Statistica

L. Freddi

Statistica

La **statistica** è un insieme di metodi e tecniche per:

- raccogliere **informazioni** su un fenomeno
- **sintetizzare** l'informazione (elaborare i dati)
- **generalizzare** i risultati ottenuti

Perché è importante?

- In generale, perché consente di valutare criticamente tutte le informazioni basate su rilevazioni e sondaggi
- In particolare, perché è elemento essenziale nell'applicazione del **metodo scientifico**

Perché è importante?

Il prodotto XXX è particolarmente efficace contro il raffreddore; infatti su 100 pazienti trattati, ben 95, pari quindi al 95% dei casi, ha mostrato completa remissione della malattia entro una settimana di cura

Perché è importante?

il 75% delle persone intervistate si è dichiarata favorevole al tal partito politico

In ambito biologico

l'uso di appropriati metodi statistici consente di

- pianificare in modo adeguato la sperimentazione
- tenere sotto controllo l'errore sperimentale
- valutare l'affidabilità dei risultati ottenuti

Una prima distinzione

Generalmente si parla di

- **statistica descrittiva**: insieme di metodi e tecniche per l'esplorazione e la sintesi dell'evidenza empirica (dati)
- **inferenza statistica**: insieme di metodi e principi per inferire le caratteristiche generali di un fenomeno mediante l'osservazione di un insieme limitato di manifestazioni dello stesso

Le due categorie differiscono principalmente per gli obiettivi che l'analisi dei dati si pone.

Popolazioni

definizione

Si chiama popolazione statistica l'insieme di tutti gli elementi (individui, geni, cellule, ecc...) che si vogliono studiare.

Esempi:

- l'insieme dei lupi del Parco Nazionale d'Abruzzo,
- l'insieme degli abitanti di Milano,
- l'insieme dei valori di temperatura rilevati a Roma alle ore 14 dal 1/6/1998 al 31/5/2001,
- le altezze degli alunni di una classe di 30 bambini di 6 anni,

sono popolazioni statistiche composte da un numero finito di elementi.

Popolazioni e campioni

Invece

- gli esseri umani sulla terra,
- le larghezze delle corolle dei fiori,
- gli atomi o le molecole di un gas,

sono elementi di insiemi finiti, ma il loro numero è così grande che, a volte, sarà utile considerare la popolazione come infinita.

definizione

Si chiama campione qualunque sottoinsieme della popolazione, selezionato in modo opportuno.

Variabili

definizione

Fissata una popolazione, si chiamano variabili statistiche tutte quelle caratteristiche che variano al variare dei componenti della popolazione.

Esempi:

- il *colore* bianco, fulvo, nero, ecc..., della pelliccia degli esemplari di una certa specie,
- il *sesso* (maschio o femmina),

sono variabili statistiche **qualitative** (dette anche **attributi**).

Variabili quantitative

- l'*età* in mesi degli esemplari di lupo del Parco degli Abruzzi,
- il *numero* di cuccioli nati da ogni femmina,

sono variabili **quantitative discrete**.

- la temperatura di Roma rilevata alle ore 14 del primo Giugno di ogni anno

è una variabile **quantitativa continua**.

In generale le variabili discrete possono assumere solo un numero finito o una infinità numerabile di valori mentre quelle continue possono assumere tutti i valori compresi in un intervallo.

Matematicamente...

... le variabili statistiche sono funzioni

$$X : C \to M$$

dove C è il campione studiato e M è l'insieme dei valori osservati (*determinazioni o modalità*).

Spesso considereremo variabili discrete in cui $M = \{x_1, x_2, \ldots, x_n\}$ (se la variabile è continua M sarà un intervallo).

Altre lettere usate: Y, Z, y_i, z_i

Scale di misura

Le scale di misura più comunemente usate sono

$$\text{var. qualitative} \begin{cases} \text{nominale} \\ \text{ordinale} \end{cases} \qquad \text{var. quantitative} \begin{cases} \text{di intervallo} \\ \text{di rapporto} \end{cases}$$

Scala nominale. Se una variabile è misurata su scala nominale, si possono instaurare solo le seguenti relazioni tra le modalità

$$x_i = x_j \text{ oppure } x_i \neq x_j$$

Esempi: genere, gruppo sanguigno, sopravvivenza.

Scale di misura

Scala ordinale. Se una variabile è misurata su scala ordinale, si possono instaurare le seguenti relazioni tra le modalità

$$x_i \leq x_j \text{ oppure } x_i \geq x_j$$

Le modalità della variabile possono quindi essere ordinate.

Esempi: titolo di studio, grado di soddisfazione, lunghezze. Il giudizio sull'effetto di un fitofarmaco può essere espresso secondo la scala seguente:

1. peggioramento;
2. nessuna variazione;
3. lieve miglioramento;
4. deciso miglioramento;
5. guarigione.

Scale di misura

Scala di intervallo. Si misurano così le variabili quantitative per le quali lo zero è *convenzionale* (arbitrario). In tal caso non ha senso rapportare le misure ottenute, ed è invece corretto confrontare per differenze.

Esempio tipico: temperatura. In tre giorni diversi sono state rilevate le seguenti temperature:

Giorno	T $^\circ C$	Diff. $^\circ C$	T $^\circ F$	Diff. $^\circ F$
1	6		42,8	
		3		5,4
2	9		48,2	
		6		10,8
3	15		59	

La variazione tra il secondo ed il terzo giorno è doppia di quella tra il primo ed il secondo, indipendentemente dalla scala utilizzata.

Scale di misura

Scala di rapporto. Si misurano così le variabili quantitative per le quali lo zero è naturale.

Esempi: peso, concentrazione, lunghezza.

In questo caso le modalità possono essere confrontate per rapporto.

- la concentrazione di atrazina in un campione d'acqua è doppia rispetto a quella in un altro campione
- il peso specifico di un oggetto significa considerare il rapporto tra il peso dell'oggetto e quello di un equivalente volume di acqua a $4°C$.

Statistica descrittiva

La **statistica descrittiva** è un'insieme di metodi e tecniche per sintetizzare l'informazione contenuta nei dati.

Gli strumenti di sintesi sono essenzialmente di tre tipi:

- tabelle
- rappresentazioni grafiche
- indici sintetici

Attenzione! Quando sintetizziamo l'informazione contenuta nei dati, ne perdiamo una parte.

Gli strumenti di sintesi devono essere scelti in modo tale da:

- preservare, per quanto possibile, l'informazione rilevante per il problema analizzato
- eliminare l'informazione non necessaria

Distribuzioni di frequenza

La **frequenza** misura quante volte una certa modalità è stata osservata nel campione studiato.

Tipica rappresentazione tabellare per variabili qualitative o per variabili quantitative discrete. Nella tabella sono riportate:

- le **modalità** della variabile
- le **frequenze** associate a ciascuna modalità

Esempio: su 50 soggetti è stato rilevato il gruppo sanguigno. I risultati sono stati riportati nella tabella seguente

Gruppo	n_i	p_i
A	20	$0,40$
B	5	$0,10$
AB	2	$0,04$
0	23	$0,46$
$Tot.$	50	$1,00$

Frequenza assoluta

definizione

Siano
C un campione di una popolazione Ω costituito da N elementi,
$M = \{x_1, \ldots, x_k\}$ un insieme finito di modalità,
$X : C \to M$ una variabile statistica (ovviamente discreta).

