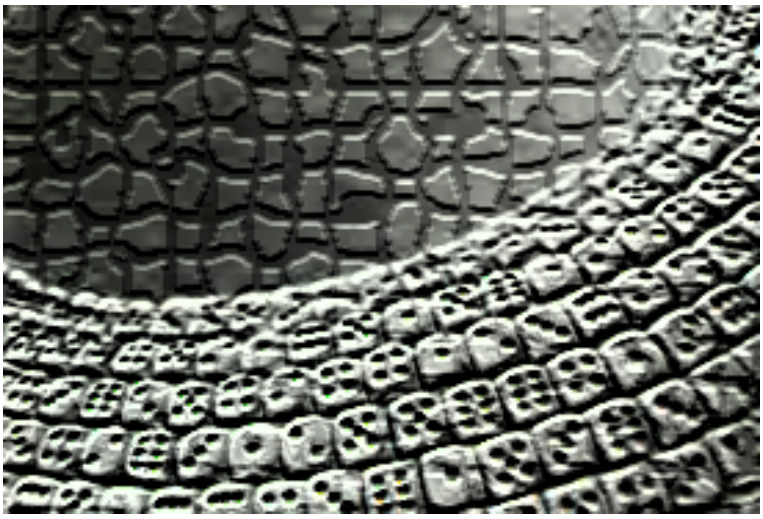
STATISTICA DESCRITTIVA

Dipartimento di Matematica



ITIS V.Volterra
San Donà di Piave

Versione [2015-16]

Indice

Ringraziamenti

Il presente lavoro si è avvalso di parecchi documenti di varia natura; in particolare si è approfittato molto del Cap. 1 della dispensa *Statistica* del Prof. Claudio Agostinelli dell'Università Ca' Foscari di Venezia e anche di vario altro materiale che ci ha lasciato in occasione del notevole corso tenuto nell'ambito del Progetto Lauree Scientifiche 2011-2012 e titolato : *La Statistica con R*, del quale ancora lo ringraziamo. Il paragrafo 1.10 sulla bontà dei dati è una libera interpretazione di un capitolo analogo del prezioso libro di David J. Hand, *STATISTICS A Very Short Introduction*, Oxford University Press.

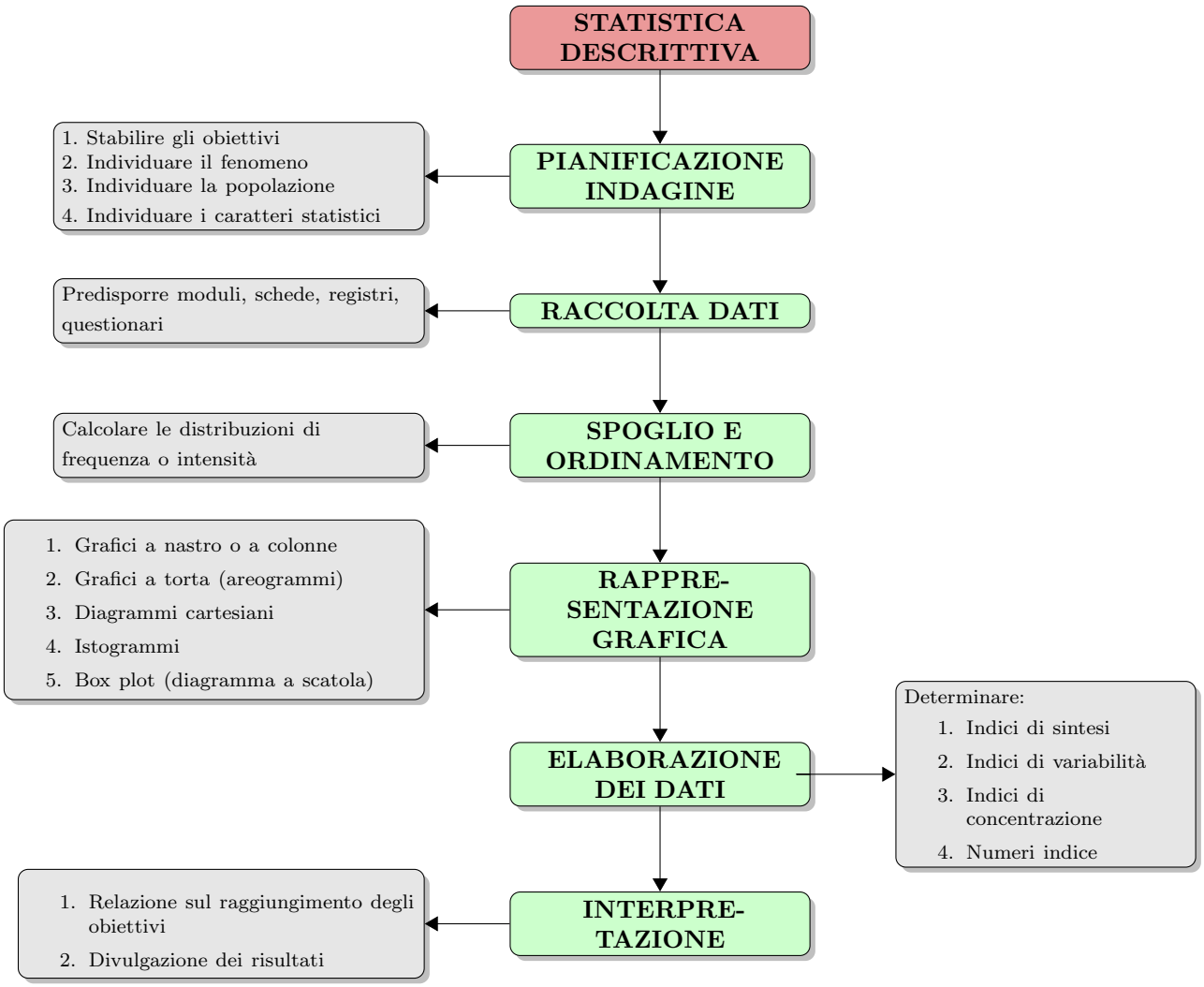
Capitolo 1

Generalità

La Statistica riguarda i metodi scientifici utilizzati per raccogliere, organizzare, sintetizzare, analizzare e presentare i dati, ma riguarda anche la possibilità di trarre conclusioni valide e di prendere decisioni ragionevoli sulla base di tali analisi.

In questa prima parte ci occuperemo degli aspetti descrittivi della Statistica. La figura alla pagina seguente riassume schematicamente le principali fasi in cui una indagine statistica si può scomporre.

Nei primi paragrafi ci occuperemo di dare alcune definizioni che saranno utili per descrivere in modo più preciso le grandezze coinvolte nella disciplina.


STATISTICA DESCRITTIVA
PIANIFICAZIONE INDAGINE
1. Stabilire gli obiettivi
2. Individuare il fenomeno
3. Individuare la popolazione
4. Individuare i caratteri statistici
RACCOLTA DATI
Predisporre moduli, schede, registri, questionari
SPOGLIO E ORDINAMENTO
Calcolare le distribuzioni di frequenza o intensità
RAPPRE-SENTAZIONE GRAFICA
1. Grafici a nastro o a colonne
2. Grafici a torta (areogrammi)
3. Diagrammi cartesiani
4. Istogrammi
5. Box plot (diagramma a scatola)
ELABORAZIONE DEI DATI
Determinare:
1. Indici di sintesi
2. Indici di variabilità
3. Indici di concentrazione
4. Numeri indice
INTERPRE-TAZIONE
1. Relazione sul raggiungimento degli obiettivi
2. Divulgazione dei risultati

1.1 Statistica e popolazioni

Definizione 1.1.1. Diciamo *popolazione* un qualsiasi insieme di oggetti (persone,cose) che si possano considerare omogenei rispetto ad una o più caratteristiche comuni. Se l'insieme è troppo grande per essere studiato interamente, allora un raggruppamento opportuno degli elementi dell'insieme si dirà *campione*. Un singolo elemento di una popolazione si dice *unità statistica*.

Definizione 1.1.2. Diciamo *carattere* o *variabile* di una popolazione una qualsiasi proprietà o caratteristica degli individui della popolazione che sia omogenea per tutti gli individui e determinabile per ciascuno di essi.

Definizione 1.1.3. Diciamo *modalità* di una variabile l'insieme dei valori distinti che può assumere.

Esempio 1.1.1. L'insieme degli studenti di questa classe costituisce una popolazione; la distanza della scuola da casa (arrotondata al metro) è un carattere che come modalità ha i numeri interi compresi fra 0 e 50000 (supponendo che nessuno abiti più lontano). Un singolo studente è una unità statistica.

Esempio 1.1.2. *Esempi di possibili popolazioni.* Gli studenti di questa classe rispetto ai loro voti in matematica. Gli stessi studenti ripetto ai loro voti in matematica e in italiano. Gli stessi rispetto alla loro provenienza. Gli stessi rispetto al loro sesso. L'insieme di tutti i tavoli su cui siete seduti rispetto al loro stato di manutenzione (buono, mediocre, cattivo).

Se analizziamo gli studenti di questa classe rispetto ai loro voti in matematica per trarne conclusioni su tutti gli studenti della scuola rispetto ai loro voti in matematica avremmo scelto un *campione* (poco significativo per la verità).

Esercizio 1.1.1. Descrivere una possibile popolazione statistica rispetto a, ripettivamente, uno, due, molti caratteri.

Definizione 1.1.4. Analizzando un campione, se ci proponiamo solo di descrivere le caratteristiche salienti della corrispondente popolazione, allora si parlerà di *statistica descrittiva*, se ci si propone invece di trarre importanti conclusioni sulla popolazione si parlerà di *statistica induttiva* o di *statistica inferenziale*. Dato che le inferenze non possono mai essere certe, allora esse sono spesso espresse in termini di un particolare linguaggio matematico che si chiama *probabilità* .

1.2 Dati

I dati si presentano in molte forme diverse. Distinguiamo due casi fondamentali:

Definizione 1.2.1. Diciamo che un dato è *numerico* o *quantitativo* se le variabili assumono valori numerici e *qualitativi* se le variabili assumono come valori qualità non numeriche. Essi si distinguono in *connessi o ordinabili* se si possono comunque ordinare (es. titolo di studio) e in *sconnessi o non ordinabili* se non hanno alcun ordine sensato (es. colore degli occhi). Una variabile si dirà *continua* se può assumere qualsiasi valore compreso fra due numeri dati (anche infiniti) e *discreta* negli altri casi. Se una variabile asssume valori qualitativi allora si dice anche una *mutabile*.

Per indicare variabili numeriche, spesso si usano le lettere X,Y ecc.

