

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
(Objetivo)

-----> Herramientas
(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?

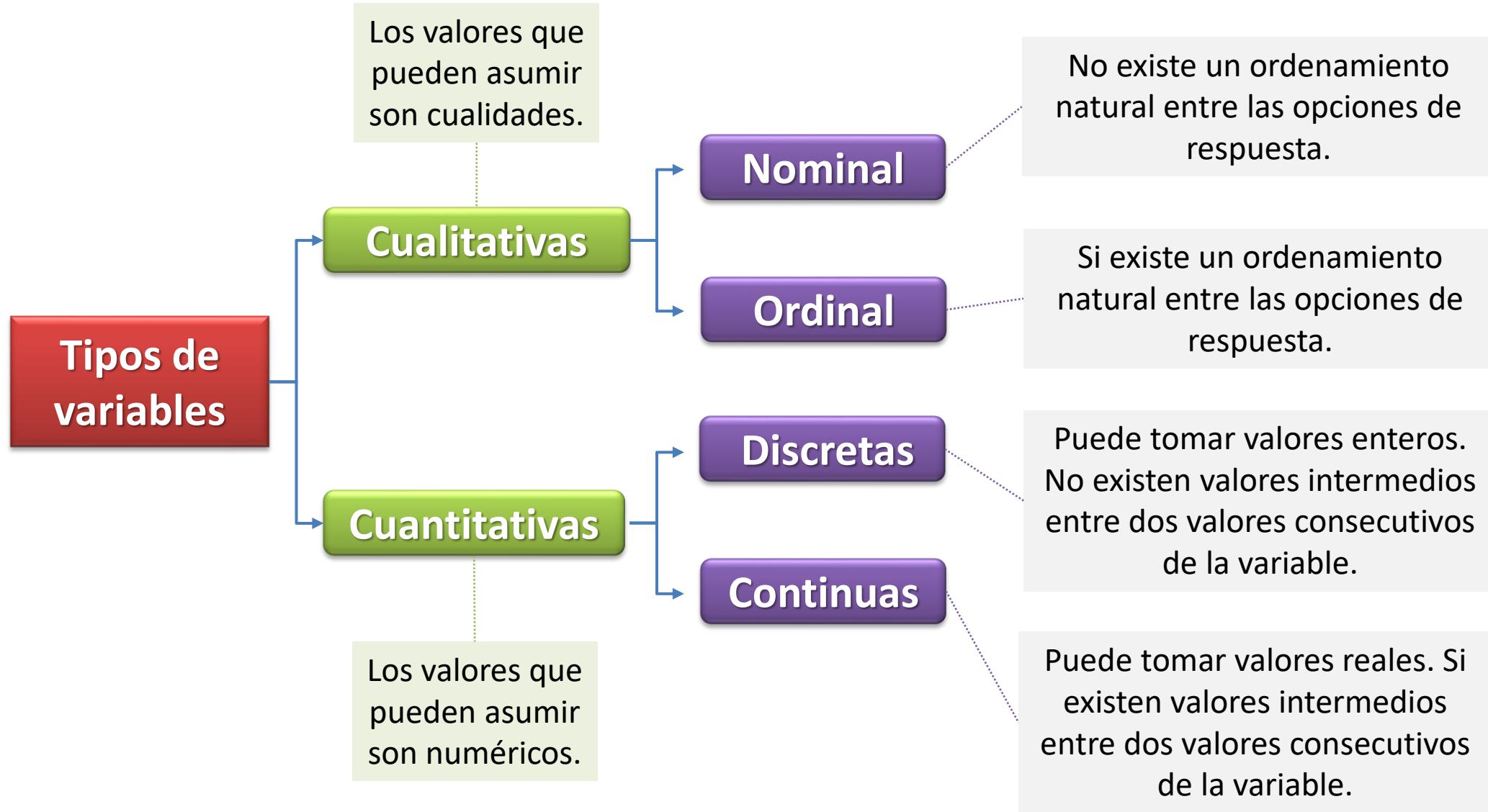
- EDAD
- PESO
- ESTATURA
- INGRESOS
- NRO. DE HIJOS

VARIABLES
CUANTITATIVAS



- SEXO
- USA LENTES
- ESTADO CIVIL
- COLOR DE OJOS
- NIVEL DE ESTUDIO

VARIABLES
CUALITATIVAS



Ejemplos:

Variable	Tipo de Variable
Concentración de oxígeno 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 Dispersion (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

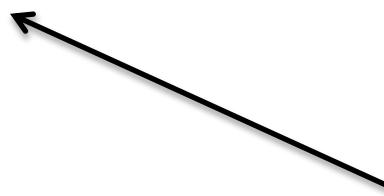
Variables Cualitativas / Cuantitativas discretas

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 NO Si

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



Muestra Bruta = Datos

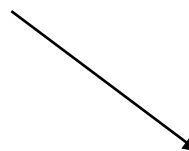
¿Que puede decir usted acerca de los resultados obtenidos?



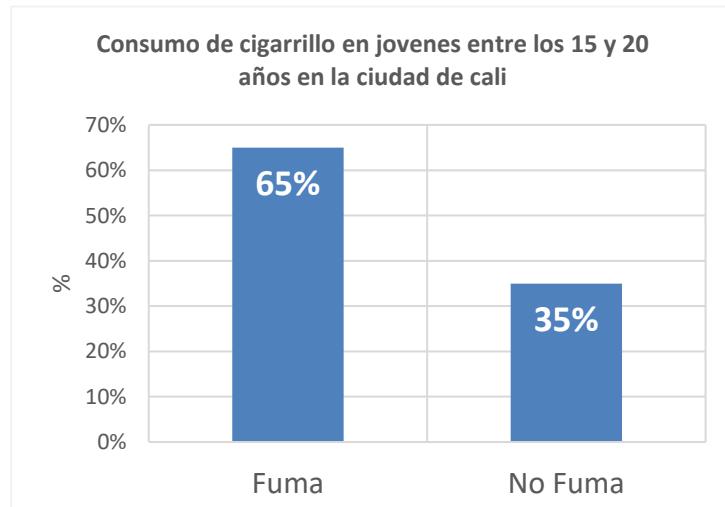
Variables cualitativas o cuantitativas discretas

Tablas de Frecuencia

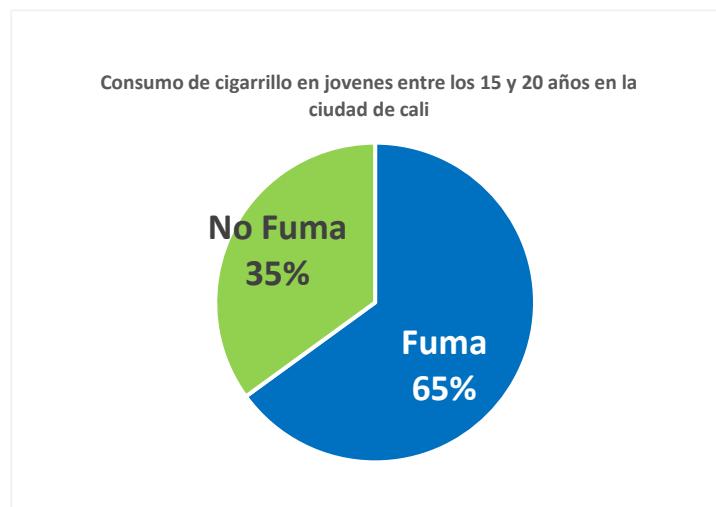
FUMA	Nro. Casos	%
SI	26	65%
NO	14	35%
Total	40	100%



