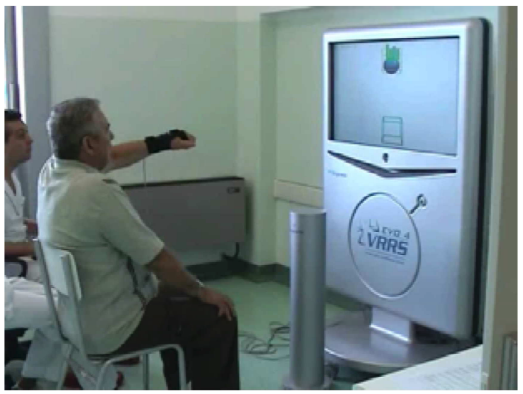
Statistica descrittiva univariata: Dati, grafici, misure di sintesi

Laura Ventura
Dipartimento di Scienze Statistiche
Università degli Studi di Padova
ventura@stat.unipd.it

Materiale a cura di Laura Ventura e Alessandra Salvan
Cagliari, Novembre 2019

Caso di studio:
Terapie di riabilitazione per l'apprendimento motorio del braccio

- **Dataset**: misurazioni relative ad uno studio sull'apprendimento motorio di un gruppo di pazienti, esposti al trattamento con realtà virtuale (IRCCS San Camillo, Lido di Venezia).
- **Variabile di interesse**: `FIM` (*Functional Indipendence Measure*), scala dell'autonomia del paziente con valori da 0 (non autosufficienza completa) a 130 (completa autonomia).
- Si hanno anche **due trattamenti**: 27 pazienti sono stati sottoposti ad una terapia di riabilitazione in un ambiente virtuale (casi, `TRATTAMENTO=1`) e 20 pazienti sono stati sottoposti ad una terapia convenzionale (controlli, `TRATTAMENTO=2`).
- La variabile `FIM` è stata misurata sia prima (`FIMPRE`) che dopo (`FIMPOST`) la terapia ricevuta, subito dopo un infarto.



Caso di studio (i dati)

- **Il dataset**:

	TRATTAMENTO	FIMPRE	FIMPOST
1	2	124	124
2	2	108	110
	...		
	...		
46	1	111	113
47	1	99	108

- **Obiettivo:** Si intende verificare se c'è un miglioramento della performance motoria dell'arto a seguito della terapia, e se il gruppo trattato con la realtà virtuale ha un miglioramento superiore rispetto al gruppo di controllo.

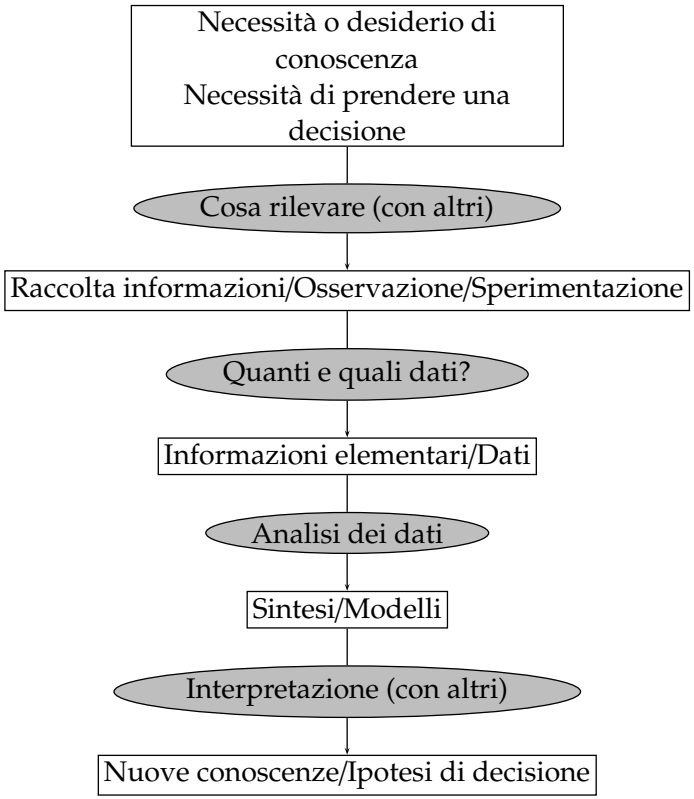
Un passo indietro:
Che cos'è la Statistica?

Di che cosa si occupa la Statistica?

- Fisica: fenomeni naturali
- Sociologia: fenomeni sociali
- Geologia: fenomeni che riguardano la crosta terrestre
- Astronomia: fenomeni celesti
- Biologia: fenomeni della vita (biologici)
- Medicina: fenomeni che riguardano lo stato di salute
- Economia: fenomeni di gestione delle risorse
- Chimica: fenomeni sulla composizione e trasformazioni della materia
- . . . . . .
- La Statistica si occupa di fenomeni reali!
  Si presta dunque a tutte le altre discipline.
  La Statistica studia i dati.

Giocare nel giardino dei vicini

Analisi statistica



La STATISTICA è un insieme di metodi rigorosi per raccogliere i dati ed estrarne informazione, utilizzando gli strumenti della Matematica.

Come opera la Statistica?

