CAPITOLO 3

STATISTICA DESCRITTIVA

1. GENERALITA' SUL METODO STATISTICO

La **Statistica** è la scienza che si occupa di come si raccolgono le informazioni su un fenomeno collettivo e di come si organizzano, si analizzano e si interpretano, utilizzando metodi e strumenti matematici. La statistica si divide in:

- **statistica descrittiva** con lo scopo di raccogliere ed elaborare i dati per descrivere fenomeni collettivi o di massa;
- **statistica induttiva**, o **inferenza statistica**, che si occupa dei metodi che permettono di stimare le caratteristiche di un fenomeno collettivo partendo dall'analisi delle caratteristiche di un campione.

Il metodo statistico è basato sull'analisi di fenomeni collettivi, allo scopo di ricavare le leggi che li governano, o almeno di ricavare eventuali regolarità in modo da poter trarre previsioni sul comportamento futuro di tale fenomeno.

Le fasi di un'indagine statistica sono essenzialmente due:

- **rilevazione dei dati**, che consiste nel raccogliere i dati relativi al fenomeno in esame;
- **organizzazione dei dati,** che consiste nel raggruppare i dati in tabelle, affinché possano essere letti, analizzati ed interpretati agevolmente;
- **elaborazione ed interpretazione dei dati**.

Si considera un insieme omogeneo (di persone, animali, automobili, ecc.) detto **universo** o **popolazione** ed in esso si esaminano quanti elementi presentano il carattere in esame. Tali caratteri si suddividono in:

- **caratteri quantitativi**: espressi da numeri risultanti da misurazioni (es. altezza degli studenti di una classe) o da enumerazioni (es. numero dei vani degli appartamenti di un comune). Essi si suddividono ulteriormente in :
  - **continui**: se sono espressi da numeri reali e possono assumere tutti i valori di un intervallo (es. altezze, pesi, aree, ecc.), in tal caso i valori si raggruppano in classi, che possono essere di ampiezze uguali o differenti;
  - **discreti**: se sono espressi da numeri naturali (es. numero di componenti delle famiglie, numero vani di un'abitazione, ecc.).
- **caratteri qualitativi**: espressi da attributi o espressioni verbali (es. nazionalità della popolazione europea, tipo di animali allevati nelle diverse regioni, colore dei capelli, titolo di studio, ecc.). Essi si suddividono ulteriormente in:

- **ordinati**: se fra le varie modalità si può stabilire una relazione d'ordine (es. titolo di studio);
- **sconnessi**: se non si può stabilire un ordine (es. nazionalità).

I dati rilevati vengono riportati su tabelle che possono essere a semplice entrata, a doppia entrata o composte a seconda che vengano rilevati un solo carattere, due caratteri collegati fra loro o più caratteri.

2. DISTRIBUZIONI STATISTICHE

In una tabella a semplice entrata costituita da due colonne, la prima riporta le varie modalità del carattere qualitativo o quantitativo da esaminare, la seconda le frequenze o i valori rilevati: si parla di *distribuzione statistica*. Se il carattere è qualitativo si parla di **serie statistica**, in caso di carattere quantitativo si parla di **seriazione statistica**. Nel caso in cui i dati da esaminare siano frequenze, si parla di **distribuzione di frequenze**. Per **frequenza assoluta** si intende il numero di unità statistiche aventi una determinata caratteristica; per **frequenza relativa** si intende il rapporto fra la frequenza assoluta e la somma di tutte le frequenze. A volte le frequenze possono essere espresse mediante percentuali, in quanto sono di più facile lettura. A volte è utile raggruppare i dati in **classi**, determinando la frequenza di ogni classe; in tal caso è necessario calcolare il **valore centrale** di ciascuna classe come semisomma degli estremi della classe stessa. Si definiscono infine **frequenze relative cumulate** la somma della frequenza relativa considerata con le frequenze dei valori precedenti.

Una **variabile statistica** è definita dall'insieme dei valori osservati di un carattere quantitativo e dalle frequenze ad essi associate. Una **mutabile statistica** è definita dall'insieme delle modalità osservate di un carattere qualitativo e dalle frequenze ad esse associate.

Se nell'analisi di un campione si rileva più di un carattere (ad esempio altezza e peso), si formano **tabelle composte**.

ESEMPI

1. Titolo di studio su un campione di 30 persone.

TITOLO DI STUDIO	N° PERSONE
Senza titolo	1
Licenza elementare	3
Scuola media inferiore	5
Scuola media superiore	12
Laurea	9
TOTALE	30

Si tratta di una serie statistica (il carattere è qualitativo) e di una distribuzione di frequenze (in questo caso assolute) data dal numero di persone.

2. Famiglie di un comune divise per numero di componenti.

N° COMPONENTI	N° FAMIGLIE
1	3.204
2	350
3	72
4	58
5	21
6	5
7 e più	2
TOTALE	3.712

Si tratta di una seriazione rispetto ad un carattere quantitativo discreto e di una distribuzione di frequenze (in questo caso assolute).

3. Studenti di una classe divisi per altezza.

CLASSI DI ALTEZZA (IN CM)	N° STUDENTI
fino a 150	1
150 —┤ 160	4
160 —┤ 165	6
165 —┤ 170	8
170 —┤ 180	3
oltre 180	1
TOTALE	23

Si tratta di una seriazione rispetto ad un carattere quantitativo continuo e di una distribuzione di frequenze (in questo caso assolute). In questa tabella ad ogni intervallo appartiene l'estremo destro e non il sinistro (si potrebbe effettuare anche l'altra scelta).

4. Diplomati di un istituto superiore.

ANNI	N° STUDENTI
2007	105
2008	102
2009	112
2010	120
2011	104
2012	122
TOTALE	665

Si tratta di una **serie storica** in quanto i dati sono riferiti ad intervalli di tempo (anni, mesi, ecc.)

5. In riferimento all'esempio 1, si calcolini le frequenze relative, le frequenze relative cumulate e le frequenze percentuali.

TITOLO DI STUDIO	N° PERSONE (FREQUENZE ASSOLUTE)	FREQUENZE RELATIVE	FREQUENZE RELATIVE CUMULATE	FREQUENZE PERCENTUALI
Senza titolo	1	0.0333	0.0333	3.33%
Licenza elementare	3	0.1	0.1333	10%
Scuola media inferiore	5	0.1667	0.3	16.67%
Scuola media superiore	12	0.4	0.7	40%
Laurea	9	0.3	1	30%
TOTALE	30	1		100%

3. RAPPRESENTAZIONI GRAFICHE

I dati raccolti in tabelle si possono rappresentare graficamente. Le rappresentazioni grafiche sono molto più espressive di una tabella di valori ed offrono molti vantaggi poiché descrivono il fenomeno in forma visiva, permettono di esaminarne l'andamento in modo globale e di confrontare caratteri diversi dello stesso fenomeno e le sue variazioni nel tempo. E' molto importante la scelta della unità di misura che deve essere fatta di volta in volta tenendo conto sia dei valori minimi sia dei valori massimi che si devono rappresentare. Considereremo di seguito le più importanti rappresentazioni grafiche.

