Elementi di statistica

1 Generalità

Le prime statistiche erano una specie di contabilità della popolazione umana e dei dati economici: numero delle nascite e delle morti per anno, percentuale di maschi tra i nuovi nati, ripartizione della popolazione attiva secondo i diversi tipi di attività ecc.
Con il passare del tempo la Statistica è diventata un metodo di indagine che permette di individuare leggi fondamentali in fenomeni di massa apparentemente governati dal caso. I campi di applicazione del metodo statistico sono numerosissimi e ad essi sono dedicati importanti istituti scientifici nazionali.
I campi tradizionali della demografia e dell'economia sono particolarmente studiati dall'Istituto Nazionale di Statistica; i campi collegati alla salute e alla sperimentazione medica sono coordinati dall'Istituto Superiore di Sanità; i campi collegati alla sismologia e in generale alle scienze geologiche fanno riferimento all'Istituto Nazionale di Geofisica.

La Statistica studia dunque i metodi per interpretare i dati raccolti e le informazioni che da questi si possono dedurre per trarre conclusioni sull'andamento dei fenomeni studiati.

In senso più ristretto, con il termine "statistica" si usa denotare l'insieme dei dati raccolti: si parla cioè di "statistica della popolazione con un titolo di studio di scuola media superiore", oppure di "statistica degli abbonati alle partite di calcio di serie A", o ancora di "statistica dei biglietti aerei venduti in un anno nelle tratte nazionali" ecc.

Terminologia statistica

I termini **popolazione** e **individuo**, che originariamente avevano il senso letterale, hanno acquistato con il tempo un carattere generale per indicare, rispettivamente, l'intero insieme e ogni suo singolo elemento.

Esempi

1. Gli studenti che frequentano una data scuola costituiscono la popolazione statistica di cui ogni studente è un individuo.
2. La totalità delle famiglie italiane è una popolazione statistica; ogni famiglia italiana è un individuo di tale popolazione.
3. Tutti i partecipanti a un dato concorso costituiscono la popolazione considerata, della quale ciascun candidato è individuo.
4. Tutte le monete coniate dalla Zecca italiana in un anno costituiscono una popolazione, della quale ciascuna moneta è individuo.
5. Le aziende metalmeccaniche italiane rappresentano una popolazione, della quale ogni singola azienda è individuo.

Una parte della popolazione è detta **campione**: da un punto di vista insiemistico un campione è un sottoinsieme della popolazione. Per esempio, gli alunni di una sezione costituiscono un campione degli alunni della scuola, le famiglie residenti in una città formano un campione delle famiglie italiane, le aziende metalmeccaniche dell'Umbria sono un campione delle aziende metalmeccaniche italiane.
I dati statistici possono provenire da varie fonti, come osservazioni dirette, esperimenti, pubblicazioni specializzate oppure possono essere raccolti per mezzo di questionari.
È raro poter avere a disposizione i dati provenienti dall'intera popolazione, in quanto è quasi sempre impossibile testare tutti gli individui. Pertanto è spesso opportuno prendere un campione della popolazione e ottenere i dati da questo. Se vogliamo trarre da questi dati conclusioni valide per l'intera popolazione, il campione deve essere scelto con grande cura. Noi supporremo che il campione preso in considerazione sia **casuale** (random); ciò significa che ogni individuo della popolazione ha la stessa probabilità di essere scelto per far parte del campione.
Se il campione è abbastanza numeroso, esso ha proprietà simili a quelle dell'intera popolazione e noi possiamo ragionevolmente confidare che i risultati della nostra ricerca fondata sul campione possano essere riconosciuti validi per l'intera popolazione.

Una proprietà che si possa osservare o studiare in ogni individuo è detta **carattere** o **attributo**. Un carattere che possa assumere diversi valori è detto **statistico**.
Un carattere statistico permette di stabilire all'interno della popolazione delle classi di equivalenza ponendo nella stessa classe tutti gli individui per i quali il carattere prende lo stesso valore. Per esempio, sui lavoratori di una certa azienda si possono considerare attributi quali l'età, il titolo di studio, la qualifica raggiunta: i lavoratori possono quindi essere classificati per età, per titolo di studio, per qualifica raggiunta ecc.

I caratteri misurati con dei numeri si dicono **quantitativi**, altrimenti **qualitativi**. Sono quantitativi l'età, il peso, la statura; sono qualitativi il colore degli occhi, la professione, la religione praticata.
La misura di un carattere è indicata anche come **intensità del carattere**.

2 Frequenze statistiche

La raccolta iniziale di dati produce tabelle di scarso interesse statistico, quasi sempre colossali archivi difficilmente leggibili.
Per esempio, l'Ufficio del Catasto di Roma possiede l'enorme elenco dei proprietari di immobili nella Capitale, elenco che include, fra l'altro, a fianco di ogni nome, il numero dei metri quadri posseduti.
Fare statistica ovviamente non vuol dire pubblicare, ammesso che le leggi sulla protezione della privacy lo consentano, tale elenco: fare statistica può, per esempio, voler dire calcolare quanti cittadini siano proprietari di immobili di meno di 80 metri quadri, quanti possiedano superfici tra 80 e 100 metri quadri ecc.
Determinare *quanti individui* di un elenco possiedano un *certo carattere* significa determinare le **frequenze** di tale carattere nell'elenco.
La determinazione delle frequenze è la prima, fondamentale operazione statistica.

Frequenza assoluta

Esempi

6 In un certo giorno alla visita di leva si sono presentati 165 giovani; di ognuno di questi è stata rilevata l'altezza.
Nella tabella a semplice entrata riportata a fianco sono indicati a sinistra l'intensità del carattere, a destra la frequenza assoluta. Si osservi che i dati sono stati distribuiti in classi e a ciascuna classe è stata attribuita una certa frequenza. Nella prima classe sono raccolte le altezze inferiori a 1,50 m, la colonna di destra mostra che nessun giovane è alto meno di 1,50 m.
La notazione $1{,}50 \vdash 1{,}60$ indica la classe in cui vengono conteggiate le altezze maggiori o uguali a 1,50 m e inferiori a 1,60, nella colonna di destra si legge che 15 giovani sono in questa classe.
Il valore 1,55 m, che è il valore medio della classe $1{,}50 \vdash 1{,}60$, viene detto **valore centrale** della classe stessa.

Altezza in metri	Frequenza assoluta
$< 1{,}50$	–
$1{,}50 \vdash 1{,}60$	15
$1{,}60 \vdash 1{,}70$	62
$1{,}70 \vdash 1{,}80$	58
$1{,}80 \vdash 1{,}90$	28
$\geq 1{,}90$	2
Totale	**165**

7 I punteggi, da 1 a 10, ottenuti dai 42 concorrenti a un concorso fotografico, in ordine di iscrizione al concorso, sono riportati qui di seguito:

7 4 5 7 5 4 3 6 8 4 9 3 5 5 6 7 2 1 6 6 2
5 3 8 8 7 5 4 6 3 2 9 1 10 3 4 7 9 1 7 6 5

I risultati sono raccolti nella tabella in basso.

Punteggio	1	2	3	4	5	6	7	8	9	10	
Frequenza assoluta	3	3	5	5	7	6	6	3	3	1	**Totale 42**

Osservazione 1

Le frequenze assolute sono numeri interi compresi tra zero e il numero totale di individui della popolazione.
La somma delle frequenze assolute dei valori di uno stesso carattere equivale al numero totale di individui della popolazione.

Frequenza relativa e frequenza percentuale

Esempio

8 Nell'esempio 6 la classe degli individui con altezza nell'intervallo 1,60 m ⊢ 1,70 m ha frequenza relativa

$$f = \frac{62}{165} = 0,38$$

Osservazione 2

Le frequenze relative sono sempre numeri compresi tra 0 e 1.
La somma delle frequenze relative dei vari valori di uno stesso carattere è uguale a 1.

È consuetudine ricondursi a collettivi di 100 unità: per far questo basta moltiplicare per 100 la frequenza relativa. Si ha così la **frequenza percentuale**.
Le due tabelle relative agli esempi 6 e 7 possono essere così completate:

Esempio 6

Altezza in metri	Frequenza assoluta	Frequenza relativa	Frequenza percentuale
< 1,50	–	–	–
1,50 ⊢ 1,60	15	0,091	**9,1%**
1,60 ⊢ 1,70	62	0,376	**37,6%**
1,70 ⊢ 1,80	58	0,351	**35,1%**
1,80 ⊢ 1,90	28	0,170	**17,0%**
≥ 1,90	2	0,012	**1,2%**
Totale	**165**	**0,999 ≅ 1**	**99% ≅ 100%**

Esempio 7

Punteggio	Frequenza assoluta	Frequenza relativa	Frequenza percentuale
1	3	0,071	7,1%
2	3	0,071	7,1%
3	5	0,119	11,9%
4	5	0,119	11,9%
5	7	0,167	16,7%
6	6	0,143	14,3%
7	6	0,143	14,3%
8	3	0,071	7,1%
9	3	0,071	7,1%
10	1	0,024	2,4%
Totale	**42**	**0,999 ≅ 1**	**99,9% ≅ 100%**

Si osservi che, a causa delle approssimazioni nel calcolo delle divisioni, la somma delle frequenze relative è un numero prossimo a 1 e la somma delle frequenze percentuali è prossima a 100. Esprimendo le frequenze relative come frazioni, la loro somma è esattamente uguale a 1.

Rappresentazioni grafiche di una distribuzione di frequenze

Per rappresentare graficamente una distribuzione di frequenze si può far uso di istogrammi o di poligoni di frequenze.

