MÓDULO 2: MÉTODOS ESTADÍSTICOS PARA IA

OBJETIVO GENERAL: INTRODUCIR LOS FUNDAMENTOS ESTADÍSTICOS NECESARIOS PARA ANALIZAR DATOS Y EVALUAR MODELOS DE INTELIGENCIA ARTIFICIAL, FAVORECIENDO LA TOMA DE DECISIONES BASADAS EN EVIDENCIA.

- CLASE 1: NOCIONES DE PROBABILIDAD. ESTADÍSTICA DESCRIPTIVA.
- CLASE 2: TEST DE HIPÓTESIS.
- CLASE 3: REGRESIÓN.
- CLASE 4: ESTADÍSTICA NO PARAMÉTRICA.

MÓDULO 2: MÉTODOS ESTADÍSTICOS PARA IA

OBJETIVO GENERAL: INTRODUCIR LOS FUNDAMENTOS ESTADÍSTICOS NECESARIOS PARA ANALIZAR DATOS Y EVALUAR MODELOS DE INTELIGENCIA ARTIFICIAL, FAVORECIENDO LA TOMA DE DECISIONES BASADAS EN EVIDENCIA.

- CLASE 1: NOCIONES DE PROBABILIDAD. ESTADÍSTICA DESCRIPTIVA.
- CLASE 2: TEST DE HIPÓTESIS.
- CLASE 3: REGRESIÓN.
- CLASE 4: ESTADÍSTICA NO PARAMÉTRICA.

Clase 1 - Parte 2: Métodos Estadísticos para IA

Estadística Descriptiva.

Objetivos de esta presentación

- Técnicas de Estadística Descriptiva.
- Tablas de Frecuencias.
- Medidas de posición y de dispersión.
- Principales gráficos de la Estadística Descriptiva.

Resumen

Estadística descriptiva

Construcción de Tablas de Frecuencias

Descripción numérica

Descripción gráfica

Estadística descriptiva

- La **estadística descriptiva** es un conjunto de técnicas numéricas y gráficas para obtener, organizar, presentar y describir un conjunto de datos con el propósito de facilitar el uso, sin extraer conclusiones (inferencias) sobre la población a la que pertenecen.

- Algunos ejemplos de técnicas descriptivas básicas que permiten realizar la descripción de datos son:

  - Construcción de tablas de frecuencias,
  - Descripción numérica,
  - Descripción gráfica.

Estadística descriptiva

- Para aplicar una técnica descriptiva, numérica o gráfica, será necesario analizar previamente el tipo de variable con la que se está trabajando.
- **Variable estadística**: es una característica de la muestra bajo estudio.
- **Datos**: son los valores que se obtienen para cada variable.
- **Observaciones:** es el conjunto de las mediciones obtenidas para una determinada muestra.

Tipo	Clases	Ejemplo
Cualitativa	Nominal	Sexo, raza, color de ojos,...
	Ordinal	Grado de contaminación, calificación,...
Cuantitativa	Discreta	Nº de hermanos, nº de materias, ...
	Continua	Peso, altura, ...

Estadística descriptiva

- **Tipos de variables estadísticas**: se distinguen dos tipos de variables.
- **Variables cualitativas o categóricas:** no se pueden expresar a través de una cantidad numérica. Representan cualidades o categorías.
  - **Variables cualitativas nominales**: las diferentes categorías no se pueden ordenar bajo ningún criterio.
  - **Variables cualitativas ordinales**: las diferentes categorías admiten un orden jerárquico.
- **Variables cuantitativas:** se pueden expresar a través de un número.
  - **Variables cuantitativas discretas**: solo pueden tomar un número finito de valores.
  - **Variables cuantitativas continuas**: pueden tomar cualquier valor.

Resumen

Estadística descriptiva

Construcción de Tablas de Frecuencias

Descripción numérica

Descripción gráfica

Distribución de frecuencias

- Una distribución de frecuencias es una tabla en la que se organizan los datos en categorías o grupos de valores que describen una característica de los individuos bajo estudio.

- Las **tablas de frecuencias** son una de las técnicas básicas para el resumen de información a partir de una muestra de datos.

- Su construcción es sencilla pero en conjuntos de datos de un tamaño moderado o grande su cálculo puede resultar laborioso, aunque se pueden obtener utilizando cualquier paquete estadístico (R, Python, etc.).

Clases Estatura	Frecuencia Absoluta	Frecuencia Relativa	Frecuencia Absoluta Acumulada	Frecuencia Relativa Acumulada
1.40 - 1.45	4	0.08	4	0.08
1.45 - 1.50	5	0.1	9	0.18
1.50 - 1.55	6	0.12	15	0.3
1.55 - 1.60	10	0.2	25	0.5
1.60 - 1.65	5	0.1	30	0.6
1.65 - 1.70	11	0.22	41	0.82
1.70 - 1.75	2	0.04	43	0.86
1.75 - 1.80	7	0.14	50	1

Distribución de frecuencias

► Construcción de una **tabla de frecuencias**:

- *Intervalos de clase*
- *Frecuencia absoluta*
- *Frecuencia absoluta acumulada*
- *Frecuencia relativa*
- *Frecuencia relativa acumulada*

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

Distribución de frecuencias

- ***Intervalos de clase:*** para variables cuantitativas continuas, se agrupan los distintos valores obtenidos en la muestra en intervalos que pueden ser de igual longitud.

- Inicialmente se debe conocer cuál es el rango de variación de los datos (diferencia entre el máximo y el mínimo), y construir los intervalos de manera que cubran todo el rango.

- En general, para seleccionar el número de intervalos, se puede usar la regla de considerar el entero más próximo a la raíz cuadrada de n, donde n es el tamaño de la muestra observada.

- El número de intervalos suele estar entre 5 y 20.

- Los extremos de los intervalos de clase son los ***límites de clase inferior y superior.*** El punto medio de cada intervalo de clase es la ***marca de clase***. La longitud de cada intervalo de clase es el ***ancho de clase***.

Distribución de frecuencias

- ***Frecuencia absoluta (f_i):*** es el número de observaciones que caen dentro de cada intervalo.

- ***Frecuencia absoluta acumulada (F_i):*** es la suma de las frecuencias absolutas de una clase o grupo de la muestra con la anterior.

- ***Frecuencia relativa (h_i):*** es el número de veces que se repite un número en un conjunto de datos respecto al total. Es decir, es igual a la frecuencia absoluta dividido por el número total de datos (f_i / N).

- ***Frecuencia relativa acumulada (H_i):*** es el cociente entre la frecuencia absoluta acumulada de un determinado valor y el número total de datos (F_i / N).

Distribución de frecuencias

- Construcción de una **tabla de frecuencias**:

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%		

Resumen

Estadística descriptiva

Construcción de Tablas de Frecuencias

Descripción numérica

Descripción gráfica

Medidas descriptivas

- Del mismo modo que las gráficas pueden mejorar la presentación de los datos, las descripciones numéricas también tienen gran valor. Presentaremos aquí varias medidas numéricas importantes para describir las características de los datos.

- Sea X la variable estadística de interés. Sea x_i la observación en el individuo i, veremos a continuación las principales medidas características para describir la información contenida en una muestra x_1, x_2, …, x_n de tamaño n.