Diagramas de Barra

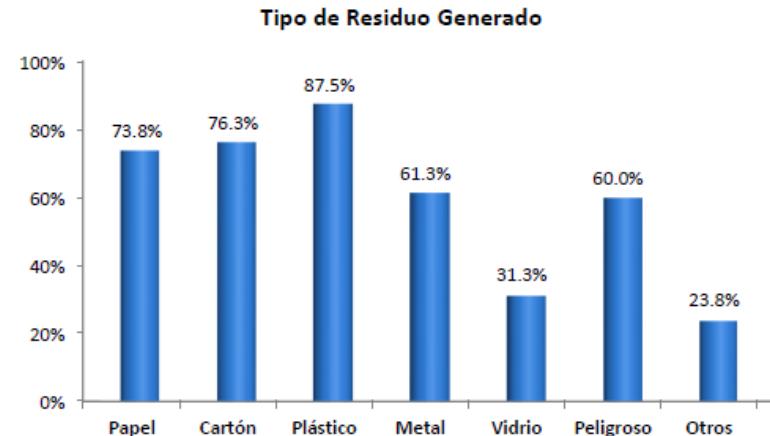


Diagramas de sectores



Funciones en R

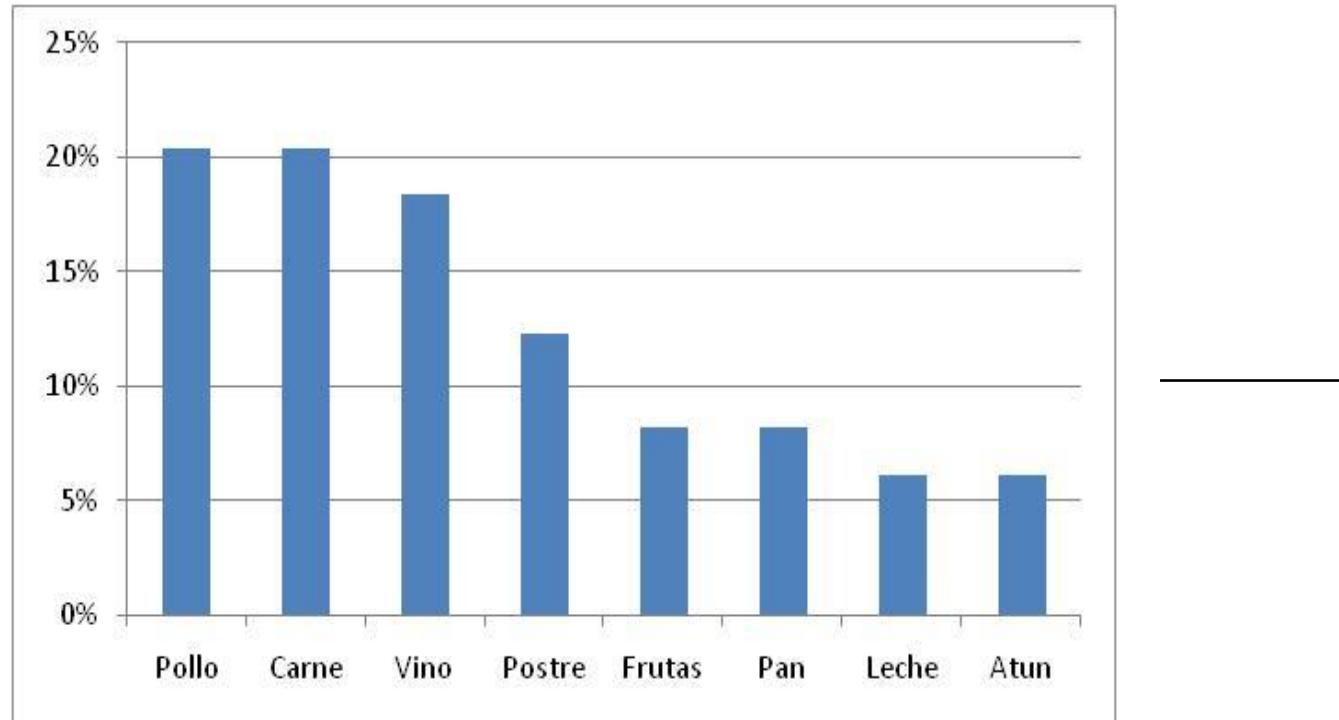
- `table(x)`, `prop.table(x, margin)`
- `barplot(table)`
- `pie(table)`



Características con muchas opciones de respuesta

Buscando la fuente de una Intoxicación grupal:

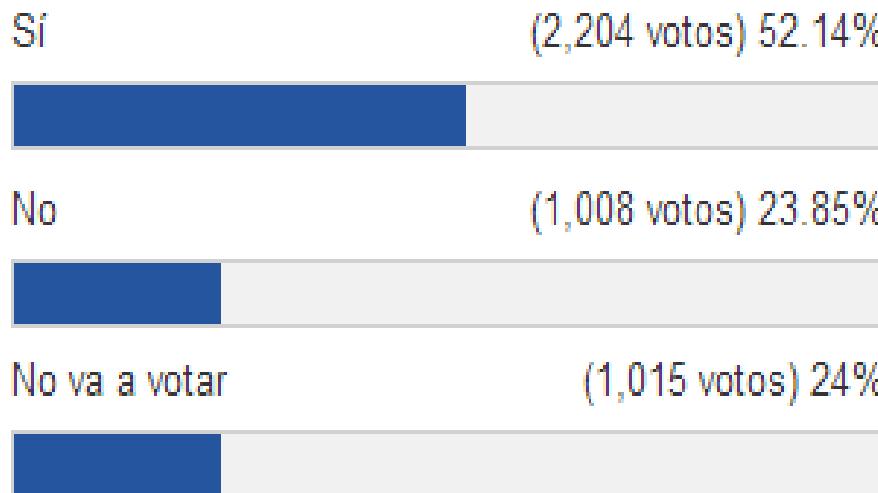
¿Consumió alguno de los siguientes alimentos?



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

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



- 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?

