EDEM

Máster en Data Analytics

Estadística con Python I

Miguel Rua del Barrio

Experiencia

Biostatistician
AMS Advanced Medical Services GmbH · Jornada completa
may. 2023 - actualidad · 10 meses
Valencia/València, Comunidad Valenciana / Comunitat Valenciana, España · En remoto

Bioestadística

University Professor
EDEM Escuela de Empresarios · Jornada parcial
ene. 2023 - actualidad · 1 año 2 meses
Valencia/València, Comunidad Valenciana / Comunitat Valenciana, España

Professor in subjects related with Statistics such as Statistics II for the BSc in Engineering and Management and Statistics with Python for the Master in Data Analytics for Business

Bioestadística

Statistician
Generalitat Valenciana
sept. 2020 - may. 2023 · 2 años 9 meses
Valencia/València, Comunidad Valenciana / Comunitat Valenciana, España

Python, Estadística y 4 aptitudes más

Researcher
BCAM - Basque Center for Applied Mathematics
ene. 2020 - sept. 2020 · 9 meses
Bilbao, País Vasco / Euskadi, España

Estadística bayesiana, Inferencia bayesiana y 5 aptitudes más

Data Scientist
Intelligent Data Analysis Laboratory (Universidad de Valencia)
ene. 2018 - jun. 2018 · 6 meses
Valencia

Public healthcare project (collaborating with Hospital Clínic Universitari) where models of ML were implemented in order to predict and classify the evolution of pneumonia in several patients.

Python, Aprendizaje automático y 3 aptitudes más

EDEM
Escuela de Empresarios

Universitat de València
Master's degree, Bioestadística
2018 - 2020

Series Temporales, Modelos Jerárquicos Bayesianos, Gestión de bases de datos, Minería de datos, Estadística espacial etc

Universitat de València
Grado, Matemáticas
2013 - 2018

Aptitudes: Estadística · Bioestadística

Karlsruher Institut für Technologie (KIT)
Grado, Matemáticas
2015 - 2016

Universidad Nacional de Educación a Distancia - U.N.E.D.
Máster, Formación de profesorado
sept. 2020 - sept. 2021

AMS ADVANCED MEDICAL SERVICES - PROGRESS

Your Competent Partner in the Pharmaceutical / Healthcare Industry since 1997

An **interdisciplinary team** of 300+ permanent employees and a **global network of long-term partners** and consultants enable us to offer the full spectrum of **services throughout the life cycle of your medicinal products / medical devices.**

800+
Clinical Studies Completed

250+
NIS/RWE

300+
AMNOG Projects

35000+
Sites

300+
Permanent Employees

93 %
Repeated Business

150+
Indications Experience

27+
Years in Business

EDEM
Escuela de Empresarios

AMS Services in Biostatistics include:

- Biostatistical consulting and writing for clinical trials and observational research studies
- Sample size calculations
- Randomization (with optional integration of the randomization list into the eCRF)
- Biostatistical review and contributions to study protocols/ study reports/posters/manuscripts
- Comprehensive and detailed statistical analysis plans (SAP)
- SAS Programming for the statistical analysis, of review listings, periodic safety reports or trial results reporting to EudraCT database
- Implementation of CDISC SDTM / ADaM standards.
- Health Economics, Patient surveys, Interim analysis, Adaptive designs, Burden of Disease studies, cost-effectiveness analyses, quality of life evaluation

Enfoque del módulo:

EDEM
Escuela de Empresarios

Contenidos:

- Nociones básicas de Estadística.
- Variables aleatorias y tipos.
- Estadística descriptiva (tablas, gráficos, estadísticos) .
- Definición y tipos de distribuciones de probabilidad.
- Contrastes de Hipótesis.
- Análisis de la Varianza (ANOVA).

Objetivos:

- Obtener fundamentos estadísticos necesarios.
- Analítica práctica de datos y representación gráfica univariada y bivariada.
- Uso de Python como herramienta para el análisis estadístico.
- Elaboración de informes estadísticos con Google Colab.

Metodología:

- 20h.: 5 Sesiones ES - 4 Sesiones FS
- Teoría-Práctica y trabajo del alumno.

Introducción

EDEM
Escuela de Empresarios

Estadística: Rama de las matemáticas encargada del estudio de los datos.

- Recogida.
- Manipulación.
- Análisis.
- Inferencia.
- Interpretación.

❖ Principal objetivo: Obtener conclusiones para una población a través del estudio de un subconjunto de ella.

Podemos distinguir dos tipos: **Descriptiva** e **inferencial**.

Nociones básicas

EDEM
Escuela de Empresarios

- Población: Conjunto de todos los individuos de nuestro interés. Denotado por N.
- Muestra: Subconjunto de la población. Tiene que ser representativa.
- Parámetro: Característica específica de la población. Calculada con datos poblacionales.
- Estadístico: Característica específica de la muestra. Calculada con datos muestrales. Servirán para estudiar los parámetros.

Ejemplo:

Estudiar la altura media de todos los españoles (**población**). Dicha media poblacional es el **parámetro**. Tomamos aleatoriamente una **muestra** de 1000 personas. Calculamos su media muestral (**estadístico**).

