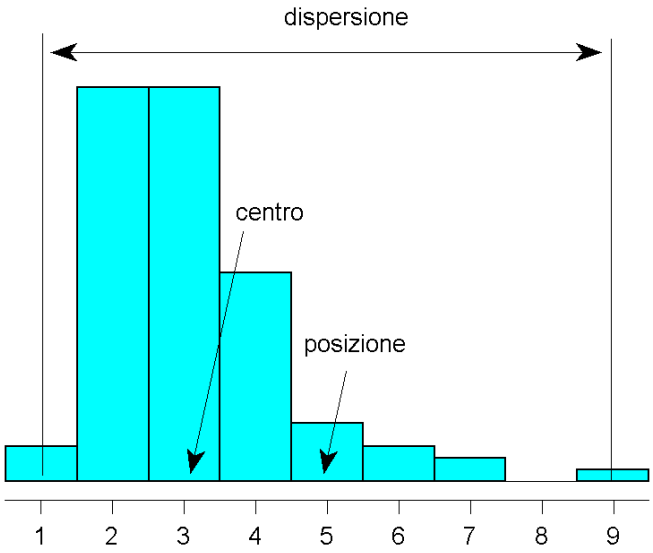
# Indicatori numerici descrittivi

Gli **indicatori numerici descrittivi** sono valori numerici che riassumono le principali caratteristiche di un set di dati; esempi di indicatori numerici descrittivi sono la *media* e la *deviazione standard* della distribuzione delle frequenze. La media è un indicatore di posizione centrale della distribuzione delle frequenze, mentre la deviazione standard è un indicatore di dispersione della variabile attorno al valore medio. Gli indicatori numerici descrittivi vengono classificati come indicatori di: **tendenza centrale**, **dispersione** e **posizione** e riassumono le caratteristiche della distribuzione delle frequenze.

**Esempio:** Ad un campione scelto a caso di famiglie con figli iscritti alle scuole elementari della città è stato chiesto di indicare quanti figli hanno. Da questo è stato ricavato l'istogramma con la distribuzione delle frequenze. E possibile evidenziare alcuni valori dell'istogramma che riassumono efficacemente la distribuzione delle frequenze, fra i valori più utilizzati si utilizza:

- Il numero di bambini presenti in una famiglia "tipo".
- Il numero di bambini più frequente.
- Il numero minimo e massimo di bambini per famiglia,
- Il numero di famiglie che hanno un numero di bambini compreso fra 2 e 5.
- Il numero di che famiglie hanno "esattamente 5" bambini.
- La percentuale di famiglie con più/meno di ….




Gli **indicatori** che riassumono la distribuzione delle frequenze sono detti anche:

1. misure **tendenza centrale**: media, mediana, moda
2. misure di **dispersione**: range, varianza, deviazione standard
3. misure di **posizione**: quartili, range interquartile (IQR), percentile e rango percentile.

Le misure numeriche riassuntive o indicatori della distribuzione delle frequenza, della frequenza relativa o della densità, sono utilizzate per rispondere a domande mirate – la percentuale di osservazioni il cui valore è minore di …, - per confrontare campioni diversi (bambini.xls).

Gli indicatori di tendenza centrale, dispersione e posizione sono definiti per dati quantitativi raggruppati e non raggruppati.

# Misure di tendenza centrale per dati Discreti o Non Raggruppati[1].

Le misure di tendenza centrale sono valori numerici tipici che danno informazioni sulla posizione centrale di un istogramma o sulla curva di distribuzione delle frequenze. La **media**, la **mediana** e la **moda** sono misure di tendenza centrale.

## **Media Aritmetica** o **Medi**a di dati discreti.

$$Media = \frac{Somma\ di\ tutti\ i\ valori\ del\ campione}{Numero\ di\ valori\ presenti\ nel\ campione}$$

La media di dati discreti si ottiene dividendo la somma di tutti i valori per il numero di valori nel data set

$$Media\ di\ una\ popolazione \quad \mu = \frac{\sum x}{N}$$

$$Media\ di\ un\ campione \quad \bar{x} = \frac{\sum x}{n}$$

dove μ (la lettera greca *m*) viene utilizzata per indicare la media di una popolazione e il simbolo $\bar{x}$ (x-barrato) per indicare la media di un campione di dati o media campionaria. I valori di *N* ed *n* indicano rispettivamente le dimensioni della popolazione e del campione.

**Esempio 3-1.** Imponibile dell'anno 1998 di 5 celebrità del cinema e della televisione.

**Tabella 3.1**

| Celebrità | Imponibile 1998 Milioni $ |
|---|---|
| Mel Gibson | 55 |
| Jerry Seinfield | 267 |
| Steven Spielberg | 175 |
| John Travolta | 47 |
| Oprah Winfrey | 125 |
| **Somma** | **669** |

$$\sum x = x_1 + x_2 + x_3 + x_4 + x_5 = 669$$

$$\bar{x} = \frac{\sum x}{n} = \frac{669}{5} = 133.8\,\text{M\$}$$

**Esempio 3-2.** I valori della v.a. X={53, 32, 61, 27, 39, 44, 49, 57} riportano l'età degli impiegati di una piccola azienda. Si calcoli la media della popolazione X e dei campioni $X_1$={32, 39, 57} e $X_2$={53, 27, 44}.

[1] Ricordiamo dal Cap.1 che per *dati discreti* o *non raggruppati* si intendono i valori di un data set che danno una informazione completa su ciascun elemento del campione o della popolazione.

*Media della popolazione* $\mu = \frac{\sum x}{N} = \frac{362}{8} = 42.25;$

*Media del campione* $X_1$ $\bar{x}_1 = \frac{\sum x}{n} = \frac{32+39+57}{3} = 42.67$

*Media del campione* $X_2$ $\bar{x}_2 = \frac{\sum x}{n} = \frac{53+27+44}{3} = 41.33$

Importante. La media della popolazione è costante; mentre la *media campionaria varia* da campione a campione e dipende dai valori che fanno parte del campione.

## Valori estremi (Outliers)

A volte un data set può contenere alcuni valori **molto piccoli** o **molto grandi**, che si differenziano notevolmente dalla maggior parte dei valori. Questi valori sono detti **Outliers o valori estremi.**

La media dipende fortemente dalla presenza di outliers e per questa ragione non è sempre possibile usare questo indicatore come misura di tendenza centrale di un campione.

**Esempio 3-2.** In tabella sono riportati gli abitanti dei cinque stati USA posti sulla costa del Pacifico (dati provenienti dal censimento eseguito nel 1997); la popolazione della California è molto più grande della popolazione degli altri stati ed è perciò un Outlier.

**Tabella 3.2**

| Stati | Popolazione Migliaia |
|---|---|
| Washington | 5610 |
| Oregon | 3243 |
| Alaska | 609 |
| Hawaii | 1187 |
| California | **32268** |

$$media = \frac{5610+3243+609+1187}{4} = 2662.25$$

$$media = \frac{5610+3243+609+1187+32268}{5} = 8583.4$$

Importante. I valori della media calcolati in *presenza dell'outlier* non descrivono la tendenza centrale del campione poiché il solo dato della California indica un valore molto superiore dei restanti valori. Questo porta ad avere una certa cautela nell'utilizzare la media come misura di tendenza centrale.

## **Mediana** di un data set discreto

*La* **Mediana** *è il valore situato nella posizione centrale di un data set ordinato secondo un ordine crescente*. La mediana divide il data set in due parti uguali.

Dato un campione di *n* valori la mediana di un set di dati discreto si ottiene in due fasi:

(a) Si ordina il campione in ordine crescente

(b) Si calcola la posizione del valore in posizione centrale con la relazione[2]

$$\text{posizione del valore centrale} = \frac{n+1}{2},$$

se il data set rappresenta una popolazione al numeratore è presente il valore N.

