Le statistiche

Alcuni modelli per la rappresentazione dei dati

Scheda **3**

Lo sviluppo corporeo

0. Introduzione

«Giovanni è basso», «Maria è troppo alta», … . A volte sono semplici osservazioni, altre volte sono giudizi un po' maligni. Ma … che cosa vuol dire "basso", che cosa vuol dire "alta"? In base a quale valutazione riusciamo a distinguere quando una persona è alta, bassa o di altezza normale?

Sicuramente siamo in grado di esprimere con un numero l'altezza di una persona («Giovanni è alto 155 cm»). C'è un modello matematico che ci permetta di stabilire quando l'altezza di una persona è normale?

Non si può rispondere nettamente con un "sì" o con un "no". Possiamo tuttavia affermare che la matematica ci permette di affrontare la questione e di metterne in luce la complessità. *Questa scheda* sarà dedicata a questo argomento.

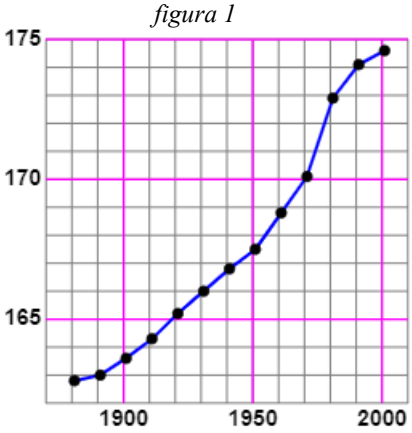
Vedremo che non può esistere una definizione assoluta di "normalità" ma che esistono degli strumenti matematici per valutare la relazione tra l'altezza di una persona e quella del complesso delle altre persone e, più in generale, per valutare la relazione tra un particolare aspetto di un certo oggetto (ad esempio il peso di un uovo) e il modo in cui tale aspetto si manifesta nella collettività di cui quell'oggetto fa parte (ad esempio il complesso delle uova prodotte dall'allevamento da cui l'uovo considerato proviene).

1. Istogrammi di distribuzione

Abbiamo già visto (→ scheda 1, §5) che per rappresentare con un unico numero come si manifesta un fenomeno collettivo si può ricorrere alla media aritmetica dei dati relativi ai singoli soggetti.

Il grafico di *figura 1* riporta la *altezza media dei maschi ventenni italiani* in vari anni, nel corso di più di un secolo.

figura 1



1 Nel 1881 l'altezza media dei maschi ventenni era di 162.8 cm, nel 1981 era di 172.9 cm. Qual è stato l'aumento medio annuo in questo intervallo di tempo? ..... mm/anno

In cent'anni l'altezza media è aumentata più di 10 cm. La crescita è stata particolarmente rapida negli anni 70, cioè per gli uomini nati negli anni 50 e che hanno trascorso la loro infanzia negli anni della ripresa e dello sviluppo economico che sono seguiti alla seconda guerra mondiale (dal 1971 al 1981 vi è stato un aumento medio di 2.7 mm/anno). Negli ultimi anni la crescita tende a rallentare; probabilmente si stabilizzerà, centimetro più centimetro meno, intorno ai 175 cm. Un fenomeno analogo (forte crescita nel XX secolo, con rallentamento negli ultimi decenni) si è verificato in tutti i paesi industrializzati, anche tra le donne.

L'aumento dell'altezza media è dovuto essenzialmente al miglioramento delle condizioni di vita, soprattutto nell'alimentazione (per ricordare alcuni dati, nel 1880 l'"italiano medio" ha consumato 15 kg di carne, 29 nel 1960 e 54 nel 1970), ma anche nell'assistenza sanitaria e nell'attività fisica (si pensi all'elevamento dell'obbligo scolastico e alla progressiva riduzione del fenomeno del lavoro minorile): questi miglioramenti hanno fatto sì che i bambini e gli adolescenti abbiano avuto sempre più modo di sfruttare al massimo le potenzialità di crescita presenti nel patrimonio genetico ereditato dai genitori. Il miglioramento nell'assistenza sanitaria ha inciso su questo aumento anche in altri modi; ad es. le donne longilinee un tempo incontravano più difficoltà nel parto e quindi mediamente avevano meno figli; pian piano questo "svantaggio" è stato colmato ed è aumentata la trasmissione del patrimonio genetico da parte delle donne più alte.

2 Abbiamo dunque visto un primo aspetto che rende *relativo* il significato di "essere basso": l'altezza media è variata nel tempo. Un maschio nato nel 1941 (cioè ventenne nel '61) e alto 165 cm di quanto è sotto all'altezza media dei suoi coetanei? E un maschio della stessa altezza nato nel 1971?

Ma non basta calcolare la distanza dell'altezza di una persona dall'altezza media. Bisogna anche vedere se, ad esempio, sono molte o sono poche le persone nate nel 1971 e con altezza inferiore di 9 o più centimetri rispetto all'altezza media. Per fare valutazioni di questo tipo possiamo riferirci agli istogrammi della *figura 2*, che rappresentano le percentuali dei ventenni maschi le cui altezze cadono in alcuni intervalli di misure.

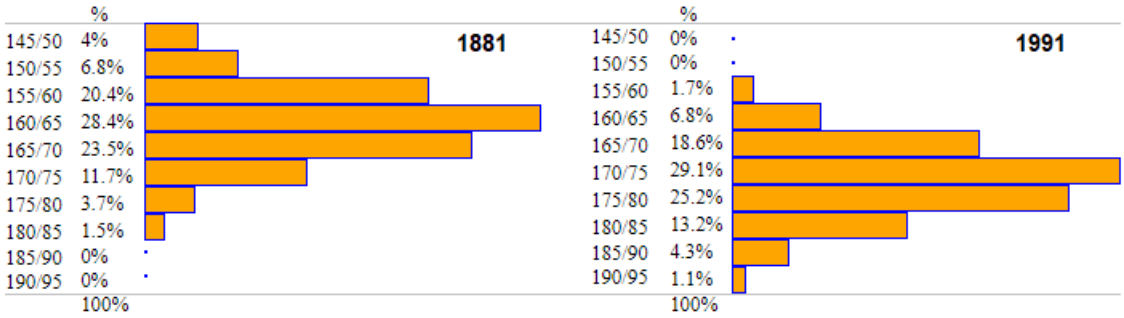


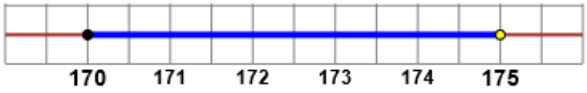
figura 2

Questi istogrammi sono stati realizzati impiegando dati pubblicati dall'Istat e già classificati negli intervalli di altezza indicati

Gli istogrammi man mano si sono spostati verso le altezze maggiori, ma hanno mantenuto più o meno la stessa forma. Ciò visualizza il fatto che le diversità genetiche all'interno della popolazione si sono mantenute e che il miglioramento delle condizioni di vita ha fatto sì che tutti, ciascuno con le potenzialità ereditate, sviluppassero maggiormente l'altezza.

3 L'intervallo di altezze più *frequente* (cioè in cui cade la maggiore percentuale di misure di altezza) nel 1881 è quello da 160 a 165 cm. Qual è quello nel 1991?

In figura 2 tra 170 e 175 cm ho classificato le misure i cui valori troncati ai centimetri sono 170, 171, 172, 173 o 174, cioè le misure che vanno da 170.0… cm a 174.9… cm. Nel disegno a fianco sono i valori che cadono tra i due pallini, cioè i valori maggiori o uguali a 170.000… e minori di 175.000….



Quando di un ***intervallo*** di valori numerici si vogliono descrivere esattamente gli estremi si usano scritture come la seguente: [170,175). Essa indica l'insieme dei numeri che sono maggiori o uguali a 170 e che sono minori di 175; cioè l'insieme dei numeri x tali che $170 \le x < 175$.

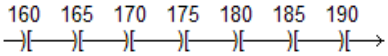
Si usa anche la scrittura: 170 175; il pallino pieno [vuoto] indica che l'estremo è [non è] compreso.

Nel caso in cui avessimo voluto includere 175 avremmo scritto [170,175] o 170 175.

4 (**a**) Come rappresenteresti l'insieme dei numeri x tali che $170 < x \le 175$?
(**b**) e l'insieme dei numeri x tali che $14 < x < 17$?
(**c**) Come completeresti questa frase "l'insieme dei numeri x tali che …" in modo da descrivere l'intervallo rappresentabile con [4.1,4.3]?
(**d**) e in modo da descrivere l'intervallo rappresentabile con (4.1,4.3]?
(**e**) Se 48 cm è la lunghezza *arrotondata* ai centimetri di un oggetto, in quale tra i seguenti intervalli puoi concludere che cade la lunghezza "esatta"?
(48, 49] (47.5, 48.5) [48, 49) [47.5, 48.5) (47, 48]

Tornando a ➡ figura 2, *come sono state ottenute le percentuali rappresentate mediante gli istogrammi*?