Istogrammi e poligoni delle frequenze

Un **istogramma** viene di solito usato con dati raggruppati in classi ed è costituito da un insieme di rettangoli, ciascuno dei quali è così costruito:

1) la base, posta sull'asse orizzontale, ha il centro nel valore centrale della classe ed è proporzionale all'ampiezza della classe; i rettangoli possono quindi avere basi differenti;
2) l'area del rettangolo è proporzionale alla frequenza della classe.

Se le classi hanno tutte la stessa ampiezza, l'altezza dei rettangoli è proporzionale alle frequenze delle classi; pertanto l'altezza può essere presa uguale alle frequenze delle classi.

Il **poligono delle frequenze** è una spezzata che unisce i punti aventi per ascissa i punti centrali delle classi e per ordinata la relativa frequenza. Costruito l'istogramma, il poligono delle frequenze unisce i punti medi dei lati superiori dei rettangoli.

Esempi

9 La tabella a fianco riporta la distribuzione delle frequenze degli stipendi dei 123 impiegati di una azienda.

Stipendio (in euro)	Frequenza assoluta	Frequenza relativa	Frequenza percentuale
1700 ⊢ 1850	36	0,29	29%
1850 ⊢ 2000	42	0,34	34%
2000 ⊢ 2150	15	0,12	12%
2150 ⊢ 2300	24	0,20	20%
2300 ⊢ 2450	6	0,05	5%
Totale	**123**	**1**	**100%**

L'istogramma con il relativo poligono delle frequenze è riportato in figura 1.
Si osservi che si sono considerate anche le due classi di frequenza zero contigue alle classi estreme; in tal caso la somma delle aree dei rettangoli dell'istogramma è uguale all'area totale racchiusa dal poligono delle frequenze.

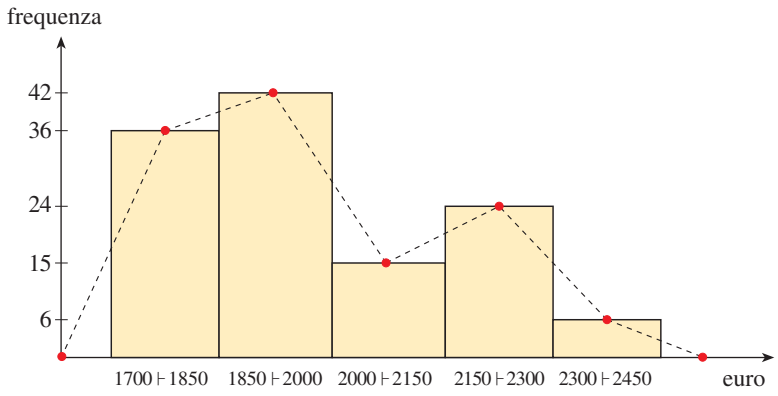


Figura 1 Stipendi del personale.

10 Vengono intervistati 600 ragazzi dai 16 ai 18 anni sull'ammontare delle loro spese mensili in giornali e riviste. I dati vengono raccolti nella tabella A.
Le classi di questa tabella hanno tutte la stessa ampiezza, pertanto le basi dei rettangoli dell'istogramma che rappresenta tale distribuzione di frequenze sono uguali e le altezze devono essere proporzionali alle frequenze. Poiché le frequenze oscillano tra 3 e 194, il rettangolo che corrisponde alla frequenza maggiore deve essere circa $65 \cong \frac{194}{3}$ volte più alto di quello che corrisponde alla frequenza minore.

È evidente che un istogramma così fatto è difficile da disegnare. È perciò utile raggruppare le classi estreme come per esempio nella tabella B, in modo da ottenere un istogramma più bilanciato.
In tal caso l'oscillazione delle frequenze varia tra 60 e 194. Per disegnare l'istogramma che rappresenta tale distribuzione occorre tener conto che le colonne estreme hanno ampiezza la prima 4 volte, l'ultima 3 volte più grande dell'ampiezza delle colonne intermedie. Poiché le aree dei rettangoli sono proporzionali alle frequenze e l'altezza dei rettangoli di base 1 (per esempio 4 ⊢ 4,99) è proporzionale alla rispettiva frequenza, allora:

- l'altezza del primo rettangolo di base 4 ha altezza proporzionale a $\frac{60}{4} = 15$;

TABELLA *A*

Spesa in euro	Frequenza assoluta
0 ⊢ 0,99	4
1 ⊢ 1,99	7
2 ⊢ 2,99	14
3 ⊢ 3,99	35
4 ⊢ 4,99	98
5 ⊢ 5,99	125
6 ⊢ 6,99	194
7 ⊢ 7,99	107
8 ⊢ 8,99	13
9 ⊢ 9,99	3
Totale	**600**

TABELLA *B*

Spesa in euro	Frequenza assoluta
0 ⊢ 3,99	60
4 ⊢ 4,99	98
5 ⊢ 5,99	125
6 ⊢ 6,99	194
7 ⊢ 9,99	123
Totale	**600**

- l'altezza dell'ultimo rettangolo di base 3 ha altezza proporzionale a $\frac{123}{3} = 41$.

Con questi dati l'istogramma è quello riportato in figura 2.

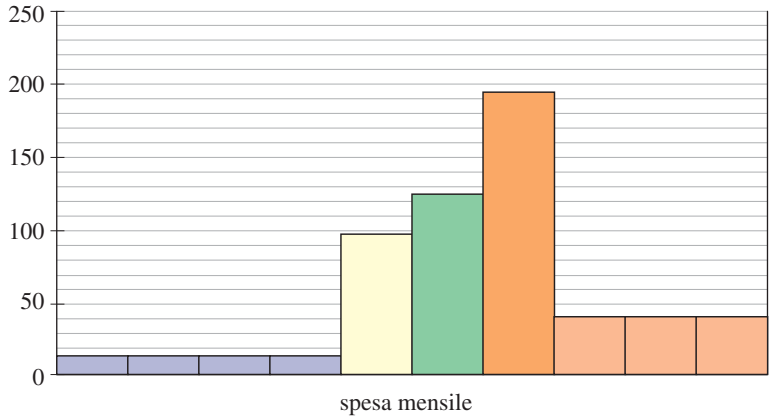


Figura 2 Spesa per giornali e riviste.

Osservazione 3

I due esempi ora illustrati ci permettono di concludere che:

- se i rettangoli di un istogramma hanno tutti la stessa base, si paragonano le frequenze delle classi paragonando le altezze dei rettangoli corrispondenti;
- se i rettangoli di un istogramma hanno basi differenti, si paragonano le frequenze delle classi paragonando le aree dei rettangoli corrispondenti: se per esempio due rettangoli hanno basi una doppia dell'altra e la stessa altezza, allora la frequenza della classe di base maggiore è doppia dell'altra.

Diagrammi a settori circolari o "a torta"

Una distribuzione di frequenze si può anche rappresentare mediante un **diagramma a settori circolari**, detto anche "*a torta*". Si divide il cerchio in 100 settori uguali, ciascuno di ampiezza $\frac{1}{100} \cdot 360° = 3{,}6°$; ciascuno di essi è l'1% dell'intero cerchio.

Una percentuale del $p\%$ è rappresentata da un settore di ampiezza $p \cdot 3{,}6°$.

Il diagramma riportato in figura 3 rappresenta la distribuzione di frequenze della tabella dell'esempio 6.

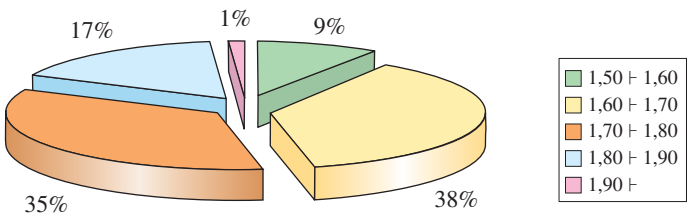


Figura 3 Altezza alla leva.

I diagrammi delle figure 4 e 5 si riferiscono alla spesa media mensile delle famiglie italiane suddivisa per alimenti e complessiva per aree geografiche relativa all'anno 1997 (dati ISTAT).

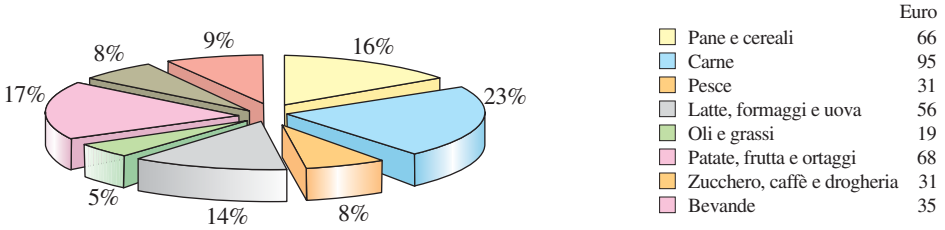


Figura 4 Spesa alimentare familiare mensile (1997).

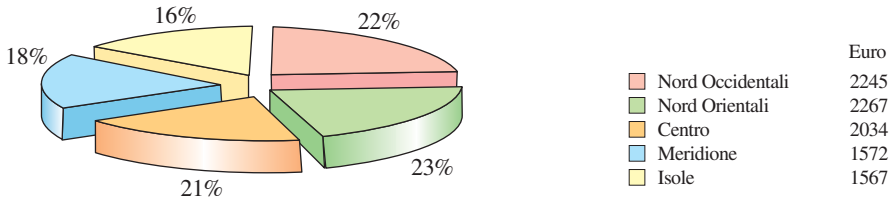


Figura 5 Spesa familiare mensile per aree geografiche (1997).

