,679	160,832	172,031	238,263	254,300	398,529	822,989	1280,59	4825,1	
110,449	116,277	131,399	152,870	184,158	203,605	305,500	353,028	734,850	1657,00	4364,1	
14,534	121,007	130,067	149,850	172,533	216,436	287,826	348,263	883,315	1369,29	3890,1	
101,697	122,751	134,079	160,227	185,842	205,348	294,900	429,046	681,405	1704,75	4781,1	
10,220	121,137	137,037	162,362	178,469	235,593	258,583	396,044	970,940	1589,00	2029,1	
1,903	118,368	131,169	162,447	170,339	212,484	294,653	421,063	650,104	1032,07	2530,1	
108,862	119,709	135,323	150,424	184,567	239,609	251,582	351,439	700,056	1339,27	3081,1	
109,646	120,951	137,438	163,282	168,396	229,204	275,650	425,423	792,997	1152,71	2704,1	
110,280	115,866	141,597	145,680	198,797	240,179	311,564	339,350	505,555	1926,64	3243,1	
15,604	115,436	133,965	152,683	180,632	224,139	250,205	328,088	659,109	1545,44	3301,1	
1,983	122,798	135,281	155,125	193,460	248,04	255,158	454,778	870,994	1470,1	3005,1	

“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?**

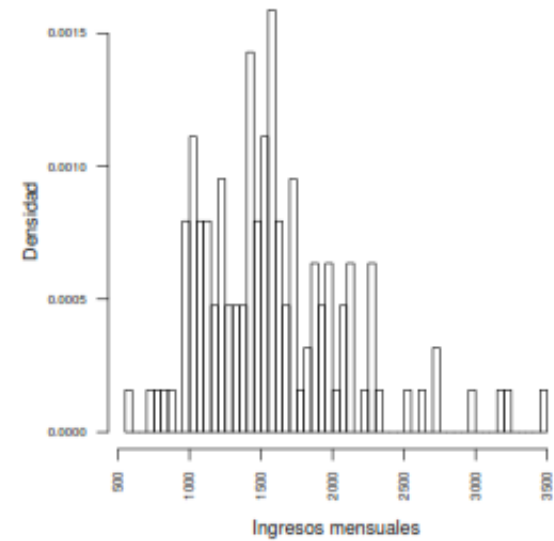
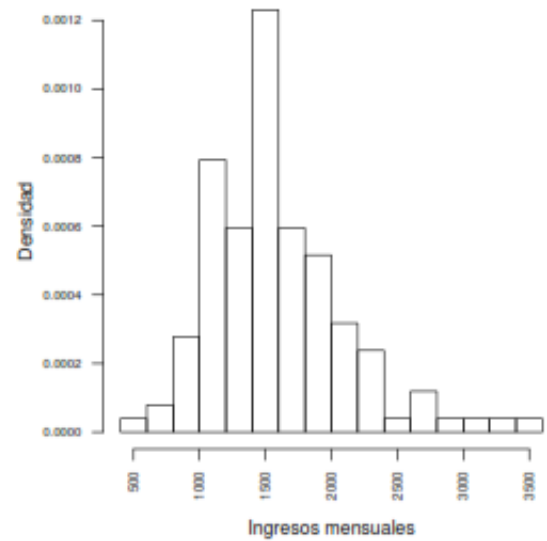
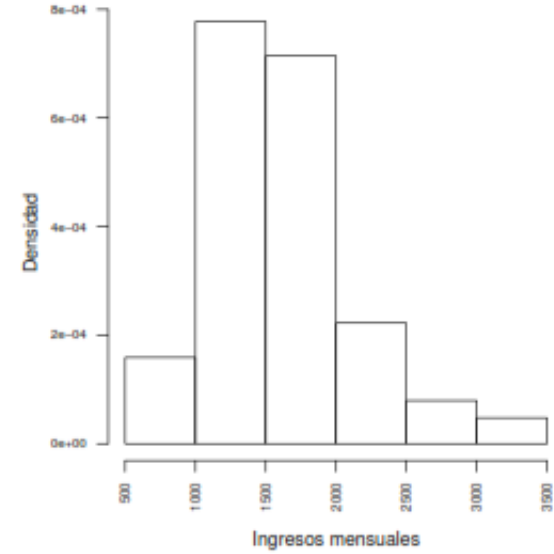
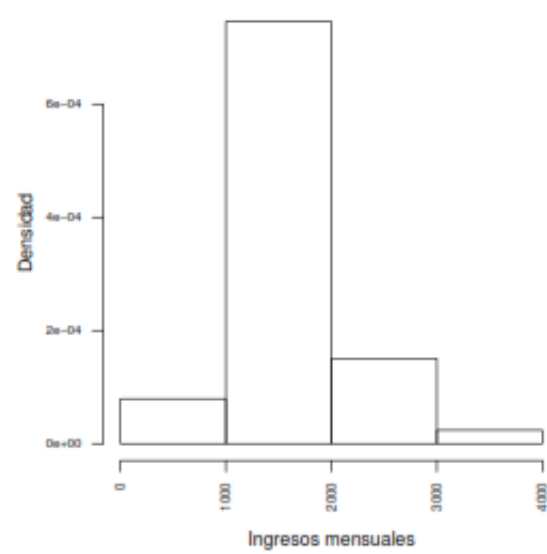
Siempre que se agrupan los datos en intervalos de clase se produce pérdida de información.