Le altezze dei ventenni sono state classificate negli intervalli raffigurati a fianco. Il termine ***classificare*** in questo caso non significa "mettere in graduatoria, assegnare un posto della classifica", ma significa "ripartire in *classi* (cioè collezioni, insiemi, aggregati, …) opportunamente definite". Le classi in cui vengono distribuiti i dati vengono spesso chiamate anche ***modalità***.



Per fare un altro esempio, se si volesse fare una statistica sul quartiere di provenienza degli alunni di una scuola, le modalità sarebbero i vari quartieri.

Il numero delle altezze che cade in un certo intervallo viene chiamato frequenza di quell'intervallo. Nel caso dell'indagine sulla provenienza degli alunni la frequenza di un quartiere è il numero degli alunni che proviene da esso. Più in generale, se considero un certo insieme di "oggetti" (ventenni, alunni di una scuola, …) e per ciascuno di essi raccolgo una particolare informazione (altezza, quartiere di provenienza, …), la ***frequenza*** di una modalità è il numero delle informazioni che vengono classificate in quella modalità o, in altre parole, è il numero delle volte che quella modalità si manifesta.

alunno	*sport*	*alunno*	*sport*
Anna	tennis	Giorgio	pallacanestro
Barbara	nessuno	Irene	salto in alto
Bruno	calcio	Laura	pallavolo
Carlo	ping-pong	Luciano	nessuno
Clara	calcio	Manuela	tennis
Dario	nuoto	Nicola	calcio
Davide	pallavolo	Paola	nessuno
Elena	nessuno	Roberta	judo
Enrico	judo	Sabina	pallacanestro
Fabrizio	nessuno	Valerio	pallanuoto

5 Nella tabella a fianco per ogni alunno è indicato lo sport maggiormente praticato. Classifica queste informazioni secondo le quattro modalità indicate nella tabella sotto a sinistra: in ogni casella scrivi (in piccola dimensione) i nomi degli alunni che verificano sia la proprietà "orizzontale" che la proprietà "verticale".
Indica, quindi, le corrispondenti frequenze nella tabella a destra, calcolando anche i totali per riga e per colonna.

Classificazione:

	fare uno sport praticabile in squadra	*non* fare uno sport praticabile in squadra

Frequenze:

	fare …	*non* …	*totale*

fare uno sport praticabile individualmente		
non fare uno sport praticabile individualmente		

fare …			
non …			
totale			20

Dopo aver classificato i dati e stabilito la frequenza delle varie modalità, per calcolare le percentuali rappresentate in istogrammi come quelli di figura 2, ogni frequenza viene divisa per il numero totale dei dati.

Nel caso di ➡ figura 2 la frequenza di ogni intervallo è stata divisa per il numero totale dei ventenni ed espressa in forma percentuale.

Un rapporto di questo genere, cioè il rapporto tra la frequenza di una modalità e il numero totale delle informazioni classificate, viene chiamato ***frequenza relativa***; infatti non esprime direttamente il numero delle volte con cui la modalità si è verificata ma lo "relativizza", ne esprime la relazione quantitativa con il totale delle informazioni classificate.

Quando la frequenza relativa è espressa in forma percentuale viene chiamata anche *frequenza percentuale*.

Nel caso della provenienza degli alunni dire che per il quartiere X si è ottenuta una frequenza relativa del 29% significa che il rapporto tra gli alunni provenienti da X e il totale degli alunni è 0.29.

Per meglio distinguerla dalla frequenza relativa, la frequenza (non relativizzata) viene spesso chiamata ***frequenza assoluta***.

frequenza assoluta di una modalità = quantità delle informazioni classificate in tale modalità

$$\textit{frequenza relativa}\text{ di una modalità} = \frac{\text{frequenza assoluta di tale modalità}}{\text{totale delle informazioni classificate}}$$

6 (**a**) Qual è la frequenza relativa della modalità "fare uno sport praticabile sia in squadra che individualmente" di cui al ➡ quesito 5? (esprimila in forma percentuale)
(**b**) Qual è la frequenza relativa dell'intervallo di altezze (in cm) [165,170) nel 1991 (➡ fig. 2)?

Una tabella che associ ad ogni modalità le corrispondenti frequenze con cui si manifesta un certo fenomeno viene detta ***distribuzione di frequenza*** (o più semplicemente *distribuzione*) di quel fenomeno (rispetto alle modalità scelte).

Ad esempio la *tabella (1.1)* è la distribuzione di frequenza degli sport praticati dagli alunni del quesito 5 rispetto alle modalità indicate (*I* sta per "praticabile individualmente", *S* sta per "praticabile a squadra").

La *tabella (1.2)* è la distribuzione di frequenza delle altezze degli italiani maschi ventenni nel 1991 rispetto agli intervalli indicati. Per essere più precisi in questo caso dovremmo parlare di *distribuzione di frequenza relativa* o di *distribuzione percentuale*.

Gli istogrammi di figura 2 vengono quindi chiamati ***istogrammi di distribuzione*** (*percentuale*).

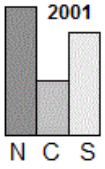
(**1.1**)	sport che è sia I che S	sport che è I ma non S	sport che è S ma non I	nessuno sport
frequenza	4	3	8	5

(**1.2**)	[0,160)	[160,165)	[165,170)	[170,175)	[175,180)	[180,185)	[185,190)	[190,∞)
freq. relativa	2%	7%	18%	29%	25%	13%	5%	1%

Il simbolo "∞" (che si legge "*infinito*") impiegato per l'ultimo intervallo indica una quantità infinita, cioè [190,∞) rappresenta l'intervallo costituito da tutti i numeri maggiori o uguali a 190.

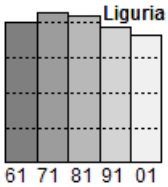
Anche gli istogrammi relativi ai consumi impiegati nella scheda 1 sono istogrammi di distribuzione: gli "oggetti" sono le lire o euro spesi in consumi, le "informazioni" sono i beni o i servizi per cui le varie lire sono state spese, le modalità sono le categorie di beni e di servizi considerate.

Si parla di istogrammi di ripartizione (o distribuzione) assoluta dei consumi se (sulla scala orizzontale come in quelli di figura 2, o verticale, come nei seguenti) sono rappresentati i dati assoluti , di istogrammi di ripartizione percentuale se sono rappresentate le percentuali.

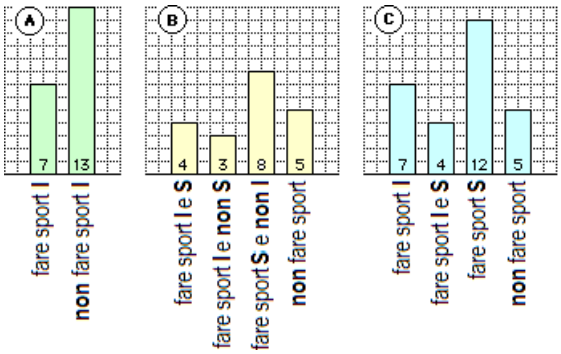


Gli istogrammi possono essere usati per visualizzare il confronto tra due o più quantità, ma non sempre si tratta di istogrammi di distribuzione.

A differenza del caso raffigurato a sinistra (ripartizione della popolazione italiana nelle tre zone geografiche), a destra (popolazione ligure in vari anni) non siamo di fronte a un istogramma di distribuzione: i rettangoli non rappresentano le parti che compongono un totale (un abitante conteggiato nel 2001 può essere stato conteggiato anche nel 1991, nel 1981, …).



7 Quali (o quale) dei tre istogrammi a fianco (➡ quesito 5, ➡ tabella 1.1) sono istogrammi di distribuzione, cioè in quali casi l'area complessiva dei rettangoli rappresenta un *totale* e le aree dei vari rettangoli rappresentano parti disgiunte (= "senza elementi in comune") del totale?



2. Media, moda, mediana

Tra la distribuzione delle altezze degli italiani (figura 2) e quella delle zone di provenienza degli alunni di una classe vi è una diversità di fondo. In un caso abbiamo *modalità di tipo numerico* (valori numerici che vengono classificati in intervalli di numeri), nell'altro no (località che vengono classificate in quartieri).

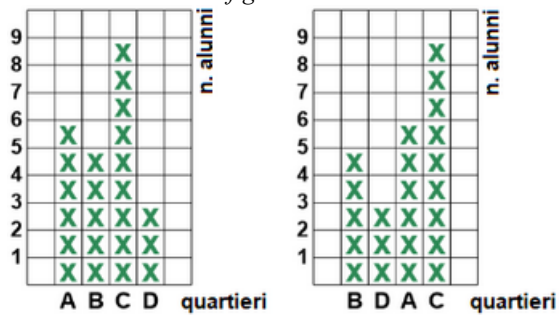
Nel primo caso quindi sull'istogramma le modalità devono essere rappresentate con un certo ordine, nel secondo caso l'ordine non è particolarmente significativo: i due istogrammi di distribuzione di frequenza assoluta della *figura 3* possono essere considerati equivalenti.

