Generalitat de Catalunya
Departament d'Educació
i Formació Professional

institut sa palomera

ANNEX
ESTADÍSTICA DESCRIPTIVA

Autor: Generada ChatGPT

Curs d'Especialització en Intel·ligència Artificial i BigData

MP5072 Sistemes d'aprenentatge automàtic

Robert Ventura Vall-llovera

Continguts

Introducció

Introducció

La ciència de dades és una fusió de diferents disciplines, entre les que s'inclouen l'estadística i les tecnologies de la informació.

En 1962, John W. Tukey va proposar una reforma de l'estadística i va deixar entreveure una nova disciplina anomenada anàlisis de dades (*Data Analysis*). Si haguéssim de d'exposar un llibre aquest seria: *Exploratory Data Analysis [Tukey, 1977]*. Tukey va presentar diagrames senzills (gràfics de caixa o dispersió) que juntament amb estadístics bàsics (mitjana, mediana, quantils, etc...), ajuden a dibuixar la imatge d'un conjunt de dades.

***Il·lustració 1** - John W. Tukey - Wikipedia*

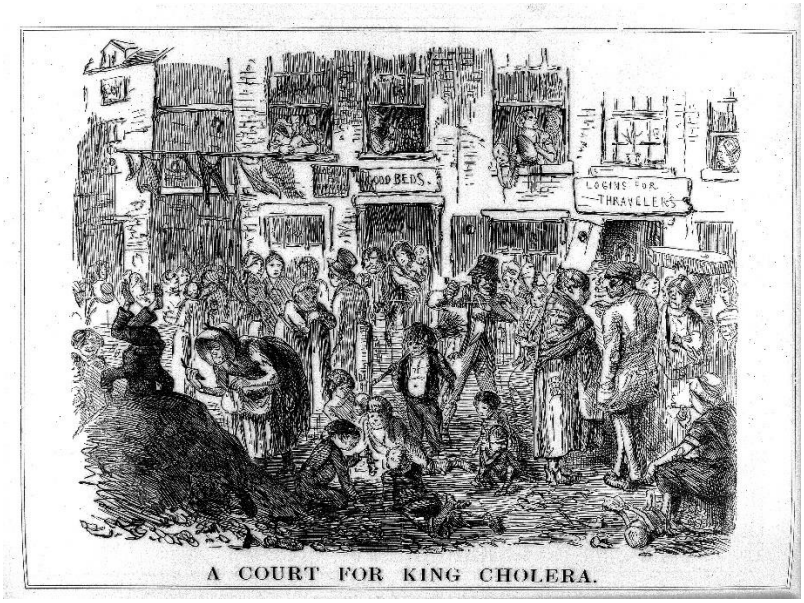
David Donoho, professor d'estadística de la Universitat d'Stanford i exalumne de Tukey reafirma que l'inici de la ciència de dades en el treball de Tukey.

L'anàlisi de dades i la seva importància en la història: El cas de John Snow i el còlera de 1854

L'anàlisi de dades ha estat clau en molts avenços científics i socials al llarg de la història. Un dels exemples més paradigmàtics i influents va ser el cas del metge anglès **John Snow**, considerat el pare de l'epidemiologia moderna, qui el 1854 va utilitzar dades per demostrar la propagació del còlera a Londres i va salvar nombroses vides amb les seves conclusions.

Durant el segle XIX (1801-1900), el còlera era una malaltia devastadora i poc compresa. En aquella època,

la teoria dominant era la del **miasma**, que suggeria que les malalties es transmetien per l'aire contaminat.



Il·lustració 2 - A court for King Cholera - Wikipedia

Quan un brot de còlera va esclatar al barri de Soho, a Londres, el 1854, Snow va decidir investigar d'una manera diferent de la resta de metges de l'època. Va començar a recopilar dades sobre els casos de còlera, registrant on vivien els malalts i traçant un mapa detallat de la zona afectada. Aquest anàlisi li va permetre identificar un patró clar: la majoria dels casos estaven concentrats al voltant d'una bomba d'aigua pública situada a Broad Street (actual Broadwick Street).

Number of Deaths in Houses supplied with Water by:						
Week Ending	Southwark Company	Kent Company	Lambeth Company	Pumps, Wells & other sources	Unknown Sources	Total
September 2	399	38	45	72	116	670
September 9	580	45	72	62	213	972
September 16	524	48	66	44	174	856
September 23	432	28	72	62	130	724
September 30	228	19	25	24	87	383
October 7	121	10	14	9	46	200
	2284	188	294	273	766	3805

TABLE III,

Showing the mortality from Cholera, and the Water Supply, in the Districts of London, in 1849.
The Districts are arranged in the order of their Mortality from Cholera.