Grafici a bastoni

La distribuzione dei punteggi dell'esempio 7 è relativa a una variabile che assume valori interi; graficamente essa può essere ben rappresentata da un grafico a bastoni (fig. 6). La lunghezza di ciascun segmento è proporzionale alla frequenza del relativo punteggio.

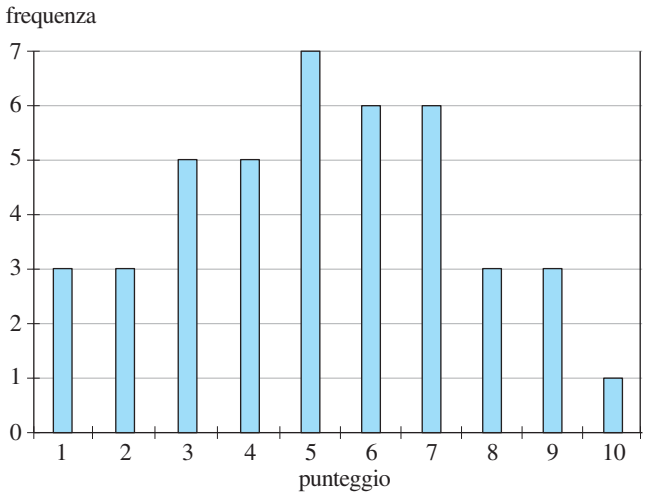


Figura 6 Punteggi del concorso fotografico.

Frequenza	13	16	16	22	18	15
Peso (kg)	65	66	67	68	69	70

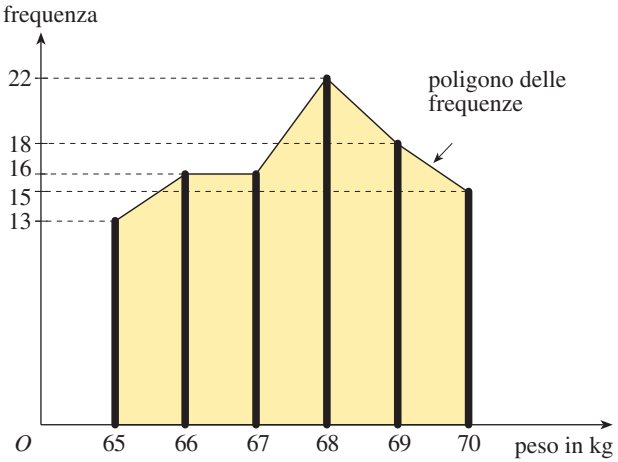


Figura 7

Nella tabella a sinistra è riportata la distribuzione delle frequenze del peso rilevato su 100 ragazzi di 16 anni. Poiché i dati sono discreti, si può usare una rappresentazione grafica a bastoni come nella figura 7.

In generale, se i dati sono discreti, il grafico a bastoni può essere considerato una rappresentazione nel piano cartesiano: sull'asse delle ascisse si riportano le modalità del carattere (nell'esempio il peso) e sull'asse delle ordinate le relative frequenze. La spezzata che unisce tali punti, cioè il **poligono delle frequenze**, è il **diagramma cartesiano**.

Frequenza cumulata

DEFINIZIONE Si dice **frequenza cumulata assoluta**, o rispettivamente **relativa** (o **percentuale**), di un valore, o di un intervallo di valori, la somma delle frequenze assolute, o rispettivamente relative (o percentuali), dello stesso carattere relative a tutti i valori, o agli intervalli di valori, minori o uguali al valore considerato.

Esempi

11 Riprendiamo l'esempio 9, relativo alle classi di stipendio.

Stipendio (in euro)	Frequenza assoluta	Frequenza cumulata assoluta	Frequenza percentuale	Frequenza cumulata percentuale
1700 ⊢ 1850	36	36	29%	29%
1850 ⊢ 2000	42	78 = 42 + 36	34%	63%
2000 ⊢ 2150	15	93 = 15 + 78	12%	75%
2150 ⊢ 2300	24	117 = 24 + 93	20%	95%
2300 ⊢ 2450	6	123 = 6 + 117	5%	100%
Totale	**123**		**100%**	

Dalla tabella si può ricavare che la percentuale degli impiegati il cui stipendio è inferiore a 2150 euro è 75%.
Si possono utilizzare per le frequenze cumulate, sia assolute sia percentuali, diagrammi analoghi a quelli già visti. Utilizziamo per esempio un istogramma, riportando sull'asse verticale le frequenze percentuali cumulate (fig. 8).

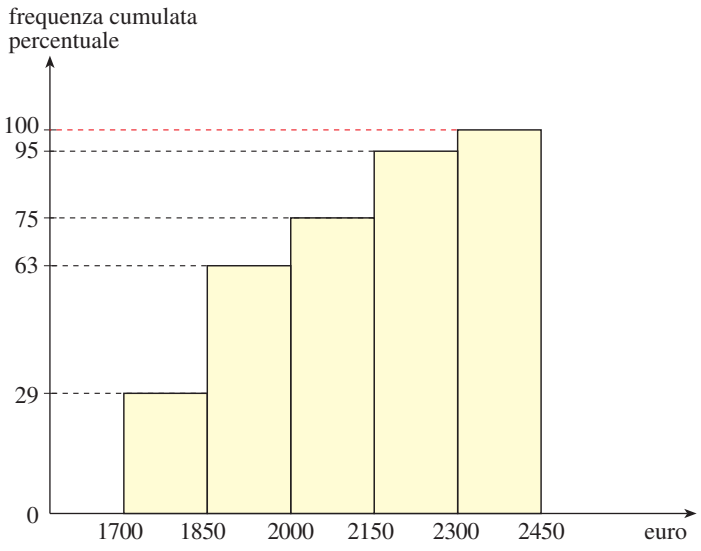


Figura 8

12 Riprendiamo l'esempio 6 e completiamo la tabella con le frequenze cumulate assolute e percentuali.

Altezza (in metri)	Frequenza assoluta	Frequenza cumulata assoluta	Frequenza percentuale	Frequenza cumulata percentuale
$< 1{,}50$	–	–	–	–
$1{,}50 \vdash 1{,}60$	15	15	9,1%	9,1%
$1{,}60 \vdash 1{,}70$	62	77	37,6%	46,7%
$1{,}70 \vdash 1{,}80$	58	135	35,1%	81,8%
$1{,}80 \vdash 1{,}90$	28	163	17,0%	98,8%
$\geq 1{,}90$	2	165	1,2%	100%
Totale	**165**		**100%**	

La tabella ci permette di dire che tra i 165 giovani il 46,7% ha altezza inferiore a 1,70 m e che il 98,8% ha altezza inferiore a 1,90 m (fig. 9).

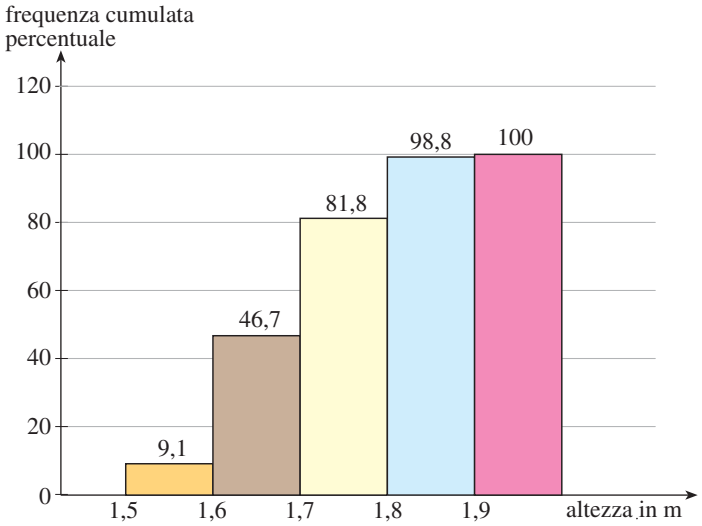


Figura 9

Osservazione 4

Le frequenze cumulate assolute o percentuali costituiscono una fila di valori non decrescente: la prima frequenza cumulata coincide con la frequenza del primo valore, la seconda frequenza cumulata coincide con la somma delle frequenze del primo e del secondo ecc. La frequenza cumulata assoluta dell'ultimo valore coincide con il numero di individui di tutta la popolazione.
La frequenza cumulata percentuale corrispondente all'ultimo valore è 100.

3 Progettare un questionario

In questo paragrafo diamo alcune indicazioni che risulteranno utili nel caso in cui si decida di effettuare un sondaggio per mezzo di un questionario.
Prima di iniziare a raccogliere i dati è importante avere un'idea chiara di quali dati occorrono e come si intende usarli. Si può investigare mediante un sondaggio su vari aspetti della vita sociale: il numero dei componenti delle famiglie di un quartiere, i mezzi di trasporto utilizzati per raggiungere la scuola, gli sport o gli svaghi preferiti dagli studenti di una scuola, il rendimento nelle materie scolastiche e tanto altro ancora.
Nel progettare un questionario si devono tenere a mente i seguenti punti:

- l'inchiesta deve raccogliere tutte le informazioni necessarie;
- il numero delle domande deve essere limitato, altrimenti gli intervistati perdono concentrazione e sono tentati di rispondere a caso;
- le domande non devono essere ambigue, cioè devono essere chiare e comprensibili;
- deve essere chiaro come rispondere alle domande, o attraverso un SÌ o un NO, oppure prevedere una serie di differenti risposte su cui apporre una crocetta; è opportuno non lasciare la risposta libera altrimenti gli intervistati scelgono modalità diverse che è poi difficile catalogare;
- offrire un ragionevole numero di risposte;
- non chiedere alle persone valutazioni soggettive tipo se la loro casa è " grande " o "piccola" oppure se i loro conoscenti sono "giovani" o "vecchi".

