ANÁLISIS INTELIGENTE DE DATOS

MAESTRÍA EN EXPLORACIÓN DE DATOS Y DESCUBRIMIENTO DEL CONOCIMIENTO

PROFESORA: MÓNICA CANTONI

AYUDANTES: CECILIA OLIVA

FABIANA ROSSI

PAMELA PAIRO

CLASE 01 – ANÁLISIS EXPLORATORIO Y DESCRIPTIVO DE DATOS

AGENDA

- Datos estructurados
- Atributos
- Medidas descriptivas univariadas
- Visualización de datos

DATOS ESTRUCTURADOS

Ejemplo: datos de calidad del vino

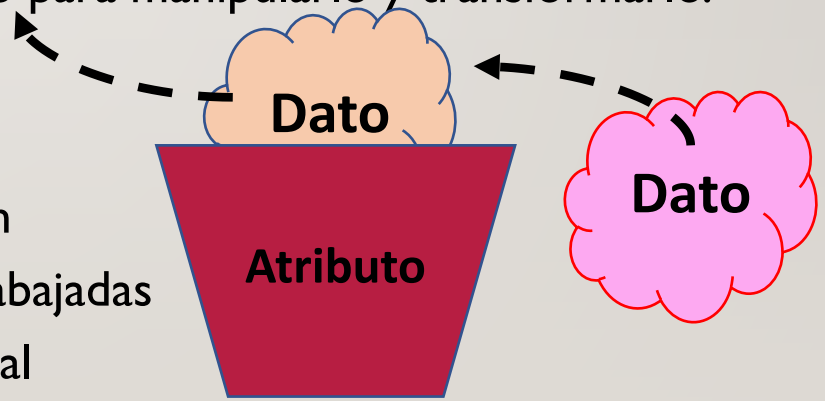
Atributos

Instancias

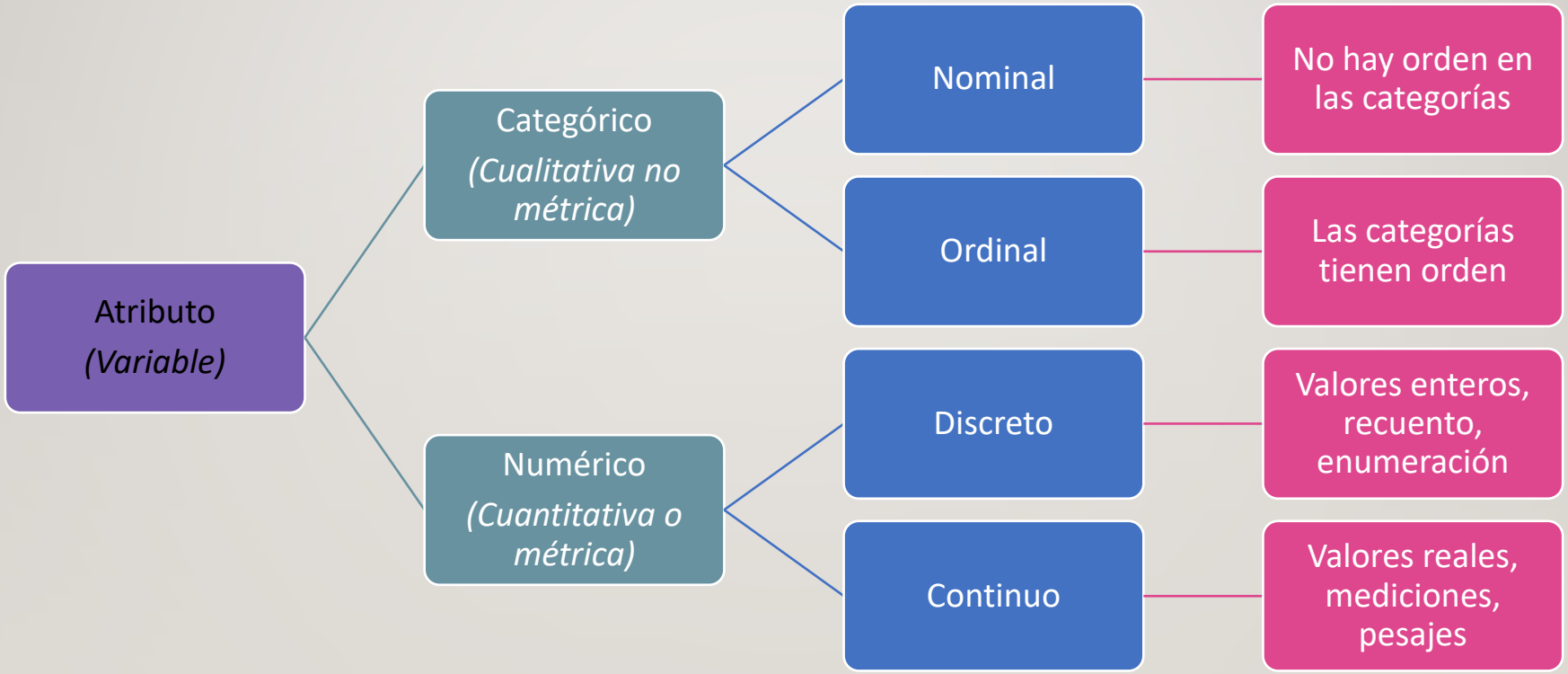
acidez fija	acidez volátil	ácido cítrico	azúcar residual	cloruros	anhídrido sulfuroso libre	anhídrido sulfuroso total	densidad	pH	sulfatos	alcohol	calidad	variedad
7	0.27	0.36	20.7	0.045	45	170	1.001	3	0.45	8.8	6	blanco
6.3	0.3	0.34	1.6	0.049	14	132	0.994	3.3	0.49	9.5	6	blanco
8.1	0.28	0.4	6.9	0.05	30	97	0.9951	3.26	0.44	10.1	6	blanco
7.2	0.23	0.32	8.5	0.058	47	186	0.9956	3.19	0.4	9.9	6	blanco
7.2	0.23	0.32	8.5	0.058	47	186	0.9956	3.19	0.4	9.9	6	blanco
8.1	0.28	0.4	6.9	0.05	30	97	0.9951	3.26	0.44	10.1	6	blanco
6.2	0.32	0.16	7	0.045	30	136	0.9949	3.18	0.47	9.6	6	blanco
7	0.27	0.36	20.7	0.045	45	170	1.001	3	0.45	8.8	6	blanco
6.3	0.3	0.34	1.6	0.049	14	132	0.994	3.3	0.49	9.5	6	blanco
8.1	0.22	0.43	1.5	0.044	28	129	0.9938	3.22	0.45	11	6	blanco
8.1	0.27	0.41	1.45	0.033	11	63	0.9908	2.99	0.56	12	5	blanco
8.6	0.23	0.4	4.2	0.035	17	109	0.9947	3.14	0.53	9.7	5	blanco
7.9	0.18	0.37	1.2	0.04	16	75	0.992	3.18	0.63	10.8	5	blanco
6.6	0.16	0.4	1.5	0.044	48	143	0.9912	3.54	0.52	12.4	7	blanco
8.3	0.42	0.62	19.25	0.04	41	172	1.0002	2.98	0.67	9.7	5	blanco
6.6	0.17	0.38	1.5	0.032	28	112	0.9914	3.25	0.55	11.4	7	blanco
6.3	0.48	0.04	1.1	0.046	30	99	0.9928	3.24	0.36	9.6	6	blanco
6.2	0.66	0.48	1.2	0.029	29	75	0.9892	3.33	0.39	12.8	8	blanco
7.4	0.34	0.42	1.1	0.033	17	171	0.9917	3.12	0.53	11.3	6	blanco

ATRIBUTOS: CONCEPTO Y EJEMPLOS

- Atributo
  - Denominado "estructura de datos" en programación y "variable" en estadística
  - Es una forma de identificar de manera sencilla un dato.
  - Es un contenedor en el que se almacena un dato. El dato puede cambiar.
  - Un atributo nos permite acceder al dato para manipularlo y transformarlo.
- Ejemplos
  - Edad, Género, Peso, Altura, Nacionalidad
  - Lugar de residencia, Nivel de instrucción
  - Categoría laboral, Cantidad de horas trabajadas
  - Profesión, Ingreso mensual, Código Postal



TIPOS DE ATRIBUTO (VARIABLE). CLASIFICACIÓN SEGÚN TIPO DE DATO



EJEMPLOS DE ATRIBUTOS CUALITATIVA (NO MÉTRICA)

Nominal	**Ordinal**
Color del vino	Nivel de instrucción
Tipo de comida	Orden de mérito en un concurso académico
País de origen	Categoría laboral
Estado civil	Nivel de inglés
Idioma	Nivel de satisfacción de un cliente con el servicio
Nacionalidad	

EJEMPLOS DE ATRIBUTOS. CUANTITATIVA (MÉTRICA)