Importante. Se il data set ordinato contiene un *numero dispari* di valori la mediana è il valore che occupa la posizione centrale; se il numero di *osservazioni è pari* la mediana si calcola come valore medio dei due valori posti in posizione centrale.

**Esempio 3-4.** I valori che seguono riportano la perdita di peso in kg di cinque membri di club salutista dopo due mesi di iscrizione al club: vogliamo calcolare la mediana.

4.9 2.4 9.2 3.9 1.7

Per prima cosa ordiniamo i dati in ordine crescente

1.7 2.4 **3.9** 4.9 9.2

Successivamente individuiamo il numero di valori presenti nel data set, cioè $n$=5 e sostituiamo questo valore nella formula della posizione del valore centrale

$$\text{posizione del valore centrale} = \frac{n+1}{2} = \frac{5+1}{2} = 3$$

Quindi la mediana è il valore di perdita di peso sito nella terza posizione del data set, ovvero 3.9 kg; il valore in evidenza della precedente tabella.

**Esempio 3-5.** La Tabella 3-3 riporta il danaro guadagnato nel 1998 in milioni di dollari dai presidenti di 10 aziende americane, che sono classificati come i più pagati. Vogliamo calcolare la mediana del data set.

**Table 3.3**

| Chief Executive and Company Name | 1998 Total Pay (millions of dollars) |
|---|---|
| Craig Barrett, Intel | 116.5 |
| Stephen Case, America Online | 159.2 |
| Michael Eisner, Walt Disney | 575.6 |
| M. Douglas Ivester, Coca-Cola | 57.3 |
| Mel Karmazin, CBS | 201.9 |
| L. Dennis Kozlowski, Tyco International | 65.3 |
| Henry Schacht, Lucent Technologies | 67.0 |
| Henry Silverman, Cendant | 63.9 |
| Sanford Weill, Citigroup | 167.1 |
| John Welch, General Electric | 83.7 |

Gli $n$=10 valori sono riportati in una unica riga

116.5 159.2 575.6 57.3 201.9 65.3 67.0 63.9 167.1 83.7

e ordinati in ordine crescente

57.3 63.9 65.3 67.0 83.7 116.5 159.2 167.1 201.9 575.6

[2] Il valore della mediana è il medesimo se il campione sia ordinato o in ordine decrescente.

Il data-set è composto da un numero pari di valori, $n$=10, e sostituendo nella formula della posizione centrale otteniamo

$$\text{posizione del valore centrale} = \frac{n+1}{2} = \frac{10+1}{2} = 5.5$$

Quindi la mediana si calcola come valore medio dei dati in posizione 5 e 6, precisamente

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 57.3 | 63.9 | 65.3 | 67 | **83.7** | **116.5** | 1592.2 | 167.1 | 201.9 | 575.6 |
| | | | | ↑ | | | | | |

$$\text{Mediana} = \frac{83.7 + 116.3}{2} = 100.1 \text{ Milioni di dollari}$$

Importante. La mediana è meno influenzata rispetto al valore medio dalla presenza nel data set di **valori limite** (**outliers)**. Di conseguenza, è da considerarsi la misura di tendenza centrale più indicata per un data set in cui sono presenti outliers. Nell'esempio riportato in Tabella 3.2 il numero di abitanti della California è chiaramente un valore limite o outlier e condiziona pesantemente la media. In questo caso la mediana è un indicatore di tendenza centrale molto più efficiente poiché non risentendo della sua presenza indica la popolazione dell'Oregon (3,243 milioni di abitanti) quale mediana, cioè il valore esattamente in posizione centrale del data set che è composto da $n$=5 osservazioni.

## **Moda** di un data-set discreto

*In Statistica la* **Moda** *è il valore che appare con frequenza più elevata nel data set.*

**Esempio 3.6**. Il data set che segue indica la velocità in Km/h di un campione di 8 conducenti recentemente multati per eccesso di velocità in centro abitato. Il valore 74Km/h appare due volte ed è il valore più frequente o moda del campione.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 77 | 69 | 74 | 81 | 71 | 68 | 74 | 73 |

Un data-set può non avere Moda oppure può averne più di una. Ad esempio un data set dove ciascun valore appare una sola volta non ha moda; un data-set dove solo un valore appare con frequenza maggiore di uno ha una sola moda ed è detto **unimodale**. Un data-set che presenti due valori con frequenza elevata (non necessariamente la medesima) è detto **bimodale**; se ha al suo interno tre o più valori di frequenza elevata è detto **multimodale**.

**Esempi**. Data set di tipo quantitativo.

- Reddito in dollari di un campione di 5 famiglie scelte a caso. Nessuna moda. {26.15, 65.75, 34.98, 47.49, 13.74}
- Campione con il costo di un particolare tipo di televisore venduto in otto magazzini. Due mode. {495, 486, 503, 495, 470, 505, 470, 499}
- Campione contenete l'età di 10 studenti scelti a caso in un corso universitario. Tre mode. {21, 19, 27, 22, 29, 19, 25, 21, 22, 30}

Importante. La moda può essere individuata anche in un data-set qualitativo.

- La composizione del "Consiglio della Didattica" della Facoltà di Medicina e Chirurgia prevede la presenza di uno studente iscritto a ciascuno dei corsi di Laurea della Facoltà: Medicina e Chirurgia, Odontoiatria e Scienze Motorie ai quali si aggiunge uno studente per ciascuna delle 12 lauree ad indirizzo sanitario. Il data set non quindi nessuna moda.
- L'organico della Facoltà di Medicina e Chirurgia è composto da 56 Professori di I fascia (Ordinari), 68 Professori di II fascia (Associati) e 51 Ricercatori. I professori associati è la categoria che appare più di frequente nella popolazione e pertanto rappresentano la moda del campione.

Per un data-set di tipo qualitativo come quelli degli esempi sopra riportati non è possibile calcolare il valore medio e la mediana.

**Conclusione.** Non è possibile dire quale delle tre misure di tendenza centrale media, mediana e moda, sia in assoluto la migliore per caratterizzare la posizione centrale del data-set, ed in situazioni diverse una di queste può essere la più indicata. La media è sicuramente la misura di tendenza centrale più utilizzata seguita dalla mediana e nel suo calcolo sono inclusi tutti i valori del data set. La mediana è la misura di tendenza centrale migliore in presenza di outliers. La moda è facilmente individuabile ma poco utile per applicazioni pratiche.

| City | Apartment Rent | Price of House | Phone Bill |
|---|---|---|---|
| Anchorage (AK) | $ 788 | $187,337 | $14.58 |
| Phoenix (AZ) | 664 | 141,484 | 19.21 |
| San Diego (CA) | 936 | 244,000 | 16.89 |
| Denver (CO) | 740 | 179,740 | 21.05 |
| Bloomington (IN) | 574 | 131,848 | 21.30 |
| Boston (MA) | 1264 | 246,975 | 23.00 |
| Manchester (NH) | 882 | 167,450 | 21.13 |
| Albuquerque (NM) | 741 | 140,875 | 23.66 |
| Buffalo (NY) | 542 | 123,095 | 25.55 |
| Charlotte (NC) | 548 | 148,500 | 17.97 |
| Salem (OR) | 561 | 169,823 | 19.92 |
| Charleston (WV) | 602 | 142,660 | 28.08 |

**Esercizio 3-11**. Calcolare la media, la mediana e la moda delle variabili riportate in tabella.

**Esercizio 3-20**. Il data set che segue indica l'ingaggio (M$) dei giocatori della squadra di Baseball Atlanta Braves nell'anno 1999. Si calcoli la media e la mediana. Le osservazioni della variabile "Ingaggio" hanno una moda? perché si/ perché no?.

| | | | | | | |
|---|---|---|---|---|---|---|
| .23 | .20 | 7.00 | .26 | 10.60 | .20 | .23 |
| 2.93 | .20 | 1.10 | .22 | .78 | 7.75 | .95 |

Per avere altri esercizi chiedere al professore copia in modulo "pdf" del libro di testo.