Esempio

13 *Un questionario tra i* 154 *studenti delle prime classi di una scuola superiore circa le preferenze in fatto di film è stato così formulato:*

a) Quali tipi di film preferisci? *(si può dare una sola preferenza)*

☐ Avventura ☐ Commedia ☐ Comici ☐ Horror
☐ Drammatici ☐ Musicali ☐ Fantascienza

b) Quanti film hai visto negli ultimi sei mesi?

☐ da 0 a 5 ☐ da 6 a 10 ☐ da 11 a 15 ☐ più di 15

Osserviamo subito che il primo quesito riguarda un carattere qualitativo, mentre il secondo si riferisce a un carattere quantitativo.
Il sondaggio ha dato i risultati riportati nella tabella.

a) *Quali tipi di film preferisci?*

Film preferiti	Frequenza
Avventura	39
Commedia	24
Comici	60
Horror	5
Drammatici	3
Musicali	7
Fantascienza	16
Totale	**154**

I risultati possono essere visualizzati per mezzo di un diagramma a torta (fig. 10) oppure per mezzo di un diagramma a bastoni (fig. 11).

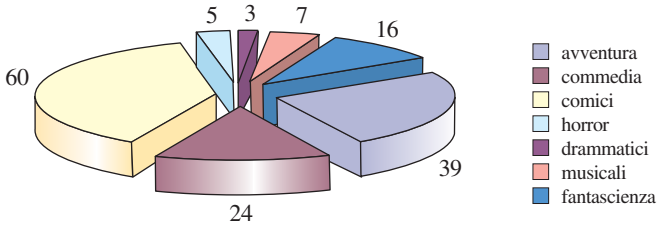


Figura 10 Film preferiti.

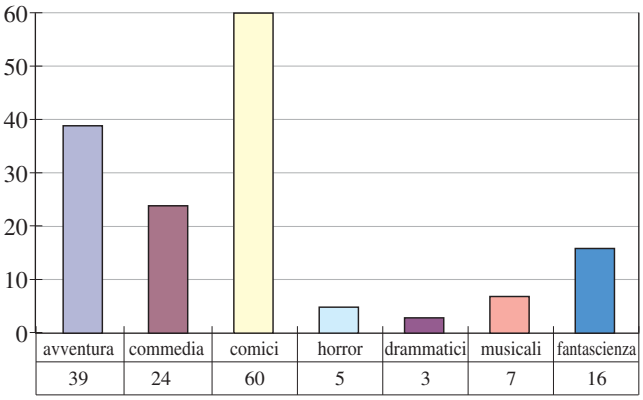


Figura 11 Film preferiti.

b) Quanti film hai visto negli ultimi sei mesi?

Numero di film	0-5	6-10	11-15	più di 15	
Risposte	11	34	63	46	**Totale 154**

I risultati possono essere visualizzati per mezzo di un istogramma (fig. 12).

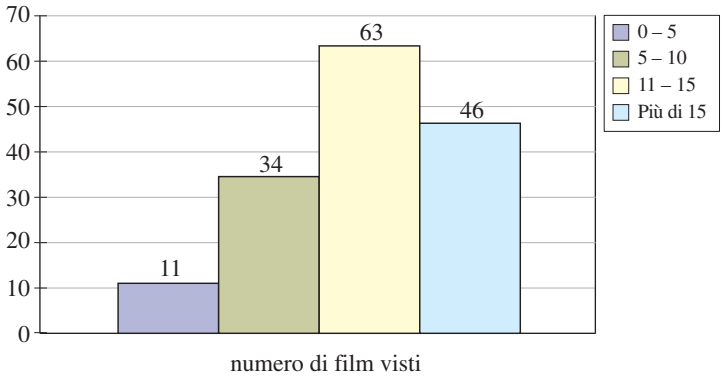


Figura 12

4 Media aritmetica. Moda. Mediana

In una indagine statistica, dopo aver rilevato i dati e averli raccolti in una tabella che contenga le relative frequenze, è spesso utile far ricorso a un valore che li rappresenti nel loro insieme dando in sintesi una informazione sul carattere delle rilevazioni.
Per esempio, supponiamo di voler confrontare le altezze degli studenti di due classi parallele. Le due tabelle con le relative frequenze non permettono di evidenziare immediatamente eventuali differenze. È necessario allora associare a ciascuna tabella due valori che le riassumano e che consentano di effettuarne il confronto.
Tali valori di sintesi prendono il nome di **medie**.

Media aritmetica

Il numero che più spesso si associa agli n valori $x_1, x_2, \ldots, x_n$ di un certo carattere quantitativo posseduto dagli n individui di una popolazione è la loro **media aritmetica**.

DEFINIZIONE Si chiama **media aritmetica** di n valori $x_1, x_2, \ldots, x_n$ il numero:

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

Se i valori $x_1, x_2, \ldots, x_k$ compaiono con frequenze rispettivamente $f_1, f_2, \ldots, f_k$ tali che $f_1 + f_2 + \ldots + f_k = n$ allora:

$$\mu = \frac{x_1 f_1 + x_2 f_2 + \ldots + x_k f_k}{n} \qquad [1]$$

Esempi

14 Si voglia determinare la media dei voti in sessantesimi riportati da 28 partecipanti a un concorso. La distribuzione è la seguente:

Voti	36	38	40	42	44	45	47	48	50	52	54	60
Frequenze	2	1	1	3	3	3	1	2	2	3	2	5

La media aritmetica è:

$$\mu = \frac{36\cdot 2 + 38 + 40 + 42\cdot 3 + 44\cdot 3 + 45\cdot 3 + 47 + 48\cdot 2 + 50\cdot 2 + 52\cdot 3 + 54\cdot 2 + 60\cdot 5}{28} = \frac{1350}{28} \cong 48{,}21$$

15 Riferendoci ora alla tabella dell'esempio 9, calcoliamo lo stipendio medio di un impiegato.
Per calcolare la media ci si riconduce a un carattere discreto, cioè si sostituisce a ogni classe il suo valore centrale assumendo l'ipotesi che i dati siano distribuiti in modo uniforme all'interno di ogni classe.
Pertanto, i valori centrali di ogni classe sono:

1775 1925 2075 2225 2375

Tenuto conto delle frequenze assolute indicate, la media aritmetica dei 123 stipendi è:

$$\mu = \frac{36\cdot 1775 + 42\cdot 1925 + 15\cdot 2075 + 24\cdot 2225 + 6\cdot 2375}{123} = \frac{243525}{123} = 1979{,}88 \text{ (euro)}$$

La media aritmetica è un **indice di posizione** *che non coincide in generale con un valore della variabile*, è facile da determinarsi e tiene conto di tutti i valori della serie di dati. La media ha significato se i valori sono diffusi in modo bilanciato. Non è invece un buon indice dei dati se sono presenti valori estremi, evidentemente anomali.

Esempio

16 *I consumi settimanali di pane di quattro famiglie sono (espressi in kg):*

3 4 4,2 22

Si calcoli la media aritmetica dei consumi settimanali.
La media aritmetica è un indice significativo del consumo di pane?

Si ha:

$$\mu = \frac{3+4+4,2+22}{4} = 8,3$$

Come si può osservare, in questo caso la media di 8,3 kg non è un buon indice per il consumo medio di pane delle quattro famiglie prese in considerazione, per la presenza del valore 22 che si discosta notevolmente dagli altri valori della serie.

In casi come quello esaminato nell'esempio precedente o la media aritmetica viene calcolata escludendo valori estremi (nell'esempio 22 kg) oppure si fa uso di altre medie, come la **moda** o la **mediana**, che definiremo più avanti, che non tengono conto di valori troppo diversi dalla maggioranza dei valori della serie statistica.

A volte si associano ai valori $x_1, x_2, \ldots, x_k$ certi pesi $p_1, p_2, \ldots, p_k$, secondo l'importanza che si stabilisce di attribuire ai valori stessi; in tal caso si definisce **media ponderata** il valore

$$\mu = \frac{x_1 p_1 + x_2 p_2 + \ldots + x_k p_k}{p_1 + p_2 + \ldots + p_k}$$

La formula [1] può quindi essere considerata una media ponderata con pesi $f_1, f_2, \ldots, f_k$.

Esempio

17 Un esame è composto da tre prove e le votazioni (in centesimi) riportate sono: per la prova scritta 65, per la prova orale 80 e per la prova pratica 85. Se alla prova scritta e alla prova orale si attribuiscono pesi tre volte superiori a quello attribuito alla prova pratica, la media ponderata dei voti è:

$$\mu = \frac{65 \cdot 3 + 80 \cdot 3 + 85}{3+3+1} \cong 74,28$$

DEFINIZIONE Le differenze:

$$x_1 - \mu \qquad x_2 - \mu \qquad \ldots \qquad x_n - \mu$$

tra i singoli valori $x_1, x_2, \ldots, x_n$ e il loro valore medio si chiamano **scarti** della serie di valori dalla media.

Proprietà fondamentali della media aritmetica

1 La somma degli scarti è nulla:

$$(x_1 - \mu) + (x_2 - \mu) + \ldots + (x_n - \mu) = 0$$

2 Detti min e Max il minimo e il massimo degli n valori $x_1, x_2, \ldots, x_n$, si ha:

$$\text{min} + \text{min} + \ldots + \text{min} \leq x_1 + x_2 + \ldots + x_n \leq \text{Max} + \text{Max} + \ldots + \text{Max}$$

quindi, dividendo per n membro a membro risulta:

$$\text{min} \leq \mu \leq \text{Max}$$

cioè, la media aritmetica μ è sempre un numero compreso tra il più piccolo e il più grande dei valori dati.

