Statistica I - Lezione 1.

Def: un campione/database è un insieme di punti dati

$$\{x_1, x_2, \dots, x_n\}$$

dove $n \in \mathbb{N}$ è il numero di punti dato (o numerosità del database).

Esempio: Compleanni

$\{09052004, 08062004, \dots\}$ $\quad n \approx 250$

Esempio: Salari (in migliaia di euro)

$\{60, 123, 1350, \dots\}$ $\quad n \approx 250$

Esempio: Test PCR

$\{-, -, +, -, +, \dots\}$

Esempio: Tempi di vita di lampadine prodotte da una fabbrica

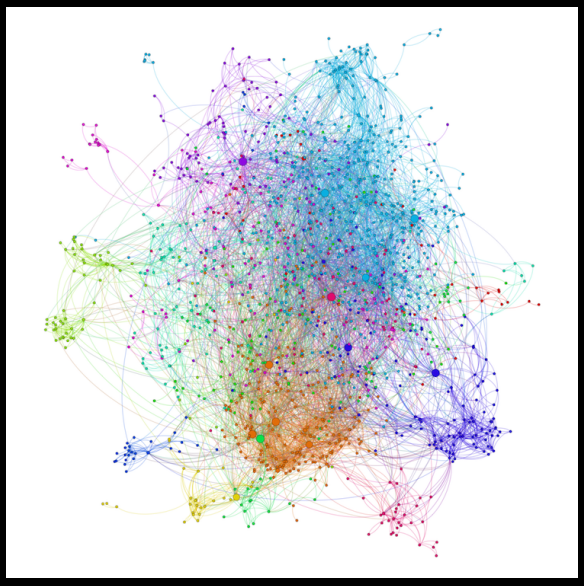
$n = 200$

TABLE 2.3 *Life in Hours of 200 Incandescent Lamps*

Item Lifetimes									
1,067	919	1,196	785	1,126	936	918	1,156	920	948
855	1,092	1,162	1,170	929	950	905	972	1,035	1,045
1,157	1,195	1,195	1,340	1,122	938	970	1,237	956	1,102
1,022	978	832	1,009	1,157	1,151	1,009	765	958	902
923	1,333	811	1,217	1,085	896	958	1,311	1,037	702
521	933	928	1,153	946	858	1,071	1,069	830	1,063
930	807	954	1,063	1,002	909	1,077	1,021	1,062	1,157
999	932	1,035	944	1,049	940	1,122	1,115	833	1,320
901	1,324	818	1,250	1,203	1,078	890	1,303	1,011	1,102
996	780	900	1,106	704	621	854	1,178	1,138	951
1,187	1,067	1,118	1,037	958	760	1,101	949	992	966
824	653	980	935	878	934	910	1,058	730	980
844	814	1,103	1,000	788	1,143	935	1,069	1,170	1,067
1,037	1,151	863	990	1,035	1,112	931	970	932	904
1,026	1,147	883	867	990	1,258	1,192	922	1,150	1,091
1,039	1,083	1,040	1,289	699	1,083	880	1,029	658	912
1,023	984	856	924	801	1,122	1,292	1,116	880	1,173
1,134	932	938	1,078	1,180	1,106	1,184	954	824	529
998	996	1,133	765	775	1,105	1,081	1,171	705	1,425
610	916	1,001	895	709	860	1,110	1,149	972	1,002

Esempio: Instagram network

{(@andreamenconi, 10000),
(@vlodimirp, 2),
(@figliolellestelle2, 50)}



Esempio: Temperature

[In Fahrenheit degrees. Airport data except as noted. Based on standard 30-year period, 1961 through 1990]