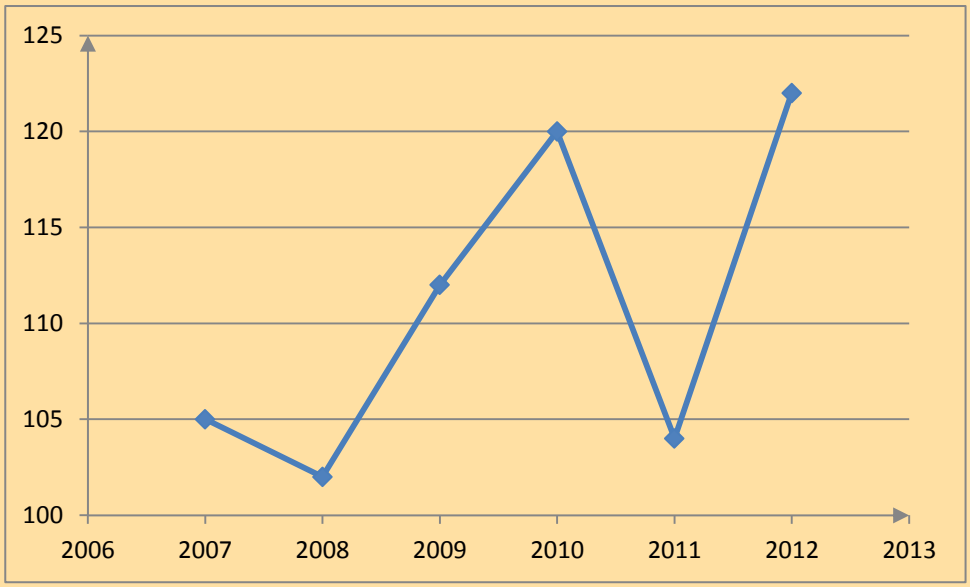
DIAGRAMMI CARTESIANI

Sono utilizzati principalmente per rappresentare serie storiche e seriazioni discrete. Le unità di misura sui due assi sono generalmente diverse. Si è soliti collegare con una spezzata i punti che rappresentano le coppie di valori. Nel caso di serie storiche la spezzata mette in evidenza l'evoluzione del fenomeno nel tempo.

ESEMPIO

Si consideri l'esempio 4 precedente e si rappresentino graficamente i dati.

ANNI	N° STUDENTI
2007	105
2008	102
2009	112
2010	120
2011	104
2012	122
TOTALE	665



ISTOGRAMMI

Si usano gli istogrammi per rappresentare seriazioni continue con dati raggruppati in classi. Fissato un sistema di assi cartesiani ortogonali, sull'asse delle ascisse si riportano tanti intervalli consecutivi quante sono le classi; su questi intervalli si costruiscono dei rettangoli **le cui aree sono proporzionali alle frequenze**. Per determinare le altezze dei rettangoli si distinguono due casi:

1. se le classi hanno tutte la stessa ampiezza, le altezze dei rettangoli sono proporzionali alle frequenze;
2. se le classi hanno ampiezze diverse, le altezze dei rettangoli si ottengono dividendo ogni frequenza per l'ampiezza della relativa base; questo rapporto è detto **densità di frequenza**.

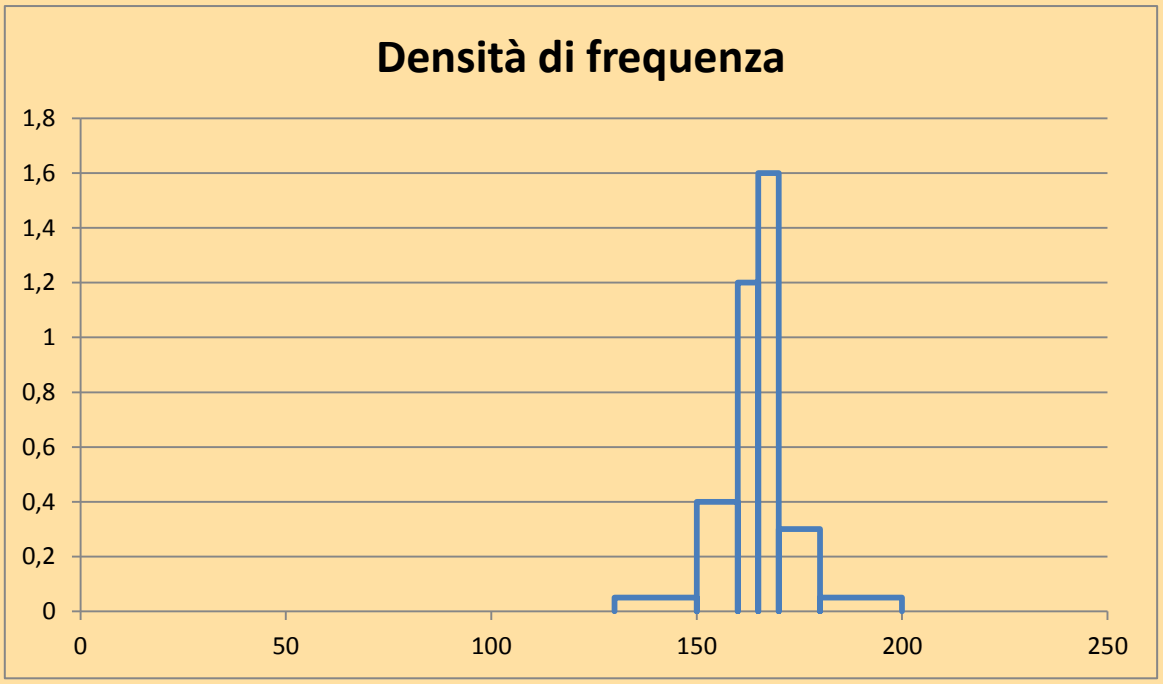
In caso di classi aperte (la prima o l'ultima), o si chiudono scegliendo un valore estremo logico, o si trascurano se la frequenza relativa è relativamente piccola.

ESEMPIO

Si consideri l'esempio 3 precedente e si rappresentino graficamente i dati.

CLASSI DI ALTEZZA (IN CM)	N° STUDENTI	DENSITÀ DI FREQUENZA
fino a 150	1	0.05
150 ⊣ 160	4	0.4
160 ⊣ 165	6	1.2
165 ⊣ 170	8	1.6
170 ⊣ 180	3	0.3
oltre 180	1	0.05
TOTALE	23	

La classe aperta "oltre 180" viene chiusa a 200; la classe aperta "fino a 150" viene chiusa a 130.