Esempio 1.2.1. Tipi di dati:

1. Il colore degli occhi degli alunni di questa classe: carattere qualitativo.
2. Numero di azioni vendute in un giorno alla Borsa di Milano: carattere numerico discreto.
3. Tempo di vita di un hard-disk: carattere numerico continuo.
4. Le capitali europee: carattere qualitativo.
5. Numero di teste nel gettare una moneta: carattere numerico discreto.

Esempio 1.2.2. Modalità:

1. Il colore degli occhi degli alunni di questa classe: uno dei possibili colori dello spettro.
2. Numero di azioni vendute in un giorno alla Borsa di Milano: numero intero da 0 a $+\infty$.
3. Tempo di vita di un hard-disk: numero decimale compreso fra 0 $+\infty$ (?).

4. Le capitali europee: uno fra i valori: Roma, Parigi, Vienna, Berlino, ...
5. Numero di teste nel gettare una moneta: numero intero compreso fra 0 e numero delle volte che si è gettata la moneta.

Esercizio 1.2.1. Stabilire il tipo di dato nei seguenti casi:

1. Temperatura misurata nelle varie località italiane
2. Nuovi iscritti alla nostra scuola negli ultimi 10 anni
3. Numero di auto che passano sulla A14 nei giorni di una settimana
4. Quantità di merci che transitano sulla A14 nei giorni della settimana
5. Millimetri di pioggia caduta a S.Donà nei vari mesi dell'anno
6. Numero delle banconote da 10 euro circolanti nei vari giorni dell'anno

Esercizio 1.2.2. Risolvere:

1. Descrivere una popolazione e una variabile per ognuno dei seguenti casi:
   (a) variabile qualitativa
   (b) variabile quantitativa discreta
   (c) variabile quantitativa continua
2. Per ciascuna delle popolazioni dell'esercizio 1., stabilire quali sono le possibili modalità.

1.3 Frequenze

I dati raccolti per una indagine statistica sono normalmente in forma grezza (raw) senza alcun ordine particolare se non quello derivato dal metodo di raccolta dei dati stessi. Se i dati sono numerici, conviene ordinarli in modo crescente o decrescente e disporli in quello che si chiama vettore; questo permette di individuare subito il valore massimo e il valore minimo mentre la differenza dei due ci darà la gamma o intervallo di variazione (range).

Definizione 1.3.1. Il numero di unità statistiche che presentano una stessa modalità, uno stesso valore, si dice *frequenza assoluta* di quella modalità. Se dividiamo una frequenza assoluta per il numero totale di unità statistiche otteniamo quella che si chiama *frequenza relativa*; rispetto alle assolute queste ultime hanno il vantaggio di poter confrontare anche distribuzioni basate su numeri diversi di unità statistiche. Se dividiamo le modalità in intervalli (regolari o non) e, per ogni intervallo, sommiamo le frequenze che ricadono nell'intervallo otteniamo una distribuzione in *classi* o *categorie*. In ogni caso l'insieme delle coppie ordinate (modalità,frequenza) o (intervallo,frequenza) si dice *distribuzione di frequenze*. Il raggruppamento in classi è assolutamente necessario per le variabili continue.

Esempio 1.3.1. Voti di maturità (sorgente: invenzione):

72 70 55 94 89 84 82 85 73 73
80 76 76 72 85 63 89 74 77 65
65 72 86 77 85 82 67 91 69 63
62 78 67 74 64 68 86 73 91 84
73 69 79 76 66 88 91 75 94 82

In questa tabella sono raccolti i voti di maturità di 50 studenti; ricordiamo che i voti vengono espressi in 100-esimi, che il voto minimo per essere promossi è 60 e che il voto minimo è 30 poichè, con voto inferiore, non si verrebbe ammessi all'orale e quindi alla valutazione finale. Vogliamo rispondere alle seguenti domande:

- Qual è il voto minimo
- Qual è il voto massimo

- Qual è la gamma o intervallo di variazione (range)
- Quanti studenti hanno ottenuto più di 85 punti
- Quanti studenti hanno ottenuto un punteggio compreso fra 60 e 70 escluso
- Quale percentuale di studenti sono stati promossi (voto ≥ 60)
- Quali punteggi non compaiono affatto nella tabella

I voti sono disposti in modo disordinato (per esempio l'ordine alfabetico degli studenti) e non è facile rispondere. Ordiniamo i dati in un vettore:

55 65 68 72 73 76 79 84 86 91
62 65 69 72 74 76 80 84 86 91
63 66 69 73 74 77 82 85 88 91
63 67 70 73 75 77 82 85 89 94
64 67 72 73 76 78 82 85 89 94

In questo vettore i dati sono ordinati in modo crescente e si riconoscono i seguenti fatti:

- Il valore minimo è 55
- Il valore massimo è 94
- La gamma o intervallo (range) è $94 - 55 = 39$
- I punteggi che non compaiono nella tabella sono: 56,57,58,59,60,61,71,81,83,87,89,90,92,93

Nonostante la migliore leggibilità dei dati, è comunque difficile rispondere alle restanti domande. Cerchiamo di suddividere i dati in classi (di voto).

[55,60) 1
[60,65) 4
[65,70) 8
[70,75) 10
[75,80) 8
[80,85) 6
[85,90) 8
[90,95] 5

Osserviamo, ad esempio, che fra 55 (compreso: notare la parentesi quadra) e 60 (escluso: notare la parentesi rotonda) vi è un solo caso, mentre vi sono 5 studenti con voto superiore o uguale al 90.

Questo tipo di classificazione è stato ottenuto suddividendo i dati in classi di ampiezza costante = 5 voti. In molti casi può essere conveniente avere intervalli con estremi diversi; nel nostro esercizio, per esempio:

[30,60) 1
[60,70) 12
[70,80) 18
[80,85) 6
[85,90) 8
[90,100] 5

In questa classificazione l'intervallo è di 10 voti tranne il primo che rileva gli studenti che sono stati respinti (non è possibile prendere meno di 30 e si è respinti con un voto minore di 60) e l'intervallo 80-90 che è stato scomposto in 80-85, 85-90 ad hoc per rispondere ad una specifica domanda. Siamo in grado, ora, di rispondere alle restanti domande:

- Gli studenti che hanno ottenuto più di 85 punti sono 8+5=13
- Gli studenti che hanno ottenuto un voto compreso fra 60 e 70 escluso sono 12
- La percentuale di studenti promossi = promossi/totale * 100 = 49/50*100 = 98%

Per rispondere all'ultima domanda e a domande simili, avremmo potuto servirci di una tabella di frequenze relative e percentuali:

Intervallo	**Frequenza**	**Freq. Rel.**	**Freq.Perc.**
$[30, 60($	1	0.02	2%
$[60, 70($	12	0.24	24%
$[70, 80($	18	0.36	36%
$[80, 85($	6	0.12	12%
$[85, 90($	8	0.16	16%
$[90, 100]$	5	0.1	10%
Totali	50	1.00	100%

Nella prima colonna le frequenze assolute, nella seconda le relative e nella terza le percentuali. Si può notare che la somma della prima colonna è 50 = numero delle unità statistiche, la somma della seconda è 1 = totale frequenze relative e la somma dell'ultima è 100 = totale percentuali. Osserviamo che è stato respinto solo il 2% degli studenti e quindi promosso il restante 98%.

Se dovessimo confrontare questi dati con un'altra statistica condotta, per esempio, in un'altra scuola su un numero di unità statistiche diverso, poniamo 100 studenti, è ovvio che le frequenze assolute sarebbero di difficile confronto mentre quelle relative si potrebbero confrontare direttamente.

Spesso risulta utile *accumulare* le frequenze per avere informazioni ulteriori sull'andamento dei dati; ovviamente ciò ha senso solo per dati ordinabili (variabili numeriche o connesse); allora diamo la seguente:

Definizione 1.3.2. La frequenza totale di tutti i valori minori od uguali dell'estremo superiore di una classe si dice *frequenza cumulativa*. Una tabella che presenti queste frequenze cumulate si dice *distribuzione cumulativa*. Se pensiamo alla frequenza cumulativa come a una funzione calcolata nel punto x si ha:

$$\text{frequenza cumulativa calcolata in } x = \sum \text{numero delle osservazioni minori od uguali a } x$$

se dividiamo la frequenza cumulata per il totale delle osservazioni si ha la *funzione di ripartizione o cumulata relativa*:

$$\text{funzione di ripartizione empirica calcolata in } x = \sum \frac{\text{numero delle osservazioni minori od uguali a } x}{\text{numero totale delle osservazioni}}$$

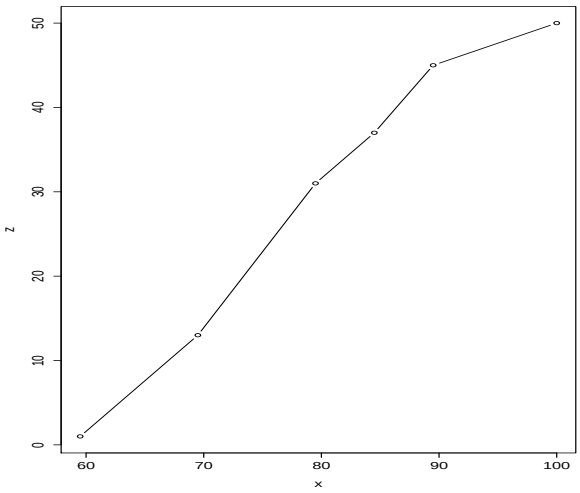
Esempio 1.3.2. Riprendiamo l'esercizio precedente: la distribuzione cumulativa diventa:

x	Frequenza cumulata
59.5	1
69.5	13
79.5	31
84.5	37
89.5	45
100	50

mentre la funzione di ripartizione empirica è:

x	Funzione ripartizione
59.5	0.02
69.5	0.26
79.5	0.62
84.5	0.74
89.5	0.90
100	1.00

possiamo anche riportare i dati in grafico della funzione di ripartizione empirica:



notiamo che la funzione di ripartizione o la frequenza cumulata permettono di rispondere a quesiti del tipo: quanti studenti hanno ottenuto un voto inferiore o ugale a . . . , quale percentuale di studenti ha ottenuto un voto inferiore o uguale a . . . .

Esercizio 1.3.1. Per ciascuno dei seguenti insiemi numerici:

1. 52 -27 36 46 13 33 60 38 41 16 16 28 95 26 21
2. 16 -18 -75 -14 18 0 -42 44 -9 -7 2 -69 8 -39 -23
3. 0.08794 -1.08294 0.33255 0.05802 -1.26005 -0.21573 -1.09427 -0.76137 -0.26398 -0.21164 0.54629 -0.36293 0.27844 0.04288 -0.18899

riordinarli in un vettore in modo crescente e determinarne la gamma di variazione

Esercizio 1.3.2. La seguente tabella rappresente i voti di laurea di 100 studenti in una università; i voti sono in 110-esimi.