Si chiama frequenza assoluta della modalità x_i il numero

$$n_i = \#\{c \in C \ : \ X(c) = x_i\} = \#X^{-1}(x_i), \quad i = 1, 2, \ldots, k.$$

Si chiama frequenza relativa il rapporto

$$p_i = \frac{n_i}{N}(\times 100), \quad i = 1, 2, \ldots, k.$$

Si ha

$$\sum_{i=1}^{k} n_i = N \qquad \text{e} \qquad \sum_{i=1}^{k} p_i = 1$$

Esempio

Su 50 soggetti è stato rilevato il gruppo sanguigno. I risultati sono stati riportati nella tabella seguente

Gruppo	n_i	p_i
A	20	$0,40$
B	5	$0,10$
AB	2	$0,04$
0	23	$0,46$
$Tot.$	50	$1,00$

Frequenza cumulata

definizione

Si chiama frequenza cumulata assoluta della modalità x_i il numero

$$N_i = \#\{c \in C \,:\, X(c) \leq x_i\} = \#X^{-1}(]-\infty, x_i]), \quad i = 1, 2, \ldots, n.$$

Si chiama frequenza cumulata relativa il rapporto

$$P_i = \frac{N_i}{N}(\times 100), \quad i = 1, 2, \ldots, n.$$

Esempio

Nella tabella seguente è riportata la distribuzione dei giudizi all'esame di licenza media rilevati su un gruppo di studenti

Giudizio	n_i	p_i	N_i	P_i
Suff.	8	0,1111	8	0,1111
Buono	29	0,4028	37	0,5139
Distinto	30	0,4167	67	0,9306
Ottimo	5	0,0694	72	1,0000
Tot.	72	1,00		

Esempio

Numero di pizze difettose (troppo grandi) prodotte da una pressa in un'ora (6 giorni di osservazione)

Giorno	n_i	p_i	N_i	P_i
1	4	0.10	4	0.10
2	10	0.25	14	0.35
3	12	0.30	26	0.65
4	6	0.15	32	0.80
5	4	0.10	36	0.90
6	4	0.10	40	1.00
Tot.	40	1.00		

Distribuzioni di frequenza

Vantaggi e svantaggi delle distribuzioni di frequenza:

+ Non si perde informazione rilevante (solo l'ordine di rilevamento va perduto)
- Scarso potere di sintesi se le modalità sono numerose
- Non utilizzabile per variabili continue.

In realtà l'ultimo punto non è del tutto vero ...

Distribuzione di frequenza per variabili continue

Se siamo disposti a rinunciare ad ulteriore informazione, la distribuzione di frequenza può essere costruita anche per variabili continue. Generalmente si opera nel modo seguente:

- si suddivide linsieme dei valori che la variabile può assumere in intervalli, detti **classi**;
- si determina il numero di osservazioni che cadono all'interno di ciascuna classe.

Esempio

Aziende agricole secondo la superficie agricola totale. Provincia di Udine.

Superficie	n_i	p_i
$0 \dashv 1$	2406	0.085
$1 \dashv 2$	3404	0.120
$2 \dashv 3$	2857	0.101
$3 \dashv 5$	4415	0.155
$5 \dashv 10$	6856	0.241
$10 \dashv 20$	5708	0.201
$20 \dashv 30$	1365	0.048
$30 \dashv 50$	751	0.026
$50 \dashv 100$	410	0.014
> 100	238	0.008
Totale	28410	1.000

Esempio

100 piante da fiore classificate in base alla larghezza della corolla

$x_i \dashv x_{i+1}$	n_i	p_i	N_i	P_i
$59,5 \dashv 62,5$	5	$0,05$	5	$0,05$
$62,5 \dashv 65,5$	18	$0,18$	23	$0,23$
$65,5 \dashv 68,5$	42	$0,42$	65	$0,65$
$68,5 \dashv 71,5$	27	$0,27$	92	$0,92$
$71,5 \dashv 74,5$	8	$0,08$	100	$1,00$

Come costruire le classi?

Non esistono regole assolute per la costruzione delle classi. In generale è buona norma:

- evitare di costruire classi con frequenze molto basse;
- modulare l'ampiezza delle classi in funzione della disponibilità di informazione "locale";
- *se possibile*, non variare l'ampiezza di classe (semplifica l'interpretazione).

Diagramma a barre - Popolazione Paesi UE 1993

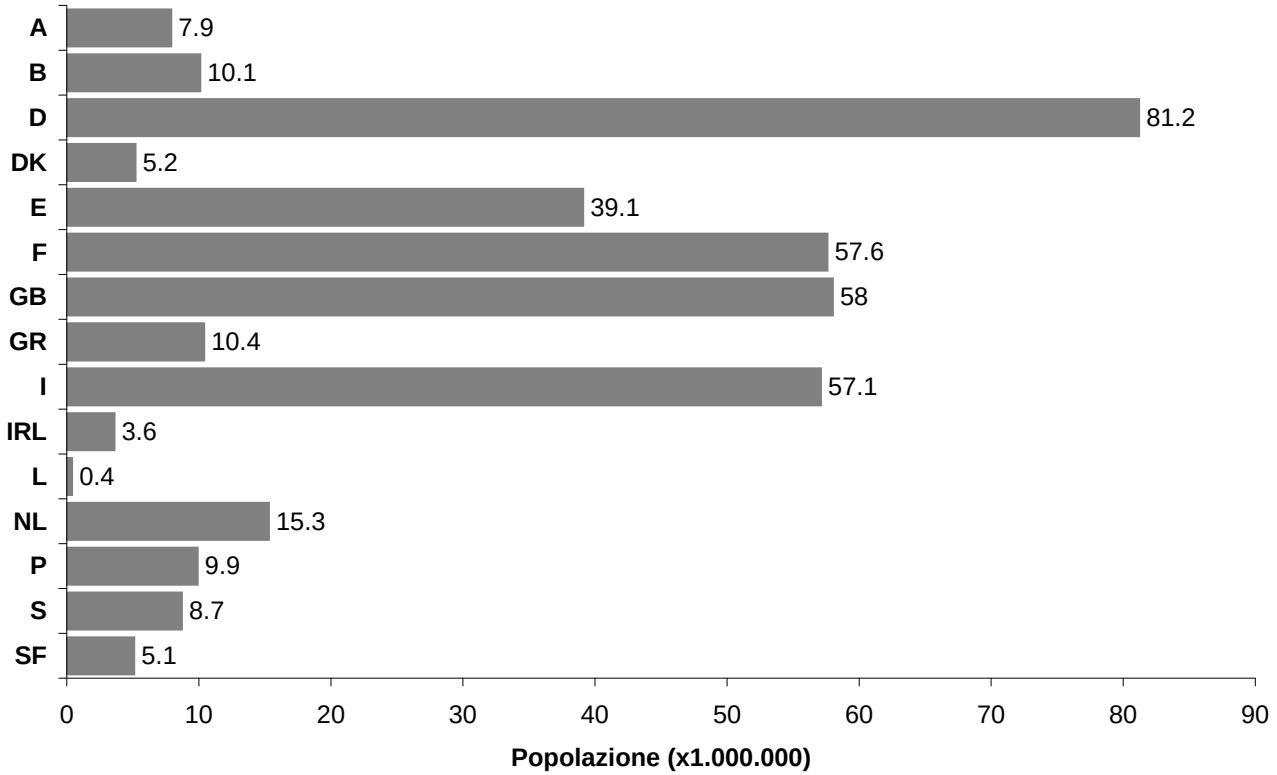
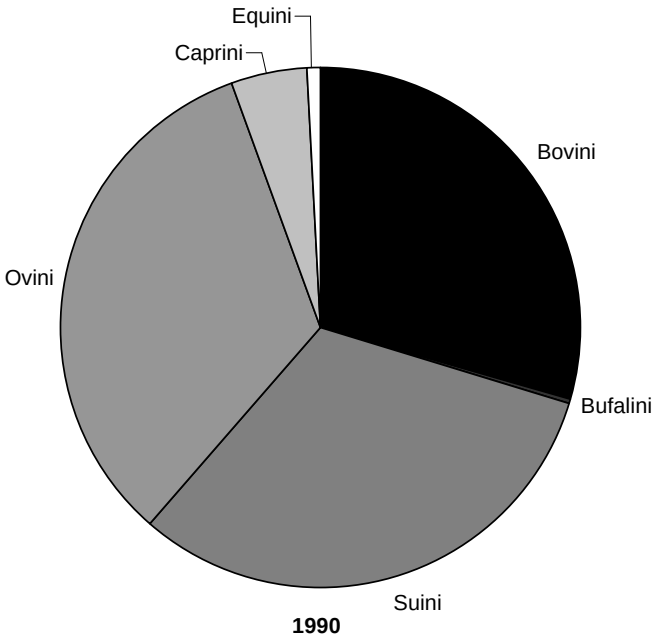