Discreta

Edad en años

Cantidad de materias aprobadas

Continua

Peso de la persona

Altura de la persona

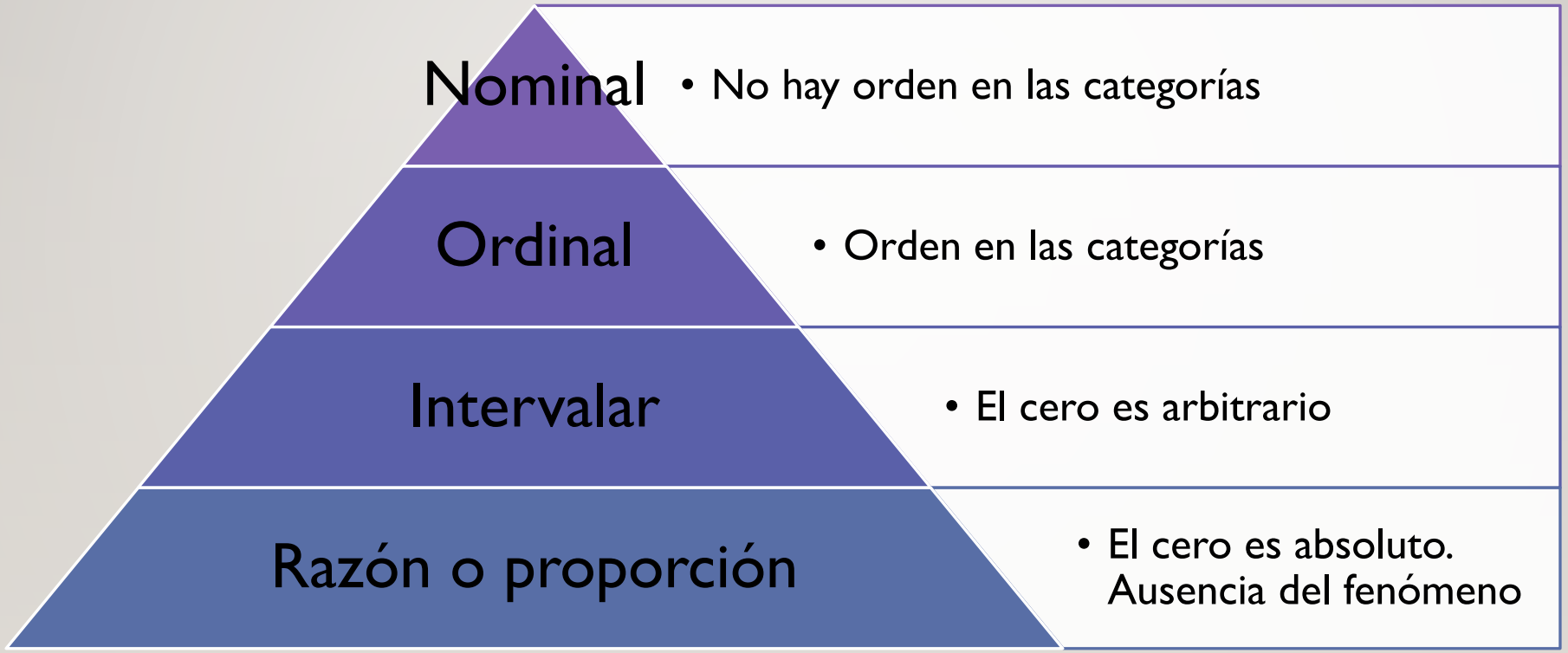
Superficie de una vivienda

Ingresos mensuales

Saldo cuenta bancaria

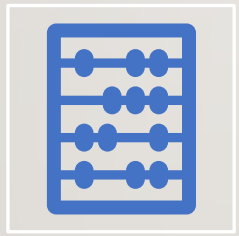
Tarifa por noche de hospedaje

TIPO DE ATRIBUTO (VARIABLE). CLASIFICACIÓN SEGÚN NIVEL DE MEDICIÓN



MEDIDAS DESCRIPTIVAS UNIVARIADAS

- Atributo Cualitativo



Frecuencia

Porcentaje

Moda

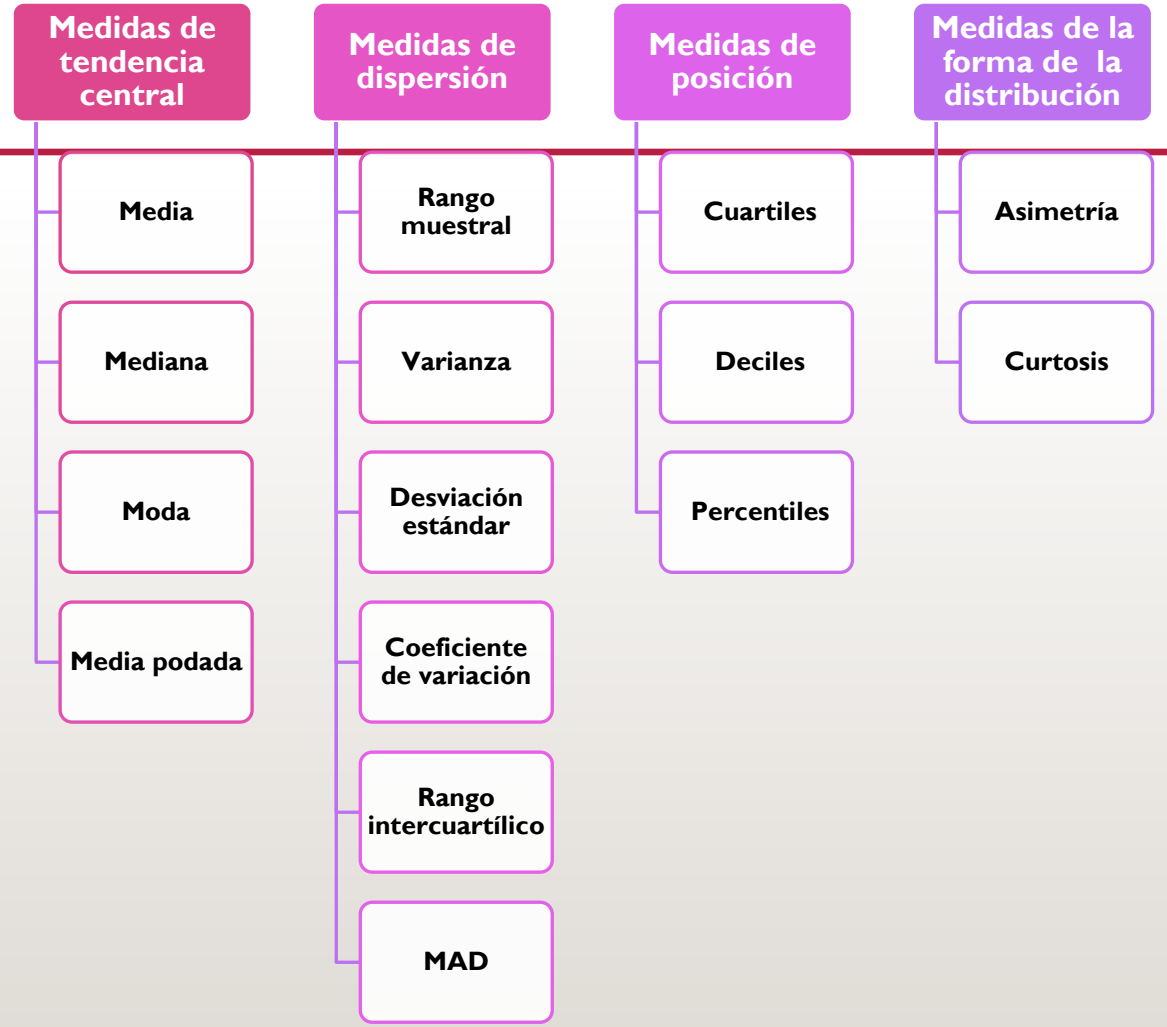
MEDIDAS DESCRIPTIVAS UNIVARIADAS CUALITATIVO

- Ejemplos de frecuencia, porcentaje y moda. Datos de cursos ofrecidos por Udemy.

Categoría	Frecuencia	Porcentaje
Moda → Informática y software	2741	22.44
Desarrollo	2659	21.77
Negocios	1774	14.52
Desarrollo personal	1704	13.95
Diseño	1534	12.56
Finanzas y contabilidad	878	7.19
Productividad en la oficina	576	4.71
Marketing	350	2.86
	12216	**100**

MEDIDAS DESCRIPTIVAS UNIVARIADAS

- **Atributo Cuantitativo**



MEDIDAS DESCRIPTIVAS UNIVARIADAS
MEDIDAS DE TENDENCIA CENTRAL

- Media
- Mediana
- Moda
- Media podada

MEDIA O PROMEDIO

- La **media** es el promedio aritmético de los valores del atributo. Dicho atributo debe ser numérico.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Ejemplo para atributo edad en años:

28 30 34 34 36 40 44 45 46 47 48 50 51 52 60 65 120

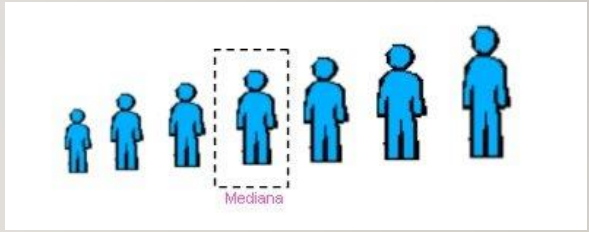
$$\bar{x} = \frac{28+30+34+34+36+44+45+46+47+50+51+52+60+65+120}{17} = 48{,}82$$

28 30 34 34 36 40 44 45 46 47 48 50 51 52 60 65

$$\bar{x} = \frac{28+30+34+34+36+44+45+46+47+50+51+52+60+65}{16} = 44{,}38$$

MEDIANA

- La mediana es el valor del atributo que divide a la serie de datos en dos partes iguales. De manera que el 50% de las observaciones está por debajo de la mediana y el 50% está por encima de la mediana.



- Antes de calcularla deben ordenarse los valores del atributo.
- El atributo debe ser numérico.
- Ejemplo para edad:
- Si la cantidad de datos es impar es el valor que se encuentra en la posición central:

28 30 34 34 36 40 44 45 **46** 47 48 50 51 52 60 65 120

- Si la cantidad de datos es par es la media de los dos valores centrales:

28 30 34 34 36 40 44 **45 46** 47 48 50 51 52 60 65

$\tilde{x} = 45.5$ La mediana es 45,5

MODA

- La moda es el valor del atributo que aparece con mayor frecuencia. Por lo tanto, puede determinarse para atributos cualitativos y cuantitativos.