87 90 94 97 88 89 92 87 87 94
110 88 76 83 86 83 91 99 94 103
99 85 104 101 105 85 91 94 95 88
93 107 97 83 84 82 101 102 82 98
93 89 96 104 94 87 90 85 89 83
84 87 96 99 74 97 100 88 97 110
91 95 90 97 77 102 92 72 95 83
98 97 95 79 91 88 99 92 92 89
83 87 92 80 85 105 72 95 87 91
85 104 93 104 92 79 92 96 86 75

Tenendo conto dell'esempio 1.3.1, rispondere alle seguenti domande:

1. Qual è il voto minimo
2. Qual è il voto massimo
3. Qual è la gamma o intervallo di variazione (range)
4. Quanti studenti hanno ottenuto più di 85 punti
5. Quanti studenti hanno ottenuto un punteggio compreso fra 90 e 100 escluso

6. Quale percentuale di studenti hanno ottenuto un voto superiore al 100
7. Quali sono i punteggi dei 5 migliori studenti
8. In quale classe vi è frequenza più alta
9. In quale classe vi è la frequenza più bassa
10. Vi è una classe con frequenza relativa superiore al 50%

Esercizio 1.3.3. La seguente tabella rappresenta il peso (in kg arrotondato al grammo, il punto indica la virgola dei decimali) di 50 studenti di classe III in un Istituto Tecnico.

53.833 62.752 79.682 69.82 58.581
42.841 57.996 81.399 85.438 50.478
77.387 63.522 58.299 78.851 92.878
51.594 61.696 75.356 53.705 77.727
61.128 48.409 60.9 54.339 54.369 50.719
89.409 70.224 87.019 59.2
87.248 62.867 50.846 58.155 52.682
60.223 61.209 27.279 73.737 35.481
70.571 67.878 77.647 70.528 65.326
56.063 67.629 60.969 44.019 85.169

Costruire una tabella di distribuzioni di frequenze assolute; creare una tabella di distribuzioni assolute, relative e cumulate suddividendo in opportune classi a propria scelta; scrivere le proprie conclusioni sulle abitudini nutrizionali di questa classe.

Esercizio 1.3.4. La seguente tabella contiene le prime 200 cifre di π compresa la parte intera.

3 1 4 1 5 9 2 6 5 3 5 8 9 7 9 3 2 3 8 4
6 2 6 4 3 3 8 3 2 7 9 5 0 2 8 8 4 1 9 7
1 6 9 3 9 9 3 7 5 1 0 5 8 2 0 9 7 4 9 4
4 5 9 2 3 0 7 8 1 6 4 0 6 2 8 6 2 0 8 9
9 8 6 2 8 0 3 4 8 2 5 3 4 2 1 1 7 0 6 7
9 8 2 1 4 8 0 8 6 5 1 3 2 8 2 3 0 6 6 4
7 0 9 3 8 4 4 6 0 9 5 5 0 5 8 2 2 3 1 7
2 5 3 5 9 4 0 8 1 2 8 4 8 1 1 1 7 4 5 0
2 8 4 1 0 2 7 0 1 9 3 8 5 2 1 1 0 5 5 5
9 6 4 4 6 2 2 9 4 8 9 5 4 9 3 0 3 8 1 9

Studiare la distribuzione di frequenze delle singole cifre.

1.4 Grafici

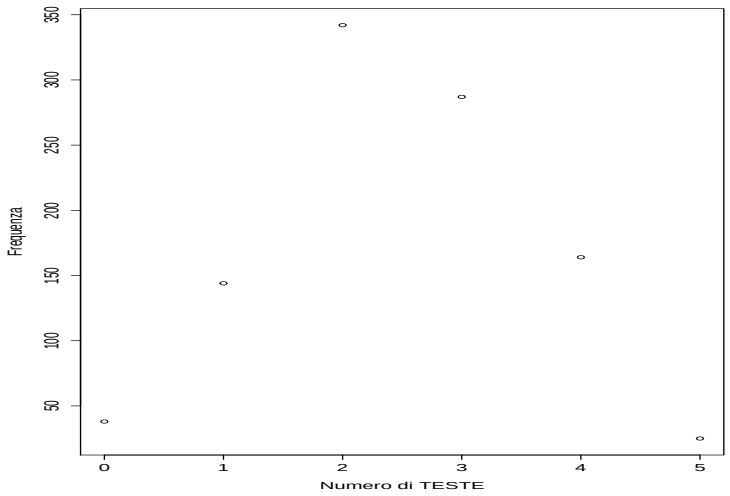
I grafici hanno lo scopo di rappresentare i dati ma la loro utilità maggiore consiste nel far cogliere le differenze nelle distribuzioni di frequenza. Vi sono molte tipologie di grafici possibili, ciascuna con aspetti positivi e negativi: la scelta dipenderà dagli obiettivi posti e dalla tipologia di dati. Vediamone alcune.

Esempio 1.4.1. Grafici cartesiani

Numero di teste	**Frequenza**
0	38
1	144
2	342
3	287
4	164
5	25
Totali	1000

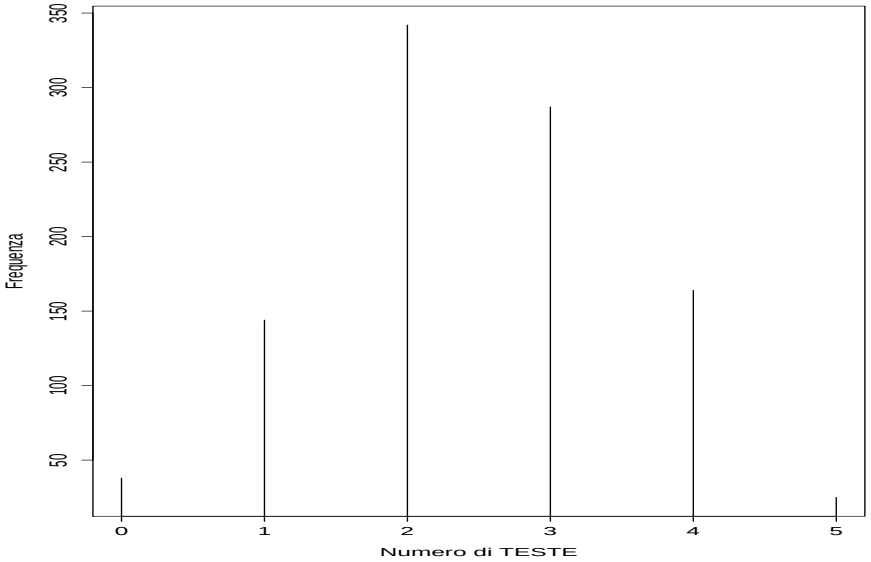
In questa tabella abbiamo i risultati del lancio di 5 monete perfettamente uguali (si presume) per 1000 volte; nella prima colonna il numero di teste da osservare, nellla seconda il numero di di teste effettivamente registrato. Per esempio, sono uscite 3 teste (e quindi 2 croci) 287 volte su 1000.

Il primo grafico che possiamo fare è un semplice *grafico cartesiano*.



Si può osservare che il grafico corrisponde perfettamente al grafico di un prodotto cartesiano che avete già studiato in precedenza. Ogni punto (o cerchietto) corrisponde ad una coppia (x, y) con $x \in$ prima colonna e $y \in$ seconda colonna. Possiamo osservare che, nella maggior parte dei casi, sono uscite 2 o 3 teste su 5.

Il grafico che segue è simile e si dice *grafico (o diagramma) a bastoncini* o a *colonne*.



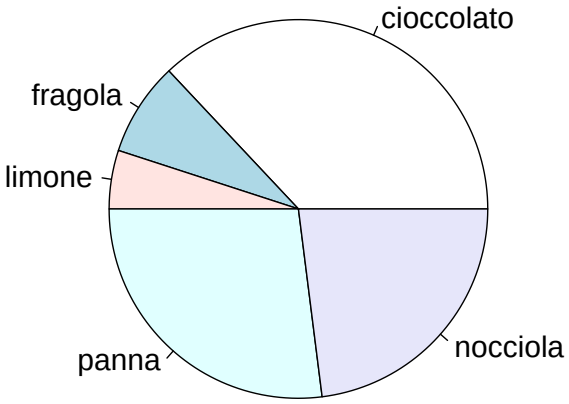
La sostanza del grafico non cambia ma la leggibilità è molto maggiore.

In presenza di distribuzioni di frequenze relative o percentuali è opportuno utilizzare il *grafico o diagramma a torta* , in quanto esso evidenzia bene la proporzione tra le varie parti rispetto al totale. Se i dati sono qualitativi allora risulta spesso più opportuno un *grafico o diagramma a barre* .

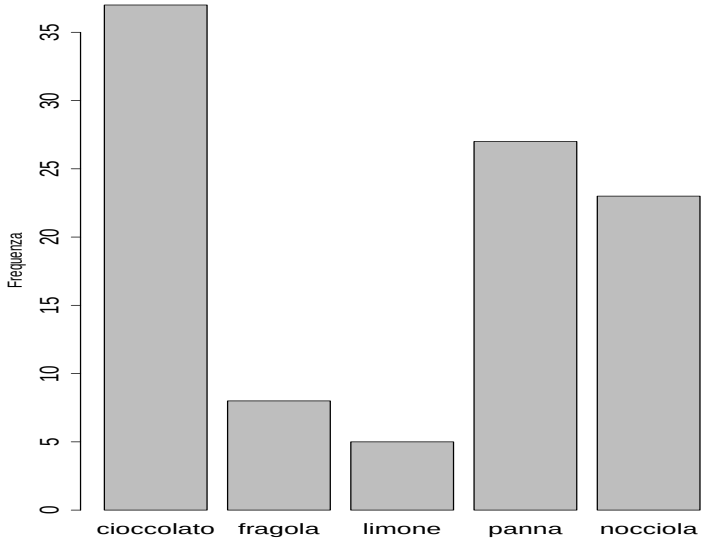
Esempio 1.4.2. Grafici a torta e barre

Gelato	Frequenza
cioccolato	37
fragola	8
limone	5
panna	27
nocciola	23
Totali	100

Essi rappresentano le preferenze di gusto nel gelato da parte di 100 ragazzi. Evidentemente è possibile produrre un grafico cartesiano o un diagramma a bastoncini, ma, in questo caso, è appropriato un diagramma a torta:



L'area del cerchio è suddivisa in modo proporzionale alle frequenze della caratteristica da rappresentare che può essere di natura qualsiasi. È decisamente più chiaro un diagramma a barre:



Notiamo che il grafico a barre è del tutto simile a quello a colonne o bastoncini poichè l'altezza delle barre è proporzionale alla frequenza.