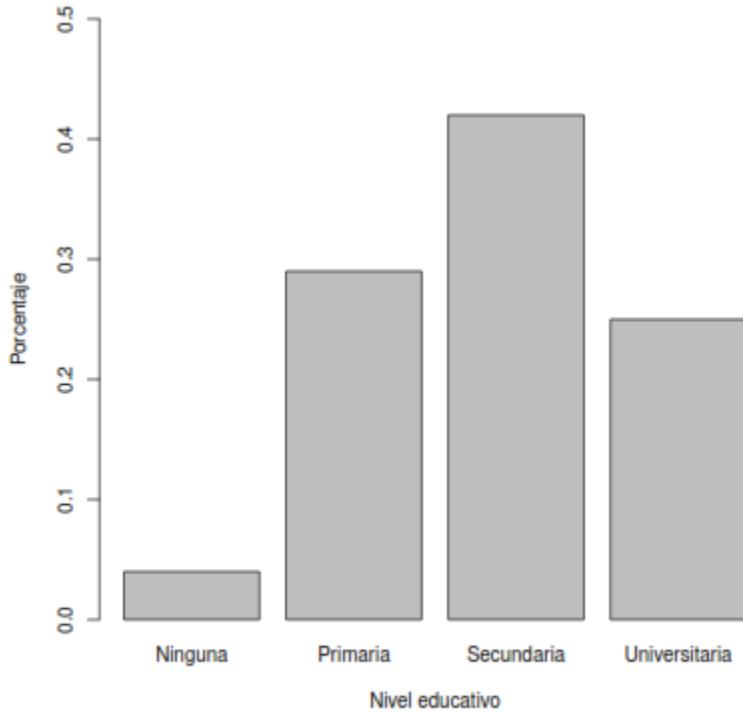


Figura: Diagrama de Barras

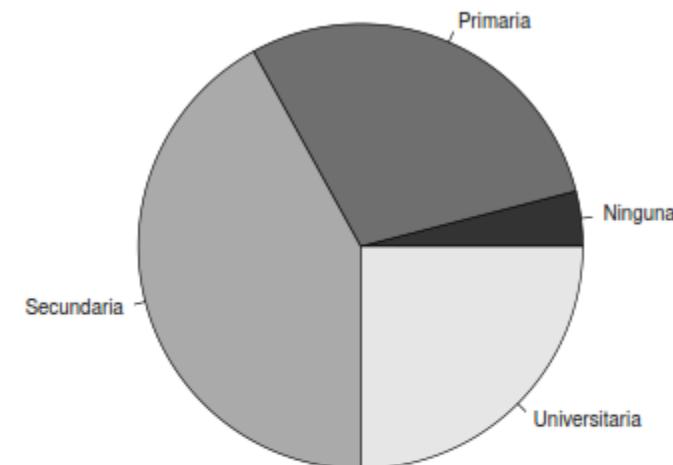
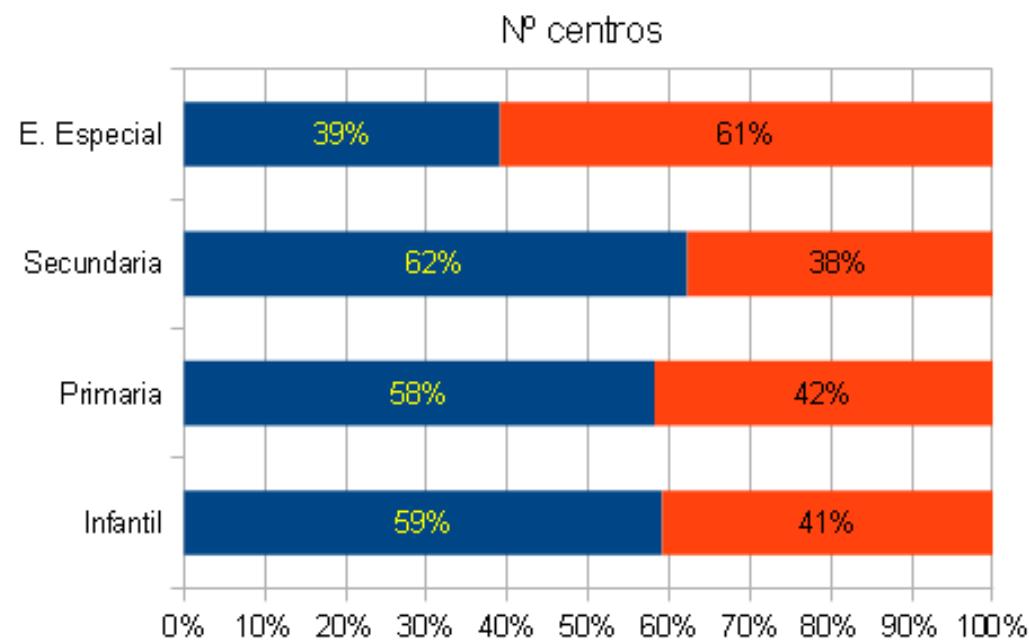
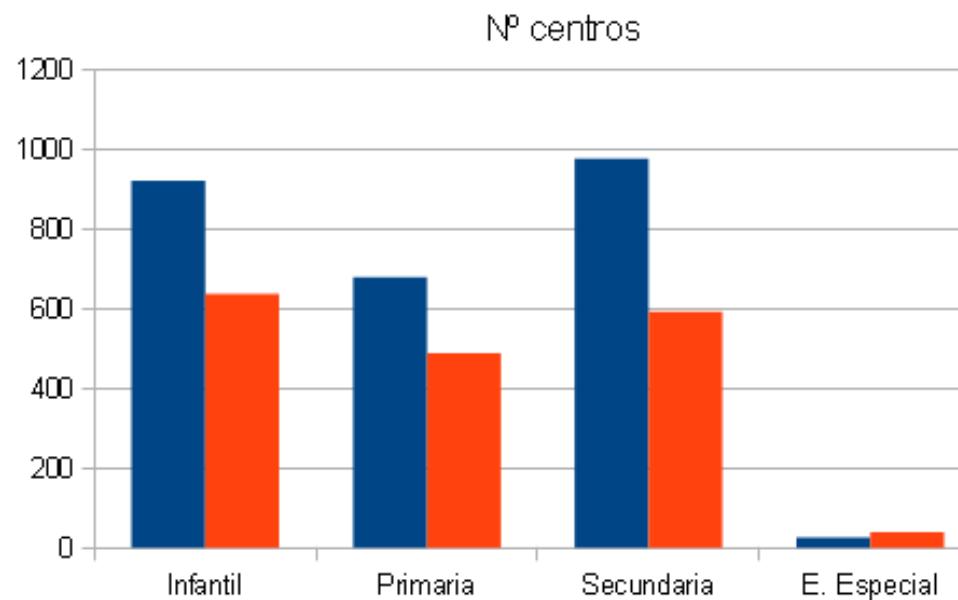
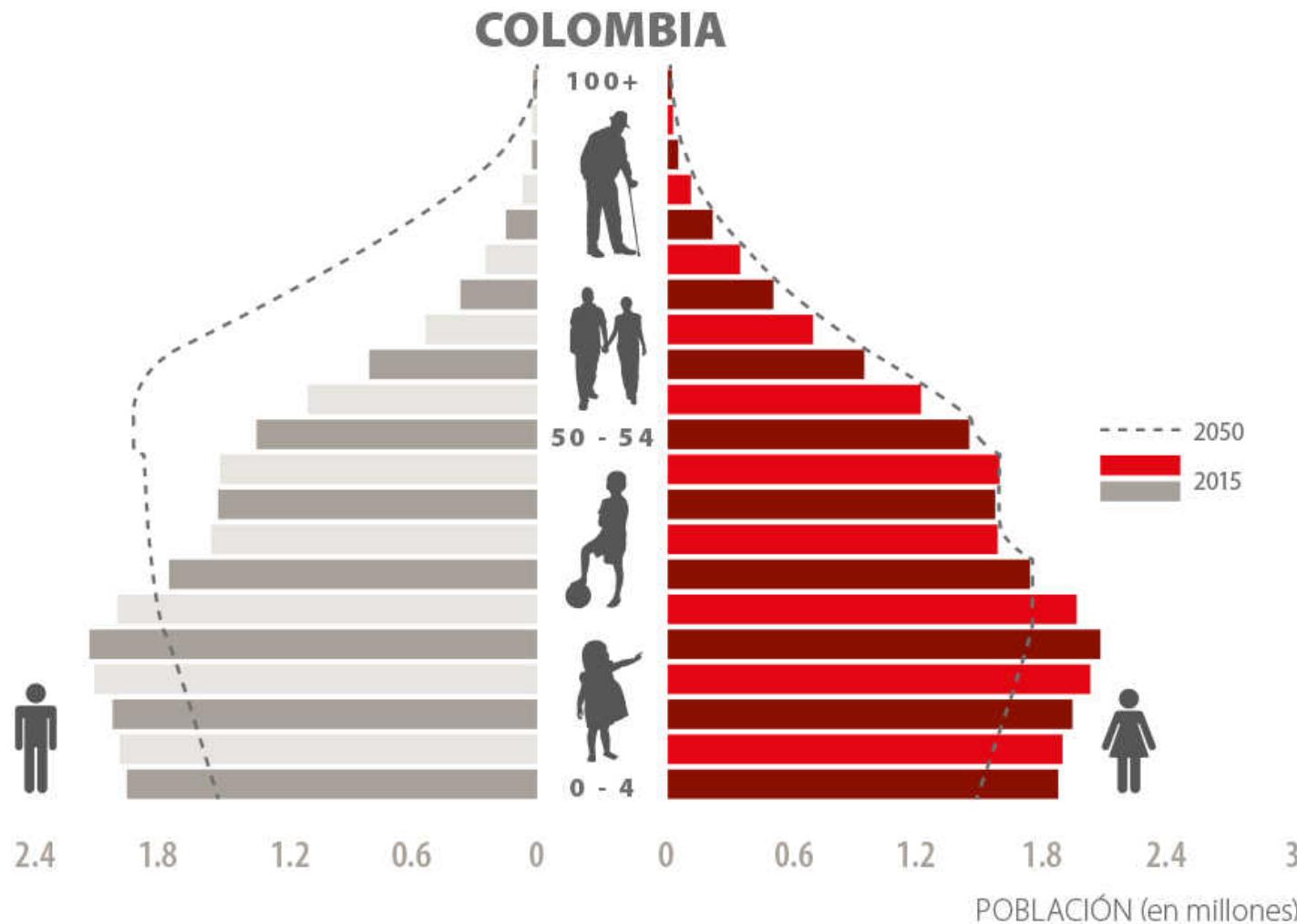


Figura: Diagrama de pastel

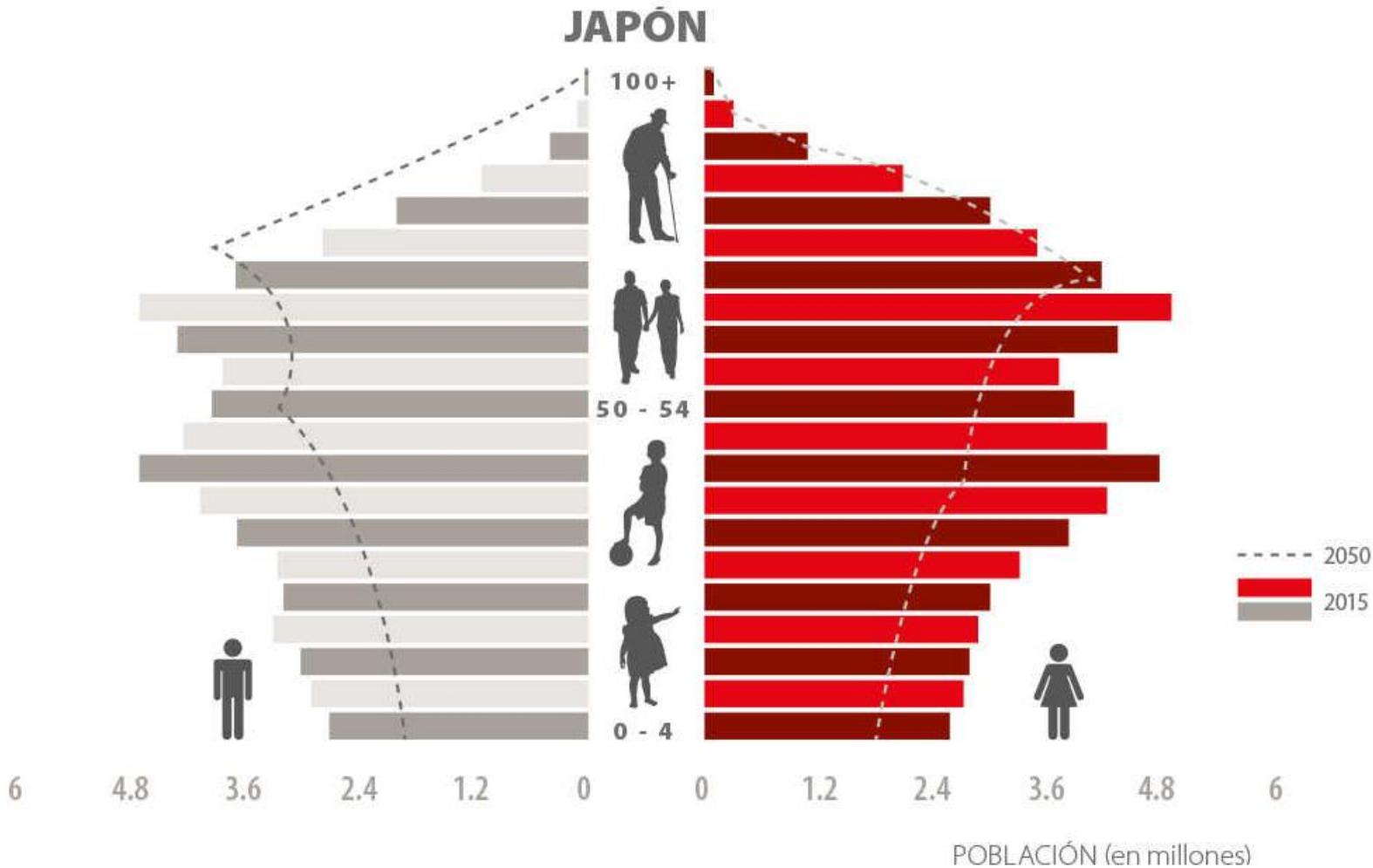
Diagramas de Barras



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?

