
PI DATA STRATEGY & CONSULTING

Melina Solovey

Data Scientist, Pi Data Strategy & Consulting

DON'T BE MEAN... BE MEDIAN, MODE OR RANGE

1er Encuentro de Estadística para DS

PI DATA STRATEGY & CONSULTING

Agenda

- Introducción
- Tipos de Estadísticas
- Variable y sus tipos
- Conceptos Básicos
- Organización de los datos
- Representación gráfica

Introducción – ¿Qué es la estadística?

Es el estudio científico que consiste en la obtención o recopilación, organización, representación y análisis de observaciones, con el fin de describir el conjunto de datos obtenidos, así como inferir generalizaciones acerca de las características de todas las observaciones a partir de una parte de estas; para tomar las decisiones más acertadas en el campo de su aplicación.

Introducción – Etapas del proceso estadístico

Tipos de Estadísticas

Estadística descriptiva:

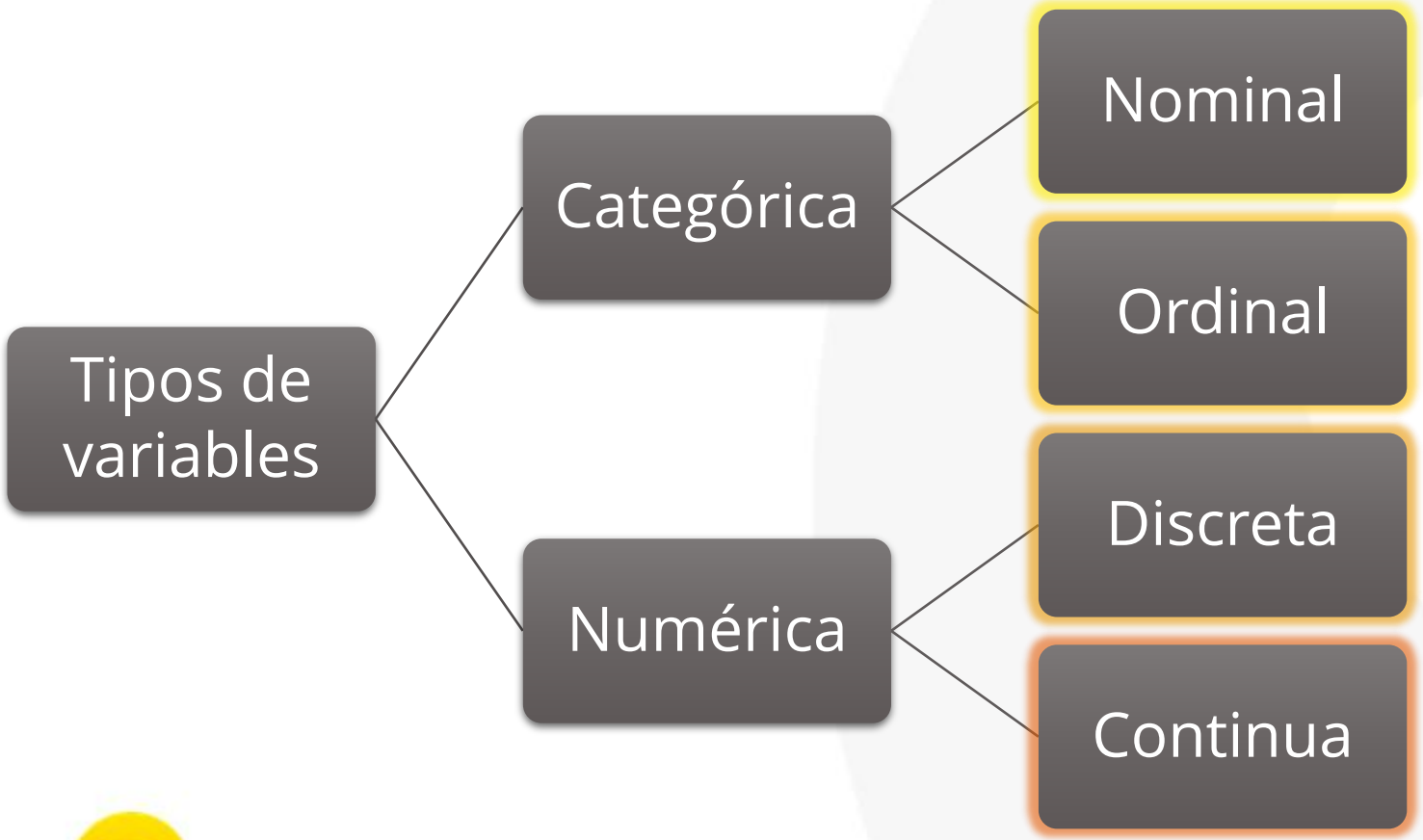
Describir el conjunto de datos obtenidos

Estadística inferencial:

Inferir generalizaciones acerca de las características de todas las observaciones a partir de un subconjunto de estas

Variable

Característica propia del individuo u objeto de estudio



Conceptos Básicos

- Individuo - Población – Muestra
- Medidas de Tendencia central
- Medidas de Dispersión
- Medidas de Forma

Medidas de resumen

Conceptos Básicos

Individuo

Cada uno de los elementos que compone la población.

Población

Conjunto de elementos a los que se les estudia una o más características.

Parámetros

- Media: μ
- Desv. Estándar: σ
- Varianza: σ^2
- Proporción: π

Muestra

Subconjunto representativo de la población.