In presenza di variabili di tipo continuo è indispensabile usare il grafico seguente:

Definizione 1.4.1. Un *Istogramma* consiste in un insieme di rettangoli che hanno:

- la base sull'asse orizzontale (asse X) di lunghezza proporzionale alla dimensione dell'intervallo
- area proporzionale alle frequenze delle classi o alle frequenze relative.

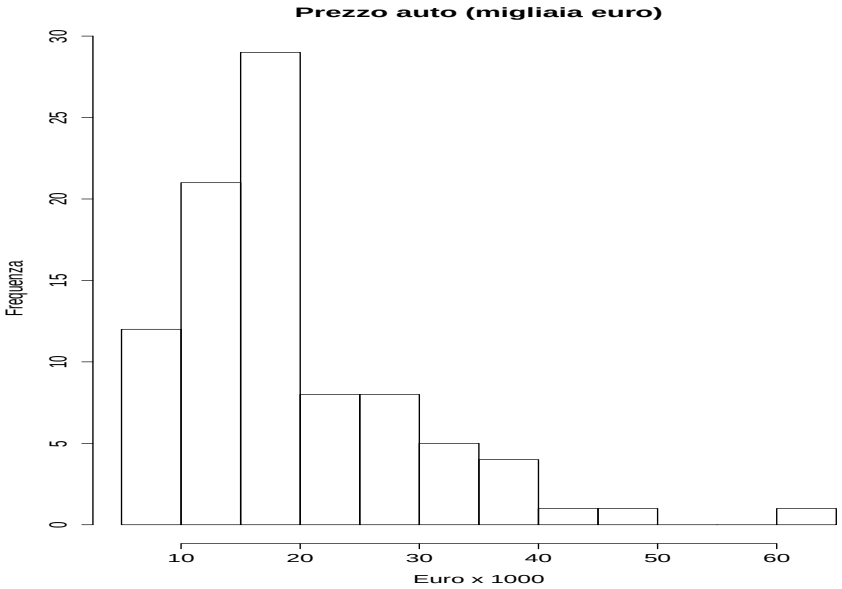
Se gli intervalli di classe hanno ampiezza costante, allora le altezze dei rettangoli sono proporzionali alle frequenze delle classi e quindi si assumono come valori sull'asse verticale proprio le frequenze. Se gli intervalli sono di ampiezza diversa allora si assume sull'asse verticale la *densità*, definita come

$$(\text{densità di un intervallo}) = \frac{\text{frequenza dell'intervallo}}{\text{lunghezza dell'intervallo}}$$

in modo che l'area del rettangolo rappresenti sempre la frequenza osservata.

Esempio 1.4.3. Istogrammi: La seguente tabella contiene i prezzi di alcune automobili in migliaia di euro:

15.9 40.1 15.8 7.4 19.8 34.3 14.9 20.7 10.9
33.9 13.4 29.5 10.1 12.1 36.1 10.3 14.4 19.5
29.1 11.4 9.2 11.3 17.5 8.3 26.1 9 8.6
37.7 15.1 11.3 15.9 8 11.6 11.8 11.1 9.8
30 15.9 13.3 14 10 16.5 15.7 17.7 18.4
15.7 16.3 19 19.9 10 19.1 19.1 18.5 18.2
20.8 16.6 15.6 20.2 13.9 32.5 21.5 24.4 22.7
23.7 18.8 25.8 20.9 47.9 31.9 13.5 28.7 9.1
26.3 38 12.2 8.4 28 61.9 16.3 11.1 19.7
34.7 18.4 19.3 12.5 35.2 14.1 19.5 8.4 20



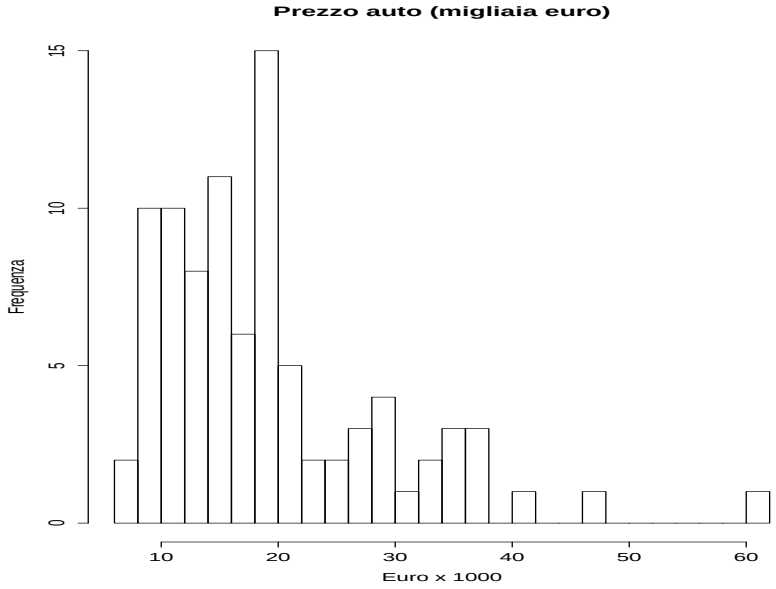
Sull'asse delle ascisse compaiono i prezzi organizzati in 24 intervalli di 5 migliaia di euro ciascuno; la scala va da 5 a 65 mila perché il minimo valore in tabella è 7,4 e il massimo è 61,9. Sull'asse delle ordinate compaiono le frequenze assolute. Osserviamo che il maggior numero di auto si colloca nella fascia di prezzo compresa fra 10 e 20 mila euro.

Esercizio 1.4.1. Verificare che il diagramma è corretto. Sono necessari i seguenti passi:

1. ordinare i dati in un vettore
2. suddividere in 12 classi e calcolare le rispettive frequenze
3. verificare che le altezze dei rettangoli sono proproziaonali alle frequenze

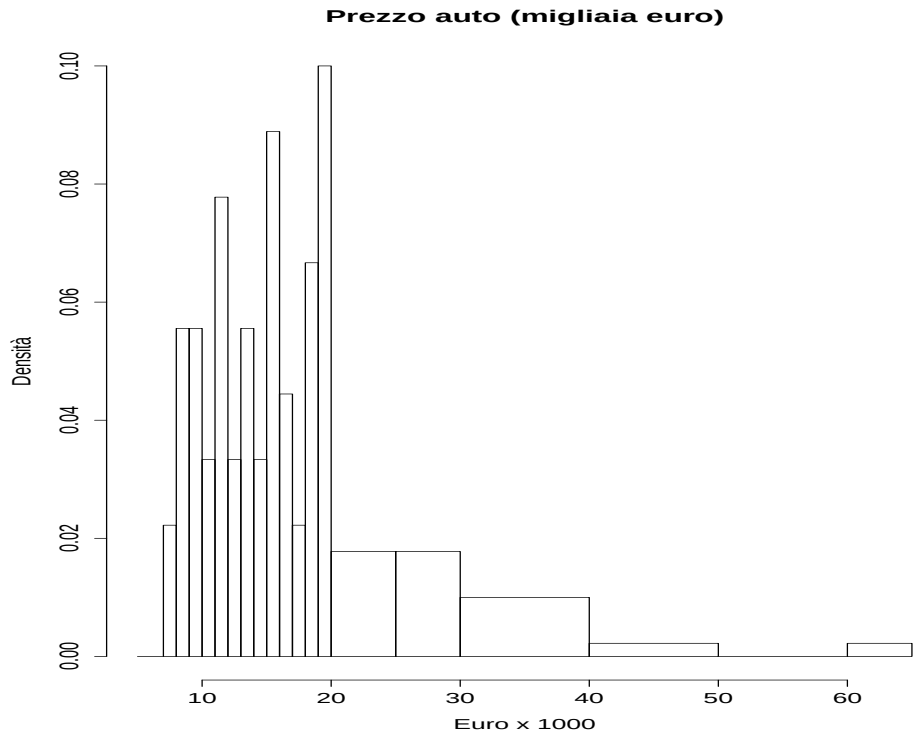
La scelta del numero delle classi è determinata dalla leggibilità e utilità del grafico corrispondente: in altre parole, se necessario, si suddividono i dati in classi più numerose o meno numerose.

Esempio 1.4.4. Continua esercizio precedente



In questo caso gli intervalli sono 24 e il corrispondente andamento dei prezzi è più chiaro: anche le auto di prezzo compreso nella fascia 5-10 mila euro sono in numero elevato; nel grafico precedente non si notava perché il dato veniva mediato dal basso valore della fascia precedente. Concludiamo che una suddivisione più fine degli intervalli può evidenziare proprietà dei dati che non sono immediatamente evidenti.

Vi è anche la possibilità di suddividere in intervalli di larghezza variabile, magari per evidenziare aree che si ritengono interessanti senza avere troppi intervalli da visualizzare. Nel nostro esempio:



La fine suddivisione degli intervalli nell'area 5-20 mila euro è compensata da una più ampia nelle altre zone di minor interesse. Notare come sull'asse delle ordinate non compaiono più le frequenze assolute ma le densità e il motivo è evidente.

Esercizio 1.4.2. La tabella seguente:

Colore	**Frequenza**
nero	18
biondo	5
castano	24
rosso	3
Totali	50

rappresenta il colore dei capelli di una polazione di maschi italiani. Costruire un grafico:

1. cartesiano
2. a torta
3. a bastoncini

Esercizio 1.4.3. La tabella seguente:

Lunghezza (cm)	Frequenza
1.2	4
1.3	7
1.4	10
1.5	12
1.6	10
1.7	6
1.8	1
Totali	50

rappresenta il risultato di 50 misurazioni dello stesso pezzo meccanico. Dopo aver suddiviso opportunamente le misure in classi di lunghezza uguale, costruire un grafico:

1. cartesiano
2. istogramma

Costruire un istogramma delle frequenze relative.

1.5 Sommatoria

In questo paragrafo facciamo una piccola disgressione sulla notazione matematica utilizzata per indicare le somme.

In generale, se dobbiamo indicare una somma lo facciamo elencandone tutti i termini: per esempio

$$a = 1 + 2 + 3 + 4 + 12 + 23$$

ma se gli addendi non sono noti e quindi sono indicati con lettere e in numero variabile, come ad esempio

$$a = f_1 + f_2 + f_3 + \cdots + f_n$$

questa notazione è utile ed è usata frequentemente sia in matematica che in statistica ma ha un problema: se dobbiamo indicare molte somme del genere e dobbiamo combinarle insieme in formule più complesse, allora la cosa diventa difficile da manipolare e da capire. Introduciamo allora la seguente:

Definizione 1.5.1.

$$a = \sum_{i=1}^{n} f_i = f_1 + f_2 + f_3 + \cdots + f_n$$

che si legge: a è uguale alla sommatoria (o somma) per i che va da 1 a n di f con i. In altre parole: il simbolo $\sum$, che si legge *sommatoria*, rappresenta la somma degli elementi f_i iniziando dal valore indicato sotto il simbolo, sino al valore indicato sopra.