Diagramma a torta - Bestiame da allevamento per specie



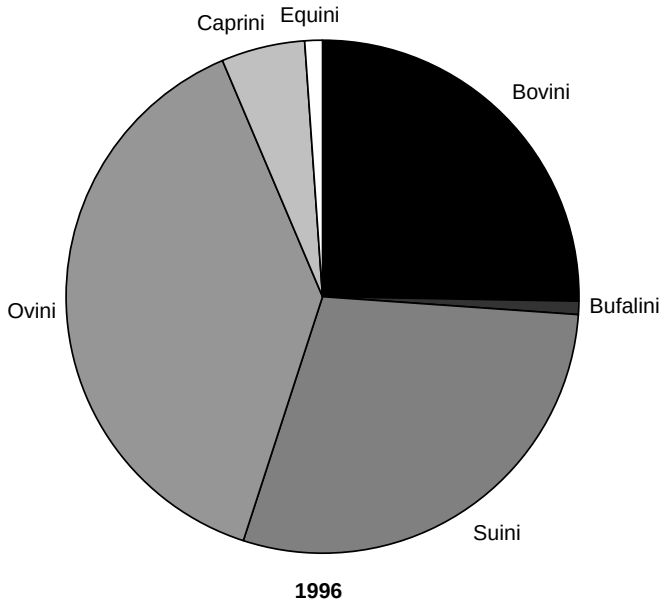
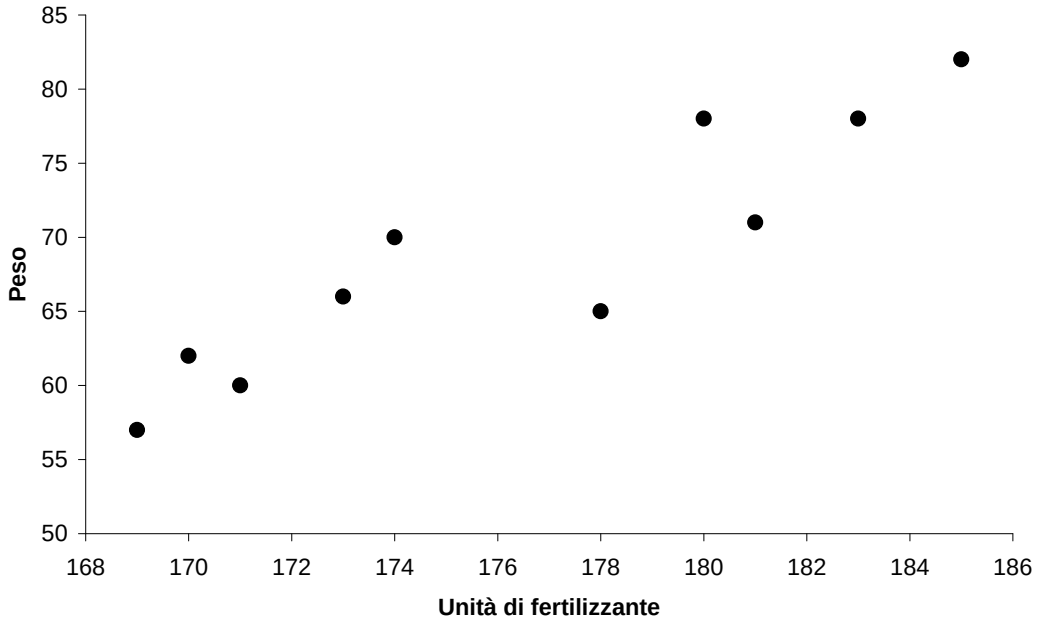
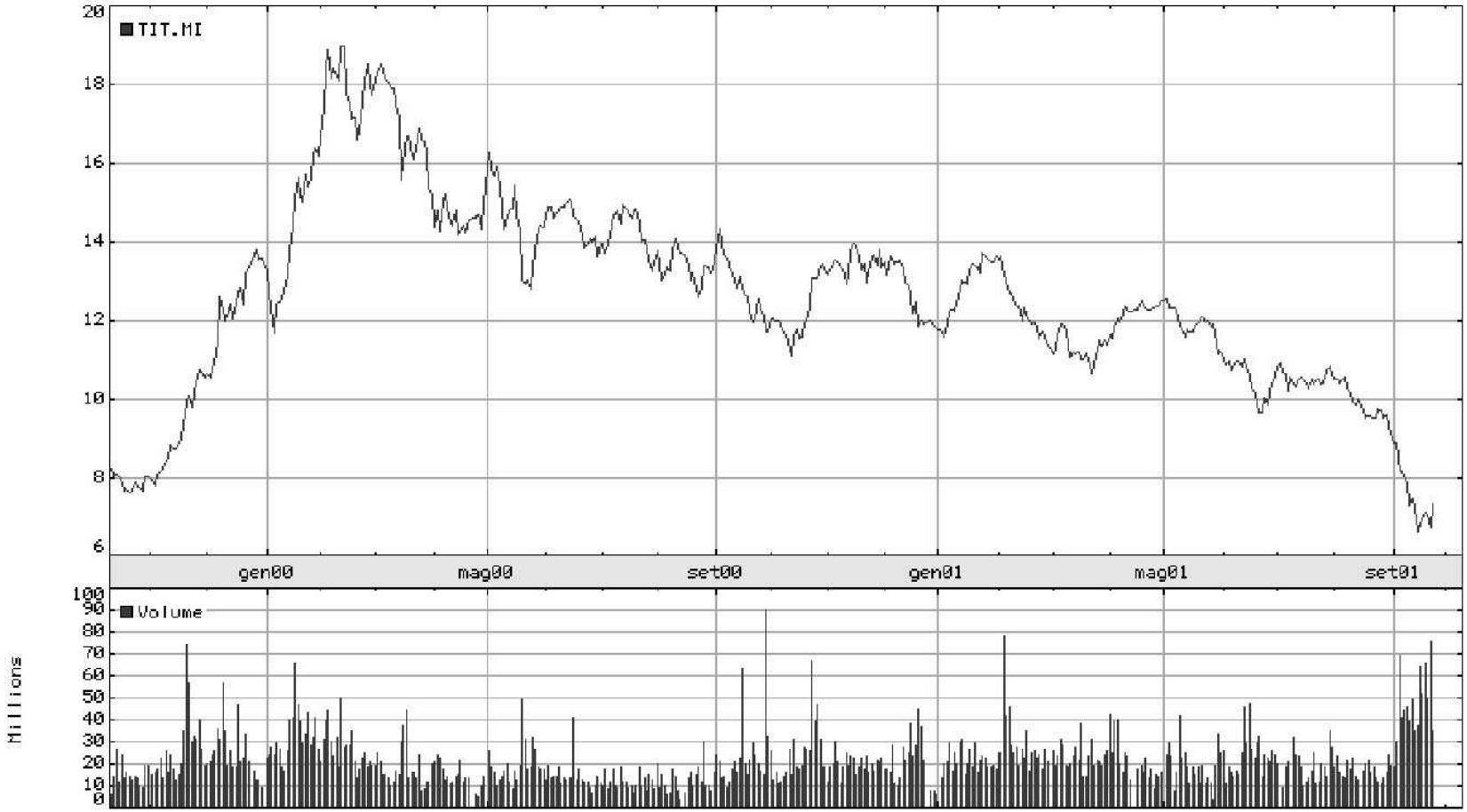