State	Station	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.	Annual avg.
AL	Mobile	40.0	42.7	50.1	57.1	64.4	70.7	73.2	72.9	68.7	57.3	49.1	43.1	57.4
AK	Juneau	19.0	22.7	26.7	32.1	38.9	45.0	48.1	47.3	42.9	37.2	27.2	22.6	34.1
AZ	Phoenix	41.2	44.7	48.8	55.3	63.9	72.9	81.0	79.2	72.8	60.8	48.9	41.8	59.3
AR	Little Rock	29.1	33.2	42.2	50.7	59.0	67.4	71.5	69.8	63.5	50.9	41.5	33.1	51.0
CA	Los Angeles	47.8	49.3	50.5	52.8	56.3	59.5	62.8	64.2	63.2	59.2	52.8	47.9	55.5
	Sacramento	37.7	41.4	43.2	45.5	50.3	55.3	58.1	58.0	55.7	50.4	43.4	37.8	48.1
	San Diego	48.9	50.7	52.8	55.6	59.1	61.9	65.7	67.3	65.6	60.9	53.9	48.8	57.6
	San Francisco	41.8	45.0	45.8	47.2	49.7	52.6	53.9	55.0	55.2	51.8	47.1	42.7	49.0
CO	Denver	16.1	20.2	25.8	34.5	43.6	52.4	58.6	56.9	47.6	36.4	25.4	17.4	36.2
CT	Hartford	15.8	18.6	28.1	37.5	47.6	56.9	62.2	60.4	51.8	40.7	32.8	21.3	39.5
DE	Wilmington	22.4	24.8	33.1	41.8	52.2	61.6	67.1	65.9	58.2	45.7	37.0	27.6	44.8
DC	Washington	26.8	29.1	37.7	46.4	56.6	66.5	71.4	70.0	62.5	50.3	41.1	31.7	49.2
FL	Jacksonville	40.5	43.3	49.2	54.9	62.1	69.1	71.9	71.8	69.0	59.3	50.2	43.4	57.1
	Miami	59.2	60.4	64.2	67.8	72.1	75.1	76.2	76.7	75.9	72.1	66.7	61.5	69.0
GA	Atlanta	31.5	34.5	42.5	50.2	58.7	66.2	69.5	69.0	63.5	51.9	42.8	35.0	51.3
HI	Honolulu	65.6	65.4	67.2	68.7	70.3	72.2	73.5	74.2	73.5	72.3	70.3	67.0	70.0
ID	Boise	21.6	27.5	31.9	36.7	43.9	52.1	57.7	56.8	48.2	39.0	31.1	22.5	39.1
IL	Chicago	12.9	17.2	28.5	38.6	47.7	57.5	62.6	61.6	53.9	42.2	31.6	19.1	39.5
	Peoria	13.2	17.7	29.8	40.8	50.9	60.7	65.4	63.1	55.2	43.1	32.5	19.3	41.0
IN	Indianapolis	17.2	20.9	31.9	41.5	51.7	61.0	65.2	62.8	55.6	43.5	34.1	23.2	42.4
IA	Des Moines	10.7	15.6	27.6	40.0	51.5	61.2	66.5	63.6	54.5	42.7	29.9	16.1	40.0
KS	Wichita	19.2	23.7	33.6	44.5	54.3	64.6	69.9	67.9	59.2	46.6	33.9	23.0	45.0
KY	Louisville	23.2	26.5	36.2	45.4	54.7	62.9	67.3	65.8	58.7	45.8	37.3	28.6	46.0
LA	New Orleans	41.8	44.4	51.6	58.4	65.2	70.8	73.1	72.8	69.5	58.7	51.0	44.8	58.5
ME	Portland	11.4	13.5	24.5	34.1	43.4	52.1	58.3	57.1	48.9	38.3	30.4	17.8	35.8
MD	Baltimore	23.4	25.9	34.1	42.5	52.6	61.8	66.8	65.7	58.4	45.9	37.1	28.2	45.2
MA	Boston	21.6	23.0	31.3	40.2	49.8	59.1	65.1	64.0	56.8	46.9	38.3	26.7	43.6
MI	Detroit	15.6	17.6	27.0	36.8	47.1	56.3	61.3	59.6	52.5	40.9	32.2	21.4	39.0
	Sault Ste. Marie	4.6	4.8	15.3	28.4	38.4	45.5	51.3	51.3	44.3	36.2	25.9	11.8	29.8
MN	Duluth	−2.2	2.8	15.7	28.9	39.6	48.5	55.1	53.3	44.5	35.1	21.5	4.9	29.0
	Minneapolis-St. Paul	2.8	9.2	22.7	36.2	47.6	57.6	63.1	60.3	50.3	38.8	25.2	10.2	35.3
MS	Jackson	32.7	35.7	44.1	51.9	60.0	67.1	70.5	69.7	63.7	50.3	42.3	36.1	52.0
MO	Kansas City	16.7	21.8	32.6	43.8	53.9	63.1	68.2	65.7	56.9	45.7	33.6	21.9	43.7
	St. Louis	20.8	25.1	35.5	46.4	56.0	65.7	70.4	67.9	60.5	48.3	37.7	26.0	46.7
MT	Great Falls	11.6	17.2	22.8	31.9	40.9	48.6	53.2	52.2	43.5	35.8	24.3	14.6	33.1

Source: U.S. National Oceanic and Atmospheric Administration, Climatography of the United States, No. 81.