❖ Hay toda una ciencia detrás para recoger buenas muestras y realizar experimentos estadísticos. La calidad del dato es primordial. (Muestreo, DOE)

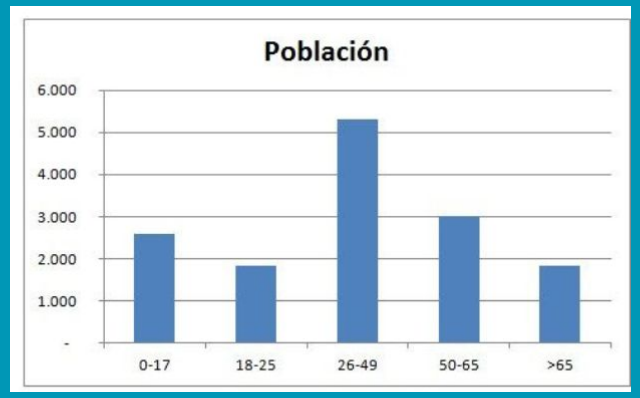
Estadística descriptiva

EDEM
Escuela de Empresarios

Concepto

La estadística descriptiva consiste en describir nuestra muestra con la ayuda de herramientas como gráficos o cálculos de estadísticos, entre otros:

- Representación de datos (tablas, gráficos)
- Recogida de datos
- Resumen de datos con estadísticos (media, mediana, moda, ...)



Edad	Población
0-17	2.592
18-25	1.834
26-49	5.314
50-65	3.012
>65	1.839

Source: *ayudahispano-3000.blogspot.com*

Nota	Alumnos
Suspenso	7
Aprobado	12
Notable	14
Sobresaliente	4

Estadística Inferencial

EDEM
Escuela de Empresarios

Concepto

En la estadística inferencial, se utilizan los resultados de los datos para hacer predicciones, obtener conclusiones y contribuir obtener información para la toma de decisiones.

Señalamos dos aplicaciones:

- Estimación del parámetro de una población usando una muestra.
- Probar una hipótesis sobre un parámetro de la población (media de altura =1,8m)

Es decir, *la inferencia estadística es el proceso de obtener conclusiones de una población basadas en los resultados de la muestra.*

Variables aleatorias

EDEM
Escuela de Empresarios

Concepto

Una variable aleatoria es una formalización matemática de una cantidad o objeto que depende de eventos aleatorios.

En otras palabras, la característica que queremos estudiar de la población será probablemente una variable aleatoria.

Las variables aleatorias están sujetas al azar. Pueden tomar diferentes valores.

- Algunos ejemplos de variables aleatorias:

  - color de ojos
  - nacionalidad
  - altura
  - fumador

Tipos de variables

EDEM
Escuela de Empresarios

Clasificación

VARIABLES CATEGÓRICAS

representan un atributo

→ **Nominativas:** nombran, etiquetan o categorizan atributos particulares que están midiendo

→ **Ordinales:** atributos con un orden entre ellos

VARIABLES CUANTITATIVAS:

representan una cantidad

→ **Discretas:** toman valores naturales. Conteo. No pueden tomar valores más finos.

→ **Continuas:** pueden ser reducidas a niveles siempre más finos. Infinitos valores posibles.

Nota: la misma variable puede ser clasificada en diferentes grupos dependiendo de cómo la medimos.

Ejercicios

EDEM
Escuela de Empresarios

→ Clasifica estas variables en sus diferentes tipos

a) Notas en una asignatura
b) Estado de Salud
c) Precio de gasolina
d) Número de hijos
e) Número de DNI
f) Fumador (medido en 0 y 1)
g) Grados que habéis estudiado

Librerías de Python

EDEM
Escuela de Empresarios

En este caso, los principales módulos que Python nos ofrece para trabajar con probabilidad y estadística son:

NumPy el popular paquete matemático de Python, se utiliza tanto que mucha gente ya lo considera parte integral del lenguaje. Nos proporciona algunas funciones estadísticas que podemos aplicar fácilmente sobre los *arrays* de Numpy.

- **scipy.stats:** Este submódulo del paquete científico Scipy es el complemento perfecto para Numpy, las funciones estadísticas que no encontremos en uno, las podemos encontrar en el otro.
- **statsmodel**: Esta librería nos brinda un gran número de herramientas para explorar datos, estimar modelos estadísticos, realizar pruebas estadísticas y muchas cosas más.

matplotlib Es la librería más popular en Python para visualizaciones y gráficos. Ella nos va a permitir realizar los gráficos de las distintas distribuciones de datos.

- **seaborn**: Esta librería es un complemento ideal de matplotlib para realizar gráficos estadísticos.
- **pandas**: Esta es la librería más popular de manejo de data frames.

Descriptiva de variables categóricas

EDEM
Escuela de Empresarios

Tabla de Frecuencias

Conceptos:

- Frecuencia absoluta
- Frecuencia relativa
- Porcentaje
- Frecuencias acumuladas

X_i	Frecuencia absoluta (n_i)	Frecuencia absoluta acumulada (N_i)	Frecuencia relativa (fi = ni/N)	Frecuencia relativa acumulada (Fi=Ni/N)
1	7	7	0,06	0,06
2	19	26	0,15	0,21
3	25	51	0,20	0,41
4	12	63	0,10	0,50
5	23	86	0,18	0,69
6	15	101	0,12	0,81
7	8	109	0,06	0,87
8	16	125	0,13	1,00
Total	125	125	1	1