- Es posible que la mayor frecuencia corresponda a varios valores diferentes, lo que da lugar a más de una moda.
- Los conjuntos de datos con uno, dos o tres modas se denominan unimodal, bimodal y trimodal, respectivamente.
- En general, un conjunto de datos con dos o más modas es multimodal.
- Si cada valor de los datos ocurre sólo una vez, entonces no hay moda.
- Ejemplo para atributo numérico. Edad en años

28 30 34 34 36 40 44 45 46 47 48 50 51 52 60 65

- La moda es **34**
- Ejemplo para atributo categórico. Color de fondo de la imagen

Blanco – negro – rojo – rojo – rojo - azul – verde – gris – gris – celeste – naranja

- La moda es **rojo**

MEDIA PODADA

- Promedio de los datos centrales recortado en un cierto porcentaje.
- Ejemplo. Para una serie de 20 valores calculamos la media podada al 10%

28	30	34	34	36	40	44	45	46	46	47	48	48	48	49	50	51	52	60	65

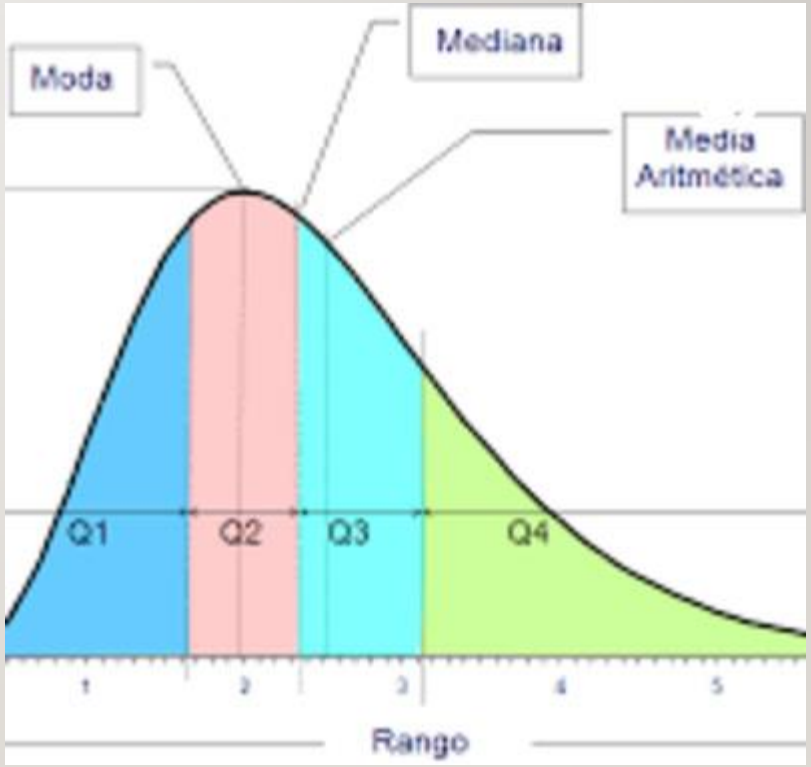
$$\bar{x}_{0.10} = \frac{34 + 34 + 36 + 40 + 44 + 45 + 46 + 46 + 47 + 48 + 48 + 48 + 49 + 50 + 51 + 52}{16} = 44{,}875$$

MEDIDAS DESCRIPTIVAS UNIVARIADAS MEDIDAS DE POSICIÓN O ESTADÍSTICAS DE ORDEN

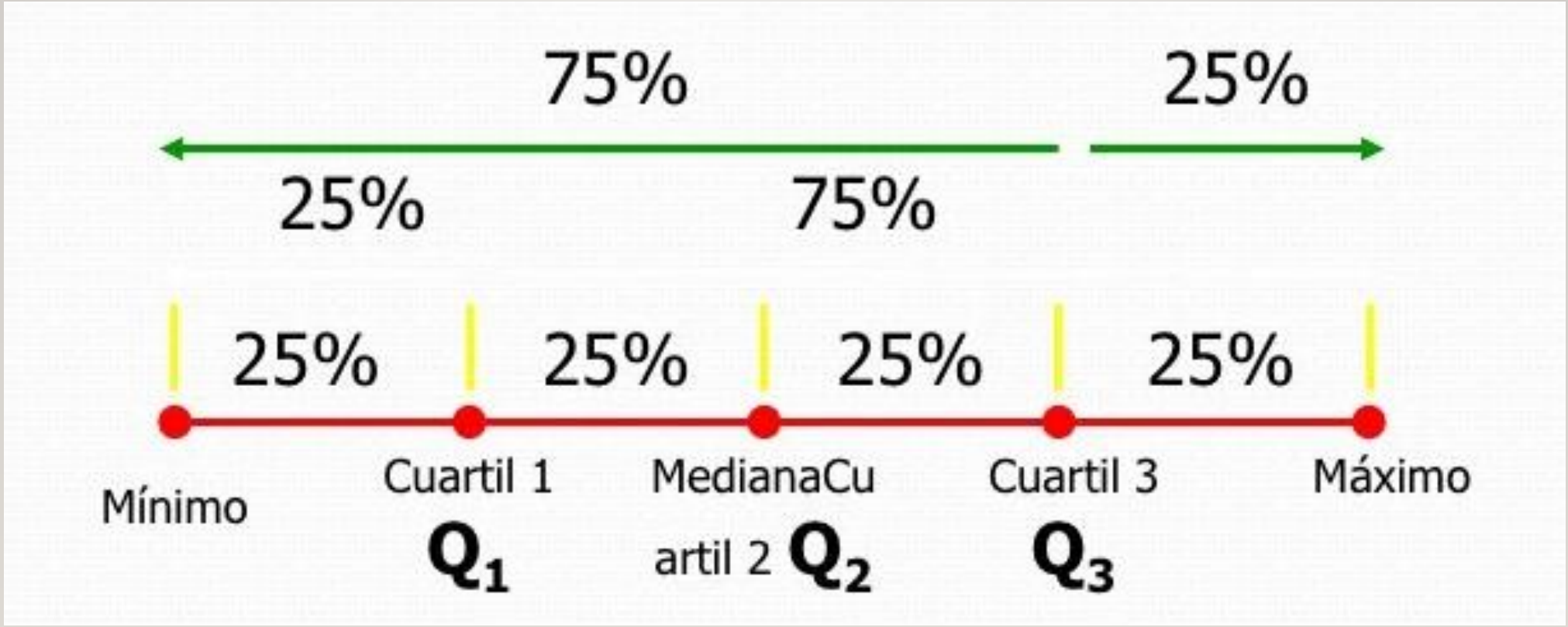
Cuantiles

Para calcular la posición:

- Cuartiles $\frac{1}{4}k(n+1), k = 1,2,3,4$
- Deciles $\frac{1}{10}k(n+1), k = 1,2,\ldots,10$
- Percentiles $\frac{1}{100}k(n+1), k = 1,2,..,100$



CUARTILES


75%
25%
25%
75%
25%
25%
25%
25%
Mínimo
Cuartil 1
Q_1
MedianaCu
artil 2 Q_2
Cuartil 3
Q_3
Máximo

EJEMPLO DE DECILES

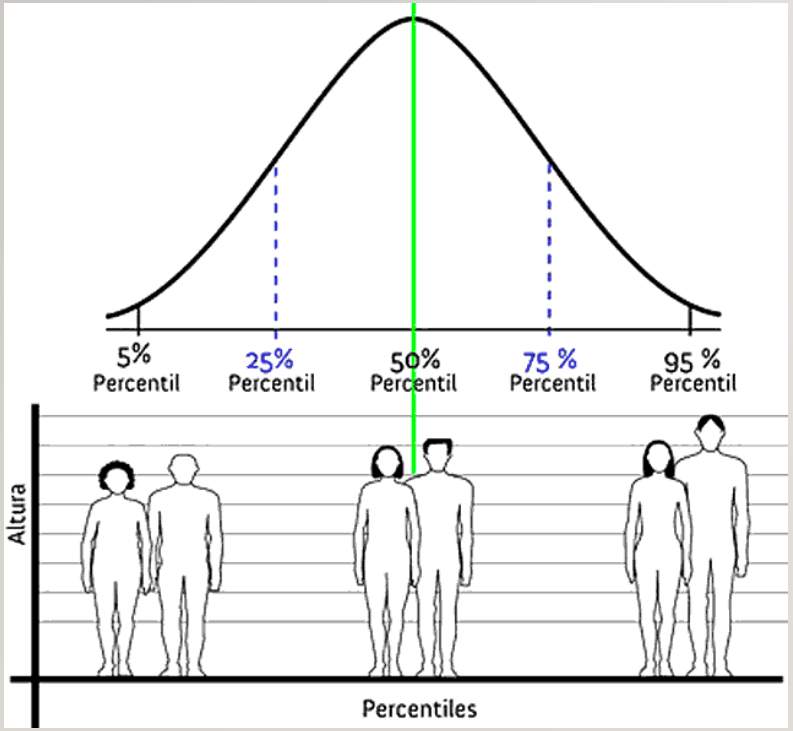
Cuadro 1. **Población según escala de ingreso per cápita familiar. Total 31 aglomerados urbanos. Segundo trimestre de 2022**

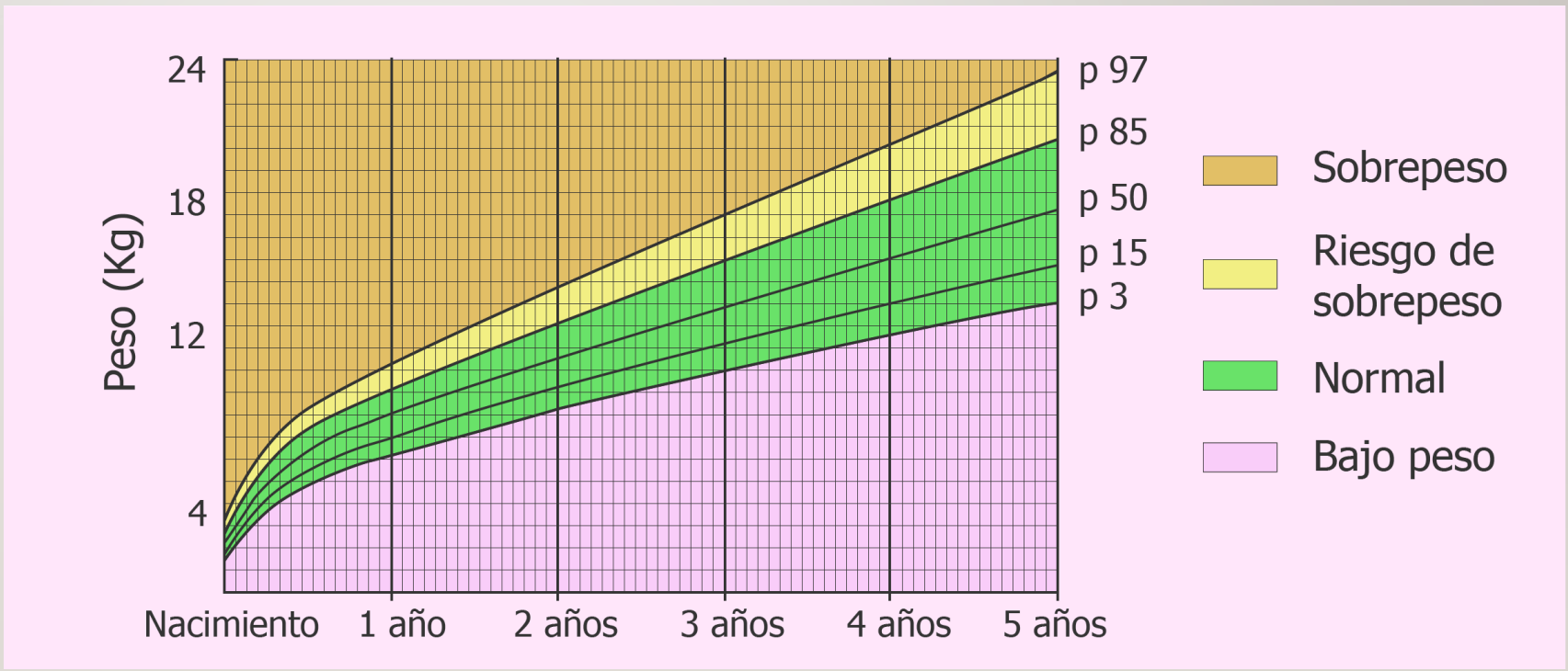
Decil	Escala de ingreso		Población		Ingreso per cápita familiar				
	Desde	Hasta	Población por decil	Porcentaje de personas	Ingreso total por decil (en miles)	Porcentaje del ingreso	Ingreso medio por decil	Ingreso medio por estrato	Mediana por decil
	$	$		%	$	%	$	$	$
1	0	11.800	2.913.811	10,0	22.852.435	1,9	7.843		8.360
2	11.800	16.400	2.913.752	10,0	41.340.673	3,4	14.188		14.250
3	16.400	20.867	2.913.741	10,0	54.428.854	4,5	18.680		18.778
4	20.867	26.000	2.913.763	10,0	68.333.727	5,6	23.452	16.041	23.400
5	26.000	31.125	2.913.612	10,0	82.892.156	6,8	28.450		28.400
6	31.167	37.500	2.914.433	10,0	99.397.370	8,2	34.105		34.000
7	37.500	46.667	2.914.302	10,0	122.147.600	10,1	41.913		41.667
8	46.667	57.875	2.912.393	10,0	150.899.168	12,4	51.813	39.069	51.750
9	57.875	80.000	2.914.092	10,0	198.341.941	16,4	68.063		67.500
10	80.000	1.164.000	2.913.318	10,0	372.242.053	30,7	127.773	97.914	106.667
Población total (1)			**29.137.217**	**100,0**	**1.212.875.977**	**100,0**	**41.626**	**41.626**	**31.125**