- Se distinguen dos tipos de medidas: **de posición** y **de dispersión**.

- Dichas medidas se utilizan para resumir la información atendiendo a tres aspectos principales: alrededor de qué valores se encuentran los datos, cuánto se dispersan y si se distribuyen de manera similar a una *campana de Gauss*, que será el modelo que se tome como referencia.

Medidas de posición

- Las **medidas de posición o localización** proporcionan valores alrededor de los cuales se distribuyen los datos observados en la muestra.

- Se distinguen **medidas de localización de tendencia central** y **de tendencia no central**.

- **Medidas de posición de tendencia central:**

  - **media aritmética o media muestral**
  - **mediana**
  - **moda**

- **Medidas de posición de tendencia no central:**

  - **cuartiles**
  - **percentiles**

Medidas de posición

- **Medidas de posición de tendencia central:**

  - **Media aritmética o media muestral** $(\bar{x})$

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

- ✓ La **media aritmética** es el valor promedio de la muestra.
- ✓ El valor de la media aritmética no tiene por qué pertenecer al conjunto de posibles valores de la variable.
- ✓ Uno de los problemas que presenta la media es que no es una medida robusta, es decir, su valor se ve influenciado por datos anormalmente altos o bajos (valores atípicos).

Medidas de posición

- **Medidas de posición de tendencia central:**

  - **Mediana** $(\tilde{x})$

- Si los datos de la muestra están ordenados de menor a mayor, la **mediana** es el valor hasta el cual se encuentran el 50 % de los casos.
- Por tanto, la mediana dejará la mitad de las observaciones por debajo de su valor y la otra mitad por encima.
- A diferencia de la media, la mediana es una medida robusta, ya que su valor se ve poco afectado por la presencia de datos atípicos.
- Si de una muestra se obtienen la media y la mediana y sus valores difieren sustancialmente, esto será indicativo de la presencia de datos atípicos.

  - **Moda**

- La **moda** es la observación que más se repite en la muestra.
- Puede existir más de una moda.

Medidas de posición

► **Medidas de posición de tendencia no central:**

❑ **Cuartiles**

✓ Si los datos de la muestra están ordenados de menor a mayor, los **cuartiles** Q1, Q2 y Q3 dividen la muestra en cuatro partes iguales.

✓ El primer cuartil o cuartil inferior (Q1) es un valor que tiene aproximadamente la cuarta parte (25%) de las observaciones por debajo de él, y el 75% restante, por encima de él.

✓ El segundo cuartil (Q2), tiene aproximadamente la mitad (50%) de las observaciones por debajo de él. El segundo cuartil coincide con la mediana.

✓ El tercer cuartil o cuartil superior (Q3), tiene aproximadamente las tres cuartas partes (75%) de las observaciones por debajo de él.

❑ **Percentiles:**

✓ Si los datos de la muestra están ordenados de menor a mayor, y se los divide en cien partes iguales, los puntos de división reciben el nombre de **percentiles**. (Puede verse que $p_{0.25} = q_1$, $p_{0.5} = q_2$, $p_{0.75} = q_3$).

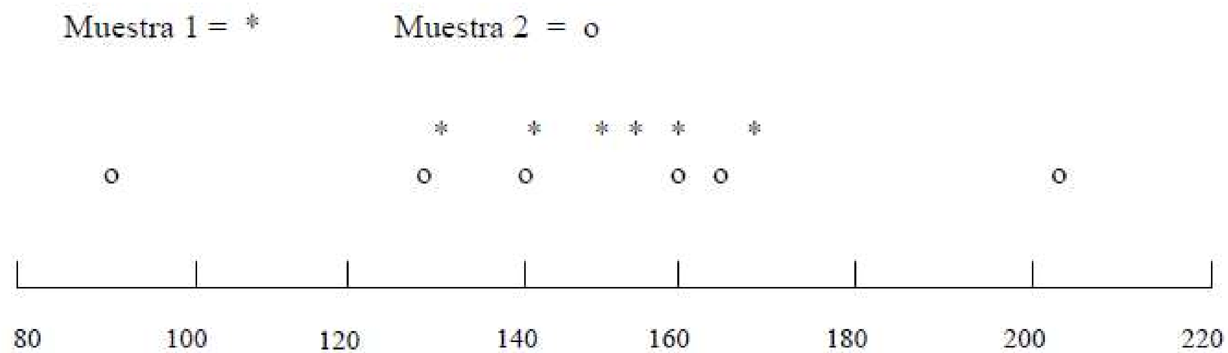
Medidas de dispersión

► Las medidas de posición no necesariamente proporcionan información suficiente para describir datos de manera adecuada.

Muestra 1: 130, 150, 145, 158, 165, 140

Muestra 2: 90, 128, 205, 140, 165, 160

Media 1 = 148 y Mediana 1 = 147.5
Media 2 = 148 y Mediana 2 = 150



► Las **medidas de dispersión o variabilidad** proporcionan una descripción más precisa de los datos. Indican cual es la dispersión que presentan. Entre ellas se encuentran:

- **rango de la muestra**
- **rango intercuartílico**
- **varianza muestral**
- **desviación estándar muestral**
- **coeficiente de variación**

Medidas de dispersión absolutas

Medida de dispersión relativa

Medidas de dispersión

► **Medidas de dispersión o variabilidad absoluta**:

✓ Dependen de las unidades en las que se miden las observaciones.

- **rango muestral** $R = \text{máx}\{x_i\} - \text{mín}\{x_i\}$

✓ Diferencia entre las observaciones más grande y más pequeña. Puede verse afectado por la presencia de datos atípicos. A mayor rango, mayor variabilidad en los datos.

- **rango intercuartílico** $RIC = Q_3 - Q_1.$

✓ Diferencia entre el cuartil superior e inferior. El rango intercuartílico es menos sensible a los valores extremos de la muestra que el rango muestral.

✓ Se lo suele conocer también como IQR (*Interquartile range*).

Medidas de dispersión

► **Medidas de dispersión o variabilidad absoluta**:

- **varianza muestral** $$s^2 = \frac{(x_1-\bar{x})^2+\ldots+(x_n-\bar{x})^2}{n} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2.$$

✓ Las unidades de medida son las de la variable al cuadrado.

- **desviación estándar muestral** $$s = +\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}.$$

✓ Está expresada en las mismas unidades de medida de las observaciones.

► **Medidas de dispersión o variabilidad relativa**:

✓ No dependen de las unidades de los datos. Son útiles para comparar dos o más muestras que difieren de manera considerable en la magnitud de las observaciones.