Inoltre, mentre nel primo caso ha senso parlare di media aritmetica dei dati, nel secondo non ha senso parlare di quartiere medio di provenienza.

In entrambi i casi si può considerare la modalità più frequente. Essa viene detta ***moda*** o ***classe modale***.

Nel caso dei quartieri di provenienza la moda è il quartiere C. Nel caso delle altezze abbiamo già individuato le classi modali nel ➡ quesito 3.

figura 3



8 Nel caso della distribuzione rappresentata dalla ➡ tabella (1.1) trova, se è possibile, la moda e la media aritmetica.

Nelle *situazioni*, come quella delle altezze, *in cui le modalità sono numeri o intervalli numerici*, la moda indica un *valore medio*, così come la media aritmetica, cioè un valore (o un intervallo di valori) che riassume, caratterizza quantitativamente il modo complessivo in cui si è manifestato il fenomeno in questione. Ad esempio per il 1991 possiamo dire (esprimendosi in cm) sia che l'altezza media dei ventenni era di 174.1, sia che la classe modale è [170,175) (➡ figura 2).

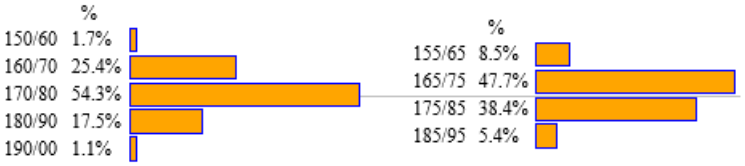


figura 4

A differenza della media, la *moda* (e più in generale la forma dell'istogramma) non *dipende* solo dai dati ma anche dalla *scelta degli intervalli* in cui classificare i dati. Ad esempio in *figura 4* sono riportati due istogrammi della distribuzione percentuale delle altezze dei ventenni nel 1991 alternativi a quello della figura 2.

9 Nella *tabella (2.1)* sono riportate le altezze (arrotondate ai cm) delle 19 alunne diciassettenni di una scuola. I dati sono riportati secondo l'ordine alfabetico dei nomi delle alunne (al posto dei nomi delle alunne abbiamo indicato il numero d'ordine).
Qual è la moda se si prendono come modalità direttamente le misure in centimetri (cioè i valori: … , 150, 151, … , 169, 170, …)?
Qual è prendendo come modalità gli intervalli: 150-154, 155-159, … ? [osserva la tabella a destra in cui sono stati raccolti i dati]
La moda è sempre unica?

(**2.1**)

1	156	6	157	11	157	16	160
2	168	7	170	12	165	17	163
3	162	8	157	13	163	18	162
4	150	9	159	14	165	19	155
5	167	10	164	15	166		

```
15 | 0
15 | 677975
16 | 243032
16 | 87556
17 | 0
```

La tabella a destra ci dà immediatamente un piccolo istogramma. Essa è costruibile molto facilmente: basta scrivere via via (meglio se su carta quadrettata), riportando solo la cifra finale, i dati nello stesso ordine con cui sono scritti, mettendoli nelle riga giusta (150-154, 155-159, 160-164, …):

```
15 |      15 |      15 |      15 | 0    15 | 0    15 | 0
15 | 6    15 | 6    15 | 6    15 | 6    15 | 6    15 | 67
16 |      16 |      16 | 2    16 | 2    16 | 2    16 | 2   …
16 |      16 | 8    16 | 8    16 | 8    16 | 87   16 | 87
17 |      17 |      17 |      17 |      17 |      17 |
```

Nota. Questo tipo, comodissimo, di rappresentazione ha un nome: "**stem-and-leaf plot**". Per ogni riga, le cifre indicate nella prima colonna sono il "gambo" (stem) del singolo dato, la cifra rimanente (dopo "|") è la "foglia" (leaf); ci possono essere più foglie (dati classificati nella stessa classe) ma un gambo può anche essere spoglio.

10 I dati relativi alle altezze delle alunne del quesito 9 (156, 168, 162, 150, 167, 157, 170, 157, 159, 164, 157, 165, 163, 165, 166, 160, 163, 162, 155) possono essere elaborati con gli script "**grande CT**" e "**ordina**" ottenendo i seguenti esiti:

```
n = 19
min = 150
```

```
max = 170
median = 162
1^,3^ quartile, diff.: 157 165 8
mean = 161.3684210526316
```

```
150, 155, 156, 157, 157, 157, 159, 160, 162, 162, 163, 163, 164, 165, 165, 166, 167, 168, 170
```

Spiega a parole come si deve fare (sul significato di "mediana" e "quartili" ci soffermeremo fra poco).

La situazione analizzata nei quesiti 9 e 10 mette in luce alcuni problemi.

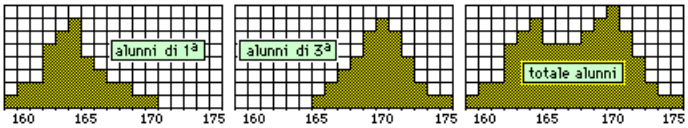
Un *primo problema* è che l'istogramma nel testo del quesito 9 ha un andamento abbastanza diverso da quello degli istogrammi di fig. 2. In questo caso ciò è dovuto al fatto che abbiamo considerato solo le informazioni relative alle diciassettenni di una particolare scuola mentre nel caso di fig. 2 avevamo a disposizione la totalità dei ventenni. Se la scuola fosse stata di dimensioni molto maggiori si sarebbero ottenuti istogrammi dall'andamento simile a quello degli istogrammi di fig. 2.

Un *secondo problema* è che ci possono essere più mode: la scelta del numero degli intervalli, influenzando la forma dell'istogramma, può anche condizionare la quantità delle classi modali che si ottengono.

Val la pena di osservare che vi sono situazioni in cui la forma dell'istogramma è diversa da quelle "a campana" degli istogrammi di fig. 2 per motivi di fondo, non perché sono poche le informazioni raccolte o perché non si sono scelti in modo opportuno gli intervalli. Ad esempio in *figura 5* sono riportati gli istogrammi di distribuzione (di frequenza assoluta) delle altezze degli alunni (maschi) delle classi - due prime e due terze - presenti nella succursale di una scuola secondaria superiore. L'istogramma a sinistra si riferisce agli alunni delle prime, quello al centro agli alunni delle terze, quello a destra al totale degli alunni.

11 Discutete la relazione tra la forma dell'istogramma relativo all'intera succursale e quella degli altri due.

figura 5



Un *terzo problema* è che l'altezza media delle alunne del quesito 9 (161 cm, arrotondando) non cade nella moda 162-164 cm. In questo caso ciò dipende dal numero delle alunne, piccolo rispetto al totale delle diciassettenni. Ma vi sono fenomeni che danno comunque luogo a istogrammi di distribuzione con moda molto diversa dalla media.

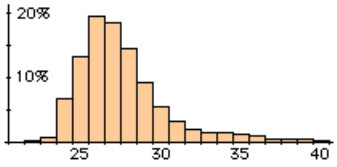


figura 6

Ad esempio nel caso della distribuzione dell'età di laurea presso l'Università di Genova nel triennio 1984-86 (*figura 6*) la media è 28 anni mentre la moda è 26 anni (in seguito, a causa dell'introduzione di due successivi livelli di laurea, questi valori si sono alzati di circa un anno). Infatti il valore della media subisce l'influenza della "coda" costituita dalle persone che si laureavano con grande ritardo (studenti lavoratori, "perdigiorno" mantenuti dalla famiglia benestante, …). E questa coda, che sta alla destra della classe modale, fa aumentare il valore della media rispetto a quello della moda.

Se nella scuola del quesito 9 l'alunna alta 150 cm si ritira e, contemporaneamente, si iscrive una diciassettenne spilungona, brava giocatrice di pallacanestro, alta 182 cm, la distribuzione delle altezze non cambia particolarmente, ma il nuovo valore, anomalo rispetto alla altre altezze, influisce non poco sul valore della media, che aumenta di quasi 2 cm. Invece il valore del dato al **centro dell'elenco dei dati** ordinati si modifica di poco: da 162 cm passa a 163 cm; e non cambierebbe se la superspilungona fosse ancora più alta.

```
15 | 0
15 | 567779
16 | 022334
16 | 55678
17 | 0
17 |
18 | 0
```

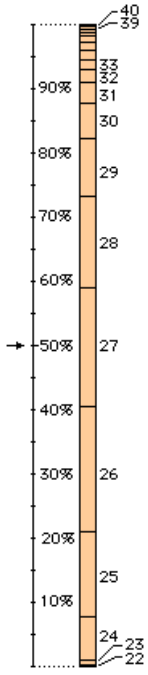
```
15 |
15 | 567779
16 | 022334
16 | 55678
17 | 0
17 |
18 | 2
```

Questo esempio e quello relativo all'età di conclusione degli studi universitari mettono in luce che la *media aritmetica* è un valore medio che *non* è *sempre significativo*. In situazioni come queste può essere utile impiegare un ulteriore tipo di valore medio, come abbiamo appena fatto per le altezze delle alunne: il valore del dato al centro dell'elenco dei dati ordinati, o ***mediana***.