Diagramma di dispersione - Relazione dose-risposta



Serie storica



Grafici per variabili continue

Come rappresentare la distribuzione di frequenza di una variabile continua?

Se le classi sono di ampiezza diversa, le frequenze *non sono* direttamente *confrontabili*.

Per costruire un grafico che rappresenti in modo adeguato l'informazione è necessario eliminare l'effetto dell'ampiezza di classe.

Densità di frequenza

Il rapporto tra la frequenza e l'ampiezza (indicata con Δ_i) di una classe è detto densità di frequenza.

$$d_i = \frac{pi}{\Delta_i}$$

Le densità di frequenza *sono* fra loro *confrontabili*. La densità di frequenza è assoluta o relativa a seconda del tipo di frequenza utilizzato nel calcolo.

Istogramma di frequenza

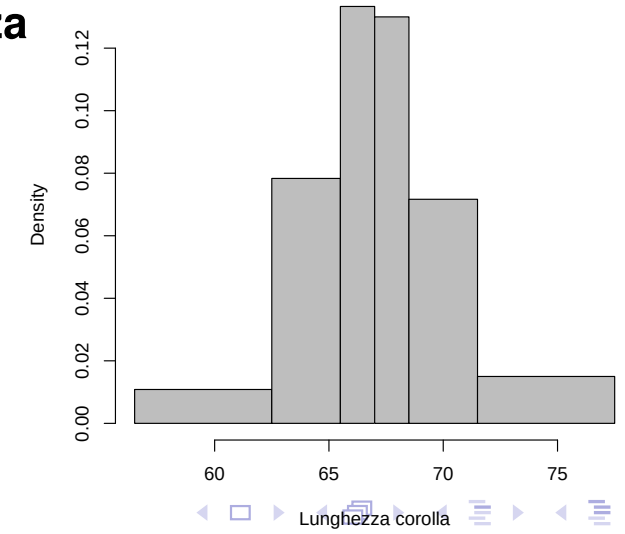
In un istogramma di frequenza ad ogni classe è associato un rettangolo:

- la base del rettangolo è pari all'ampiezza di classe;
- l'altezza del rettangolo è pari alla densità di frequenza;
- l'area del rettangolo è per costruzione la frequenza (assoluta o relativa) associata alla classe;

Distribuzione di piante in base alla lunghezza della corolla

$x_i \dashv x_{i+1}$	n_i	p_i	Δ_i	d_i
56,5 ⊣ 62,5	13	0,065	6,0	0,0108
62,5 ⊣ 65,5	47	0,235	3,0	0,0783
65,5 ⊣ 67,0	40	0,200	1,5	0,1333
67,0 ⊣ 68,5	39	0,195	1,5	0,1300
68,5 ⊣ 71,5	43	0,215	3,0	0,0717
71,5 ⊣ 77,5	18	0,090	6,0	0,0150

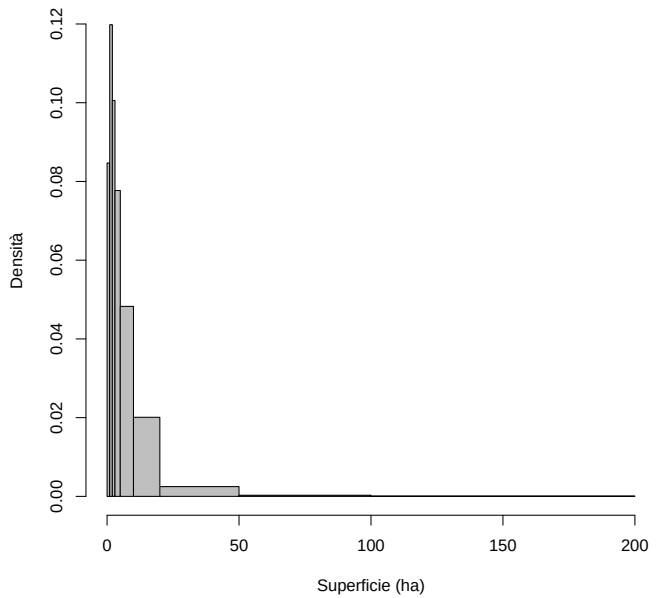
Istogramma - Piante in base alla lunghezza della corolla



Distribuzione delle aziende agricole per superficie agricola

Sup.	n_i	p_i	Δ_i	d_i
0 ⊣ 1	2406	0.085	1	0.08500
1 ⊣ 2	3404	0.120	1	0.12000
2 ⊣ 3	2857	0.101	1	0.10100
3 ⊣ 5	4415	0.155	2	0.07750
5 ⊣ 10	6856	0.241	5	0.04820
10 ⊣ 20	5708	0.201	10	0.02010
20 ⊣ 30	1365	0.048	10	0.00480
30 ⊣ 50	751	0.026	20	0.00130
50 ⊣ 100	410	0.014	50	0.00028
100+	238	0.008	100	0.00008
Totale	28410	1.000		

Istogramma - Aziende agricole per superficie agricola

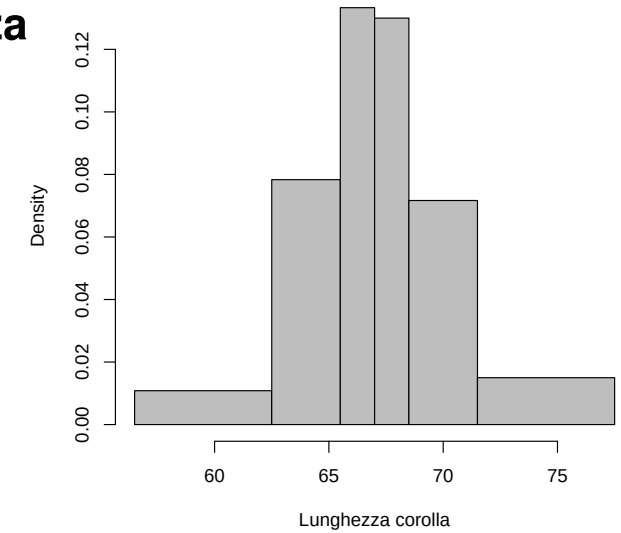


Distribuzione di piante in base alla lunghezza della corolla

$x_i \dashv x_{i+1}$	n_i	p_i	Δ_i	d_i
56,5 ⊣ 62,5	13	0,065	6,0	0,0108
62,5 ⊣ 65,5	47	0,235	3,0	0,0783
65,5 ⊣ 67,0	40	0,200	1,5	0,1333
67,0 ⊣ 68,5	39	0,195	1,5	0,1300
68,5 ⊣ 71,5	43	0,215	3,0	0,0717
71,5 ⊣ 77,5	18	0,090	6,0	0,0150