District.	Population in the middle of 1849.	Deaths from Cholera.	Deaths by Cholera to 10,000 inhabits.	Annual value of House & Shop room to each person in £.	Water Supply.
Rotherhithe	17,208	352	205	4.238	Southwark and Vauxhall Water Works, Kent Water Works, and Tidal Ditches.
St. Olave, Southwark	19,278	349	181	4.559	Southwark and Vauxhall.
St. George, Southwark	50,900	836	164	3.518	Southwark and Vauxhall, Lambeth.
Bermondsey	45,500	734	161	3.077	Southwark and Vauxhall.
St. Saviour, Southwark	35,227	539	153	5.291	Southwark and Vauxhall.
Newington	63,074	907	144	3.788	Southwark and Vauxhall, Lambeth.
Lambeth	134,768	1618	120	4.389	Southwark and Vauxhall, Lambeth.
Wandsworth	48,446	484	100	4.839	Pump-wells, Southwark and Vauxhall, river Wandle.
Camberwell	51,714	504	97	4.508	Southwark and Vauxhall, Lambeth.
West London	28,829	429	96	7.454	New River.
Bethnal Green	87,263	789	90	1.480	East London.
Shoreditch	104,122	789	76	3.103	New River, East London.
Greenwich	95,954	718	75	3.379	Kent.
Poplar	44,103	313	71	7.360	East London.
Westminster	64,109	437	68	4.189	Chelsea.
Whitechapel	78,590	506	64	3.388	East London.
St. Giles	54,062	285	53	5.635	New River.
Stepney	106,988	501	47	3.319	East London.
Chelsea	53,379	247	46	4.210	Chelsea.
East London	43,495	182	45	4.823	New River.
St. George's, East	47,334	199	42	4.753	East London.
London City	55,816	207	38	17.676	New River.
St. Martin	24,557	91	37	11.844	New River.
Strand	44,254	156	35	7.374	New River.
Holborn	46,134	161	35	5.883	New River.
St. Luke	53,234	183	34	3.731	New River.
Kensington (except Paddington)	110,491	260	33	5.070	West Middlesex, Chelsea, Grand Junction.
Lewisham	32,299	96	30	4.824	Kent.
Belgrave	37,918	105	28	8.875	Chelsea.
Hackney	55,152	139	25	4.397	New River, East London.
Islington	87,761	187	22	5.494	New River.
St. Pancras	160,122	360	22	4.871	New River, Hampstead, West Middlesex.
Clerkenwell	63,499	121	19	4.138	New River.
Marylebone	153,960	261	17	7.586	West Middlesex.
St. James, Westminster	36,426	57	16	12.669	Grand Junction, New River.
Paddington	41,267	35	8	9.349	Grand Junction.
Hampstead	11,572	9	8	5.804	Hampstead, West Middlesex.
Hanover Square & May Fair	33,196	26	8	16.754	Grand Junction.
London	2,280,282	14137	62	—	

TABLE VII.

The mortality from Cholera in the four weeks ending 5th August.

Sub-Districts.	Population in 1851.	Deaths from Cholera in the four wks. ending 5th August.	Water Supply.				
			Southwark & Vauxhall.	Lambeth.	Pump-wells.	River Thames, ditches, &c.	Unascertained.
St. Saviour, Southwark	19,709	26	24	—	—	2	—
St. Olave, Southwark	8,015	19	15	—	—	2	2
St. John, Horsleydown	11,360	18	17	—	—	1	—
St. James, Bermondsey	18,899	29	23	—	—	6	—
St. Mary Magdalen	13,934	20	19	—	—	1	—
Leather Market	15,295	23	23	—	—	—	—
Rotherhithe	17,805	26	17	—	—	9	—
Battersea	10,560	13	10	—	1	2	—
Wandsworth	9,611	2	—	—	—	2	—
Putney	5,280	1	—	—	1	—	—
Camberwell	17,742	19	19	—	—	—	—
Peckham	19,444	4	4	—	—	—	—
Christchurch, Southwk.	16,022	3	2	1	—	—	—
Kent Road	18,126	8	7	1	—	—	—
Borough Road	15,862	21	20	1	—	—	—
London Road	17,836	9	5	4	—	—	—
Trinity, Newington	20,922	14	14	—	—	—	—
St. Peter, Walworth	29,861	20	20	—	—	—	—
St. Mary, Newington	14,033	5	5	—	—	—	—
Waterloo Road (1st)	14,088	5	5	—	—	—	—
Waterloo Road (2nd)	18,348	5	5	—	—	—	—
Lambeth Church (1st)	18,409	5	2	1	—	1	1
Lambeth Church (2nd)	26,784	10	7	2	—	—	1
Kennington (1st)	24,261	11	9	1	1	—	—
Kennington (2nd)	18,848	3	3	—	—	—	—
Brixton	14,610	1	—	1	—	—	—
Clapham	16,290	5	4	—	1	—	—
St.George,Camberwell	15,849	9	7	2	—	—	—
Norwood	3,977	—	—	—	—	—	—
Streatham	9,023	—	—	—	—	—	—
Dulwich	1,632	—	—	—	—	—	—
Sydenham	4,501	—	—	—	—	—	—
	486,936	334	286	14	4	26	4

TABLE VIII.