Vediamo come interpretare graficamente la mediana nel caso delle età di laurea.

In questo caso non dispongo dei dati dei singoli studenti ma solo dell'istogramma di distribuzione di fig. 6. Non posso quindi procedere come ho fatto per l'altezza media delle alunne. Posso tuttavia individuare la mediana, seguendo due diversi procedimenti:

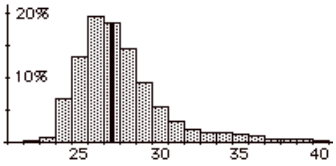
figura 7



(1) Sommo le frequenze percentuali delle varie età a partire dall'età minore (cioè dalla colonna più a sinistra dell'istogramma) e mi fermo quando raggiungo il 50%. Mentre nel caso delle altezze delle alunne si sono ordinati i dati in una tabella e si è presa la casella centrale, qui è come se appilassi i rettangolini che formano l'istogramma (passando a *figura 7*) e considerassi quello che sta a metà della striscia ottenuta, cioè il rettangolino per cui passa la quota che indica il 50%: l'età mediana di laurea è dunque di 27 anni. In altre parole il 50% degli studenti si laurea entro i 27 anni e l'altro 50% si laurea a un'età non inferiore ai 27 anni.

(2) Opero sull'istogramma di figura 6: la linea di divisione verticale che lo taglia in due parti di area uguale (*figura 8*) passa per l'intervallo che rappresenta i 27 anni.

figura 8



3. Campionamento

Facciamo un'ultima osservazione a proposito della ➡ tabella (2.1). Essa non è il frutto di un esame che ha riguardato tutte le ragazze diciassettenni italiane, ma solo una parte di esse. Si tratta, comunque, di un numero non troppo piccolo di ragazze, che ci consente di fare delle deduzioni su come è distribuita l'altezza del complesso delle diciassettenni italiane.

Quando per studiare un certo aspetto di un particolare insieme di "oggetti" (persone, animali, prodotti, …) si compiono osservazioni solo su una parte di essi, questa parte "estratta" dall'insieme totale degli oggetti viene chiamata ***campione***; l'analisi statistica così effettuata viene chiamata ***indagine campionaria***; il procedimento con cui si sono "estratti" gli oggetti di cui raccogliere le informazioni, viene chiamato *campionamento*.

Un famoso esempio di indagine *non* campionaria sulla popolazione italiana è costituito dai *censimenti*, che vengono effettuati ogni dieci anni (…, 1971, 1981, 1991, …) intervistando attraverso opportuni questionari tutti gli italiani.

12 Se chiedeste a ciascuno studente delle classi prime della vostra scuola quale numero di scarpa porta e analizzaste i dati così raccolti, che cosa realizzereste?

un'indagine campionaria sugli studenti delle classi 1^{e} della vostra scuola	☐	un'indagine campionaria sui ragazzi italiani di 14-15 anni	☐
un'indagine "completa" sugli studenti delle classi 1^{e} della vostra scuola	☐	un'indagine "completa" sui ragazzi italiani di 14-15 anni	☐

13 Supponiamo che con l'indagine del quesito 12 si voglia effettuare un'analisi statistica sui ragazzi italiani di 14-15 anni. Il campione scelto ti sembra "rappresentativo", cioè adeguato a fornire informazioni estendibili all'intera popolazione italiana di 14-15 anni?

I termini "campionaria", "campionamento", … derivano dalla parola "campione" intesa come "esemplare rappresentativo" (pensa al rappresentante che mostra campioni dei beni prodotti dalle ditte per cui lavora). La parola è stata poi estesa al significato statistico di "parte rappresentativa" di un certo insieme di soggetti.

E` importante fissare l'attenzione sull'aggettivo ***rappresentativa***: non basta prendere un po' di soggetti e fare su questi i calcoli per ottenere delle informazioni significative sulla totalità dei soggetti.

Supponiamo che l'Istat voglia analizzare un particolare aspetto delle condizioni di vita degli italiani tra un censimento e l'altro, ad esempio il numero dei componenti delle famiglie, e non abbia il tempo e i mezzi per fare un'indagine completa su tutti gli italiani. Può estrarre un campione di famiglie e analizzare i dati di queste. Ma deve fare l'estrazione non privilegiando una zona geografica, una fascia di età dei genitori, una condizione economica, … rispetto ad altre: infatti il fenomeno si presenta in maniera diversa al variare della regione, dell'epoca e dell'età in cui si sono sposati i genitori, delle condizioni sociali ed economiche, …; un campione che fosse fatto quasi tutto di famiglie dell'Italia centrale o che privilegiasse le famiglie di recente formazione rappresenterebbe poco fedelmente il complesso delle famiglie italiane.

Inoltre il campione deve essere *sufficientemente numeroso*. Ad esempio se una fabbrica di dischetti per calcolatori vuole fare un'indagine sulla quantità di letture/registrazioni che si possono fare sui dischetti prodotti prima che questi si danneggino (e, ovviamente, non sottopone ad una prova di durata tutti i dischetti: così facendo distruggerebbe tutta la propria produzione!) deve decidere quanti dischetti prendere "a caso" durante, ad esempio, una particolare giornata di produzione: prenderne il 10% sarebbe troppo dispendioso (occorrerebbe impiegare troppi dispositivi di lettura/scrittura su disco magnetico); prenderne lo 0.5% è *sufficiente*? Non è facile rispondere a questa domanda: occorre tener conto di altri fattori e utilizzare concetti matematici che per adesso non abbiamo ancora affrontato.

Riprenderai il problema del campionamento più avanti nel corso degli studi, dopo che avrai imparato i primi elementi di *calcolo delle probabilità*, cioè della parte della matematica che si occupa dei fenomeni casuali.

4. Percentili e "normalità"

Da ➡ fig.7 posso ricavare che il 10% degli studenti si laurea entro i 25 anni (e il 90% si laurea dopo il compimento dei 25 anni) e che il 75% degli studenti si laurea entro i 29 (e il 25% si laurea avendo già compiuto i 29 anni). Infatti tagliando il diagramma a striscia alle quote 10% e 75% vado a cadere nei rettangoli che rappresentano le età di 25 anni e 29 anni, rispettivamente.

14 Usando fig.7 completa la seguente *tabella (4.1)*, dove *età* indica l'età che separa il primo *p*% degli studenti (ordinati per età al momento della laurea) dai rimanenti.

(**4.1**)

p%	5%	10%	25%	50%	75%	90%	95%
età (in anni)		25			29		

Il valore corrispondente a una frequenza cumulata del *p*% viene detto *p*-esimo ***percentile*** o percentile di *ordine p*. Ad es. nel nostro caso il 50° percentile (cioè la mediana) è 27, il 10° percentile è 25, il 75° è 29.

Tabelle come (4.1), o quelle che si ottengono con una diversa scelta delle percentuali, possono essere considerate un'alternativa agli istogrammi di distribuzione percentuale.

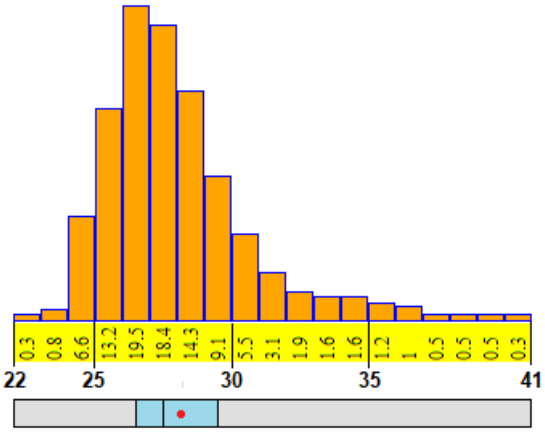
Ad esempio la forma allungata verso destra dell'istogramma di ➡ fig.8 trova corrispondenza nel fatto che il 40% che segue la mediana (cioè gli studenti che vanno dal 50° al 90° percentile) spaziano dai 27 ai 31 anni, mentre il 40% che precede la mediana (cioè gli studenti che vanno dal 10° al 50° percentile) spaziano in un intervallo molto più piccolo, dai 25 ai 27 anni.

La differenza tra l'intervallo che va dal 5° al 50° percentile e quello che va dal 50° al 95° è ancora maggiore: nel primo caso si spazia su 4 anni di età, dall'età di 24 anni a quella di 27, nel secondo si spazia su 8 anni, dall'età di 27 a quella di 34.

15 Secondo voi è normale che, alla fine degli anni '80, uno studente si laureasse a 28 anni (mentre ci sono studenti che si laureavano a 22 e 23 anni)? Secondo voi è basso un adulto alto 168 cm (mentre l'altezza media dei maschi che avevano 20 anni nel 1991 è 174 cm - ➡ figura 2)?