Allora, per esempio

$$b = g_1 + g_2 + g_3 + \cdots + g_k$$

si scrive

$$b = \sum_{i=1}^{k} g_i$$

Esempio 1.5.1. Alcune sommatorie:

$$\sum_{j=1}^{5} j = 1 + 2 + 3 + 4 + 5 = 15$$

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

Vediamo alcune proprietà utili della sommatoria:

Teorema 1.5.1.

$$\sum_{i=1}^{k}(y_i + x_i) = \sum_{i=1}^{k} y_i + \sum_{i=1}^{k} x_i$$

In sostanza è come una specie di proprietà distributiva del simbolo $\sum$ sulla somma. In pratica è una applicazione delle proprietà associativa e commutativa.

Dimostrazione. Abbiamo

$$\sum_{i=1}^{k}(y_i + x_i) = (y_1 + x_1) + (y_2 + x_2) + \cdots + (y_k + x_k) = \quad \text{applico commutativa e associativa più volte}$$

$$= (y_1 + y_2 + \cdots + y_k) + (x_1 + x_2 + \cdots + x_k) = \sum_{i=1}^{k} y_i + \sum_{i=1}^{k} x_i$$

□

Teorema 1.5.2.

$$\sum_{i=1}^{k} a y_i = a \sum_{i=1}^{k} y_i$$

Una costante si può portare fuori dalla sommatoria.

Dimostrazione. Esercizio. □

Esercizio 1.5.1. Trovare il valore delle seguenti somme:

1. $$\sum_{i=1}^{10} 2$$
2. $$\sum_{i=1}^{10} i^2$$
3. $$\sum_{i=1}^{10} 2^i$$
4. $$\sum_{i=3}^{6} (i-1)$$
5. $$\sum_{i=-2}^{2} i^2$$

1.6 Indici di sintesi

Nel valutare una certa distribuzione di frequenze ma più spesso nel confrontare due diverse distribuzioni di frequenze, si è portati a cercare qualche elemento che permetta una valutazione immediata delle diverse posizioni dei dati. In sostanza sarebbe utile avere un unico numero che permetta un confronto diretto della *posizione* della distribuzione. Alcuni di questi numeri sono:

- media aritmetica
- mediana
- moda
- quantili

Media aritmetica

Definizione 1.6.1. Supponiamo che in una statistica i valori rilevati siano indicati con $y_1, y_2, \cdots y_n$. Allora la *media aritmetica* (ingl. *mean*) dei dati è il numero

$$\bar{y} = \frac{y_1 + y_2 + \cdots + y_n}{n} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

La media aritmetica è solo una delle possibili medie ma è la più usata e quindi viene spesso chiamata semplicemente la *media*.

Esempio 1.6.1. La media aritmetica:
La media della sequenza $8, 3, 5, 12, 10$ è

$$\bar{y} = \frac{8+3+5+12+10}{5} = \frac{38}{5} = 7.6$$

Esempio 1.6.2. Consideriamo la sequenza:

$$5, 5, 5, 8, 8, 6, 6, 6, 6, 2$$

la media è

$$\bar{y} = \frac{5+5+5+8+8+6+6+6+6+2}{10} = \frac{57}{10} = 5.7$$

considerando che i valori sono ripetuti, possiamo sintetizzarli nella tabella:

Valori	**Frequenza**
5	3
8	2
6	4
3	1

quindi possiamo calcolare la media raggruppando i valori con la stessa frequenza:

$$\bar{y} = \frac{5\cdot 3 + 8\cdot 2 + 6\cdot 4 + 2}{3+2+4+1} = \frac{15+16+24+2}{10} = \frac{57}{10} = 5.7$$

L'ultimo esempio ci suggerisce la seguente:

Definizione 1.6.2. Se in una statistica i valori $y_1, y_2, \cdots, y_n$ compaiono con le rispettive frequenze $f_1, f_2, \cdots, f_n$, allora la media aritmetica dei dati è il numero

$$\bar{y} = \frac{f_1x_1 + f_2x_2 + \cdots + f_nx_n}{f_1 + f_2 + \cdots + f_n} = \frac{\sum_{i=1}^{n} f_ix_i}{\sum_{i=1}^{n} f_i} = \frac{\sum_{i=1}^{n} f_ix_i}{n}$$

dove

$$n = \sum_{i=1}^{n} f_i$$

La media, ovviamente, si calcola solo in presenza di dati numerici; osserviamo anche che, in presenza di dati estremi molto diversi dalla maggioranza dei valori rilevati, la media perde gran parte della sua utilità, come si può dedurre dal seguente sempio:

Esempio 1.6.3. Consideriamo la sequenza:

$$1, 1, 1, 1, 1, 1, 1, 1, 1, 1000$$

la media è

$$\bar{y} = \frac{1009}{10} = 100.9$$

questo valore, pur corretto dal punto di vista della definizione, non è rappresentativo della maggioranza dei valori della sequenza. In questi casi, è necessario usare la *mediana*, indice non influenzato dai valori estremi.

Mediana

Definizione 1.6.3. Si definisce *mediana* di una distribuzione quel valore che si colloca a metà fra i valori ordinati della distribuzione stessa.

In altre parole, la mediana è un valore che è maggiore o uguale del 50% delle osservazioni e minore o uguale del restante 50%.

Se il numero delle osservazioni è dispari allora la mediana sarà il valore centrale mentre se sono in numero pari la mediana è la media dei due valori centrali.

Esempio 1.6.4. La mediana:
La mediana della sequenza $3, 4, 4, 5, 6, 8, 8, 8, 10$ è 6
La mediana della sequenza $5, 5, 7, 9, 11, 12, 15, 18$ è $\frac{9+11}{2} = 10$

Moda

Definizione 1.6.4. Si definisce *moda* di una distribuzione quel valore che compare con più frequenza nelle osservazioni, cioè il valore più comune.

Se la variabile è continua, la moda è la classe con maggiore densità di frequenza.

La moda però potrebbe non esistere oppure non essere un valore unico e quindi non è una misura molto usata.

Esempio 1.6.5. La moda:
La moda della sequenza $2, 3, 3, 3, 4, 4, 5, 5, 6, 12$ è 3
La moda della sequenza $2, 5, 7, 9, 14$ non esiste o sono tutti i valori osservati.
La moda della sequenza $2, 5, 7, 7, 8, 9, 10, 10$ ha moda 7 e 10 ed è chiamata *bimodale*.

Quantili

Se è interessante conoscere il valore che divide a metà le osservazioni, è molto più interessante conoscere il valore che divide le osservazioni in quarti o in frazioni anche inferiori.

Definizione 1.6.5. Si definisce *primo quartile* di una distribuzione ordinata quel valore che divide i dati in 25% e 75%. Il *secondo quartile* divide la distribuzione in due parti uguali e quindi è la mediana. Si definisce *terzo quartile* di una distribuzione ordinata quel valore che divide i dati in 75% e 25%. In altre parole, i quartili dividono la distribuzione in quarti. In modo analogo si definiscono i *percentili* che sono ovviamente 99. In generale, queste suddivisioni dei dati in parti uguali vengono dette *quantili*.

Esercizio 1.6.1. Trovare la media, mediana e i tre quartili dei dati relativi all'esercizio 1.3.1.

Esercizio 1.6.2. Trovare la media, mediana e i tre quartili dei dati relativi agli esercizi 1.3.1.

Esercizio 1.6.3. Dei dati relativi all'esercizio 1.3.1, trovare anche il percentile corrispondente al 30% (indicato a volte con P_{30}).

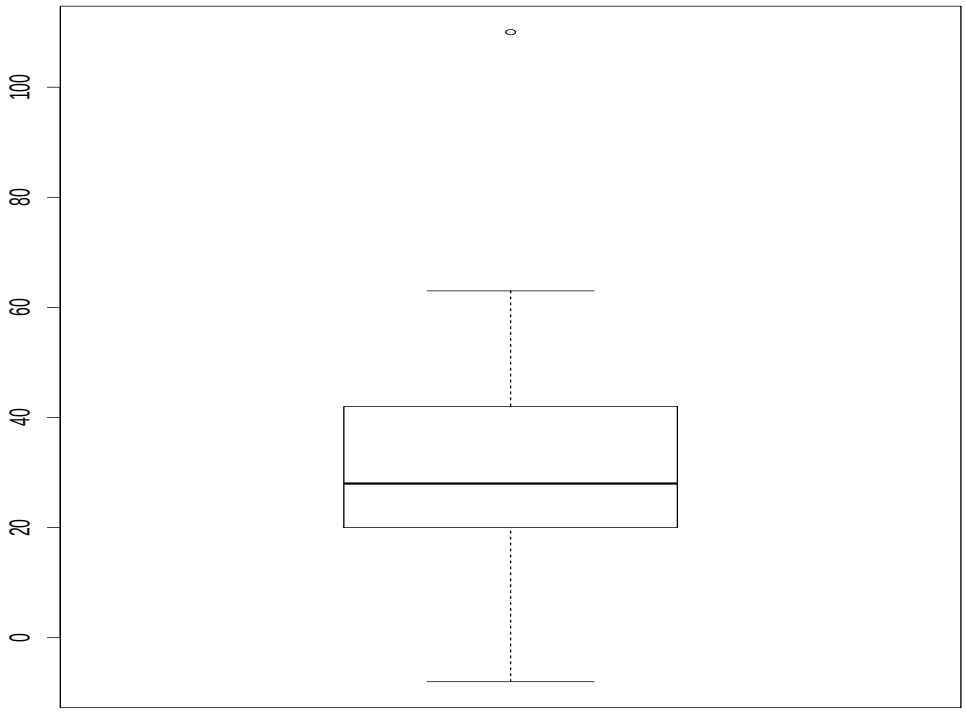
1.7 Diagramma a scatola (boxplot)

Questo diagramma, detto anche *diagramma a scatola con baffi*, riassume tutti gli indici di sintesi importanti di cui abbiamo parlato: valore massimo, valore minimo, primo quartile, mediana, terzo quartile.