- **coeficiente de variación** $$CV = \frac{s}{\bar{x}}.$$

Ejemplos

- El ejemplo_3.ipynb y el ejemplo_4.ipynb muestran ejemplos sobre el cálculo de algunas medidas de posición y medidas de dispersión. En el primer ejemplo, a partir de generar n números con distribución chi-cuadrado con k grados de libertad. Y en el segundo ejemplo, para diferentes datasets.
- En ambos casos, los principales valores pueden obtenerse con la función **describe**. La función **describe()** pertenece a la librería **pandas** y sirve para obtener un **resumen estadístico rápido** de un DataFrame o Serie.

```
datos_mios.describe()
```

	personas.pre	personas.post
count	5.000000	5.000000
mean	11.600000	6.200000
std	7.300685	8.408329
min	5.000000	1.000000
25%	8.000000	2.000000
50%	10.000000	2.000000
75%	11.000000	5.000000
max	24.000000	21.000000

Resumen

Estadística descriptiva

Construcción de Tablas de Frecuencias

Descripción numérica

Descripción gráfica

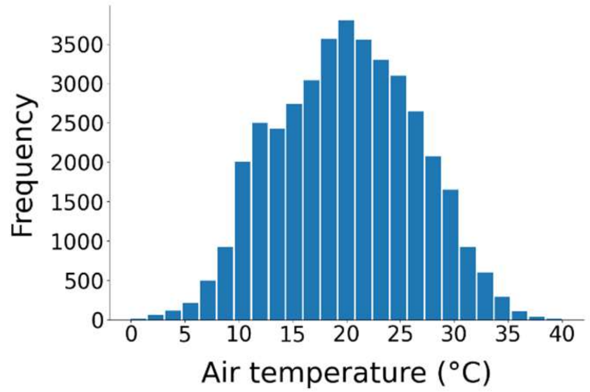
Descripción gráfica

- Otra manera de describir la información presente en datos es la gráfica. Existen muchos tipos de gráficos en la Estadística Descriptiva. Comentaremos sólo algunos de ellos:

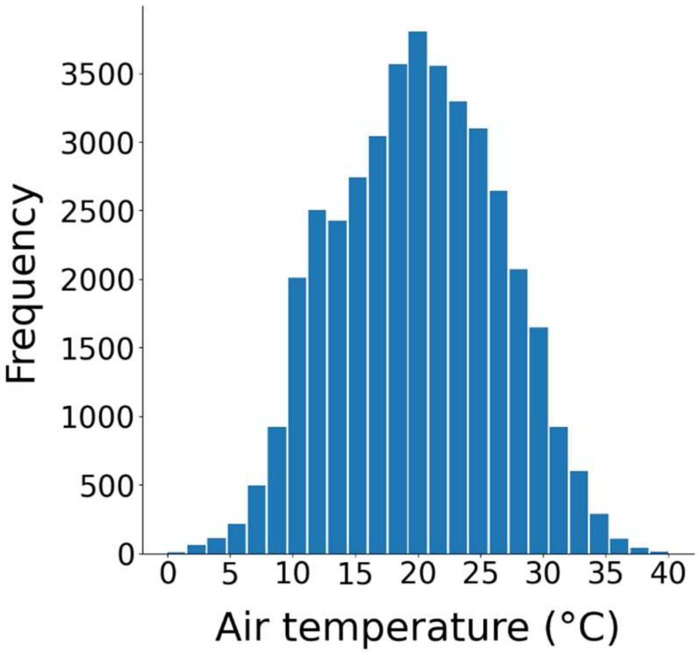
    - Histogramas
    - Distribución acumulada empírica
    - QQplots
    - Diagrama de cajas

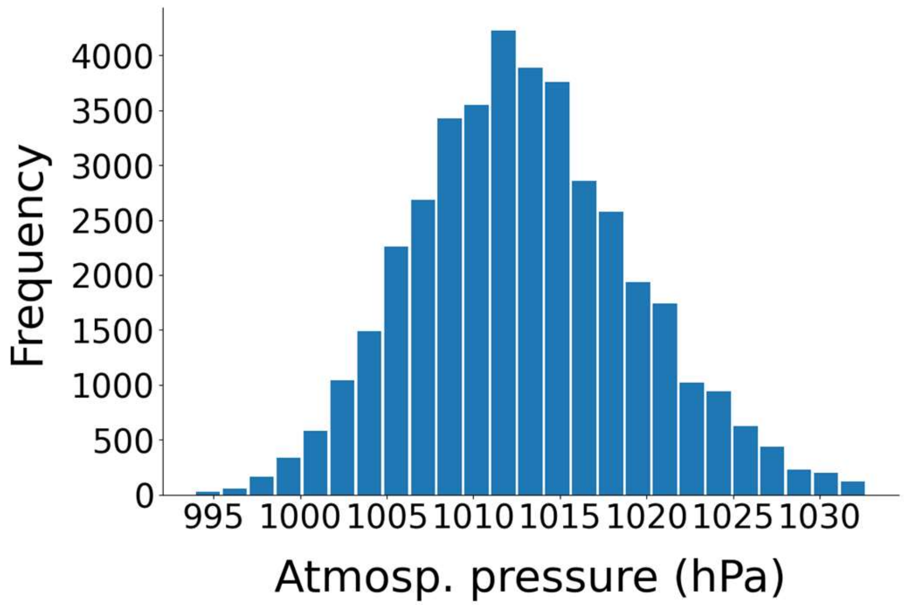
Histograma

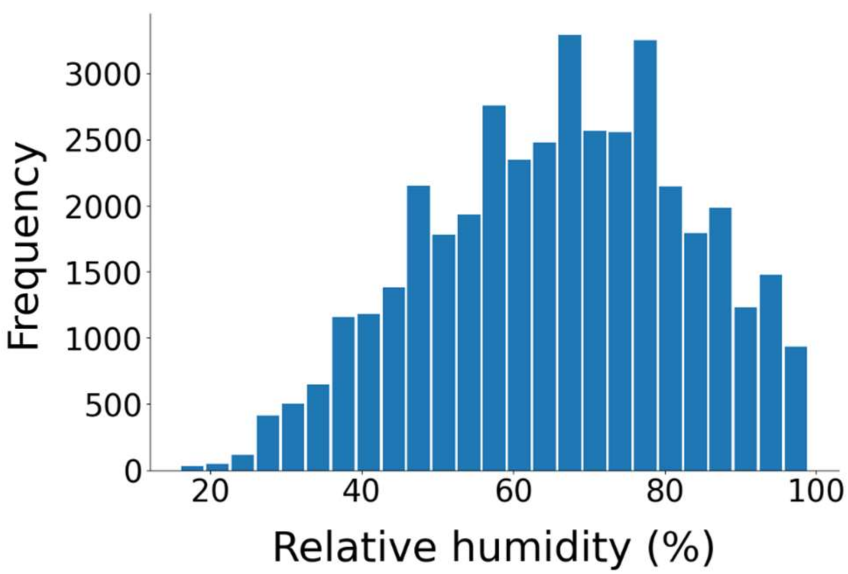
- Un **histograma** es una representación de la distribución de los datos.
- Permite visualizar dónde están situados los datos y tener una idea de su distribución.
- Para su **construcción** se debe partir el recorrido de la variable (valor mínimo y máximo) en *n* intervalos (**bins**) iguales (no necesariamente). Luego, se debe contar cuántos datos caen en cada intervalo.
- Posteriormente se dibujan rectángulos. La base de cada rectángulo es el intervalo de clase y la altura es el número de frecuencia (absoluta o relativa).