- 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:

$$2^k > n$$

$$1 + 3.3 \log_{10}(n)$$

¿Cuántas clases utilizar?



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	<u>4.5</u>	1.6
3.1	2.5	4.2	1.2	3.2
2.6	1.7	3.5	2.2	4.4
2.7	<u>0.3</u>	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

Pasos para construir una distribución de frecuencia en datos agrupados

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

$$2^k > n \rightarrow 2^6 = 64 > 50 \rightarrow \mathbf{k = 6}$$

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

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

$$Rango = Max(x_i) - Min(x_i)$$

$$R = 4.5 - 0.3 = \mathbf{4.2}$$

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



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

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]	4,4	6	0,12	50	1,00
	Total	50	1.0		

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

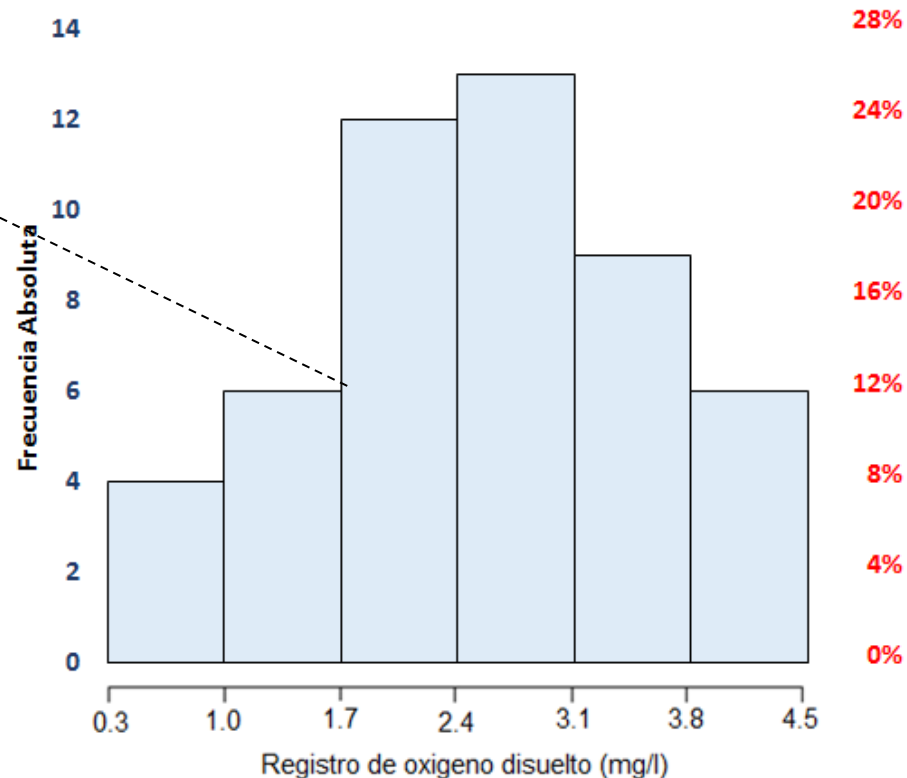
El 88% de las mediciones presentaron un registro de oxígeno disuelto menor o igual a 3.8 mg/l.