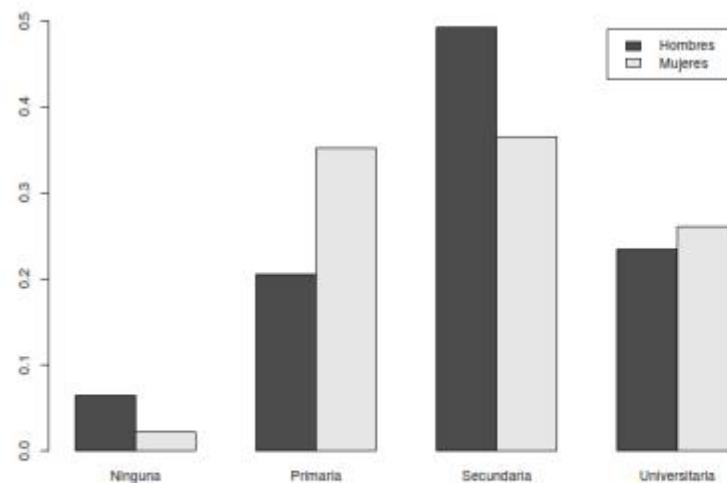


Figura: Buena representación

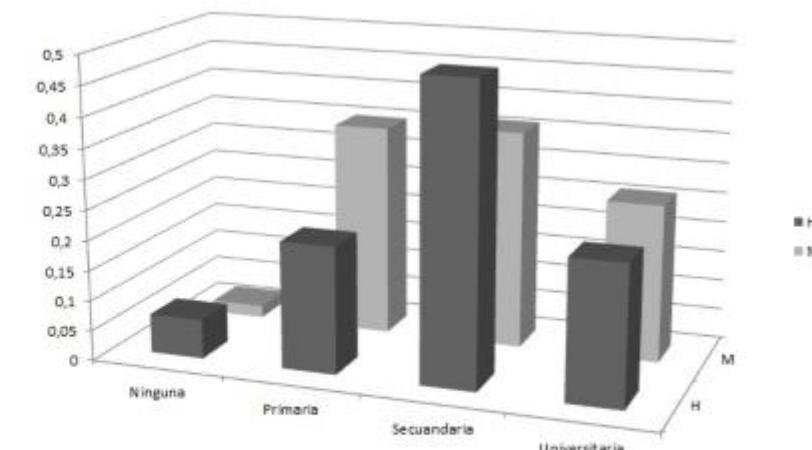
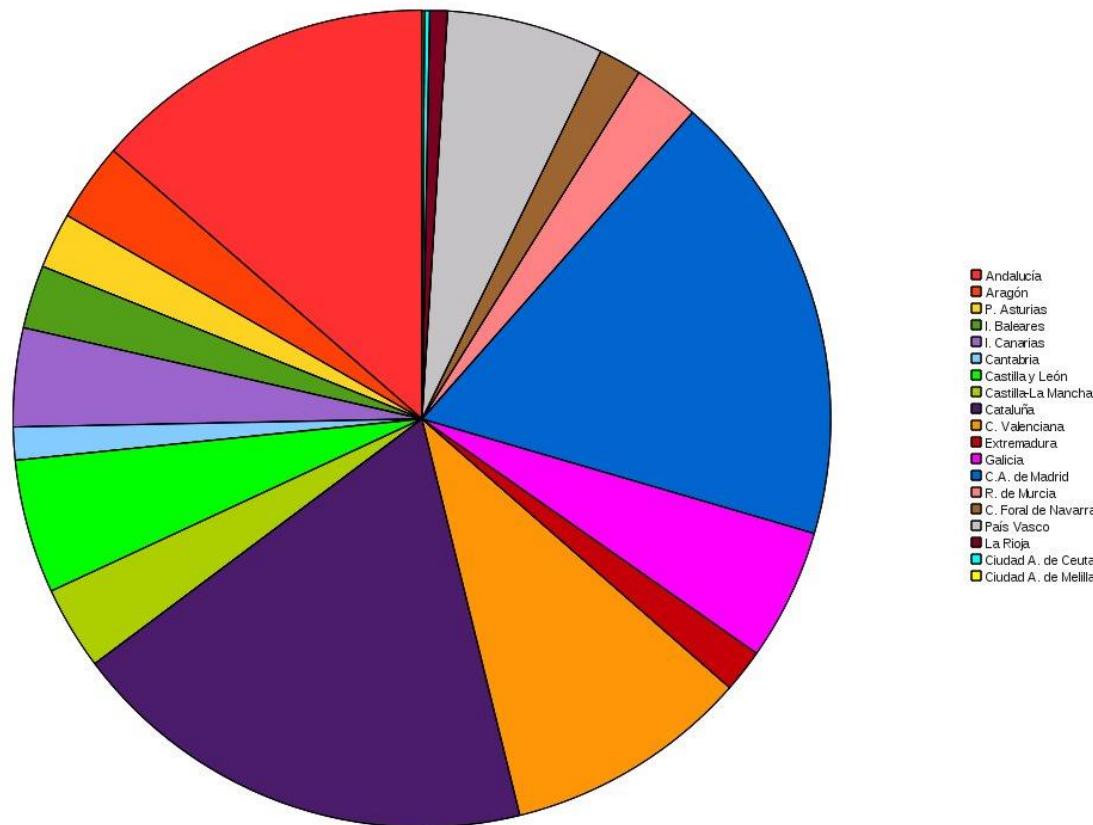


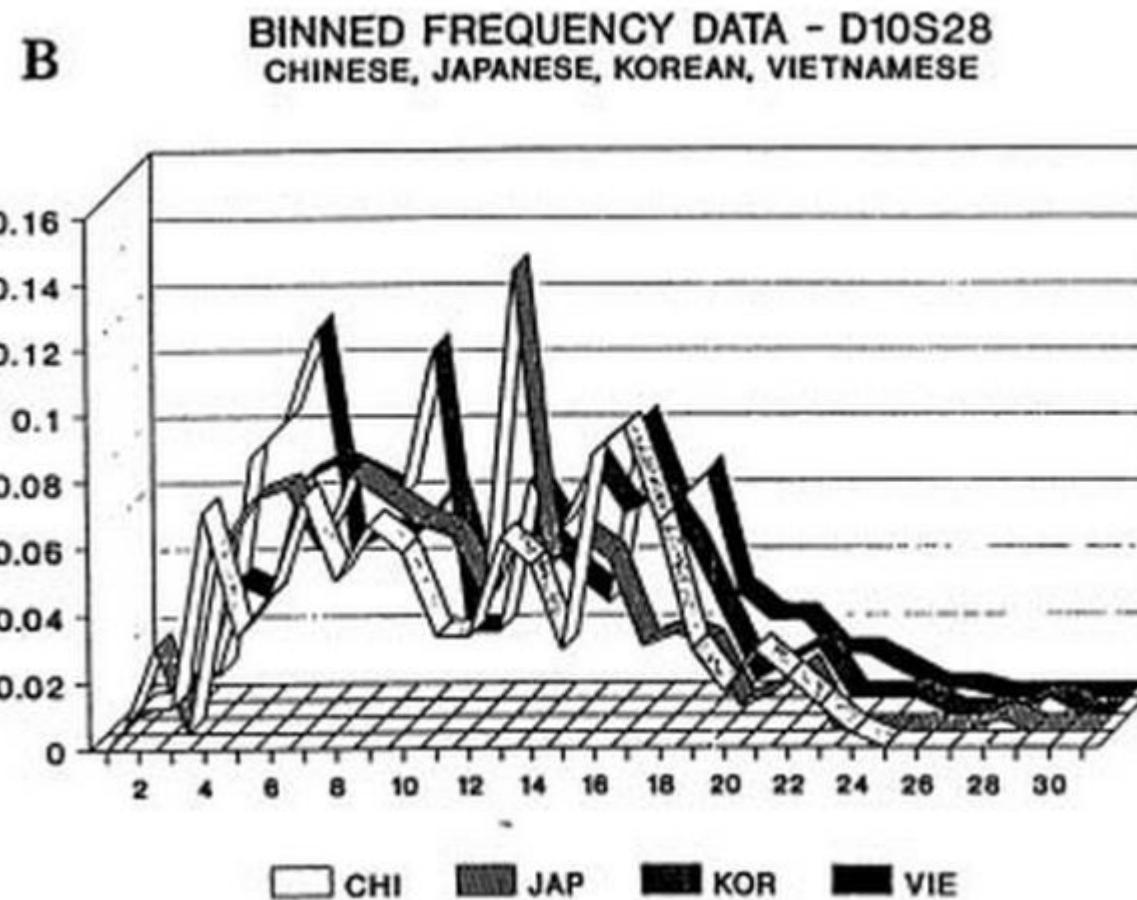
Figura: Mala representación

Un grafico vale más que mil palabras!

