Statistica

La statistica può essere vista come la scienza che organizza ed analizza dati numerici per fini descrittivi o per permettere di prendere delle decisioni e fare previsioni.

Statistica descrittiva: dalla mole di dati numerici a disposizione trae degli indicatori sintetici che possano riassumere le proprietà salienti dell'intera distribuzione.

Statistica inferenziale: utilizza dati statistici per pevisioni di tipo probabilistico su situazioni future *(incerte)*, su popolazioni più ampie . . .

POPOLAZIONE: *serie di dati*, che rappresenta linsieme che si vuole indagare *(reali, sperimentali, matematici)*

CAMPIONE: *serie di dati*, che rappresenta una porzione della popolazione *(campione rappresentativo)*

VARIABILI: qualitative, quantitative *(continue, discrete)*

Distribuzione di Frequenza - Esempio

Supponiamo di avere un campione di $n = 200$ famiglie, di cui rileviamo il carattere *titolo di studio del capo–famiglia*.

Questo carattere può presentare $m = 5$ differenti realizzazioni *(categorie)*.

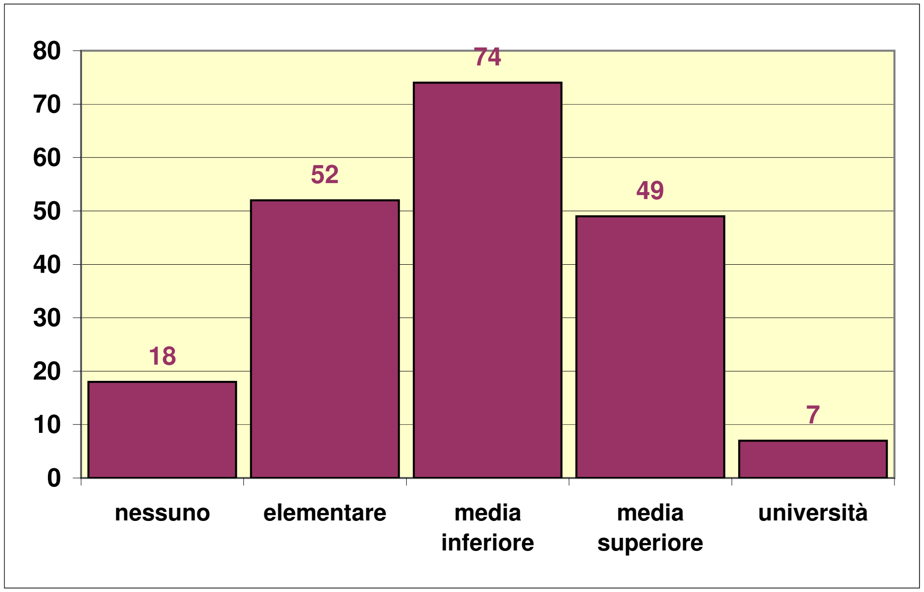
Costruiamo la tabella della *distribuzione di frequenza*:

	f_i	f_i/n	F_i	F_i/n
Laurea	18	0.090	18	0.090
Diploma scuola media superiore	52	0.260	70	0.350
Diploma scuola media inferiore	74	0.370	144	0.720
Licenza elementare	49	0.245	193	0.965
Nessun titolo	7	0.035	200	1.000
	200	1.000		

Distribuzione di Frequenza - Esempio

Rappresentiamo i dati riportati nella tabella della distribuzione delle frequenze con un *istogramma delle frequenze*

titolo di studio	fi	fi / n
nessuno	18	0,09
elementare	52	0,260
media inferiore	74	0,370
media superiore	49	0,245
università	7	0,035
	200	1,000



- ogni rettangolo rappresenta un carattere
- l'area del rettangolo e proporzionale alla frequenza di quel carattere

Distribuzione di Frequenza

dati raggruppati in classi o categorie $\quad (x_i\,,\,f_i)_{i=1...m}$

FREQUENZA ASSOLUTA f_i: è numero di *osservazioni* che ricadono in ciascuna classe. Numero totale di osservazioni $n = \sum_{i=1}^{m} f_i$

FREQUENZA RELATIVA f_i/n: è il rapporto tra la frequenza assoluta e il numero totale n di osservazioni e rappresenta la percentuale di osservazoni in ogni classe o categoria.

FREQUENZA ASSOLUTA CUMULATA F_i: $\quad F_i = \sum_{k=1}^{i} f_k$

FREQUENZA RELATIVA CUMULATA F_i/n: $\quad \frac{1}{n}\sum_{k=1}^{i} f_k$

Statistica Descrittiva

misure, indici (numerici) che descrivono le caratteristiche della distribuzione di una o più variabili in modo sintetico

- **indici di posizione o centralità :**

  valore centrale, medie algebriche, mediana, moda

  (detti anche *misure di intensità, centri . . .)*

- **indici di dispersione o variabilità :**

  intervallo di variazione, varianza, varianza stimata, deviazione standard, deviazione standard stimata

- **indici di simmetria o asimmetria :**

  . . .