Estadísticos

- Media: $\bar{X}$
- Desv. Estándar: S
- Varianza: S^2
- Proporción: p

Conceptos Básicos: Medidas de Tendencia Central

$$Median = middle\ value\ (when\ the\ data\ are\ arranged\ in\ order)$$

$$Mean = \frac{sum\ of\ all\ values}{total\ numer\ of\ values}$$

$$Mode = most\ common\ value$$

Escalas ordinales

Distribuciones sesgadas

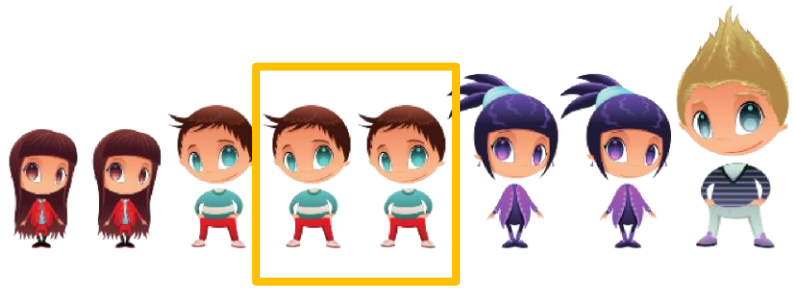
Variables Categóricas

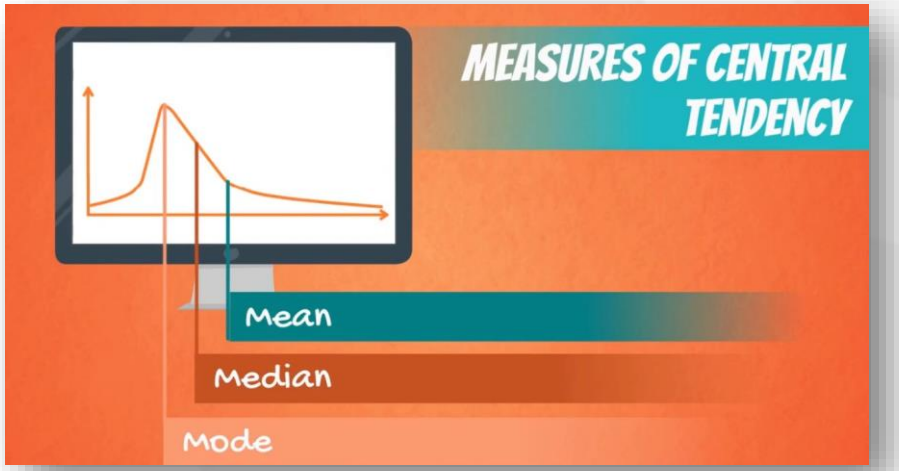
4 – 4 – 5 – 6 – 4 → me = 4,6 / md = 4

4 – 4 – 5 – 6 – 4 – 100 → me = 20,5 / md = 4,5

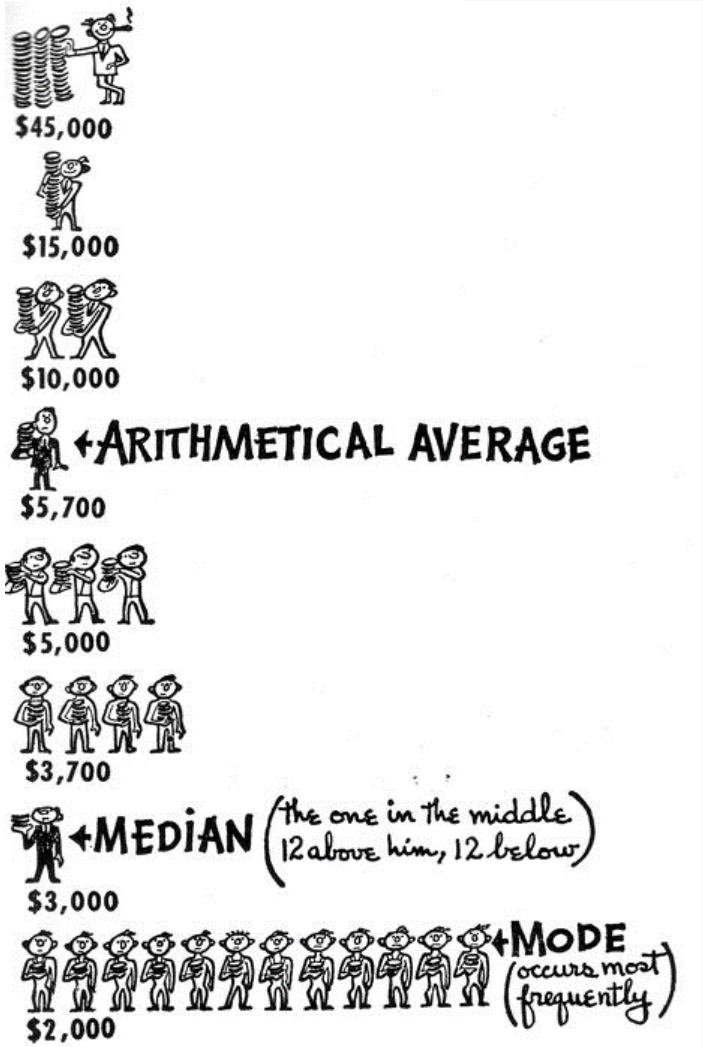
8 – 7 – 6 – 10 – 9 → me = 8 / md = 8

8 – 7 – 6 – 10 – 9 – 1 → me = 6,8 / md = 7,5





Conceptos Básicos: Medidas de Tendencia Central



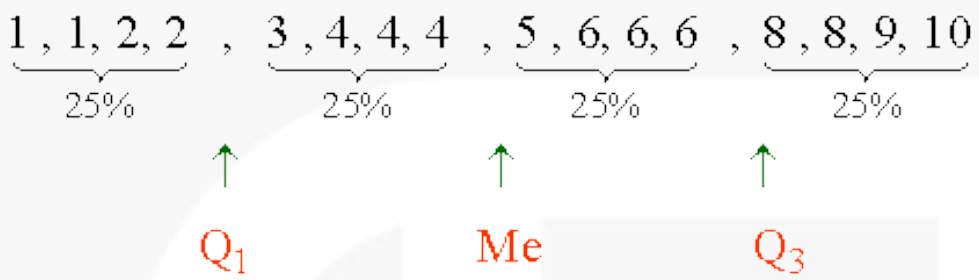
Conceptos Básicos: Medidas de Dispersión