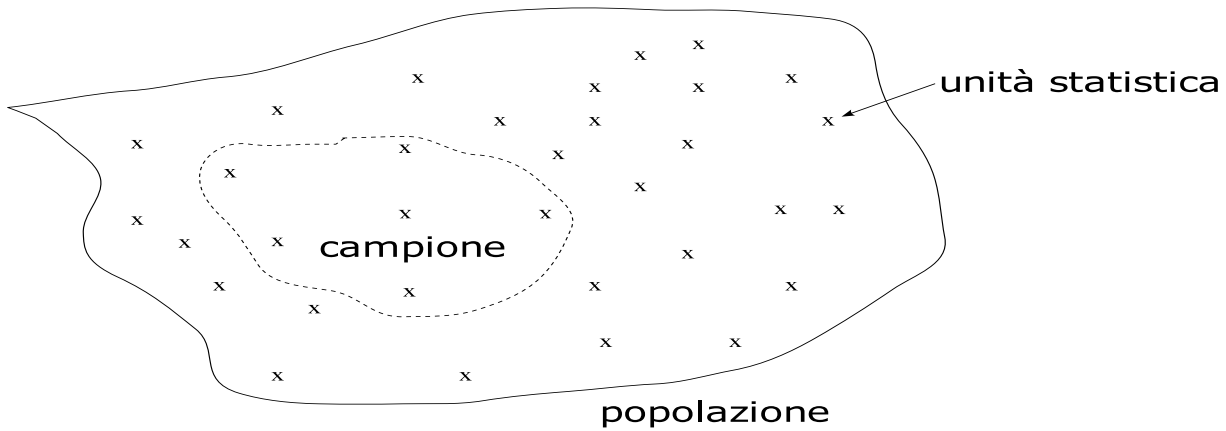
- Il punto di partenza di una indagine statistica è costituito da un insieme (che chiamiamo popolazione di riferimento), disomogeneo all'interno (ovvero non tutti gli elementi sono uguali tra di loro) e che costituisce la parte del mondo che ci interessa.
- Gli elementi di questo insieme (che di volta in volta nei problemi concreti saranno persone, animali, batteri, immagini raccolte da un satellite,...) vengono convenzionalmente indicati come unità statistiche.
- In genere, i ricercatori studiano un sottoinsieme della popolazione relativamente piccolo (campione) e desiderano trarre conclusioni circa l'intera popolazione.
- Inferenza: come utilizzare le informazioni nel campione per trarre conclusioni sulla distribuzione delle variabili di interesse nella popolazione.
- È importante anche poter associare alle analisi condotte su un campione una valutazione dell'affidabilità dei risultati.

Statistica Descrittiva vs Inferenza Statistica

- Statistica descrittiva: analisi esplorativa dei dati.
- Inferenza statistica: vogliamo utilizzare le informazioni del campione per fare delle affermazioni sulle caratteristiche di tutta la popolazione.
- Tra Statistica Descrittiva ed Inferenza Statistica esiste una ovvia fratellanza e, nelle applicazioni, i problemi di inferenza vengono normalmente affrontati in accordo allo schema:

$$\begin{array}{ccc} \text{descrizione} & & \text{affermazioni sulle} \\ \text{caratteristiche} & \rightarrow & \text{caratteristiche} \\ \text{del campione} & & \text{della popolazione} \end{array}$$

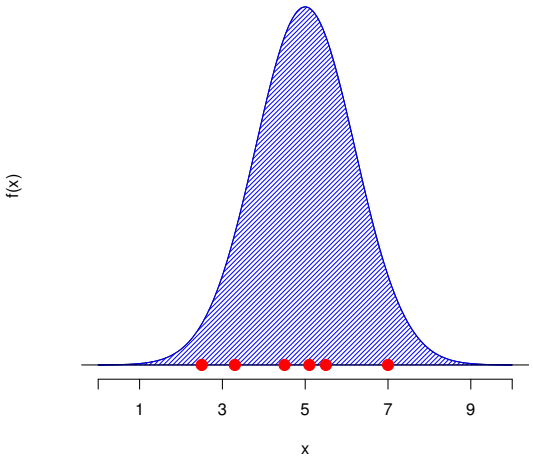
Con le informazioni rilevate sulle unità statistiche appartenenti al campione vogliamo produrre affermazioni su tutta la popolazione.



Inferenza Statistica e Probabilità

- Il trucco alla base dell'inferenza statistica consiste nel descrivere la relazione tra popolazione e campione utilizzando il Calcolo delle Probabilità.
- Interpreteremo i risultati sperimentali (i dati) come uno dei tanti risultati che un meccanismo probabilistico (esperimento casuale) poteva fornirci.

Si può considerare un modello matematico per il processo generatore dei dati: i dati sono visti come risultato di un esperimento casuale.



- In Statistica, il punto di vista è rovesciato: si dispone dei dati, e da questi, assunta la loro generazione da una distribuzione di probabilità, si cerca di risalire alla legge di probabilità (in parte ignota) corrispondente.

I dati e le variabili (caratteri)

Analisi esplorativa

- La prima fase di ogni analisi statistica è rappresentata dall'organizzazione e dalla sintesi dei DATI, le informazioni raccolte sulle UNITÀ STATISTICHE che compongono il CAMPIONE.
- Concetti e strumenti fondamentali dell'analisi esplorativa sono:
  - Variabili e tipi di variabili (qualitative sconnesse o ordinali, quantitative discrete o continue).
  - Frequenze (assolute, relative, percentuali, cumulate) e tabelle.
  - Grafici (a torta, a barre, istogramma).
  - Misure di posizione (media, mediana, moda, quantili).
  - Misure di variabilità (varianza, scarto interqurtile, campio di variazione).

Un po' di terminologia: i dati e le variabili

- Prima di fare cose "divertenti" con i dati, è necessario conoscere un po' di gergo per chiamare le cose con il nome giusto.
- I DATI sono una raccolta di informazioni (espresse in forma numerica).
- Le entità (individui, ore del giorno, ...) che vengono osservate nello studio sono dette UNITÀ STATISTICHE (casi).
- L'insieme di tutte le unità statistiche di interesse per lo studio è detto POPOLAZIONE di riferimento.
- Invece, un sottoinsieme di unità statistiche selezionate (spesso casualmente) da una popolazione è detto CAMPIONE. La dimensione del campione può variare da poche unità a molte migliaia di osservazioni.
- Una quantità di interesse nella popolazione è detta parametro, mentre la quantità calcolata sul campione è detta statistica.