## Relazione fra Media Mediana e Moda.

Le misure di tendenza centrale media mediana e moda sono fra loro in relazione diversa se l'istogramma o la distribuzione delle frequenze sono simmetrici o asimmetrici. Questo significa che la conoscenza del valore medio, della mediana e della moda può dare indicazioni sulla forma della distribuzione.

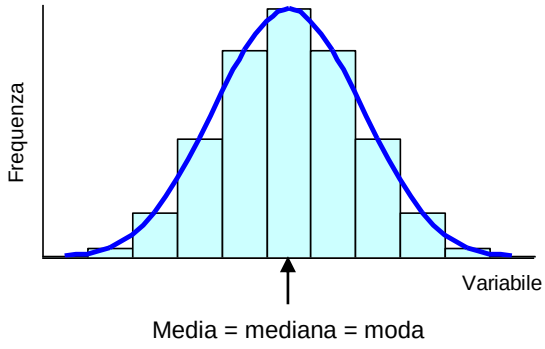



**Fig.3.2** Media, mediana e moda di un istogramma e una distribuzione delle frequenza simmetrica.

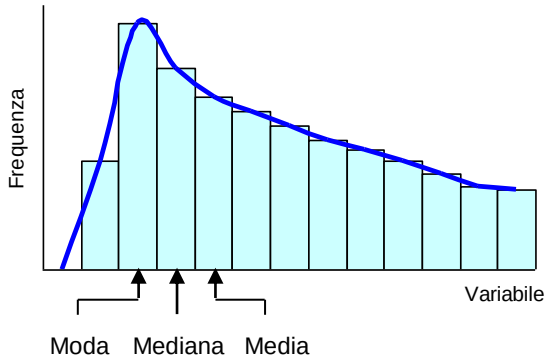



**Fig.3.3** Media, mediana e moda di un istogramma e una distribuzione delle frequenza asimmetrica verso destra.

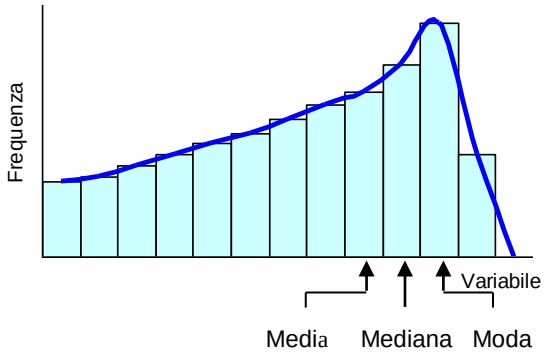



**Fig.3.4** Media, mediana e moda di un istogramma e una distribuzione delle frequenza asimmetrica verso sinistra.

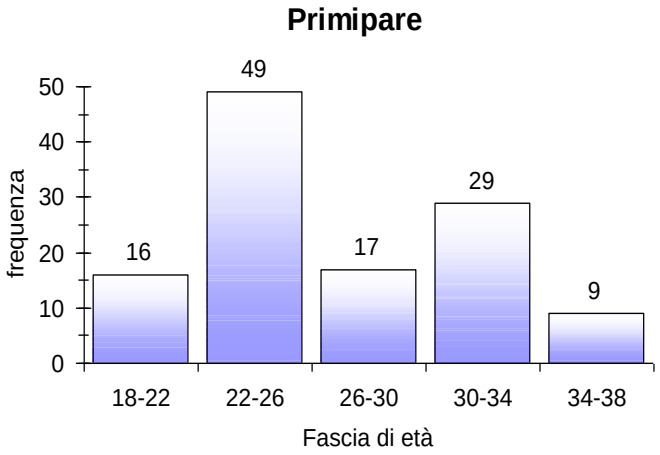
Nel caso in cui l'istogramma o la distribuzione delle frequenze siano simmetrici con un solo picco il valore di Media, Mediana e Moda coincidono (Fig.3.2). Per un istogramma o una distribuzione delle frequenze simmetrica verso destra (Fig.3.3) la media è il valore più grande, la moda il più piccolo e la mediana ha un valore intermedio. Il valore medio elevato è dovuto alla presenza di outliers nella coda destra della distribuzione. Per un istogramma o una distribuzione delle frequenze simmetrica verso sinistra (Fig.3.4) la media, la mediana e la moda hanno valori crescenti ed il piccolo valore medio è dovuto alla presenza di outliers nella coda destra della distribuzione.

## Media di dati raggruppati

Ricordiamo dal Cap.1 che per *dati numerici raggruppati* si intendono i valori di un data set che per semplicità di rappresentazione sono stati suddivisi in classi, e per ciascuna classe è data la relativa frequenza. Un data set numerico raggruppato è indicato direttamente con la tebella di distribuzione delle frequenze.

**Esempio.** La tabella che segue riporta la distribuzione delle frequenze dell'età delle primipare, cioè l'età delle donne partorienti per la prima volta. Il campione esaminato ha dimensioni $n$=120 e le osservazioni della variabile età sono state raccolte direttamente raggruppando i valori in classi di ampiezza costante. I valori della frequenza sono riportati nell'istogramma a lato della tabella[3].

| Classe (anni) | Frequenza *f* |
|---|---|
| 18-22 | 16 |
| 22-26 | 49 |
| 26-30 | 17 |
| 30-34 | 29 |
| 34-38 | 9 |
| *n* | 120 |




Alcuni indicatori di tendenza centrale quali la mediana e la moda sono direttamente calcolabili dalla tabella della frequenze: la distribuzione è bimodale con punte di frequenza per le classi 22-26 e 30-34, e alla classe 22-26 anni appartiene anche la mediana. Volendo calcolare il valore medio della distribuzione è evidente che non è possibile utilizzare le formule definite per i dati non raggruppati. Per il calcolo del valore medio si utilizza il valore centrale di ciascuna classe $m$ scelto ad approssimazione di tutti i valori appartenenti a quella classe.

La media di un set di dati **raggruppati in classi** si ottiene dividendo la somma dei prodotti $m_i \times f_i$ per il numero di valori nel data set.

$$\textit{Media di una popolazione} \quad \mu = \frac{1}{N}\sum m_i \times f_i$$

$$\textit{Media campionaria} \quad \bar{x} = \frac{1}{n}\sum m_i \times f_i$$

dove $m_i$ ed $f_i$ sono rispettivamente il *valore centrale* (class mark) e la *frequenza* di ciascuna classe.

La media campionaria risultante dalle formule indicate è pari $\bar{x} = 26.8$ che per il campione esaminato è un indicatore di tendenza centrale corretto.

[3]L'istogramma descrive una distribuzione delle frequenze simmetrica rispetto alla classe centrale e bimodale, cioè due classi mostrano una frequenza nettamente maggiore delle altre. Questa distribuzione della età delle primipare è tipica delle società occidentali dove l'età del primo parto per le donne è conseguente al grado di scolarizzazione da esse conseguito. Le donne laureate che assumono posizioni di responsabilità nel mondo del lavoro decidono di avere il primo figlio ad una più avanzata (30-34 anni) rispetto a coloro che terminano gli studi con la scuola di secondo grado.

Importante. L'approssimazione dell'intervallo di ciascuna classe con il valore centrale della tessa ha un significato statistico importante. I dati raggruppati in classi perdono il loro valore originale a vantaggio di una rappresentazione più compatta – distribuzione delle frequenze o istogramma – del data set. I valori contenuti all'interno di una singola classe sono compresi fra un massimo ed un minimo che ne rappresentano gli estremi, il valore centrale della classe è quindi il valore medio di tutte le osservazioni che cadono all'interno della classe. Questo significa che il valore medio di un set di dati raggruppati si calcola sostituendo i valori reali delle osservazioni in una data classe con $f$ – volte il valore medio della classe, dove $f$ è la frequenza della classe.