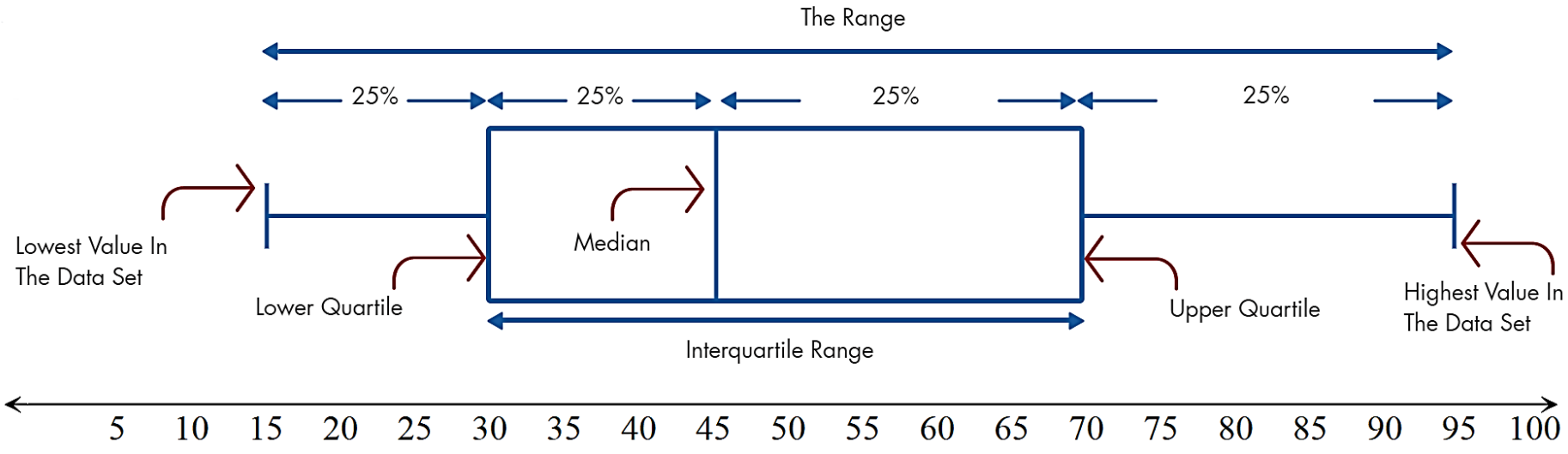
Rango

Rango Intercuartílico

Desviación estándar

Varianza

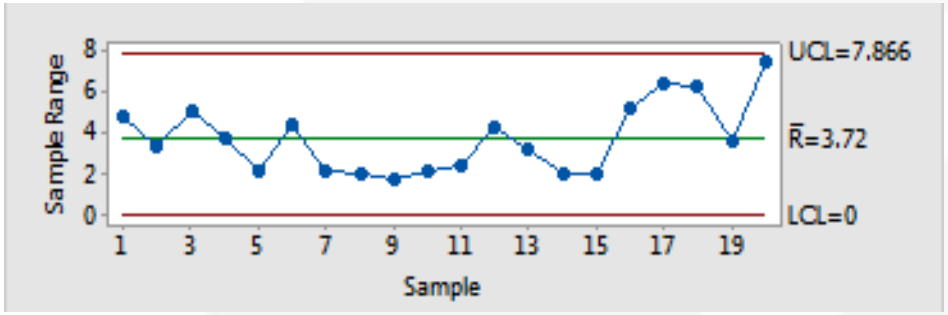


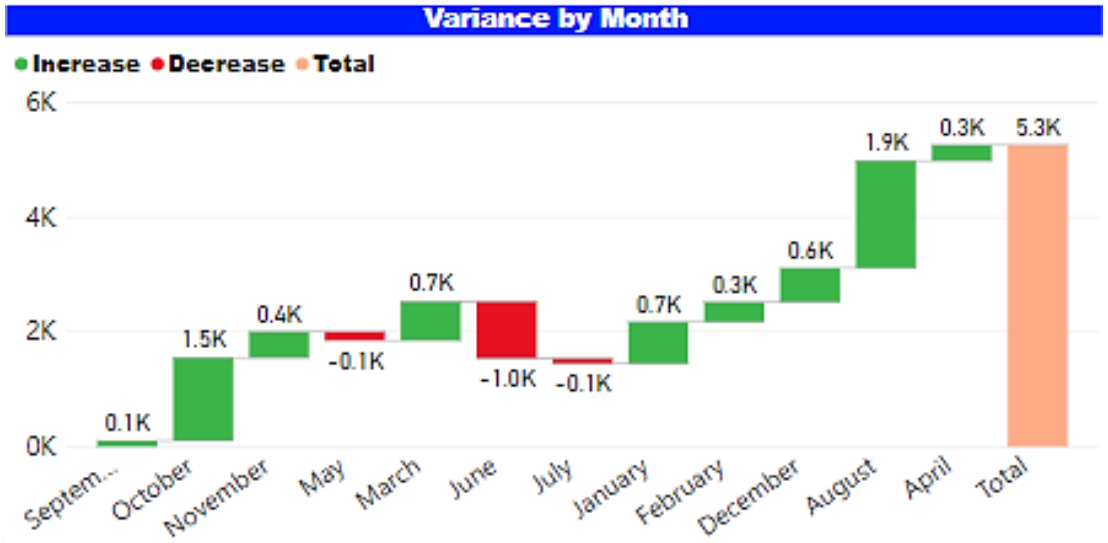


Conceptos Básicos: Medidas de Dispersión

Rango

$$R = X_{max} - X_{min}$$





Conceptos Básicos: Medidas de Dispersión

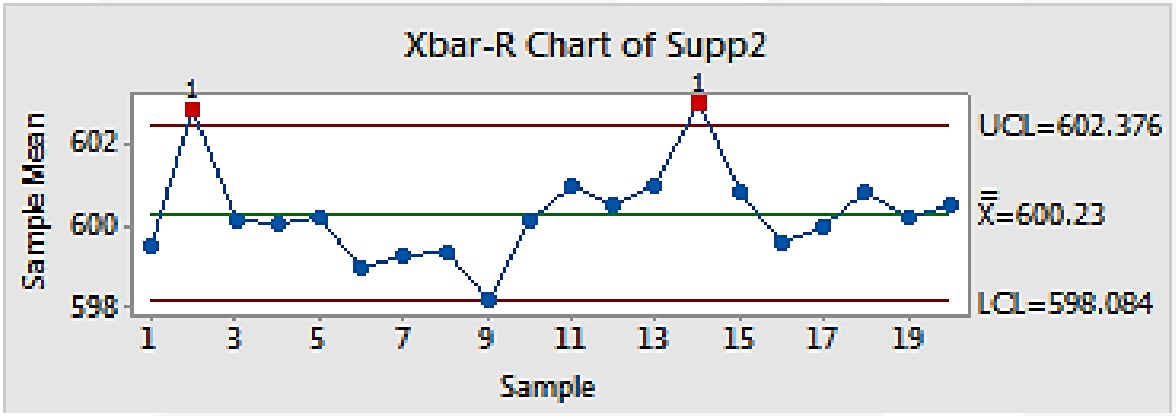
Varianza

Analicemos...