Esempio Topi

	Germ-Free Mice
1	58, 92, 93, 94, 95
2	02, 12, 15, 29, 30, 37, 40, 44, 47, 59
3	01, 01, 21, 37
4	15, 34, 44, 85, 96
5	29, 37
6	24
7	07
8	00

	Conventional Mice
1	59, 89, 91, 98
2	35, 45, 50, 56, 61, 65, 66, 80
3	43, 56, 83
4	03, 14, 28, 32

2.2 Organizzazione e descrizione dei dati

Per leggere, interpretare e capire il contenuto di un database e le caratt. globali dobbiamo presentare il suo contenuto in maniera intelleggibile.

⟹ rappresentiamo le frequenze (freq. assolute) nel database, cioè il numero di volte che un valore è osservato.
(efficiente se il numero di valori distinti è basso)

Queste possono essere rappresentate in diversi modi:

- Tabelle:

valore	v_1	v_2	v_3	v_4	...
frequenza	f_1	f_2	f_3	f_4	...

Esempio. (Stipendi)

Tabella 2.1 Stipendi annuali iniziali. Dati in migliaia di dollari.

Stipendio iniziale	Frequenza
27	4
28	1
29	3
30	5
31	8
32	10
34	5
36	2
37	3
40	1

$n = 42$

- Grafico a bastoncini / Grafico a barre.

Esempio (Stipendi)

Grafico a bastoncini

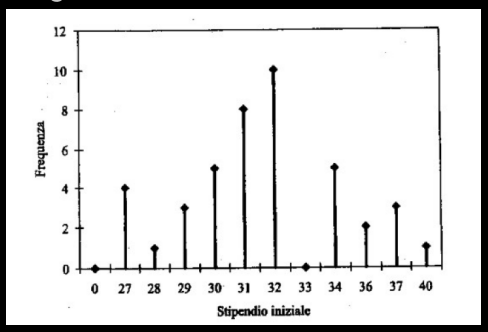


Grafico a bare

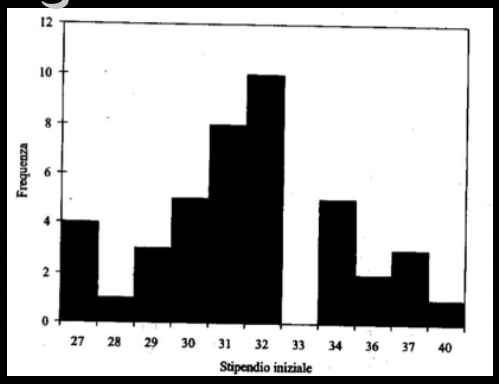
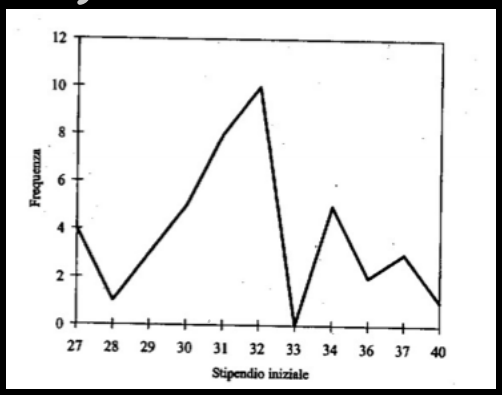
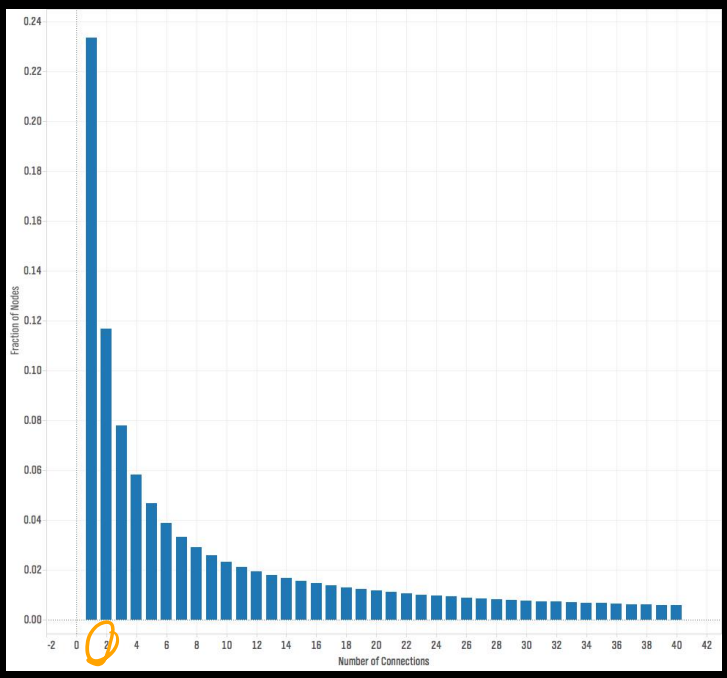