(1) Las diferencias en los totales de población y de ingresos entre los distintos cuadros se deben al uso de los diferentes ponderadores correspondientes en cada caso.

Fuente: INDEC, Dirección de Encuesta Permanente de Hogares.

PERCENTILES





MEDIDAS DESCRIPTIVAS UNIVARIADAS
MEDIDAS DE DISPERSIÓN

- Rango muestral
- Varianza muestral
- Desviación estándar muestral
- Coeficiente de variación
- Rango intercuartílico
- MAD.

RANGO MUESTRAL: MÍNIMO Y MÁXIMO

- Para atributo numérico edad en años veamos como se determina:

28 30 34 34 36 40 44 45 46 47 48 50 51 52 60 65 120

- Mínimo es el valor mas pequeño del atributo en el conjunto de datos
- Mínimo es **28**
- Máximo es el valor más grande del atributo en el conjunto de datos
- Máximo es **120**

$$rg_{(x)} = x^{(n)} - x^{(1)}$$

$$rg_{(x)} = 120 - 28 = 92$$

VARIANZA MUESTRAL

- La varianza muestral es el promedio de las distancias cuadráticas entre el valor del atributo y la media.

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

28 30 34 34 36 40 44 45 46 47 48 50 51 52 60 65 120

- La varianza es 436,78
- **Ejemplo 2.** Edad en años con 16 observaciones

28 30 34 34 36 40 44 45 46 47 48 50 51 52 60 65

- La varianza es 107,05

DESVIACIÓN ESTÁNDAR MUESTRAL

- La desviación estándar muestral es el promedio de las distancias entre el valor del atributo y la media.

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

- **Ejemplo 1**. Edad en años con 17 observaciones

28 30 34 34 36 40 44 45 46 47 48 50 51 52 60 65 120

- La desviación estándar es 20,90
- **Ejemplo 2.** Edad en años con 16 observaciones

28 30 34 34 36 40 44 45 46 47 48 50 51 52 60 65

- La desviación estándar es 10,35

COEFICIENTE DE VARIACIÓN

$$CV(x) = \frac{S_x}{\bar{x}}$$

Ejemplo 1.

$$CV(x) = \frac{S_x}{\bar{x}} = \frac{20.9}{48.82} = 0.428\ (42{,}8\%)$$

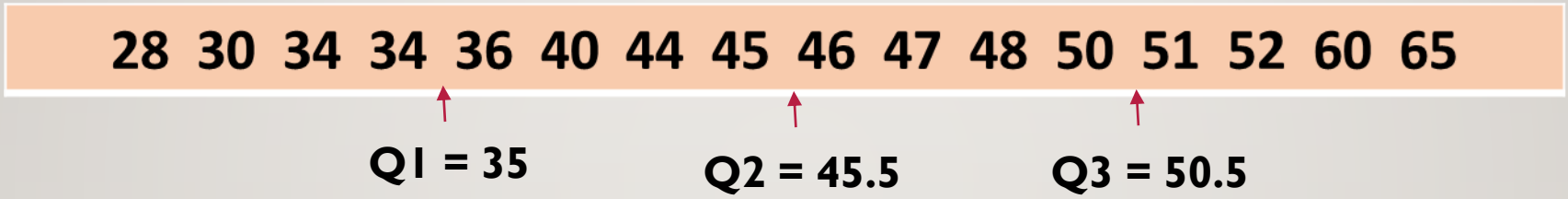
Ejemplo 2.

$$CV(x) = \frac{S_x}{\bar{x}} = \frac{10.35}{44.38} = 0.233\ (23{,}3\%)$$

Si se multiplica por 100 se interpreta como un porcentaje.

RANGO INTERCUARTÍLICO

- $RI = Q3 - Q1$



$RI = Q3 - Q1 = 50.5 - 35 = 15.5$

MAD

Mediana de los desvíos absolutos respecto de la mediana

Veamos un ejemplo: **2 – 3 – 5 – 8 – 13 – 27**

Mediana $\tilde{x} = \frac{1}{2}(5 + 8) = 6.5$

Los desvíos respecto de la mediana son:

-4.5, -3.5, -1.5, 1.5, 6.5, 20.5

Los valores absolutos de los desvíos en forma creciente:

1.5, 1.5, 3.5, 4.5, 6.5, 20.5

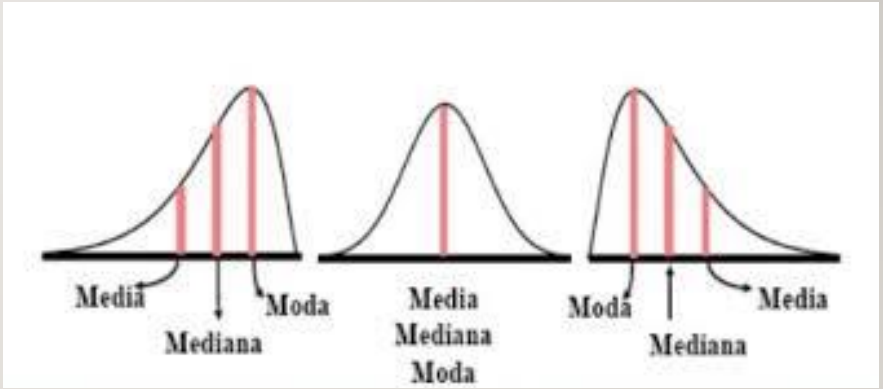
La mediana de los desvíos absolutos:

$$MAD = \frac{3.5 + 4.5}{2} = 4$$

MEDIDAS DE LA FORMA DE LA DISTRIBUCIÓN ASIMETRÍA

- Coeficiente de asimetría muestral
- Fisher $skf(x) = \frac{\sqrt{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^{3/2}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^3}{ns_x^3}$

$$Si\ skf(x) = 0\ Simétrica\ o\ insesgada$$
$$Si\ skf(x) > 0\ Asimétrica, sesgada\ a\ la\ derecha$$
$$Si\ skf(x) < 0\ Asimétrica, sesgada\ a\ la\ izquierda$$

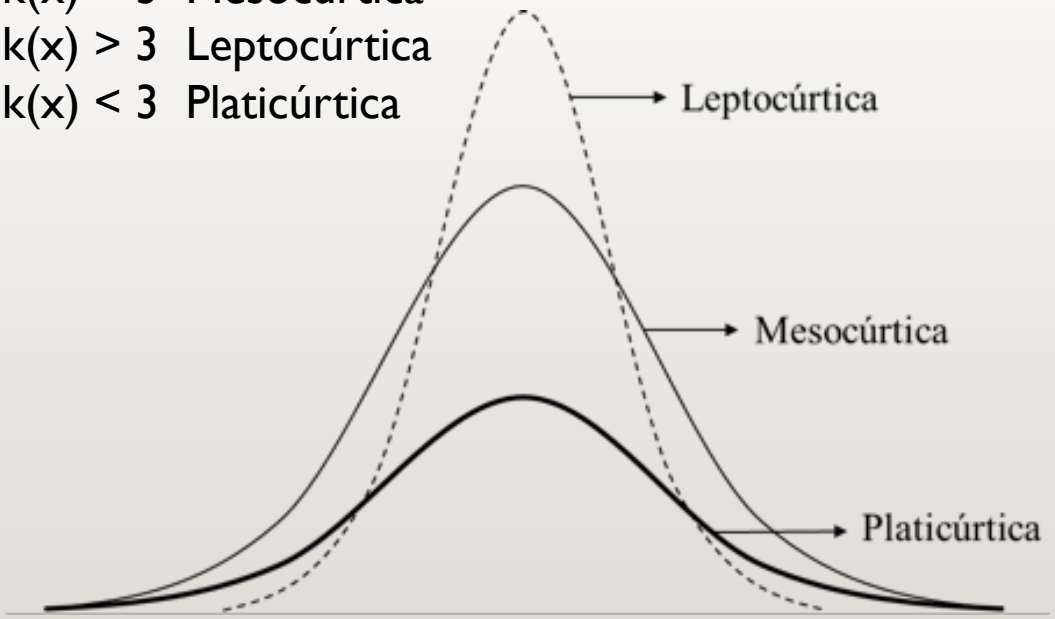


- Pearson $skp(x) = \frac{\bar{x}-Mo(x)}{s_x}$
- Bowley $skb(x) = \frac{(q3-q2)+(q1-q2)}{q3-q1} = \frac{q3+q1-2\tilde{x}}{q3-q1}$