$$S^2 = \frac{\sum (X_i - \mu)^2}{n - 1}$$

desvío

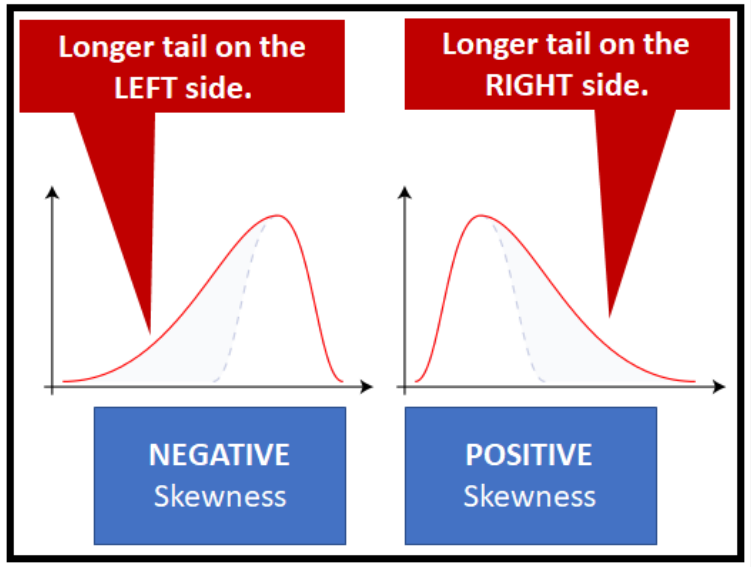
SCe

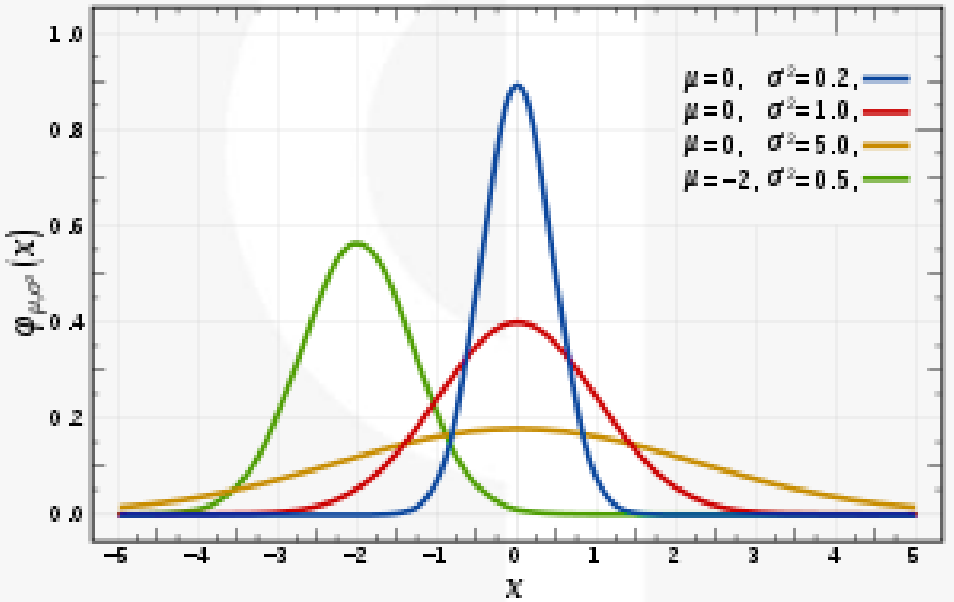


Conceptos Básicos: Medidas de Forma

Asimetría

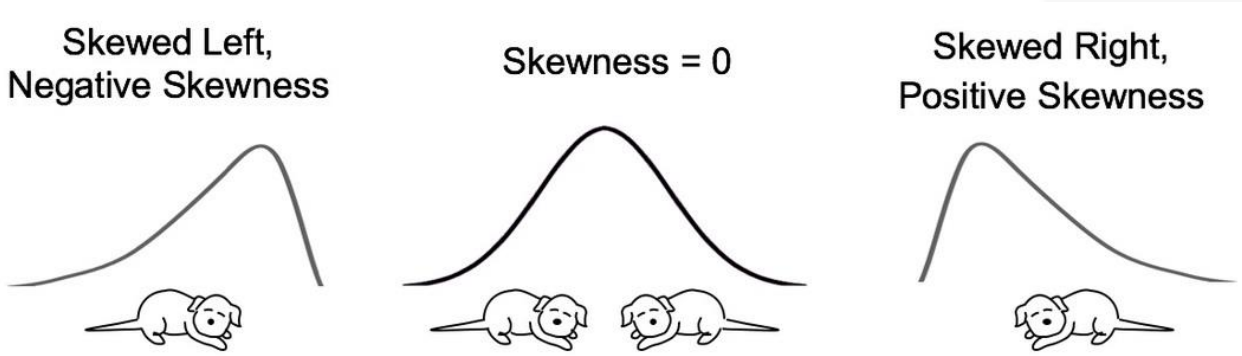
Curtosis





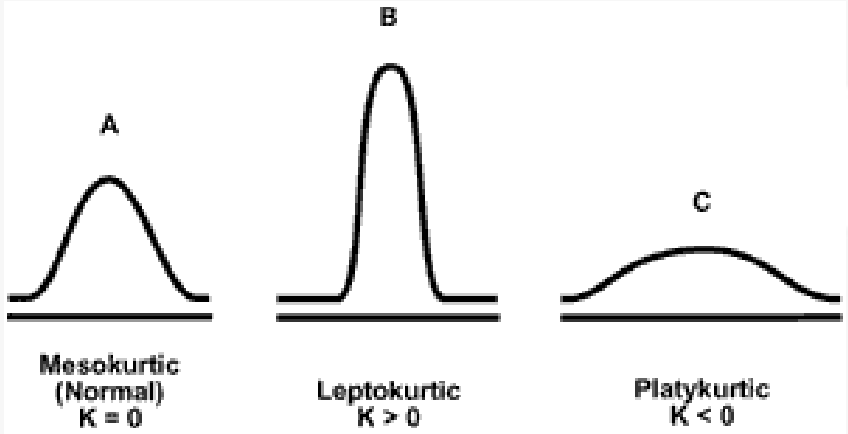
Conceptos Básicos: Medidas de Forma

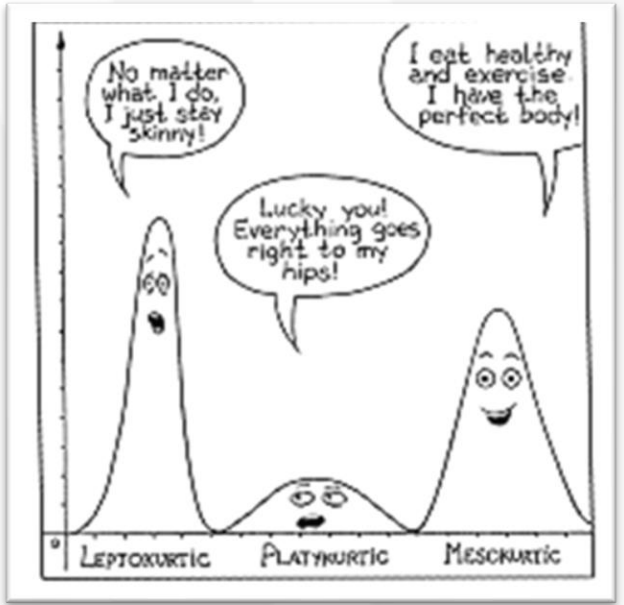
Asimetría



Reproduced by permission of John Wiley and Sons
from the book, Statistics from A to Z -- Confusing Concepts Clarified

Curtosis





Organización de los datos

Tabla de Frecuencia

Type of Pet	Tally	Frequency
Dog	𝍸 𝍸 ||	12
Cat	𝍸 ||	7
Goldfish	𝍸 |	6
Budgie	|||	3
Hamster	||	2
Lizard	|	1
Snake	|	1
Rabbit	|||	3

Class (Rs.)	Tally Marks	Frequency Students
20 - 30	𝍸	5
30 - 40	𝍸 |||	8
40 - 50	𝍸 ||||	9
50 - 60	𝍸 𝍸	10
60 - 70	𝍸 |	6
70 - 80	||	2
Total		**40**

Number of Children y	*Frequency f*	*Relative Frequency* $rf = f/n$	*Percentage Frequency* $p = 100 * rf$
0	1	0.10	10.00
1	0	0.00	00.00
2	2	0.20	20.00
3	1	0.10	10.00
4	2	0.20	20.00
5	4	0.40	40.00
	$n = 10$	1.00	100.00%

Organización de los datos

Tabla de Contingencia

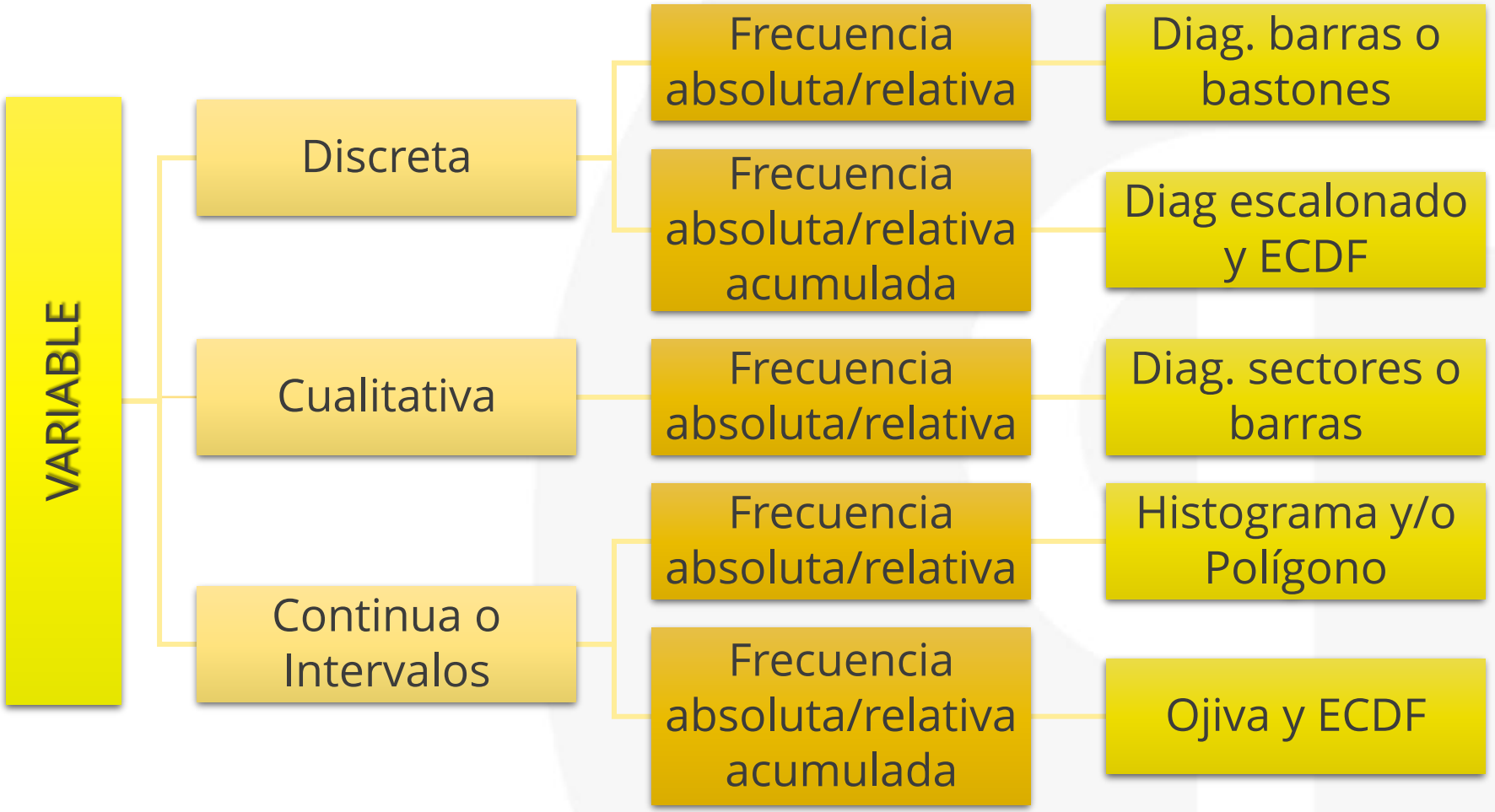
What flavor of ice cream would you pick?			
	Chocolate	Vanilla	Neither
Children	40	22	15
Teens	12	16	45
Adults	55	54	10
Total	107	92	70