Mortality from Cholera in the seven weeks ending 26th August.

Sub-Districts.	Population in 1851.	Deaths from Cholera in the seven weeks ending 26th August.	Water Supply.				
			Southwark & Vauxhall.	Lambeth.	Pump-wells.	River Thames and ditches.	Unascertained.
*St. Saviour, Southwark	19,709	125	115	—	—	10	—
*St. Olave, Southwark	8,015	53	43	—	—	5	5
*St. John, Horsleydown	11,360	51	48	—	—	3	—
*St. James, Bermondsey	18,899	123	102	—	—	21	—
*St. Mary Magdalen	13,934	87	83	—	—	4	—
*Leather Market	15,295	81	81	—	—	—	—
*Rotherhithe	17,805	103	68	—	—	35	—
*Battersea	10,560	54	42	—	4	8	—
Wandsworth	9,611	11	1	—	2	8	—
Putney	5,280	1	—	—	1	—	—
*Camberwell	17,742	96	96	—	—	—	—
*Peckham	19,444	59	59	—	—	—	—
Christchurch, Southwk.	16,022	25	11	13	—	—	1
Kent Road	18,126	67	52	5	—	—	—
Borough Road	15,862	71	61	7	—	—	3
London Road	17,836	29	21	8	—	—	—
Trinity, Newington	20,922	68	52	6	—	—	2
St. Peter, Walworth	29,861	90	84	4	—	—	2
St. Mary, Newington	14,033	21	19	1	1	—	—
Waterloo Road (1st)	14,088	10	9	1	—	—	—
Waterloo Road (2nd)	18,348	36	25	8	1	2	—
Lambeth Church (1st)	18,409	18	6	9	—	1	2
Lambeth Church (2nd)	26,748	53	34	13	1	—	5
Kennington (1st)	24,261	71	63	5	3	—	—
Kennington (2nd)	18,848	38	34	3	1	—	—
Brixton	14,610	9	5	2	—	—	2
*Clapham	16,290	24	19	—	5	—	—
St. George, Camberwell	15,849	42	30	9	2	—	1
Norwood	3,977	8	—	2	1	5	—
Streatham	9,023	6	—	1	5	1	—
Dulwich	1,632	—	—	—	—	—	—
Sydenham	4,501	4	—	1	2	—	1
	486,936	1614	1233	98	29	102	22

Snow **va desafiar** les creences dominants de l'època i va **proposar que la còlera es transmetia a través de l'aigua contaminada**. L'any 1854, va començar a investigar un gran brot de còlera al districte de Soho, a Londres, on treballava. Va obtenir una llista de persones que havien mort per còlera i va analitzar on vivien**. Va descobrir que la majoria residia al voltant d'una bomba d'aigua a Broad Street, d'on la gent extreia aigua per beure**.

Curiosament, **també va notar que hi havia dues zones del barri on gairebé ningú es posava malalt**. Va descobrir que aquestes **persones utilitzaven altres fonts d'aigua potable**. A partir d'aquestes observacions, Snow va concloure que l'aigua contaminada de la bomba de Broad Street era la causa de la malaltia.

Tot i que les seves idees van ser rebudes amb escepticisme, Snow va aconseguir convèncer les autoritats de retirar la maneta de la bomba perquè no es pogués utilitzar. Poc després, el brot es va acabar. **Investigacions posteriors van revelar que la bomba extreia aigua d'una zona contaminada per matèria fecal procedent d'una fossa sèptica propera, utilitzada per recollir aigües residuals**.

Snow va continuar fent servir dades per demostrar com es transmetia la còlera. La seva feina pionera va establir les bases de l'epidemiologia moderna—l'estudi de com i per què les malalties es propaguen entre diferents grups de persones.

El seu mapa de la còlera és un exemple clàssic del valor de la cartografia per entendre i controlar la propagació de les malalties. Amb els avenços moderns en sistemes d'informació geogràfica, la cartografia continua sent una eina epidemiològica molt potent.