Grafico a linee



Esempio (twitter):

valore			
# followers	2	3	...
frequenza	10000	7000	



Talvolta il numero assoluto f di realizzazioni di un valore ci interessa meno della sua frequenza relativa (frazione di volte che un valore appare in un database di ampiezza n), data da

$$\frac{f_i}{n}$$

f_i → frequenza assoluta dell'i-esimo valore

n → numerosità del dataset

Come nel caso delle frequenze assolute, le frequenze relative si possono rappresentare in

- tabelle
- grafici a barre
- grafici a bastoncini
- grafici a linee

Esempio

$n = 42$

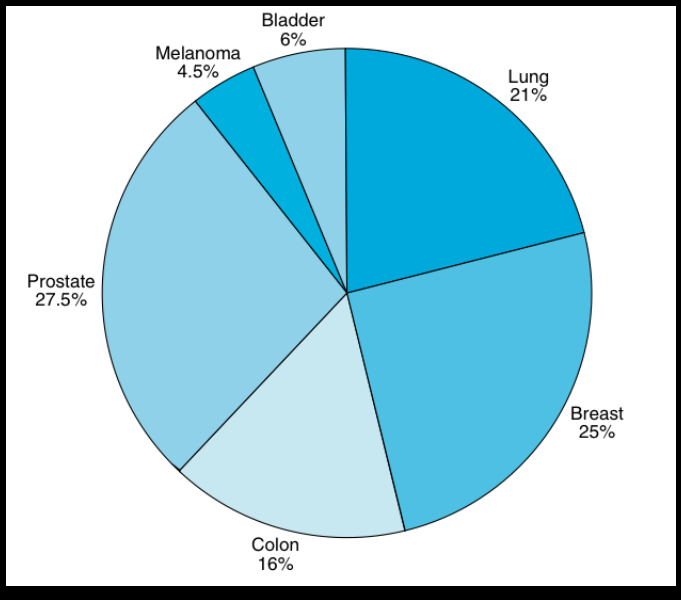
Tabella 2.2 Redditi annuali iniziali. Dati in migliaia di dollari.

Stipendio iniziale	Frequenza relativa
27	4/42 ≈ 0.0952 = 9.52%
28	1/42 ≈ 0.0238 = 2.38%
29	3/42 ≈ 0.0714 = 7.14%
30	5/42 ≈ 0.1190 = 11.90%
31	8/42 ≈ 0.1905 = 19.05%
32	10/42 ≈ 0.2381 = 23.81%
34	5/42 ≈ 0.1190 = 11.90%
36	2/42 ≈ 0.0476 = 4.76%
37	3/42 ≈ 0.0714 = 7.14%
40	1/42 ≈ 0.0238 = 2.38%

Inoltre, un'altra rappresentazione possibile è quella tramite diagramma a torta

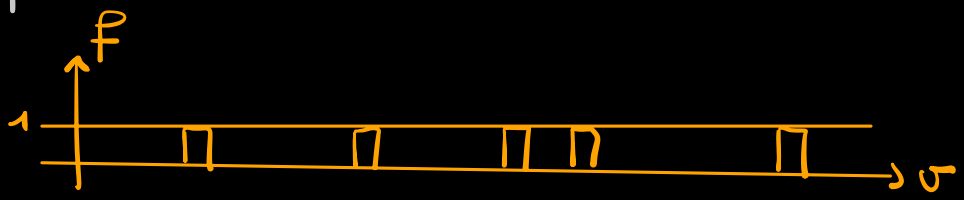
Esempio

Type of Cancer	Number of New Cases	Relative Frequency
Lung	42	.21
Breast	50	.25
Colon	32	.16
Prostate	55	.275
Melanoma	9	.045
Bladder	12	.06