Superpower	Male	Female	TOTAL
Flight	26	11	37
Invisibility	14	31	45
Other	10	8	18
TOTAL	50	50	100

Tabla de contingencia Interes ordenada * edad optimizada a P1

			edad optimizada a P1					
			< 27	27 - 36	36 - 47	47 - 72	72 - P66	Total
Interes ordenada	Nada	Recuento	90	110	148	427	361	1136
		% dentro de edad optimizada a P1	6,0%	6,5%	8,1%	15,4%	32,5%	12,8%
	Poco	Recuento	335	414	467	699	288	2203
		% dentro de edad optimizada a P1	22,2%	24,6%	25,6%	25,2%	25,9%	24,7%
	Bastante	Recuento	634	764	804	1202	323	3727
		% dentro de edad optimizada a P1	41,9%	45,3%	44,1%	43,3%	29,1%	41,9%
	Mucho	Recuento	453	397	404	446	138	1838
		% dentro de edad optimizada a P1	30,0%	23,6%	22,2%	16,1%	12,4%	20,6%
Total		Recuento	1512	1685	1823	2774	1110	8904
		% dentro de edad optimizada a P1	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%

Representaciones gráficas



Representaciones gráficas

Gráficos Estadísticos

Gráfico de bastones

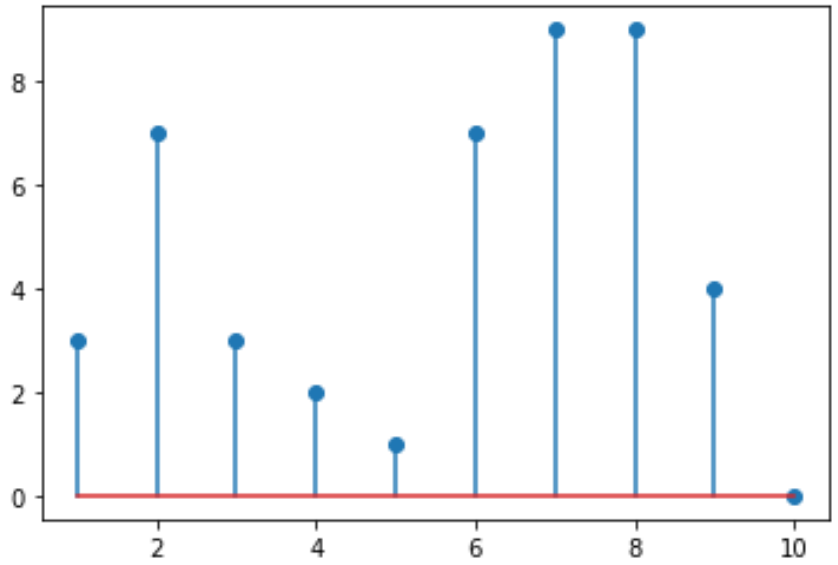
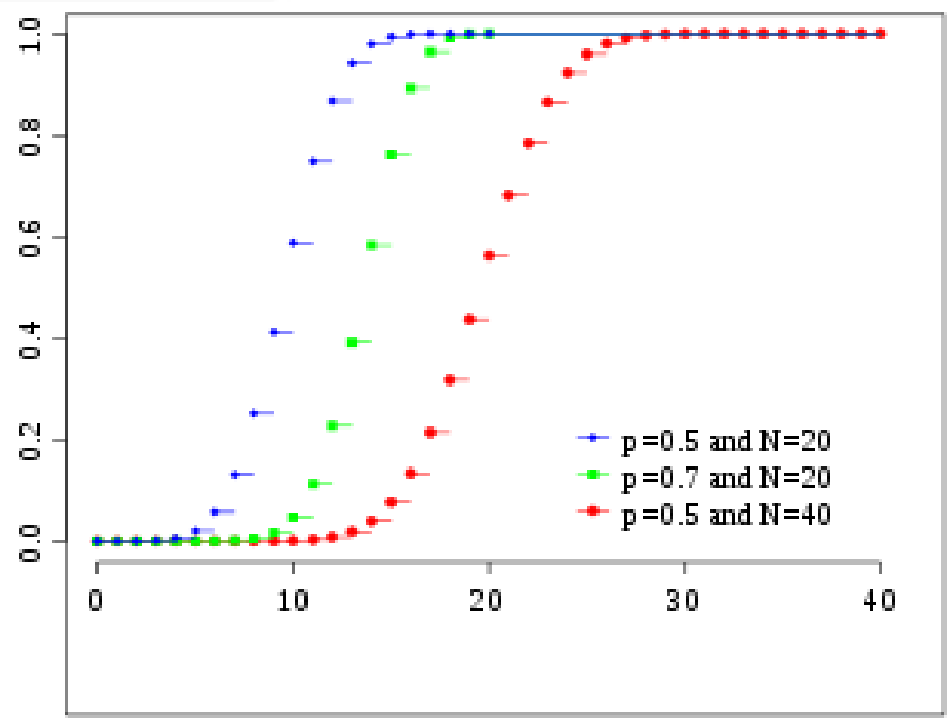
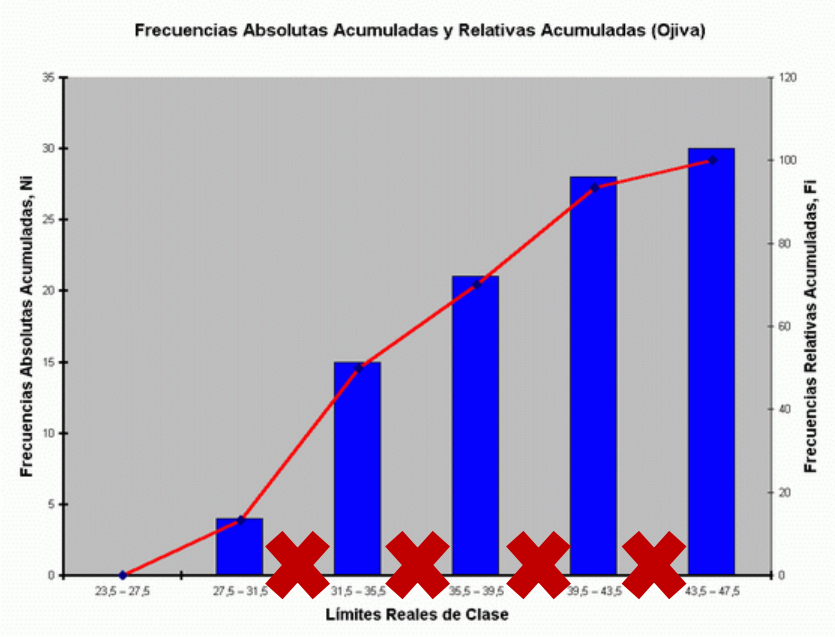


Gráfico escalonado (ECDF discreta)