Vediamo come rappresentare i dati relativi alla distribuzione dell'età di laurea in modo da ottenere la figura seguente.

figura 9



Potremmo usare gli script per tracciare istogrammi introdotti nelle schede precedenti (e usati anche per alcuni istogrammi di questa scheda), ma è più comodo lo script **histo**, che consente di tracciare istogrammi anche di dati non già classificati.

In questo caso sappiamo che le frequenze percentuali nelle varie classi (tra 22 e 23 anni, tra 23 e 24 anni, …) sono quelle indicate nella figura. Introduciamo come dati i **centri dei vari intervalli** (22.5, 23.5, …) moltiplicati per le rispettive frequenze (0.3, 0.8, …).

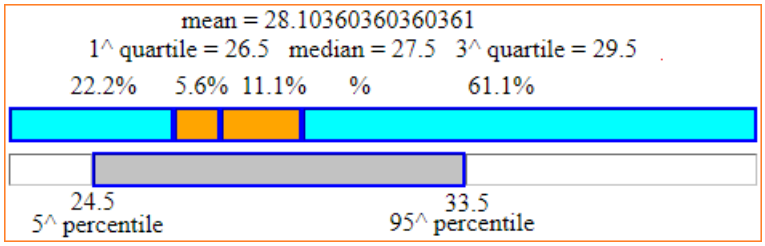
16 Apri lo script, copia e incolla nella casella dei dati quanto segue;
22.5*0.3, 23.5*0.8, 24.5*6.6, 25.5*13.2, 26.5*19.5, 27.5*18.4, 28.5*14.3, 29.5*9.1, 30.5*5.5, 31.5*3.1, 32.5*1.9, 33.5*1.6, 34.5*1.6, 35.5*1.2, 36.5*1.0, 37.5*0.5, 38.5*0.5, 39.5*0.5, 40.5*0.3
Clicca [Enter], poi [min], [max], [median], [mean]. Poi scegli 22 come [A], 41 come [B] e 19 come [numero intervalli] (perché?).

Ottieni l'istogramma sopra raffigurato, ruotato di 90°, e le uscite seguenti:

n = 99.89999999999998 **min = 22.5 max = 40.5 median = 27.5 1^ | 3^ quartile = 26.5 | 29.5 mean = 28.10360360360361**

Come **n** non si ottiene 100 perché le percentuali sono arrotondate (poi, appare 99.899… invece di **99.9** a casua del fatto che il computer non opera in base 10, come abbiamo già osservato).

Il rettangolino sotto all'istogramma rappresenta la collocazione della mediana, del 25º e del 75º percentile, chiamati anche primo e terzo **quartile** (invece dei centesimi vengono usati i quarti); il pallino rosso rappresenta la media. La sua rappresentazione può essere ottenuta con lo script **boxplot**, introducendo la stessa sequenza di dati usata con **histo**:



Il nome "boxplot" richiama il fatto che il nome per esteso di questa rappresentazione è *box-and-whiskers-plot* ("diagramma a scatola e baffi"). È una figura "lineare" (si sviluppa solo orizzontalmente, non in due dimensioni, come gli istogrammi) che sintetizza in modo efficace come si distribuiscono i dati.

Il box (scatola) rappresenta il 50% centrale dei dati, la tacca dentro al box rappresenta la mediana. Il fatto che il box sia spostato verso sinistra (cioè che il baffo sinistro sia molto più corto di quello destro) corrisponde all'allungamento verso destra dell'istogramma. Sotto sono rappresentati anche il 5° e il 95° percentile.

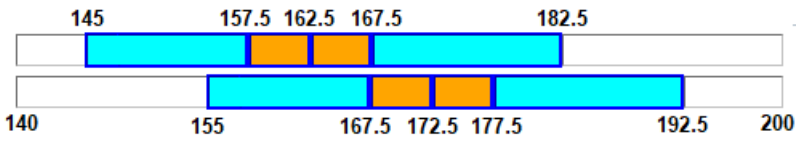
Come ci saremmo aspettati, essendo l'istogramma allungato verso destra (ricorda quanto discusso dopo figura 6 a proposito delle altezze delle alunne), la mediana è inferiore della media.

Per la **media** si è ottenuto 28.10360360360361; ma tutte queste cifre non hanno ovviamente senso; come dobbiamo **arrotondare** questo valore? Analogamente nel caso del quesito 10 abbiamo ottenuto la media 161.3684210526316; come arrotondarla?

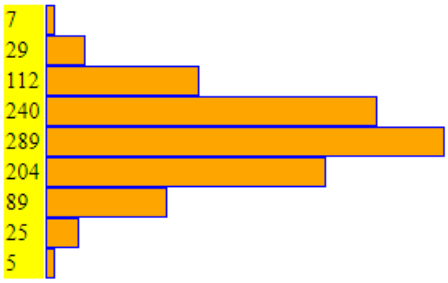
La prima media è stata ottenuta da dati arrotondati agli interi, e i dati erano migliaia; possiamo arrotondare la media con due cifre in più, ai centesimi: 28.10. Anche nel secondo caso i dati erano arrotondati agli interi, ma erano poco più di una decina; possiamo arrotondare con una sola cifra in più: 161.4.

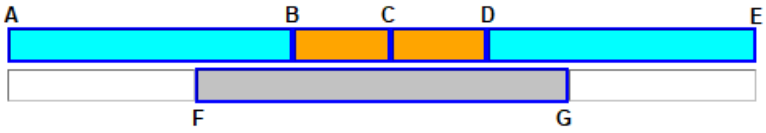
Questa scelta possono essere motivate con considerazioni di calcolo delle probabilità che, per ora, non siamo in grado di affrontare.

I boxplot consentono di confrontare facilmente andamenti dello stesso fenomeno in anni diversi. Ecco ad esempio rappresentati, in modo approssimato, i boxplot relativi alle altezze dei ventenni nel 1881 e nel 1991:



17 Nel 1976 su 1000 ventenni italiani 7 avevano altezza in centimetri in [150,155), 29 in [155,160), 112 in [160,165), 240 in [165,170), 289 in [170,175), 204 in [175,180), 89 in [180,185), 25 in [185,190), 5 in [190,195).
Introducendo opportunamente 152.5*7, 157.5*29, 162.5*112, 167.5*240, 172.5*289, 177.5*204, 182.5*89, 187.5*25, 192.5*5 tracciane l'istogramma e il boxplot in modo da ottenere le figure a destra e sottostante.
Scrivi quali sono i valori A, B, …, F e il valore medio.





I percentili permettono di affrontare in modo serio questioni come: «*che cosa vuol dire essere di altezza normale?*». Ad esempio che cosa si intende dicendo che una persona è bassa? Che la sua altezza è inferiore all'altezza mediana? Ma in tal caso le persone si dividerebbero quasi tutte in alte o basse, e sarebbero normali solo poche persone.

Per dare un significato "oggettivo" alla valutazione dell'altezza dobbiamo fissare delle convenzioni. Ad esempio potremmo dire che sono "nella media" le altezze che cadono entro il 50% centrale dei dati, cioè tra il 25° e il 75° percentile, e che sono "basse" quelle inferiori al 25° percentile e "alte" quelle che superano il 75° percentile. Considerazioni analoghe si potrebbero fare per l'età di laurea.

Si tratta, comunque, sempre di valutazioni statistiche basate su scelte convenzionali e che devono essere riferite a valutazioni più generali della situazione che si sta considerando.

Facciamo due esempi.

(***1***) Se ritenessimo "statisticamente" normale laurearsi tra il 25° e il 75° percentile, cioè, nel caso considerato, tra i 26 e i 29 anni, non potremmo certo considerare "anormale" (nel senso di "tipo strano") uno studente che si laurea a 25 anni o ritenere che chi si iscrive all'università possa preventivare come "normale" (cioè come obiettivo "accettabile") la conclusione degli studi a 29 anni.

(***2***) Se un pediatra dispone della tabella a fianco dei percentili per le altezze delle bambine di 12 mesi e, visitando una bimba di 1 anno, trova che è alta circa 68 cm, può supporre che vi sia qualche ritardo nella crescita. Infatti la sua altezza è inferiore al 3° percentile: il 97% delle bimbe della sua età ha un' altezza superiore.

3°	10°	25°	50°	75°	90°	97°
69	71	72	74	76	77	79

Ovviamente in questa valutazione il pediatra deve tener conto dell'altezza dei genitori: se anche la loro altezza cadesse tra i primi percentili il fenomeno non sarebbe particolarmente preoccupante.

Inoltre deve effettuare la misura con cura, eventualmente ripetendola più volte: già con un adulto da una misurazione all'altra ci può essere lo scarto di un paio di centimetri (sulla misura incidono la posizione della colonna vertebrale, che può variare anche in relazione alla stanchezza della persona, la posizione della testa, la cura con cui viene letta la scala graduata, …); con un bimbo piccolo, che è difficile da tener fermo, lo scarto può essere anche maggiore.