I metodi esposti sopra perdono il loro senso se il numero di valori diversi è comparabile all'ampiezza del campione:

Controesempio



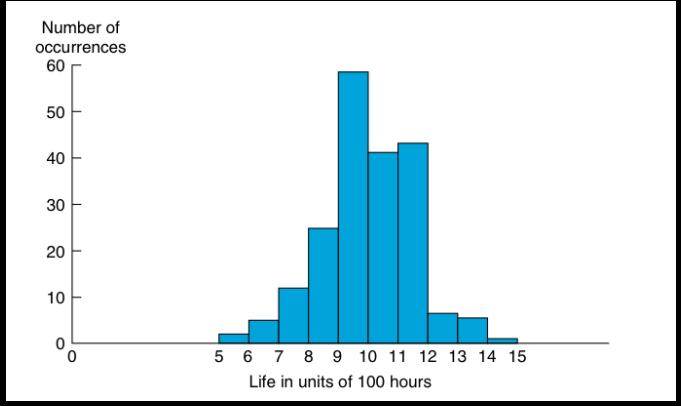
Se questo è il caso, possiamo dividere i dati in gruppi di valori contigui (classi) ed associare ogni punto dato a una di esse.

TABLE 2.3 *Life in Hours of 200 Incandescent Lamps*

Item Lifetimes									
1,067	919	1,196	785	1,126	936	918	1,156	920	948
855	1,092	1,162	1,170	929	950	905	972	1,035	1,045
1,157	1,195	1,195	1,340	1,122	938	970	1,237	956	1,102
1,022	978	832	1,009	1,157	1,151	1,009	765	958	902
923	1,333	811	1,217	1,085	896	958	1,311	1,037	702
521	933	928	1,153	946	858	1,071	1,069	830	1,063
[illegible]	807	954	1,063	1,002	909	1,077	1,021	1,062	1,157
999	932	1,035	944	1,049	940	1,122	1,115	833	1,320
901	1,324	818	1,250	1,203	1,078	890	1,303	1,011	1,102
996	780	900	1,106	704	621	854	1,178	1,138	951
1,187	1,067	1,118	1,037	958	760	1,101	949	992	966
824	653	980	935	878	934	910	1,058	730	980
844	814	1,103	1,000	788	1,143	935	1,069	1,170	1,067
1,037	1,151	863	990	1,035	1,112	931	970	932	904
1,026	1,147	883	867	990	1,258	1,192	922	1,150	1,091
1,039	1,083	1,040	1,289	699	1,083	880	1,029	658	912
1,023	984	856	924	801	1,122	1,292	1,116	880	[illegible]
1,134	932	938	1,078	1,180	1,106	1,184	954	824	529
998	996	1,133	765	775	1,105	1,081	1,171	705	1,425
610	916	1,001	895	709	860	1,110	1,149	972	1,002

TABLE 2.4 *A Class Frequency Table*

Class Interval	Frequency (Number of Data Values in the Interval)
500–600	2
600–700	5
700–800	12
800–900	25
900–1000	58
1000–1100	41
1100–1200	43
1200–1300	7
1300–1400	6
1400–1500	1



⚠ La definizione delle classi gioca un ruolo importante!

- se troppo poche ⟶ una colonne
- se troppe ⟶ troppe colonne "basse".

2.3 Le grandezze che sintetizzano i dati

Spesso i database a disposizione sono così grandi (milioni di punti dati, centinaia di categorie, ...) che risulta necessario rappresentarli/sintetizzarli ulteriormente.

Esempio (Temperat)

TABLE 2.5 *Normal Daily Minimum Temperature — Selected Cities*

[In Fahrenheit degrees. Airport data except as noted. Based on standard 30-year period, 1961 through 1990]