La media di popolazione e campionaria per dati raggruppati può essere calcolata anche tramite le formule che seguono, ottenute da una semplice trasformazione algebrica

$$\textit{media campionaria} \qquad \bar{x} = \sum_{i=1}^{n} m_i \frac{f_i}{n}$$

$$\textit{media di popolazione} \qquad \mu = \sum_{i=1}^{N} m_i \frac{f_i}{N}$$

I valori $f_i/N$ e $f_i/n$ indicano la frequenza relativa di ciascuna classe, rispettivamente per la media di popolazione e campionaria.

Queste espressioni hanno una valenza generale e possono essere utilizzate per il calcolo della media di dati non raggruppati e raggruppati poiché utilizzano la **definizione empirica** di probabilità.

**Esempio.** Il data set X = $\{x_i, i=1,\ldots,10\}$ = {12, 8, 7, 11, 12, 5, 8, 9, 6, 8} è costituito da una successione di dati discreti dove uno o più valori $x_i$ appaiono più volte. Se a ciascun valore $x_i$ assegniamo la frequenza relativa $f=1/n$ dove $n$ = 10 è il numero di eventi nel data set, e indichiamo con $m$ la frequenza con cui appare ciascun valore la media campionaria diviene:

$$\textit{media campionaria} \qquad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \sum_{j=1}^{p} x_j \frac{m_j}{n}$$

dove con $p$ indichiamo i valori distinti presenti ne data set X e con $m_j/n$ la frequenza relativa di ciascun valore. Nel data set in esame la media campionarie è pari a $\bar{x} = 8.6$, ed è stata calcolata sommando gli 8 valori distinti moltiplicato ciascuno per la propria frequenza relativa.

Come sarà dimostrato nel capitolo relativo alla definizione del concetto di probabilità la frequenza relativa è una definizione empirica della probabilità, la probabilità che un determinato valore $x_j$ appaia come risultato di un dato esperimento.

L'uso della frequenza relativa consente di utilizzare una unica espressione per il calcolo della media campionaria di un set di dati raggruppato e non raggruppato.

# Misure di dispersione per dati discreti o non raggruppati

Le misure di posizione media, mediana e moda non sono sufficienti per descrivere il data set. Due data set diversi possono avere la medesima media ma avere una dispersine molto diversa, la variazione dei valori presenti nel primo data set può essere molto più grande o più piccola della dispersine del secondo[4].

I due data set che seguono riportano l'età degli impiegati di due piccole aziende dove si può osservare che l'età media dei dipendenti è la medesima, 40 anni.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| I Azienda: | 47 | 38 | 35 | 40 | 36 | 45 | 39 |
| II Azienda: | | 70 | 33 | 18 | 52 | 27 | |

Se non conoscessimo l'età di ciascuno saremmo portati a concludere erroneamente che la distribuzione dei valori in ciascun data set è la medesima, mentre valori del I data set ha valori molto concentrati mentre il II secondo molto dispersi.

Questo significa che la media, la mediana e la moda non sono sufficienti per descrivere il profilo della distribuzione dei valori del data set, ma è necessario utilizzare anche misure numeriche che diano indicazione della dispersine dei valori: le **misure di dispersione.** Le misure di tendenza centrale e di dispersione forniscono una descrizione della posizione e della dispersione dei valori nel data set. Le misure di dispersine descritte in questa sezione sono il **range**, la **varianza** e la **deviazione standard**.

## Range.

Il range è la misura di dispersione più immediata da calcolare ed è pari alla differenza fra il valore maggiore ed il valore minore.

**Range di un data set discreto**

Range = Valore massimo – valore minimo

**Esempio 3.11.** La tabella indica la superficie in Miglia al quadrato degli stati americani del sud.ovest.

| Stato | Superficie (miglia$^2$) |
|---|---|
| **Arkansas** | **53182** |
| **Lousiana** | **49651** |
| **Oklaoma** | **69903** |
| **Texas** | **267277** |

La dispersione dei valori di superficie di questo data set calcolata tramite il range è pari **a**

Range = Valore massimo – valore minimo = 217626 miglia quadrate

L'esempio precedente dimostra che il **range** è influenzato dalla presenza di outliers; infatti, il suo valore si riduce a 20252 miglia quadrate se nel data set viene tolto lo stato del Texas, e non rappresenta una buona misura di dispersione per data set che contengono valori estremi. Inoltre, come misura di dispersione è scarsamente indicativa poiché utilizza i soli valori massimo e minimo trascurando qualsiasi indicazione su come sono dispersi i restanti valori.

[4] In statistica i termini dispersione e variazione ed il termine inglese *spread* hanno il medesimo significati e vengono indifferentemente utilizzate per descrivere la distribuzione dei valori di un data set.

## Varianza e Deviazione Standard.

La **deviazione standard** è la misura di dispersione usata più di frequente e fornisce una indicazione sul fatto che i valori del data set siano più o meno raggruppati attorno alla media.

Valori molto raggruppati attorno alla media ➔ piccola deviazione standard

Valori molto dispersi attorno alla media ➔ grande deviazione standard

*La deviazione standard si calcola dalla radice quadrata della* **varianza**. La varianza di una popolazione e di un campione si indicano rispettivamente con i simboli greco $\sigma^2$ (sigma quadro) e $s^2$ (esse quadro), e la deviazione standard con i simboli $\sigma$ e $s$ rispettivamente per una popolazione ed un campione. La varianza e la deviazione standard di un campione sono dette varianza campionaria e deviazione standard campionaria.

$$\sigma^2 = \frac{\sum (x-\mu)^2}{N} \quad e \quad s^2 = \frac{\sum (x-\bar{x})^2}{n-1}$$

Si osservi che il denominatore nella formula per la varianza di popolazione è N mentre nella formula per il campione è *n*-1[5].

Gli addendi al numeratore delle formule $(x-\mu)$ e $(x-\bar{x})$ sono le *deviazioni* dei valori *x* rispetto al valore medio. La somma delle deviazione rispetto al valore medio è sempre zero; cioè $\sum (x-\mu) = 0$ e $\sum (x-\bar{x}) = 0$, come è dimostrato nell'esempio numerico seguente dove sono calcolate le deviazione del set di dati che riporta il punteggio in centesima della prova in itinere di statistica, il cui valore medio è $\bar{x} = 84$.

Tabella 3.5

| $x$ | $x-\bar{x}$ |
|---|---|
| 82 | 95 – 84 = + 11 |
| 95 | 82 – 84 = – 2 |
| 67 | 67 – 84 = – 17 |
| 92 | 92 – 84 = + 8 |
| | $\sum (x-\bar{x}) = 0$ |

La varianza di popolazione e campionaria si calcolano facilmente con le seguenti formule semplificate[6], dalle quali si calcolano i valori della deviazione standard.

$$\sigma^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{N}}{N} \quad e \quad s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1}$$

deviazione standard di popolazione: $\sigma = \sqrt{\sigma^2}$

deviazione standard campionaria: $s = \sqrt{s^2}$

[5] La varianza campionaria è una stima della varianza di popolazione e se il denominatore è posto uguale ad *n* la varianza di popolazione è sottostimata.

[6] L'uso delle formule semplificate consente il calcolo della varianza anche in assenza di un computer, tuttavia gli odierni calcolatori hanno già programmate al loro interno le formule per il calcolo della varianza di popolazione e campionaria.