ESEMPIO: La popolazione oggetto di studio è l'insieme di tutti i pazienti affetti da patologia simile, anche in futuro (si tratta di una popolazione virtuale).
Il campione è costituito dai $n = 47$ pazienti che sono entrati nell'esperimento.

Variabili e modalità

DEF: Una VARIABILE (o CARATTERE) è una caratteristica di interesse rilevata sulle unità statistiche (ad esempio, età, peso, trattamento, ...).

Il termine 'variabile' evidenzia che la caratteristica di interesse può assumere una pluralità di valori. L'insieme dei valori possibili si può pensare noto, ma prima di fare l'osservazione su una unità statistica, non sappiamo quale valore si osserverà.

DEF: I valori distinti assunti da una variabile sono detti MODALITÀ della variabile. Le modalità si presumono note preliminarmente.

Esempio: nello studio sul trattamento con la realtà virtuale, la variabile *FIM* può assumere valori nell'intervallo (0, 130). Le modalità sono dunque tutti i numeri reali appartenenti a questo intervallo.

Esempio: in uno studio sulla biodiversità, si può osservare la variabile *numero di esemplari di lupo* avvistati in una settimana da un certo punto di osservazione. Le modalità sono i valori $0, 1, 2, 3, \ldots$ (i numeri naturali), anche se difficilmente si osserveranno valori grandi.

Tipi di variabili

Una variabile può essere:

• QUALITATIVA o CATEGORIALE quando le sue modalità sono espresse in forma verbale (*sesso*, *livello di istruzione*, *trattamento*, ...).

A sua volta una variabile qualitativa può essere:
– SCONNESSA o NOMINALE se non esiste nessun ordinamento tra le modalità.
Esempi:

la variabile *sesso* con modalità M e F;

la variabile *modo di somministrazione* con modalità ORALE, ENDOVENA, ...

– ORDINALE se è possibile individuare un ordinamento naturale delle modalità.
Esempi:

la variabile *livello di istruzione* con modalità ELEMENTARE, MEDIA INFERIORE, MEDIA SUPERIORE, ...;

la variabile *giudizio* con modalità INSUFFICIENTE, SUFFICIENTE, DISCRETO, OTTIMO.

Tipi di variabili

• Se le modalità sono solo due si parla di variabili DICOTOMICHE o BINARIE (*sesso*, *presenza*, ...). A volte le due modalità sono espresse con valori numerici (0,1, oppure 1,2,...), ma il valore del numero non vuol dire assolutamente nulla!!

Oppure, una variabile può essere:

• QUANTITATIVA (o NUMERICA) quando le modalità sono espresse da numeri (età, peso,...). A sua volta una variabile quantitativa può essere:

– DISCRETA quando l'insieme delle modalità è finito o numerabile (stessa cardinalità dell'insieme dei naturali). Esempi:
- la variabile *numero di 'teste' in 10 lanci di una moneta*, con modalità 0,1, ..., 10;
- le variabili *numero di sedute*, *numero di figli*, ... con modalità 0, 1, 2, ...;

– CONTINUA quando l'insieme delle modalità è un intervallo, ossia un sottoinsieme, eventualmente illimitato, dei numeri reali. Esempi:
- la variabile *peso* (in kg) che ha come modalità possibili tutti i valori positivi,
- la variabile *dose* di un dato farmaco (in mg) con modalità da zero a 1000mg.
- eventuale suddivisione in classi.

- VARIABILI QUALITATIVE vs QUANTITATIVE
  - A seconda del tipo di variabili osservate, sono possibili diverse analisi statistiche.
  - Ci sono degli strumenti statistici appositi per studiare tipi diversi di variabili.
  - Tra le varie tipologie di dati è implicita una gerarchia (le variabili quantitative possono essere discretizzate, le variabili quantitative discrete possono essere tradotte in variabili qualitative ordinali, quelle ordinali possono essere considerate nominali). Le analisi statistiche sono più ricche, per così dire, ascendendo la gerarchia.
- DATI UNIVARIATI vs MULTIVARIATI
  - Le analisi univariate considerano una sola variabile rilevata sulle unità.
  - Nello studio congiunto di due variabili si parla di analisi bivariata.
  - Lo studio congiunto di due o più variabili è detto analisi multivariata (ovviamente il multivariato include il bivariato).

Esempio sui trattamenti (per fissare la terminologia)

Vogliamo studiare quale tra due trattamenti, 1 e 2, è migliore.

$\Downarrow$

La popolazione di riferimento è l'insieme di tutti i pazienti affetti da quella particolare patologia (oggi, ma anche domani, ...). Le unità statistiche sono i pazienti. In questo caso la popolazione è virtuale e si osserva un campione.

Campione di 47 unità, 27 unità sono trattate con 1 e 20 unità con 2.
All'inizio e alla fine della terapia si valuta, per ogni unità, la `FIM`.

$\Downarrow$

DATI				VARIABILI	TIPO
unità	trattamento	FIMPRE	FIMPOST	*trattamento*:	dicotomica
1	1	124	124		
2	1	108	110	`FIM`:	quantitativa
...	...	...			
27	1	111	113		
...	...	...			
47	2	99	108		

Esercizi

(1) La media della pressione sistolica sanguigna in un maschio adulto sano è ritenuta pari a circa 129. Si è misurata la pressione di 100 maschi adulti sani appartenenti ad una comunità le cui abitudini dietetiche potrebbero essere causa dell'aumento dei valori della pressione. La pressione sistolica sanguigna è un esempio di variabile continua, discreta o dicotomica? L'unità statistica è la comunità, il singolo maschio o il singolo valore della pressione?