Valore Centrale

VALORE CENTRALE : dato l'insieme di valori $\{x_1, x_2, \ldots x_n\}$, considera solo i due valori estremi *(non tiene conto di tutti i valori)*

$$\frac{x_{max} + x_{min}}{2}$$

$$x_{max} = \max\{x_1, x_2, \ldots x_n\} \quad \text{e} \qquad x_{min} = \min\{x_1, x_2, \ldots x_n\}$$

ESEMPIO : {3, 20, 27, 25, 30, 310}

$$\frac{x_{max} + x_{min}}{2} = \frac{310 + 3}{2} = 156.5$$

Media Aritmetica - 1

MEDIA SEMPLICE : dato l'insieme di valori $\{x_1, x_2, \ldots x_n\}$

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

MEDIA PONDERATA *(dati raggruppati)*: dato l'insieme di valori $\{x_1, x_2, \ldots x_m\}$ con le rispettive frequenza assolute $\{f_1, f_2, \ldots f_m\}$

$$\bar{x} = \frac{\sum_{i=1}^{m} f_i x_i}{\sum_{i=1}^{m} f_i} = \frac{1}{n} \cdot \sum_{i=1}^{m} f_i x_i = \frac{f_1 x_1 + f_2 x_2 + \ldots + f_n x_n}{n}$$

Media Aritmetica - 2

- la media può non appartenere all'insieme dei dati
- insiemi di dati diversi possono avere la stessa media
- utilizza tutti i dati
- centro di gravità dei dati
- riduce l'effetto dei dati estremi *(outlier)*

PROPRIETÀ:

- se applico una trasformazione lineare ai dati
$$y_i = a\,x_i + b \Rightarrow \bar{y} = a\,\bar{x} + b$$
- la somma degli scarti dalla media è nulla $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$
- la somma dei quadrati degli scarti dalla media è minima
$\sum_{i=1}^{n}(x_i - x)^2$ *assume il valore minimo per* $x = \bar{x}$

Media Aritmetica - 3

- La somma degli scarti dalla media è nulla.

$$\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} = \sum_{i=1}^{n} x_i - n\,\bar{x} = \sum_{i=1}^{n} x_i - n \cdot \frac{1}{n} \cdot \sum_{i=1}^{n} x_i = 0$$

- La somma dei quadrati degli scarti dalla media è minima.

  Nel caso di due soli dati $\{x_1, x_2\}$ tale somma è $F(x) = (x_1 - x)^2 + (x_2 - x)^2$

  Determino i punti di massimo e di minimo relativo

$$F'(x) = -2\,(x_1 - x) - 2\,(x_2 - x) = 0 \;\Rightarrow\; x = \frac{1}{2}(x_1 + x_2)$$

  lo studio del segno di $F'(x)$ mi dice che la somma dei quadrati degli scarti ha un punto di minimo in

$$\bar{x} = \frac{1}{2}(x_1 + x_2) \quad \textit{media aritmetica}$$

Media Aritmetica - Esercizi

ESERCIZIO 1 : dato l'insieme di valori $\{12, 25, 37, 41, 0, 53\}$

$$\bar{x} = \frac{1}{6} \cdot (12 + 25 + 37 + 41 + 0 + 53) = 28$$

ESERCIZIO 2 :dato l'insieme di valori $\{28, 28, 28, 28, 28, 28\}$

$$\bar{x} = \frac{1}{6} \cdot (28 + 28 + 28 + 28 + 28 + 28) = 28$$

ESERCIZIO 3 *(dati raggruppati)* : In un campione di 200 persone si sa che 20 pesano 50 kg, 30 pesano 55kg, 50 pesano 60kg, 70 pesano 65kg, 20 pesano 75 Kg e 10 pesano 80kg. Calcolare il peso medio.

$$\bar{x} = \frac{20 \cdot 50 + 30 \cdot 55 + 50 \cdot 60 + 70 \cdot 65 + 20 \cdot 75 + 10 \cdot 80}{200} =$$

$$= \frac{20}{200} \cdot 50 + \frac{30}{200} \cdot 55 + \frac{50}{200} \cdot 60 + \frac{70}{200} \cdot 65 + \frac{20}{200} \cdot 75 + \frac{10}{200} \cdot 80 = 62.5\,\text{Kg}$$

Media Geometrica

MEDIA SEMPLICE : dato l'insieme di valori $\{x_1, x_2, \ldots x_n\}$

$$x_g = \sqrt[n]{\prod_{i=1}^{n} x_i} = \sqrt[n]{x_1 \cdot x_2 \cdots x_n} \quad \Rightarrow \quad \log x_g = \frac{1}{n} \cdot \sum_{i=1}^{n} \log x_i$$