I aquesta és la història de John Snow i la seva lluita contra la còlera

Conceptes generals i tipus de dades

En qualsevol anàlisi estadístic l'objectiu és extreure conclusions respecte un col·lectiu d'interès anomenat **població**. El tamany de la població (formada per **individus**) pot fer inabordable l'estudi individualitzat de les característiques de cadascun d'ells. Per exemple si volem realitzar un estudi sobre el nivell de glucèmia dels homes adults a Catalunya seria impossibles realitzar una mostra de glucèmia en cadascun d'ells. Per aquest motiu, les mostres es realitzen respecte una **mostra**.

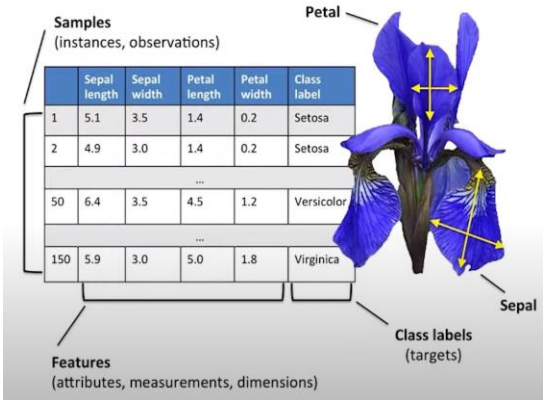
- **Població:** Conjunt complet d'individus, objectes o fenòmens. Exemple: el conjunt d'estudiants de la universitat.
- **Mostra:** subconjunt de la població. Exemple: Els estudiants de Psicologia.
- **Mostreig:** procediment de selecció de mostres.
- **Representativitat de la mostra:** les mostres són part de les poblacions, però no són tota la població. En conseqüència, les característiques de les mostres no són exactament semblants a les de les poblacions. La representativitat de la mostra és el grau en què la mostra s'aproxima a la població.
- **Biaix:** grau de diferència o discrepància entre les característiques de les mostres i les de la població.
- **Anàlisi univariable:** analitza les variables separadament. És més senzill, més fàcil de comprendre i d'interpretar-ne els resultats, però suposa una simplificació molt forta de la realitat. El comportament és un fenomen complex, l'explicació del qual requereix establir com interactua el conjunt de variables rellevant.
- **Anàlisi multivariable:** analitza les dades i incorpora les interrelacions entre les variables. És més complex i de comprensió més dificultosa, però dona una informació més completa de l'objecte estudiat.

Tipus de dades

Els tipus de dades que considerarem són:

- **Dades qualitatives o categòriques**: Expressen una qualitat del individu. No es poden expressar mitjançant una quantitat numèrica. Podríem trobar-hi sinònims com enumeracions i factors.
    - **Nominals:** No té cap sentit la seva ordenació. Per exemple el sexe, color ulls, l'espècie, el país d'origen.
    - **Ordinals**: Es poden ordenar de manera natural. Per exemple la nota (excel·lent, aprovat, suspens), grau de contaminació.
    - **Binàries:** Es consideren un cas especial que només poden tenir dos valors, 0/1, true/false, sí/no.
- **Dades quantitatives o numèriques**: Es refereixen a mesures que expressen una escala numèrica, com edats, longituds, etc...
    - **Discretes**: No es poden dividir. Per exemple el nº de germans, nº d'assignatures.
    - **Contínues**: Les que podem definir amb decimals. Per exemple pes, alçada, velocitat del vent....

Per exemple si mirem el DataFrame de Iris tenim que les 4 primeres columnes són dades quantitatives i en canvi la 5a és una dada qualitativa o un factor.



	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

Estadístics de localització

Quan explorem les dades volem obtenir un valor típic per cada característica a on es trobi representada la majoria de les dades, és a dir, una tendència central.

Cal dir que els estadística han desenvolupat estimacions alternatives a la mitjana ja que en molts caso aquesta mesura no sempre és la millor per representar un valor central.

- **Moda (Mo):** és el valor més freqüent. Si hi ha més de dues modes direm que són bimodals, si no hi ha cap moda és una distribució amodal.
    - Exemple X: 4,5,4,5,4,5,5,5,5,5,7,8,7 => $Mo = 5$

- **Mediana (Md) :** És el valor central d'una llista de dades ordenades.
    - Si el nombre de dades és senar la mediana és el valor central que separa la llista de dades en dues mitats
        - Exemple X: 6,7,8,9,10 => $Md = 8$
    - Si el nombre de dades és parell. La mediana és la mitjana dels valors que divideix les dades ordenades en dues mitats.
        - Exemple X: 6,7,8,9 => $Md = 7.5 = \frac{(7+8)}{2}$

- **Mitjana (M):** És el valor mig del grup de dades. S'obté sumant totes les dades i dividint el resultat per la quantitat de les mateixes.