A questo punto dovrebbe essere chiaro che il concetto di ***normalità*** è convenzionale e dipende dal contesto. Ad esempio se un regista cerca per una parte un uomo né troppo alto né troppo basso può dare l'incarico di cercare un uomo la cui altezza rientri in quella della maggioranza degli uomini. Questa espressione informale può essere tradotta dai collaboratori del regista nella ricerca di una persona la cui altezza cada nel 50% centrale delle altezze, cioè tra il 25° e il 75° percentile.

In altre situazioni si possono assumere come altezze "normali" intervalli più piccoli (ad es. tra il 30° e il 70° percentile, cioè il 40% centrale dei dati) o più grandi (ad es. dal 3° al 97° percentile, cioè il 94% centrale).

Veniamo, infine, a dati che vi riguardano più da vicino.

Le *tabelle (4.2)* e *(4.3)* contengono alcuni percentili relativi alle *altezze a varie età dei ragazzi e delle ragazze italiane nate intorno al 1980*.

(4.2) M

età	3°	10°	25°	50°	75°	90°	97°
14	148	154	159	165	170	174	179
15	153	160	164	170	175	178	184
16	157	163	168	173	177	181	186
17	159	165	170	174	178	182	187
18	160	166	170	174	179	183	188
19	160	166	170	174	179	183	188

(4.3) F

14	149	153	156	160	164	167	171
15	150	155	158	161	165	168	172
16	151	156	159	162	166	169	172
17	151	156	159	162	166	169	172

18 Discutete le principali differenze tra maschi e femmine messe in luce dalle tabelle (4.2) e (4.3).

Le figure 1 e 2 e le tabelle (4.2) e (4.3) sono riferite al complesso degli italiani. In *zone diverse* del paese la distribuzione delle altezze si può manifestare in maniera piuttosto differente. Ad esempio l'altezza media dei maschi ventenni nel 1976, che sul totale dell'Italia era 172.0 cm, in Sardegna era 168.5 cm, in Abruzzo 171.1 cm e in Friuli-Venezia Giulia 175.6 cm.

L'altezza di una ragazza o di un ragazzo che risiede in Abruzzo (regione che presenta una distribuzione delle altezze quasi uguale a quella del complesso dell'Italia) ma ha i genitori originari della Sardegna o del Friuli dovrebbe essere riferita più ai dati di questa regione che a quelli nazionali, cioè a dati che sono slittati in un caso di quasi 4 cm in meno, nell'altro di quasi 4 cm in più rispetto a quelli delle tabelle (4.2)-(4.3).

Le tabelle (4.2)-(4.3) sono da interpretare tenendo conto oltre che di questo aspetto anche del fatto che i *tempi dello sviluppo* dell'altezza possono variare da individuo a individuo. Vi può essere il ragazzo alto 170 cm a 15 anni (oltre il 50° percentile) e che negli

anni successi non cresce più (scendendo sotto al 25° percentile) e quello che a 15 anni è alto 160 cm (sotto al 25° percentile) ma che continua a crescere e a 18 anni raggiunge i 175 cm (oltre il 50° percentile).

I tempi dello sviluppo dell'altezza sono cambiati nel corso degli anni: oltre all'altezza media (→ fig. 1) è cambiata anche l'età in cui ciascuno raggiunge la propria altezza massima. Attualmente in Italia praticamente tutti i maschi (→ tabella (4.2)) oltre i 18 anni non aumentano più in altezza e praticamente tutte le femmine (→ tabella (4.3)) a 16 hanno già raggiunto l'altezza massima. Agli inizi del Novecento queste età erano spostate in avanti di 5 o 6 anni.

Differenze tra maschi e femmine, tra individuo e individuo e tra epoche diverse analoghe a quelle osservate per lo sviluppo dell'altezza valgono anche per lo *sviluppo sessuale*. Ad esempio nel 1890 in Europa una donna era in grado di procreare figli mediamente a partire dai 16 anni; nel 1990 questa età media era scesa a 13 anni. Per i maschi queste età vanno spostate in avanti di circa 2 anni.

Pure in questo caso si tratta di valori medi: anche per queste età si potrebbero considerare istogrammi di distribuzione o tabelle di percentili. Ad esempio vi può essere la ragazza che è sessualmente "adulta" a 11 anni e quella che lo diventa a 16.

5. Concludendo

Con questa scheda abbiamo visto ulteriori *modelli matematici* usati per fare statistiche e abbiamo esaminato alcuni problemi relativi al loro impiego.

Le ultime osservazioni sul campionamento offrono lo spunto per sottolineare che l'uso dei modelli statistici è soggetto a interpretazioni erronee o distorte più di altri modelli matematici.

Il motivo risiede nel fatto che con essi spesso non si rappresentano tanto le caratteristiche di un particolare oggetto o persona quanto le condizioni che riguardano una *collettività*, le caratteristiche essenziali dell'*andamento* complessivo di un fenomeno che varia nel tempo, ….

Il modo in cui vengono raccolte le informazioni (su tutta la popolazione o su quanta parte di essa? ogni quanto tempo? con quale modalità di rilevamento? …) e il fatto che le caratterisitiche delle persone o degli eventi singoli possono discostarsi molto dalla valutazione complessiva che emerge, introducono notevoli elementi di approssimatività.

Alcuni degli esercizi seguenti offrono occasioni per esemplificare e approfondire questa riflessione.

6. Approfondimenti

Un'indagine

Proponiamoci di fare anche noi un'indagine statistica, ad esempio su due aspetti: le altezze dei ragazzi e delle ragazze tra i 14 e i 18 anni, per operare un confronto con i dati delle tabelle (4.3) e (4.4), e sulla lunghezza dei capelli dei ragazzi e delle ragazze della vostra età.

19 Precisate meglio gli obiettivi della vostra indagine e discutete come organizzarla affinché si possano ottenere informazioni utili e attendibili.

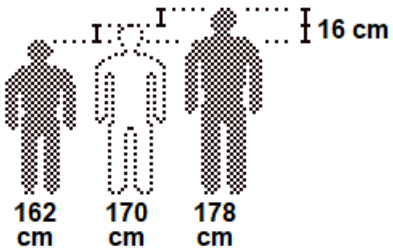
Per adesso potrete accontentarvi di prendere come campione i ragazzi delle classi della vostra scuola, restringendovi alla sola vostra classe per quanto riguarda la lunghezza dei capelli. Eventualmente potrete confrontare i risultati della vostra indagine con quelli ottenuti con un'indagine simile da alunni di altre scuole e con i risultati che si ottengono mettendo insieme tutti i dati.

20 Raccolti i dati, registrateli e analizzateli opportunamente. Se fate copia dei vostri dati e la stessa operazione viene fatta da altre classi, mettendo poi insieme i dati raccolti otterrete un campione più numeroso su cui ripetere l'analisi.

7. Esercizi

e1 Nel caso delle rappresentazioni "procapite" (kg di carne consumata per abitante, m² di superficie per abitante, m³ di spazio abitativo per famiglia, € di reddito per lavoratore, …) la ***media*** può essere interpretata come rapporto tra due grandezze: un *totale* espresso in una data unità di misura (kg, €, m², m³, …) e una "popolazione" (di persone, famiglie, …).

Nel caso dell'altezza media questa interpretazione non ha senso: è vero che faccio la somma delle altezze e la divido per il numero delle persone, ma questa somma non la posso interpretare come "altezza totale" delle persone! non posso dire che l'altezza media è di 174 cm per abitante!

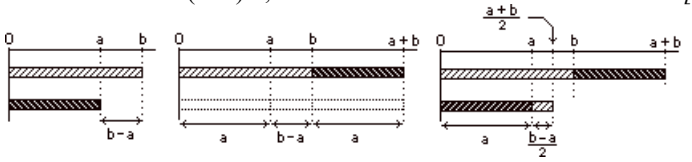


Posso tuttavia dare anche questa interpretazione:
l'altezza media di due persone è pari all'altezza di una terza persona che abbia lo stesso dislivello dalla prima e dalla seconda.

Ad es. 170 cm è la media di 162 e 178 cm; infatti (162+178)/2=340/2= 170. Ma 170 è anche il *valore a metà* tra 162 e 178: 162+8=170, 178–8=170.

- *Osserva* la figura seguente, che illustra due modi per trovare la lunghezza media M di due segmenti lunghi *a* e *b*:

uno è usare la formula: M = (*a*+*b*)/2; l'altro è usare: M = *a*+..... *[completa]*



- Prova a calcolare a mente lo stipendio medio mensile (*m*) di una famiglia composta solo da marito e moglie, *lei* con stipendio di 2 milioni e 400 mila (*x*), *lui* con stipendio di 2 milioni e 500 mila lire (*y*), usando le formule:

(1) $m = (x+y)/2$ (2) $m = x+(y–x)/2$

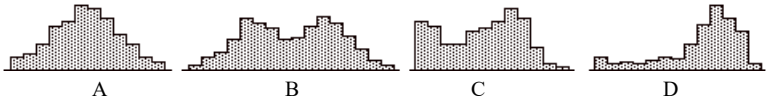
Quale procedimento trovi più conveniente? Perché?