Descriptiva de Variables Categóricas

Diagrama de barras

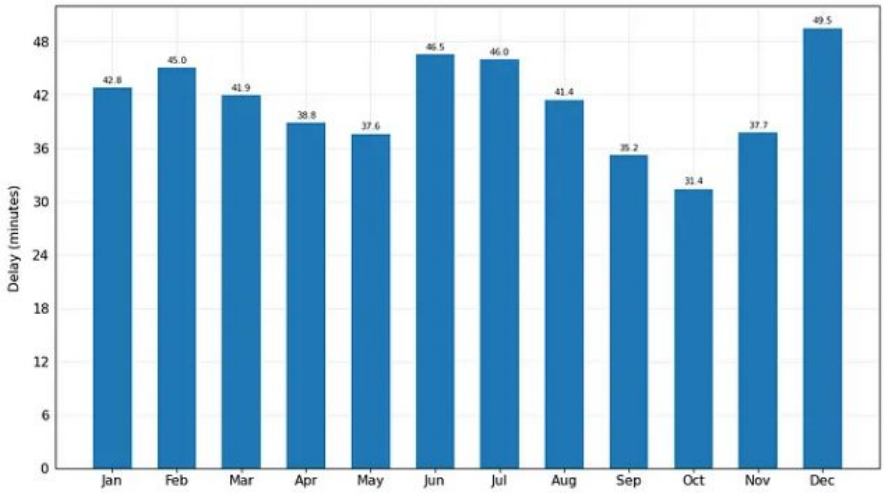
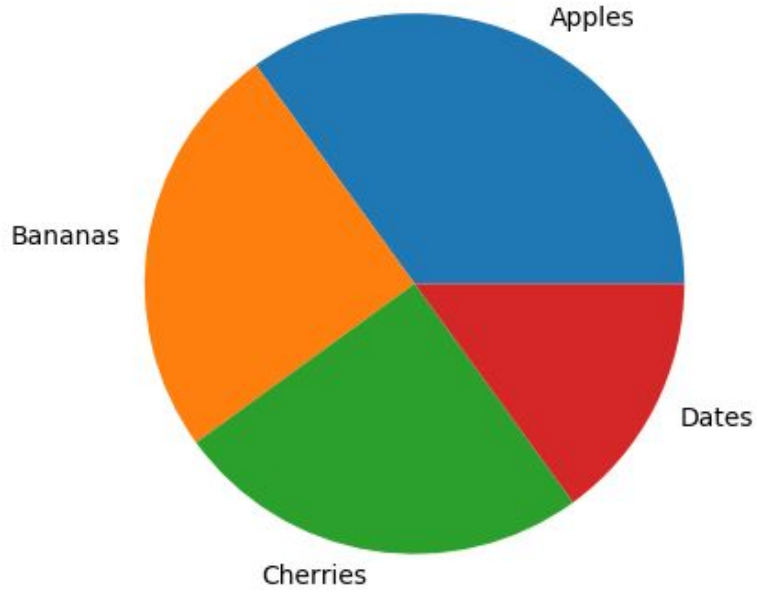


Diagrama de sectores



EDEM
Escuela de Empresarios

Descriptiva de Variables Cuantitativas Continuas

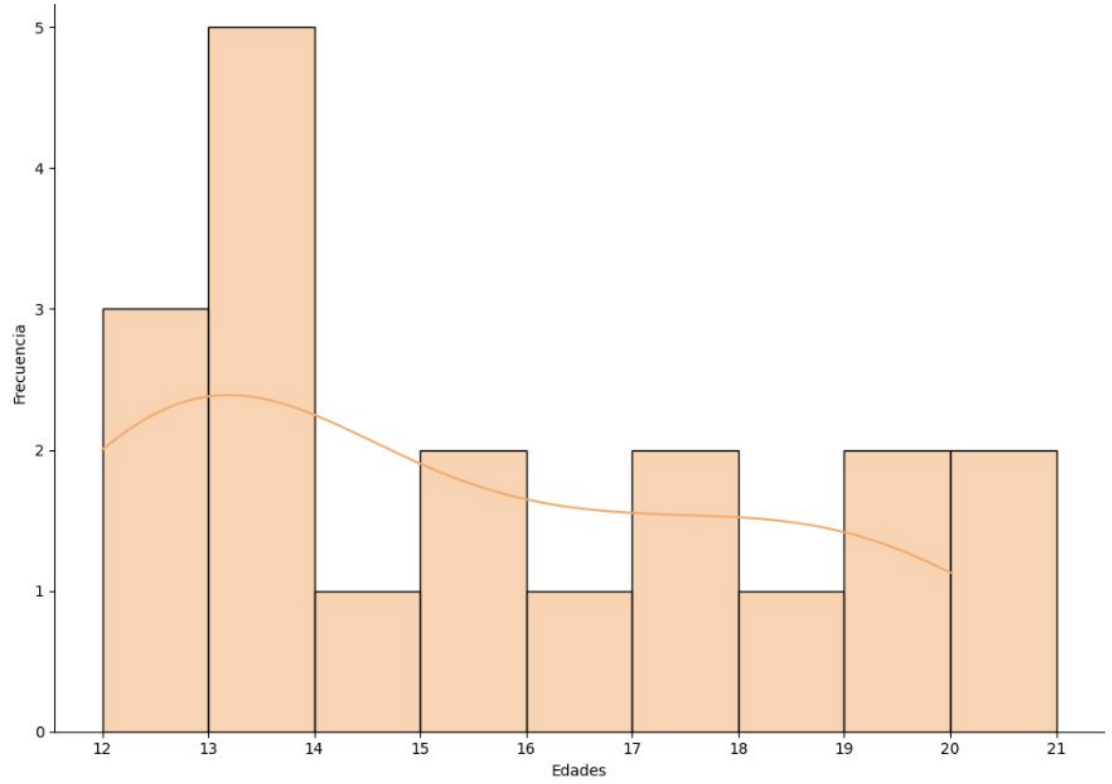
i) Tabla de frecuencias con intervalos

Edad (x)	Marca de Clase (X_i)	Frecuencia absoluta (f_i)	Frecuencia absoluta acumulada (F_i)	Frecuencia relativa (f_r)		Frecuencia relativa acumulada (F_r)	
[10 - 19)	14.5	5	5	0.1	10%	0.1	10%
[19 - 28)	23.5	11	16	0.22	22%	0.32	32%
[28 - 37)	32.5	8	24	0.16	16%	0.48	48%
[37 - 46)	41.5	5	29	0.1	10%	0.58	58%
[46 - 55)	50.5	8	37	0.16	16%	0.74	74%
[55 - 64)	59.5	6	43	0.12	12%	0.86	86%
[64 - 73]	68.5	7	50	0.14	14%	1	100%
	Total	50	Total	1	100%		