**Esempio 3-12**. La tabella riporta il guadagno lordo ottenuto nel 1998 di cinque personalità dello spettacolo. Si vuole calcolare la deviazione standard dei valori del data set con le formula semplificate.

**Tabella**

| Celebrità | Imponibile 1998 (Milioni di $) |
|---|---|
| Mel Gibson | 55 |
| Jerry Seinfield | 267 |
| Steven Spielberg | 175 |
| John Travolta | 47 |
| Oprah Winfley | 125 |

I valori di $s^2$ ed $s$ sono il risultato di un calcolo che si esegue con una sequenza di quattro operazioni: le prime due sono riassunte nella Tabella 3.6, le successive risultano dalla formula semplificata.

1. *Calcolo di* $\sum x$ : si sommano i valori $x$ presenti nel data-set
2. *Calcolo di* $\sum x^2$ : si elevano al quadrato i valori di $x$ e si calcola la somma.
3. *Calcolo della varianza* $\mathbf{s}^2$ *con la formula semplificata*.
4. *Calcolo della deviazione standard* **s** *come radice quadrata della varianza.*

**Tabella 3.6**

| x | $x^2$ |
|---|---|
| 55 | 3025 |
| 267 | 71289 |
| 175 | 30625 |
| 47 | 2209 |
| 125 | 15625 |
| $\Sigma x = 669$ | $\Sigma x^2 = 122773$ |

$$s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{122773 - \frac{(669)^2}{5}}{5-1} = 8315.2$$

$$s = \sqrt{s^2} = \sqrt{8315.2} = 91.187718 \quad \text{M\$}$$

Ciascun passaggio può essere è facilmente calcolato con l'aiuto di un foglio elettronico dove in colonne successive vengono riprodotti i passi 1-4 citati in precedenza (esempi di calcolo in Excel). Il risultato indica un valore della deviazione standard pari a circa 91.2M$

Importante.

1. I valori della varianza e della deviazione standard sono sempre positivi
2. L'unità di misura della varianza è pari al quadrato della unità di misura dei dati originali.

## Parametri di una popolazione e statistiche campionarie.

Le misure descrittive di posizione (media, mediana, moda) e di dispersione (range, varianza, deviazione standard) calcolate per una data-set che rappresenta una **popolazione** sono dette **parametri di popolazione** o **parametri**.

Le misura descrittive calcolate da un data-set scelto a **campione** di una popolazione sono dette **statistiche campionarie** o statistiche.

La media $\mu$ e deviazione standard $\sigma$ di una popolazione ➔ **Parametri** della popolazione

La media s e deviazione standard s di un campione ➔ **Statistiche** campionarie

**Esempio 3-13**. Il data set che segue indica il guadagno dei 6 impiegati di una piccola compagnia.

29.50 16.20 35.45 21.35 49.70 24.60

Si vuole calcolare la media e la varianza di questi dati applicando i passi 1-4 indicati sopra.

**Table 3.7**

| $x$ | $x^2$ |
|---|---|
| 29.50 | 870.2500 |
| 16.20 | 262.4400 |
| 35.45 | 1256.7025 |
| 21.35 | 455.8225 |
| 49.70 | 2470.0900 |
| 24.60 | 605.1600 |
| $\Sigma x = 176.80$ | $\Sigma x^2 = 5920.4650$ |

Dalle formule per il calcolo della varianza e deviazione standard di popolazione si ottiene

$$\sigma^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{N}}{N} = \frac{5920.4650 - \frac{(176.80)^2}{6}}{6} = \mathbf{118.4597}$$

$$\sigma = \sqrt{118.4597} = \mathbf{\$10.884\ thousand = \$10,884}$$

**Esercizio 3-43**. I dati che seguono indicano il numero di furti d'auto denunciati nelle ultime due settimane.

6 3 7 11 4 3 8 7 2 6 15 9

Si calcoli la media, la mediana, la moda se presente, il range, la varianza e la deviazione standard [r = 6.8, 7, 3, 13, 15.2, 3.9].

# Media, varianza e Deviazione Standard per dati raggruppati

I dati sono detti raggruppati quando sono rappresentati sotto forma di una tabella dove è indicata la distribuzione delle frequenze. Questo significa che i valori della variabile non sono noti e per il calcolo del valore medio (o della deviazione standard) non è possibile utilizzare formule del tipo $\Sigma x/N$ (o $\Sigma(x- \underline{x})^2 /N$ ) utilizzate per i dati non raggruppati.

**MEAN FOR GROUPED DATA**

Mean for population data: $\mu = \dfrac{\Sigma mf}{N}$

Mean for sample data: $\bar{x} = \dfrac{\Sigma mf}{n}$

where $m$ is the midpoint and $f$ is the frequency of a class.

## Media di dati raggruppati

Per il calcolo della media di dati raggruppati si individua il **punto centrale** di ciascuna classe $m$ e lo si moltiplica per la corrispondente frequenza $f$. La somma dei prodotti $\Sigma mf$ è una approssimazione della somma dei valori di $x$ non noti. Il valore medio si calcola dividendo la somma per il numero di osservazioni.

**Esempio 3-14**. La tabella indica il tempo impiegato nel percorso casa lavoro dagli impiegati (25) di una piccola azienda. Noto il valore del punto centrale $m$ di ciascuna classe si vuole calcolare il valore medio.

**Table 3.9**

| Daily Commuting Time (minutes) | $f$ | $m$ | $mf$ |
|---|---|---|---|
| 0 to less than 10 | 4 | 5 | 20 |
| 10 to less than 20 | 9 | 15 | 135 |
| 20 to less than 30 | 6 | 25 | 150 |
| 30 to less than 40 | 4 | 35 | 140 |
| 40 to less than 50 | 2 | 45 | 90 |
| | $N = 25$ | | $\Sigma mf = 535$ |

il valore medio diviene:

$$\mu = \frac{\Sigma mf}{N} = \frac{535}{25} = \mathbf{21.40\ minutes}$$

Si osservi che Per ciascuna classe il valore centrale $m$ indica il tempo medio impiegato dagli impiegati che indicano un tempo di percorrenza compreso fra gli estremi della classe. Infatti, se ipotizziamo che i 4 impiegati della prima classe impieghino un tempo variabile fra 0 e 10 minuti per raggiungere il luogo di lavoro, il valore $m$=5 indica il tempo medio impiegato da ciascuno di essi e il prodotto 5x4=20 è approssimativamente il tempo totale impiegato da tutti e quattro gli impiegati per raggiungere il luogo di lavoro.

**Esempio 3-15**. Distribuzione delle frequenze degli ordini ricevuti da una piccola azienda durante i 50 giorni lavorativi precedenti l'indagine. Si osservi che stiamo esaminando un campione di valori estratto dalla popolazione rappresentata dai tutti i giorni lavorativi dell'anno, dei precedenti ecc.

**Table 3.11**

| Number of Orders | $f$ | $m$ | $mf$ |
|---|---|---|---|
| 10–12 | 4 | 11 | 44 |
| 13–15 | 12 | 14 | 168 |
| 16–18 | 20 | 17 | 340 |
| 19–21 | 14 | 20 | 280 |
| | $n = 50$ | | $\Sigma mf = 832$ |

Il numero medio di ordini ricevuti nei 50 gg indagati o media campionaria è pari a

$$\bar{x} = \frac{\Sigma mf}{n} = \frac{832}{50} = \mathbf{16.64\ orders}$$

## Varianza e deviazione standard per dati raggruppati

A partire dal valore $m$ del punto centrale di ciascuna classe le formule per il calcolo della varianza di popolazione e campionaria sono rispettivamente

$$\sigma^2 = \frac{\Sigma f(m-\mu)^2}{N} \quad \text{and} \quad s^2 = \frac{\Sigma f(m-\bar{x})^2}{n-1}$$

I rispettivi valori della deviazione standard $\sigma$ e $s$ si ottengono dalla radice quadrata positiva della varianza.