Aportación autonómica al PIB(%)

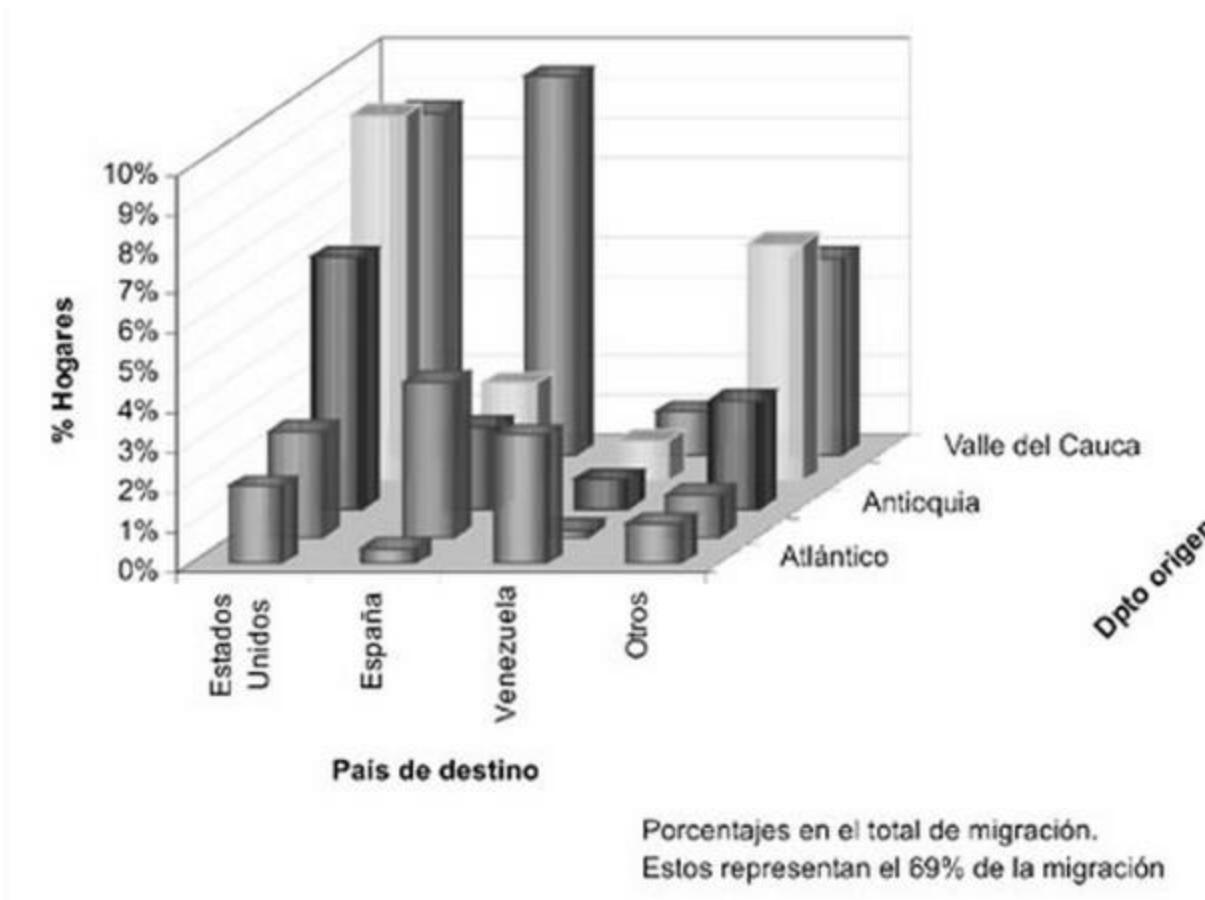


Un grafico vale más que mil palabras!



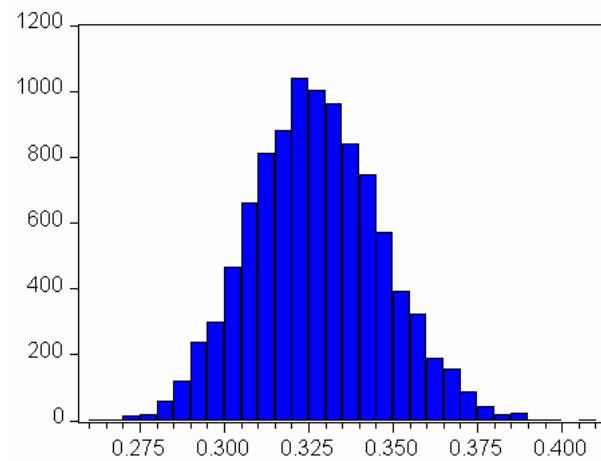
“Los gráficos no deben ser más complejos que los datos que describe”
(evite efectos 3D).

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	 	8
3	 	8
4		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

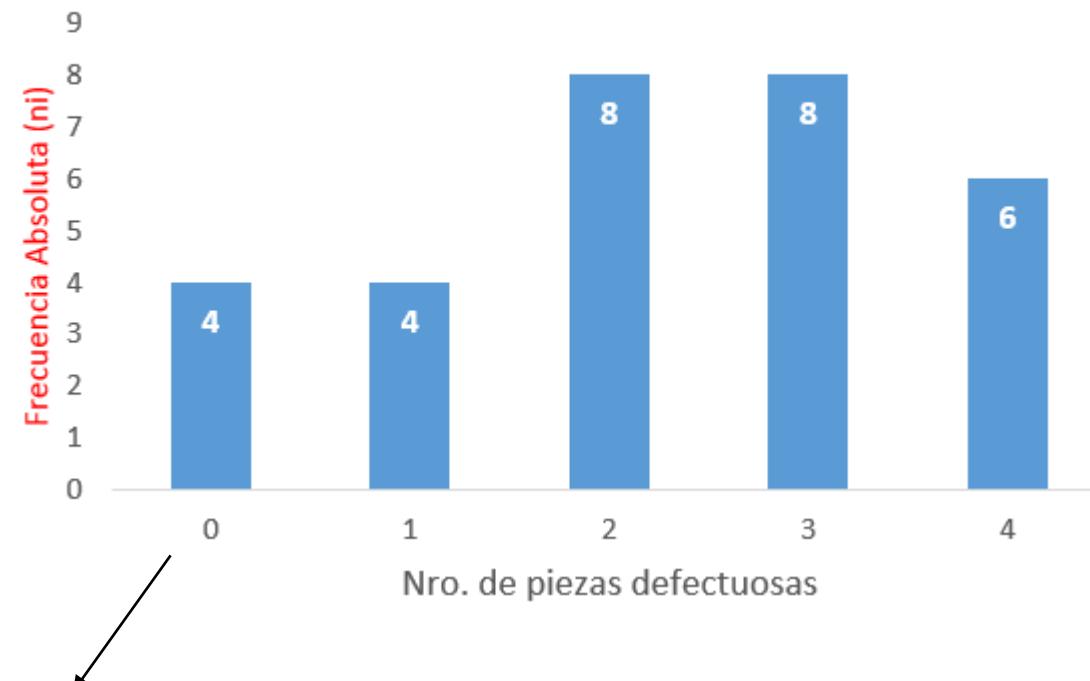
Caso Discreto

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

Distribución del nro. de piezas defectuosas

Puede ser la
frecuencia absoluta o
relativa (%)