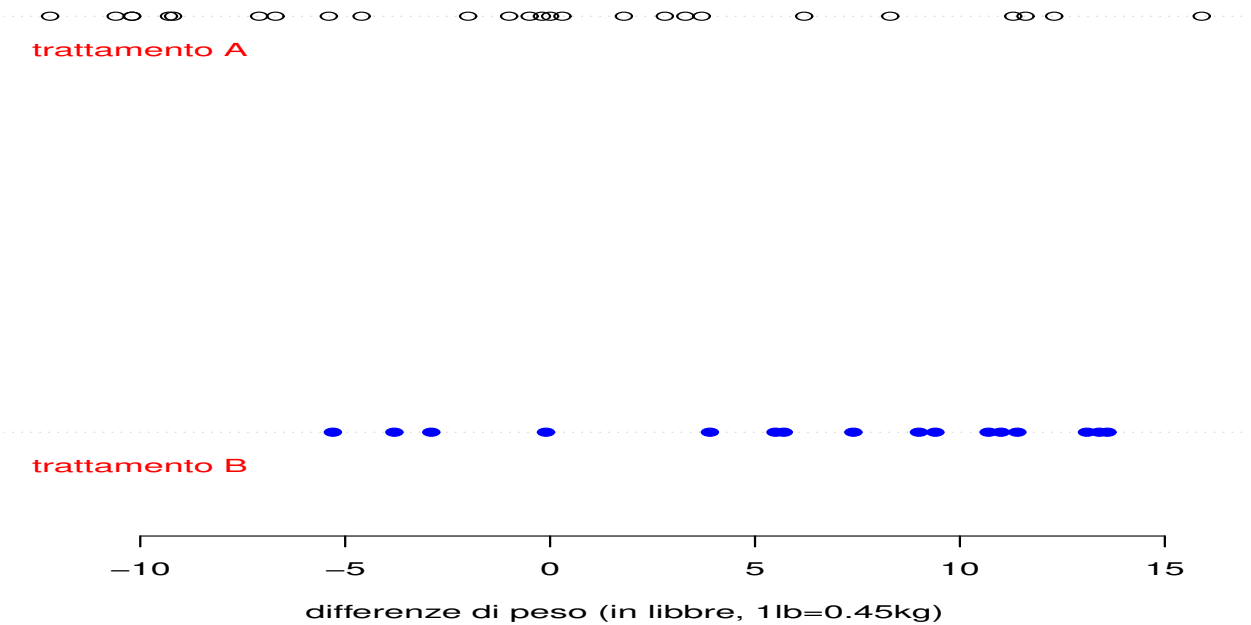
(2) 100 appezzamenti di terreno di uguale dimensione e coltivati con un certo ortaggio sono stati divisi in 4 gruppi di 25 appezzamenti ciascuno. Ciascun gruppo è stato poi fertilizzato usando 4 diverse dosi di una certa sostanza (dose 1 = 1hg, dose 2 = 2hg, dose 3 = 3hg, dose 4 = 4hg). La variabile "dose di fertilizzante" è un esempio di variabile continua, discreta o categoriale?

(3) Indicare la natura e le modalità delle seguenti variabili: Sesso, Numero di figli, Reddito familiare (in euro), Prezzo all'ingrosso di una merce (in euro), Corso di laurea frequentato, Altezza (in cm), Religione.

Presentazione dei dati: Frequenze, Tabelle e Grafici

Caso di Studio

- Confronto di trattamenti (Medicina): 43 pazienti anoressiche, divise a caso in 2 gruppi di numerosità 26 e 17, sottoposti a 2 diversi trattamenti (A e B). Si misura la variazione di peso tra prima e dopo il trattamento.



Presentazione dei dati: Frequenze e Tabelle

Le tabelle sono uno dei metodi migliori per presentare un insieme di dati.

Può sembrare banale, ... MA ci vuole un po' di attenzione!

Lo scopo di una tabella è sintetizzare un insieme di osservazioni, e di trasformare un insieme di dati in un formato facile da capire.

ESEMPIO: I dati grezzi nei due gruppi di pazienti anoressiche sono riportati di seguito. Si tratta delle differenze di peso (in libbre) nei due gruppi le pazienti trattate con la terapia A (placebo) e le pazienti trattate con la terapia B (farmaco).

trattam. A	-0.5	-9.3	-5.4	12.3	-2.0	-10.2	-12.2	11.6	-7.1	6.2
	-0.2	-9.2	8.3	3.3	11.3	0.0	-1.0	-10.6	-4.6	-6.7
	2.8	0.3	1.8	3.7	15.9	-10.2				
trattam. B	11.4	11.0	5.5	9.4	13.6	-2.9	-0.1	7.4	21.5	-5.3
	-3.8	13.4	13.1	9.0	3.9	5.7	10.7			

Dai dati grezzi alle tabelle di frequenza

I DATI GREZZI sono costituiti da tutte le misurazioni effettuate su ognuna delle unità statistiche prese in esame.

I dati grezzi, soprattutto in presenza di molte osservazioni, non permettono di cogliere in maniera sintetica le caratteristiche del fenomeno in esame.

Sono in genere "troppi" per cercare di capire qualcosa solamente "guardandoli".

Nell'esempio, non si capisce se vi sono misurazioni che possono essere sbagliate nei dati osservati o quale trattamento sembra preferibile. E la situazione è tanto più complicata quanto più grande è il numero n di osservazioni.