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

    - Exemple X: 1,5,6,6,7,12 => $M = \bar{X} = \frac{1+5+6+6+7+12}{6} = \frac{37}{6} = 6.1\hat{6}$

La **mitjana no és un estadístic robust** ja que està molt influenciat per valors atípics (*outliers*). Per exemple, en un cas d'un jugador de futbol que en un partit ha fet 5 gols, aquests 5 gols tindran un pes molt gran al resultat de la mitjana quan la majoria de valors per aquell jugador és de 0,1 gols.

- **Mitjana truncada (Mt):** La mitjana truncada és una variació de la mitjana ignorant un número fix, en cada extrem, dels valors ordenats i a continuació es calcula la mitjana dels valors restants.
  La mitjana truncada elimina la influència dels valors extrems. Per exemple en el certes disciplines s'eliminen les puntuacions mínimes i màximes d'uns quants jutges i la puntuació final és la mitjana de les puntuacions del jutges restants. Això permet que un sol jutge no pugui manipular la puntuació per afavorir o desafavorir un concursant.

$$Mt = \bar{X} = \frac{\sum_{i=p+1}^{n-p} x_i}{n - 2p}$$

- **Exemple** X: 1,5,6,6,7,12 => $Mt_{(p=1)} = \bar{X} = \frac{5+6+6+7}{4} = \frac{24}{4} = 6$

- **Mitjana ponderada (Mp):** La mitjana ponderada es calcula multiplicant cada valor de les dades per un pes especificat i dividint la seva suma per la suma de les ponderacions..

$$Mp = \bar{X}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

  - **Exemple** X: 5(3),5(3),7(8) => $Mp = \bar{X}_w = \frac{5 \cdot 3+5 \cdot 3+7 \cdot 8}{3+3+8} = \frac{15+15+56}{14} = \frac{86}{14} = 6.14$

  La mitjana ponderada és útil quan sabem que en les observacions hi ha variables que han de tenir un pes major o inferior. Per exemple, si agafem la mitjana de diferents valors de sensors i un d'aquests és menys precís, podríem reduir la ponderació de les dades d'aquest sensor.
  Un cas típic és quan un professor ha d'avaluar l'alumnat mitjançant 3 treballs, però vol donar més pes a un d'aquets tres treballs per reflectir millor les hores dedicades a aquest.

Estadístics de dispersió o variabilitat

La localització que s'ha vist anteriorment és només una dimensió per veure el resum d'una característica. Una segona dimensió és la dispersió.

Les mesures de dispersió, també anomenades mesures de variabilitat (*variability*), mostren el grau d'agrupació o dispersió dels valors d'un conjunt de mostres per mitjà d'un nombre. Ens diuen si els diferents valors d'una variable estan molt o poc allunyades de la mitjana.

Rang

El rang o recorregut estadístic és la diferència entre el valor **mínim** i el valor **màxim** de les observacions que tenim.

Per exemple si tenim dos mostres que ens indiquen el pes d'un grup de persones amb un edat determinada:

Pesos(Kg) Persones 40 anys									
Mostra A	105	100	60	65	80	85	77	82	79
Mostra B	82	83	85	78	82	80	77	74	81

- Mostra A: Rang [105-60] = **45**
- Mostra B: Rang [85-74] = **11**

Amb el rang ja podem veure que la mostra B mostra menys dispersió que la mostra A ja que el valor dels rangs són molt diferents.

Desviacions

Les desviacions són les diferències entre els valors observats i la mesura de la localització utilitzada.

Per exemple si agafem la Mostra A del cas anterior i agafem la mesura de la mitjana com a localització de la mostra tindrem les següents desviacions.

Pesos(Kg) Persones 40 anys										**Avg**	**Med**
Mostra A	105	100	60	65	80	85	77	82	79	81,4	80
Desviacions (mitjana)	24	19	21	16	1	4	4	1	2	10.2	-
Desviacions (mediana)	25	20	20	15	0	5	3	2	1	10.1	-

Variància

La variància d'unes dades és la mitjana aritmètica del quadrat de les desviacions respecte a la pròpia mitjana. Amb aquest valor podem veure quan de lluny estem de la mitjana aritmètica.

Es simbolitza com a σ^2 (sigma minúscula grega al quadrat)

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

es pot simplificar

$$\sigma^2 = \frac{\sum_{i=1}^{N} {x_i}^2}{N} - \bar{x}^2$$

La variància ens determina la dispersió del conjunt d'observacions. Una variància gran significa que les dades estan molt disperses. En canvi una variància petita significa que els valors estan, en general, pròxims a la mitjana.