State	Station	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.	Annual avg.
AL	Mobile	40.0	42.7	50.1	57.1	64.4	70.7	73.2	72.9	68.7	57.3	49.1	43.1	57.4
AK	Juneau	19.0	22.7	26.7	32.1	38.9	45.0	48.1	47.3	42.9	37.2	27.2	22.6	34.1
AZ	Phoenix	41.2	44.7	48.8	55.3	63.9	72.9	81.0	79.2	72.8	60.8	48.9	41.8	59.3
AR	Little Rock	29.1	33.2	42.2	50.7	59.0	67.4	71.5	69.8	63.5	50.9	41.5	33.1	51.0
CA	Los Angeles	47.8	49.3	50.5	52.8	56.3	59.5	62.8	64.2	63.2	59.2	52.8	47.9	55.5
	Sacramento	37.7	41.4	43.2	45.5	50.3	55.3	58.1	58.0	55.7	50.4	43.4	37.8	48.1
	San Diego	48.9	50.7	52.8	55.6	59.1	61.9	65.7	67.3	65.6	60.9	53.9	48.8	57.6
	San Francisco	41.8	45.0	45.8	47.2	49.7	52.6	53.9	55.0	55.2	51.8	47.1	42.7	49.0
CO	Denver	16.1	20.2	25.8	34.5	43.6	52.4	58.6	56.9	47.6	36.4	25.4	17.4	36.2
CT	Hartford	15.8	18.6	28.1	37.5	47.6	56.9	62.2	60.4	51.8	40.7	32.8	21.3	39.5
DE	Wilmington	22.4	24.8	33.1	41.8	52.2	61.6	67.1	65.9	58.2	45.7	37.0	27.6	44.8
DC	Washington	26.8	29.1	37.7	46.4	56.6	66.5	71.4	70.0	62.5	50.3	41.1	31.7	49.2
FL	Jacksonville	40.5	43.3	49.2	54.9	62.1	69.1	71.9	71.8	69.0	59.3	50.2	43.4	57.1
	Miami	59.2	60.4	64.2	67.8	72.1	75.1	76.2	76.7	75.9	72.1	66.7	61.5	69.0
GA	Atlanta	31.5	34.5	42.5	50.2	58.7	66.2	69.5	69.0	63.5	51.9	42.8	35.0	51.3
HI	Honolulu	65.6	65.4	67.2	68.7	70.3	72.2	73.5	74.2	73.5	72.3	70.3	67.0	70.0
ID	Boise	21.6	27.5	31.9	36.7	43.9	52.1	57.7	56.8	48.2	39.0	31.1	22.5	39.1
IL	Chicago	12.9	17.2	28.5	38.6	47.7	57.5	62.6	61.6	53.9	42.2	31.6	19.1	39.5
	Peoria	13.2	17.7	29.8	40.8	50.9	60.7	65.4	63.1	55.2	43.1	32.5	19.3	41.0
IN	Indianapolis	17.2	20.9	31.9	41.5	51.7	61.0	65.2	62.8	55.6	43.5	34.1	23.2	42.4
IA	Des Moines	10.7	15.6	27.6	40.0	51.5	61.2	66.5	63.6	54.5	42.7	29.9	16.1	40.0
KS	Wichita	19.2	23.7	33.6	44.5	54.3	64.6	69.9	67.9	59.2	46.6	33.9	23.0	45.0

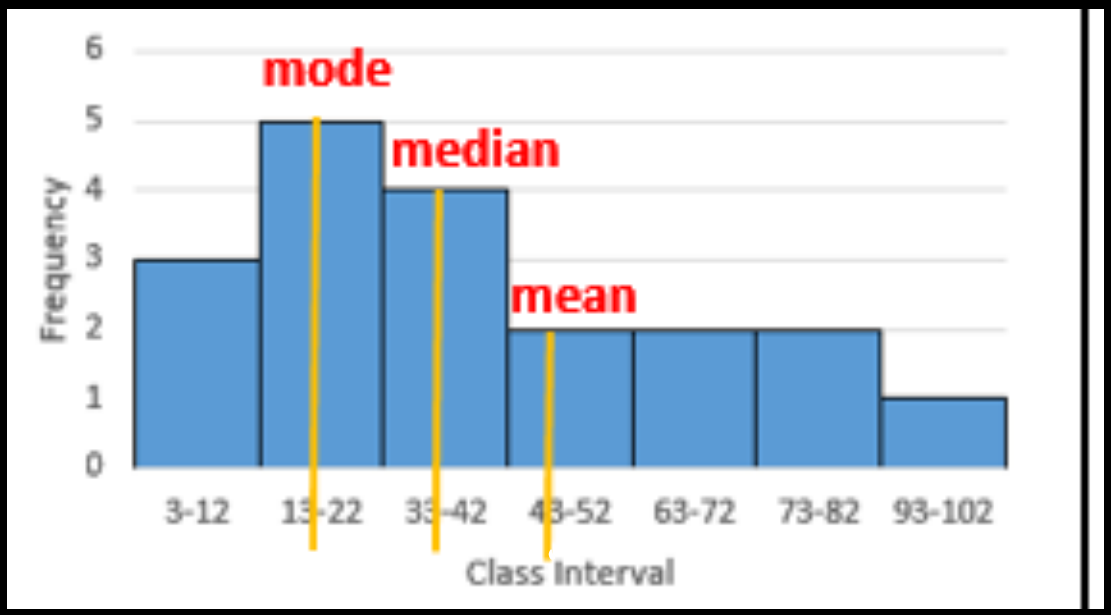
In questi casi vale la pena sintetizzare i dati tramite statistiche
⟶ grandezze calcolate a partire dai dati.