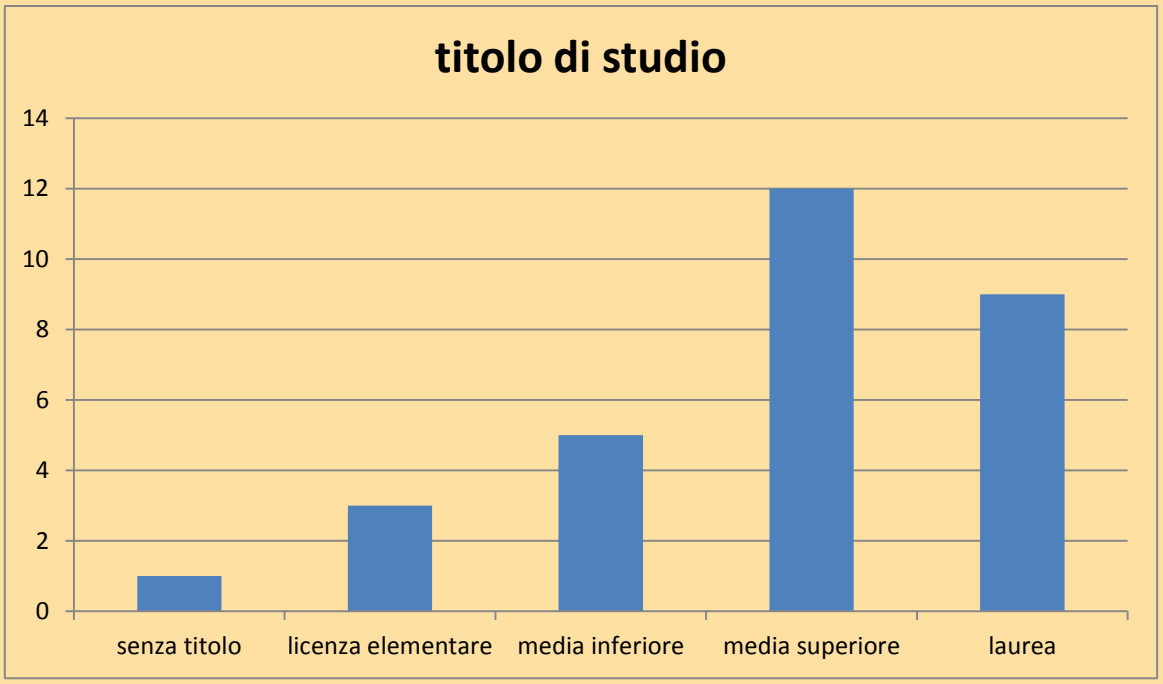
ORTOGRAMMI (O DIAGRAMMI A CANNE D'ORGANO)

Si usano gli ortogrammi per rappresentare serie statistiche, dati relativi a caratteri di tipo qualitativo o quantitativo discreto. Sono rappresentati da rettangoli di basi uguali ed altezze proporzionali alla frequenza del fenomeno da analizzare.

ESEMPIO

Si consideri l'esempio 1 precedente e si rappresentino graficamente i dati.

TITOLO DI STUDIO	N° PERSONE
Senza titolo	1
Licenza elementare	3
Scuola media inferiore	5
Scuola media superiore	12
Laurea	9
TOTALE	30



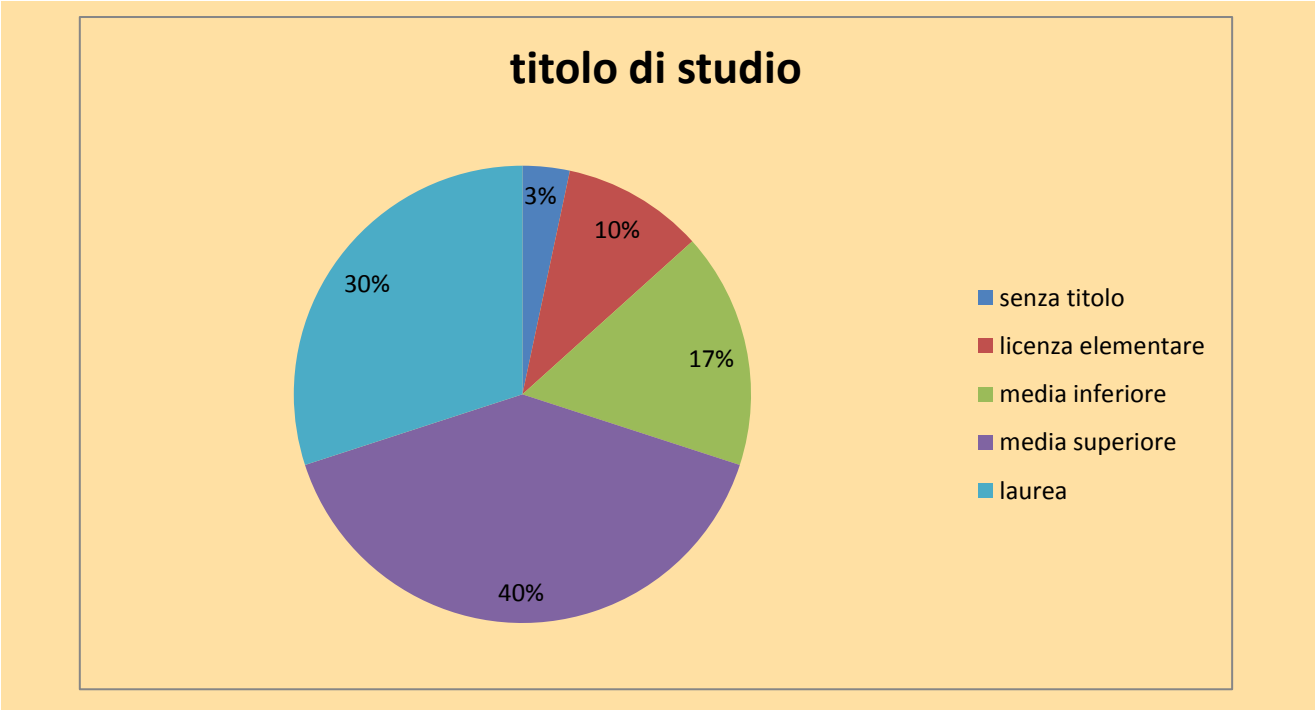
DIAGRAMMI A SETTORI CIRCOLARI O AEROGRAMMI

Tali diagrammi, detti comunemente "a torta", si utilizzano per rappresentare distribuzione di frequenze o di intensità, in cui la totalità del fenomeno viene ripartita in settori circolari i cui angoli al centro sono proporzionali alle intensità del fenomeno in esame.

ESEMPIO

Si consideri ancora l'esempio 1 precedente e si rappresentino graficamente i dati, utilizzando un diagramma "a torta".

TITOLO DI STUDIO	N° PERSONE
Senza titolo	1
Licenza elementare	3
Scuola media inferiore	5
Scuola media superiore	12
Laurea	9
TOTALE	30



CARTOGRAMMI

Servono per rappresentare serie territoriali. Si realizzano riportando su una carta geografica, mediante simboli o colorazioni diverse, le frequenze di una rilevazione nelle varie parti in cui è diviso un territorio.

IDEOGRAMMI

Sono rappresentazioni mediante figure di grandezza diversa, con aree proporzionali all'intensità del fenomeno da esaminare.

4. MEDIE STATISTICHE

Una volta che si è fatta la rilevazione dei dati di un certo fenomeno, è necessario riassumere la distribuzione tramite valori che la caratterizzino in modo da poterla confrontare con la distribuzione di fenomeni analoghi osservati in tempi e luoghi differenti. Uno di questi valori è dato dal valore medio.
In statistica si distinguono due tipi di medie:

- **medie di calcolo (o ferme)**, che utilizzano tutti i valori della distribuzione, ad esempio la media aritmetica, geometrica, quadratica, armonica;
- **medie di posizione (o lasche)**, che utilizzano solo alcuni valori, ad esempio la moda e la mediana.