- La località C è esattamente a metà strada tra una località posta al 34° km (x) di una certa strada statale e una località B posta al 112° km (y). Calcola a quale chilometro (m) si trova C.
  Quale procedimento tra (1) e (2) trovi più conveniente? Perché?

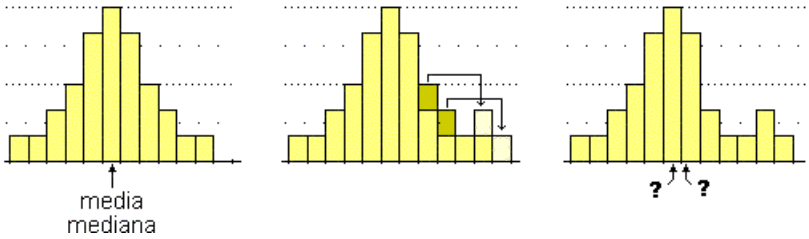
e2 Per controllare attraverso un procedimento "numerico" la dimostrazione "geometrica" dell'equivalenza delle formule (1) e (2) (➡ quesito e1) per il calcolo della media tra x e y, *completate* i seguenti passaggi:

$$x + \frac{y-x}{2} = \frac{x \cdot 2}{2} + \frac{y-x}{2} = \frac{x \cdot 2 + y - x}{2} = \dots$$

e3 Indica tra i seguenti istogrammi quale può rappresentare la distribuzione: (1) dell'età dei morti in un paese sviluppato, (2) dell'età dei morti in un paese sottosviluppato, (3) dell'altezza delle femmine adulte di una città, (4) delle altezze degli adulti (maschi e femmine) di una città.

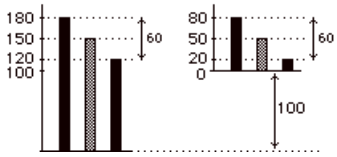


e4 Ho un istogramma di distribuzione dalla forma simmetrica, in cui media e mediana cadono entrambe nella classe centrale. Se tolgo pezzi da colonne a destra della colonna centrale e li sposto più a destra, quale tra mediana e media resta immutata? quale aumenta? perché?



e5 Tra gli istogrammi raffigurati nel quesito e3 quale ha sicuramente la media inferiore alla mediana; quale può avere media e mediana che cadono nella classe modale; quale può avere media e mediana che cadono in una stessa classe, diversa dalla classe modale?

e6 Hai visto nel quesito e1 che la media tra due numeri coincide con il valore che sta a metà tra essi.



Il disegno a fianco suggerisce che per trovare la media tra 120 e 180 posso operare su 20 e 80: la distanza tra 120 e 180 è uguale alla distanza tra 20 e 80 (ottenuti togliendo 100), per cui posso trovare il valore che sta a metà di questi ultimi e poi riaggiungere 100:

(media tra 120 e 180) = (media tra 20 e 80) + 100

Tale procedimento (togliere uno stesso numero da tutti i valori di cui si fa la media e poi riaggiungerlo al risultato) può essere esteso al calcolo della media di più di due valori.

Applicalo per calcolare la media di ciascuno dei seguenti insiemi di dati:

(a) 253, 254, 259, 256 (b) 2.5, 2.1, 2.3 (c) 1037, 1045, 1000, 1002

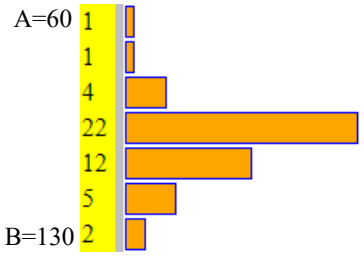
e7 Completa la seguente formula in modo che rappresenti il procedimento descritto nel quesito precedente:

$$\frac{x_1 + x_2 + x_3 + \dots + x_n}{n} = \frac{(x_1 - h) + (x_2 - h) + (x_3 - h) + \dots + (x_n - h)}{n} + \dots$$

e8 Nelle gare di corsa non particolarmente "importanti" (e, fino a qualche decennio fa, in tutte le gare) i tempi non vengono rilevati con apparecchiature elettroniche, ma a mano, con dei cronometri. I cronometri, così come tutti gli odierni orologi al quarzo, sono precisissimi: sgarrano di pochi secondi al mese. Quindi, se un orologio è dotato di un pulsante "start/stop" e visualizza i centesimi di secondo, siamo sicuri che il tempo che intercorre tra due successive pressioni del pulsante è rappresentato correttamente, troncato ai centesimi di secondo, dal numero che viene visualizzato.

Nei cronometraggi delle gare, tuttavia, non viene impiegato un unico cronometro, ma i tempi vengono misurati contemporaneamente da più cronometristi. Poi vengono presi come tempi i valori medi, troncati ai centesimi, dei tempi registrati dai diversi cronometristi.

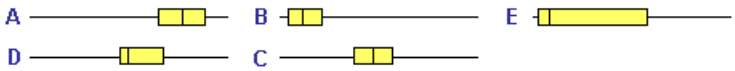
Discuti questa scelta alla luce dell'analisi dei valori in centesimi di secondo che una persona ("normale", non un allenato cronometrista) ha ottenuto misurando ripetutamente con un orologio A il tempo che impiega un altro orologio B a scattare in avanti di 1 s (ad esempio la persona ha dato lo Start sull'orologio A appena l'orologio B ha visualizzato 15:31:08 e ha dato lo Stop appena B ha visualizzato 15:31:09, e ha trascritto il tempo visualizzato da A; poi ha fatto lo stesso per esempio dalla visualizzazione di 15:31:46 a quella di 15:31:47; ecc.). Ecco i valori, **troncati** ai centesimi di secondo: 111, 103, 109, 97, 99, 110, 99, 103, 109, 106, 109, 112, 98, 110, 90, 98, 96, 90, 96, 90, 109, 103, 96, 97, 96, 96, 109, 103, 98, 94, 115, 78, 85, 96, 89, 121, 98, 103, 97, 88, 98, 129, 96, 68, 91, 80, 102 Utilizzando la **grande CT** aggiungo ai dati 0.5 (devo fare il calcolo con i **centri dei vari intervalli**: vedi) e quindi analizzo i dati così ottenuti, usando **histo**. ***Scrivi***, arrotondata ai centesimi, la media ottenuta.



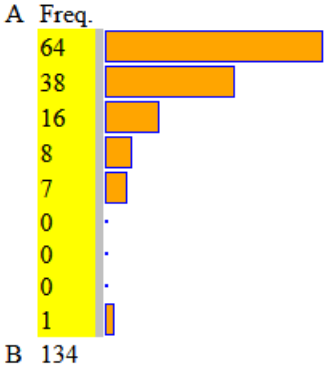
intervals = 7 their width = 10
n=47 min=68.5 max=129.5
median=98.5 1^ | 3^ quartile = 96.5 | 109.5

mean= ...

Tra i diagrammi schizzati a lato, ***qual*** è il box-plot dei tempi registrati? Perché? Verifica la tua risposta usando lo script **boxplot**.



e9 Un ente pubblico ordina alla ditta SifanStat, specializzata in indagini statistiche, lo studio dei tempi di arrivo degli utenti ai propri sportelli. Un dipendente della SifanStat si piazza all'ingresso del locale in cui sono collocati gli sportelli e per circa un'ora, in un orario di punta, misura il tempo che intercorre tra l'arrivo di un utente e il successivo, contando complessivamente l'arrivo di 134 utenti. I tempi che ha rilevato (troncati ai secondi) sono: 15, 29, 4, 99, 30, 11, 45, 7, 61, 67, 59, 8, 37, 30, 5, 53, 18, 22, 173, 6, 79, 16, 16, 6, 10, 22, 13, 16, 39, 7, 25, 9, 24, 5, 10, 84, 89, 38, 14, 16, 9, 3, 34, 9, 35, 34, 16, 13, 37, 29, 10, 51, 19, 72, 15, 9, 21, 13, 51, 2, 8, 2, 40, 28, 5, 22, 51, 6, 37, 21, 4, 4, 33, 35, 81, 21, 8, 19, 42, 9, 44, 12, 9, 53, 15, 35, 96, 35, 27, 34, 4, 13, 57, 11, 20, 25, 25, 3, 48, 17, 72, 14, 13, 31, 94, 2, 2, 38, 11, 9, 7, 76, 96, 35, 9, 21, 28, 47, 78, 26, 41, 16, 3, 15, 1, 56, 26, 3, 8, 17, 25, 40, 15, 60



A = 0 B = 180 intervals = 9 their width = 20
n = 134 min = 1.5 max = 173.5
median = 21.5 1^ | 3^ quartile = 9.5 | 38.5
mean = 29.55223880597015

(**a**) Ai dati è stato aggiunto 1/2. Perché?