3 Se tutti i termini della serie subiscono un incremento (o un decremento) uguale a b anche la loro media aritmetica subisce lo stesso incremento (o decremento) b;
se tutti i termini della serie vengono moltiplicati (o divisi) per lo stesso numero a anche la loro media aritmetica risulta moltiplicata (o divisa) per a.
Pertanto, se su ciascuno degli n valori $x_1, x_2, \ldots, x_n$ si opera la trasformazione:

$$y_i = ax_i + b$$

si ha:

$$\mu_Y = a\mu + b$$

avendo indicato con μ_Y la media aritmetica degli $y_1, \ldots, y_n$.

Moda

DEFINIZIONE Si chiama **moda** degli n elementi $x_1, x_2, \ldots, x_n$ l'elemento (o gli elementi) che ha la frequenza più alta.

Esempi

18 Sia: $n = 3 \quad x_1 = 4 \quad x_2 = 5 \quad x_3 = 5$

Il valore 4 ha frequenza 1 (uno dei valori $x_1, x_2, \ldots, x_n$ vale 4);
Il valore 5 ha frequenza 2 (due dei valori $x_1, x_2, \ldots, x_n$ valgono 5);
la moda è 5.

19 Sia: $n = 4 \quad x_1 = 1 \quad x_2 = 7 \quad x_3 = 7 \quad x_4 = 1$

Il valore 1 ha frequenza 2 (due dei valori $x_1, x_2, \ldots, x_n$ valgono 1);
Il valore 7 ha frequenza 2 (due dei valori $x_1, x_2, \ldots, x_n$ valgono 7);
1 e 7 sono mode.

20 Sia: $n = 5 \quad x_1 = 2, \quad x_2 = 4, \quad x_3 = 3, \quad x_4 = 6, \quad x_5 = 10$

In questo caso la moda non esiste non essendoci alcun numero che ha frequenza maggiore degli altri.

Se i dati non sono discreti ma sono raggruppati in classi di uguale ampiezza, la classe a cui corrisponde la massima frequenza viene detta **classe modale**.

Esempio 21

A una gara, il cui punteggio massimo è 30, partecipano 35 persone. I risultati, raggruppati in classi di uguale ampiezza con le relative frequenze assolute, sono riportati nella tabella e nell'istogramma di figura 13.

Punteggio	$0 \vdash 5$	$5 \vdash 10$	$10 \vdash 15$	$15 \vdash 20$	$20 \vdash 25$	$25 \dashv 30$
Frequenza	2	3	9	12	6	5

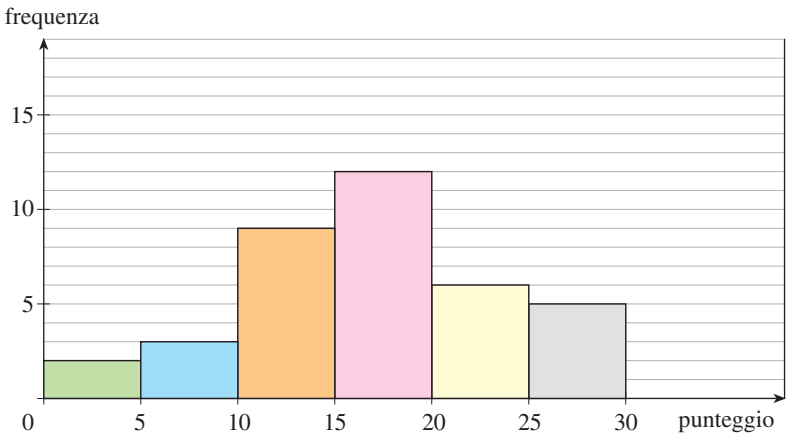


Figura 13

La *classe modale* è ovviamente la classe $15 \vdash 20$, alla quale corrisponde la frequenza massima uguale a 12.

Si può osservare che *la moda è un* **indice di posizione** *facile da trovare, il cui valore non è affetto da valori estremi e può essere usato per dati non numerici*. Per contro, la moda non tiene conto di tutti i valori e può non esistere.
La moda ha significato quando la sua frequenza è nettamente superiore alle frequenze degli altri elementi della serie. Per esempio, nel questionario proposto nell'esempio 13, la moda è rappresentata dai film comici.

Mediana

DEFINIZIONE Si chiama **mediana** di una successione di n numeri ogni valore x_m tale che i numeri della successione minori di x_m sono tanti quanti quelli maggiori di x_m.

In altri termini, rappresentati gli $x_1, x_2, \ldots, x_n$ come punti di una retta, x_m è un punto "centrale" rispetto agli $x_1, x_2, \ldots, x_n$: ne cadono tanti alla sua sinistra quanti alla sua destra.
Per determinare la mediana, consideriamo due diversi tipi di distribuzioni di dati:

a) distribuzioni i cui dati hanno tutti frequenza 1;
b) distribuzioni in cui alcuni dati hanno frequenza maggiore di 1.

Esaminiamo i due casi separatamente.

a) Supponiamo che i dati, tutti di frequenza uguale a 1, siano disposti in ordine crescente (decrescente). Allora:
- se il numero di dati è *dispari*, la mediana è il dato che occupa il posto centrale;
- se il numero dei dati è *pari* come mediana si può scegliere un valore qualsiasi compreso tra i due valori centrali della successione; in genere si sceglie come mediana la media aritmetica dei due valori centrali.

Esempi

22 $n = 3$ se $x_1 < x_2 < x_3$, la mediana è necessariamente x_2.

$n = 4$ se $x_1 < x_2 < x_3 < x_4$, la mediana è un qualsiasi numero compreso tra x_2 e x_3.

Generalmente si sceglie $x_m = \frac{x_2 + x_3}{2}$.

23 *Determinare la mediana dell'insieme dei seguenti dati:* 5, 7, 2, 10, 23, 9, 11

I dati sono in numero dispari, quindi, scritti in ordine crescente:

2, 5, 7, **9**, 10, 11, 23

si deduce che la *mediana* è **9**, essendo il valore centrale della serie.

24 *Determinare la mediana dell'insieme dei seguenti dati:* 22, 65, 13, 45, 34, 16.

Scriviamo i dati in ordine crescente:

13, 16, 22, 34, 45, 65

Poiché sono in numero pari non c'è un valore centrale, pertanto come mediana si può scegliere un valore qualunque compreso fra i due valori centrali 22 e 34; generalmente si sceglie la loro media aritmetica, quindi:

$$mediana = \frac{22 + 34}{2} = \mathbf{28}$$

b) Consideriamo, ora, una distribuzione in cui alcuni dati abbiano una frequenza maggiore di 1.

Esempio

25 *Calcolare la mediana della seguente serie di* $n = 20$ *dati:*

5 5 6 2 3 6 3 2 5 5 6 5 3 2 2 2 3 5 5 6

Si può procedere in due modi.
Il primo consiste, come abbiamo già visto, nel disporre in ordine crescente i 20 dati

2 2 2 2 2 3 3 3 3 **5** **5** 5 5 5 5 5 6 6 6 6

Poiché i dati sono in numero pari, la mediana è la media aritmetica dei due valori centrali:

$$mediana = \frac{5 + 5}{2} = \mathbf{5}$$

Il secondo metodo, utile nei casi in cui siano presenti dati con alta frequenza, consiste innanzi tutto nel costruire una tabella contenente nella prima colonna i valori dei dati in ordine crescente, nella seconda colonna le relative frequenze, nella terza le frequenze cumulate (vedi tabella a fianco).

Dati	Frequenze	Frequenze cumulate
2	5	5
3	4	9
5	7	16
6	4	20

Poiché la mediana è il valore centrale, cioè quello che lascia alla sua sinistra un numero di valori uguale a quello che lascia alla sua destra, nella tabella la mediana è il valore che corrisponde alla prima frequenza cumulata che supera 10, pari alla metà degli elementi.
Ritroviamo come *mediana* il valore **5** al quale corrisponde la frequenza cumulata $16 > 10$.

Esaminiamo ora il caso in cui i dati siano distribuiti in classi. Negli esempi che seguono mostriamo come si calcola la mediana servendosi della curva delle frequenze cumulate percentuali.

Esempi

26 Nel caso di una distribuzione in classi, come le classi di stipendio dell'esempio 11, il calcolo delle frequenze cumulate relative rende più semplice l'individuazione della classe mediana. Infatti a tale classe deve corrispondere la frequenza cumulata relativa 0,5.
Riconsideriamo la tabella dell'esempio 11.

Stipendio (in euro)	Frequenza assoluta	Frequenza cumulata assoluta	Frequenza percentuale	Frequenza cumulata percentuale
1700 ⊢ 1850	36	36	29%	29%
1850 ⊢ 2000	42	78	34%	63%
2000 ⊢ 2150	15	93	12%	75%
2150 ⊢ 2300	24	117	20%	95%
2300 ⊢ 2450	6	123	5%	100%
Totale	**123**		**100%**	

Costruiamo in un sistema di riferimento, in cui in ascisse riportiamo le classi e in ordinate le frequenze percentuali cumulate, una poligonale, detta **curva delle frequenze percentuali cumulate crescenti**, oppure **curva di ripartizione**, avente per estremi i punti (1700; 0), (1850; 29), (2000; 63), …, (2450; 100) (fig. 14).