Un tipo importante di statistica è quello che mira ad identificare il valore "tipico" del campione, attorno al quale si "accumulano" i dati.

⚠ non esiste una definizione unica di tale statistica ma dipende dalla nostra interpretazione di "tipico" e "accumulano".

Infatti esistono (almeno) 3 tali statistiche:

Esempio



Teniamo quindi, per un campione $\{x_1, \dots x_n\}$ di n elementi

Def1 Si dice <u>media campionaria</u> (denotata con $\bar{x}$)

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{j=1}^{n} x_j$$

Richiamo (Sommatoria): $\sum_{j=1}^{10} x_j = x_1 + x_2 + \cdots + x_{10}$

Esempio: $\sum_{j=0}^{5} 2^j = 2^0 + 2^1 + 2^2 + 2^3 + 2^4 + 2^5 = 63$

<u>Note utili</u>: se $a, b \in \mathbb{R}$

- $\sum_{j=1}^{10} a \cdot x_j = (ax_1 + ax_2 + \cdots + a \cdot x_{10}) = a \cdot (x_1 + x_2 \cdots + x_{10}) = a \cdot \sum_{j=1}^{10} x_j$
- $\sum_{j=1}^{10} (x_j + b) = (x_1 + b + x_2 + b + \cdots + x_{10} + b) = (x_1 + x_2 + \cdots x_{10}) + 10 \cdot b = \sum_{j=1}^{10} x_j + 10b$

Esempio (Punteggio medio al campionato di golf):

$\{284, 280, 277, 282, 279, 285, 281, 278, 277\}$

x_1

possiamo calcolare la media

$$\bar{x} = \frac{1}{9} \cdot (284 + 280 + \dots + 277) = 280.\overline{3}$$

Nota: nel caso in cui l'insieme di dati è fornito tramite le frequenze (assolute) dei suoi valori, si calcola la media campionaria pesando i valori per la loro frequenza

Esempio:

Età	15	16	17	18	19	20
freq.	2	5	11	9	14	13

$n = f_1 + f_2 + \dots f_6 = 54$

abbiamo $\bar{x} = \frac{1}{54}(\underbrace{2}_{f_1} \cdot \underbrace{15}_{v_1} + 5 \cdot 16 + 11 \cdot 17 + 9 \cdot 18 + 14 \cdot 19 + 13 \cdot 20) = 18.24$

In generale: per un insieme di dati con valori

$\{v_1, \dots, v_k\}$ e relative frequenze $\{f_1, \dots, f_k\}$

(il valore v_k ha freq. assolute f_k)

- il numero complessivo di dati è $n = f_1 + \dots + f_k = \sum_{j=1}^{k} f_j$
- la media campionaria è

$$\bar{x} = \frac{1}{n} \sum_{j=1}^{k} v_j \cdot f_j = \frac{f_1 \cdot v_1 + f_2 v_2 + \dots + f_k v_k}{n}$$

(peso → f_1, valore → v_1)

Passiamo ora alla definizione di **mediana**, cioè il valore dell'elemento "al centro della lista".

Def 2: Assegnato un insieme di dati di ampiezza n lo si ordini dal minore al maggiore. La **mediana campionaria** è data da

- il valore del dato in posizione $\frac{n+1}{2}$ se n è dispari
- la media aritmetica tra i valori dei dati in posizione $\frac{n}{2}$ e $\frac{n}{2}+1$ se n è pari.