Tabelle di frequenza

Il tipo di tabella comunemente usato per riassumere i dati è la tabella di frequenza.

Supponiamo di aver osservato una sola variabile (analisi statistica univariata) su n unità.

Per variabili qualitative o quantitative discrete con poche modalità si conta il numero di volte che ciascuna modalità è stata osservata.

Per variabili quantitative si procede raggruppando i valori della variabile in una serie di intervalli non sovrapposti. Quindi si conta quante osservazioni cadono nei vari intervalli.

DEF: La FREQUENZA (assoluta) di una modalità, o di un intervallo di valori, è il numero di unità che presentano quella modalità, o un valore della variabile nell'intervallo.

DEF: Una tabella di frequenza consiste nella lista di modalità osservate, o intervalli, e nella frequenza corrispondente a ciascuna di esse.

Tabelle di frequenza

Chiamiamo X la variabile rilevata sulle n unità statistiche osservate.

E chiamiamo $x_1, x_2, \ldots, x_k$ le diverse modalità osservate
(possono essere in forma verbale o numerica).

La tabella di frequenza associa alle modalità della variabile X, qualitativa o quantitativa, le corrispondenti frequenze assolute $n_1, \ldots, n_k$,
ossia la DISTRIBUZIONE DI FREQUENZA assoluta.

X	Freq. Ass.
x_1	n_1
x_2	n_2
$\ldots$	$\ldots$
x_j	n_j
$\ldots$	$\ldots$
x_k	n_k
Totale	n

Per caratteri quantitativi continui:

– Si raggruppano i valori in intervalli.

– Alcune informazioni vengono perse ma si guadagna in maneggiabilità.

– Gli intervalli non devono essere troppi né troppo pochi, altrimenti i vantaggi offerti dalla sintesi vengono persi.

– L'ampiezza degli intervalli è quasi sempre uguale per poter facilitare il confronto tra classi diverse.

Nota ... Non n righe ma k righe, con $k =$ numero di modalità distinte di X.

Tabella di frequenza per una variabile quantitativa continua

Nell'ESEMPIO sul trattamento dell'anoressia, utilizziamo intervalli di ampiezza 5.

Considerando solo gli intervalli con frequenza positiva in almeno uno dei due gruppi, si ottiene la seguente tabella di frequenza:

Aumento di peso	A	B
	Freq. Ass.	Freq. Ass.
$(-15, -10]$	4	0
$(-10, -5]$	5	1
$(-5, 0]$	6	3
$(0, 5]$	5	1
$(5, 10]$	2	5
$(10, 15]$	3	6
$(15, 20]$	1	0
$(20, 25]$	0	1
Totale	26	17

Nota . . . in generale le pazienti trattate con il farmaco (B) presentano un aumento di peso più elevato rispetto alle pazienti trattate con placebo (A).
MA: Attenzione ai totali! SONO DIVERSI!!

Frequenze relative

È più utile talvolta conoscere la proporzione di osservazioni per ciascuna modalità, o intervallo, piuttosto che il numero assoluto.

Dividendo le frequenze assolute per il numero totale n di unità statistiche, si ottengono le FREQUENZE RELATIVE, ovvero

$$\text{frequenza relativa} = \frac{\text{frequenza assoluta}}{\text{numero totale di osservazioni}} \quad \Leftrightarrow p_j = \frac{n_j}{n}$$

Le frequenze relative permettono di confrontare tabelle di frequenza di una variabile calcolate per insiemi di unità statistiche di diversa numerosità complessiva.

Nota ... Spesso si moltiplica per 100 per avere le frequenze relative percentuali.

Frequenze relative percentuali

Frequenze relative (%)

Le frequenze relative % sono la percentuale del numero di osservazioni che soddisfano una data caratteristica (che appartengono ad un certo intervallo).

Nell'ESEMPIO si ottiene la seguente tabella:

Aumento di peso	A	B
	Freq. Rel. %	Freq. Rel. %
$(-15, -10]$	15	0
$(-10, -5]$	19	6
$(-5, 0]$	23	18
$(0, 5]$	19	6
$(5, 10]$	8	29
$(10, 15]$	12	35
$(15, 20]$	4	0
$(20, 25]$	0	6
Totale	100	100

Nota ... le pazienti trattate con il farmaco (B) hanno una proporzione più elevata di osservazioni superiori a 10, mentre le pazienti trattate con placebo (A) presentano una proporzione più elevata al di sotto di questo valore.

Frequenze (assolute e relative) cumulate

Solo per variabili le cui modalità sono ordinate, si misura la frequenza con cui si presentano modalità di valore inferiore o uguale ad una certa modalità.

Si ottengono "cumulando" progressivamente le frequenze (ossia sommando le frequenze della modalità, o dell'intervallo, specificata a quelle di tutte le modalità, o intervalli, precedenti).

Possono essere assolute o relative. Ma per confrontare gruppi di numerosità diverse occorre utilizzare le frequenze relative.

Aumento di peso	A	B
	Freq. Rel. Cum. %	Freq. Rel. Cum. %
$(-15, -10]$	15	0
$(-10, -5]$	34	6
$(-5, 0]$	57	24
$(0, 5]$	76	30
$(5, 10]$	84	59
$(10, 15]$	96	94
$(15, 20]$	100	94
$(20, 25]$	100	100

Altre rappresentazioni dei dati in tabelle ...

Quando si osserva una variabile nel tempo, si ottiene una tabella che prende il nome di serie storica o temporale. In questo caso le unità statistiche sono i diversi tempi in cui si effettua l'osservazione.