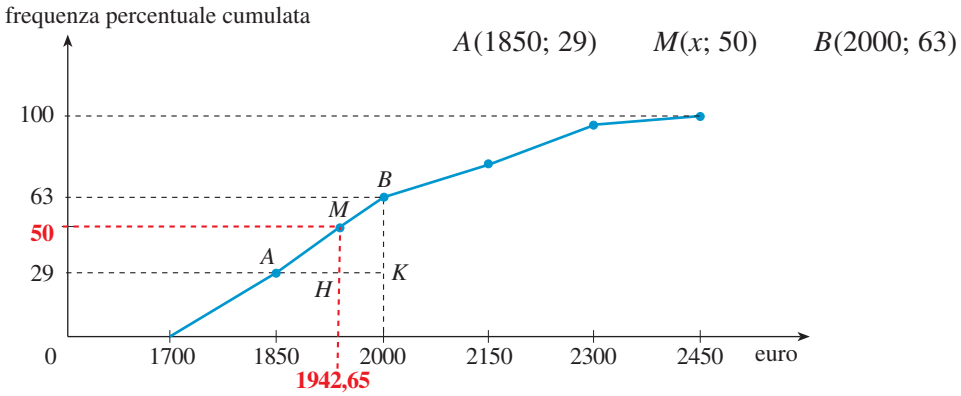


Figura 14

La mediana è l'ascissa x del punto M la cui ordinata è 50%. Calcoliamone il valore considerando i triangoli simili AMH e ABK. Si ha:

$$\frac{\overline{AH}}{\overline{AK}} = \frac{\overline{MH}}{\overline{BK}} \quad \Rightarrow \quad \frac{x-1850}{2000-1850} = \frac{50-29}{63-29}$$

da cui si ottiene la mediana:

$$mediana = 1942{,}65$$

Sulla figura 14 si può valutare un valore approssimato della mediana.

27 *Un'indagine fatta su* 40 *studenti per sapere quanti minuti impiegano ad arrivare a scuola ha prodotto la seguente tabella:*

Tempo in minuti	0 ⊢ 5	5 ⊢ 10	10 ⊢ 15	15 ⊢ 20	20 ⊢ 25
Frequenza	5	12	6	10	7
Frequenze cumulate	5	17	23	33	40

Si costruisca il grafico delle frequenze cumulate e si determini la mediana.

Il grafico delle frequenze cumulate (fig. 15), è una poligonale crescente, avente per estremi i punti:

(0; 0), (5; 5), (10; 17),

(15; 23), (20; 33), (25; 40)

Sulla figura si legge l'ascissa del punto della poligonale avente per ordinata 20, cioè la metà delle frequenze cumulate totali: la *mediana* è **12,5**, infatti la metà degli studenti impiega meno di 12,5 minuti per arrivare a scuola.

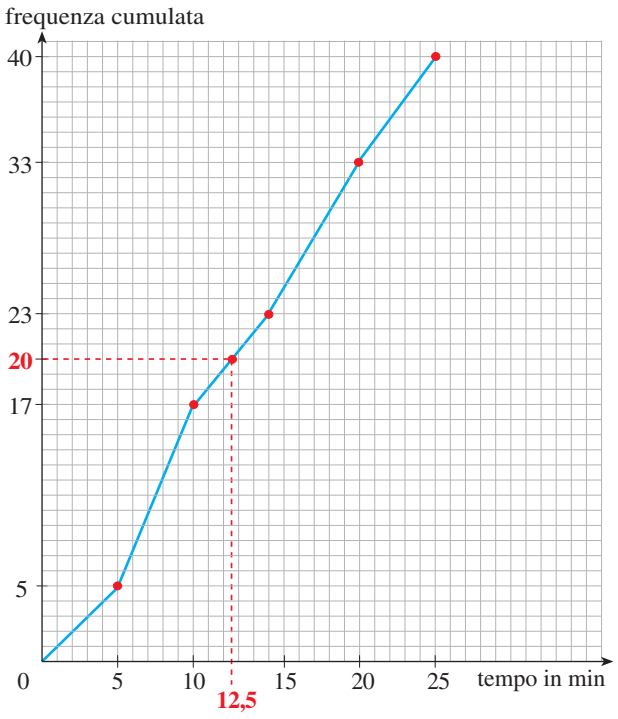


Figura 15

Osserviamo che *la mediana è un* **indice di posizione** *che non è affetto dai valori estremi*. La mediana è facile da trovare se i dati sono discreti; se invece i dati sono raggruppati in classi non è sempre ottenibile con semplici calcoli. La mediana non tiene conto di tutti i valori quindi dà maggiori informazioni nel caso di serie di dati con valori estremi.

Osservazione 5

Il termine *mediana* deriva dal latino "medium" cioè "ciò che sta in mezzo".
Non essendo influenzata dai valori estremi, essa è particolarmente utile quando i valori estremi sono in qualche modo sospetti o quando vogliamo ridurre il loro peso.

ESEMPI

1. Nel risolvere un problema proposto a una classe di 15 alunni, 14 di essi danno la soluzione entro un'ora, mentre uno di essi impiega più di tre ore, in tal caso considerata la distribuzione dei tempi, il valore di tempo più rappresentativo è la mediana.

2. Se in una fabbrica vengono installate 10 000 lampadine, la loro vita media può essere facilmente trovata annotando dopo quanto tempo esattamente la metà di esse è stata sostituita; infatti tale valore rappresenta la mediana della distribuzione dei tempi.

3. La mediana è spesso usata in alcuni tipi di ricerche mediche. Per esempio, per paragonare la potenza di differenti tipi di veleni, il ricercatore annota quale dosaggio di ciascun veleno causa la morte di esattamente la metà delle cavie.

Quartili

Rappresentati i valori $x_1, x_2, \ldots, x_n$ come punti di una retta, tre punti q_1, q_2 e q_3 si dicono **primo**, **secondo** e **terzo quartile** se un quarto degli $x_1, x_2, \ldots, x_n$ cadono a sinistra di q_1, un quarto tra q_1 e q_2, un quarto tra q_2 e q_3 e infine l'ultimo quarto a destra di q_3.
Come nel caso della mediana, la scelta dei quartili può essere obbligata in uno degli elementi x_i o libera in un certo intervallo $[x_h; x_k]$. In quest'ultimo caso si usa scegliere $\frac{x_h + x_k}{2}$.

Esempi

28 Se $n = 100$ e $x_1 = 1, x_2 = 2, \ldots, x_{99} = 99, x_{100} = 100$ sarà $q_1 = 25$, $q_2 = 50$, $q_3 = 75$

29 *Il punteggio massimo di un concorso è di* 50 *punti prevedendo votazioni che comprendano anche i decimi, cioè del tipo:* 25,7, 41,8...
I voti ricevuti dai 30 *concorrenti vengono ripartiti in classi e riportati nella seguente tabella.*

Voti	0 ⊢ 10	10 ⊢ 20	20 ⊢ 30	30 ⊢ 40	40 ⊢ 50
Frequenze	2	6	12	6	4
Frequenze cumulate	2	8	20	26	30

Si determinino la mediana e i quartili.

Costruiamo il poligono delle frequenze cumulate (fig. 16) e consideriamo sull'asse delle ordinate i punti di ordinata 7,5, 15 e 22,5, che dividono l'intervallo [0; 30] in quattro parti uguali.

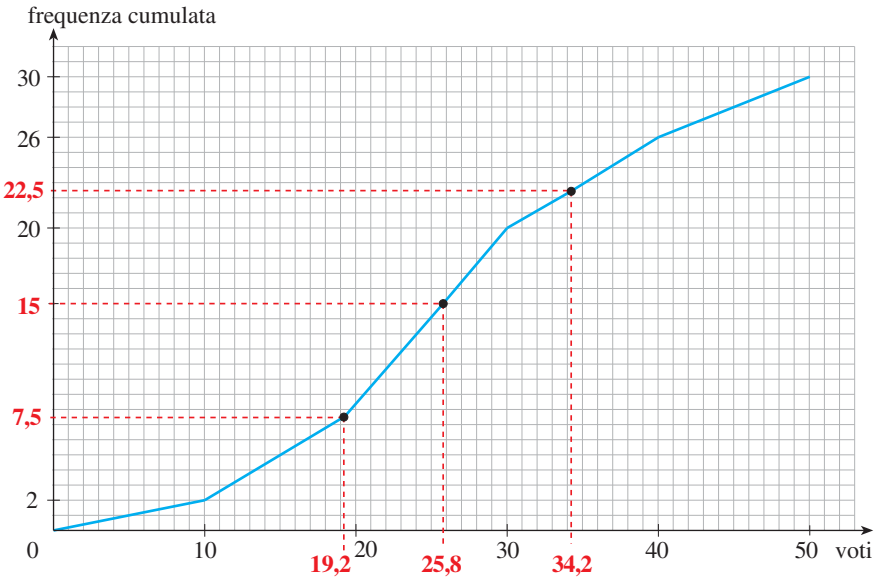


Figura 16

Tracciando per tali punti le parallele all'asse x fino a incontrare la poligonale, otteniamo tre punti aventi per ascisse rispettivamente

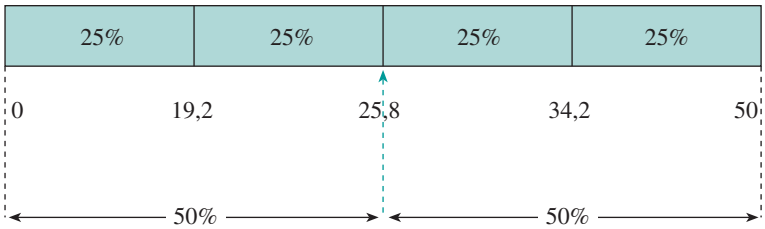
19,2 25,8 34,2

I valori vengono così divisi in quattro classi contenenti lo stesso numero di elementi, cioè il 25% del totale.
I valori che operano questa suddivisione si dicono *quartili*:

19,2 è il *primo quartile*: il 25% dei concorrenti ha ricevuto un voto inferiore a 19,2;
25,8 è il *secondo quartile*: il 50% dei concorrenti ha ricevuto un voto inferiore a 25,8;
34,2 è il *terzo quartile*: il 75% dei concorrenti ha ricevuto un voto inferiore a 34,2.