Esempio

Età	15	16	17	18	19	20
freq.	2	5	11	9	14	13

$n = 54 \rightarrow \frac{n}{2} = 27$

$\frac{n}{2}+1 = 28$

27 esimo elemento ha valore 18

28 esimo elemento ha valore 19

$\Longrightarrow$ mediana $= \frac{18+19}{2} = 18.5$

Esempio (Topi)

Conventional Mice

1	59, 89, 91, 98
2	35, 45, 50, 56, 61, 65, 66, 80
3	43, 56, 83
4	03, 14, 28, 32

Germ-Free Mice

1	58, 92, 93, 94, 95
2	02, 12, 15, 29, 30, 37, 40, 44, 47, 59
3	01, 01, 21, 37
4	15, 34, 44, 85, 96
5	29, 37
6	24
7	07
8	00

Troviamo che le medie sono

$\begin{cases} \bar{x}_{\text{steile}} = 344.07 \\ \bar{x}_{\text{normale}} = 292.32 \end{cases}$

Per le mediane invece abbiamo

$\begin{cases} n_{\text{steile}} = 29 \\ n_{\text{normale}} = 19 \end{cases}$

$\begin{cases} \text{mediana}_{\text{steile}} = 259 \\ \text{mediana}_{\text{normale}} = 265 \end{cases}$

⚠ le medie campionarie sono molto diverse a causa di pochi valori grandi nel primo campione (outliers)! La mediana non "percepisce" questi dati.

Giungiamo infine alla terza statistica: la moda

Def 3: La moda campionaria di un insieme di dati (se esiste) è l'unico valore che ha frequenza massima

Se non esiste un unico valore con frequenza massima ciascuno di essi è detto valore modale.

Esempio

Età	15	16	17	18	19	20
freq.	2	5	11	9	14	13

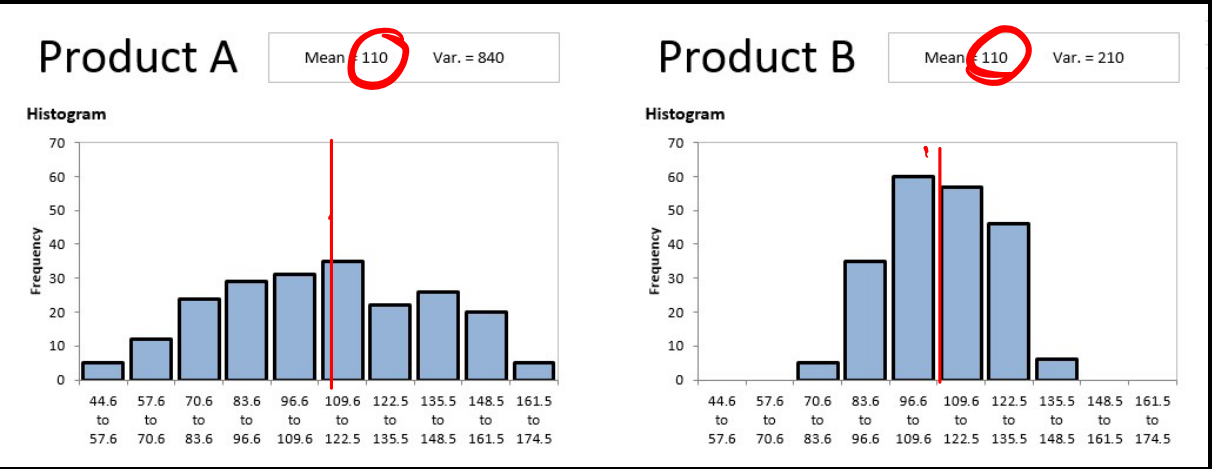
↖ 19 è la moda del dataset.

Varianza e deviazione standard

Invece di cercare il "centro" o valore "tipico" di un campione, a volte vogliamo sintetizzare quanto il campione sia "disperso" attorno a tale valore.
In altre parole vorremmo quantificare la "distanza tipica" dei punti dati dal loro centro.

Esempio

Salari per clienti di due prodotti.



Idea 1: (media delle differenze.) $\frac{1}{n}\sum_{j=1}^{n}(x_j-\bar{x})$?

$$\frac{1}{n}\sum_{j=1}^{n}(x_j-\bar{x}) = \frac{1}{n}\sum x_j - \frac{1}{n}\cdot n\cdot\bar{x} = \frac{1}{n}\sum_{j=1}^{n}x_j - \bar{x} = 0$$

$\bar{x}$ è la media campionaria

Def: Assegnato un insieme di dati $\{x_1,\dots,x_n\}$ si dice varianza campionaria la quantità

$$s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j-\bar{x})^2$$

dove $\bar{x}$ è la media campionaria.