La scelta del tipo di media da usare dipende dal problema che bisogna risolvere.

MEDIA ARITMETICA

Si definisce **media aritmetica** M di più numeri quel valore che, sostituito ai dati, lascia invariata la loro somma. Indicati con $x_1, x_2, \dots, x_n$ i numeri dati, si ottiene la **media aritmetica semplice**:

$$M = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Se i valori $x_1, x_2, \ldots, x_n$ hanno frequenze diverse fra loro, ad esempio frequenze rispettivamente pari a $y_1, y_2, \ldots, y_n$, si ricava la **media aritmetica ponderata**, poiché le frequenze vengono anche dette **pesi**:

$$M = \frac{x_1 y_1 + x_2 y_2 + \ldots + x_n y_n}{y_1 + y_2 + \ldots + y_n} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} y_i}$$

OSSERVAZIONI:

- Nel caso di serie statistiche si calcola la media aritmetica semplice; nel caso di seriazioni statistiche si applica la media aritmetica ponderata.
- In caso di seriazioni con valori raggruppati in classi, è necessario calcolare il valore centrale di ogni classe come semisomma dei valori estremi dell'intervallo, e poi procedere al calcolo della media aritmetica ponderata.
- Nel caso in cui i dati del problema siano rappresentate da numeri naturali (ad esempio persone, stanze, …) e la media non sia un numero intero, è necessario arrotondarla al numero intero più prossimo.
- E' la media più utilizzata, ad esempio quando si vuole determinare una media delle spese, dei consumi, dei redditi, dei voti, delle temperature, ecc.

ESEMPI

1. Si consideri la serie storica seguente (esempio 4 precedente), e si calcoli la media aritmetica di studenti diplomati.

ANNI	N° STUDENTI
2007	105
2008	102
2009	112
2010	120
2011	104
2012	122
TOTALE	665

Il valor medio sarà dato dalla media aritmetica semplice:

$$M = \frac{105 + 102 + 112 + 120 + 104 + 122}{6} = \frac{665}{6} = 110{,}83 \approx 111$$

In media dal 2007 al 2012 si sono diplomati 111 studenti.

2. Si consideri la seriazione seguente (esempio 2 precedente), e si calcoli la media aritmetica.

N° COMPONENTI	N° FAMIGLIE
1	3.204
2	350
3	72
4	58
5	21
6	5
7 e più	2
TOTALE	3.712

Il valor medio sarà dato dalla media aritmetica ponderata; quindi:

N° COMPONENTI (X)	N° FAMIGLIE (Y)	XY
1	3.204	3.204
2	350	700
3	72	216
4	58	232
5	21	105
6	5	30
7	2	14
TOTALE	3.712	4.501

$$M = \frac{4.501}{3.712} = 1{,}21$$

Quindi, nel campione esaminato, le famiglie sono composte in media da una persona.

MEDIA GEOMETRICA

Può essere calcolata solo nel caso in cui i valori siano tutti positivi. Si definisce **media geometrica** G di più numeri positivi quel valore che, sostituito ai dati, lascia invariato il loro prodotto. Indicati con $x_1, x_2, \dots, x_n$ i numeri dati, si ottiene la **media geometrica semplice**:

$$G = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$$

Se i valori $x_1, x_2, \dots, x_n$ hanno frequenze diverse fra loro, ad esempio frequenze rispettivamente pari a $y_1, y_2, \dots, y_n$, si ricava la **media geometrica ponderata:**

$$G = \sqrt[N]{x_1^{y_1} \cdot x_2^{y_2} \cdot \ldots \cdot x_n^{y_n}} \qquad con\ N = \sum_{i=1}^{n} y_i$$

OSSERVAZIONI:

- In caso di seriazioni con valori raggruppati in classi, è necessario calcolare il valore centrale di ogni classe come semisomma dei valori estremi dell'intervallo, e poi procedere al calcolo della media geometrica ponderata.
- Si deve calcolare la media geometrica quando si vuole calcolare il tasso medio di più tassi applicati ad uno stesso capitale in capitalizzazione composta oppure per calcolare il tasso di incremento medio o di decremento dei prezzi, oppure il tasso di accrescimento di una popolazione, o ancora per determinare una media nei cambi monetari.

ESEMPIO

Il capitale di € 10.000 è stato impiegato ad interesse composto per 15 anni. Il tasso annuo è stato del 2% per i primi 4 anni, del 2,5% per i successivi 6 anni e del 3,25% per gli ultimi 5 anni. Calcolare il tasso medio di impiego.

Il montante è dato da:

dopo 4 anni: $M_1 = 10.000 \cdot (1 + 0.02)^4 = 10.824{,}32$

dopo altri 6 anni: $M_2 = 10.824{,}32 \cdot (1 + 0.025)^6 = 12.552{,}89$

dopo altri 5 anni: $M_3 = 12.552{,}89 \cdot (1 + 0.0325)^5 = 14.729{,}71$

Quindi $M = 10.000 \cdot (1 + i)^{15} = 14.729{,}71$

Cioè: $10.000 \cdot (1 + i)^{15} = 10.000 \cdot 1.02^4 \cdot 1.025^6 \cdot 1.0325^5$

$$(1 + i)^{15} = 1.02^4 \cdot 1.025^6 \cdot 1.0325^5 \quad \Rightarrow \quad 1 + i = \sqrt[15]{1.02^4 \cdot 1.025^6 \cdot 1.0325^5}$$

$$i = \sqrt[15]{1.02^4 \cdot 1.025^6 \cdot 1.0325^5} - 1 \cong 0.0262$$

Quindi il tasso annuo medio di investimento è stato circa del 2,62%.

MEDIA QUADRATICA

Si definisce **media quadratica** Q di più numeri quel valore che, sostituito ai dati, lascia invariata la somma dei loro quadrati. Indicati con $x_1, x_2, \dots, x_n$ i numeri dati, si ottiene la **media quadratica semplice**:

$$Q = \sqrt{\frac{x_1^2 + x_2^2 + \dots + x_n^2}{n}}$$

Se i valori $x_1, x_2, \dots, x_n$ hanno frequenze diverse fra loro, ad esempio frequenze rispettivamente pari a $y_1, y_2, \dots, y_n$, si ricava la **media quadratica ponderata:**

$$Q = \sqrt{\frac{x_1^2 \cdot y_1 + x_2^2 \cdot y_2 + \dots + x_n^2 \cdot y_n}{N}} \qquad con \;\; N = \sum_{i=1}^{n} y_i$$

OSSERVAZIONI:

- La media quadratica, semplice o ponderata, è la radice quadrata della media aritmetica, semplice o ponderata, dei quadrati dei valori dei dati.
- E' utilizzata per mettere in evidenza l'esistenza di valori che si discostano molto dai valori centrali, in quanto è la media più influenzata dalla presenza di valori molto piccoli o molto grandi della distribuzione.
- Un'applicazione di tale media è lo scarto quadratico medio (paragrafo seguente).