(**b**) Quale o quali script sono stati impiegati per ottenere tali esiti

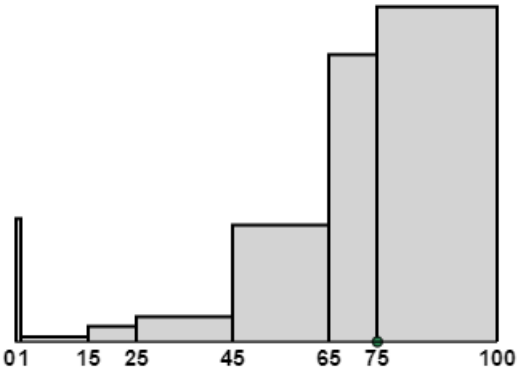
(**c**) Come arrotonderesti la media ottenuta?

(**d**) Tra i diagrammi riprodotti nel quesito precedente, *qual* è il box-plot di questo file? Perché? Verifica la tua risposta usando un opportuno script.

e10 Analizziamo dei *dati già classificati in intervalli di diversa ampiezza.* Consideriamo i dati a destra, relativi alla distribuzione dell'età dei morti in Italia nel 1990. I dati sono in centinaia di persone: ad es. sono morte 25 centinaia di persone nella fascia 1-14 anni (cioè in [1,15): avevano compiuto 1 anno e non ancora i 15). La tabella-Istat da cui sono stati riportati i dati indicava l'ultima classe come "75 e più". Si è introdotto [75,100) supponendo che, allora, fosse trascurabile la percentuale dei morti ultracentenari. Sotto la rispettiva rappresentazione grafica.

[0,1) 46 [1,15) 25 [15,25) 58
[25,45) 186 [45,65) 870 [65,75) 1071
[75,100) 3124 5380 dati

0.5*46, 8*25, 20*58, 35*186, 55*870, 70*1071, 87.5*3124



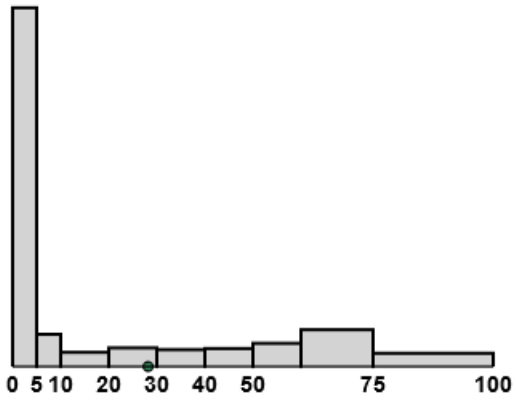
(**a**) Il valore medio dell'età di morte è rappresentato dal pallino verde. Esso è stato calcolato con la **grande CT** introducendo i dati sopra riportati. Perché sono stati introdotti questi dati? Qual è il valore medio arrotondato ai decimi?
(**b**) Se l'area dell'intero istogramma vale 100, qual è l'area del rettangolo più a destra?

e11 La tabella (7.1) contiene la distribuzione dell'età dei morti in Italia in due periodi. I dati sono in centinaia di persone. Nel caso del decennio 1881-90 per ogni fascia di età è riportato il numero medio dei morti in un anno (ad es. nell'intervallo di anni di età [5,10) vi sono stati in media 343 centinaia di morti all'anno).

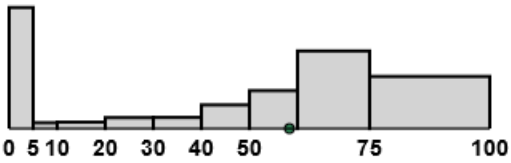
(**7.1**)

anni	0-4	5-9	10-19	20-29	30-39	40-49	50-59	60-74	75-∞
1881-90	3818	343	303	398	360	384	495	1177	708
1951	729	35	77	132	134	285	457	1401	1569

(A)



(B)



(**a**) Associa ad ogni periodo il corrispondente istogramma.
(**b**) In quali fasce di età era più alta la percentuale di morti per anno di età nei due periodi? E nel 1990? Come hai individuato le risposte?

e12 La tabella (7.2) contiene il *peso medio* di maschi e femmine di altezza e fascia di età fissate. Contiene inoltre il "*peso ideale*" di maschi e femmine di età adulta; non viene indicato un unico dato, ma un intervallo: ad es. il peso ideale delle donne alte 150 cm può andare da 44 a 54 kg, nel senso che una donna alta 150 cm con scheletro particolarmente leggero ha come peso ideale 44 kg e una con scheletro particolarmente pesante ha come peso ideale 54 kg. Il peso ideale di una certa categoria di soggetti viene definito convenzionalmente come il peso a cui corrisponde l'età media di morte più alta (i soggetti con quel peso mediamente vivono più a lungo dei soggetti con altro peso).

(**7.2**)

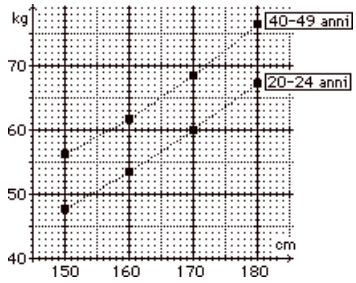
indagine campionaria
sulla popolazione italiana
(anno 1990)

		peso medio (kg)		peso ideale (kg)	
	altezza (cm)	20-24 anni	40-49 anni	*da*	*a*
M	160	59.9	65.3	53	64
	170	65.7	72.9	56	72
	180	72.8	80.5	66	80
	190	80.4	88.9	73	89
F	150	47.7	56.3	44	54
	160	53.5	61.7	48	59
	170	59.8	68.4	54	67
	180	67.3	76.4	62	75

- Una ragazza robusta alta 160 cm e pesante 60 kg legge preoccupata in una "rivista femminile", in un articolo sulle diete, che il peso ideale di una donna della sua altezza è 50 kg. Perché ciò che è scritto sulla rivista è una stupidaggine?
- Un uomo di 45 anni e alto 180 cm, che a vent'anni pesava 64 kg, ora pesa 81 kg. Da una statistica sul giornale legge che a mezz'età un uomo della sua altezza pesa mediamente 80 kg. Ritenendo, allora, di avere un peso "normale", decide di non dare più importanza alle sollecitazioni della moglie («pesi troppo: stai più attento nel mangiare!»). Ti sembra sensata questa conclusione?

e13 I grafici a lato rappresentano il peso medio *P* in funzione dell'altezza *h* nel caso delle donne tra 20 e 24 anni e nel caso di quelle tra 40 e 49 anni (vedi *ques. e16*). Il pallini sono la "traduzione" dei dati della tabella (7.3); le linee punteggiate che li congiungono consentono di trovare i pesi medi corrispondenti ad altre altezze (nell'ipotesi che tra un pallino e l'altro la variazione del peso medio sia proporzionale a quella dell'altezza).

Trova in questo modo (arrotondato ai kg) il peso medio delle donne di 20-24 anni alte 167 cm e confrontalo con quello che ottieni usando direttamente i dati della tabella e un opportuno metodo numerico.



e14 Potete effettuare delle altre indagini statistiche. Ad es. comprare qualche chilo di patate di una qualità fissata in un particolare negozio, pesare ciascuna patata e studiare come si distribuisce il peso delle patate, o fare un'indagine simile per qualche altro prodotto alimentare. Oppure potete scegliere un marciapiede di una grande strada, una direzione di cammino e misurare l'intervallo di tempo che intercorre tra il passaggio di un pedone e il successivo (scegliete un punto che non sia preceduto, a poca distanza, da un semaforo, che condizionerebbe il flusso delle persone) e studiare come si distribuiscono questi tempi. Oppure potete misurarvi (tutti gli alunni della classe non affebbrati) la temperatura corporea in più ore diverse e per più giorni consecutivi, raccogliere i dati e discutere che cosa si deve intendere come "temperatura normale".

1) Segna con l'evidenziatore, nelle parti della scheda indicate, frasi e/o formule che descrivono il significato dei seguenti termini:
intervallo di numeri (dopo ques.3), *classificare in modalità* (dopo ques.4), *frequenza assoluta, relativa e percentuale* (dopo ques.5), *distribuzione di frequenza* (dopo ques.6), *classe modale* (§2), *mediana* (dopo fig.6), *percentile* (dopo ques.14), *indagine campionaria* (§3).

2) Su un foglio da "quadernone", nella prima facciata, esemplifica l'uso di ciascuno dei concetti sopra elencati mediante una frase in cui esso venga impiegato.

3) Nella seconda facciata riassumi in modo discorsivo (senza formule, come in una descrizione "al telefono") il contenuto della scheda (non fare un elenco di argomenti, ma cerca di far capire il "filo del discorso").

script: piccola CT grande CT isto isto con % boxplot striscia 100 60 ordina Grafici