Anni	Casi di malattia
1997	2207
1998	20435
1999	20692
2000	21080
2001	21514

Numero di pazienti con una data malattia in Italia dal 1997 al 2001.

Nel caso di una variabile osservata in diverse località geografiche, si parla di serie territoriale o spaziale.

Ripartizioni	Casi di malattia
Nord	11416
Centro	4513
Sud e Isole	6286

Numero di pazienti con una data malattia per ripartizione territoriale.

Grafici ben fatti

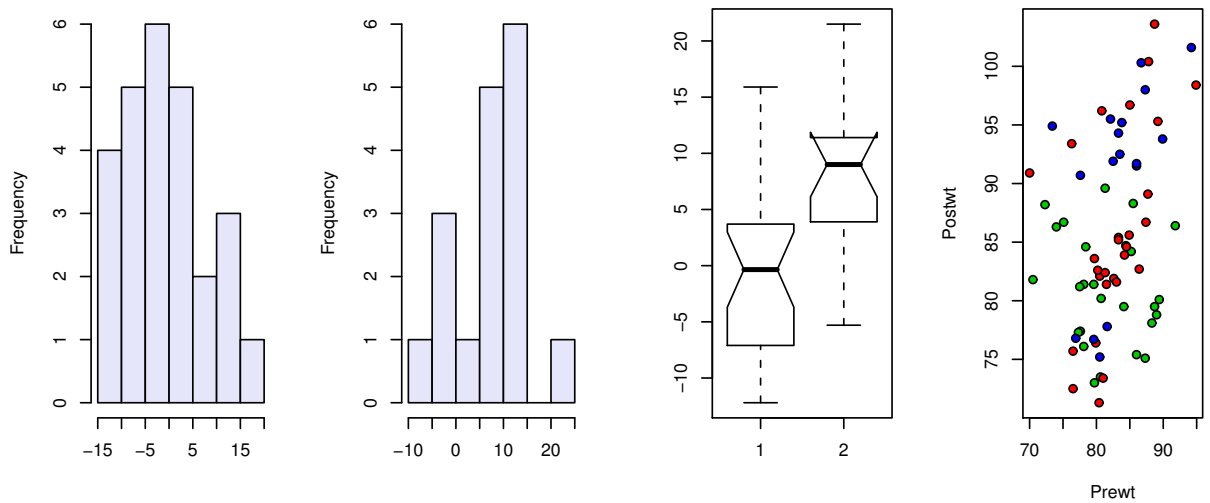
Un buon grafico permette di imparare molto su un insieme di dati con una semplice "occhiata".

Un grafico è di solito di più facile consultazione di una tabella.

Il tipo giusto di grafico dipende dal tipo di dati disponibili. Ci sono grafici diversi per variabili qualitative o quantitative, discrete o continue, serie storiche, serie geografiche ecc. Anche il numero di osservazioni è importante per scegliere il tipo di grafico.

È fondamentale l'uso del computer!

I grafici devono essere accurati, semplici, chiari, belli e ben strutturati.

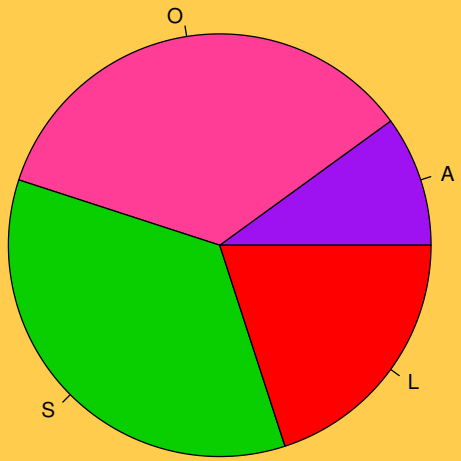



Il grafico a torta più preciso del mondo
Torta che ho mangiato
Torta che non ho ancora mangiato

Diagramma a torta

Per variabili qualitative (ordinali o meno), si possono anche usare i diagrammi a torta (o diagrammi circolari) per illustrare le frequenze relative.

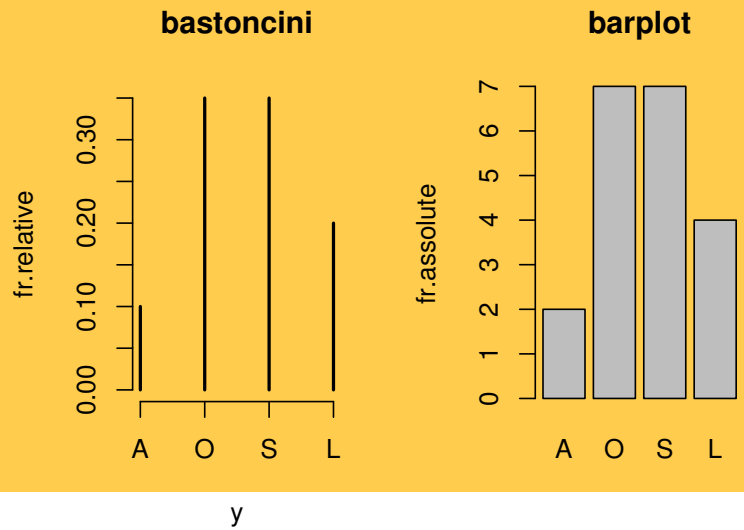
ESEMPIO: Rappresentazione della variabile grado di scolarità.
Le diverse modalità sono rappresentate da uno spicchio della torta.
L'angolo al centro è proporzionale alla frequenza relativa di quella modalità: $360^o \times p_j$.



Diagrammi a barre

Per illustrare distribuzioni di frequenza (assolute o relative) di variabili qualitative (ordinali o meno) si usano i diagrammi a barre (o a bastoncini).