24

Istogramma - Piante in base alla lunghezza della corolla



25

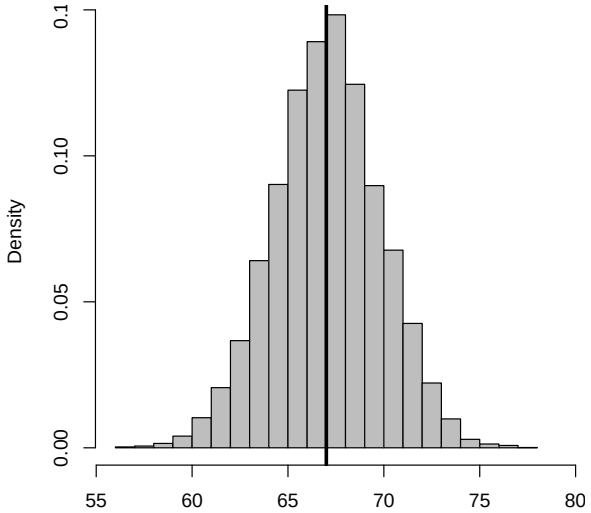
Caratteristiche dell'istogramma

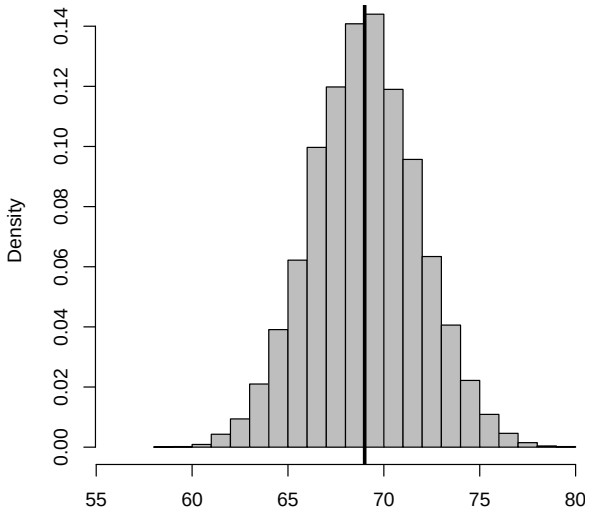
Da un istogramma è possibile desumere alcune rilevanti caratteristiche del fenomeno, per esempio:

- tendenza centrale
- dispersione
- grado di simmetria della distribuzione

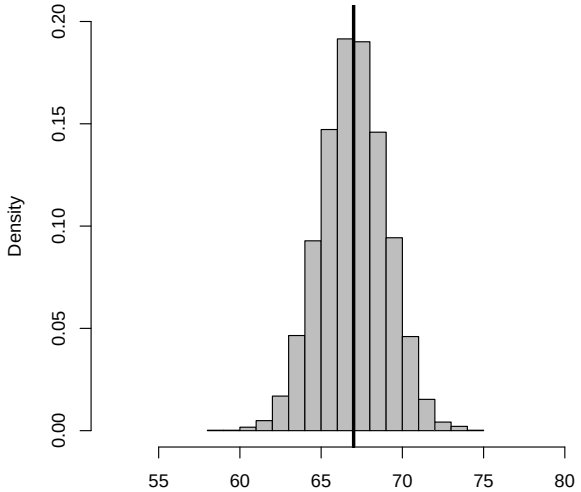
Illustriamo queste caratteristiche in alcuni esempi.