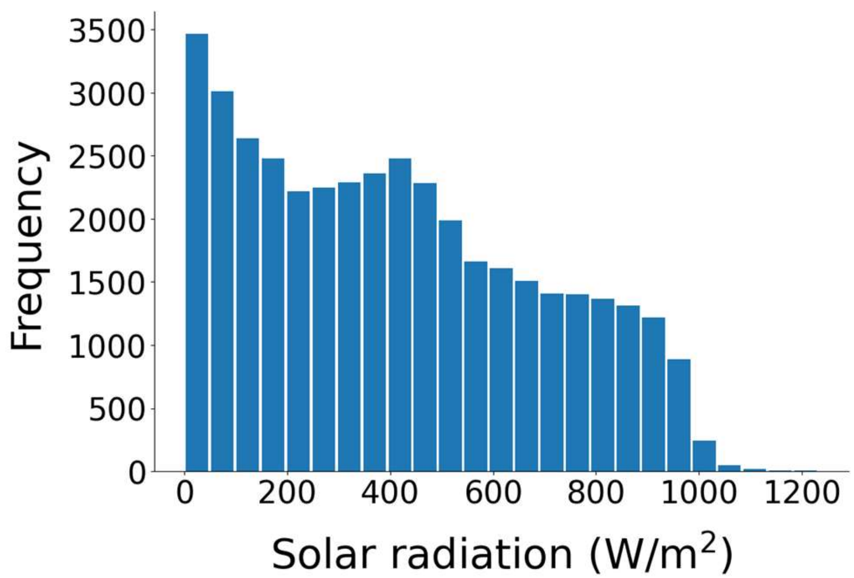


Histograma: Ejemplos









Histograma: Construcción en Python

- En el ejemplo_5.ipynb utilizamos datos de 138 personas con desorden alimenticio (patient) y 98 personas que se tomaron como control (control). Se contabilizó el ejercicio físico realizado (en horas por semana)[1].

- Seleccionamos los pacientes cuando tenían menos de 9 años y luego la variable *exercise*:

```
# 🧪 Filtrar pacientes menores de 9 años
datos_pre = datos[(datos['group'] == 'patient') & (datos['age'] < 9)]
datos_patient_menores = datos_pre['exercise']
print("Cantidad de pacientes menores de 9 años:", len(datos_patient_menores))
```

```
Cantidad de pacientes menores de 9 años: 138
```

Código/ejemplo_5.ipynb

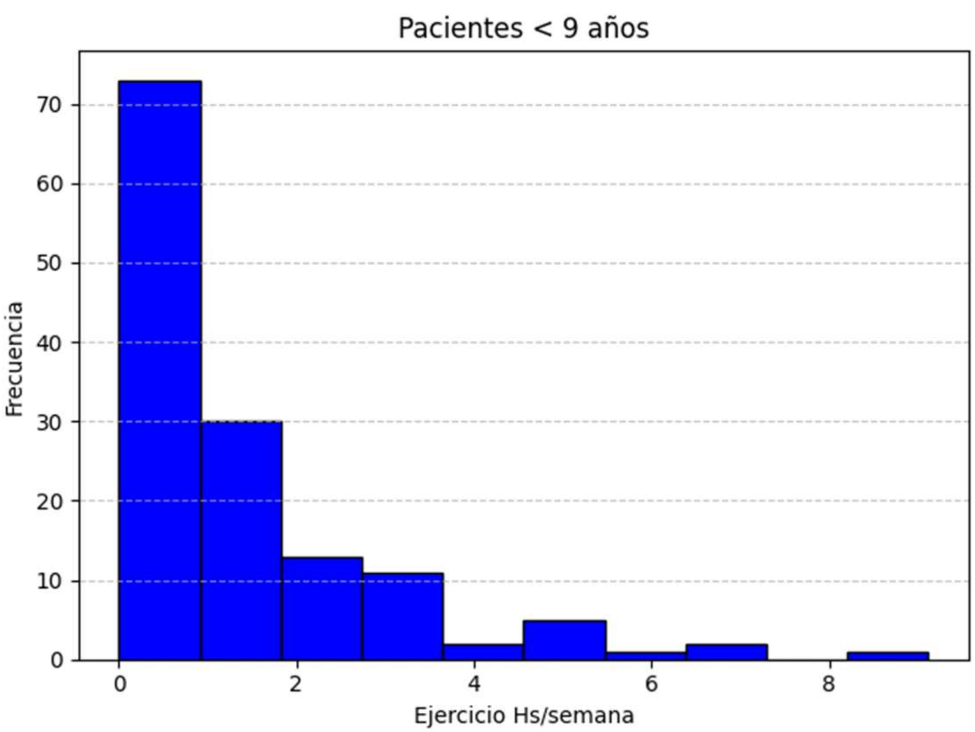
[1] Datos Blackmore.csv

Histograma: Construcción en Python

- Construimos el histograma con la función **hist** de la librería **matplotlib**.

```python
# 📊 Construcción del histograma
plt.hist(datos_patient_menores, bins=10, color='blue', edgecolor='black')
plt.title("Pacientes < 9 años")
plt.xlabel("Ejercicio Hs/semana")
plt.ylabel("Frecuencia")
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()
```

Código/ejemplo_5.ipynb



- Es un histograma sesgado positivamente (o a derecha).
- Esto también se observa en las medidas de localización (resultados de la función **describe()**), donde $\bar{x} = 1{,}334$ y $\tilde{x} = 0{,}875$, es decir para una muestra sesgada a derecha $\tilde{x} < \bar{x}$.
- También se ve que la media no es robusta a valores extremos, es decir, si existe un valor muy grande, la media crecerá mucho aunque no así la mediana.

Histograma normalizado y distribución de probabilidad

- Muchas veces suelen utilizarse **Histogramas normalizados** en conjunto con la **función de distribución de probabilidad** que siguen los datos. Esto es útil para comparar con distribuciones teóricas (ej. curva normal) o con una función de densidad dada.

- En el ejemplo_6.ipynb, generamos *n* números pseudo–aleatorios con distribuciones conocidas (normal y Chi-cuadrado) y evaluamos sus histogramas[2].

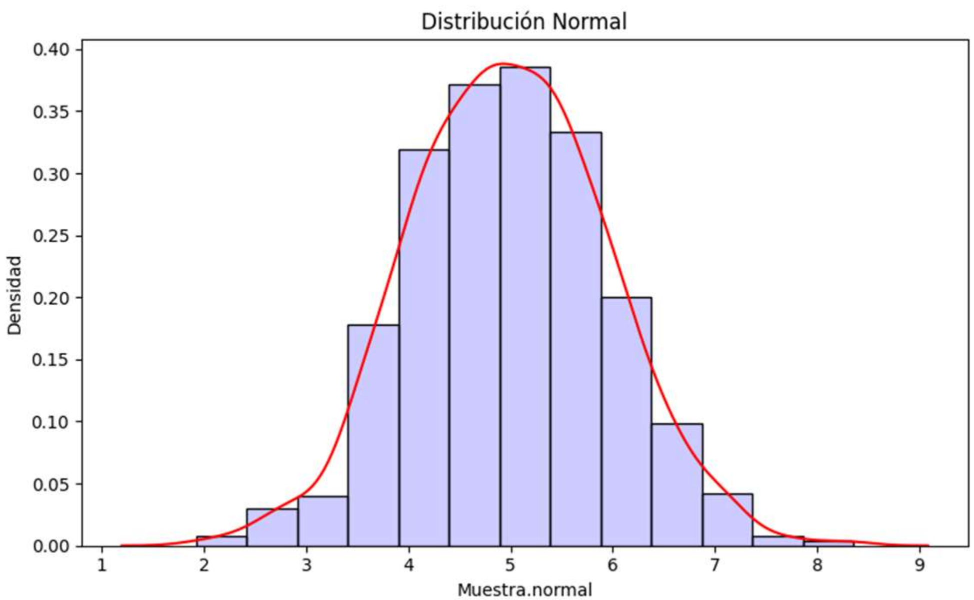
```python
# 📊 Histograma + densidad para muestra.normal
plt.figure(figsize=(8, 5))
sns.histplot(datosmios['muestra.normal'],
             binwidth=0.5,
             kde=False,
             stat='density',
             color='blue',
             edgecolor='black',
             alpha=0.2)
sns.kdeplot(datosmios['muestra.normal'], color='red')
plt.xlabel("Muestra.normal")
plt.ylabel("Densidad")
plt.title("Distribución Normal")
plt.tight_layout()
plt.show()
```