ESEMPIO: Rappresentazione della variabile grado di scolarità, con A analfabeta, O scuola dell'obbligo, S diploma di scuola superiore e L laurea.



Nota . . . I rettangoli (barre) hanno base uguale e sono separati gli uni dagli altri per non indicare alcuna continuità. Il numero di barre è pari al numero di modalità da rappresentare.

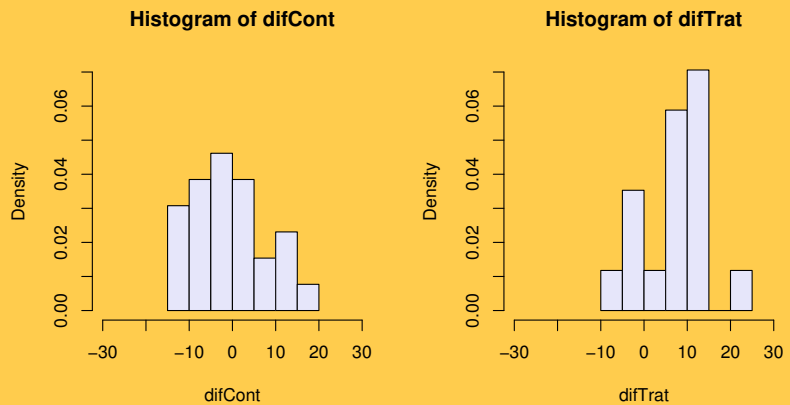
Istogramma

Per rappresentare la distribuzione di frequenza di una variabile quantitativa continua con raggruppamento in intervalli, si può usare un ISTOGRAMMA. Viene costruito ponendo:

$$\text{basi rettangoli} = \begin{pmatrix} \text{intervalli riportati nella prima} \\ \text{colonna della tabella di frequenza} \end{pmatrix}$$

$$\text{altezze rettangoli} = \text{frequenze assolute (o relative) se le basi sono uguali}$$

ESEMPIO: Istogrammi dell'aumento di peso nei due gruppi di pazienti



Nota ... Il grafico suggerisce le stesse considerazioni fatte sulla base della tabella di frequenza: la distribuzione delle pazienti trattate con il farmaco (tratt. B) è, rispetto a quella delle pazienti trattate con il placebo (tratt. A), più spostata verso destra.

Osservazioni su istogrammi

- Numero di intervalli: nella costruzione di un istogramma esiste un elemento di arbitrarietà: la scelta di quanti e quali intervalli utilizzare. È necessario fare un po' di attenzione.
- Suggerimenti: quasi sempre è conveniente fare più di un grafico. Si provano differenti lunghezze per gli intervalli e poi si sceglie.
- Il numero di intervalli deve dipendere dal numero di dati. Ripartire 1000 osservazioni in 40 intervalli può anche dare risultati sensati, ma usare gli stessi 40 intervalli per 20 dati non è appropriato.
- Se le basi hanno ampiezze diverse, in corrispondenza di ciascun intervallo si ha un rettangolo la cui area è proporzionale alla frequenza corrispondente all'intervallo stesso, ovvero area rettangolo $=$ base $\times$ altezza $=$ frequenza.

Poligono di frequenza – Grafico a linee

È un grafico molto simile all'istogramma e usa i suoi stessi assi.

È costruito congiungendo con segmenti i punti centrali dei lati superiori dei rettangoli che definiscono l'istogramma.
In genere, si aggiungono due classi terminali con frequenza zero e ampiezza pari a quella della classe adiacente. In questo modo, la somma delle aree dei rettangoli è uguale all'area sottesa al poligono di frequenza, se le classi sono di egual lunghezza.

ESEMPIO: Poligoni di frequenza dell'aumento di peso nei due gruppi di pazienti.

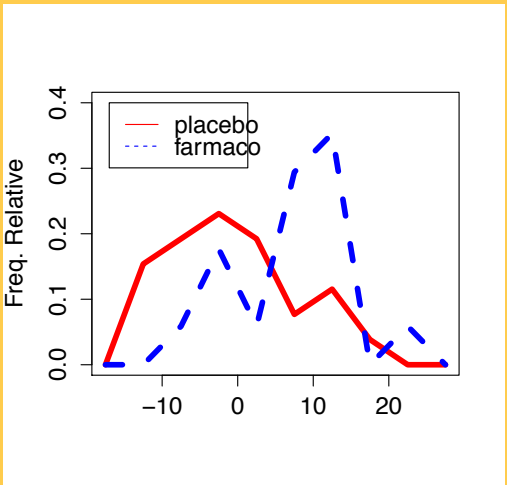


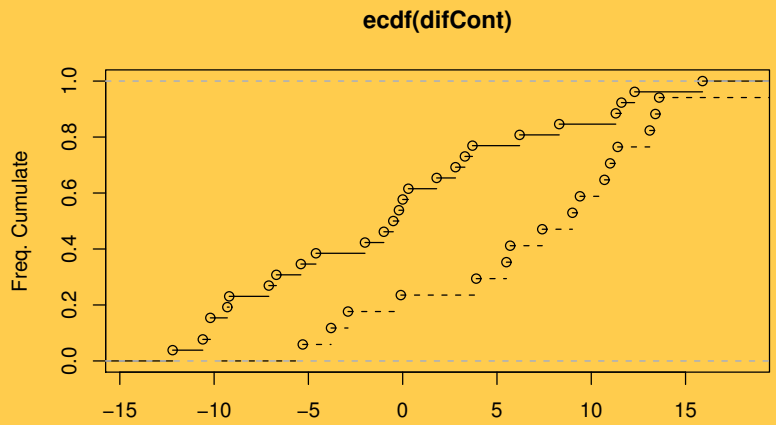
Grafico delle frequenze relative cumulate

Spesso può essere utile rappresentare graficamente le frequenze relative cumulate, ossia la proporzione di unità che presentano valore della variabile minore o uguale a una certa modalità.

$$\left(\begin{array}{c}\text{frequenza relativa}\\ \text{cumulata in } x\end{array}\right) = \frac{\left(\begin{array}{c}\text{numero di osservazioni}\\ \text{minori o uguali a } x\end{array}\right)}{\left(\begin{array}{c}\text{numero totale di}\\ \text{osservazioni}\end{array}\right)}$$

È anche detta funzione di ripartizione empirica.