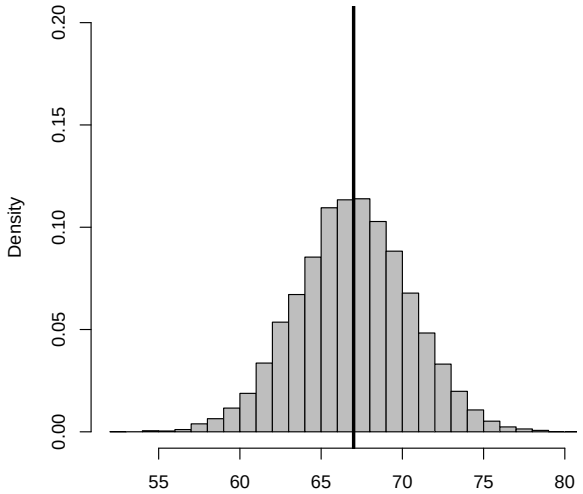
La tendenza centrale



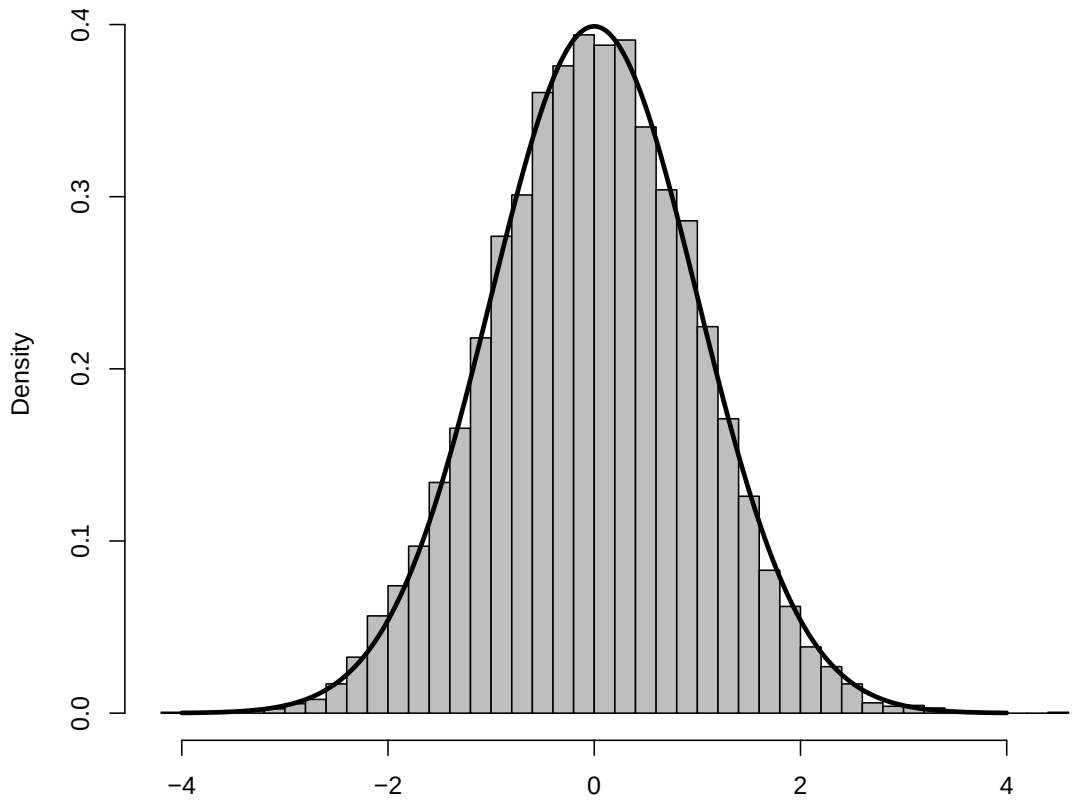


Il grado di dispersione

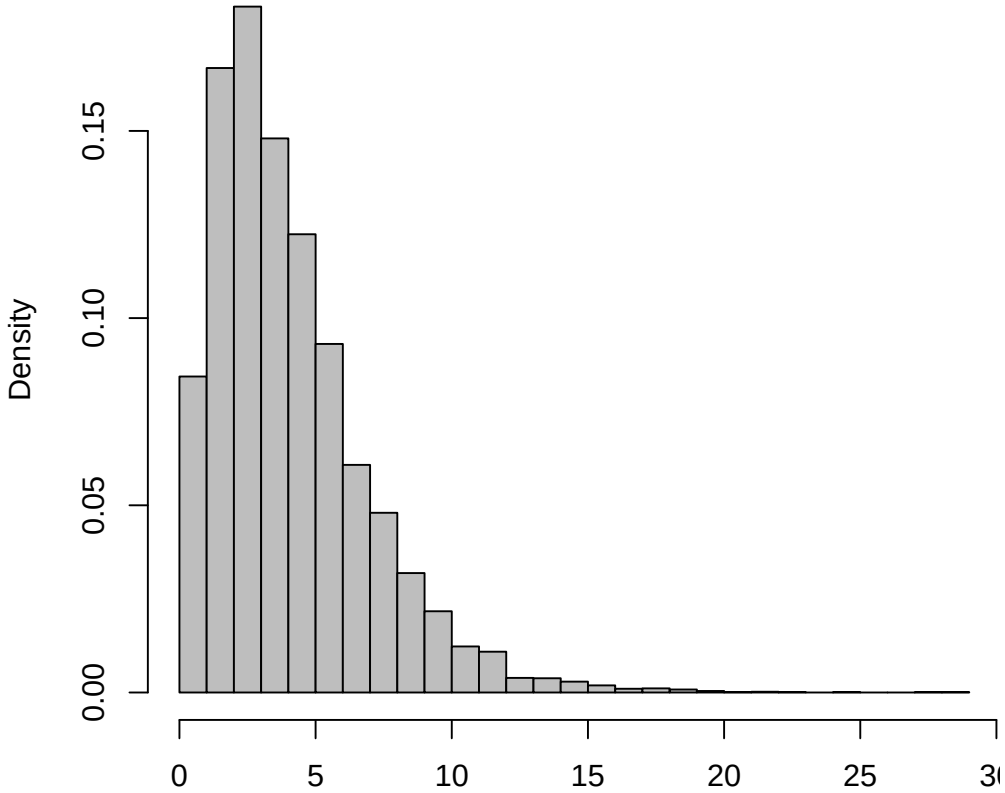




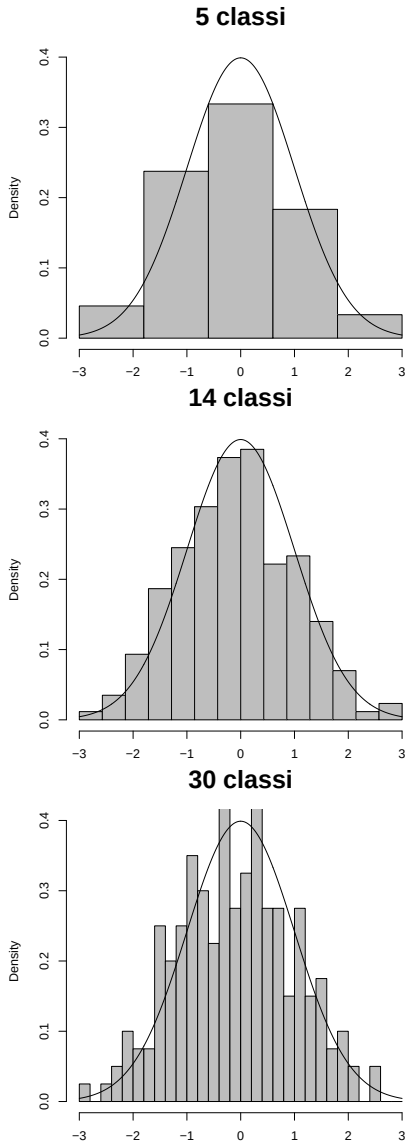
Simmetria ...