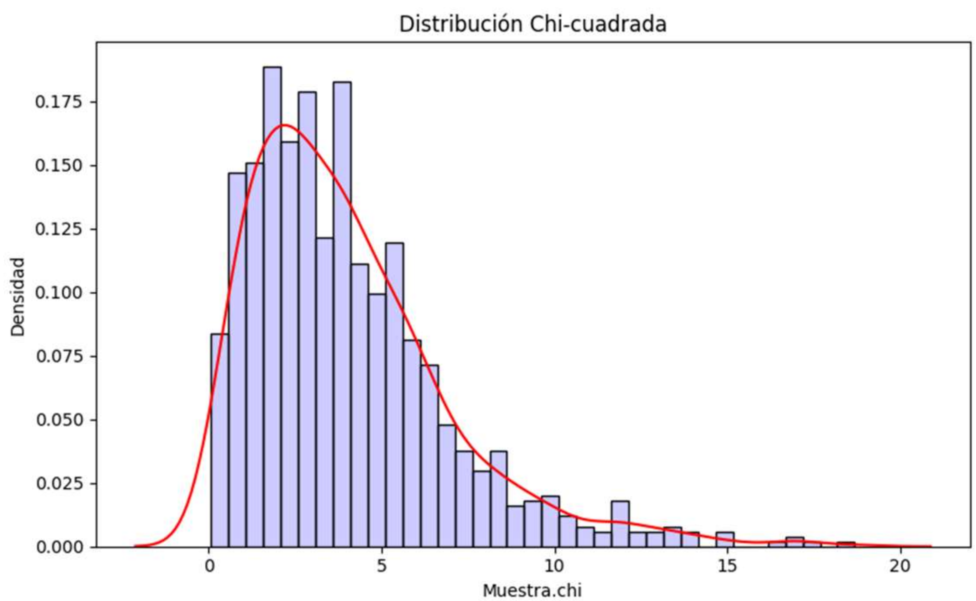
Código/ejemplo_6.ipynb

[2] Construimos los histograma con la función **histplot** de la librería **seaborn**.

Histograma: Construcción en Python

- Con `stat='density'` → el histograma se **normaliza** de modo que el área total bajo las barras sea igual a 1, convirtiéndolo en una **estimación de densidad de probabilidad**.
- La función **kdeplot** de la librería **seaborn** estima la **función de densidad de probabilidad** de manera suave y grafica una curva continua, que se denomina polígono de frecuencias suavizado (líneas que unen el punto medio de cada bin).



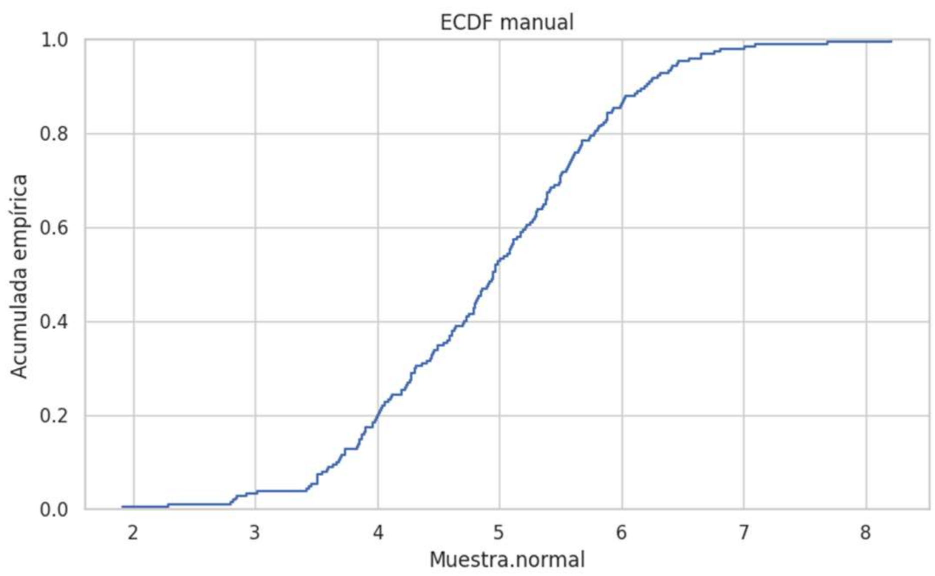


En el ejemplo_6.ipynb, compare los valores muestrales con los poblacionales (valores teóricos de la distribución de probabilidades) para ambos casos.

Distribución acumulada empírica

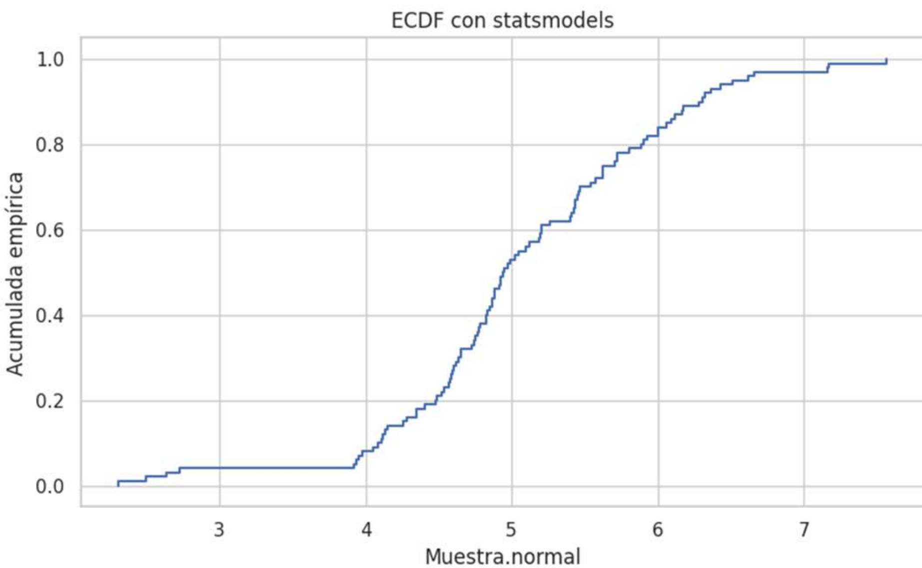
- Así como el histograma intenta describir la fdp, el gráfico de la **distribución acumulada empírica** es útil cuando se quiere estimar la **Fda** de los datos.
- Se grafica una función de *x*, definida como la fracción de datos menores o iguales que *x*.
- En el ejemplo_7.ipynb, se grafica la función de distribución acumulada empírica (ECDF) de tres maneras diferentes:

1) Construida de manera manual:

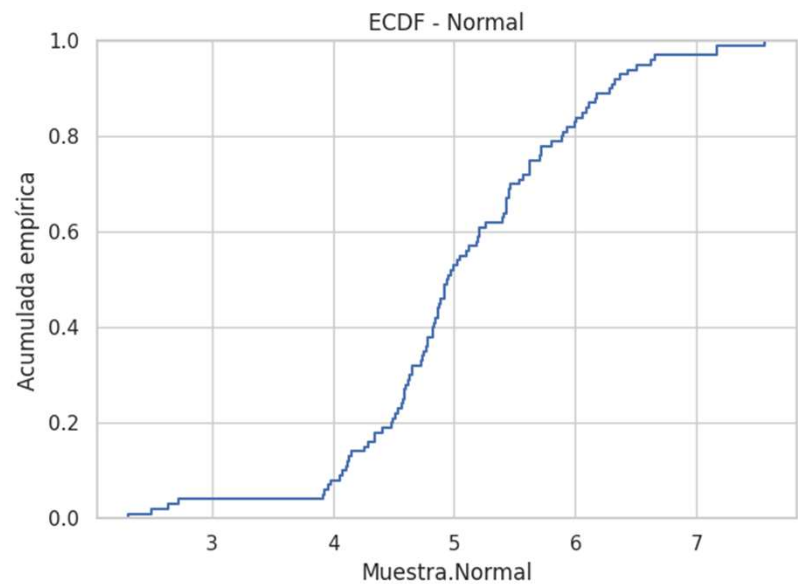


Distribución acumulada empírica

2) Construida con la función **ECDF** de la librería **statsmodels**.

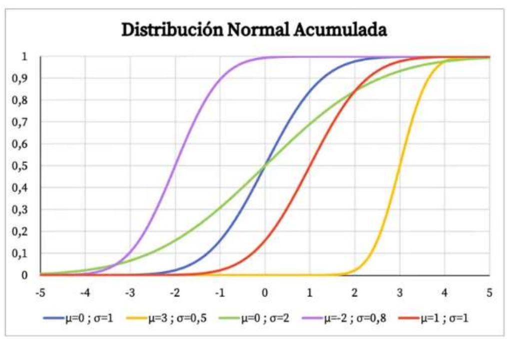


3) Construida con la función **ecdfplot** de la librería **seaborn**.



► Es muy utilizada para visualizar si la muestra tiene distribución Normal.

Recordar!!!



Distribución acumulada empírica

- Por ejemplo: en la figura se muestra la distribución acumulada empírica de una muestra de una distribución χ^2_{5-1} (con 4 grados de libertad).

- Se puede ver que se pierde la forma sigmodea (forma de S) de la distribución Normal.

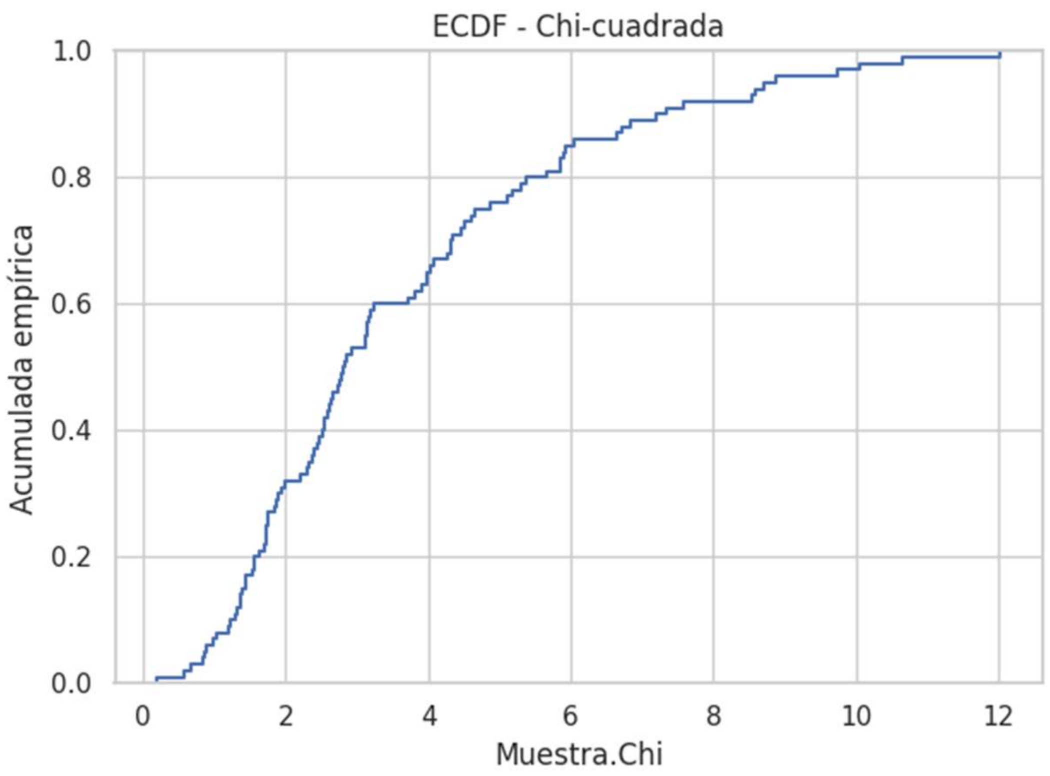
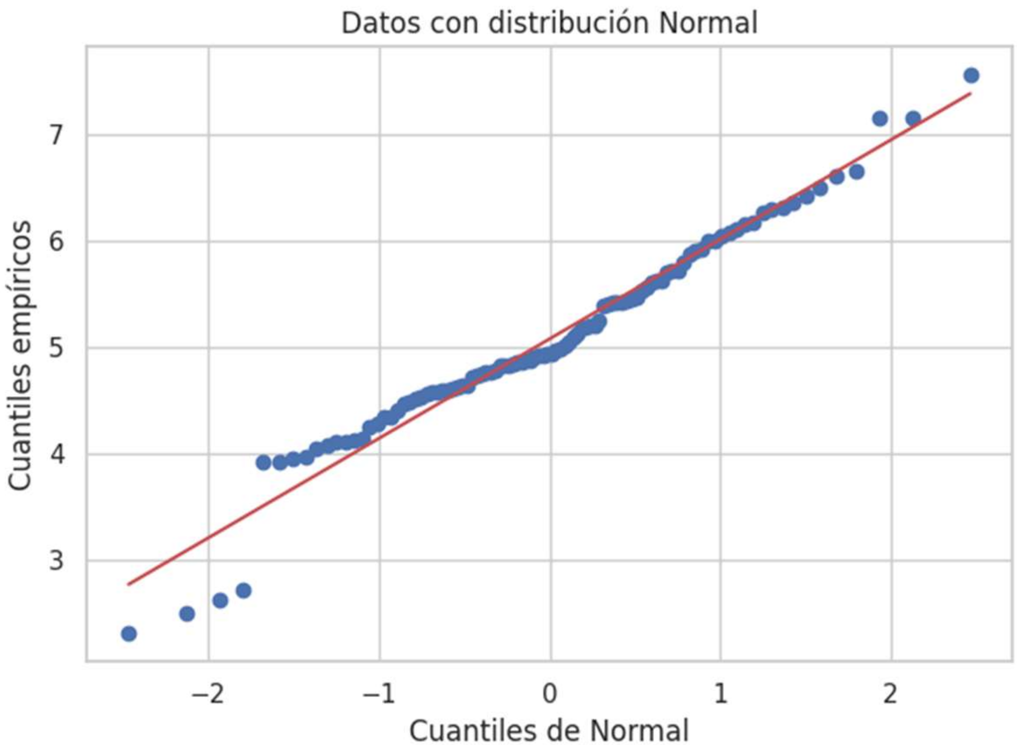