MEDIDAS DE LA FORMA DE LA DISTRIBUCIÓN CURTOSIS

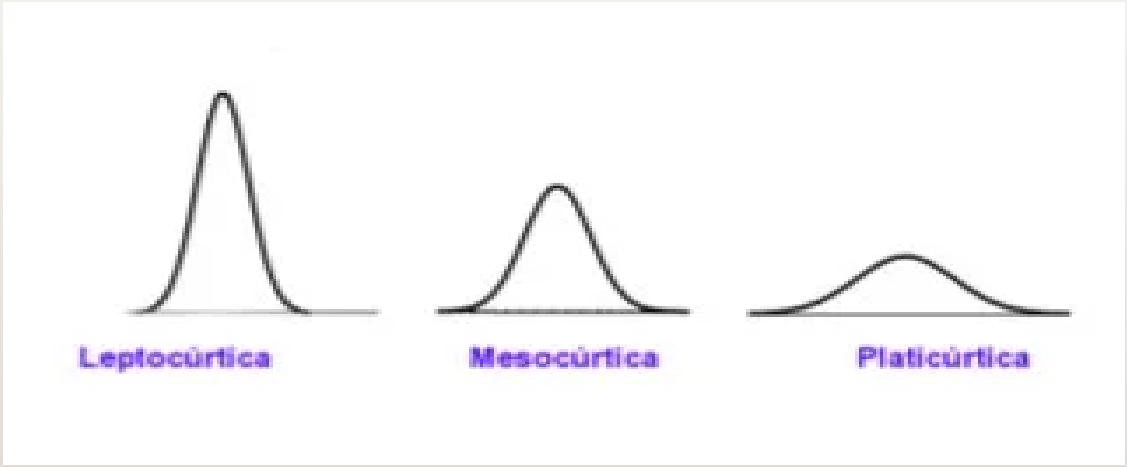
$$k(x) = \frac{n\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2}$$

Si k(x) = 3 Mesocúrtica
Si k(x) > 3 Leptocúrtica
Si k(x) < 3 Platicúrtica



$$k(x) = \frac{n\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2} - 3$$

Si k(x) = 0 Mesocúrtica
Si k(x) > 0 Leptocúrtica
Si k(x) < 0 Platicúrtica

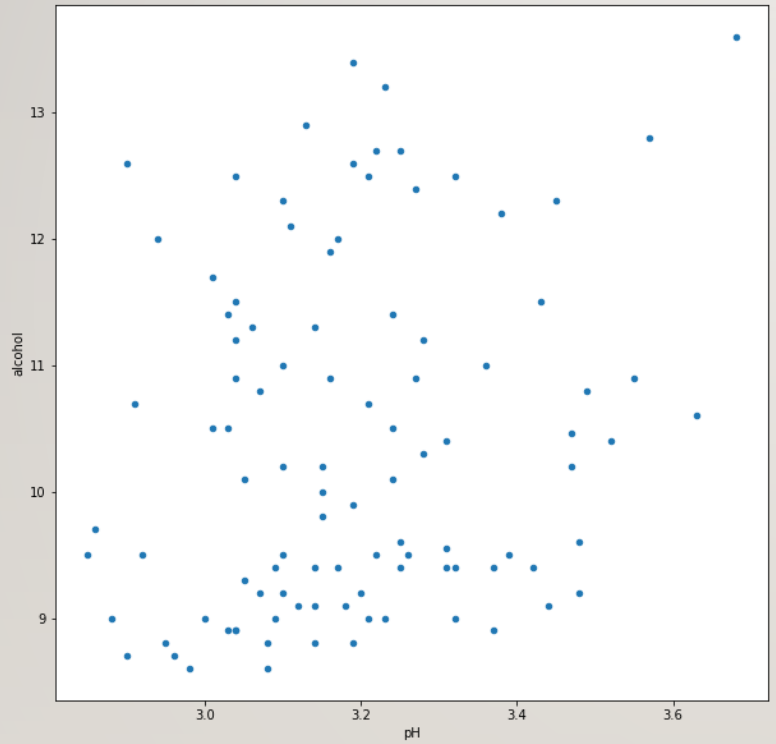


VISUALIZACIÓN DE LA INFORMACIÓN

VISUALIZACIÓN DE LA INFORMACIÓN

- Diagrama de dispersión
- Diagrama circular 2D y 3D
- Diagrama de barras
- Gráfico de barras superpuestas
- Gráfico de barras adyacentes
- Gráfico de bastones
- Diagrama de tallo y hoja
- Histograma
- Polígono de frecuencias
- Gráfico de densidad
- Función de distribución
- Función escalonada
- Boxplot
- Boxplot comparativos.