Esempio 1.7.1. Boxplot:
Consideriamo i seguenti dati:

-8 15 19 21 25 28 37 38 47 56
-5 17 20 22 26 29 37 39 49 58
1 17 20 22 26 34 38 42 49 60
7 18 20 23 27 35 38 44 51 63
15 18 21 25 28 35 38 44 56 110

Il grafico boxplot:



Il diagramma - partendo da sotto -: linea del minimo valore, base scatola = primo quartile, linea interna scatola = mediana, linea superiore scatola = terzo quartile, linea superiore = massimo, cerchietto all'estremità superiore = dato che si estende troppo lontano dalla massa dei dati (l'algoritmo di calcolo è: distanza maggiore 1,5xs dove s=scarto interquartile, cioè l'altezza della scatola) e questo anche per non avere scatole con baffi esageratamente lunghi.

1.8 Proprietà della media e della mediana

1.9 Misure di variabilità

In quasi tutte le indagini statistiche è importante misurare di quanto i dati rilevati si discostano dalla media. Il grado di questa variazione si chiama *variazione* o *dispersione* dei dati. Vi sono molti possibili indicatori di dispersione ma il più usato è il seguente:

Definizione 1.9.1. Se con $y = (y_1, y_2, \cdots, y_n)$ indichiamo i dati osservati, con n il loro numero e con $\bar{y}$ la loro media aritmetica, cioè

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

allora chiamiamo *varianza* la

$$\text{var}(y) = \text{ varianza}(y_1, y_2, \cdots, y_n) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

in altre parole, la varianza è la media dei quadrati di tutte le differenze rispetto alla media: una misura di quanto i dati sono *distanti* dalla media aritmetica. Osserviamo che le differenze sono elevate al quadrato perchè le differenze potrebbero avere segno opposto e quindi elidersi.

Definizione 1.9.2. Chiamiamo *scarto quadratico medio* la radice della varianza, cioè:

$$\text{sqm}(y) = \text{scarto quadratico medio}(y) = \sqrt{\text{var}(y)}$$

Spesso sqm() viene usato perchè ha la stessa unità di misura dei dati osservati mentre var() ha unità di misura pari al quadrato di quella dei dati.

Altre misure di variabilità (che non hanno grande interesse per noi) sono:

- Campo di variazione (range) (che abbiamo definito in precedenza)
- Scarto interquartile = terzo quartile - primo quartile (molto più resistente della varianza quando si hanno poche osservazioni)

Esercizio 1.9.1. Calcolare la varianza e lo scarto quadratico medio dei dati in tutti gli esercizi precedenti (se ha senso) e darne una interpretazione.

Esercizio 1.9.2. Formula alternativa per il calcolo della varianza. Dimostrare che vale la seguente formula:

$$\text{var}(y) = \left(\frac{1}{n}\sum_{i=1}^{n} y_i^2\right) - \bar{y}^2$$

ovvero:

$$(\text{varianza}) = \left(\frac{\text{media dei}}{\text{quadrati}}\right) - \left(\frac{\text{quadrato della}}{\textit{media}}\right)$$

1.10 Osservazioni sui dati

I dati che raccogliamo non sono sempre *buoni* dati: se sono distorti, insufficienti o corrotti (sbagliati) possono indurre a conclusioni errate.

Poniamoci la domanda: cosa sono i *cattivi dati*? La risposta può essere anche molto complessa ma noi ci limiteremo a descrivere solo due casi: dati *incompleti* o dati *scorretti*.

Dati incompleti

I dati sono incompleti se ne manca qualcuno. In una indagine telefonica qualcuno può non rispondere; in una indagine clinica, un paziente non si presenta al prelievo del sangue; in un questionario, la persona decide di non rispondere alla domanda.

L'incompletezza dei dati può essere conseguenza di un *bias di selezione*[1] cioè un errore nella selezione del campione da sottoporre a test. Per esempio, se vogliamo sapere cosa pensa la popolazione del problema dei finanziamenti alla ricerca in matematica e lo chiediamo ad un campione di laureati in matematica, abbiamo certamente dati poco significativi e, forse decisamente incompleti. Per bias di selezione non si deve intendere un errore materiale nella predisposizione dell'indagine, ma una errata valutazione del campione prescelto rispetto ai caratteri indagati.

Quali accorgimenti si applicano per ridurre il rischio di dati incompleti?

Se i dati mancanti sono pochi, questi semplicemente si tolgono dall'indagine. Ovviamente questa soluzione può presentare inconvenienti anche gravi: riduzione drastica della quantità di dati disponibile per l'analisi, oppure introduzione di un bias di selezione non presente in origine.

Esagerando si potrebbe pensare al caso in cui **tutti** i dati contengano una incompletezza: in tal caso non rimarrebbe alcun dato; oppure in una indagine sulla popolazione, risulta che tutte le risposte del campione femminile siano da rigettare: rimarrebbero solo i dati del campione maschile con gravissimo bias di selezione.

La seconda soluzione al problema dei dati incompleti consiste nell'inserire dei valori sostitutiti; per esempio il valor medio dei dati validi (es. in una indagine sui redditi, la media dei valori stipendiali presenti).

Anche in questo caso bisogna valutare attentamente la singola indagine poichè si stanno falsificando i dati. In qualche caso la cosa potrebbe non avere conseguenze significative (es. in una indagine sulla provenienza geografica degli studenti mancano alcune età anagrafiche: il dato può essere sostituito con la media o con dati generati casualmente), ma in altri casi può condurre ad errori gravi di analisi (es. in una indagine sui redditi, mancano i dati relativi ad una categoria di lavoratori; in questo caso la sostituzione sarebbe completamente arbitraria).

Se si ha il sospetto che i dati mancanti non siano casuali ma ci sia un motivo specifico che abbia indotto qualcuno a non fornirli (es. sui social network non tutte le persone forniscono dati veritieri sull'età), allora bisogna ricorrere a tecniche statistiche più elaborate, ricorrendo ad un modello statistico probabilistico.

La soluzione ottimale del problema dei dati incompleti non esiste, conviene sempre minimizzare il rischio ponendo estrema cura nella raccolta dei dati.

Dati scorretti

I dati possono essere scorretti in moltissimi modi e per moltissime ragioni diverse: lettura di strumenti sbagliata, errori nella digitazione, ecc. Oltre a questo bisogna considerare anche la propagazione degli eventuali errori: se un'azienda basa le sue scelte commerciali su un'analisi sbagliata dell'andamento del mercato, queste potrebbero indurre altri soggetti economici a prendere decisioni errate.

Da queste considerazioni appare chiaro che i dati devono essere accuratamente esaminati e ripuliti prima di ogni analisi statistica.

In presenza di un errore evidente, per esempio un dato decisamente fuori media o inaspettato (che può essere rilevato anche da un boxplot), si possono attuare le misure di correzione discusse in precedenza. É evidente che se i dati errati sono molti si può correggere la situazione solo con l'ausilio di modelli statistici sofosticati e procedure automatizzate.

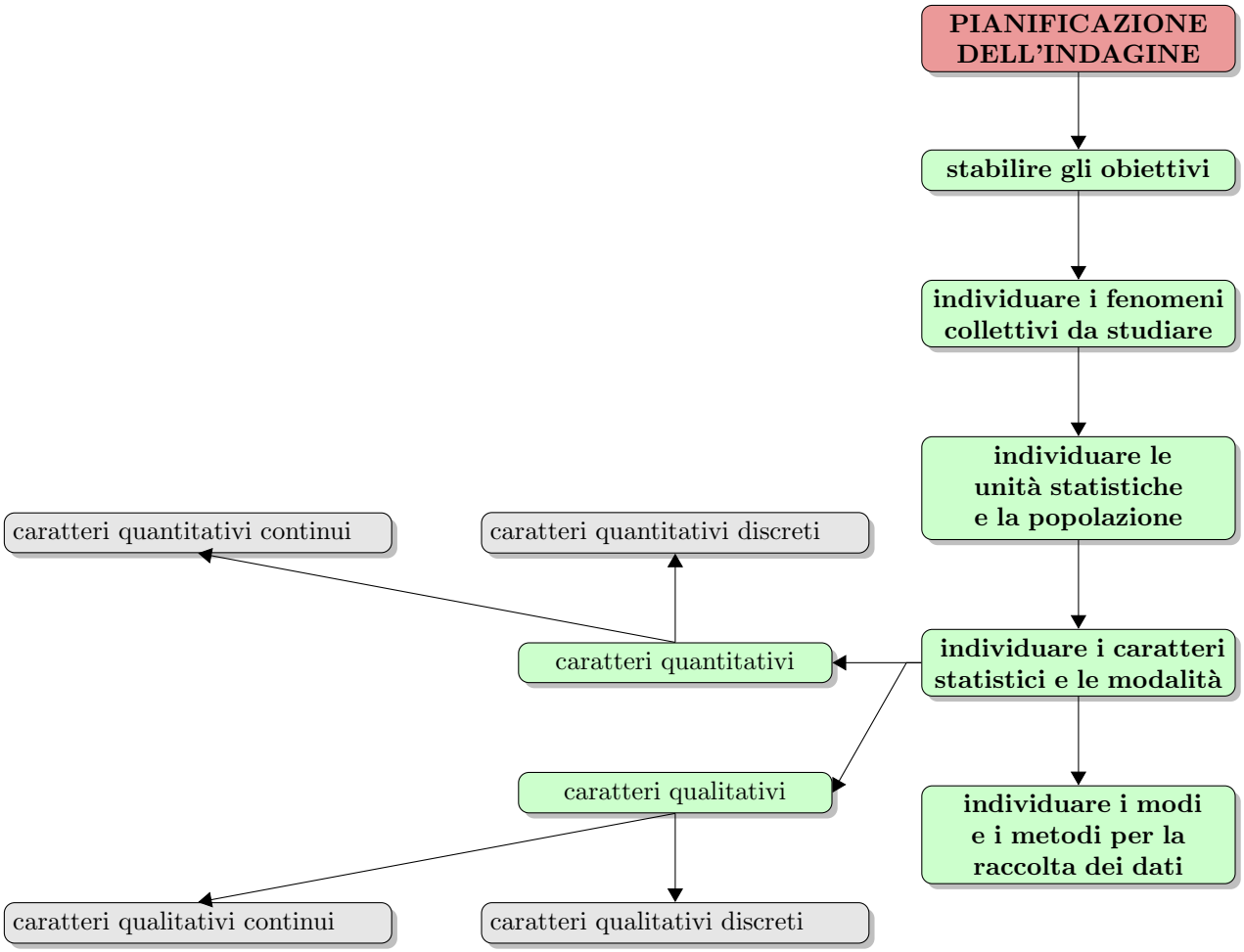
[1] La parola *bias* è di difficile traduzione ma possiamo assumere che significhi *errore*.