Anche nel caso di dati raggruppati è possibile definire formule semplificate che abbreviano il calcolo[7] della varianza

**SHORT-CUT FORMULAS FOR THE VARIANCE AND STANDARD DEVIATION FOR GROUPED DATA**

$$\sigma^2 = \frac{\Sigma m^2 f - \frac{(\Sigma mf)^2}{N}}{N} \quad \text{and} \quad s^2 = \frac{\Sigma m^2 f - \frac{(\Sigma mf)^2}{n}}{n-1}$$

where $\sigma^2$ is the population variance, $s^2$ is the sample variance, and $m$ is the midpoint of a class.

The standard deviation is obtained by taking the positive square root of the variance.

Population standard deviation: $\sigma = \sqrt{\sigma^2}$

Sample standard deviation: $s = \sqrt{s^2}$

[7] Fra le funzioni predefinite di Excel nella categoria funzioni statistiche sono presenti le funzioni VAR.POP e VAR che calcolano rispettivamente la varianza di popolazione e la varianza campionaria.

**Esempio 3-16**. La Tab. 3-16 esplicita il calcolo della varianza $\sigma^2$ e della deviazione standard $\sigma$ della distribuzione delle frequenze del tempo impiegato per il percorso casa-lavoro (Tabella 3-9) utilizzando la formula semplificata.

**Table 3.12**

| Daily Commuting Time (minutes) | $f$ | $m$ | $mf$ | $m^2f$ |
|---|---|---|---|---|
| 0 to less than 10 | 4 | 5 | 20 | 100 |
| 10 to less than 20 | 9 | 15 | 135 | 2,025 |
| 20 to less than 30 | 6 | 25 | 150 | 3,750 |
| 30 to less than 40 | 4 | 35 | 140 | 4,900 |
| 40 to less than 50 | 2 | 45 | 90 | 4,050 |
| | $N = 25$ | | $\Sigma mf = 535$ | $\Sigma m^2f = 14{,}825$ |

Individuato il centro di ciascuna classe $m$ la successione delle operazioni è la seguente

***Step 1.*** *Calculate the value of* $\Sigma mf$. $\Sigma mf = 535$

***Step 2.*** *Find the value of* $\Sigma m^2f$. $\Sigma m^2f = 14{,}825$

***Step 3.*** *Calculate the variance.*

$$\sigma^2 = \frac{\Sigma m^2f - \frac{(\Sigma mf)^2}{N}}{N} = \frac{14{,}825 - \frac{(535)^2}{25}}{25} = \frac{3376}{25} = \mathbf{135.04}$$

***Step 4.*** *Calculate the standard deviation.* $\sigma = \sqrt{\sigma^2} = \sqrt{135.04} = \mathbf{11.62\ minutes}$

I valori di calcolati corrispondono alla **varianza** ed alla **deviazione standard di popolazione** poiché l'analisi è stata eseguita su tutti gli N=25 dipendenti dell'azienda.

**Esempio 3-17**. Varianza e deviazione standard della distribuzione delle frequenze del numero degli ordini riportati in Tabella 3-10.

**Table 3.13**

| Number of Orders | $f$ | $m$ | $mf$ | $m^2f$ |
|---|---|---|---|---|
| 10–12 | 4 | 11 | 44 | 484 |
| 13–15 | 12 | 14 | 168 | 2,352 |
| 16–18 | 20 | 17 | 340 | 5,780 |
| 19–21 | 14 | 20 | 280 | 5,600 |
| | $n = 50$ | | $\Sigma mf = 832$ | $\Sigma m^2f = 14{,}216$ |

poiché i dati corrispondono ad un campione di soli 50 gg, le formule semplificate per il calcolo della **varianza** e **deviazione standard campionaria** danno i seguenti valori:

$$s^2 = \frac{\Sigma m^2f - \frac{(\Sigma mf)^2}{n}}{n-1} = \frac{14{,}216 - \frac{(832)^2}{50}}{50 - 1} = \mathbf{7.5820}$$

$$s = \sqrt{s^2} = \sqrt{7.5820} = \mathbf{2.75\ orders}$$

Importante. La formula semplificata per il calcolo dei parametri di una popolazione e delle statistiche campionarie può essere facilmente sviluppata in ambiente Excel riproducendo le Tab.3-12 e 3-13 (formulesemplificate.xls).

## Uso della deviazione standard

A partire dal valore medio e dalla deviazione standard è possibile calcolare la proporzione o percentuale delle osservazioni presenti nel data set esaminato. La percentuale k-esima di un data set di forma qualsiasi può essere calcolata utilizzando il teorema di Chebishev[8] o tramite la regola empirica che riportiamo di seguito.

La **regola empirica** può essere utilizzata solo nel caso di una *distribuzione a forma di campana* (*bell-shaped distribution*) come quella riportata in Fig. 3-9. Nel Cap. 6 vedremo che una distribuzione della forma di Fig. 3-9 è detta **distribuzione normale** o **curva normale**.

**EMPIRICAL RULE** For a bell-shaped distribution, approximately

1. 68% of the observations lie within one standard deviation of the mean
2. 95% of the observations lie within two standard deviations of the mean
3. 99.7% of the observations lie within three standard deviations of the mean

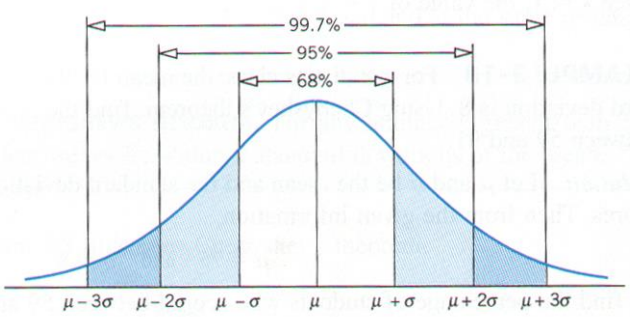



**Figure 3.9** Illustration of the empirical rule.

**Esempio 3-19.** La distribuzione dell'età di un campione di 5000 persone ha la forma a campana con media 40 e deviazione standard 12 anni. Calcolare la percentuale approssimativa delle persone che hanno una età compresa fra 16 e 64 anni. Preso a riferimento i valori

$\underline{x} = 40$ *anni* s=12 *anni*

I valori 16 e 24 anni sono alla medesima distanza da $\underline{x}$ e precisante 24 anni, dividendo 24 per *s*=12 trasformiamo la distanza dalla media in unità della deviazione standard. Poiché 24/12 = 2 l'intervallo di età 16-42 anni corrisponde all'area della distribuzione racchiusa nell'intervallo $\underline{x} - s$ e $\underline{x} + s$.

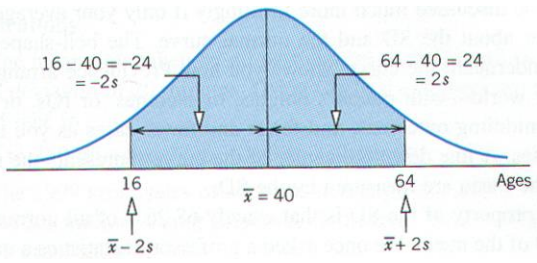



**Figure 3.10** Percentage of people who are 16 to 64 years old.

[8] Mann Capitolo 3 pp 98-100

La Fig. 3-10 illustra la regola empirica applicata al problema: poiché l'area calcolata è contenuta entro due deviazioni standard dalla media all'incirca il 95% del campione ha una età compresa fra 16 e 24 anni.