Por ser una variable discreta las barras no deben juntarse

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

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

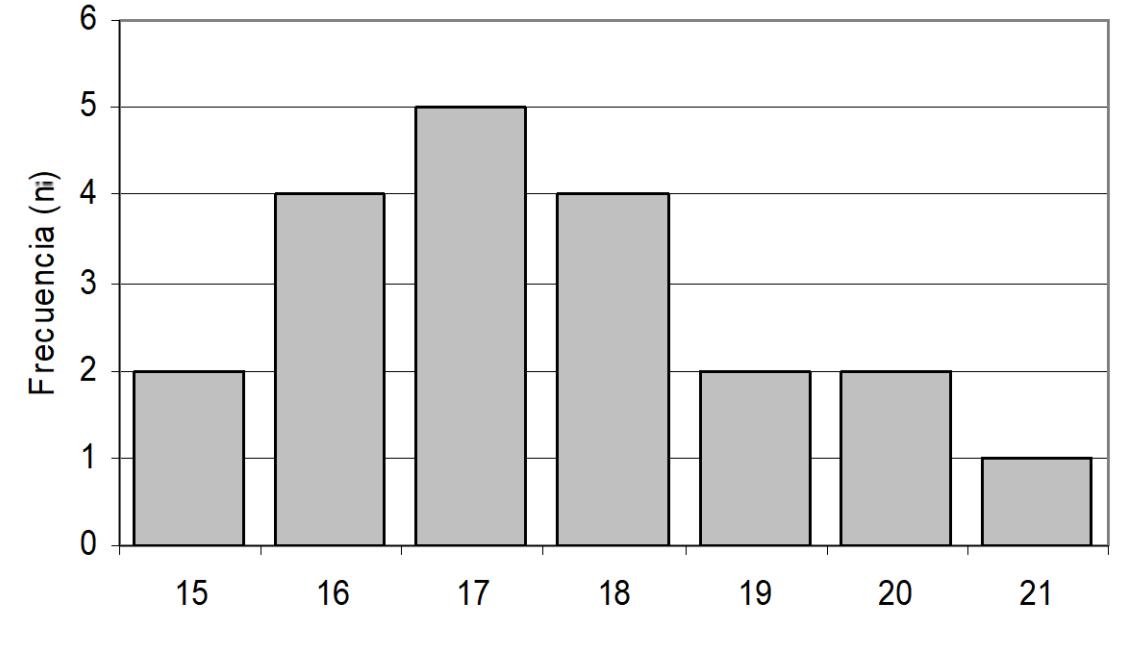
- a. Identifique el tipo de variable.
- b. Construya su respectiva distribución de frecuencias.
- c. Realice el grafico para la frecuencia absoluta simple.
- d. En que porcentaje de las ocasiones se presentan al servicio menos de 19 clientes?
- e. En que porcentaje 19 o mas clientes?

Distribución de Frecuencias

TABLA DE FRECUENCIA DEL NUMERO DE CLIENTES POR HORA

X_i Valor observado	n_i (Frecuenci a 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		

Distribución del numero de clientes por hora



Datos Cuantitativos Continuos

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

	14,91	43,70	160,278	196,561	210,145	220,015	241,401	251,145	261,015	271,885	281,755	291,625	301,495	311,365	321,235	331,105	341,975	351,845	361,715	371,585	381,455	391,325	401,195	411,065	421,935	431,805	441,675	451,545	461,415	471,285	481,155	491,025	501,895	511,765	521,635	531,505	541,375	551,245	561,115	571,985	581,855	591,725	601,595	611,465	621,335	631,205	641,075	651,945	661,815	671,685	681,555	691,425	701,295	711,165	721,035	731,905	741,775	751,645	761,515	771,385	781,255	791,125	801,995	811,865	821,735	831,605	841,475	851,345	861,215	871,085	881,955	891,825	901,695	911,565	921,435	931,305	941,175	951,045	961,915	971,785	981,655	991,525	1001,395	1011,265	1021,135	1031,005	1041,875	1051,745	1061,615	1071,485	1081,355	1091,225	1101,095	1111,965	1121,835	1131,705	1141,575	1151,445	1161,315	1171,185	1181,055	1191,925	1201,795	1211,665	1221,535	1231,405	1241,275	1251,145	1261,015	1271,885	1281,755	1291,625	1301,495	1311,365	1321,235	1331,105	1341,975	1351,845	1361,715	1371,585	1381,455	1391,325	1401,195	1411,065	1421,935	1431,805	1441,675	1451,545	1461,415	1471,285	1481,155	1491,025	1501,895	1511,765	1521,635	1531,505	1541,375	1551,245	1561,115	1571,985	1581,855	1591,725	1601,595	1611,465	1621,335	1631,205	1641,075	1651,945	1661,815	1671,685	1681,555	1691,425	1701,295	1711,165	1721,035	1731,905	1741,775	1751,645	1761,515	1771,385	1781,255	1791,125	1801,995	1811,865	1821,735	1831,605	1841,475	1851,345	1861,215	1871,085	1881,955	1891,825	1901,695	1911,565	1921,435	1931,305	1941,175	1951,045	1961,915	1971,785	1981,655	1991,525	2001,395	2011,265	2021,135	2031,005	2041,875	2051,745	2061,615	2071,485	2081,355	2091,225	2101,095	2111,965	2121,835	2131,705	2141,575	2151,445	2161,315	2171,185	2181,055	2191,925	2201,795	2211,665	2221,535	2231,405	2241,275	2251,145	2261,015	2271,885	2281,755	2291,625	2301,495	2311,365	2321,235	2331,105	2341,975	2351,845	2361,715	2371,585	2381,455	2391,325	2401,195	2411,065	2421,935	2431,805	2441,675	2451,545	2461,415	2471,285	2481,155	2491,025	2501,895	2511,765	2521,635	2531,505	2541,375	2551,245	2561,115	2571,985	2581,855	2591,725	2601,595	2611,465	2621,335	2631,205	2641,075	2651,945	2661,815	2671,685	2681,555	2691,425	2701,295	2711,165	2721,035	2731,905	2741,775	2751,645	2761,515	2771,385	2781,255	2791,125	2801,995	2811,865	2821,735	2831,605	2841,475	2851,345	2861,215	2871,085	2881,955	2891,825	2901,695	2911,565	2921,435	2931,305	2941,175	2951,045	2961,915	2971,785	2981,655	2991,525	3001,395	3011,265	3021,135	3031,005	3041,875	3051,745	3061,615	3071,485	3081,355	3091,225	3101,095	3111,965	3121,835	3131,705	3141,575	3151,445	3161,315	3171,185	3181,055	3191,925	3201,795	3211,665	3221,535	3231,405	3241,275	3251,145	3261,015	3271,885	3281,755	3291,625	3301,495	3311,365	3321,235	3331,105	3341,975	3351,845	3361,715	3371,585	3381,455	3391,325	3401,195	3411,065	3421,935	3431,805	3441,675	3451,545	3461,415	3471,285	3481,155	3491,025	3501,895	3511,765	3521,635	3531,505	3541,375	3551,245	3561,115	3571,985	3581,855	3591,725	3601,595	3611,465	3621,335	3631,205	3641,075	3651,945	3661,815	3671,685	3681,555	3691,425	3701,295	3711,165	3721,035	3731,905	3741,775	3751,645	3761,515	3771,385	3781,255	3791,125	3801,995	3811,865	3821,735	3831,605	3841,475	3851,345	3861,215	3871,085	3881,955	3891,825	3901,695	3911,565	3921,435	3931,305	3941,175	3951,045	3961,915	3971,785	3981,655	3991,525	4001,395	4011,265	4021,135	4031,005	4041,875	4051,745	4061,615	4071,485	4081,355	4091,225	4101,095	4111,965	4121,835	4131,705	4141,575	4151,445	4161,315	4171,185	4181,055	4191,925	4201,795	4211,665	4221,535	4231,405	4241,275	4251,145	4261,015	4271,885	4281,755	4291,625	4301,495	4311,365	4321,235	4331,105	4341,975	4351,845	4361,715	4371,585	4381,455	4391,325	4401,195	4411,065	4421,935	4431,805	4441,675	4451,545	4461,415	4471,285	4481,155	4491,025	4501,895	4511,765	4521,635	4531,505	4541,375	4551,245	4561,115	4571,985	4581,855	4591,725	4601,595	4611,465	4621,335	4631,205	4641,075	4651,945	4661,815	4671,685	4681,555	4691,425	4701,295	4711,165	4721,035	4731,905	4741,775	4751,645	4761,515	4771,385	4781,255	4791,125	4801,995	4811,865	4821,735	4831,605	4841,475	4851,345	4861,215	4871,085	4881,955	4891,825	4901,695	4911,565	4921,435	4931,305	4941,175	4951,045	4961,915	4971,785	4981,655	4991,525	5001,395	5011,265	5021,135	5031,005	5041,875	5051,745	5061,615	5071,485	5081,355	5091,225	5101,095	5111,965	5121,835	5131,705	5141,575	5151,445	5161,315	5171,185	5181,055	5191,925	5201,795	5211,665	5221,535	5231,405	5241,275	5251,145	5261,015	5271,885	5281,755	5291,625	5301,495	5311,365	5321,235	5331,105	5341,975	5351,845	5361,715	5371,585	5381,455	5391,325	5401,195	5411,065	5421,935	5431,805	5441,675	5451,545	5461,415	5471,285	5481,155	5491,025	5501,895	5511,765	5521,635	5531,505	5541,375	5551,245	5561,115	5571,985	5581,855	5591,725	5601,595	5611,465	5621,335	5631,205	5641,075	5651,945	5661,815	5671,685	5681,555	5691,425	5701,295	5711,165	5721,035	5731,905	5741,775	5751,645	5761,515	5771,385	5781,255	5791,125	5801,995	5811,865	5821,735	5831,605	5841,475	5851,345	5861,215	5871,085	5881,955	5891,825	5901,695	5911,565	5921,435	5931,305	5941,175	5951,045	5961,915	5971,785	5981,655	5991,525	6001,395	6011,265	6021,135	6031,005	6041,875	6051,745	6061,615	6071,485	6081,355	6091,225	6101,095	6111,965	6121,835	6131,705	6141,575	6151,445	6161,315	6171,185	6181,055	6191,925	6201,795	6211,665	6221,535	6231,405	6241,275	6251,145	6261,015	6271,885	6281,755	6291,625	6301,495	6311,365	6321,235	6331,105	6341,975	6351,845	6361,715	6371,585	6381,455	6391,325	6401,195	6411,065	6421,935	6431,805	6441,675	6451,545	6461,415	6471,285	6481,155	6491,025	6501,895	6511,765	6521,635	6531,505	6541,375	6551,245	6561,115	6571,985	6581,855	6591,725	6601,595	6611,465	6621,335	6631,205	6641,075	6651,945	6661,815	6671,685	6681,555	6691,425	6701,295	6711,165	6721,035	6731,905	6741,775	6751,645	6761,515	6771,385	6781,255	6791,125	6801,995	6811,865	6821,735	6831,605	6841,475	6851,345	6861,215	6871,085	6881,955	6891,825	6901,695	6911,565	6921,435	6931,305	6941,175	6951,045	6961,915	6971,785	6981,655	6991,525	7001,395	7011,265	7021,135	7031,005	7041,875	7051,745	7061,615	7071,485	7081,355	7091,225	7101,095	7111,965	7121,835	7131,705	7141,575	7151,445	7161,315	7171,185	7181,055	7191,925	7201,795	7211,665	7221,535	7231,405	7241,275	7251,145	7261,015	7271,885	7281,755	7291,625	7301,495	7311,365	7321,235	7331,105	7341,975	7351,845	7361,715	7371,585	7381,455	7391,325	7401,195	7411,065	7421,935	7431,805	7441,675	7451,545	7461,415	7471,285	7481,155	7491,025	7501,895	7511,765	7521,635	7531,505	7541,375	7551,245	7561,115	7571,985	7581,855	7591,725	7601,595	7611,465	7621,335	7631,205	7641,075	7651,945	7661,815	7671,685	7681,555	7691,425	7701,295	7711,165	7721,035	7731,905	7741,775	7751,645	7761,515	7771,385	7781,255	7791,125	7801,995	7811,865	7821,735	7831,605	7841,475	7851,345	7861,215	7871,085	7881,955	7891,825	7901,695	7911,565	7921,435	7931,305	7941,175	7951,045	7961,915	7971,785	7981,655	7991,525	8001,395	8011,265	8021,135	8031,005	8041,875	8051,745	8061,615	8071,485	8081,355	8091,225	8101,095	8111,965	8121,835	8131,705	8141,575	8151,445	8161,315	8171,185	8181,055	8191,925	8201,795	8211,665	8221,535	8231,405	8241,275	8251,145	8261,015	8271,885	8281,755	8291,625	8301,495	8311,365	8321,235	8331,105	8341,975	8351,845	8361,715	8371,585	8381,455	8391,325	8401,195	8411,065	8421,935	8431,805	8441,675	8451,545	8461,415	8471,285	8481,155	8491,025	8501,895	8511,765	8521,635	8531,505	8541,375	8551,245	8561,115	8571,985	8581,855	8591,725	8601,595	8611,465	8621,335	8631,205	8641,075	8651,945	8661,815	8671,685</th