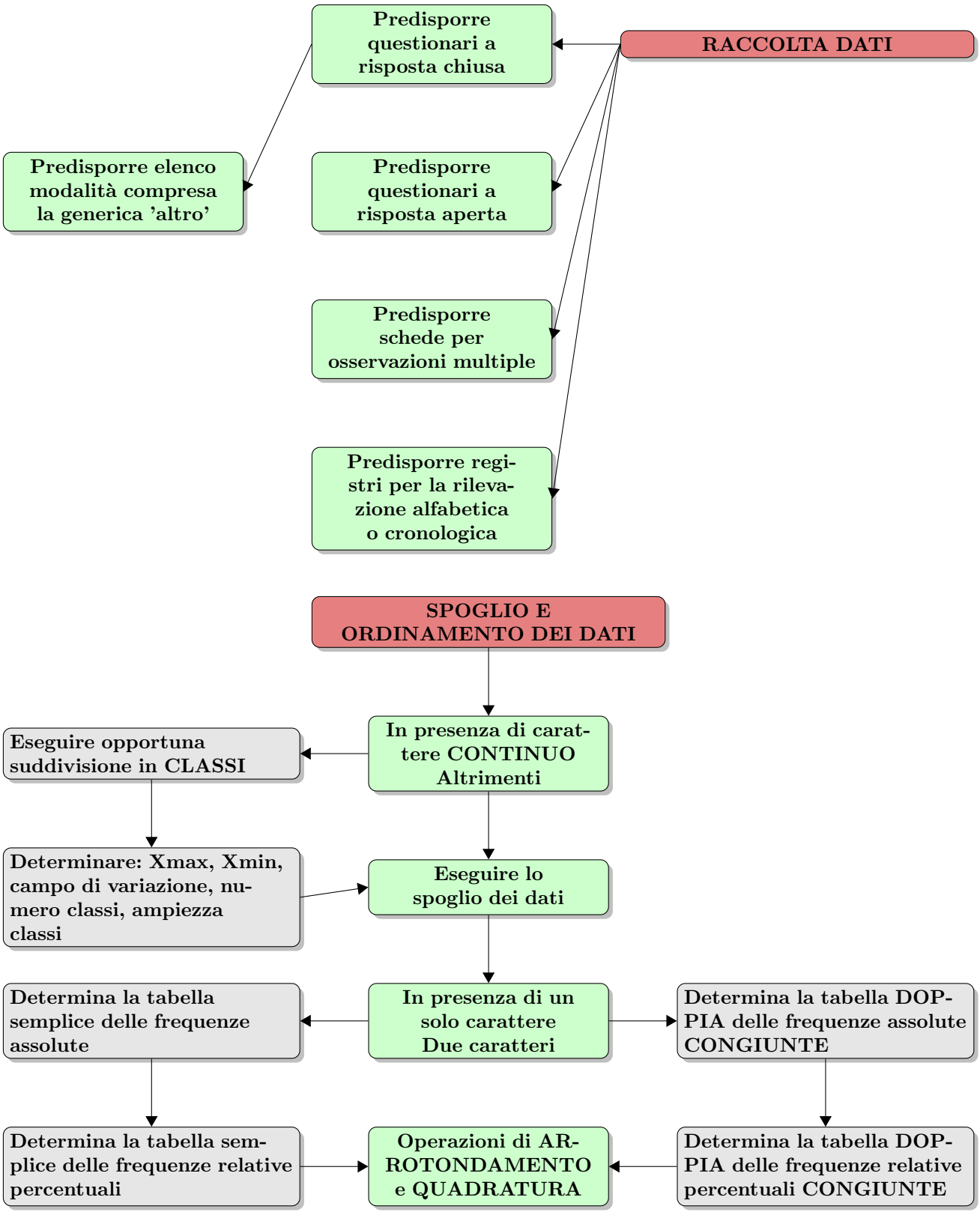
In ogni caso la conclusione finale è che conviene progettare accuratamente la raccolta dati, semplificando il più possibile le domande da porre e rendendole assolutamente non ambigue. Si dovrà curare particolarmente la raccolta dati e il loro successivo inserimento in elaboratori elettronici.

In altre parole **fare di tutto per partire con dati validi**. Come dicono abitualmente gli statistici: *se butti dentro spazzatura, puoi solo tirar fuori spazzatura.*

1.11 Schemi di lavoro

Proponiamo alcuni schemi che forniscono una guida passo-passo per la realizzazione di una ricerca di statistica descrittiva.




RACCOLTA DATI
Predisporre questionari a risposta chiusa
Predisporre elenco modalità compresa la generica 'altro'
Predisporre questionari a risposta aperta
Predisporre schede per osservazioni multiple
Predisporre registri per la rilevazione alfabetica o cronologica
SPOGLIO E ORDINAMENTO DEI DATI
In presenza di carattere CONTINUO Altrimenti
Eseguire opportuna suddivisione in CLASSI
Determinare: Xmax, Xmin, campo di variazione, numero classi, ampiezza classi
Eseguire lo spoglio dei dati
In presenza di un solo carattere Due caratteri
Determina la tabella semplice delle frequenze assolute
Determina la tabella DOPPIA delle frequenze assolute CONGIUNTE
Determina la tabella semplice delle frequenze relative percentuali
Operazioni di ARROTONDAMENTO e QUADRATURA
Determina la tabella DOPPIA delle frequenze relative percentuali CONGIUNTE

1.12 Proposte di ricerca (case study)

In questo paragrafo proponiamo alcuni esempi di ricerca statistica di una certa complessità nei quali è possibile applicare tutti gli elementi di analisi che abbiamo descritto nel testo. Nel progettare la ricerca, nel preparare la rilevazione dei dati, nel raccogliere i dati stessi, nello spoglio delle schede, nell'analisi dei dati e nella preparazione dei grafici finali, è utile seguire gli schemi forniti nel paragrafo precedente.

1.12.1 Indagine statistica sul metodo di studio

FASE UNO: PIANIFICAZIONE

Si vuole indagare come gli studenti di una certa classe affrontano lo studio delle varie discipline. É evidente che non basta porre a ciascuno studente la fatidica domanda: 'Come studi? oppure 'Che metodo di studio adotti?'.

É necessario piuttosto sviscerare la questione evidenziando alcune variabili statistiche che possano dare indicazioni dirette o indirette sul metodo di studio. Tali variabili potrebbero essere definite attraverso domande del tipo :

a) Quante ore studi mediamente al giorno?

b) Quanto di questo tempo dedichi generalmente al ripasso?

c) Ripeti a voce alta?

d) Ricopi gli appunti presi in classe?

e) Durante lo studio, ascolti musica?

f) ...

Ci sono poi delle variabili che possono influire sul metodo di studio in maniera indiretta. Per esempio:

g) Quanto dista la tua scuola da casa?

h) Quanto ci impieghi a raggiungerla?

i) Pratichi sport?

l) Se si, quale? A che livello? Per quante ore alla settimana?

j) Esci con gli amici quotidianamente? Anche dopo cena?

k) ??

E poi si potrebbero indagare alcuni aspetti motivazionali...

m) Quanto ti interessa andar bene a scuola?

n) Sei contento della scuola che hai scelto?

o) ?..

FASE DUE: RACCOLTA DATI

Si procede formulando un questionario da somministrare a ciascuno studente in forma anonima. É preferibile che le domande siano a risposta chiusa, per consentire un più agile spoglio.

Domanda a,b) carattere continuo: proporre intervalli di tempo (meno di un'ora; da 1 a 2 ore . . .)

Domanda c,d,e) variabile qualitativa sconnessa (. . . si/no)

Domanda g) carattere continuo: proporre fasce chilometriche (meno di 3km; da 3 a 10 km, . . .)

Domanda h) carattere continuo: proporre intervalli di tempo (meno di 1/4 d?ora; da 1/4 a 1/2 ora . . .)

Domanda i) carattere qualitativo sconnesso (elenco sport + modalità 'altro') Carattere discreto (2 ore, 4 ore , 6 ore, 8 ore , più di 8 . . .)

Domanda j) carattere qualitativa sconnesso (. . . si/no)

Domanda l,m) carattere qualitativo CONNESSO (per nulla, abbastanza, molto, moltissimo)

FASE TRE : SPOGLIO E ORDINAMENTO

I dati vengono ordinati in tabelle per costruire le distribuzioni di frequenze assolute e percentuali

FASE QUATTRO : RAPPRESENTAZIONE GRAFICA

Si realizzano i grafici più significativi : diagrammi a bastoncini per le distribuzioni di frequenza assoluta su caratteri discreti e qualitativi; istogrammi per i caratteri continui e grafici a torta per le percentuali.

FASE CINQUE : ELABORAZIONE DEI DATI

In questa fase si calcolano alcuni indici di sintesi e di variabilità significativi. Per i dati numerici sicuramente si determinano la media aritmetica, la moda, la mediana, il campo di variazione, lo scarto quadratico medio ed , eventualmente, i quartili, con la conseguente determinazione del grafico box-plot. Per i dati qualitativi connessi si calcolano la moda e la mediana. Per i dati sconnessi solo la moda.

FASE SEI : INTERPRETAZIONE

Dall?analisi delle tabelle, dei grafici e degli indici calcolati si possono trarre conclusioni ed osservazioni utili a formulare un?analisi sul metodo di studio degli studenti indagati.

1.12.2 UCLA Case Studies: Stock Prices

A[2] basic rule of thumb for investors in the stock market is to "diversify"; that is to spread one's money across stocks which are likely to behave differently in response to various conditions in the market. Risk to the investor is reduced because, under a given set of circumstances, some stocks in the portfolio will rise while others fall. How can one determine which stocks are similar and which are not for the purpose of diversification? The data provided are daily stock prices from January 1988 through October 1991[3], for ten aerospace companies. Given this information, the first step toward answering the question posed above is to reformulate the question in terms of these data. For example, two stocks may be considered similar if they maintain approximately the same level, vary to a similar degree, or tend to move up and down in related ways over some relevant time period. An initial analysis might use some graphical techniques to examine these aspects of the data.

[2]UCLA: University California Los Angeles, che ringraziamo per il materiale disponibile in rete all'indirizzo: http://www.stat.ucla.edu/cases/

[3]Da noi opportunamente ridotte per ragioni di spazio.

a) Make histograms of these price series.

b) What information is lost in converting the raw data into histograms ?

c) What is gained ?

Time Plots

Another simple tool for comparing price series over time is the univariate time plot. Plot stock price on day for each of the ten companies for which price series is provided.

d) Are the Y axis scales the same for all plots?

e) What advantages are there in making all scales the same?

f) What are the disadvantages?