Gráfico Q-Q

► Otros gráficos interesantes a la hora de probar normalidad de una muestra (verificar si tiene distribución Normal[3]) son los **gráficos Q–Q** (QQplots).

► La idea es graficar en el eje de las abscisas los cuantiles que se esperan si la distribución fuera Normal y en el eje de las ordenadas los cuantiles muestrales. De esta manera, si la distribución fuera Normal, los datos se concentrarían en una recta. Se suele agregar también la recta a la gráfica.



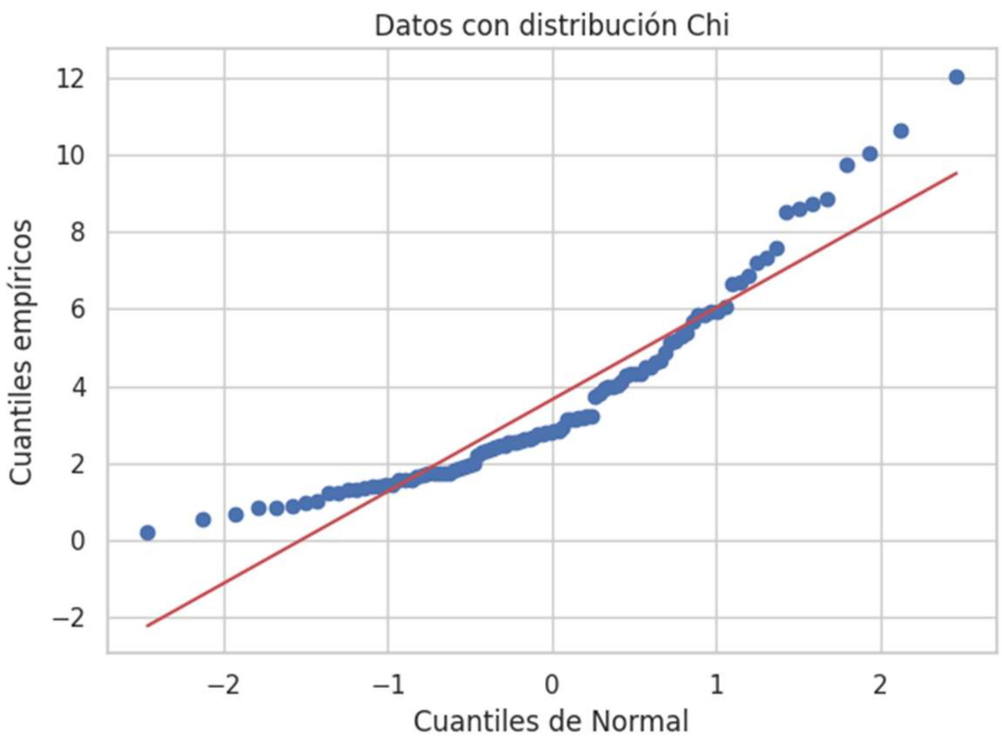
► En el ejemplo_7.ipynb se comparan los dos gráficos utilizando las muestras con distribuciones $N(\mu = 5, \sigma^2 = 1)$ y χ^2_{5-1}.

[3] Esto es importante porque algunos Métodos Estadísticos dependen fuertemente de la suposición de normalidad.

Gráfico Q-Q

- Otros gráficos interesantes a la hora de probar normalidad de una muestra (verificar si tiene distribución Normal[3]) son los **gráficos Q–Q** (Q-Q plots).

- La idea es graficar en el eje de las abscisas los cuantiles que se esperan si la distribución fuera Normal y en el eje de las ordenadas los cuantiles muestrales. De esta manera, si la distribución fuera Normal, los datos se concentrarían en una recta. Se suele agregar también la recta a la gráfica.



- En el ejemplo_7.ipynb se comparan los dos gráficos utilizando las muestras con distribuciones $N(\mu = 5, \sigma^2 = 1)$ y χ^2_{5-1}.

[3] Esto es importante porque algunos Métodos Estadísticos dependen fuertemente de la suposición de normalidad.

Gráfico Q-Q

- También podemos comparar diferentes datos, por ejemplo, utilizaremos un conjunto de datos que R tiene cargados por defecto (mtcars).

- Gracias a la librería **statsmodels**, desde Python podemos acceder a muchos **datasets clásicos de R** sin necesidad de tener R instalado.

```
import statsmodels.api as sm

data = sm.datasets.get_rdataset("nombre_dataset", "paquete").data
```

.data → lo convierte en un **DataFrame de pandas**.

- Podemos comparar por ejemplo, lo que consumen los autos en millas por galón (mpg) de acuerdo a su cilindrada (cyl).

Gráfico Q-Q

► Las diferentes inclinaciones significa que poseen diferentes varianzas, a menor pendiente menor varianza.