e asimmetria di una distribuzione



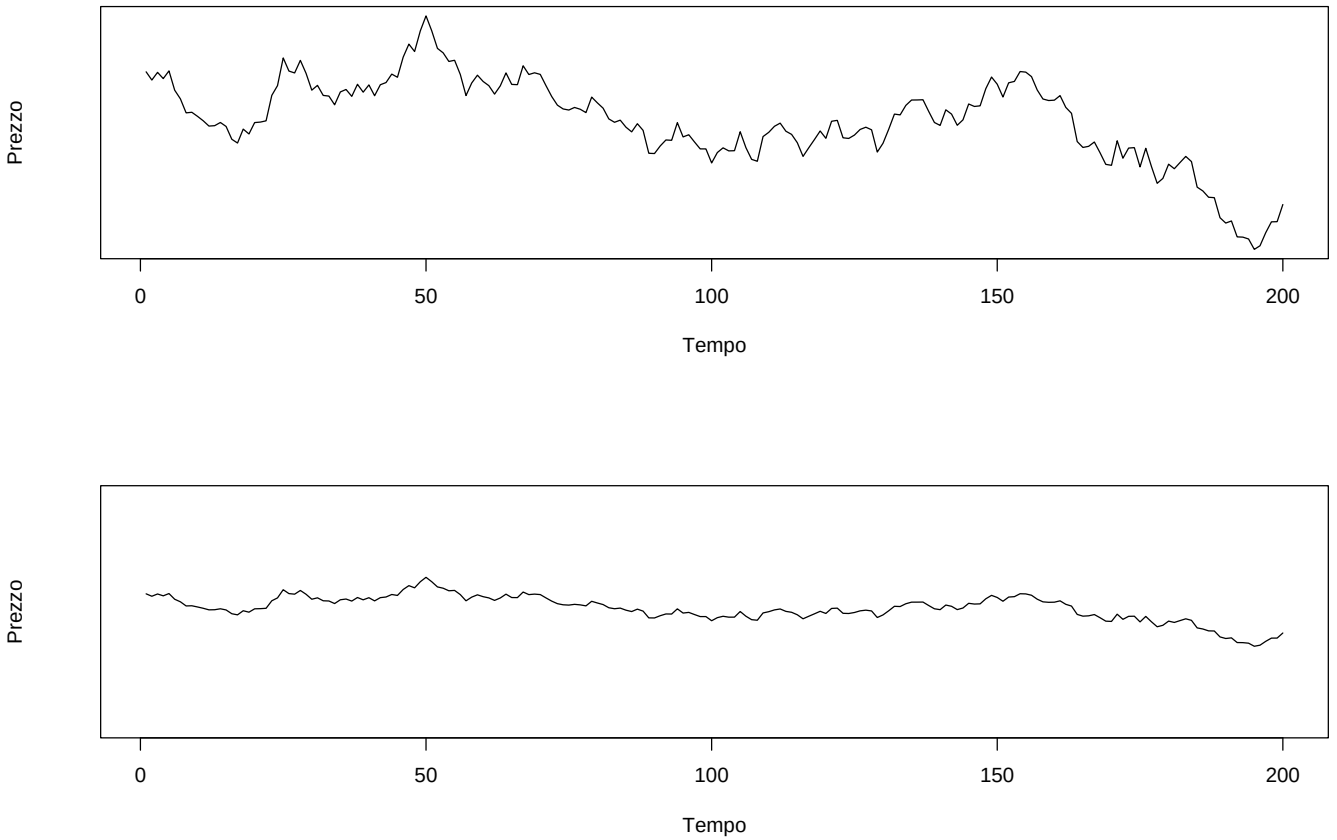
L'effetto dell'ampiezza di classe



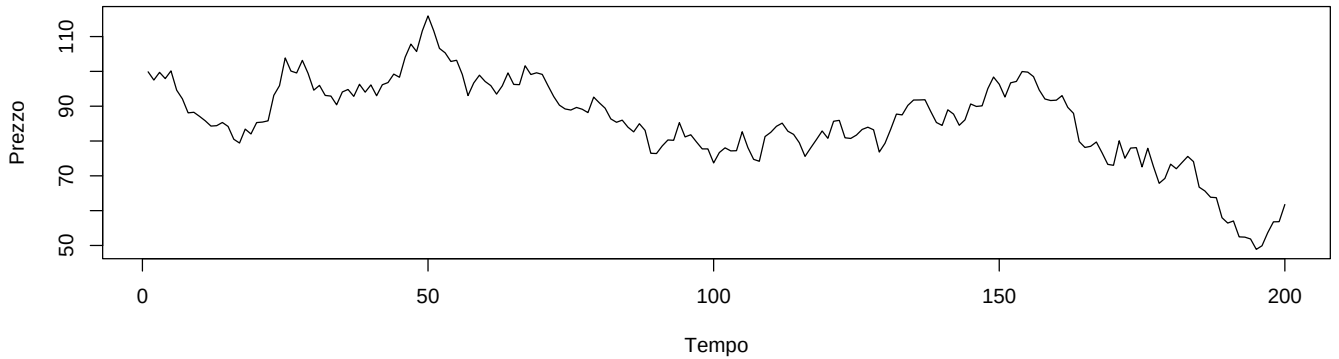
Vantaggi e svantaggi delle rappresentazioni grafiche

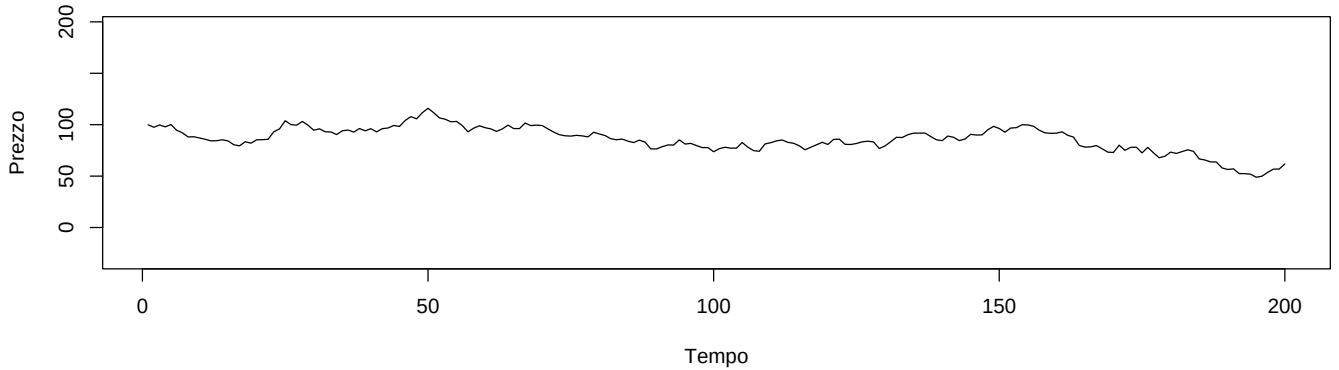
+ Conservano la maggior parte dell'informazione contenuta nei dati
+ Sono di immediata comprensione
- Nonostante la (presunta) semplicità, non sempre è chiaro quale sia la rappresentazione da utilizzare
- Possono essere usati in modo strumentale

Come mentire con un grafico



La rappresentazione corretta





Gli indici sintetici

Le caratteristiche più rilevanti di una distribuzione, per esempio

- la tendenza centrale del fenomeno
- il grado di dispersione
- la simmetria

possono essere rappresentate mediante numeri, detti **indici sintetici**.

Indici di posizione

Gli indici di posizione servono per individuare la tendenza centrale del fenomeno studiato. I più utilizzati sono:

- moda
- mediana
- media aritmetica

Moda

La **moda** di una distribuzione è la modalità più frequente (prevalente).

Qualora si utilizzi una distribuzione in classi per variabili continue, la **classe modale** è quella con la **densità** di frequenza più elevata.

Può essere utilizzata per qualunque tipo di variabile, ma è poco informativa.

Esempio

Gruppo	n_i	p_i
A	20	$0,40$
B	5	$0,10$
AB	2	$0,04$
0	23	$0,46$
Tot.	50	$1,00$

La moda (Mo) è il gruppo sanguigno 0.

Mediana

La **mediana** è il valore che occupa la posizione centrale nella distribuzione, tale che:

- metà delle osservazioni sono uguali o minori
- metà delle osservazioni sono uguali o superiori

La mediana divide in due parti di egual numero l'insieme dei valori osservati. Si può utilizzare solo per variabili misurate almeno su scala ordinale.

Calcolo della mediana

Per calcolare la mediana bisogna:

1. **ordinare** gli N valori osservati in ordine crescente
2. prendere il valore **centrale** nella graduatoria ordinata

Il modo di procedere per il secondo punto varia a seconda della numerosità del collettivo studiato.

N dispari

Se N è dispari allora esiste un unico valore che divide esattamente in due la distribuzione. Il valore centrale occupa la posizione

$$\frac{N+1}{2}$$

nella graduatoria ordinata.

$$\mathrm{Me} = X(\frac{N+1}{2}).$$

Calcolo della mediana

N pari

Se N è pari, si considerano valori centrali quelli che occupano le posizioni

$$\frac{N}{2} \text{ e } \frac{N}{2} + 1$$

Esistono quindi due mediane

$$\text{Me}_1 = X(\frac{N}{2}) \text{ e } \text{Me}_2 = x(\frac{N}{2} + 1)$$

Quando possibile (variabili *quantitative*) si usa come mediana la semisomma dei valori centrali

$$\text{Me} = \frac{X(\frac{N}{2}) + X(\frac{N}{2} + 1)}{2}$$

Esempio di calcolo

Nella tabella seguente sono riportati i giudizi (A, B, C o D) ottenuti ad un esame da 9 studenti.

Studente	1	2	3	4	5	6	7	8	9
Giudizio	*B*	*D*	*A*	*C*	*B*	*A*	*D*	*C*	*A*

Dovremo quindi ordinare i valori e scegliere come mediana quello che occupa la 5^a posizione

Posizione	1	2	3	4	5	6	7	8	9
Giudizio	*D*	*D*	*C*	*C*	$\boxed{B}$	*B*	*A*	*A*	*A*

Calcolo della mediana

Nel caso i valori osservati siano 10 (una D in più rispetto all'esempio precedente)

Posizione	1	2	3	4	5	6	7	8	9	10
Giudizio	*D*	*D*	*D*	*C*	*C*	*B*	*B*	*A*	*A*	*A*

bisogna considerare la 5^a e la 6^a posizione

$$\mathrm{Me}_1 = C, \quad \mathrm{Me}_2 = B$$

Calcolo su distribuzioni di frequenza

Qualora sia disponibile la distribuzione di frequenza cumulata, la mediana (classe mediana) corrisponde alla modalità (classe) associata alla prima frequenza cumulata relativa superiore al 50%.