Intervalos de Clase

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

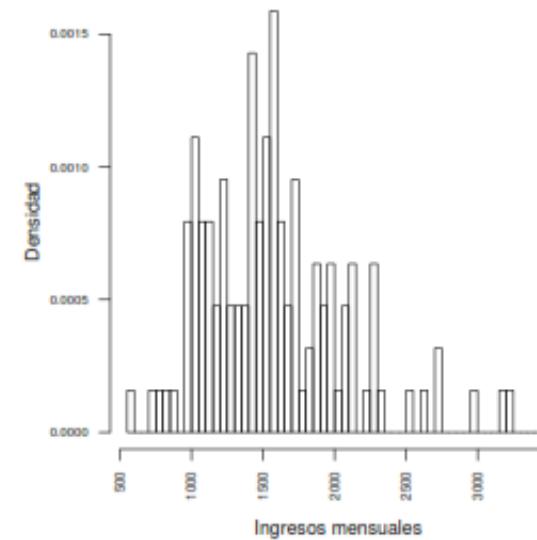
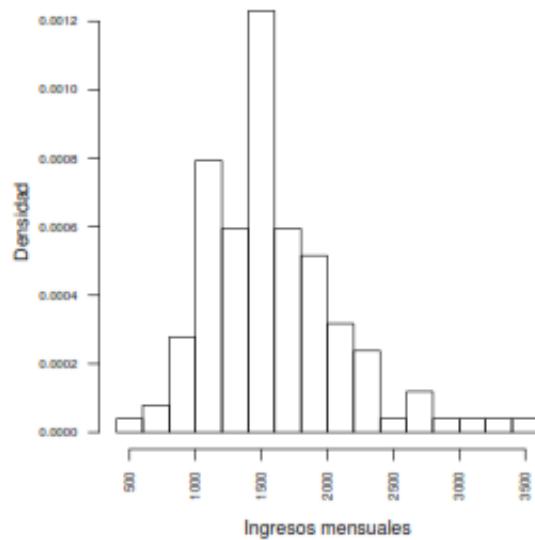
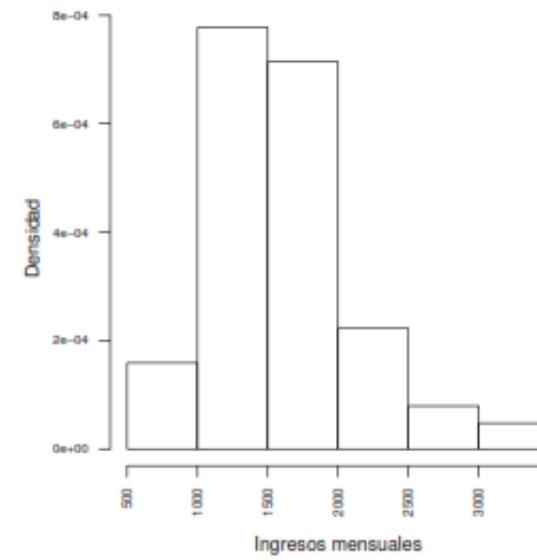
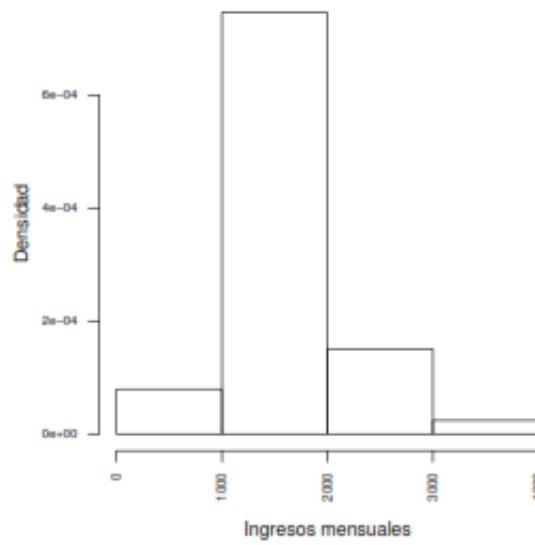
¿Cuántos intervalos?

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: $2^k > n$

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

¿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:

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

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

$$Rango = Max(x_i) - Min(x_i) \rightarrow R = 4.5 - 0.3 = 4.2$$

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

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

Distribución de Frecuencia

PASOS PARA CONSTRUIR UNA DISTRIBUCIÓN DE FRECUENCIA EN DATOS AGRUPADOS

4. Determinar los límites ($L_0, L_1, L_2, \dots, L_k$):

$$L_0 = \text{Min}$$

$$L_1 = L_0 + C$$

$$L_2 = L_1 + C$$

$$L_3 = L_2 + C$$

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

$$L_0 = 0.3$$

$$L_1 = 0.3 + 0.7 = 1.0$$

$$L_2 = 1.0 + 0.7 = 1.7$$

$$L_3 = 1.7 + 0.7 = 2.4$$

$$L_4 = 2.4 + 0.7 = 3.1$$

$$L_5 = 3.1 + 0.7 = 3.8$$

$$L_6 = 3.8 + 0.7 = 4.5$$

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 oxígeno disuelto entre 3.1 y 3.8 mg/l.