MEDIA PONDERATA : dato l'insieme di valori $\{x_1, x_2, \ldots x_m\}$ con le rispettive frequenza assolute $\{f_1, f_2, \ldots f_m\}$

$$\log x_g = \frac{1}{n} \cdot \sum_{i=}^{n} f_i \log x_i$$

Mediana

Dato l'insieme di valori ordinati $x_1 \leq x_2 \leq \cdots \leq x_{n-1} \leq x_n$ si chiama *mediana* (*valore mediano*) il valore M_e che occupa la posizione centrale

- se n è dispari c'è un unico termine mediano di posto $\frac{n+1}{2}$

$$M_e = x_{\frac{n+1}{2}}$$

- se n è pari ci sono due termini mediani di posti $\frac{n}{2}$ e $\frac{n}{2} + 1$

$$M_e = \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1})$$

Utilizza tutti i valori ma si basa soltanto sull'ordinamento degli stessi.

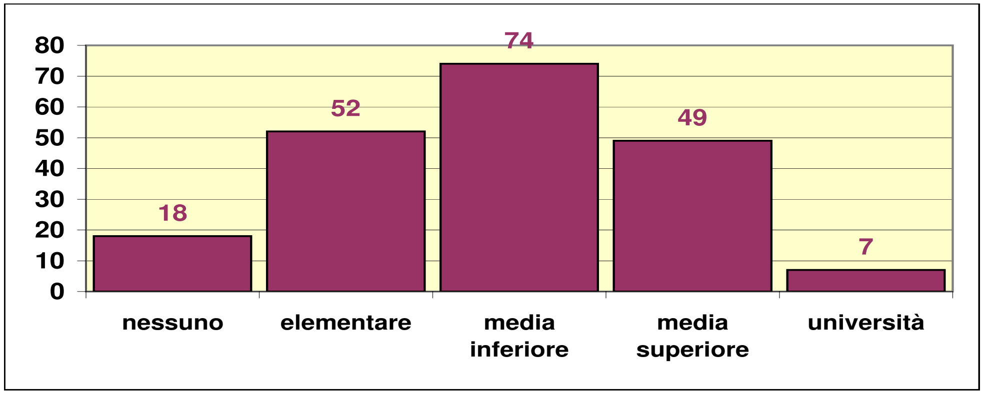
ESEMPIO 1: $\{0, 13, 25, 81, 503\} \Rightarrow M_e = 25$

ESEMPIO 2: $\{1, 2, 81, 93, 327, 503\} \Rightarrow M_e = 87$

Moda

MODA : valore (o classe) al quale associata la frequenza più alta

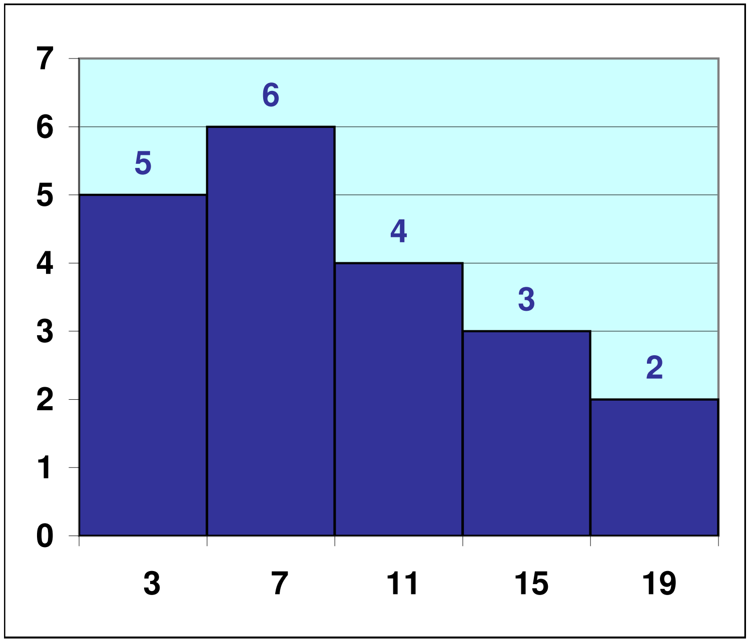
titolo di studio	fi	fi / n
nessuno	18	0,09
elementare	52	0,260
media inferiore	74	0,370
media superiore	49	0,245
università	7	0,035
	200	1,000



(si può applicare anche a dati qualitativi espressi su scala nominale)

Esempio: Media, Mediana, Moda

classe	ri	fi	fi / n
1-5	3	5	0,25
5-9	7	6	0,300
9-13	11	4	0,200
13-17	15	3	0,150
17-21	19	2	0,100
		20	1,000



- MEDIA: si calcola come media ponderata $\bar{x} = 9.2$
- MEDIANA: $M_e = 7$ è la media del decimo e dell'undicesimo termine che hanno entrambi valore 7 .
- MODA: è la classe $5 - 9$ o il suo rappresentante $r_2 = 7$, corrispondenti a $f_2 = 6$
- *moda* $<$ *mediana* $<$ *media* distribuzione obliqua a destra