**Esempio 3.80**. La vita media delle batterie per auto è di 44 mesi con una deviazione standard di 3 mesi. Nell'ipotesi che la distribuzione della vita media sia a forma di campana si vuole calcolare la percentuale delle batterie che hanno una vita compresa fra

41-47 mesi  38-50 mesi  35-53 mesi

**Esempio 3.81**. Secondo i dati forniti dal ministero dell'istruzione il salario medio annuo degli insegnanti negli anni 1996-1997 è risultato di 38336$. Assumendo che la distribuzione degli stipendi abbia una forma a campana con media 38436$ e deviazione standard 35000$, calcolare con la regola empirica la percentuale degli insegnati il cui salario è compreso fra

31436-45436$  34936-41936$  27936-48936$

**Esempio 3.83**. L'età delle auto possedute dai dipendenti di una grande industria ha una distribuzione a campana con media e varianza rispettivamente pari a 7 e 2 anni.

a. Utilizzando la regola empirica calcolare la percentuale di auto con età compresa fra

   5-9 anni  1-13 anni

b. Utilizzando la regola empirica calcolare l'intervallo di età che contiene il 95% delle auto possedute dai dipendenti dell'azienda.

# Misure di posizione

Una **misura di posizione** indica la posizione relativa di un singolo valore rispetto ai restanti valori di un campione o di una popolazione. Le misure di posizioni alle quali siamo interessati sono i quartili, i percentili e l'ordine di percentile.

## Quartili e Range Interquartile.

I **quartili** sono misure o indicatori riassuntivi che dividono un data set ordinato (campione o popolazione) in quattro parti uguali. Si indicano come **I quartile** (Q1), **II quartile** (Q2) e **III quartile** (Q3).

**QUARTILES** *Quartiles* are three summary measures that divide a ranked data set into four equal parts. The second quartile is the same as the median of a data set. The first quartile is the value of the middle term among the observations that are less than the median, and the third quartile is the value of the middle term among the observations that are greater than the median.

Figure 3.11 describes the positions of the three quartiles.

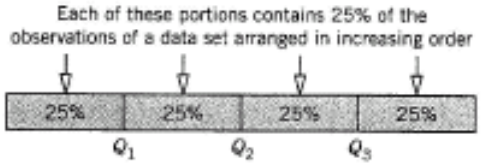



**Figure 3.11** Quartiles.

Il data set del quale si vogliono calcolare i quartili deve essere prima ordinato in ordine crescente. La differenza fra il III quartile ed il I quartile prende nome di **range interquartile** (**IQR**)

**INTERQUARTILE RANGE** The difference between the third and the first quartiles gives the *interquartile range*; that is,

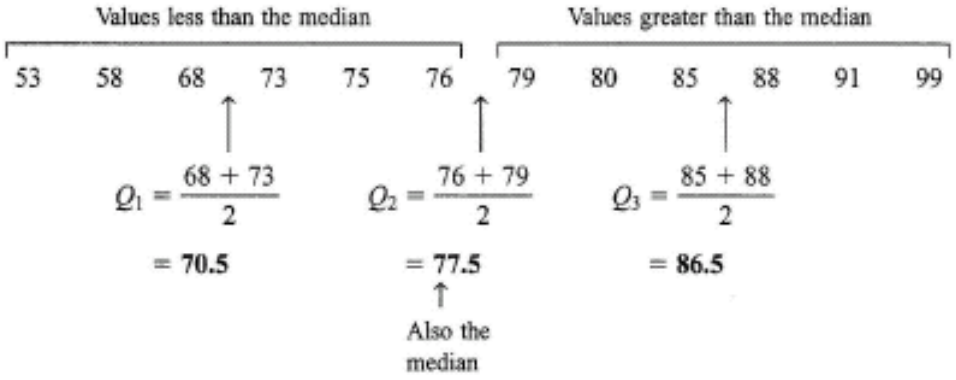
$$\text{IQR} = \text{Interquartile range} = Q_3 - Q_1$$

**Esempio 3-20**. Il data-set riporta i punteggi in 100 degli studenti che hanno sostenuto la prova scritta di matematica.

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 75 | 80 | 68 | 53 | 99 | 58 | 76 | 73 | 85 | 88 | 91 | 79 |

a. Calcolare i valori di Q1, Q2 e Q3
b. Individuare la posizione del punteggio 88/110 rispetto ai quartili
c.

Dai valori ordinati dl campione calcoliamo in successione Q1, Q2 e Q3.



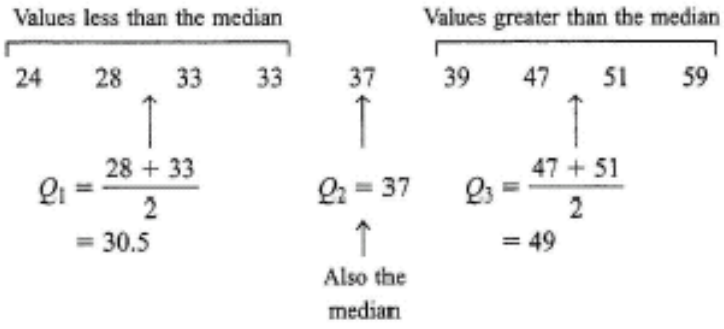

I valori di Q1, Q2, Q2 individuano il punteggio che separa il 25%, il 50% ed il 75% del data-set; Il II quartile separa il data set in due parti uguali ed è pari alla mediana. Il punteggio 88/100 fa parte dell'ultimo quarto del data set. Il valore di IQR è pari a IQR = Q3 –Q1 = 86.5 – 70.5 = 16.

**Esempio 3-21**. Il data-set riporta l'età dei novi impiegati di una piccola compagnia assicurativa.

47 28 39 51 33 37 59 24 33

a. Calcolare i valori di Q1, Q2 e Q3
b. Individuare la posizione nel data set dell'impiegato di età pari a 28 anni
c. Calcolare il valore di IQR

Ordinato il campione si ottiene:




Noti i valori dei quartili Q1, Q2 e Q3, è facile individuare che il valore "28 anni" è entro il 25% del campione, e calcolare IQR = Q3 – Q1 = 18.5.

# Percentili e Ordine di Percentile (Percentile Rank)

I **percentili** sono misure riassuntive che dividono il data set ordinato in 100 parti uguali.

Il data set è ordinato in ordine crescente è suddiviso il 100 parti uguali. Ciascuna porzione contiene l'1% delle osservazioni presenti nel data set ordinato. Il k-esimo percentile è indicato con $P_k$ , dove k è un valore intero compreso fra 1 e 99.

Il *k-esimo* percentile, $P_k$, è un valore tale che all'incirca il k% delle osservazioni presenti nel data set hanno **valore minore** del valore $P_k$, e all'incirca il (100 – k)% delle misure hanno un **valore maggiore** del valore $P_k$.

**PERCENTILES** The (approximate) value of the *k*th *percentile*, denoted by $P_k$, is

$$P_k = \text{Value of the } \left(\frac{kn}{100}\right)\text{th term in a ranked data set}$$

where $k$ denotes the number of the percentile and $n$ represents the sample size.

**Esempio 3-22**. Calcolare il 62-esimo percentile del data-set con i voti del compito di matematica, ed descrivere il significato di 62-esimo percentile

75 80 68 53 99 58 76 73 85 88 91 79

1. ordiniamo i valori in ordine crescente

53 58 68 73 75 76 79 80 85 88 91 99

2. calcoliamo il 62° percentile

$$\frac{kn}{100} = \frac{62(12)}{100} = 7.44\text{th term}$$

3. il valore 7.44 indica approssimativamente la posizione intermedia fra il 7° e 8° valore del data set, di conseguenza

$$P_{62} = 62\text{nd percentile} = \frac{79 + 80}{2} = \mathbf{79.5}$$

4.

Il valore $P_{62}$ indica che all'incirca il 62% del punteggio ottenuti dai candidati alla prova d'esame è inferiore a 79.5 centesimi, analogamente il 38% dei candidati ha ottenuto un punteggio superiore a 79.5[9].