EDEM
Escuela de Empresarios

Descriptiva de Variables Cuantitativas (Continuas)

ii) Histograma



Medidas descriptivas univariantes

EDEM
Escuela de Empresarios

- Hay números clave para describir los datos (estadísticos).
- Dada una muestra, queremos obtener información de ella a través de un estadístico adecuado.

Tipos de descriptiva univariantes:

- ➔ **Tendencia central**
- ➔ **Posición**
- ➔ **Dispersión**
- ➔ **Forma de distribución**

Medidas descriptivas univariantes

EDEM
Escuela de Empresarios

Medidas de tendencia central:

- Nos ayudan a describir los valores centrales de nuestra muestra. En particular, es de las medidas más importantes.

Las medidas de tendencia central más importantes:

- ➔ **Media**
- ➔ **Mediana**
- ➔ **Moda**

Medidas de tendencia central

EDEM
Escuela de Empresarios

Media aritmética (media simple):

- **Media muestral:** Se denota como $\bar{x}$ y se calcula con la siguiente expresión:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

- Sabemos por propiedades matemáticas que la media muestral es el estimador **máximo-verosímil** de la media poblacional.

- **Media poblacional:** Se denota como μ y es el parámetro que queremos estimar.

Pregunta:

¿Podemos calcular la media poblacional?

Medidas de tendencia central

EDEM
Escuela de Empresarios

- Es importante comentar que la media muestral cambia si escogemos distintas muestras de una misma población.

¿Cómo podremos saber si un valor de $\bar{x}$ es compatible con un valor de μ?

→ **Contraste de hipótesis** (que veremos más adelante)

Media muestral

Pros	Contras
• **Muy útil** • **Usa todos los valores** • **Fácil de calcular / computar**	• **Outliers** • **Variables no simétricas**

Mediana

EDEM
Escuela de Empresarios

Dada una muestra ordenada, la mediana es el número que divide la muestra al 50%. Es decir, la mitad de los valores serán mayores que ella y el otro 50% menores.

Para calcularla depende de si el número de muestra es par o impar.

Ejemplo:

1) Mediana de la muestra {1,2,3,3,3} es 3

2) Mediana de la muestra {1,2,3,4} es 5/2=2.5

IMPORTANTE

Es necesario ordenar la muestra para calcularla ⚠️

Mediana

EDEM
Escuela de Empresarios

Pros	Contras
• No se ve alterada por distribuciones asimétricas o outliers	• Más difícil de calcular • Habitualmente puede ser menos representativa de la medio • Peor para tratar con ella matemáticamente

- *Abrimos debate:*

¿Es mejor la media o la mediana?

Moda

EDEM
Escuela de Empresarios

La moda es el valor que se encuentra con más frecuencia. Puede haber una, varias o ninguna.

Ejemplo:

- La moda de la muestra {1,2,3,4,4,5} es 4
- Las modas de la muestra {1,2,3,3,5,5} son 3 y 5
- La muestra {1,2,3,4} no tiene moda

Pros	Contras
• **No se ve afectada por outliers** • **Muy informativa para datos categóricos**	• **Puede no haber moda, por lo tanto, poco informativa en algunos casos** • **Puede no ser única** • **Puede no ser nada representativa**

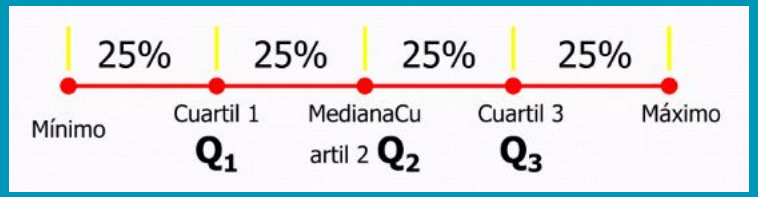
Medidas de posición

EDEM
Escuela de Empresarios

- **En este caso, las medidas siguientes nos darán información acerca de cómo está distribuida la muestra.**
- **Nos indica la posición de un valor respecto al resto de datos.**
- **Muy útil para describir conjuntos de datos grandes.**

Medidas de posición más importantes:

i) Percentiles
ii) Cuartiles
iii) Mínimo y Máximo



Medidas de posición

EDEM
Escuela de Empresarios

Percentiles:

- Los percentiles separan la muestra en 100 intervalos con la misma cantidad de valores.
- Los percentiles P25, P50 y P75 son los más interesantes.

Cuartiles:

- Los cuartiles Q1, Q2 y Q3 son los percentiles 25, 50 y 75.
- Q1 y Q2 se conocen como primer y tercer cuartil, mientras que el Q2 es la mediana.
- Los cuartiles separan nuestra muestra ordenada en 4 partes con las mismas observaciones.