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

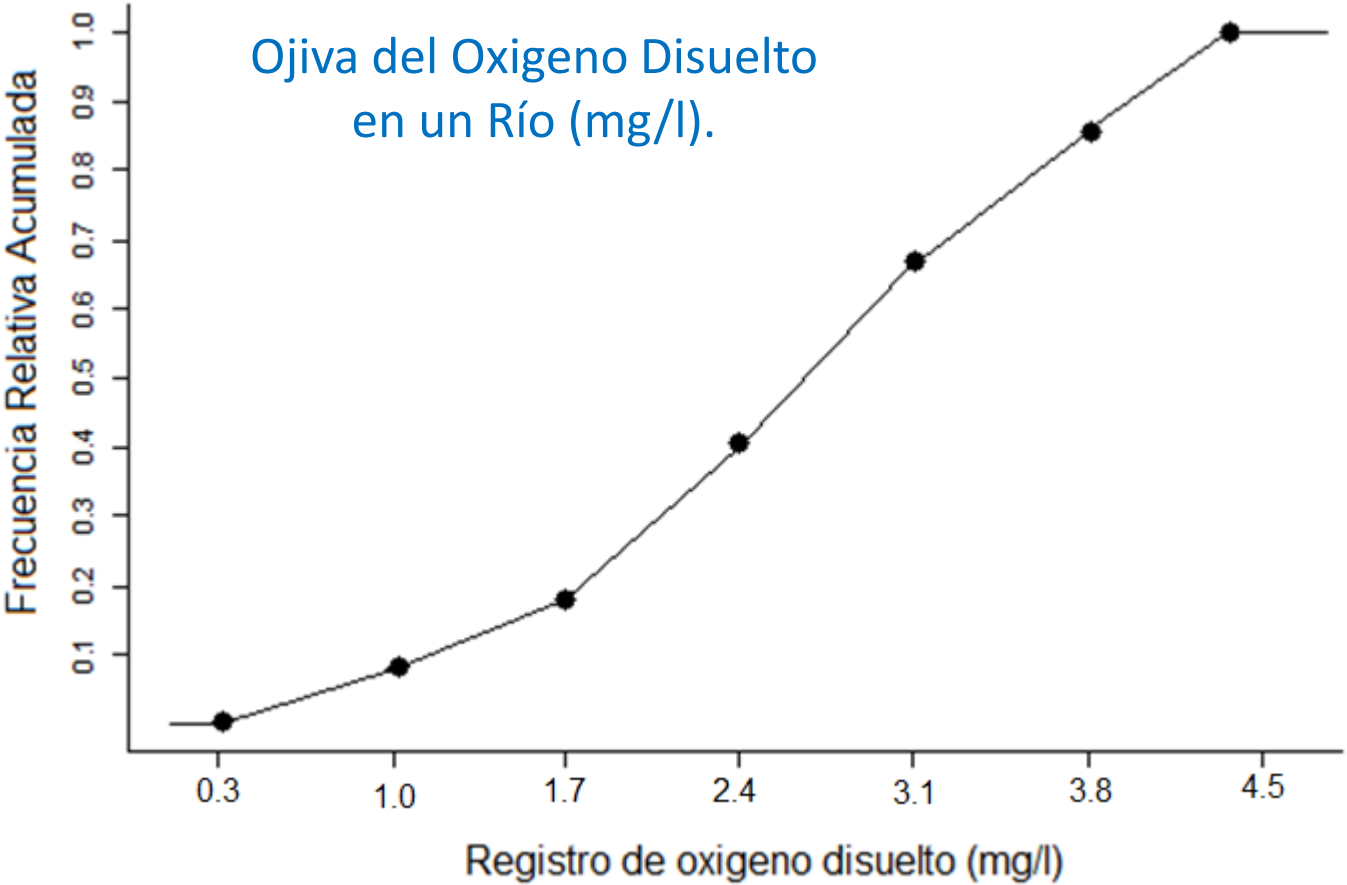
La barras se juntan por continuidad de la variable



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**.




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

Intervalos	N_i	F_i
[0.3 , 1.0]	4	0,08
(1.0 , 1.7]	10	0,20
(1.7 , 2.4]	22	0,44
(2.4 , 3.1]	35	0,70
(3.1 , 3.8]	44	0,88
(3.8 , 4.5]	50	1,00

Distribución de Frecuencia

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

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]	4,4	6	0,12	50	1,00
	Total	50	1.0		

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

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]	4,4	6	0,12	50	1,00
	Total	50	1.0		