Ovviamente **25,8** è la *mediana*, cioè il valore che divide la popolazione in due parti:

il 50% riceve un voto inferiore a 25,8, l'altra metà un voto superiore a 25,8.



30 Riprendiamo l'esempio 27 con il relativo grafico e riportiamo sull'asse delle ordinate i punti di ordinata 10 e 30 corrispondenti a un quarto e a tre quarti degli effettivi.
Conducendo da essi le parallele all'asse delle ascisse fino ad incontrare la poligonale, si ottengono due punti le cui ascisse 7 e 18,5 si possono leggere sul grafico (fig. 17).

Pertanto si ha:

7 *primo quartile*;
12,5 *secondo quartile* (mediana);
18,5 *terzo quartile.*

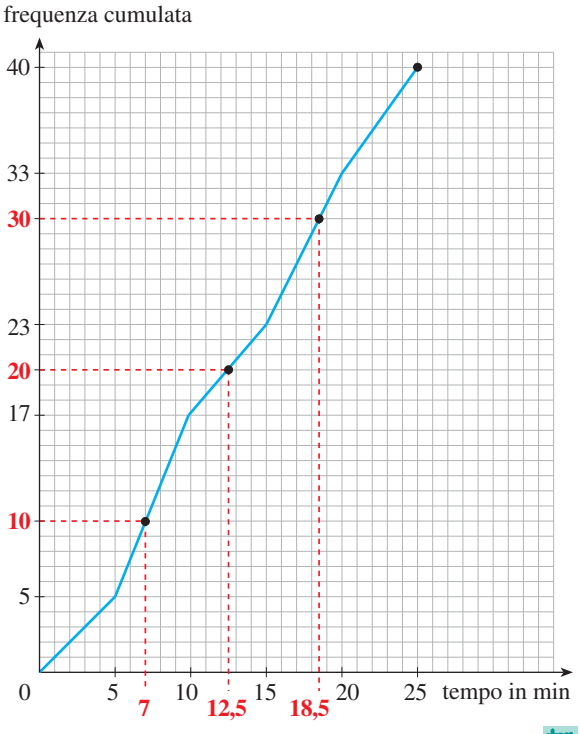


Figura 17

5 Indici di dispersione

Nello studio di dati statistici non soltanto è utile determinare un valore medio, ma è anche importante essere in grado di valutare la variabilità, detta anche **dispersione**, delle misure.

Esempio

31 Le distribuzioni delle età dei partecipanti a due diverse crociere hanno la stessa media aritmetica, ma mentre nella prima i partecipanti hanno età compresa tra 20 e 30 anni, nella seconda partecipano soprattutto famiglie con bambini accompagnati dai nonni. È evidente che un indice che dia una misura di tale differenza sarà utile per indirizzare i partecipanti all'una o all'altra crociera.

Vi sono quattro modi principali per descrivere la variabilità di una serie di dati:

- il *range* o *campo di variazione*;
- lo *scarto semplice medio*;
- lo *scarto quadratico medio*;
- lo *scarto interquartile*.

Range o campo di variazione

Si definisce **range** o **campo di variazione** di una statistica $x_1, x_2, \ldots, x_n$ la differenza tra il max e il min dei valori, cioè il numero:

$$d = \text{Max}\ \{x_1, x_2, \ldots, x_n\} - \text{min}\ \{x_1, x_2, \ldots, x_n\}$$

Esempi

32 Sia:

$$\{x_1, x_2, \ldots, x_n\} = \{35, 11, 35, 37, 34, 34, 36\}$$

Allora:

$$\text{Max} = 37 \qquad \text{min} = 11 \qquad d = 37 - 11 = 26$$

33 Riprendendo l'esempio 31, la prima crociera è caratterizzata da un range che è prossimo a 10, mentre la seconda ha un range il cui valore è certamente molto più alto.

Scarto semplice medio

Assegnata la statistica $\{x_1; x_2; \ldots; x_n\}$, di media aritmetica μ, considerati i valori assoluti degli scarti dalla media:

$$|x_1 - \mu|,\ |x_2 - \mu|,\ \ldots,\ |x_n - \mu|$$

si definisce **scarto semplice medio** il numero non negativo

$$S_1 = \frac{|x_1 - \mu| + |x_2 - \mu| + \ldots + |x_n - \mu|}{n}$$

che rappresenta la media aritmetica dei valori assoluti degli scarti.

Riportando sull'asse delle ascisse i valori x_n e la media μ della distribuzione, ciascun valore $|x_n - \mu|$ rappresenta la distanza del punto P di ascissa x_n dal punto M di ascissa μ.

Esempio

34 *Le età dei cinque membri della famiglia Rossi sono*:

8, 10, 12, 32, 38

Calcolare la media aritmetica e lo scarto semplice medio.

Si ha:

$$\mu = \frac{8+10+12+32+38}{5} = 20$$

quindi 20 anni è l'età media della famiglia.
Calcoliamo gli scarti dalla media:

$8 - 20 = -12$ $\quad 10 - 20 = -10$ $\quad 12 - 20 = -8$ $\quad 32 - 20 = 12$ $\quad 38 - 20 = 18$

i cui valori assoluti sono: 12, 10, 8, 12, 18 e rappresentano le lunghezze dei segmenti orizzontali in colore in figura 18.

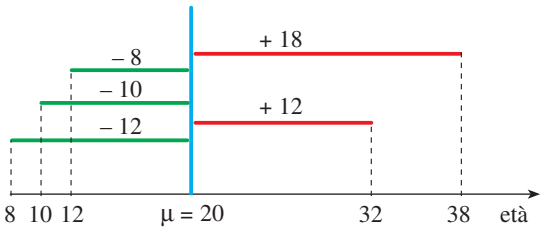


Figura 18

La loro media aritmetica

$$S_1 = \frac{12+10+8+12+18}{5} = 12$$

è lo scarto semplice medio e rappresenta la distanza media dei valori dalla media 20.

Scarto quadratico medio

Assegnata la statistica $\{x_1; x_2; \ldots; x_n\}$, di media aritmetica μ, si definisce **scarto quadratico medio** il numero non negativo

$$\sigma = \sqrt{\frac{(x_1-\mu)^2 + (x_2-\mu)^2 + \ldots + (x_n-\mu)^2}{n}}$$

detto anche **deviazione standard**. Il quadrato di tale numero

$$\sigma^2 = \frac{(x_1-\mu)^2 + (x_2-\mu)^2 + \ldots + (x_n-\mu)^2}{n}$$

è detto **varianza**, e rappresenta la media dei quadrati degli scarti.

Esempio 35

Nel caso della famiglia Rossi dell'esempio precedente, si può preparare la tabella a fianco.

Età	μ	$x - \mu$	$(x - \mu)^2$
8	20	−12	144
10	20	−10	100
12	20	−8	64
32	20	12	144
38	20	18	324
			776

Pertanto si ha che la varianza è data da:

$$\sigma^2 = \frac{776}{5} = 155,2$$

e lo scarto quadratico medio è:

$$\sigma = \sqrt{155,2} \cong 12,46$$

Osservazione 6

In fisica le operazioni di misura di una grandezza vengono ripetute di solito più volte. Infatti, per quanto effettuate con metodi esatti, con strumenti appropriati e con grande cura, sono inevitabilmente affette da errori per varie cause: errori dovuti alla sensibilità dello strumento, errori casuali che possono produrre valori per difetto o per eccesso.
Per esempio, supponiamo di voler misurare il periodo T' di un pendolo, cioè il tempo che occorre perché compia un'oscillazione completa, mediante un cronometro contasecondi che supponiamo esatto.
Iniziamo l'esperimento: facciamo compiere al pendolo 10 oscillazioni complete, misuriamo il tempo T' che intercorre tra l'inizio della prima oscillazione e la fine dell'ultima. È chiaro che ci sarà un errore piccolo ma inevitabile nell'istante in cui si fa partire il cronometro all'inizio della prima oscillazione e analogamente quando si deve arrestare alla fine dell'ultima oscillazione.
Per ogni misura di T' troviamo il corrispondente valore

$$T = \frac{T'}{10}$$

Ripetiamo l'esperimento più volte, determinando la serie di valori

$$T_1 \qquad T_2 \qquad \ldots \qquad T_n$$

La loro media aritmetica

$$T^* = \frac{T_1 + T_2 + \ldots + T_n}{n}$$

è il valore più probabile della misura del periodo del pendolo. Si può assumere come valore dell'errore di cui è affetta la misura T^* il numero, detto **semidispersione**:

$$d = \frac{T_{max} - T_{min}}{2}$$

dove T_{max} e T_{min} sono rispettivamente il più grande e il più piccolo numero della serie di valori considerata.
La misura T del periodo viene indicata con il simbolo

$$T = T^* \pm d$$

che sta a indicare che il valore di T è presumibilmente compreso tra $(T^* - d)$ e $(T^* + d)$.

6 I caratteri in un testo

Le diverse lingue si differenziano tra loro per i vocaboli, per la sintassi, per la diversa musicalità che producono. Si servono tuttavia tutte (escluse naturalmente l'arabo, il cinese, il giapponese ecc.) degli stessi caratteri, anche se letti con suoni diversi.
Un'analisi statistica interessante è l'esame delle frequenze relative dei vari caratteri alfabetici in un testo di media lunghezza:

- percentuale di "a",
- percentuale di "b",
- ecc.