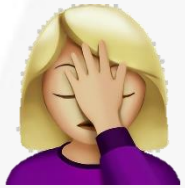


Representaciones gráficas

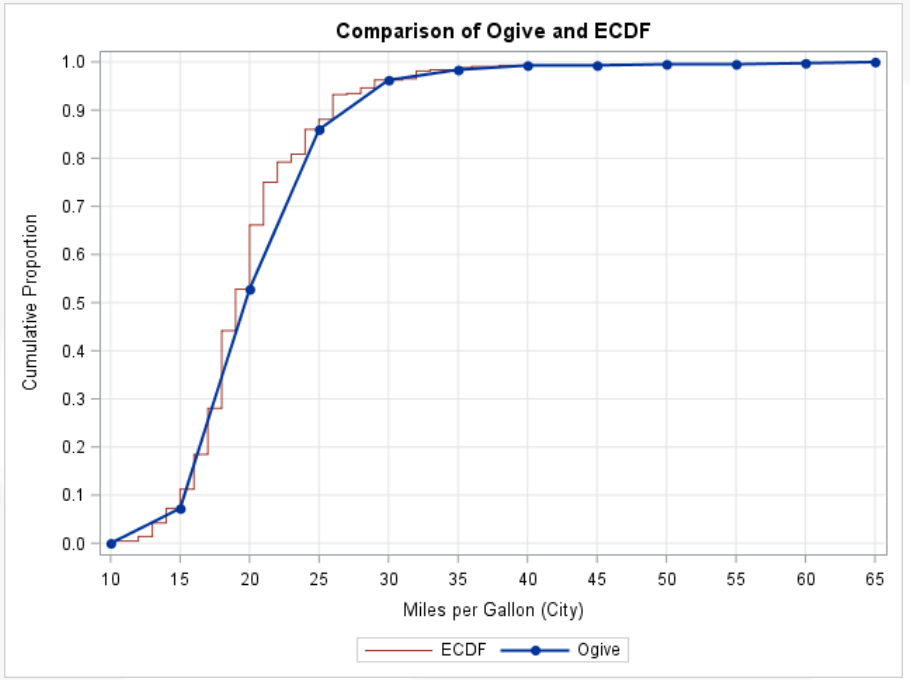
Gráficos de frecuencias acumuladas



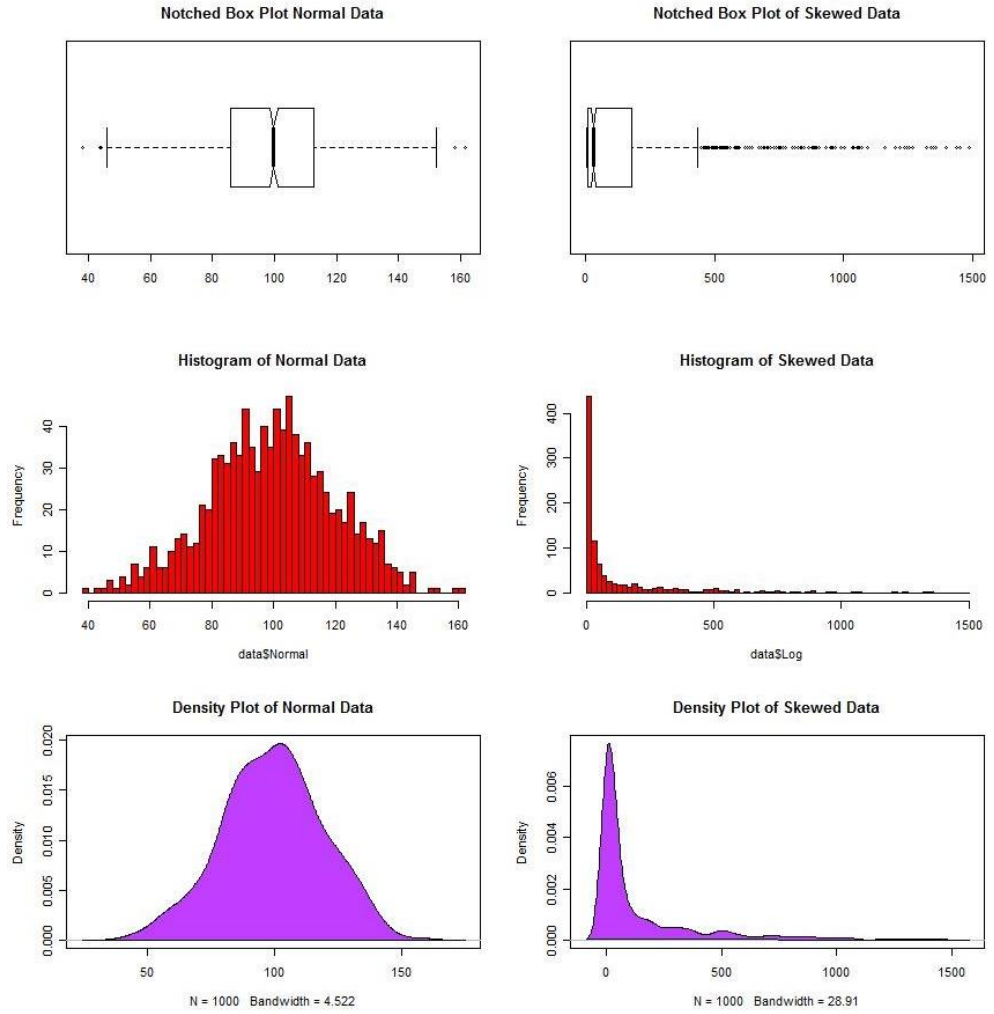
Función de densidad acumulada empírica (ECDF)

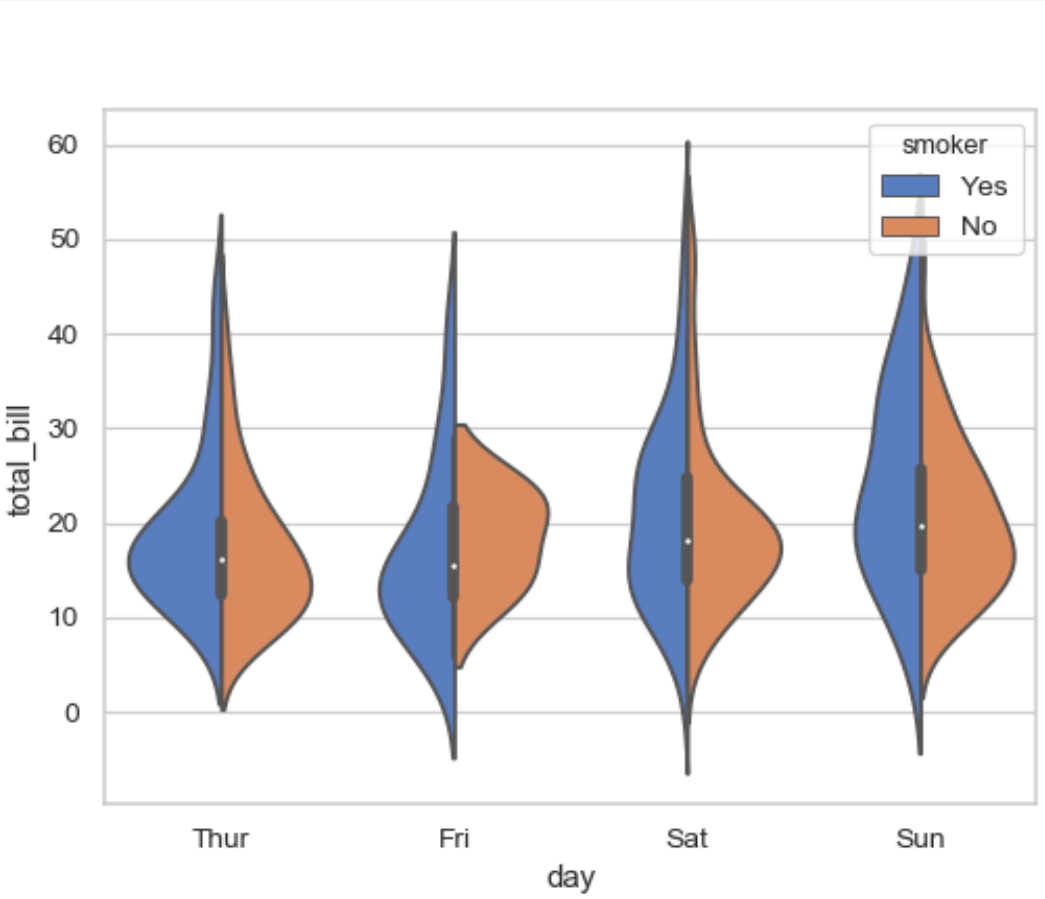


¡Cuidado!