ESEMPIO

Da una rilevazione degli incidenti sul lavoro in 60 aziende, si sono ricavati i seguenti risultati:

N° incidenti	N° aziende
10	40
15	12
50	8

Calcolare il numero medio di incidenti per azienda con la media aritmetica e quadratica e verificare che quest'ultima è maggiore di quella aritmetica.

$$M = \frac{10 \cdot 40 + 15 \cdot 12 + 50 \cdot 8}{60} = 16{,}3$$

$$Q = \sqrt{\frac{10^2 \cdot 40 + 15^2 \cdot 12 + 50^2 \cdot 8}{60}} = 21{,}1$$

Si osserva che le due medie differiscono notevolmente.

MEDIA ARMONICA

Può essere calcolata solo nel caso in cui i valori siano tutti positivi. Si definisce **media armonica** A di più numeri positivi quel valore che, sostituito ai dati, lascia invariata la somma dei loro reciproci. Indicati con $x_1, x_2, \dots, x_n$ i numeri dati, si ottiene la **media armonica semplice**:

$$A = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}}$$

Se i valori $x_1, x_2, \dots, x_n$ hanno frequenze diverse fra loro, ad esempio frequenze rispettivamente pari a $y_1, y_2, \dots, y_n$, si ricava la **media armonica ponderata:**

$$A = \frac{N}{\frac{y_1}{x_1} + \frac{y_2}{x_2} + \dots + \frac{y_n}{x_n}} \qquad con\ N = \sum_{i=1}^{n} y_i$$

OSSERVAZIONI:

- La media armonica, semplice o ponderata, è il reciproco della media aritmetica, semplice o ponderata, dei reciproci dei valori dei dati.

- E' utilizzata per determinare il potere di acquisto medio di una moneta (che è il reciproco del prezzo di una merce) come reciproco della media armonica dei prezzi.
- Si applica anche per determinare la velocità media come media armonica delle velocità (infatti il reciproco di una velocità è il tempo necessario per percorrere l'unità di spazio).

ESEMPI

1. Una merce è stata venduta nel corso di 5 anni ai seguenti prezzi unitari: € 16, 18, 20, 24, 28. Calcolare il potere medio di acquisto riferito ad un importo di € 100.

Si calcola prima il prezzo medio come media armonica dei prezzi:

$$A = \frac{5}{\frac{1}{16}+\frac{1}{18}+\frac{1}{20}+\frac{1}{24}+\frac{1}{28}} = 20{,}37$$

Quindi con € 100 si potrebbero acquistare in media $\frac{100}{20{,}37} \cong 4{,}91$ unità di merce, cioè il potere di acquisto medio per un importo di € 100 è di circa 4,91 unità di merce.

2. Un automobilista percorre 60 km alla velocità di 50 km/h, successivamente percorre 80 km alla velocità di 100 km/h ed infine 42 km alla velocità di 60 km/h. Determinare la velocità media.

Dalla fisica si sa che per velocità media si intende il rapporto tra lo spazio percorso ed il tempo impiegato a percorrerlo. Quindi per ogni tratto il tempo impiegato è:

$$t_1 = \frac{60}{50} = 1{,}2\ h \qquad t_2 = \frac{80}{100} = 0{,}8\ h \qquad t_3 = \frac{42}{60} = 0{,}7\ h$$

Il tempo complessivo è perciò di $1{,}2 + 0{,}8 + 0{,}7 = 2{,}7$ h.

Lo spazio complessivo è di $60 + 80 + 42 = 182$ km

Quindi la velocità media è:

$$v_m = \frac{182}{2{,}7} = 67{,}41\ \ km/h$$

Si sarebbe ottenuto lo stesso risultato calcolando la media armonica ponderata delle velocità con pesi uguali agli spazi percorsi:

$$v_m = A = \frac{60+80+42}{\frac{60}{50}+\frac{80}{100}+\frac{42}{60}} = \frac{182}{2{,}7} = 67{,}41\ \ km/h$$

OSSERVAZIONE

Fra le quattro medie di calcolo sussiste la relazione (di cui si tralascia la dimostrazione):

$$A \leq G \leq M \leq Q$$

Vale il segno uguale solo nel caso in cui tutti i dati siano uguali fra loro e quindi uguali a qualsiasi media.

MODA O VALORE NORMALE

Si definisce **moda (valore normale, valore modale o norma)** di una distribuzione di frequenze, e si indica con **Mo**, la modalità o il valore della variabile cui corrisponde la massima frequenza. Nel caso in cui i dati siano raggruppati in classi, si chiamerà **classe modale** la classe con densità di frequenza maggiore; se l'ampiezza delle classi è costante essa coinciderà con la classe con maggiore frequenza. Esistono variabili statistiche con più di un valore modale e prendono il nome di **distribuzioni plurimodali**.

OSSERVAZIONE:

- E' utilizzata quando è importante conoscere il valore che si presenta con maggiore probabilità nella distribuzione.

ESEMPI

1. Nell'esempio n° 1 del paragrafo 2 la moda è "scuola media superiore" cui corrisponde la massima frequenza.
2. Nell'esempio n° 2 del paragrafo 2 la moda è 1 componente poiché è il valore cui corrisponde la massima frequenza.
3. Nell'esempio n° 3 del paragrafo 2 la classe modale è 165 ⊣ 170 cui corrisponde la massima densità di frequenza come si vede dall'istogramma corrispondente.

MEDIANA

La mediana **Me** rappresenta il valore centrale di una distribuzione quando i dati sono ordinati, cioè essa è il valore che bipartisce la successione.

a. Per le serie statistiche: ordinati i valori, se il numero dei termini è dispari, la mediana è proprio il valore centrale; se il numero dei termini è pari, si assume come **mediana** la semisomma dei due termini centrali.
b. Per le distribuzioni di frequenze con valori discreti: i dati sono generalmente già ordinati, quindi:
   - si calcolano le frequenze assolute cumulate (che si ottengono associando ad ogni valore la somma della rispettiva frequenza con tutte quelle che la precedono);
   - si indica con N la somma delle frequenze;

   la **mediana** è quel valore che corrisponde alla frequenza cumulata $\frac{N}{2}$ se N è pari, alla frequenza cumulata $\frac{N+1}{2}$ se N è dispari.
c. Per le distribuzioni di frequenze con dati raggruppati in classi: si determina la classe mediana mediante le frequenze assolute cumulate e si procede come nel caso b). Per ottenere esattamente il valore mediano, si esegue un'interpolazione lineare fra i due valori estremi della classe in cui cade la

mediana. La mediana non è influenzata dai valori estremi della distribuzione, quindi anche se le classi estreme non fossero chiuse, non è necessario chiuderle. Se la distribuzione fosse molto asimmetrica, il valore mediano è più appropriato della media aritmetica per esprimere un valore sintetico della distribuzione.

ESEMPI

1. Nell'esempio n° 1 del paragrafo 2, mettendo in ordine non decrescente i dati si ottiene:

1 3 **5** 9 12

Essendo il numero di termini dispari, Me = 5.