L'idea naturale, ma ingenua, che tali percentuali dipendano dal testo considerato è sbagliata: la risposta invece, sorprendente, è che tali percentuali dipendono dalla lingua (italiano, inglese ecc.) usata nel testo.
L'istogramma delle frequenze relative con le quali i vari caratteri compaiono in un testo è una caratteristica, una sorta di DNA della lingua!
Riportiamo la distribuzione dei caratteri relativa a due classici italiani, il primo capitolo dei *Promessi Sposi* di Alessandro Manzoni e il *Discorso sopra lo stato presente dei costumi degl'Italiani* di Giacomo Leopardi (fig. 19), e a un testo attuale di un quotidiano (fig. 20).
I tre istogrammi sono pressoché identici, pur riferiti a testi diversi per stile, per contenuto, per autori.
Riportiamo poi l'istogramma dei caratteri relativo a un testo inglese attuale, *What is Mathematics* di R. Courant e H. Robbins (fig. 21) e quello di un testo in lingua francese, CABRI *Géomètre*, un'introduzione a CABRI (fig. 22).

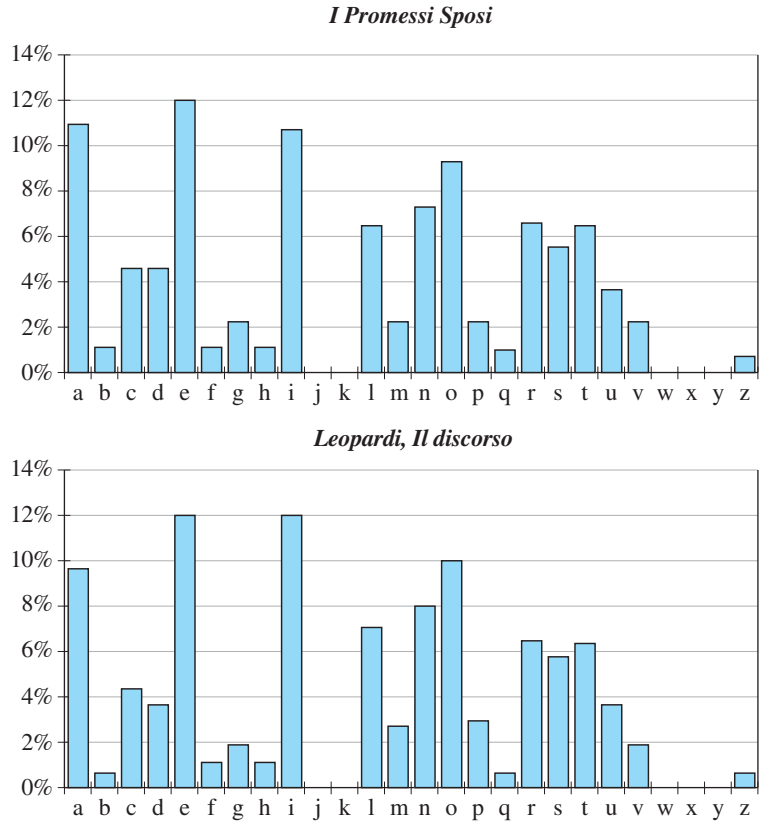


Figura 19

Un testo attuale

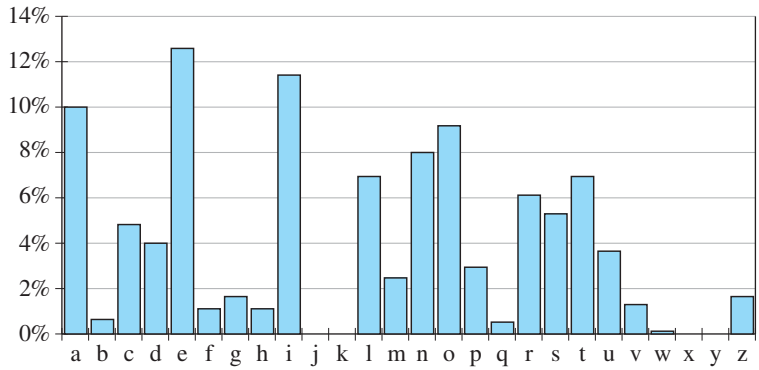


Figura 20

What is Mathematics

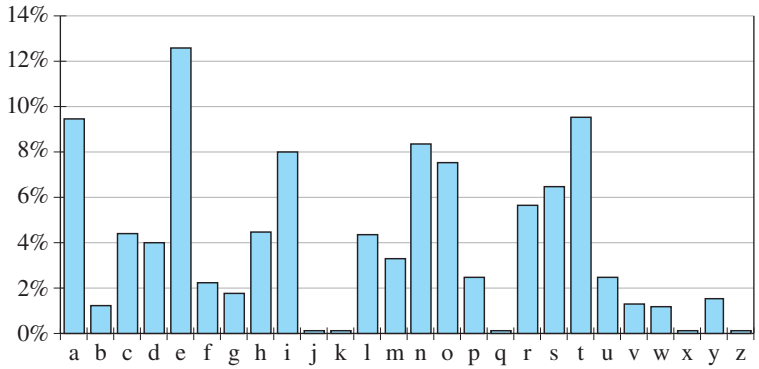


Figura 21

CABRI Géomètre

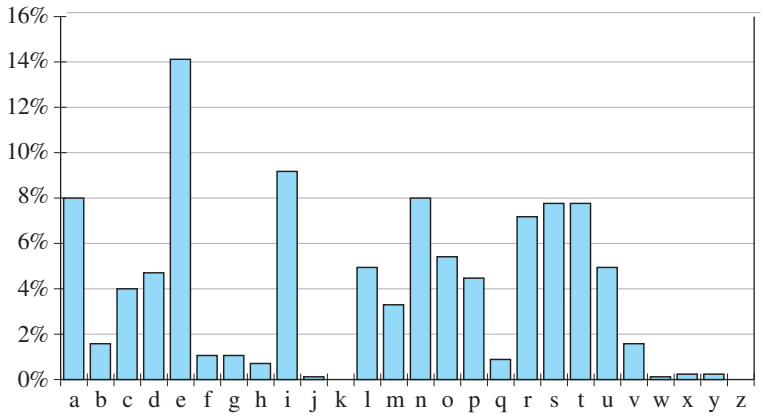


Figura 22

I due istogrammi delle figure 21 e 22 sono diversi tra loro e diversi dall'istogramma di figura 20 del testo italiano.
Somiglianze ce ne sono:

- la lettera "e" è innegabilmente la più usata in tutte e tre le lingue;
- sulla seconda lettera già appaiono differenze: la "i" in italiano e in francese, la "a" in inglese;
- al terzo posto in italiano e in francese si ha una vocale, in inglese si ha la "t";
- ecc.

Un'osservazione…

Supponiamo che un burlone abbia manipolato una stampante in modo che tutte le volte che dovrebbe scrivere "a" scriva per esempio "c", tutte le volte che dovrebbe scrivere "b" scriva "s" e così via…
Le stampe prodotte sarebbero veramente illeggibili!
Ma… se sapessimo almeno in che lingua sono scritte, forse…
Immaginiamo che cosa potremmo fare, sapendo che si tratta per esempio, di un testo in italiano:

- calcoliamo le frequenze con cui le varie lettere si incontrano sul foglio uscito dalla stampante impazzita;
- il carattere che ha la maggiore frequenza deve essere quello che corrisponde alla "e";
- quello con la frequenza immediatamente successiva deve corrispondere alla "i";
- ecc.

Una volta riconosciute le corrispondenze il testo della stampante pazza può essere tranquillamente… decifrato!

Crittografia

Il termine *crittografia* si riferisce alle tecniche di codificare e decodificare un messaggio in modo da rendere sicuro da manipolazioni il trasferimento.
Nel linguaggio della crittografia i messaggi comuni, non protetti da alcun artificio, sono detti *in chiaro*, la trasformazione di un messaggio in chiaro in uno protetto si dice *cifratura*, e il messaggio così ottenuto si dice *cifrato*. L'operazione opposta, ricavare da un messaggio cifrato il corrispondente in chiaro si dice *decifratura*.
Cifrare e *decifrare* sono i due verbi dei crittografi.

Il metodo di cifratura più semplice e ingenuo è la pura sostituzione di ciascuna lettera con una diversa ad essa corrispondente secondo un accordo tra i due interlocutori.
Lo scambio concordato potrebbe essere:

- ad "a" sostituire "b";
- a "b" sostituire "c";
- ecc.

come suggerisce la seguente tabellina

A	B	C	D	E	F	G	H	I	L	M	N	O	P	Q	R	S	T	U	V	Z
B	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*L*	*M*	*N*	*O*	*P*	*Q*	*R*	*S*	*T*	*U*	*V*	*Z*	*A*

Il messaggio in chiaro

"Vediamoci a Roma"

verrebbe cifrato in

"Zfelbnpdl b Spnb"

e naturalmente decifrato con il procedimento opposto.

Il metodo è da considerarsi ingenuo: un crittografo esperto, dopo aver esaminato un certo numero di messaggi cifrati e calcolate le frequenze relative con cui i vari caratteri figurano, riconoscerà agevolmente la corrispondenza tra lettere che c'eravamo illusi di tenere segreta…
Dopo di ciò la decifrazione dei cifrati diventa un gioco da ragazzi… anche se non si deve trascurare il tempo che un'operazione di decifratura può richiedere!
Tutti i sistemi di crittografia sono decifrabili da un bravo crittografo: solo che la decifratura può richiedere tanto tempo da divenire poco interessante…!
Il pregio degli attuali metodi crittografici consiste nei tempi inaccettabilmente lunghi per la decifratura.