Look at the overall shapes of the plots.

g) Can you group the companies according to the shapes?

h) Are these groupings a sensible answer to the question posed above concerning similarity, or should one also consider the level of activity?

i) That is, given two graphs with roughly the same shape, would you consider them similar even if one averaged about 20 dollars and the other about 65?

l) What about variability? How can you assess variability in these graphs?

m) Would a great difference in variability be enough for you to place two otherwise similar stocks in different groups?

Descriptive Statistics

It might also be useful to have one or two numbers that capture relevant characteristics of a stock's behavior. Mean and variance are two descriptive statistics often used to summarize data.

Compute the means of stock prices for Companies A through J.

n) Which company has the highest mean price? The lowest?

o) Does this mean that the company with the higher mean is a better investment than the company with the lower mean?

p) Describe the histograms of the companies with the highest and lowest means.

q) What is different?

r) What is the same?

s) Just by looking at the histogram, which company's stock looks more variable?

t) What does variability mean in the context of stock prices?

u) Two possible measures of variability are variance, and interquartile range. Compute the variance and interquartile range for each company. Which is a better measure of variability, thinking of variability as risk?

v) Do these two measures tell the same story about these two stocks?

I dati per sviluppare il case study si possono reperire nella rete oppure si può scaricare il file 'stock.txt' nel sito della scuola, nell'area riservata al materiale statistico.

La struttura dei dati si può dedurre dalla tabella seguente che può anche essere usata per svolgere l'esercizio.

04JAN88	17.219	50.500	18.750	43.000	60.875	26.375	67.750	19.000	48.750	34.875
05JAN88	17.891	51.375	19.625	44.000	62.000	26.125	68.125	19.125	48.750	35.625
06JAN88	18.438	50.875	19.875	43.875	61.875	27.250	68.500	18.250	49.000	36.375
07JAN88	18.672	51.500	20.000	44.000	62.625	27.875	69.375	18.375	49.625	36.250
08JAN88	17.438	49.000	20.000	41.375	59.750	25.875	63.250	16.500	47.500	35.500
11JAN88	18.109	49.000	19.500	41.875	59.625	26.625	66.250	17.125	47.750	34.375
12JAN88	18.563	49.375	19.125	42.500	60.750	27.250	65.750	16.875	47.875	34.000
13JAN88	18.672	50.125	19.250	43.000	61.750	28.000	66.000	16.875	47.250	34.625
14JAN88	18.563	49.750	19.000	43.250	61.750	29.000	65.750	17.125	47.000	34.875
15JAN88	19.063	50.500	19.125	43.875	61.875	29.625	66.875	17.750	47.375	36.000
18JAN88	19.000	50.250	19.625	44.000	62.125	30.000	66.500	17.375	47.750	35.625
19JAN88	19.063	49.750	20.000	44.375	61.250	29.875	66.500	16.875	48.000	35.375
20JAN88	18.719	49.250	19.000	43.500	60.375	29.000	65.875	16.500	48.000	34.500
21JAN88	18.438	49.250	18.375	43.375	60.375	29.000	65.000	16.500	47.500	34.875
22JAN88	19.063	50.250	18.375	43.500	60.375	29.125	65.750	16.375	47.875	36.625
25JAN88	20.000	50.250	18.125	44.000	60.750	30.000	67.000	16.750	49.000	37.000
26JAN88	19.891	50.125	18.250	44.625	60.875	30.000	66.250	17.000	48.125	37.375
27JAN88	19.563	50.125	18.625	46.000	61.250	29.750	66.500	16.875	48.750	37.750
28JAN88	19.891	51.000	18.750	46.500	61.875	31.375	67.375	17.625	49.000	37.875
29JAN88	20.328	52.250	18.875	47.000	63.500	32.125	67.625	17.875	49.250	38.375
01FEB88	20.563	52.625	18.875	46.500	63.375	32.125	67.000	18.000	49.000	38.625
02FEB88	20.438	53.250	19.250	46.125	63.625	32.125	66.375	18.250	48.875	38.500
03FEB88	20.500	53.750	19.250	46.000	63.250	30.750	66.500	18.000	47.750	37.500
04FEB88	20.563	53.750	19.125	45.750	63.250	30.000	67.375	18.250	47.500	37.125
05FEB88	20.328	53.500	19.000	45.500	62.375	30.000	67.375	18.375	46.000	37.375
08FEB88	19.891	52.875	18.875	45.000	61.375	29.250	67.375	17.625	44.500	36.375
09FEB88	20.391	52.500	19.000	45.125	61.625	29.250	67.500	18.000	46.000	36.375
10FEB88	20.891	52.750	19.250	45.250	62.000	29.500	68.375	18.000	46.750	37.000
11FEB88	20.891	52.125	19.000	46.000	62.000	29.875	68.500	17.625	47.000	37.000
12FEB88	21.063	52.500	19.125	47.250	62.250	29.875	68.875	18.125	47.250	37.375
16FEB88	21.281	52.750	19.125	46.875	62.000	29.375	69.250	18.250	47.125	37.625
17FEB88	21.219	53.375	18.875	46.125	61.625	28.875	69.000	18.250	47.750	38.250
18FEB88	20.891	52.375	18.625	46.375	61.375	28.875	68.375	17.750	47.750	38.000
19FEB88	21.281	52.750	19.000	46.125	62.250	28.750	69.750	18.375	47.375	38.750
22FEB88	21.328	53.000	19.125	46.375	63.000	29.125	70.500	18.875	48.250	40.250
23FEB88	21.219	53.125	20.000	46.875	63.250	28.750	70.000	18.625	48.000	39.625
24FEB88	21.281	52.625	19.875	46.750	63.500	28.375	69.875	18.125	48.500	40.625
25FEB88	21.328	52.250	19.250	46.375	63.250	27.625	69.500	18.125	49.125	40.375
26FEB88	21.109	52.250	19.375	45.750	63.375	27.500	69.625	18.375	49.500	40.250
29FEB88	21.109	52.500	19.250	46.875	63.500	28.125	70.625	18.625	49.875	40.750

1.12.3 Instructor Reputation and Teacher Ratings

Il seguente *case study* riguarda la differente valutazione che gli studenti danno della efficacia di una lezione in base alle opinioni - più o meno fondate - che hanno sentito sull'insegnante che tiene la lezione stessa. L'idea di questa indagine è molto interessante e andrebbe riprogettata per aderire alle possibilità della

nostra scuola (logistiche e psicologiche). I dati forniti nel case study, si riferiscono ad una vera statistica condotta in una scuola statunitense; in nota i referimenti del caso. Si tratta comunque di una statistica che raccoglie dati *sperimentali*.

How[4] powerful are rumors? Frequently, students ask friends and/or look at instructor evaluations to decide if a class is worth taking. Kelley (1950) found that instructor reputation has a profound impact on actual teaching ratings. Towler and Dipboye (1998) replicated and extended this study by asking (a) Does an instructor's prior reputation affect student ratings? and (b) Does the size of this effect depend on student characteristics. This case study presents only data relevant to the former question. [5]

Experimental Design

Subjects were randomly assigned to one of two conditions. Before viewing the lecture, students were give a summary of the instructors prior teaching evaluations. There were two conditions: Charismatic instructor and Punitive instructor.

Summary given in the "Charismatic instructor" condition: Frequently at or near the top of the academic department in all teaching categories. Professor S was always lively and stimulating in class, and commanded respect from everyone. In class, she always encouraged students to express their ideas and opinions, however foolish or half-baked. Professor S was always innovative. She used differing teaching methods and frequently allowed students to experiment and be creative. Outside the classroom, Professor S was always approachable and treated students as individuals.

Summary given in the "Punitive instructor" condition: Frequently near the bottom of the academic department in all important teaching categories. Professor S did not show an interest in students' progress or make any attempt to sustain student interest in the subject. When students asked questions in class, they were frequently told to find the answers for themselves. When students felt they had produced a good piece of work, very rarely were they given positive feedback. In fact, Professor S consistently seemed to grade students harder than other lecturers in the department.

Then all subjects watched the same twenty-minute lecture given by the exact same lecturer. Following the lecture, subjects rated the lecturer. Subjects answered three questions about the leadership qualities of the lecturer. A summary rating score was computed and used as the variable "rating" here.

Raw data

Condition	Rating
2	2.6667
1	1.6667
2	2.0000
1	3.0000
1	1.6667
1	2.3333
2	2.0000
2	1.3333
2	1.6667
1	4.0000
2	2.3333
1	2.3333

[4] http://onlinestatbook.com/

[5] Kelley, H. H.(1950). The warm-cold variable in first impression of persons. Journal of Personality, 18, 431-439.
Towler, A., & Dipboye, R. L. (1998). The effect of instructor reputation and need for cognition on student behavior (poster presented at American Psychological Society conference, May 1998). (Contact Annette Towler (towleraj@rice.edu) for a reprint of the article.)

2 2.6667
2 2.0000
2 1.6667
1 2.0000
1 2.6667
1 2.6667
1 2.3333
2 1.3333
1 3.3333
2 2.3333
2 2.0000
1 2.3333
2 2.3333
1 2.3333
2 2.3333
2 2.3333
1 2.6667
1 3.0000
1 2.6667
1 3.0000
2 2.3333
1 2.0000
2 1.6667
1 2.3333
2 3.6667
1 2.6667
2 2.3333
1 3.0000
2 2.6667
1 3.3333
1 3.0000
1 2.6667
2 2.0000
2 2.3333
2 2.3333
2 3.3333

Parte I

Contributi

Contributi e licenza

Erica Boatto	Algebra I - Algebra II - Insiemi - Esercizi di geometria metrica
Beniamino Bortelli	Grafici
Roberto Carrer	Coordinatore progetto - Numeri - Funzioni - Algebra Lineare - Integrazione - Matematica 5 - Statistica descrittiva - Sistemi dinamici
Morena De Poli	Laboratorio matematica
Piero Fantuzzi	Algebra I - Algebra II - Insiemi - Esercizi di geometria metrica
Caterina Fregonese	Analisi (Integrazione) - Esercizi
Carmen Granzotto	Funzioni - Analisi (Integrazione)
Franca Gressini	Funzioni - Statistica descrittiva - Teoria della probabilità I - Teoria della probabilità II - Teoria della probabilità III
Beatrice Hitthaler	Funzioni trascendenti - Geometria analitica Numeri complessi - Analisi - Matematica 5 Teoria della probabilità I - Teoria della probabilità II
Lucia Perissinotto	Funzioni trascendenti - Geometria analitica Numeri complessi - Analisi - Matematica 5 Teoria della probabilità I - Teoria della probabilità II
Pietro Sinico	Geometria I - Geometria II
STUDENTI	
Matteo Alessandrini classe VA 2012-2013	Algebra Lineare
Simone Simonella classe IVA 2014-2015	Sistemi dinamici

Dipartimento di Matematica
ITIS V.Volterra
San Donà di Piave