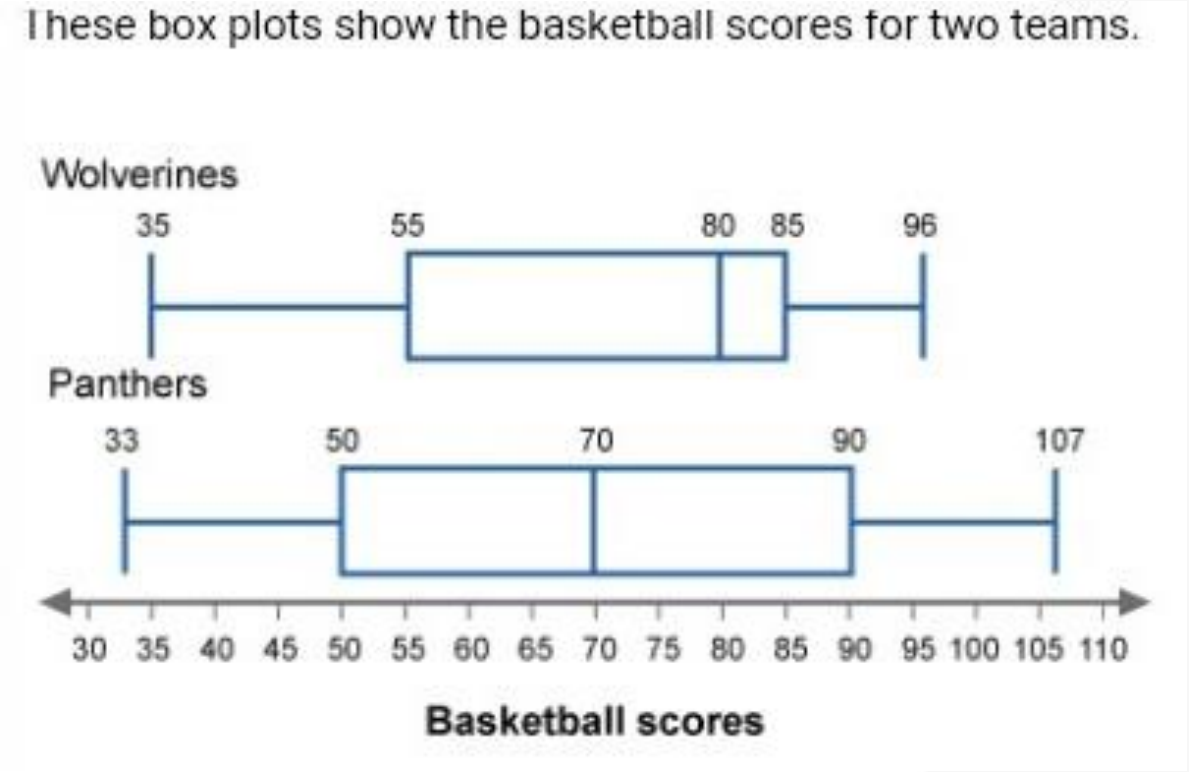


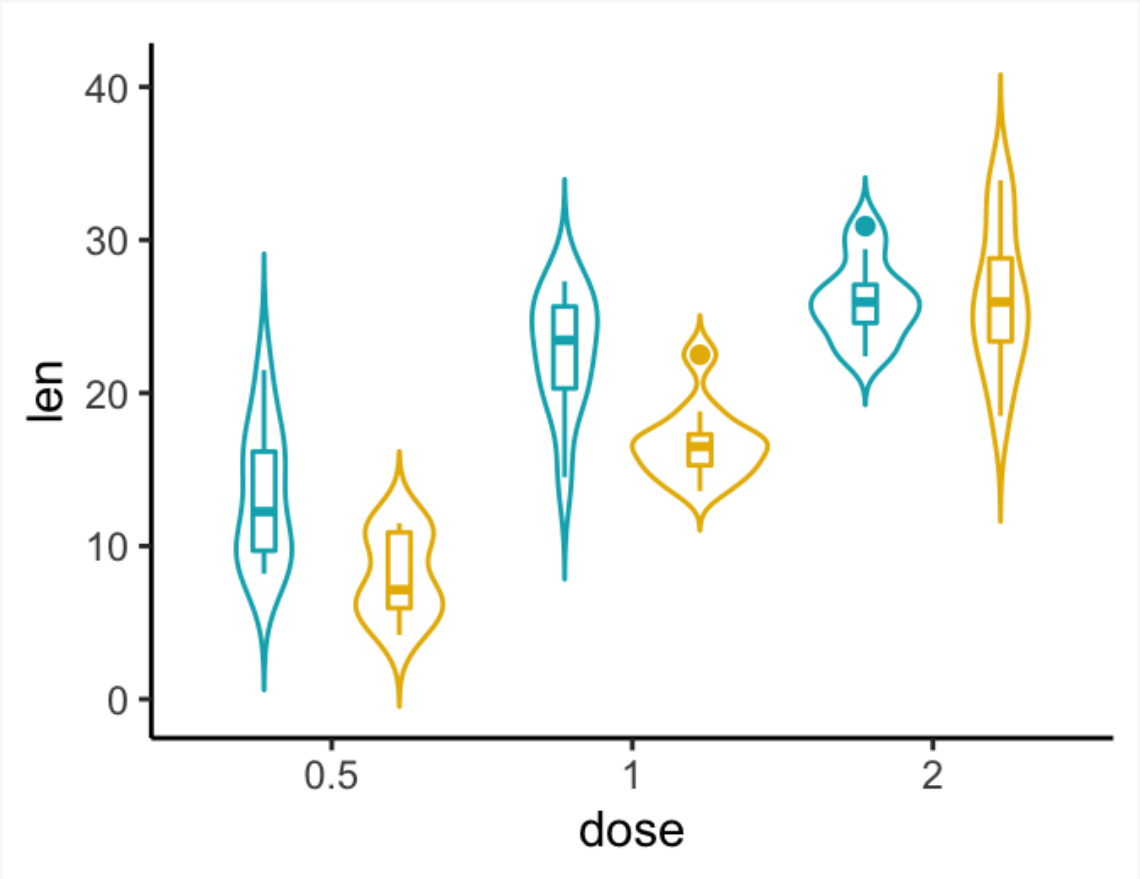
Representaciones gráficas



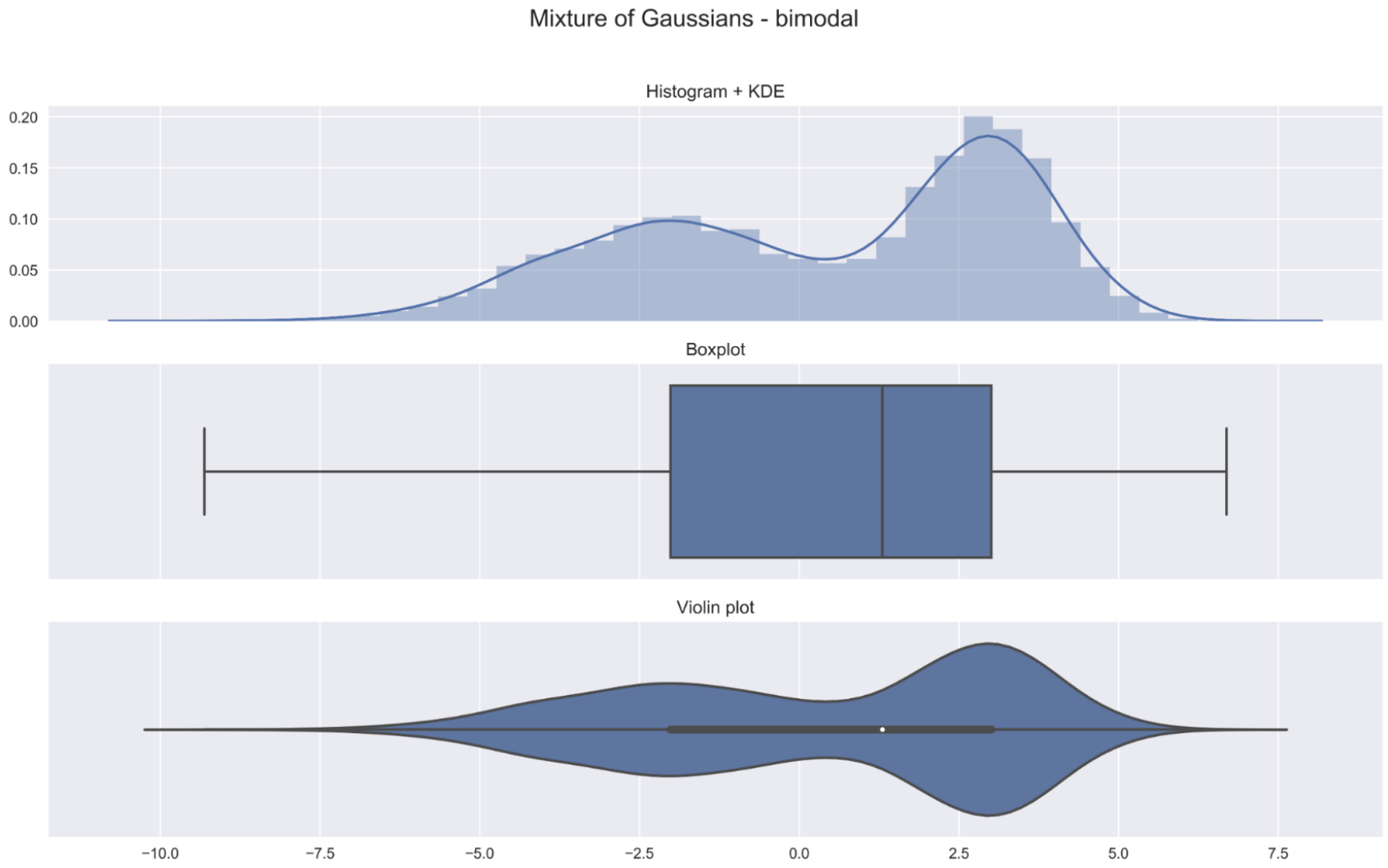


Representaciones gráficas





Representaciones gráficas



Librerías Python para Estadística

- **pandas**: librería muy conocida para análisis de datos en series y dataframes. Tiene funciones muy útiles para el análisis estadístico descriptivo y para trabajar con series de tiempo, valiéndose de las indexaciones.
- **numpy:** proporciona algunas funciones estadísticas que podemos aplicar de manera fácil sobre los arrays de Numpy.
- **matplotlib:** librería para visualizaciones, permite realizar gráficos de la distribución de los datos.
- **seaborn:** para la realización de gráficos estadísticos.
- **scipy.stats:** módulo del paquete científico Scipy, es un complemento para Numpy, de funciones estadísticas.
- **statsmodels:** pone a nuestra disposición un gran número de herramientas especializadas para la exploración de los datos, gráficas, distribuciones estadísticas, pruebas estadísticas, modelos estadísticos; es muy especializada.