Cálculo de cuartiles

EDEM
Escuela de Empresarios

Ejercicio:

- **Calcula los cuartiles de la siguiente muestra:**

13, 12, 11, 18, 16, 22, 16, 21, 17

Solución:

EDEM
Escuela de Empresarios

Example – Quartiles

Find the three quartiles of the following ordered sample:

$$11, 12, 13, 16, 16, 17, 18, 21, 22.$$

First of all, notice that the size of the sample is $n = 9$.

(Q_1) It is the value located in the position $0.25(9 + 1) = 2.5$. When this happens, the value we look for is the mean between the 2^{nd} and 3^{rd} value: $Q_1 = \frac{12+13}{2} = 12.5$.

(Q_2) It is the value located in the position $0.5(9 + 1) = 5$. Thus, $Q_2 = 16$.

(Q_3) It is the value located in the position $0.75(9 + 1) = 7.5$. So, Q_3 is equal to the mean between the 7^{th} and 8^{th} value: $\frac{18+21}{2} = 19.5$.

Medidas de posición

EDEM
Escuela de Empresarios

Resumen de las 5 medidas de posición: describe() en Python

- **Mínimo**
- **Primer Cuartil**
- **Mediana**
- **Tercer Cuartil**
- **Máximo**

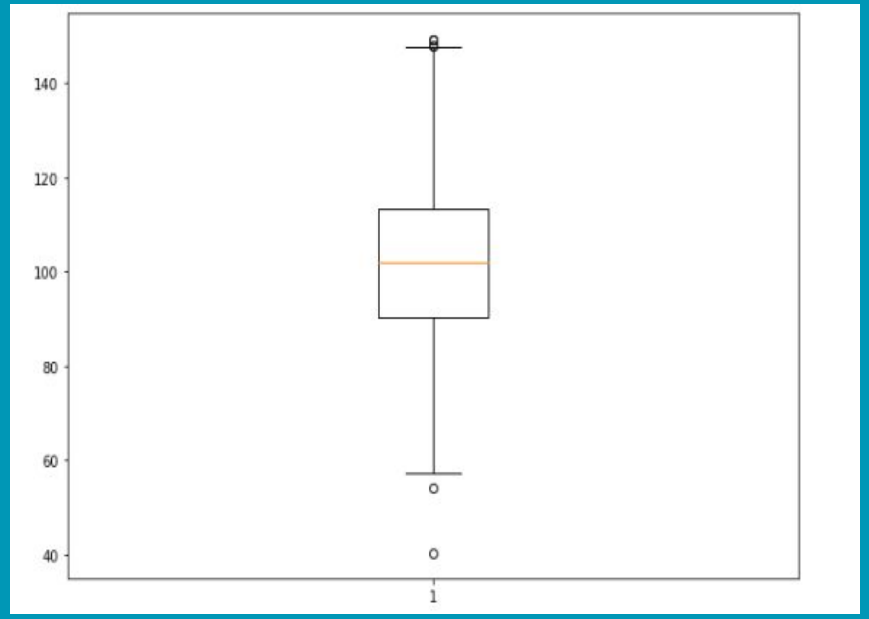
SE CUMPLE QUE:

$$(Q_0) < Q_1 < Q_2 < Q_3 < (Q_4)$$
$$\text{min} \quad 1^{st}\ \text{quart.} \quad \text{median} \quad 3^{rd}\ \text{quart.} \quad \text{max}$$

Medidas de posición

EDEM
Escuela de Empresarios

Diagramas de Box-Whisker



- **Es un gráfico que representa la forma de nuestra variable.**
- **Emplea las 5 medidas mencionadas anteriormente.**
- **La caja de dentro muestra el rango entre en Q1 y Q3, así como representada la mediana.**
- **Hay dos bigotes**
- **Los outliers se muestra con círculos (quedan fuera de los bigotes)**

Medidas de dispersión

EDEM
Escuela de Empresarios

Estas medidas nos dan información sobre la variabilidad de nuestros datos:

i) **Rango**

ii) **Rango Intercuartílico**

iii) **Varianza**

iv) **Desviación típica**

v) **Coeficiente de variación**

Medidas de dispersión

EDEM
Escuela de Empresarios

Rango y rango intercuartílico

El rango es la diferencia entre el máximo y el mínimo

El rango intercuartílico es la diferencia entre el Q3 y el Q1

Con el rango intercuartílico encontramos el 50% de los datos más centrados

El rango es sensible a los outliers mientras que el rango intercuartílico no lo es

Medidas de dispersión

EDEM
Escuela de Empresarios

Varianza

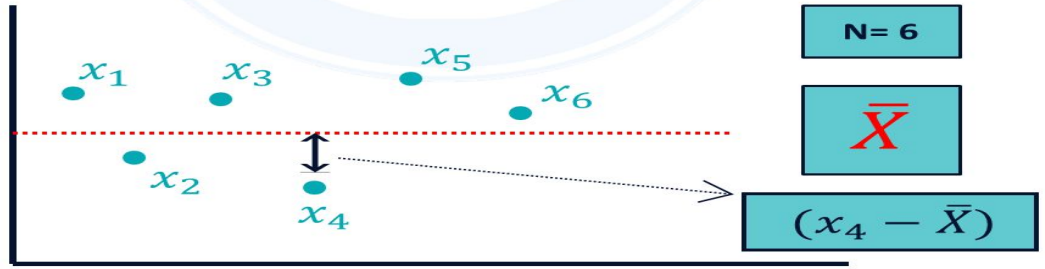
Es la media de los cuadrados de las desviaciones de la media con los valores.