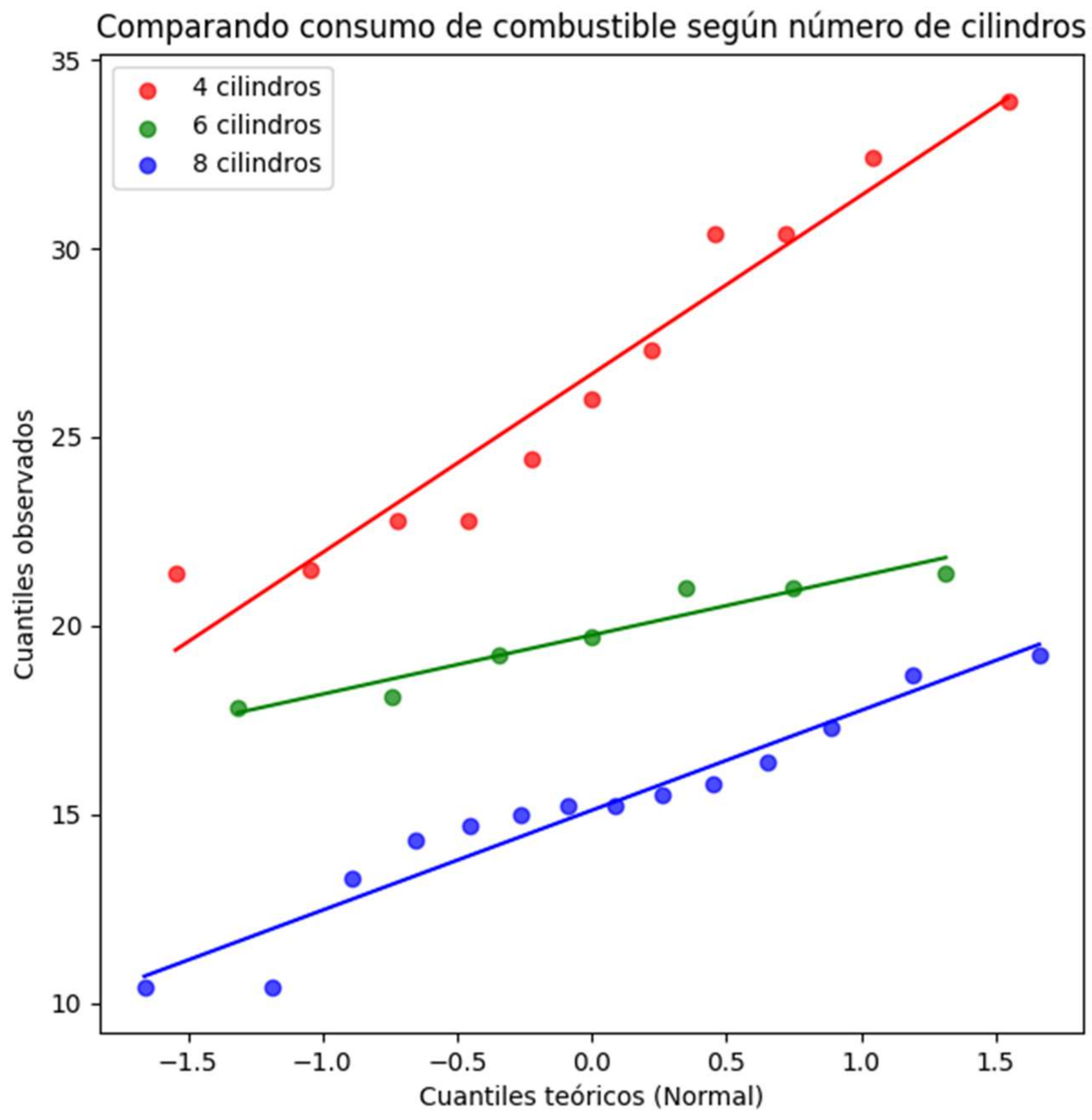


Diagrama de cajas

- A partir de las medidas descriptivas anteriores, de posición y de dispersión, se puede construir una nueva representación gráfica, denominada **diagrama de caja (o diagrama de cajas y bigotes)**. → Muy utilizada!!!

- El diagrama de caja se construye a partir de las siguientes medidas:

  - El primer y el tercer cuartil, *Q*1 y *Q*3, delimitan la caja central.
  - La longitud de la caja viene dada por el RIC (rango intercuartílico).
  - Los límites inferior y superior (bigotes) se calculan como:

$$LI = \text{máx}\{\text{mín}\{x_i\}, Q_1 - 1.5(Q_3 - Q_1)\},$$

$$LS = \text{mín}\{\text{máx}\{x_i\}, Q_3 + 1.5(Q_3 - Q_1)\}.$$

  - La mediana (*Q*2) se representa con una línea horizontal en la caja central.

Diagrama de cajas

- El diagrama de caja puede utilizarse para determinar los valores atípicos de la muestra, que son datos que difieren numéricamente de los demás.
- Formalmente, los datos atípicos son aquellos datos que quedan fuera del intervalo (*LI; LS*).

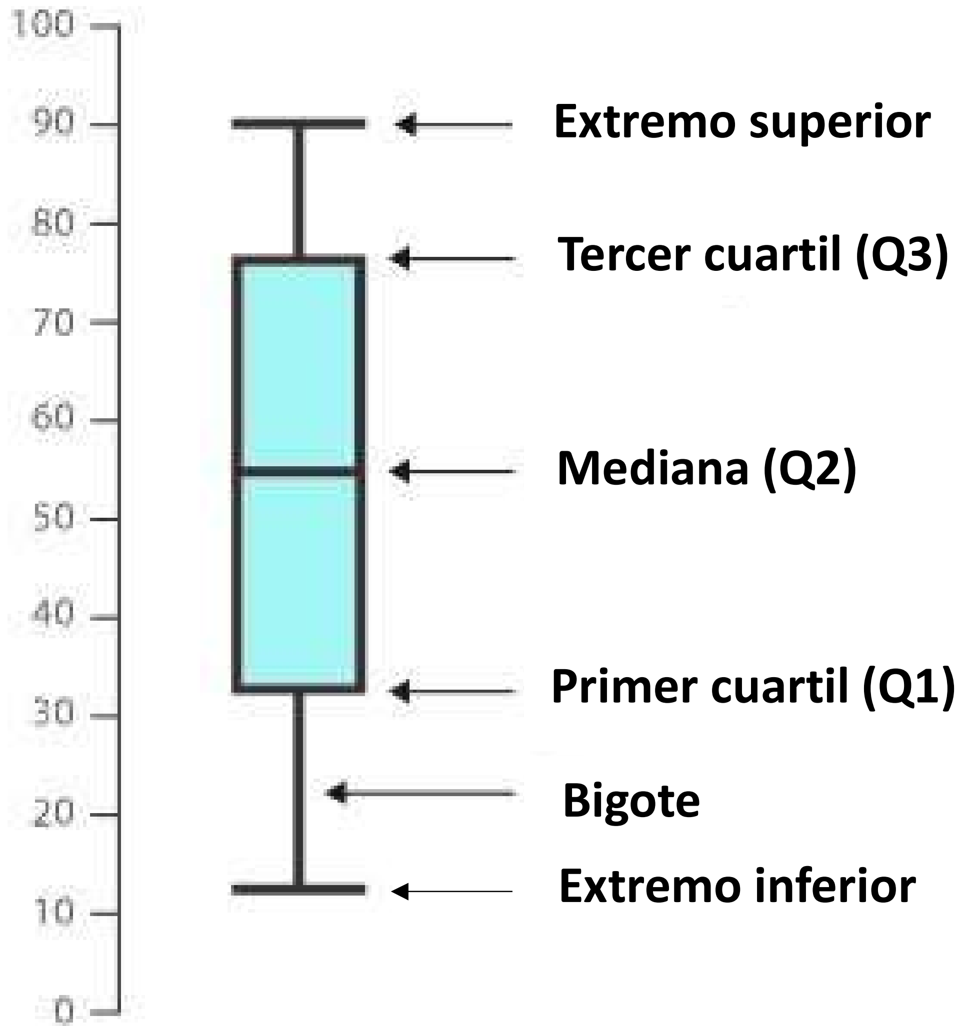


Diagrama de cajas: Ejemplo

- El ejemplo_8.ipynb muestra la comparación realizada con los Q–Qplots, ahora con el diagrama de cajas y bigotes.

- Podemos decir que a medida que aumenta la cilindrada los datos están menos dispersos.

- Además, podemos decir que hay evidencia de que son diferentes, a mayor cilindrada menor consumo.

- Note que para 8 cilindros aparecen unos puntos, estos se denominan **valores extremos o atípicos** (por encima o por debajo de (*LI; LS*).