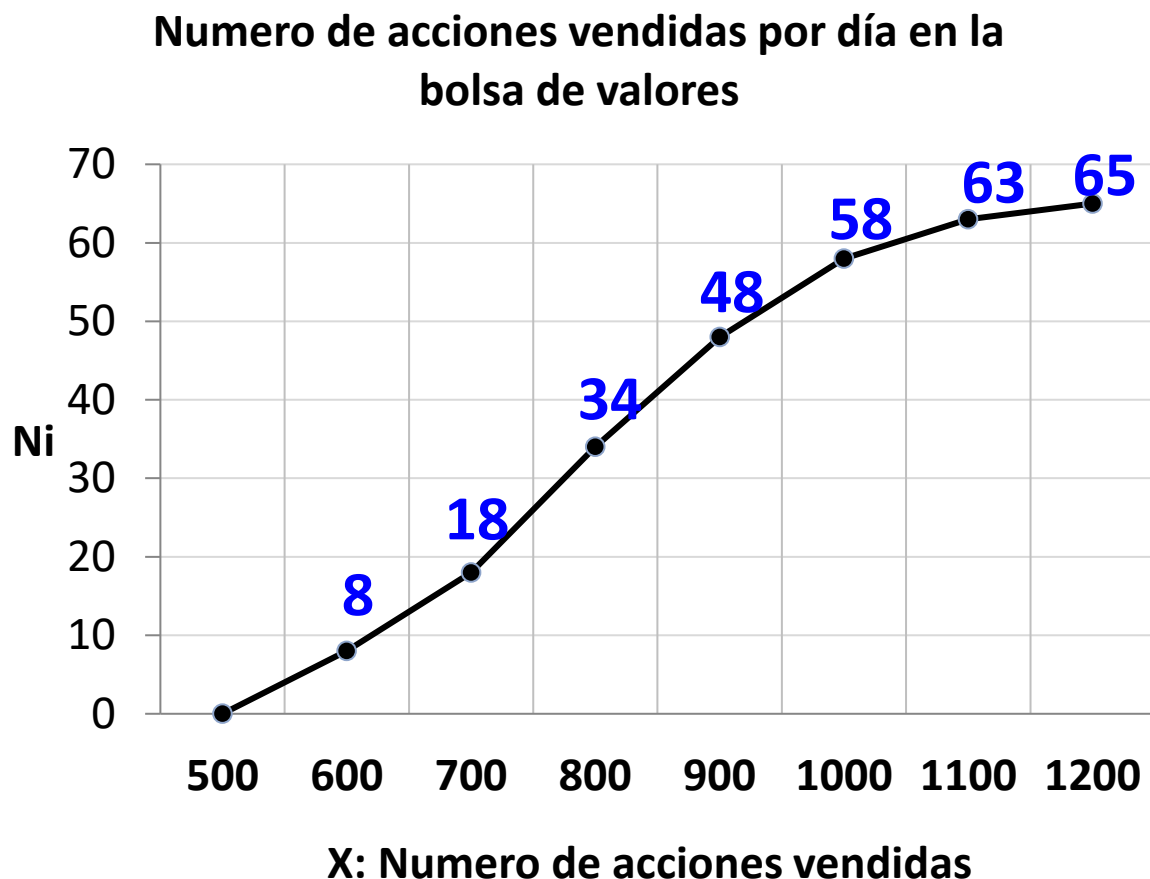
A partir de la tabla de frecuencias:

1. ¿Qué porcentaje de registros presentan niveles de OD superiores a 2.5 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?

$$x = L_{i-1} + \frac{(P-F(x_{i-1}))}{f_i} * C_i$$

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**.



- De acuerdo con la gráfica construya la **tabla de frecuencias** respectiva (trabaje con 2 decimales).
- Interprete los valores de:
 n_3 , N_4 , f_5 y F_6
- Presente gráficamente la **frecuencia relativa simple** del número de acciones de Ecopetrol vendidas por día.
- ¿**Qué porcentaje** de días presentan entre **750 y 950** acciones vendidas?
- ¿**A partir de qué número de acciones** vendidas se encuentra acumulado el **80%** de los datos?

Ejercicio

a. De acuerdo con la gráfica construya la **tabla de frecuencias** respectiva (trabaje con 2 decimales).

Li	Ls	Marca de clase	ni	fi	Ni	Fi
500	600	550	8	0.12	8	0.12
600	700	650	10	0.15	18	0.28
700	800	750	16	0.25	34	0.52
800	900	850	14	0.22	48	0.74
900	1000	950	10	0.15	58	0.89
1000	1100	1050	5	0.08	63	0.97
1100	1200	1150	2	0.03	65	1.00
Total			65	1.0		

X	F(X)
750	40.0%
950	81.5%
	41.5%

X	F(X)
940	80.0%

- b. Interprete los valores de: n_3 , N_4 , f_5 y F_6
- c. Presente gráficamente la **frecuencia relativa simple** del número de acciones de Ecopetrol vendidas por día.
- d. ¿**Qué porcentaje** de días presentan entre **750 y 950** acciones vendidas?
- e. ¿**A partir de qué número de acciones** vendidas se encuentra acumulado el **80%** de los datos?