VARIANZA

$$\sigma^2 = \frac{\sum_1^N (x_i - \bar{X})^2}{N}$$

- **X** → Variable
- $\mathbf{x_i}$ → Observación número i de la variable X.
- **N** → Número de observaciones.
- $\bar{\mathbf{x}}$ → Es la media de la variable X.

Es una medida de dispersión que representa la variabilidad de una serie de datos respecto a su media.



Medidas de dispersión

EDEM
Escuela de Empresarios

Desviación típica

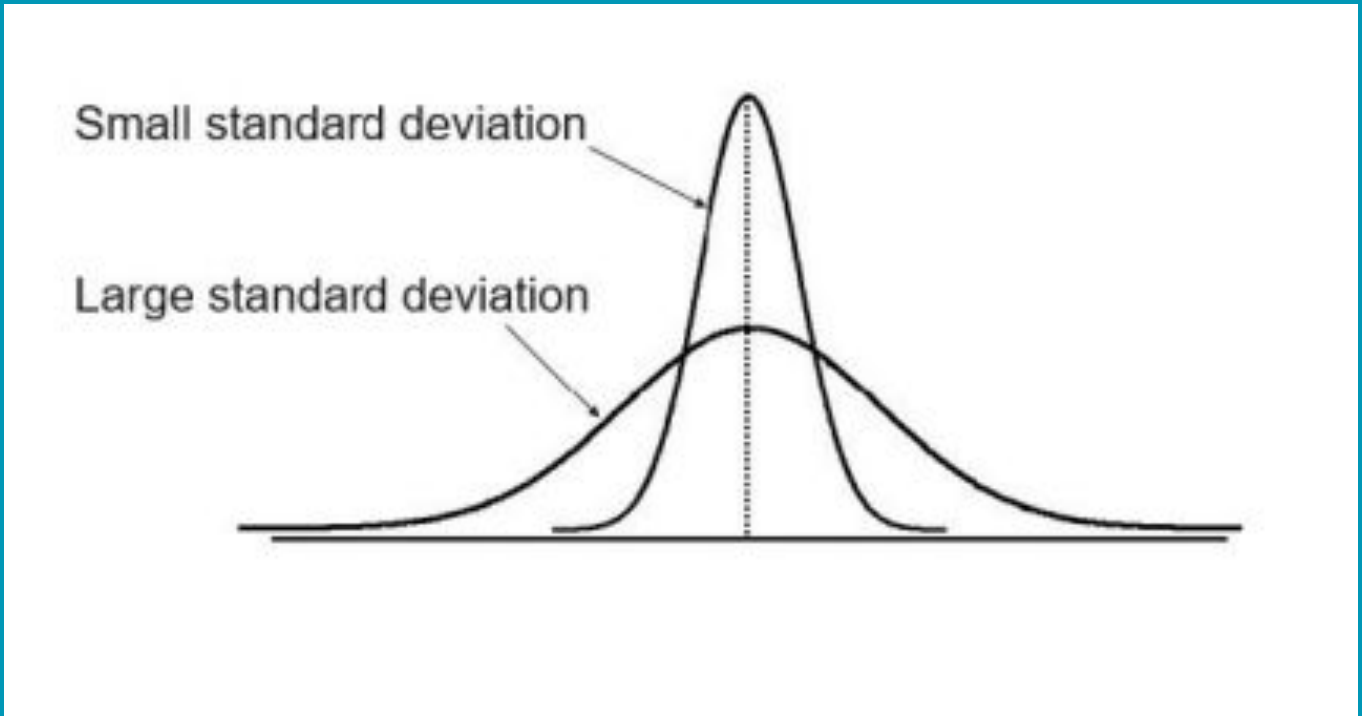
Es la raíz cuadrada positiva de la varianza. es la medida más utilizada por su fácil interpretación.

$$\sigma = \sqrt{\frac{\sum_i^N (X_i - \bar{X})^2}{N}}$$

- **También tendremos una poblacional y una muestral (s)**

Medidas de dispersión

EDEM
Escuela de Empresarios



Ejercicio:

- **Calcula la desviación típica de la siguiente muestra:**

10, 12, 14, 17, 17, 18, 18, 24

Ejercicio:

- **Calcula la desviación típica de la siguiente muestra:**

10, 12, 14, 17, 17, 18, 18, 24

$$s = \sqrt{\frac{(10-\bar{x})^2 + (12-\bar{x})^2 + (14-\bar{x})^2 + \ldots + (24-\bar{x})^2}{n-1}}$$

$$= \sqrt{\frac{(10-16)^2 + (12-16)^2 + (14-16)^2 + \ldots + (24-16)^2}{8-1}}$$

$$= \sqrt{\frac{6^2 + 4^2 + 2^2 + 1^2 + 1^2 + 2^2 + 2^2 + 8^2}{7}} = \sqrt{\frac{130}{7}} = 4.30\ldots$$

Medidas de dispersión

EDEM
Escuela de Empresarios

Coeficiente de variación

Esta medida representa la variabilidad de los datos respecto a la media.