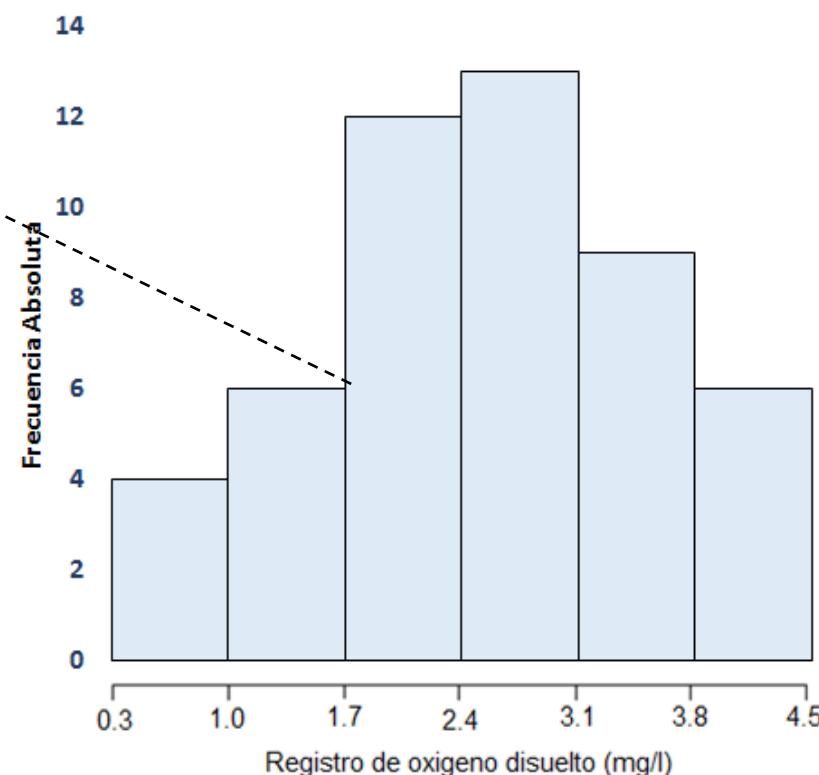
El 88% de las mediciones presentaron un registro de oxígeno disuelto menor o igual a 3.8 mg/l.

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



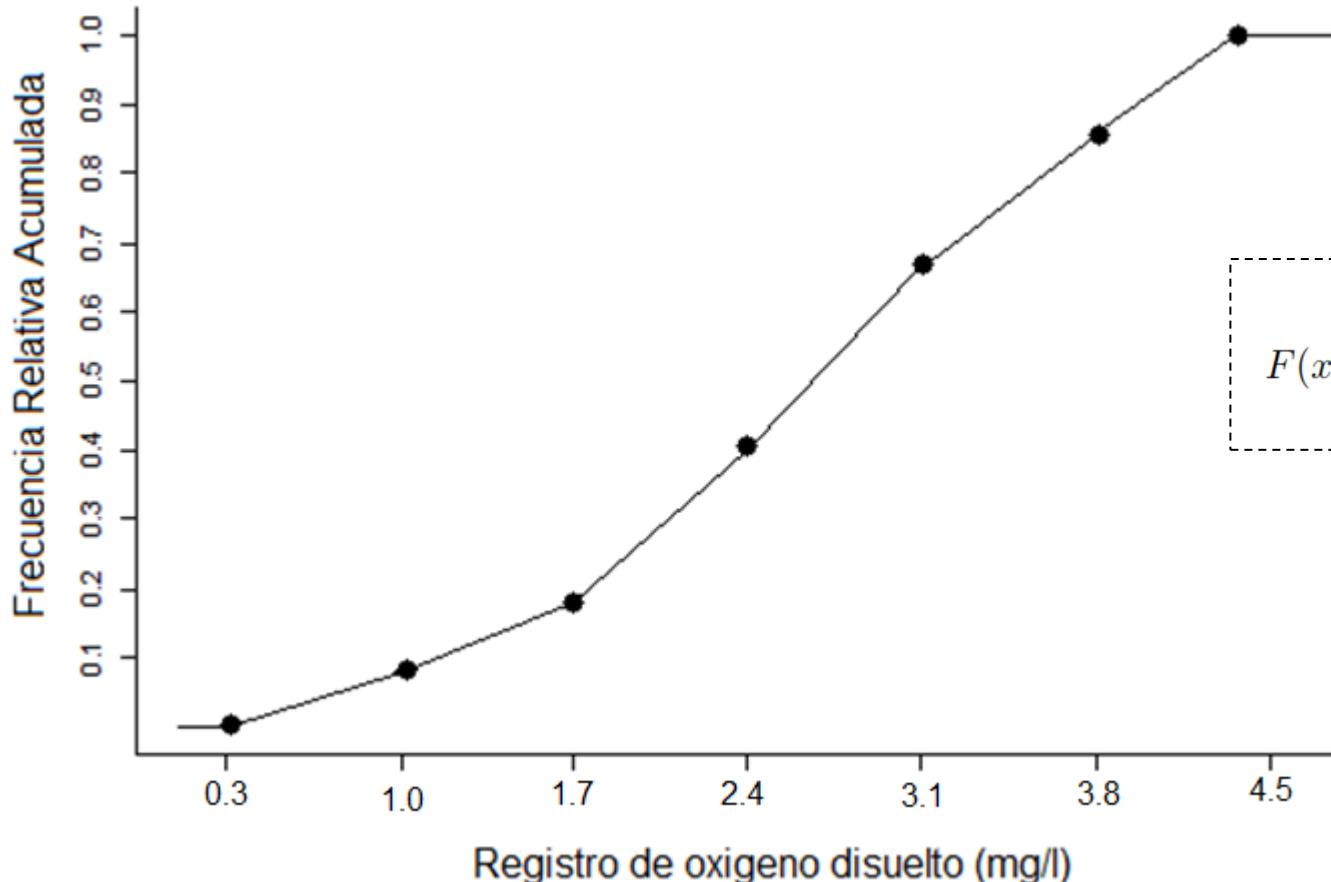
Intervalos	n_i	f_i
[0,3 , 1,0]	4	0,08
(1,0 , 1,7]	6	0,12
(1,7 , 2,4]	12	0,24
(2,4 , 3,1]	13	0,26
(3,1 , 3,8]	9	0,18
(3,8 , 4,5]	6	0,12
Total	50	1,00

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 Oxígeno 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 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.

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		

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

Ejercicio

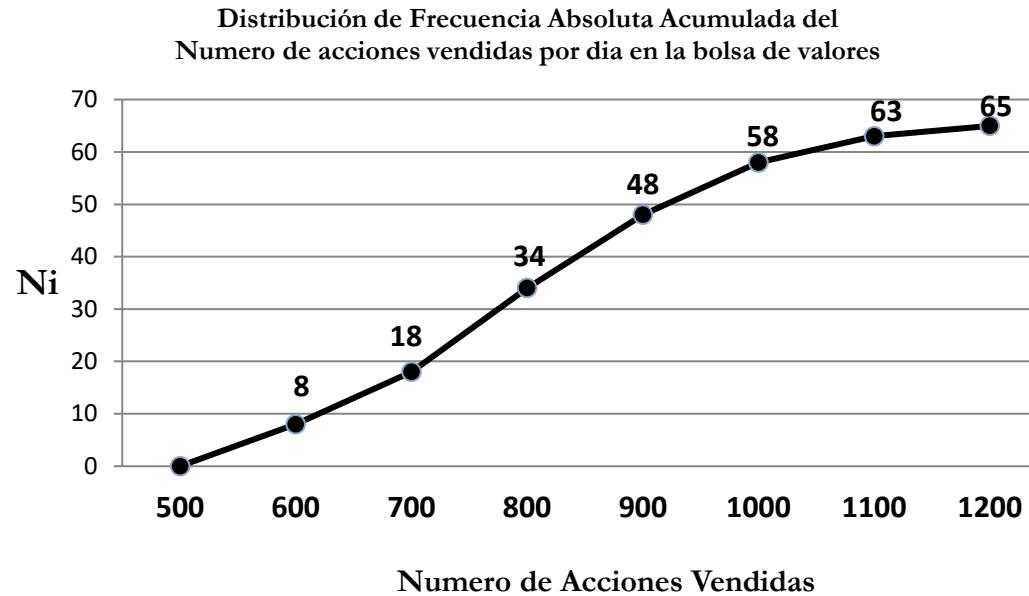
Intervalos	x'_i	n_i	f_i	N_i	F_i
[0.3 , 1.0]	0,65	4	0,08	4	0,08
(1.0 , 1.7]	1,35	6	0,12	10	0,20
(1.7 , 2.4]	2,05	12	0,24	22	0,44
(2.4 , 3.1]	2,75	13	0,26	35	0,70
(3.1 , 3.8]	3,45	9	0,18	44	0,88
(3.8 , 4.5]	4,15	6	0,12	50	1,00
	Total	50	1.0		

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 qué 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.