- **pyMC:** módulo de Python que implementa modelos estadísticos bayesianos, así como la cadena de Markov Monte Carlo(MCMC).

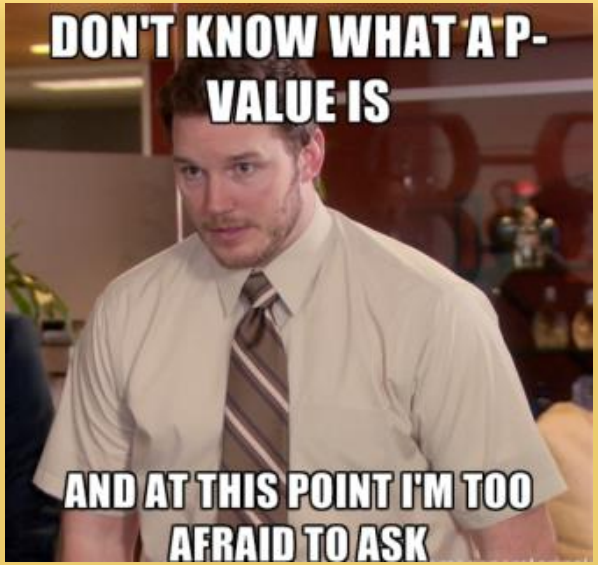
Note – it

¿Qué problema se plantea resolver?
¿Qué significan las variables?
¿Qué tipo de variables tengo?

¿Qué medidas estadísticas utilizar?

¿Qué tablas son más pertinentes?
¿Qué gráficos son más apropiados?

Diseña una metodología de análisis
...y aplica



(17-09-2020)

2do Encuentro: I don't know what a p-value is.... and now I'm too afraid to ask

Pruebas estadísticas paramétricas - variables cuantitativas.

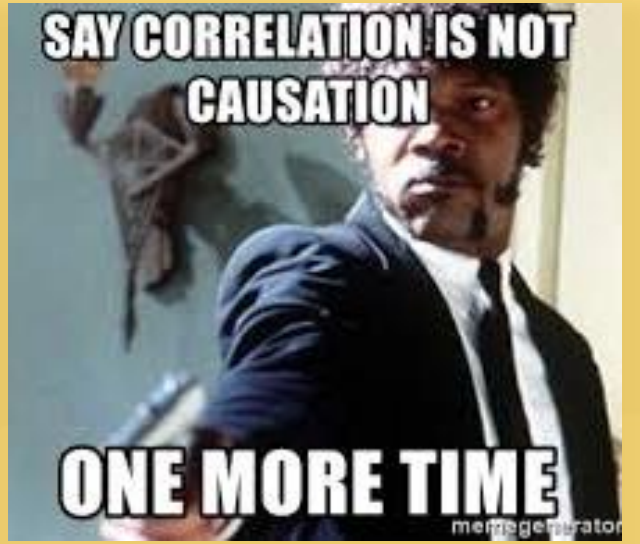
(Test - Media y comparación de medias - Varianza y comparación de varianzas - Proporción y comparación de proporciones - ANOVA – Tukey)

(01-10-2020)

3er Encuentro: Reject, or fail to reject... There is no accept

Pruebas estadísticas no paramétricas - variables cuantitativas.

(Test - U Mann-Whitney, Wilcoxon - Kruskal-Wallis, Friedman - Chi-cuadrado)



(15-10-2020)

4to Encuentro: Say correlation is not causation ... one more time

(Coeficientes de correlación – Pearson - Spearman – Kendall - Test de hipótesis - -prueba de independencia y homogeneidad).

MUCHAS GRACIAS!