Puede ser útil para comparar dos poblaciones: ¿Por qué?

$$C.V. = \frac{S}{\overline{X}} * 100$$

Medidas de forma de una distribución

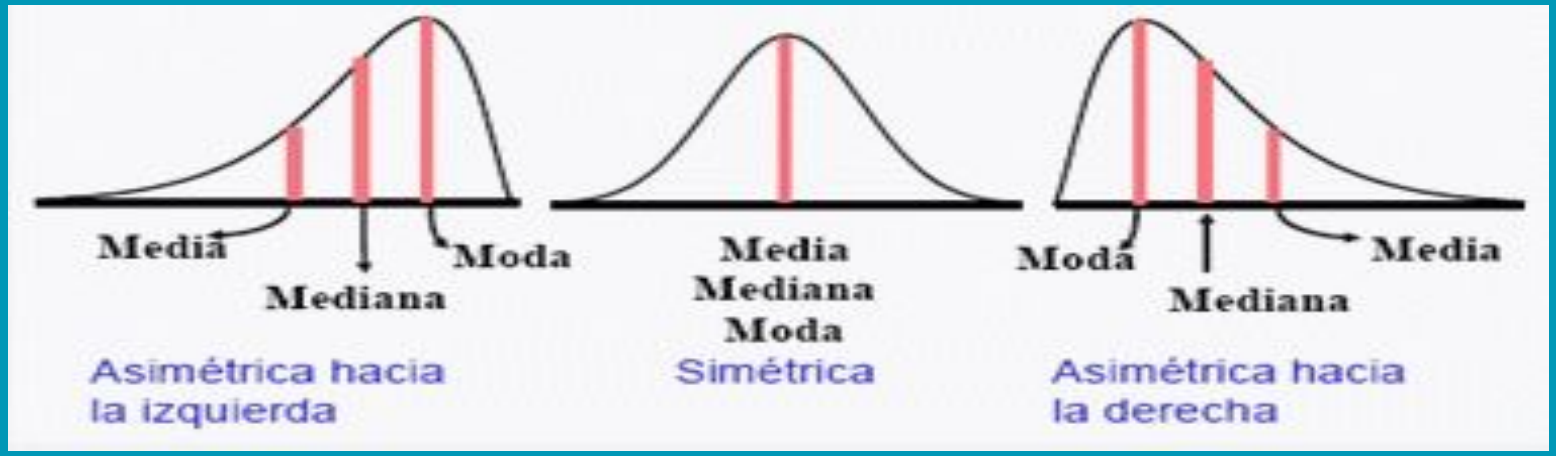
EDEM
Escuela de Empresarios

Gráficamente, podemos describir la forma de una distribución usando un histograma. El objetivo es visualizar si los datos están idénticamente distribuidos desde el centro de la muestra.

<u>Simetría</u>

Decimos que una que la forma de una distribución es simétrica cuando las observaciones están cercanamente equilibradas respecto al centro.

Asimetría (skewness) cuando los datos no son simétricos. Hay dos tipos de asimetrías.



Descriptiva bivariante

EDEM
Escuela de Empresarios

Hasta ahora sólo hemos utilizado descriptiva univariante. De ahora en adelante, consideramos dos variables X e Y del mismo tipo: cuantitativas continuas.

En este caso, vamos a estudiar dos medidas muy utilizadas:

- **Covarianza**
- **Coeficiente de correlación**

❖ **Ambas medidas se utilizarán para ver el grado de relación lineal que tienen ambas variables**

Descriptiva bivariante

EDEM
Escuela de Empresarios

Introducimos una forma clásica de describir una relación *LINEAL* entre dos variables X e Y:

Covarianza →

Mide la potencia (en términos lineales) de la relación entre dos variables. Distinguimos dos tipos:

$$Cov(X,Y) = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{n}$$

Descriptiva bivariante

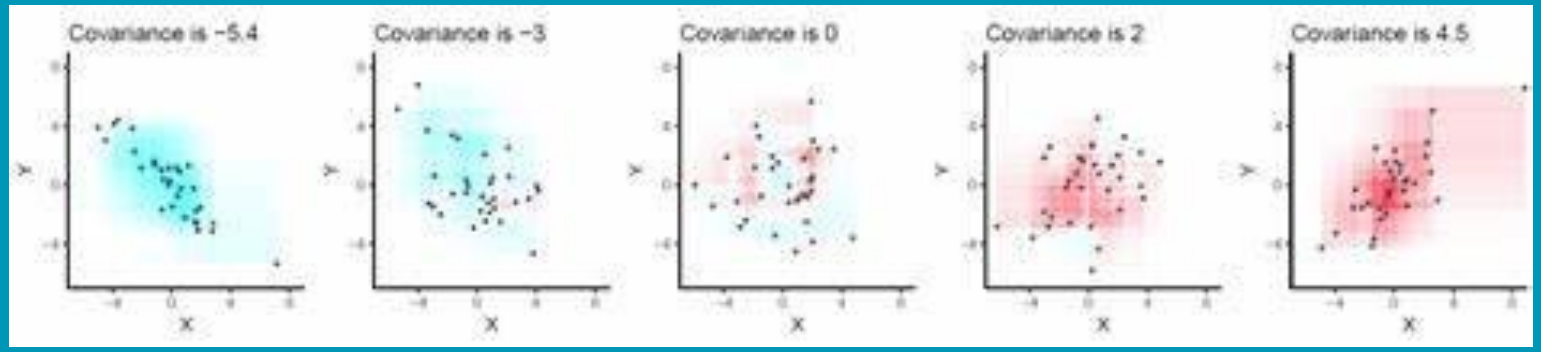
EDEM
Escuela de Empresarios

Interpretación de la covarianza:

1) **Si Cov(X,Y)>0, entonces X e Y tienden a moverse en la misma dirección**

2) **Si Cov(X,Y)<0, entonces X e Y tienden a moverse en la dirección opuesta**

3) **Si Cov(x,Y)=0, entonces es que son independientes**



Descriptiva bivariante

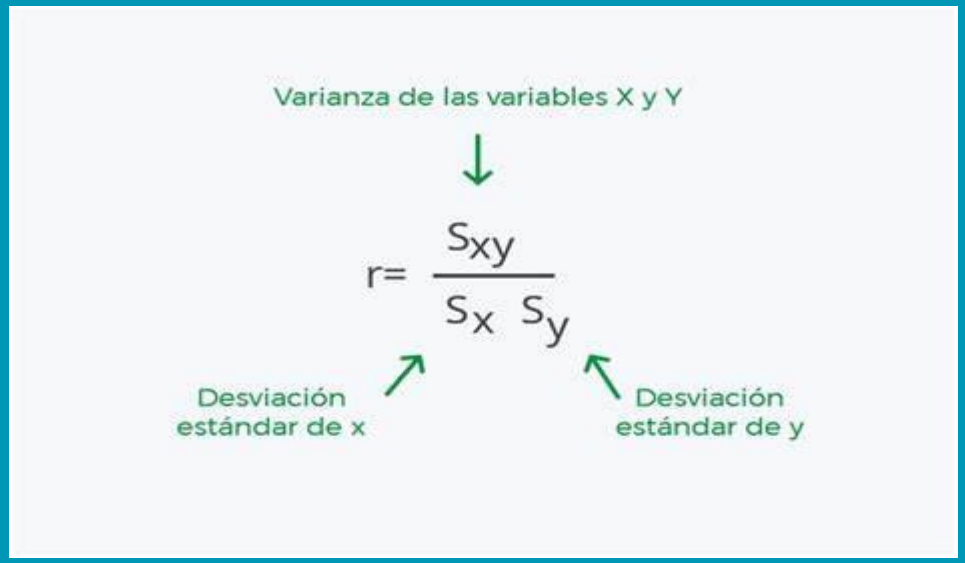
EDEM
Escuela de Empresarios

Coeficiente de Pearson:

- **Es una medida de dependencia lineal entre dos variables aleatorias cuantitativas.**
- **A diferencia de la covarianza, la correlación es independiente de la escala de medida de las variables.**

De manera menos formal, podemos definir el coeficiente de correlación de Pearson como *un índice que puede utilizarse para medir el grado de relación (lineal) de dos variables siempre y cuando ambas sean cuantitativas y continuas.*

- **Está comprendido entre -1 y 1**



Descriptiva bivariante

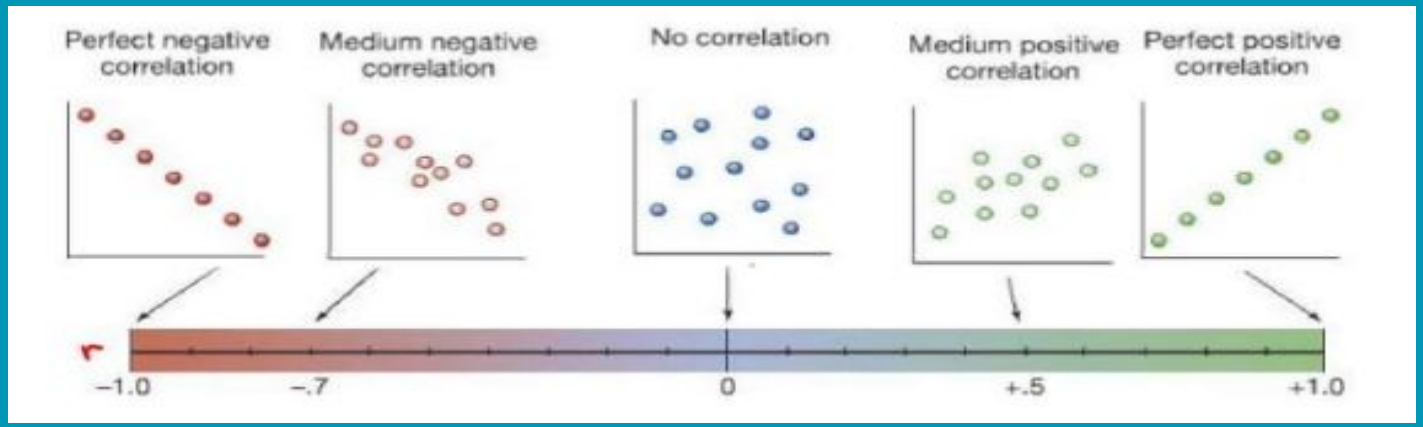
EDEM
Escuela de Empresarios

Propiedades:

1) **Cuanto más cerca de -1, más fuerte será la relación lineal negativa entre ambas variables.**

2) **Cuanto más cerca del 1, más fuerte será la relación lineal positiva entre las variables.**

3) **Cuanto más cerca de cero, menos probable que exista una relación lineal entre ambas variables.**



Práctica con Python

EDEM
Escuela de Empresarios

Parte práctica de la sesión:

1. **Descriptiva con Python**

Resumen

EDEM
Escuela de Empresarios

Durante esta unidad hemos visto:

- **Nociones básicas de estadística como su nomenclatura (población, muestra..).**
- **Definición de las variables aleatorias y sus diferentes tipos.**
- **Descriptiva univariante para cada tiempo de variable (gráficos y estadísticos)**
- **Descripción bivariante para ambos tipos de variable.**
- **Implementación de toda la descriptiva vista con Python.**
- **Realización de informes estadísticos con Google Colab.**

Siguientes objetivos:

- **Distribuciones de probabilidad.**
- **Contrastes de hipótesis.**
- **ANOVA.**

EDEM
Escuela de Empresarios

¡GRACIAS POR VUESTRA ATENCIÓN!

miguel.ruadelbarrio@ams-europe.com

linkedin.com/in/miguel-rua-del-barrio-5214661b5