Una variància igual a zero ($\sigma^2=0$) implica que tots els valors són iguals i ,per tant, també coincideixen amb la mitjana aritmètica.

La fórmula anterior és l'anomenada variància poblacional, també tenim la variància mostral que és agafant N-1 valors de la mostra. Aquest valor que extraiem és aleatori.

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

Pesos(Kg) Persones 40 anys									
Mostra A	105	100	60	65	80	85	77	82	79
Mostra B	82	83	85	78	82	80	77	74	81

Pesos Persones 40 anys	$\bar{x}$	σ^2	Veiem que la variància entre les dues mostres són diferents encara que les mitjanes siguin molt semblants. Això significa que les observacions en la Mostra A són més disperses que en la Mostra B
Mostra A	81,33	186,22	
Mostra B	80,22	10,173	

Desviació típica / estàndard

La desviació típica és l'arrel quadrada de la variància i es representa com a **σ** (sigma minúscula grega)

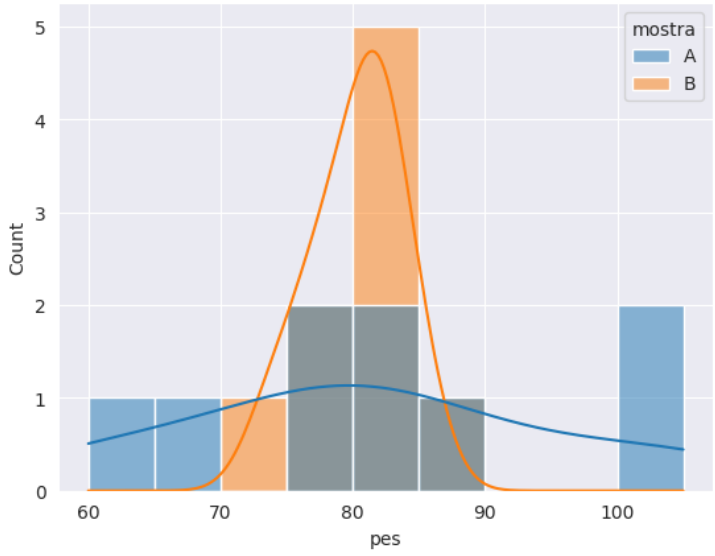
Per calcular-la el que fem és calcular la variància i li apliquem l'arrel quadrada.

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

La interpretació és semblant a la variància, però l'escala de valors és el mateix que la de les dades.

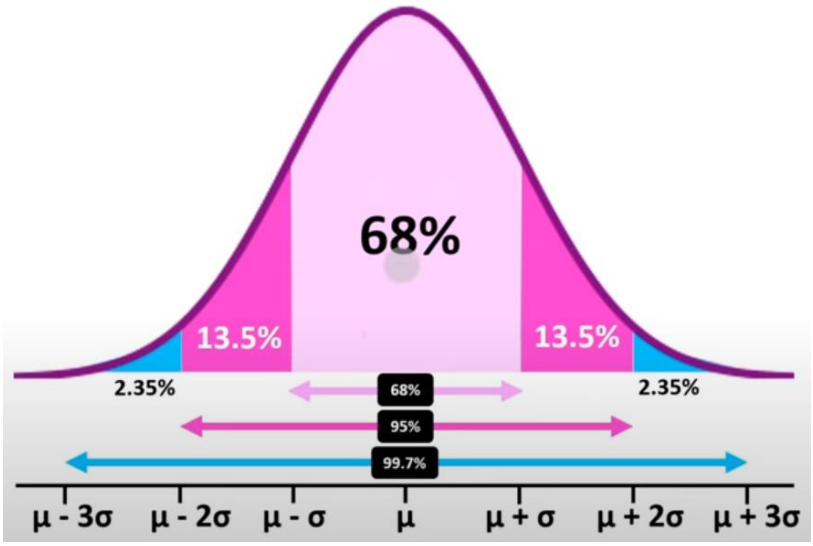
Pesos	$\bar{x}$	**σ²**	**σ**	Amb la desviació típica veiem el mateix que hem vist amb la variància, però ara ho veiem amb el mateix ordre de magnitud que el valor de les observacions i podem dir informalment que a la Mostra B hi ha una desviació d'uns 3 Kg respecte la mitjana.
Mostra A	81,33	186,22	13.646	
Mostra B	80,22	10,173	3.189	

```
import seaborn as sns
import pandas as pd

data = {'pes':[105, 100, 60, 65, 80, 85, 77, 82, 79,
               82, 83, 85, 78, 82, 80, 77, 74, 81],
        'mostra':['A' for x in range(9)]+['B' for x in range(9)]}

df = pd.DataFrame(data)
sns.set_style("darkgrid")
sns.histplot(data=df, x="pes", hue="mostra", binwidth=5, kde=True)
```