2. Nell'esempio n° 2 del paragrafo 2, per calcolare la mediana è necessario ampliare la tabella:

N° COMPONENTI	N° FAMIGLIE (FREQUENZA ASSOLUTA)	FREQUENZE ASSOLUTE CUMULATE
1	3.204	3.204
2	350	3.554
3	72	3.626
4	58	3.684
5	21	3.705
6	5	3.710
7 e più	2	3.712

Ora N = 3.712, che è pari. Quindi N/2 = 1.856. La mediana è il numero di componenti che corrisponde al termine di posto 1.856. Dalla colonna delle frequenze assolute si evince che il termine di posto 1.856 corrisponde al valore di 1 componente. Quindi Me = 1.

3. Nell'esempio n° 3 del paragrafo 2, per calcolare la mediana è necessario ampliare la tabella:

CLASSI DI ALTEZZA (IN CM)	N° STUDENTI (FREQUENZA ASSOLUTA)	FREQUENZA ASSOLUTA CUMULATA
fino a 150	1	1
150 —┤ 160	4	5
160 —┤ 165	6	11
165 —┤ 170	8	19
170 —┤ 180	3	22
oltre 180	1	23

Ora N = 23, che è dispari. Quindi (N+1)/2 = 12. La classe mediana è la classe che corrisponde al termine di posto 12. Dalla colonna delle frequenze assolute si evince che il termine di posto 12 corrisponde alla classe 165 —┤ 170.

Ora si procede ad un'interpolazione lineare:

$165 \rightarrow 11$

$Me \rightarrow 12$

$170 \rightarrow 19$

Quindi:

$$Me = \frac{12-11}{19-11}(170-165)+165 = 165{,}625$$

Perciò l'altezza mediana degli studenti è Me=165,625 cm.

5. INDICI DI VARIABILITA'

Una carattere importante dei dati statistici è la **variabilità**. Per analizzare una distribuzione, dopo aver calcolato uno o più valori medi, si cerca di evidenziare la dispersione dei dati, che caratterizza la variabilità del fenomeno. Può interessare conoscere sia di quanto i dati differiscano da un valore medio, sia di quanto i dati differiscono fra loro. Per misurare la variabilità di un fenomeno vi sono vari indici.

CAMPO DI VARIAZIONE (O ESCURSIONE)

Si definisce **campo di variazione** la differenza tra il maggiore ed il minore dei valori rilevati. Tale indice è molto semplice da calcolare, ma non ha grande importanza in quanto tiene conto solo dei valori estremi e non degli altri.

ESEMPIO

Sia data la seguente tabella che indica la produzione mondiale (in migliaia di quintali) di fibre naturali e chimiche:

ANNI	COTONE E LANA	FIBRE CHIMICHE
1995	21.448	22.207
1996	20.956	25.235
1997	21.286	27.927
1998	19.937	29.314
1999	20.250	30.376
2000	20.775	32.464

Calcolare il campo di variazione delle due serie storiche.

Lana e cotone: $21.448 - 19.937 = 1.511$ (migliaia di quintali)

Fibre chimiche: $32.464 - 22.207 = 10.257$ (migliaia di quintali)

Il campo di variazione della produzione di fibre chimiche è molto maggiore di quello della produzione di fibre naturali.

SCARTO QUADRATICO MEDIO; VARIANZA

Si definisce **scarto quadratico medio** (o **deviazione standard**) **σ** la media quadratica, semplice o ponderata, degli scarti dei valori dalla media aritmetica. Esso è tanto più piccolo quanto più i dati sono prossimi al valore medio ed è nullo se e solo se i dati sono tutti uguali fra loro. E' un indice della dispersione dei dati molto sensibile per evidenziare l'esistenza di dati che si scostano molto dal valore medio. Per il calcolo si usano le seguenti formule:

Per le serie:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - M)^2}{n}}$$

Per le seriazioni:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - M)^2 \cdot y_i}{\sum_{i=1}^{n} y_i}}$$

Il quadrato dello scarto quadratico medio, cioè $\boldsymbol{\sigma^2}$, è detto **varianza**. Si può dimostrare che la varianza è uguale alla differenza fra la media aritmetica, semplice o ponderata, dei quadrati dei termini ed il quadrato della media, cioè:

Per le serie:

$$\sigma^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} - M^2$$

Per le seriazioni:

$$\sigma^2 = \frac{\sum_{i=1}^{n} x_i^2 \cdot y_i}{\sum_{i=1}^{n} y_i} - M^2$$

ESEMPI

1. Si sono misurate le lunghezze del salto in lungo di due studenti e si sono ricavati i seguenti valori in centimetri:

1° studente: 368; 372; 373; 371; 366

2° studente: 360; 371; 385; 370; 364

Si calcoli lo scarto quadratico medio delle lunghezze dei salti.

Prima di tutto si calcola la media aritmetica:

$$M_1 = \frac{368 + 372 + 373 + 371 + 366}{5} = 370$$

$$M_2 = \frac{360 + 371 + 385 + 370 + 364}{5} = 370$$

Per entrambi gli studenti, la lunghezza media dei salti è 370 cm, ma sono molto più variabili le lunghezze dei salti del secondo studente.

Si calcola ora lo scarto quadratico medio:

$$\sigma_1 = \sqrt{\frac{(368-370)^2 + (372-370)^2 + (373-370)^2 + (371-370)^2 + (366-370)^2}{5}} =$$

$$= 2{,}61\ cm$$

$$\sigma_1 = \sqrt{\frac{(360-370)^2 + (371-370)^2 + (385-370)^2 + (370-370)^2 + (364-370)^2}{5}} =$$

$$= 8{,}51\ cm$$

Quindi, in media, gli scarti del secondo studente sono superiori a quelli del primo.

2. Calcolare lo scarto quadratico medio per la distribuzione del parco automobilistico secondo la cilindrata.

CLASSI DI CILINDRATA	N° AUTOVETTURE (IN MIGLIAIA)
0 ⊣ 800	2.291
800 ⊣ 1.200	10.828
1.200 ⊣ 1.600	11.067
1.600 ⊣ 2.000	7.712
2.000 ⊣ 2.500	1.320
2.500 ⊣ 3.000	489

Si costruisce la tabella per calcolare media aritmetica ponderata e scarto quadratico medio.

CLASSI DI CILINDRATA	N° AUTOVETTURE (IN MIGLIAIA) Y	VALORI CENTRALI X	X · Y	$(X-M)$	$(X-M)^2$	$(X-M)^2 \cdot Y$
0 ⊣ 800	2.291	400	916.400	-947,93	898.564,17	2.058.610.512
800 ⊣ 1.200	10.828	1.000	10.828.000	-347,93	121.052,67	1.310.758.345
1.200 ⊣ 1.600	11.067	1.400	15.493.800	52,07	2.711,68	30.010.115,8
1.600 ⊣ 2.000	7.712	1.800	13.881.600	452,07	204.370,68	1.576.106.671
2.000 ⊣ 2.500	1.320	2.250	2.970.000	902,07	813.737,06	1.074.132.914
2.500 ⊣ 3.000	489	2.750	1.344.750	1.402,07	1.965.810,81	961.281.486
TOTALI	33.707		45.434.550			7.010.900.045

Quindi si ha:

$$M = \frac{45.434.550}{33.707} = 1.347{,}93$$

$$\sigma = \sqrt{\frac{7.010.900.045}{33.707}} = 456{,}065$$

Perciò la cilindrata delle autovetture mediamente differisce di circa 456 cm^3 dalla cilindrata media.