È possibile calcolare l'**ordine di percentile** (rango percentile) per un particolare valore $x_i$ del data set con la formula

**PERCENTILE RANK OF A VALUE**

$$\text{Percentile rank of } x_i = \frac{\text{Number of values less than } x_i}{\text{Total number of values in the data set}} \times 100$$

**Esempio 3-23**. Vogliamo calcolare il rango percentile del punteggio 85/100, a partire dal campione ordinato

53 58 68 73 75 76 79 80 85 88 91 99

poiché esattamente 8 dei 12 valori presenti nel data set precedono il valore scelto il rango percentile è pari a

$$\text{Percentile rank of } 85 = \frac{8}{12} \times 100 = \mathbf{66.67\%}$$

Possiamo quindi dedurre che all'incirca il 67% dei candidati ha ottenuto un punteggio inferiore a 85/100 e il 33% un punteggio pari o superiore.

Importante. I quartili di percentile ed il rango percentile di un data set possono essere calcolati con la funzioni di Excel QUARTILE, PERCENTILE e PERCENTIL.RANGO. I valori calcolato con le funzioni di Excel possono essere leggermente diverse dai valori calcolati in precedenza poiché utilizzano algoritmi di interpolazione per approssimare al meglio il valore percentuale (25%, 62%, ecc.) richiesto. Esempi di calcolo dei quartili e percentili sono riportati nel foglio Excel percentile.xls.

## Box-and-Whisker plot

Il **Box-and-whiskers plot** è una rappresentazione grafica del data set costruita a partire dalle cinque misure di posizione: mediana, I quartile, III quartile ed i valori più piccolo e più grande del data set compresi fra l'**upper** e **lower inner fences**.

Il Box-and-whiskers plot permette di **visualizzare** il centro, la dispersione, la asimmetria del data set ed indica la presenza di outliers. Permette inoltre di **confrontare** data set diversi sulla base dei rispettivi Box-and-whiskers plot.

**BOX-AND-WHISKER PLOT** A plot that shows the center, spread, and skewness of a data set. It is constructed by drawing a box and two whiskers that use the median, the first quartile, the third quartile, and the smallest and the largest values in the data set between the lower and the upper inner fences.

[9] Poichè il campione esaminato ha poche osservazioni, n=12, il 62° percentile non corrisponde esattamente al 62% dei dati, infatti 7/12 corrisponde al 58% dei dati e 5/12 al 42%, rispettivamente minori e maggiori del 68% e del 32% attesi. Le stime percentuali divengono più accurate all'aumentare delle dimensioni del campione.

Esempio 3-24. I dati che seguono indicano il reddito (in migliaia di euro) di un campione di 12 famiglie. I dati sono stati ordinati in ordine crescente.

17 22 23 27 29 32 38 42 46 52 60 92

1. Calcolare la mediana Q1, Q3 e IQR

$$\text{Median} = (32 + 38)/2 = 35$$

$$Q_1 = (23 + 27)/2 = 25$$

$$Q_3 = (46 + 52)/2 = 49$$

$$\text{IQR} = Q_3 - Q_1 = 49 - 25 = 24$$

2. Calcolare i *punti* che sono posti 1.5 x IQR al di sotto di Q1 e 1.5 x IQR al di sopra di Q3. Questi due punti sono indicati come **lower** e **upper inner fences**.

$$1.5 \times \text{IQR} = 1.5 \times 24 = 36$$

$$\text{Lower inner fence} = Q_1 - 36 = 25 - 36 = -11$$

$$\text{Upper inner fence} = Q_3 + 36 = 49 + 36 = 85$$

3. Individuare i valori più grande e più piccolo del data set compresi fra il lower e upper inner fences.

Smallest value within the two inner fences = 17

Largest value within the two inner fences = 60

4. Disegnare I valori Q1, mediana e Q3 come una scatola (Box) posta al di sopra di un intervallo graduato dove sono contenuti tutti i valori del data set

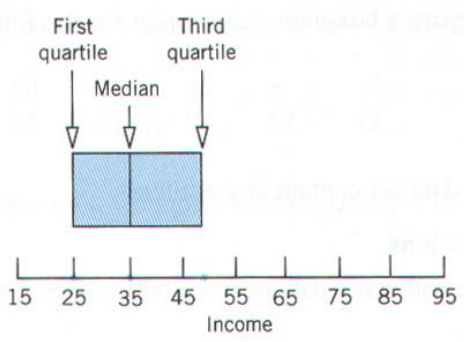



Figure 3.13

.

5. Congiungere con due linee orizzontali (**wiskers**) i valori Q1 con il lower fence (17) e Q3 con l'upper fence (60) calcolati a punto 3. I valori che cadono al di fuori dell'inner e

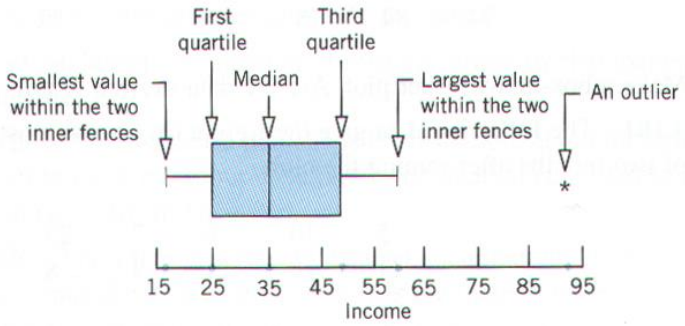



Figure 3.14

lower fenches sono evidenziati con un asterisco e sono detti **outliers**

In Fig.3-14 all'incirca il 50% dei dati cadono all'interno della scatola (Box), circa il 25% cadono a sinistra del Box ed il restante 25% cadono a destra del Box. Circa il 50% dei valori cadono a sinistra della mediana ed il restante 50% a destra. I valori del data set analizzato sono asimmetrici verso destra poiché l'iniziale 50% dei valori, al di sotto della mediana, sono distribuiti in un intervallo di dimensioni inferiori rispetto al restante 50% posto al di sopra della mediana.

## Calcolo del Box-and-wiskers plot - Work-book Excel "dbox.xls"

Per semplificare il calcolo del Box-and-wiskers plot si può utilizzare un work-book excel che consente il calcolo degli indicatori di posizione di un campione ordinato. L'esercizio che segue illustra il funzionamento del work-book

**Esercizio.** Il club locale di golf ha organizzato un torneo estivo fra giocatori di sesso maschile e femminile. I dati riportati in tabella indicano i punteggi ottenuti nel torneo a 18 buche da 17 concorrenti uomini e 15 concorrenti donne, selezionati a caso fra tutti i partecipanti. Si vuole

a. Costruire il box-and-wiskers plot di entrambi i data set con l'opportuno work-book.
b. Verificare se sono presenti outliers (valori $x \le Q1-1.5*IQR$ $x \ge Q3+1.5*IQR$).
c. Confrontare i campioni esaminati sulla base dei valori della:
   - mediana del campione,
   - il valore di IQR
   - il lower and upper fences.
d. Indicare quale gruppo ha ottenuto la migliore prestazione, sulla base del Box-plot, e degli indicatori di posizione centrale e di dispersione, e perché.

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Uomini | 87 | 86 | 92 | 97 | 110 | 76 | 81 | 92 | 112 | 95 | 107 | 102 | 97 | 108 | 85 | 101 | 99 |
| Donne | 101 | 100 | 87 | 95 | 98 | 81 | 117 | 107 | 103 | 97 | 90 | 100 | 99 | 94 | 94 | | |

I risultati dell'analisi sono nel foglio elettronico golf.xls che a sua volta richiama work-book dboxp_1.xls.