Giudizio	n_i	p_i	N_i	P_i
Suff.	8	0,1111	8	0,1111
Buono	29	0,4028	37	0,5139
Distinto	30	0,4167	67	0,9306
Ottimo	5	0,0694	72	1,0000
Tot.	72	1,0000		

La mediana della distribuzione è "Buono".

Pregi e difetti della mediana

+ è un buon indicatore della tendenza centrale
+ risente poco di ciò che accade sulle code della distribuzione (è *robusta*)
- è difficile da trattare analiticamente

La media aritmetica

La *media aritmetica* è il più importante indice di posizione. La formula per il calcolo della media è:

$$\left.\begin{array}{c}\bar{X}\\ \mu\\ M(X)\end{array}\right\} = \frac{x_1 n_1 + x_2 n_2 + \ldots + x_k n_k}{N} = \frac{1}{N}\sum_{i=1}^{k} x_i n_i = \sum_{i=1}^{k} x_i p_i$$

ossia la media è la somma dei valori osservati divisa per la numerosità del campione. Si può utilizzare *solo* per variabili *quantitative*.

Nel caso particolare $k = N$ (cioè $n_i = 1$ per ogni i) si ha

$$\bar{X} = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Proprietà della media aritmetica

- La media aritmetica è sempre compresa tra il minimo ed il massimo dei valori osservati (internalità)

$$x_{\min} \leq \bar{X} \leq x_{\max}$$

- la somma degli scarti dalla media è sempre pari a zero

$$\sum_{i=1}^{k}(x_i - \bar{X})n_i = 0$$

- è equivariante per trasformazioni lineari, cioè se X e Y sono variabili statistiche legate dalla trasformazione lineare $Y = a + bX$, con a e b costanti, si ha $M(Y) = a + bM(X)$. Inoltre, date due variabili discrete X ed Y si ha $M(X + Y) = M(X) + M(Y)$

Pregi e difetti della media

+ è un buon indicatore della tendenza centrale
+ è semplice da trattare analiticamente
- risente in misura rilevante di ciò che accade sulle code della distribuzione (è *poco robusta*)

Indici di dispersione o di variabilità

Variabilità: in quale misura i valori osservati differiscono tra loro

Dispersione: in quale misura i valori osservati differiscono da un valore di riferimento

In realtà i due concetti sono (almeno parzialmente) sovrapponibili e noi useremo i due termini come sinonimi.

Campo di variazione (range)

Il campo di variazione o range è la differenza tra il massimo ed il minimo valore osservati:

$$R = x_{\text{max}} - x_{\text{min}}$$

Il campo di variazione è poco usato perché:

- trascura la maggior parte dell'informazione disponibile
- risente eccessivamente dei valori estremi

Scarto interquartile

Per eliminare il problema dei valori estremi, talvolta si usa lo **scarto interquartile**, ossia la differenza tra il terzo ed il primo quartile.

Primo quartile: lascia alla sua sinistra il 25% delle osservazioni

Terzo quartile: Lascia alla sua sinistra il 75% delle osservazioni

Rimane inalterato il problema dello scarso sfruttamento dell'informazione

Come sfruttare tutta l'informazione?

Gli indici visti in precedenza sono poco informativi. È possibile costruire un indice che sfrutti al meglio il contenuto informativo dei dati? Il grado di dispersione delle singole osservazioni è misurato dagli scarti

$$x_i - \bar{X}$$

Un buon indice di dispersione deve essere una sintesi di queste quantità.

Devianza

La devianza è la somma degli scarti dalla media al quadrato

$$\mathrm{Dev}(X) = \sum_{i=1}^{k} |x_i - \bar{X}|^2 n_i$$

- Elevando al quadrato, trascuriamo il segno degli scarti
- La devianza dipende dalla numerosità del campione
- L'unità di misura è il quadrato di quella della variabile

Varianza

La varianza si usa per eliminare l'effetto della numerosità del campione. Si può calcolare in due modi, usando

- la numerosità del campione (*varianza campionaria*)

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{k}|x_i - \bar{X}|^2 n_i = \sum_{i=1}^{k}|x_i - \bar{X}|^2 p_i$$

- i gradi di libertà (*varianza campionaria corretta*)

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{k}|x_i - \bar{X}|^2 n_i = \frac{N}{N-1}\sum_{i=1}^{k}|x_i - \bar{X}|^2 p_i$$

Gradi di libertà

Poiché la somma degli scarti dalla media è necessariamente uguale a zero, fissata la media solo $N - 1$ scarti sono liberi di variare (ossia di assumere un qualunque valore). Lo scarto rimanente deve assumere l'unico valore che consente di soddisfare il vincolo.

Esempio di calcolo La tabella seguente si riferisce all'altezza rilevata su 10 soggetti.

X	$r(X)$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1.82	8	0.064	0.004096
1.84	10	0.084	0.007056
1.71	3	−0.046	0.002116
1.75	5	−0.006	0.000036
1.81	7	0.054	0.002916
1.72	4	−0.036	0.001296
1.82	9	0.064	0.004096
1.68	2	−0.076	0.005776
1.75	6	−0.006	0.000036
1.66	1	−0.096	0.009216
17.56			0.03664

$$\bar{x} = 1.756, \quad \mathrm{Me}_1 = 1.75,\ \mathrm{Me}_2 = 1.75, \quad S^2 = \frac{0.03664}{9} = 0.004071$$

Proprietà della varianza

- La varianza è sempre maggiore o uguale a zero
- La varianza è invariante per traslazione

$$Y = a + X \implies \mathrm{Var}(Y) = \mathrm{Var}(X)$$

- La varianza non è invariante per cambiamenti di scala

$$Y = bX \implies \mathrm{Var}(Y) = b^2 \, \mathrm{Var}(X)$$

Scarto quadratico medio

Lo *scarto quadratico medio* o *deviazione standard* è la radice quadrata della varianza

$$\sigma = \sqrt{\sigma^2} \text{ oppure } S = \sqrt{S^2}.$$

È l'indice più frequentemente utilizzato perché è espresso nella stessa unità di misura della variabile d'interesse.

Coefficiente di variazione

Il coefficiente di variazione è dato da

$$CV = \frac{\sigma}{\bar{X}}$$

- È un numero puro (adimensionale)
- Elimina l'effetto dell'intensità media del fenomeno studiato.

Serve per fare confronti.

Il calcolo della varianza

La varianza può essere calcolata mediante una formula alternativa:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{k} x_i^2 n_i - \bar{X}^2 = \overline{X^2} - \bar{X}^2$$

oppure

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{k} x_i^2 n_i - \frac{N}{N-1}\bar{X}^2$$

(dimostrazione: basta sviluppare il quadrato e usare la definizione di media aritmetica).

I vantaggi: l'uso della formula semplificata consente

- di ottenere il risultato con meno operazioni
- di ridurre gli errori dovuti ad arrotondamenti

Esempio di calcolo

La varianza dell'altezza rilevata su 10 soggetti può essere calcolata più semplicemente.

X	X^2
1.82	3.3124
1.84	3.3856
1.71	2.9241
1.75	3.0625
1.81	3.2761
1.72	2.9584
1.82	3.3124
1.68	2.8224
1.75	3.0625
1.66	2.7556
17.56	30.872

$$S^2 = \frac{1}{9} \cdot 30.872 - \frac{10}{9} \cdot 1.756^2 = 0.004071$$

Esercizi consigliati

Svolgere gli esercizi da 12.1 a 12.9 del testo consigliato.