Regla empírica del 68-95-99.7

La regla empírica ens diu que en una mostra de dades que segueixi un distribució normal tenim que a una distància d'una desviació estàndard tenim el 68% de les dades. A dues desviacions estàndards el 95% i a tres el 99,7%



Percentil

En un conjunt de dades, el percentil és un estadístic que divideix les dades ordenades en 100 parts iguals. L'n-èsim percentil és el valor per sota del qual es troba el n% de les dades.

Per exemple, per trobar el percentil 80, ordenem les dades. Llavors començant per el valor més petit continuem fins el 80% de les dades. La mediana seria el percentil 50.

És habitual informar dels quartils (percentil 25 (primer quartil), 50(segon quartil=mediana) i 75(tercer quartil)).

Tenim:

- Q1 (primer quartil): El 25% de les dades són menors o iguals
- Q2 (segon quartil): El 50% de les dades a cada costat
- Q3 (tercer quartil): El 75% de les dades són menors o iguals.

Si diem que la nota d'un alumne està al percentil 90 significa que aquella nota és millor que el 90% dels exàmens dels alumnes.

Rang interquartítlic

És una mesura de variabilitat que es calcula com la diferència entre el tercer i el primer quartil d'una sèrie de dades.

És una mesura més robusta que la desviació estàndard perquè no es veu tan afectada per els valors extrems.

Dividim les dades en 4 parts. Primer hem de ordenar les dades

Pesos(Kg) Persones 40 anys									
Mostra A	60	65	77	78	80	82	85	100	105
Mostra B	74	77	78	80	81	82	82	83	85
Quartils			Q1		Q2		Q3		

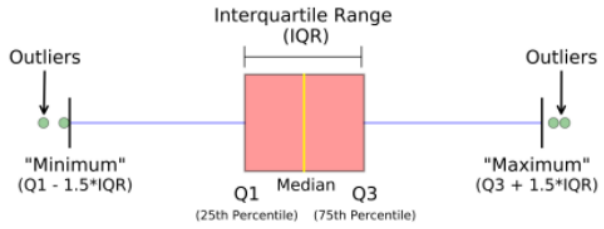
El rang interquartílic és Q3-Q1 i com hem dit es centra en la majoria de les dades i és tolerant a dades atípiques perquè ens centrem a on estan la majoria de les observacions.

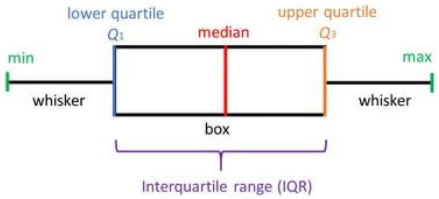
Pesos (Kg)	$\bar{x}$	σ^2	σ	Q3-Q1
Mostra A	81,33	186,22	13,646	8
Mostra B	80,22	10,173	3,189	4

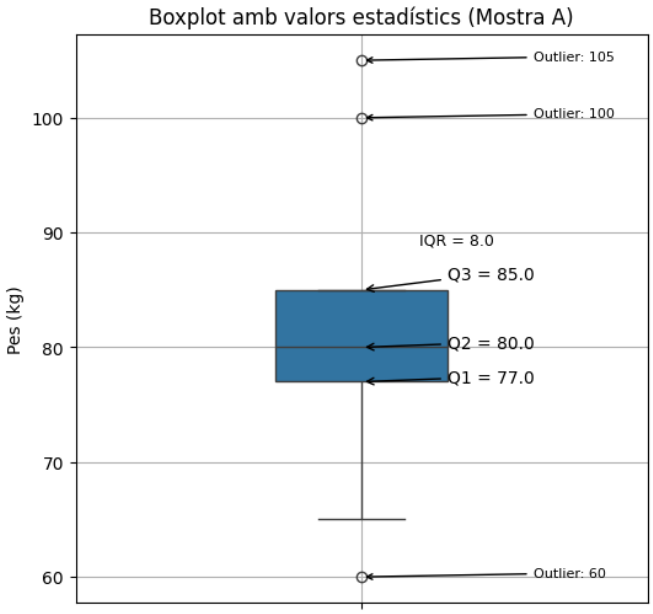
```
import numpy as np

def estudiDispersioMostra(mostra):
    print("Mitjana:", round(np.mean(mostra),3))

    print("Rang:", np.max(mostra)-np.min(mostra))

    print("Variança poblacional:", round(np.var(mostra),3))
    #print("Variança mostral:", round(np.var(mostra, ddof=1),3))

    print("Desviació estàndard poblacional:", round(np.std(mostra),3))
    #print("Desviació estàndard mostral:", round(np.std(mostra, ddof=1),3))

    #Càlcul del rang interquartílic
    Q1 = np.percentile(mostra,25)
    Q3 = np.percentile(mostra,75)

    print("Rang interquartílic:", Q3-Q1)


print ("------------------------------")
print ("--- ESTUDI DE LA MOSTRA A ---")
print ("------------------------------")
pesosA=[105,100,60,65,80,85,77,82,79]
estudiDispersioMostra(pesosA)

print ("------------------------------")
print ("--- ESTUDI DE LA MOSTRA B ---")
```

```
print ("----------------------------")
pesosB=[82,83,85,78,82,80,77,74,81]
estudiDispersioMostra(pesosB)

----------------------------
--- ESTUDI DE LA MOSTRA A ---
----------------------------
Mitjana: 81.444
Rang: 45
Variança poblacional: 185.58
Desviació estàndard poblacional: 13.623
Rang interquartílic: 8.0
----------------------------
--- ESTUDI DE LA MOSTRA B ---
----------------------------
Mitjana: 80.222
Rang: 11
Variança poblacional: 10.173
Desviació estàndard poblacional: 3.189
Rang interquartílic: 4.0
```