Lo scarto quadratico medio è espresso nelle unità di misura scelte per la rilevazione dei dati ed è quindi di difficile confronto con lo scarto quadratico medio di altre rilevazioni in unità di misura differenti. Si introduce perciò un **coefficiente di variabilità**, che è un numero puro, dato dalla seguente formula:

$$C.V. = \frac{\sigma}{M}$$

ESEMPIO

Calcolare e confrontare i coefficienti di variabilità dell'esempio 1 precedente.

Dall'esempio precedente si ricava:

$$M_1 = 370\ cm \qquad \sigma_1 = 2{,}61\ cm$$
$$M_2 = 370\ cm \qquad \sigma_2 = 8{,}51\ cm$$

Quindi:

$$C.V._1 = \frac{2{,}61}{370} = 0{,}0071 = 0.71\% \qquad C.V._2 = \frac{8{,}51}{370} = 0{,}023 = 2{,}3\%$$

Quindi è più alto il coefficiente di variabilità de secondo studente.

Data una variabile statistica X, la sua media aritmetica M e la sua varianza σ^2, assegnato un valore k, la probabilità che i valori della variabile differiscano almeno di k dal valore medio è inferiore al rapporto $\frac{\sigma^2}{k^2}$, cioè:

$$P(|X - M| \geq k) \leq \frac{\sigma^2}{k^2}$$

Tale relazione prende il nome di **teorema di Bienaymé-Čebičev**.

ESEMPIO

Considerando l'esercizio precedente, determinare quale percentuale di salti differisce dal salto medio del secondo studente di almeno 15 cm.

$$P(|X - 370| \geq 4) \leq \frac{8{,}51^2}{15^2} = 0{,}322 = 32{,}2\%$$

Pertanto meno del 32,2% dei salti ha una lunghezza che differisce da quella media di almeno 15 cm.

SCOSTAMENTO SEMPLICE MEDIO

Lo scostamento semplice medio è la media aritmetica dei valori assoluti degli scarti dei valori della distribuzione da un valore medio. Quindi se si considera come valore medio la media aritmetica, semplice o ponderata, della distribuzione, si ha:

Per le serie:

$$S_M = \frac{\sum_{i=1}^{n} |x_i - M|}{n}$$

Per le seriazioni:

$$S_M = \frac{\sum_{i=1}^{n} |x_i - M| \cdot y_i}{\sum_{i=1}^{n} y_i}$$

CONCENTRAZIONE

Un particolare aspetto della variabilità di un fenomeno è la **concentrazione**, che permette di stabilire se il fenomeno è equamente distribuito fra tutte le unità statistiche o se è concentrato in poche unità. Lo studio della concentrazione si usa per analizzare distribuzioni di redditi o di consumi tra le persone di una popolazione, distribuzione di imprese industriali per numero di addetti, e così via. Tra i metodi per

misurare la concentrazione si consideri il **metodo grafico di Lorentz** applicato allo studio della distribuzione della ricchezza e descritto di seguito:

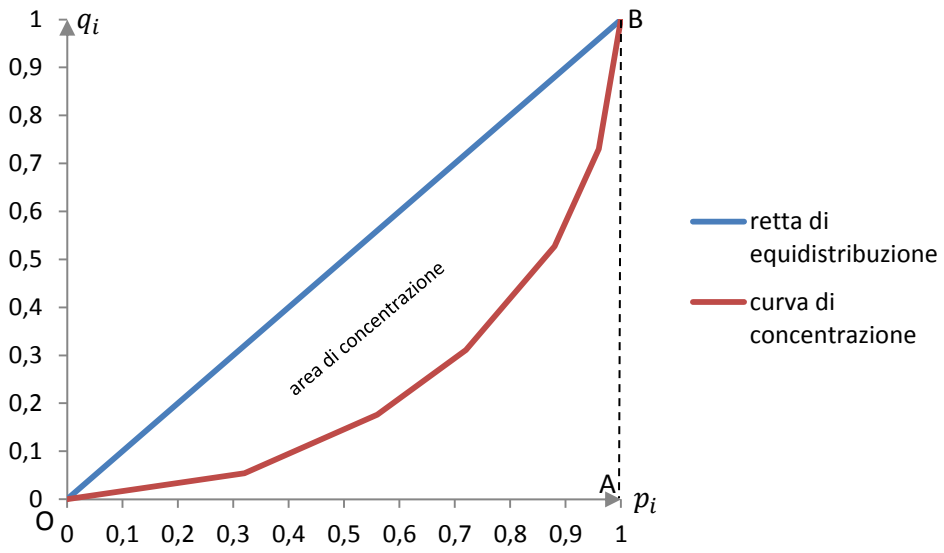
1. Sia data una variabile statistica $x_1, x_2, \dots, x_n$ i cui valori (o i valori centrali nel caso i valori siano suddivisi in classi) siano in ordine non decrescente e siano $y_1, y_2, \dots, y_n$ le relative frequenze.
2. Si calcolano i prodotti $x_i \cdot y_i$ che rappresentano **l'intensità dell'i-esimo carattere**, e la loro somma che indica **l'intensità globale del fenomeno**.
3. Si calcolano le frequenze cumulate e le intensità cumulate.
4. Si calcolano le frequenze relative cumulate, indicate con p_i, dividendo le frequenze cumulate per la somma delle frequenze.
5. Si calcolano le intensità relative cumulate, indicate con q_i, dividendo le intensità cumulate per l'intensità globale.
6. Si rappresentano su un piano cartesiano le coppie $(p_i; q_i)$, collegandole con una spezzata detta **curva di concentrazione di Lorentz**.
7. Si traggono le conclusioni.

Conclusioni:

Se le intensità relative cumulate q_i sono uguali alle frequenze relative cumulate p_i, la curva di Lorentz coincide con la bisettrice del primo quadrante, che prende il nome di **retta di equidistribuzione**, poiché il fenomeno è **equidistribuito**. Se le intensità relative cumulate q_i sono minori delle rispettive frequenze relative cumulate p_i, il fenomeno è tanto più **concentrato** quanto più le q_i sono inferiori alle relative p_i. La concentrazione è massima quando l'intensità globale è concentrata in una sola unità statistica. Il **rapporto di concentrazione** è dato da (con riferimento al grafico seguente):

$$R = \frac{area\ di\ concentrazione}{area\ del\ triangolo\ OAB} \qquad \text{e risulta} \qquad 0 \leq R \leq 1$$

Se l'area di concentrazione è nulla, la curva di Lorentz coincide con la retta di equidistribuzione, R=0 e si dice che **esiste equidistribuzione**; se l'area di concentrazione coincide con l'area del triangolo OAB, R=1 e la **concentrazione è massima**.