DIAGRAMA DE DISPERSIÓN



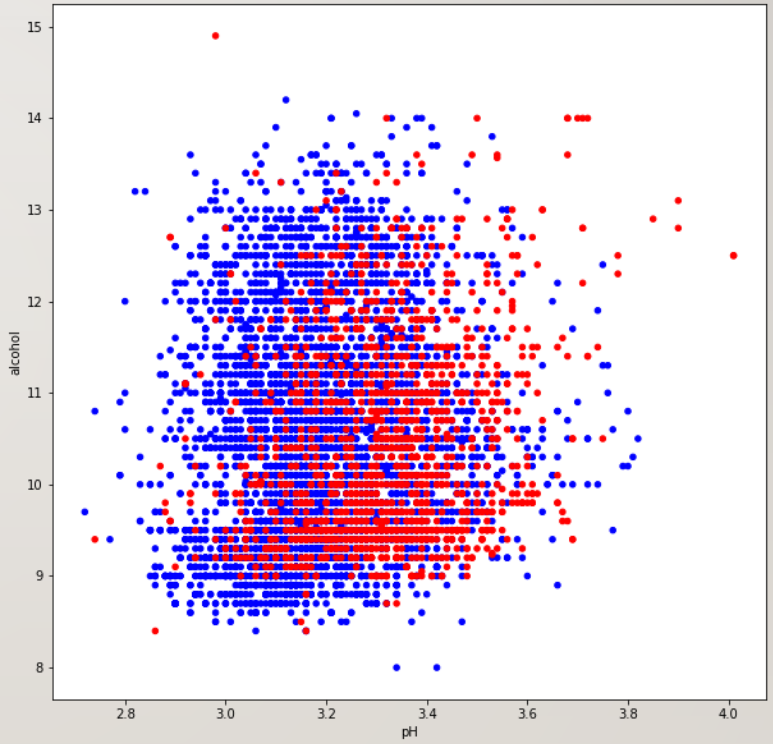


DIAGRAMA CIRCULAR 2D

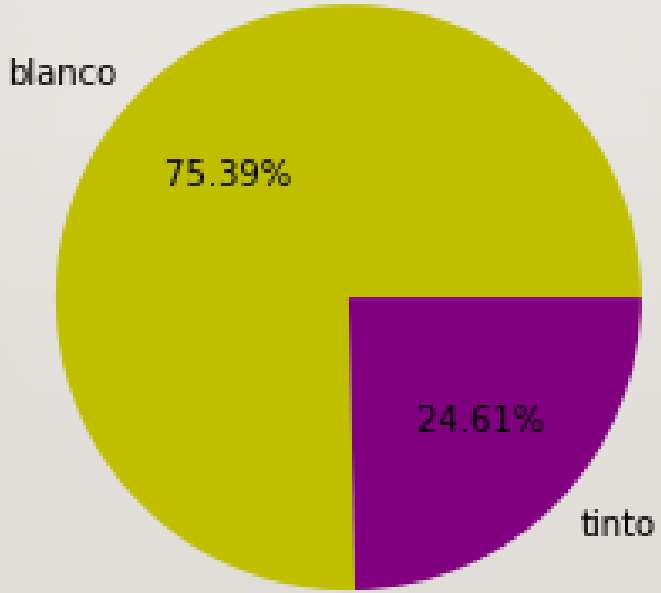


DIAGRAMA CIRCULAR 3D

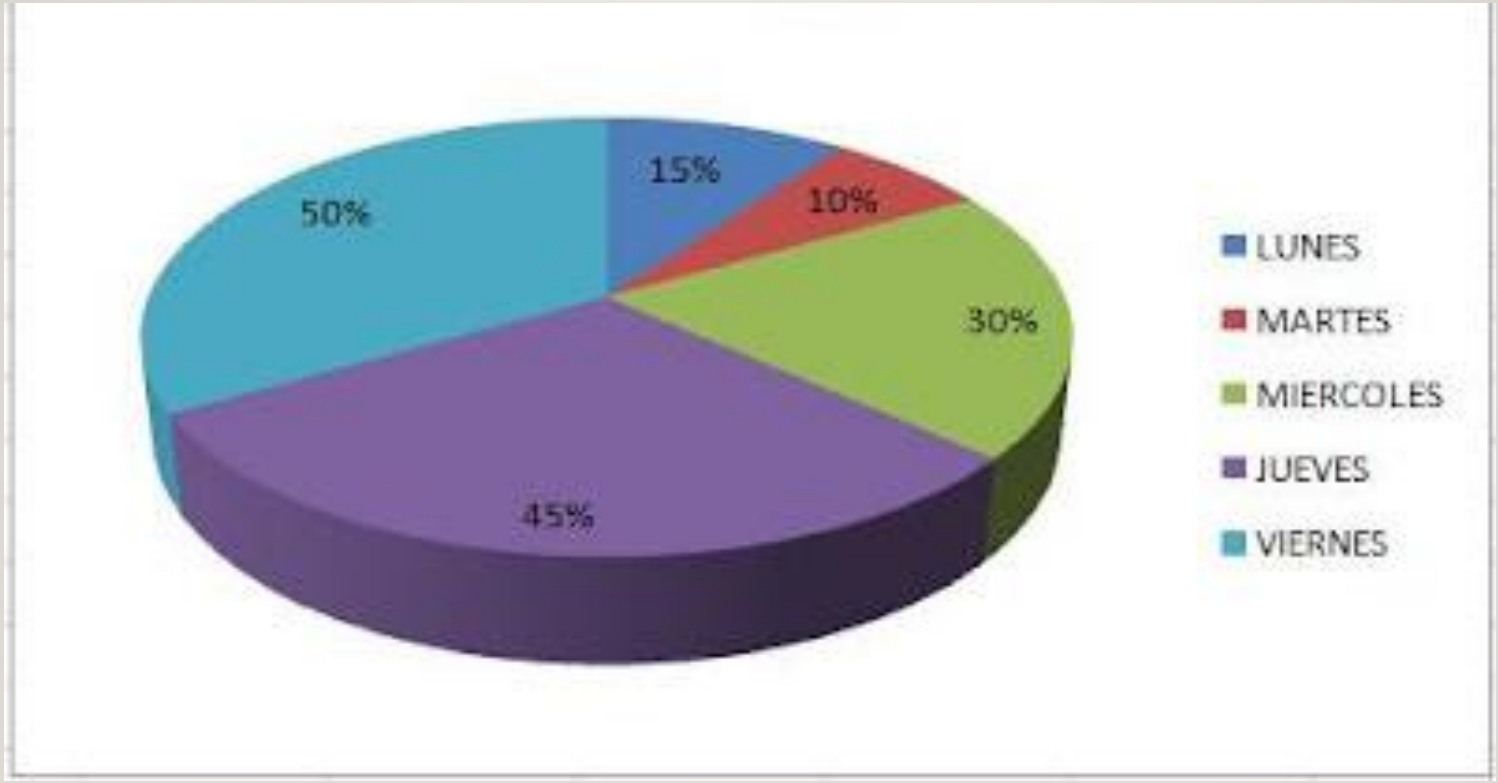


DIAGRAMA DE BARRAS

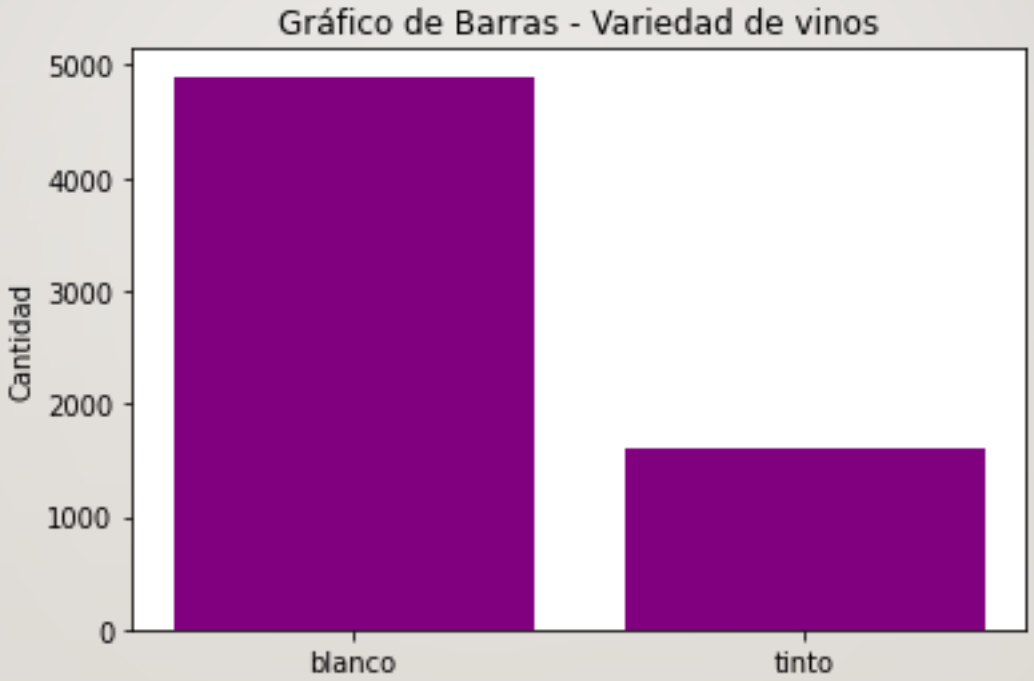


GRÁFICO DE BARRAS SUPERPUESTAS

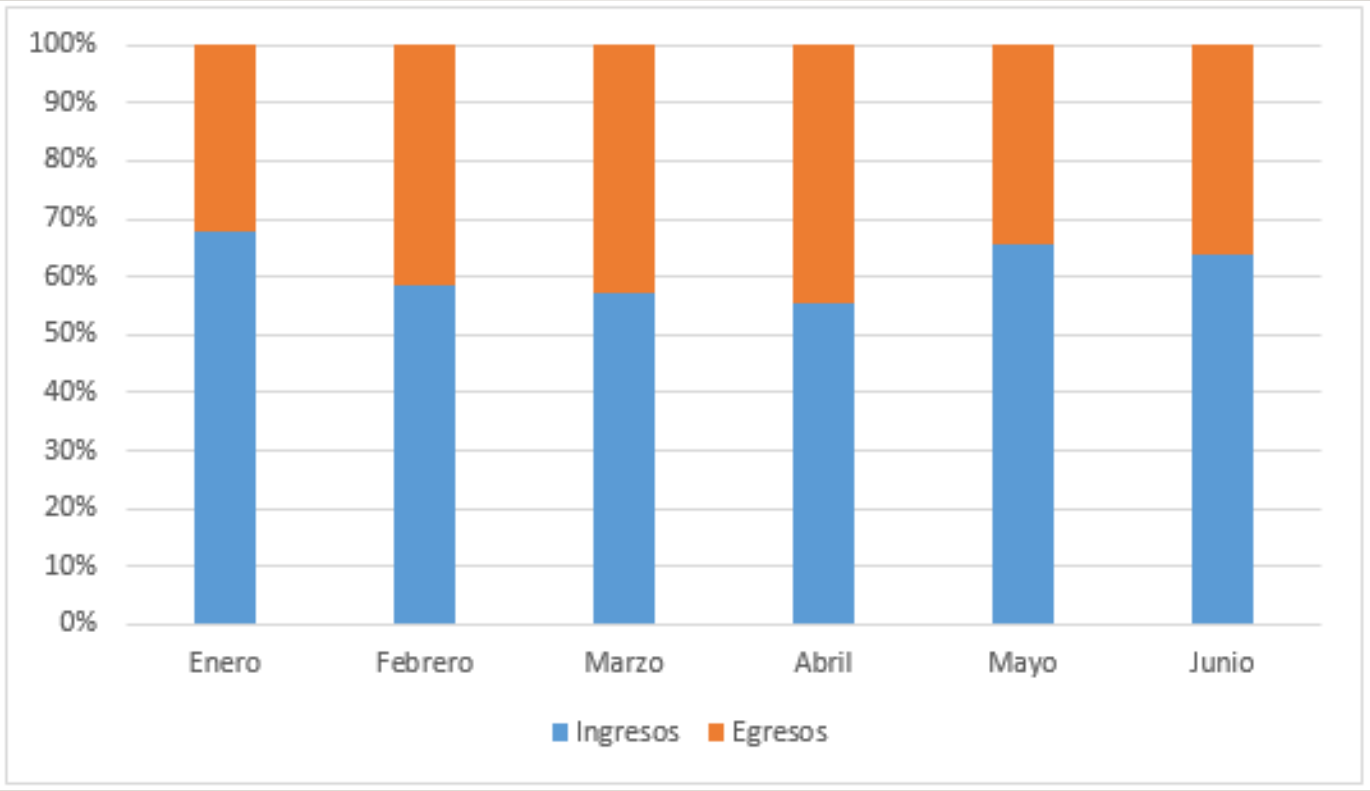