Els diagrames de caixa o boxplots, presentats per Tukey (1977), utilitzen percentils i permeten visualitzar la distribució de les dades d'una forma ràpida. La figura exposada mostra el boxplot de la Mostra A







Amb Pandas podem utilitzar `describe()` per veure els estadístics de localització i variabilitat bàsics de cada variable. En aquest dataframe només tenim la variable pes.

```
import pandas as pd

data = {
    'pes': [105, 100, 60, 65, 80, 85, 77, 82, 79,
            82, 83, 85, 78, 82, 80, 77, 74, 81],
    'mostra': ['A'] * 9 + ['B'] * 9
}
df = pd.DataFrame(data)

df_A = df[df['mostra'] == 'A'].describe()
print(df_A)

            pes
count    9.000000
mean    81.444444
std     14.449145
min     60.000000
25%     77.000000
50%     80.000000
75%     85.000000
max    105.000000
```

Exercici:

El departament de Direcció Esportiva d'un club de bàsquet us ha sol·licitat un informe per valorar el fitxatge d'un nou jugador.

S'han recopilat dades de 5 jugadors, recollint el nombre de punts anotats per cadascun d'ells en 9 partits.

Analitza les dades mitjançant els estadístics bàsics vists fins ara i ajuda't de representacions gràfiques per facilitar-ne la interpretació.

Interpreta els resultats i elabora una recomanació raonada:

- **Quin jugador fixaries i per què?**

Partits/Jugador	1	2	3	4	5	6	7	8	9
Jugador A	10	6	7	12	14	18	15	4	10
Jugador B	5	6	3	0	8	8	7	6	7
Jugador C	8	3	10	3	5	6	7	4	0
Jugador D	0	2	4	5	8	10	10	15	38
Jugador E	15	12	3	1	5	8	10	4	15
Jugador F	12	25	7	12	14	18	15	4	10

Exploració de la distribució de les dades

Taules de freqüències i histogrames

Les taules de freqüències són una de les tècniques bàsiques pel resum de la informació a partir de la mostra de les dades. La seva construcció és senzilla però en conjunts grans el càlcul pot resultar feixuc.

Donada una variable qualitativa, per cada nivell podem contar quantes dades hi ha d'aquest nivell (**freqüència absoluta**) i quina fracció del total representen (**freqüència relativa**) o la proporció.

Notació

Diferents nivells qualitatius (diferents valors que pot agafar la variable qualitativa): $l_1, l_2, \dots, l_k$

Tenim **n** observacions d'aquest tipus de dades denotem per cada observació: $X_1, X_2, \dots X_n$

Els resultats que obtenim han d'estar dins dels nivell: : $X_j \epsilon \{l_1, l_2, \dots, l_k\}$

Amb les notacions anteriors tenim:

- La **freqüència absoluta**, n_j, del nivell l_j en aquesta variable qualitativa és el número d'observacions en què X_i agafa el valor de l_j
- La **freqüència relativa** del nivell l_j en aquesta variable qualitativa és la fracció/proporció

$$f_j = \frac{n_j}{n}$$

REFERÈNCIES

- Peter Bruce, Andre Bruce & Peter Gedeck (2020). Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python. O'Reilly Media.