ESEMPIO

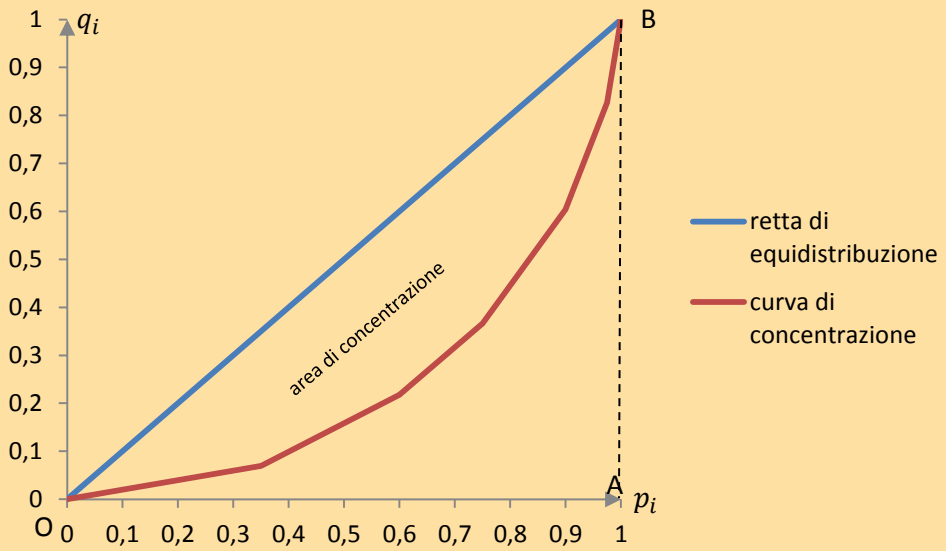
Analizzare la concentrazione della seguente distribuzione del reddito (in migliaia di euro) di 400 persone:

CLASSI DI REDDITO	N° PERSONE
0 ⊣ 2	140
2 ⊣ 4	100
4 ⊣ 6	60
6 ⊣ 10	60
10 ⊣ 20	30
20 ⊣ 50	10

Si costruisce una nuova tabella in cui sono presenti le colonne delle frequenze relative cumulate e delle intensità relative cumulate:

CLASSI DI REDDITO	VALORE CENTRALE X_i	FREQUENZE Y_i	INTENSITÀ $X_i Y_i$	FREQUENZE ASSOLUTE CUMULATE	INTENSITÀ ASSOLUTE CUMULATE	FREQUENZE RELATIVE CUMULATE P_i	INTENSITÀ RELATIVE CUMULATE Q_i
0 ⊣ 2	1	140	140	140	140	0,35	0,0693
2 ⊣ 4	3	100	300	240	440	0,6	0,2178
4 ⊣ 6	5	60	300	300	740	0,75	0,3663
6 ⊣ 10	8	60	480	360	1.220	0,9	0,604
10 ⊣ 20	15	30	450	390	1.670	0,975	0,8267
20 ⊣ 50	35	10	350	400	2.020	1	1
TOTALI		400	2.020				

Analizzando le ultime due colonne della tabella si evince che il 35% delle persone possiede il 6,93% del reddito; il 60% delle persone possiede il 21,78% del reddito e così via. Non c'è quindi equidistribuzione. Calcoliamo ora l'area di concentrazione come differenza tra l'area del triangolo OAB e l'area S sotto la curva di concentrazione, che è formata da un triangolo e da tanti trapezi rettangoli. Graficamente:



$$S = \frac{0.0693 \cdot 0.35}{2} + \frac{(0.2178 + 0.0639) \cdot (0.6 - 0.35)}{2} + \frac{(0.3663 + 0.2178) \cdot (0.75 - 0.6)}{2} + \frac{(0.604 + 0.3663) \cdot (0.9 - 0.75)}{2} + \frac{(0.8267 + 0.604) \cdot (0.975 - 0.9)}{2} + \frac{(1 + 0.8267) \cdot (1 - 0.975)}{2} = 0.24108$$

$$R = \frac{\frac{1 \cdot 1}{2} - 0.24108}{\frac{1 \cdot 1}{2}} = 0.51784$$

Quindi il reddito è abbastanza concentrato.

6. RAPPORTI STATISTICI

RAPPORTI DI COMPOSIZIONE

Sono rapporti tra dati omogenei ed indicano come la frequenza si ripartisce fra le possibili scelte. Si calcolano come rapporto fra frequenza di un fenomeno e frequenza totale.

RAPPORTI DI DERIVAZIONE

Confrontano due dati statistici il primo dei quali è conseguenza del secondo e servono per confrontare l'andamento di un fenomeno in luoghi o periodi diversi. Ad esempio: il **quoziente di mortalità** (rapporto fra il numero di morti in un anno e il totale della popolazione), il **quoziente di natalità**, il rapporto tra il numero di bocciati ed il totale degli studenti, e così via.

RAPPORTI DI DENSITÀ

Confrontano sia due dati omogenei sia dati non omogenei; ad esempio la densità di popolazione è il rapporto fra il numero di abitanti e la superficie di territorio in cui vivono, e così via.

RAPPORTI DI COESISTENZA

Sono rapporti tra le frequenze di due fenomeni diversi riferiti alle stesse unità statistiche e danno un'indicazione dello squilibrio fra dati coesistenti in uno stesso luogo ed in uno stesso intervallo di tempo.

NUMERI INDICE

Sono rapporti espressi in percentuale fra le intensità di un certo fenomeno in tempi e luoghi diversi; servono per analizzare l'andamento del fenomeno stesso. Si possono calcolare:

a **base fissa**, in tal caso si sceglie un dato come base e si dividono tutti gli altri per la base e così si evidenzia la variazione del fenomeno rispetto alla base scelta;

a **base mobile** e si divide ciascun dato per il precedente, individuando la variazione di ciascun dato rispetto al precedente.

ESEMPI

1. Si calcolino i rapporti di composizione della tabella dell'esempio 1 paragrafo 2.

TITOLO DI STUDIO	N° PERSONE	RAPPORTI DI COMPOSIZIONE (%)
Senza titolo	1	3.33
Licenza elementare	3	10
Scuola media inferiore	5	16.67
Scuola media superiore	12	40
Laurea	9	30
TOTALE	30	100

Interpretazione: il 10% delle persone esaminate possiede il titolo di licenza elementare, il 40% di scuola media superiore, ecc.

2. Si calcolino i numeri indici a base fissa e mobile dell'es. n° 4 paragrafo 2. Si sceglie come base l'anno 2007.

ANNI	N° STUDENTI	NUMERI INDICE A BASE FISSA (%)	NUMERI INDICE A BASE MOBILE (%)
2007	105	100	---
2008	102	97,14	97,14
2009	112	106,67	109,80
2010	120	114,29	107,14
2011	104	99,05	86,67
2012	122	116,19	117,31
TOTALE	665		

Interpretazione: i numeri indici superiori a 100 indicano un incremento di studenti, mentre quelli inferiori a 100 indicano una diminuzione; rispetto al 2007 per quelli a base fissa, rispetto all'anno precedente per quelli a base mobile.