GRÁFICO DE BARRAS ADYACENTES

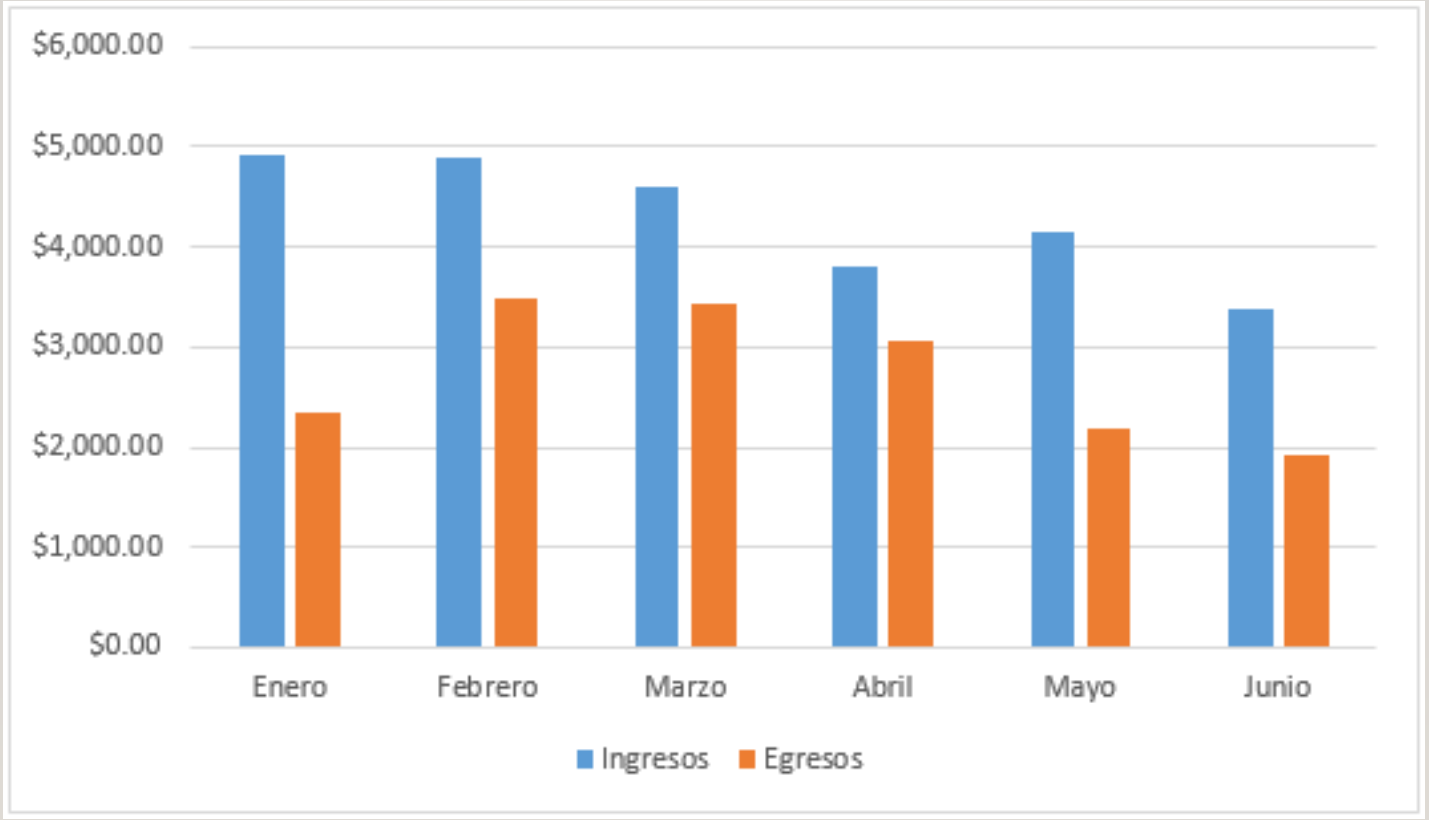
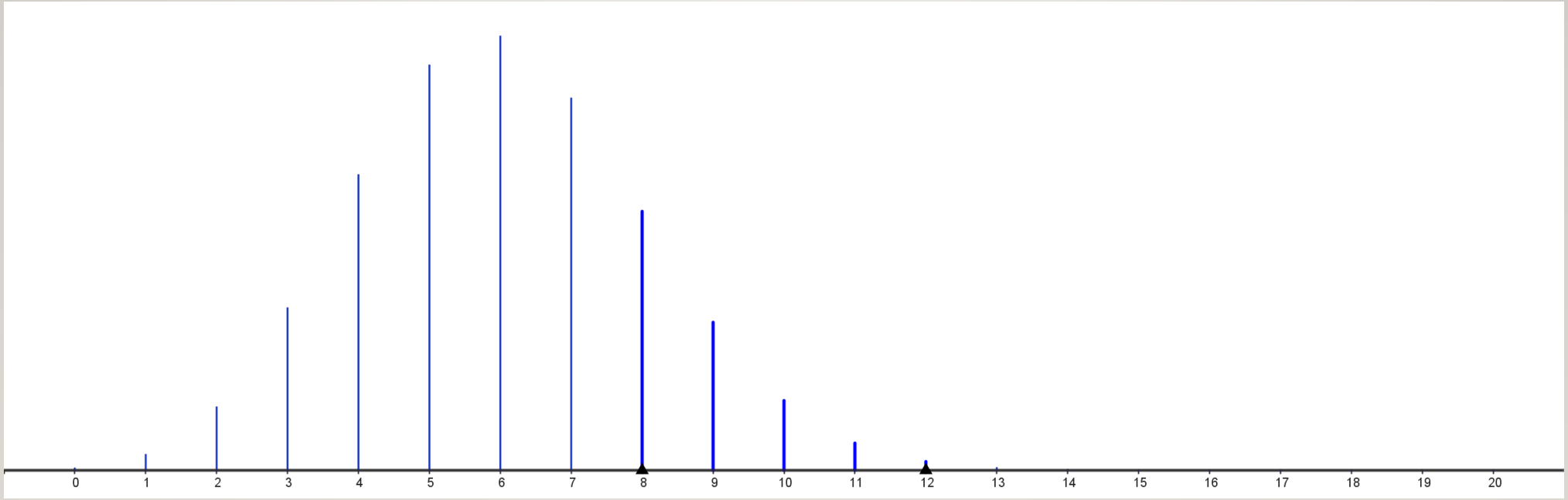
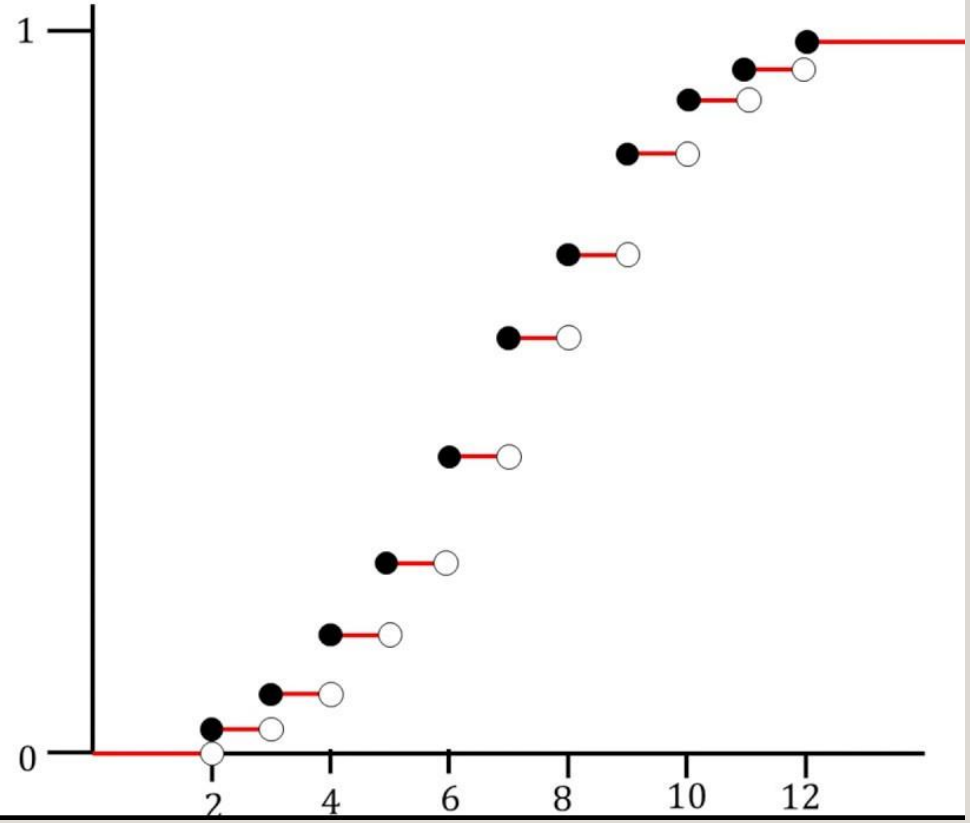


GRÁFICO DE BASTONES



FUNCIÓN DE DISTRIBUCIÓN

Valor	Probabilidad
x=2	1/36
x=3	2/36
x=4	3/36
x=5	4/36
x=6	5/36
x=7	6/36
x=8	5/36
x=9	4/36
x=10	3/36
x=11	2/36
x=12	1/36



FUNCIÓN DE DISTRIBUCIÓN ESCALONADA

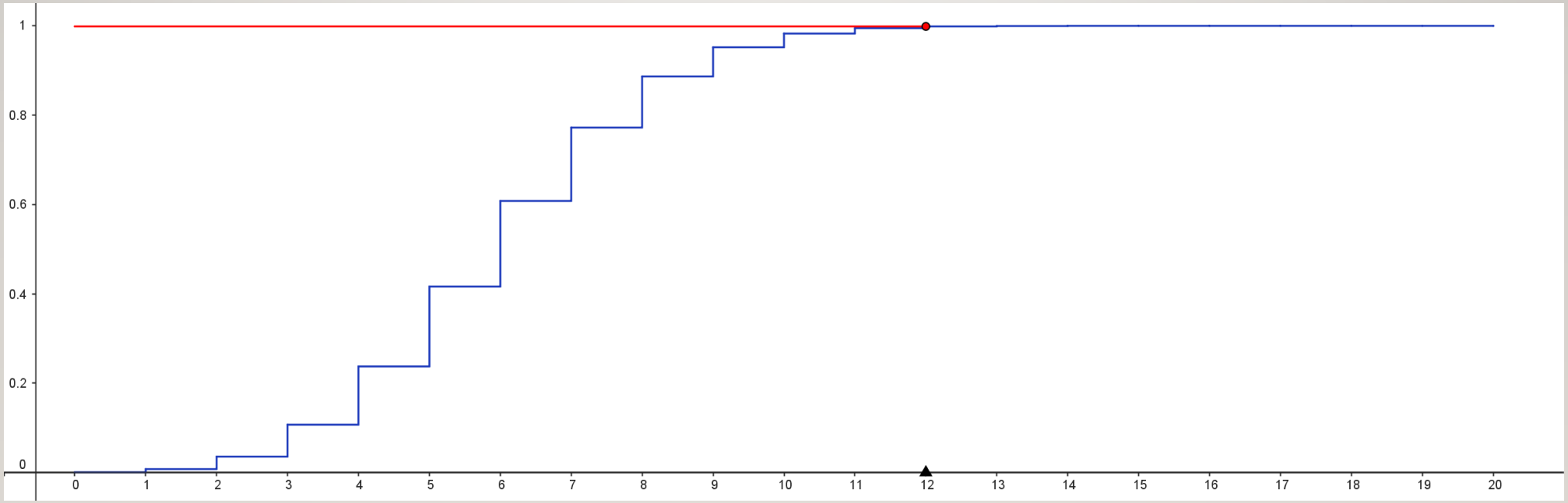


DIAGRAMA DE TALLO HOJA

134, 156, 158, 159, 160, 162, 164

Tallo | hoja

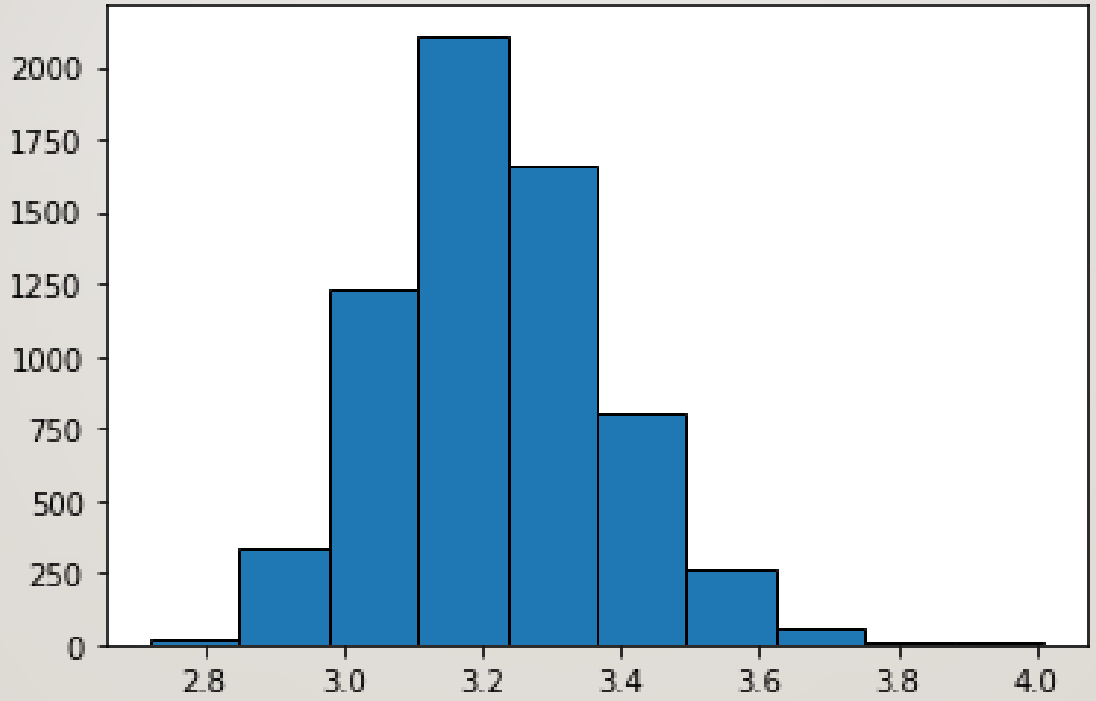
13 | 4

14 |

15 | 6 8 9

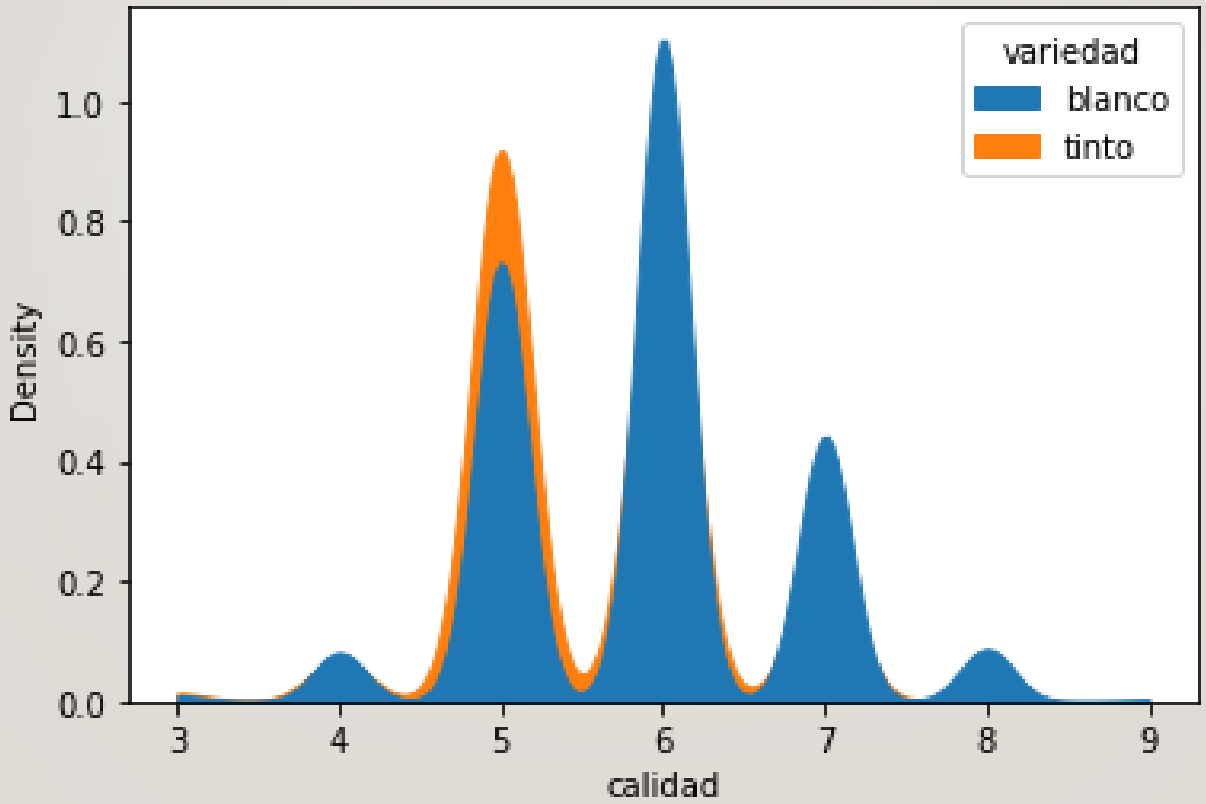
16 | 0 2 4

HISTOGRAMA

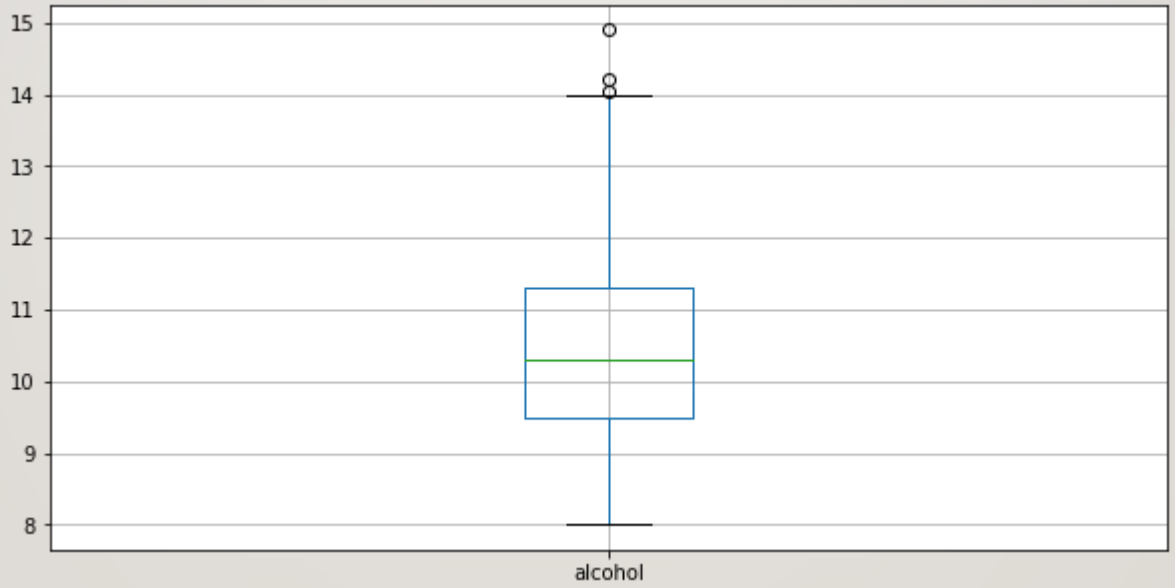


POLÍGONO DE FRECUENCIAS

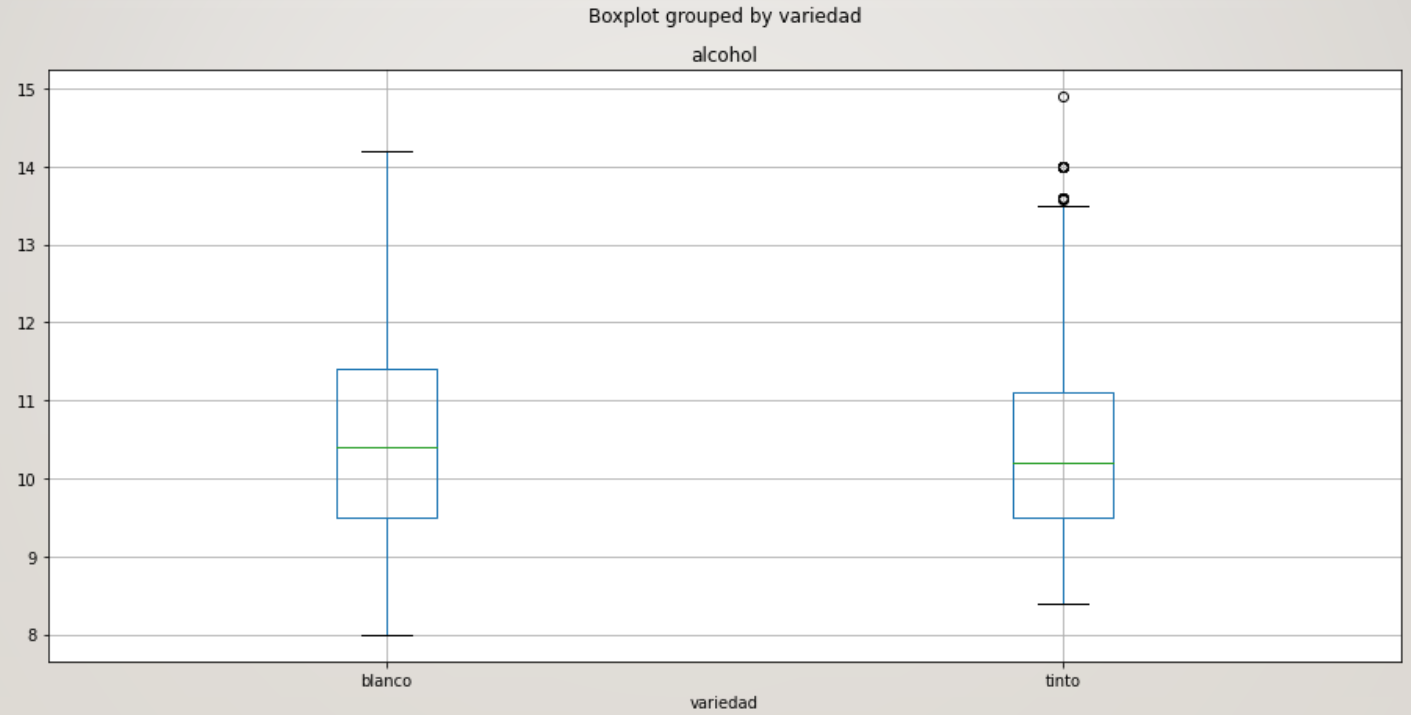
GRÁFICO DE DENSIDAD



BOXPLOT (DIAGRAMA DE CAJA Y BIGOTES



BOXPLOT COMPARATIVOS



¿PREGUNTAS?