ESEMPIO: Frequenze relative cumulate dell'aumento di peso nei due gruppi di pazienti (sui dati grezzi)



Esercizi

(1) Su un campione di $n = 20$ individui si sono ottenute le seguenti osservazioni per le variabili *età*, *sesso* e *numero di visite mediche all'anno*.

Unità	Età	Sesso	N.Vis.	Unità	Età	Sesso	N.Vis.
1	35	m	1	11	33	m	2
2	37	m	2	12	46	f	4
3	59	f	1	13	41	f	3
4	54	m	0	14	53	m	1
5	44	f	2	15	38	f	1
6	38	m	1	16	55	m	2
7	62	f	1	17	50	m	3
8	71	f	3	18	63	m	0
9	56	m	3	19	35	f	1
10	60	m	2	20	51	m	2

Si costruiscano le distribuzioni di frequenza per le tre variabili. Si costruisca la distribuzione di frequenza relativa e cumulata per l'età divisa in classi. Si rappresentino le distribuzioni di frequenza con i grafici ritenuti più idonei.

(2) La tabella di frequenza che segue riporta i voti ottenuti da una classe di studenti universitari al termine di un corso. Per il voto d'esame, si ottenga la distribuzione delle frequenze relative; si proponga una opportuna suddivisione in classi e la distribuzione di frequenza associata; l'istogramma o altri grafici utili.

Voto	18	19	20	21	22	23	24	25	26	27	28	29	30	Tot
Freq	7	2	5	1	3	2	12	1	8	4	6	1	5	57

Misure di sintesi

Misure di sintesi

Fin qui si sono studiati le tabelle e i grafici come metodi per organizzare e sintetizzare visivamente i dati.

Questi metodi non permettono, tuttavia, di formulare affermazioni sintetiche che caratterizzino una distribuzione nel suo insieme e che ne evidenzino caratteristiche essenziali.

Per variabili quantitative, è utile disporre di misure numeriche di sintesi.

Obiettivo di tali misure è:
– Descrivere sinteticamente caratteristiche di un insieme di dati;
– Mettere in evidenza le particolarità di una distribuzione di frequenza.

Tutti avete sentito parlare di una "media" (come, ad esempio, il voto medio alla maturità di una classe di liceo). Ma che cosa indica esattamente questa media? E basta da sola a descrivere un insieme di dati?

Indici di posizione (centrale)

Una caratteristica comunemente riportata per in un insieme di dati è il suo 'centro', ossia un singolo valore che si può ritenere 'centrale' rispetto alla distribuzione di frequenza.

Un indice di posizione è un valore della variabile tipico per l'insieme di dati osservato. Sarà una modalità osservata per variabili qualitative e un valore compreso tra il più piccolo e il più grande valore osservato per le variabili quantitative.
Gli indici di posizione permettono anche di confrontare distribuzioni differenti.

Alcuni indici di posizione sono:

- Media aritmetica (o valor medio o semplicemente media)
- Mediana
- Moda
- Quantili (percentili)

La media aritmetica

L'indice di posizione più comunemente usato per variabili quantitative è la media aritmetica.

Indichiamo con $x_1, x_2, \ldots, x_n$ i valori osservati su n unità di una variabile quantitativa X di interesse.

La media si calcola sommando i valori e dividendo per il numero totale delle osservazioni. In formule,

$$m_x = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Il simbolo Σ è il simbolo di sommatoria e $\sum_{i=1}^{n} x_i$ si legge somma delle x_i per i che varia da 1 a n.

Quindi $\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n$.

La media aritmetica è sempre compresa tra il più piccolo e il più grande dei valori osservati.

Esempio

Su foglie di $n = 13$ diverse specie di felci è stato misurato il tempo di insorgenza di una particolare muffa.

Le misurazioni (espresse in giorni) sono:

$$0.10, 0.25, 0.50, 4, 12, 12, 24, 24, 31, 36, 42, 55, 96$$

La media aritmetica è:

$$\begin{aligned} m_x &= \frac{(0.10 + 0.25 + 0.50 + 4 + 12 + 12 + 24 + 24 + 31 + 36 + 42 + 55 + 96)}{13} \\ &= 25.9 \text{ giorni} \end{aligned}$$

Togliendo l'ultima osservazione si trova:

$$\begin{aligned} m_x &= \frac{(0.10 + 0.25 + 0.50 + 4 + 12 + 12 + 24 + 24 + 31 + 36 + 42 + 55)}{12} \\ &= 20.1 \text{ giorni} \end{aligned}$$

Osservazione

Quando una osservazione è molto grande (o molto piccola) la media può cambiare notevolmente.

DIFETTO: la media è estremamente sensibile ai valori anomali (o insoliti o estremi). Il valore 96 potrebbe essere una misurazione errata. Ma spesso l'errore non è così evidente, o l'osservazione anomala potrebbe non essere un errore.

Esistono misure di sintesi che non sono così sensibili alle singole osservazioni.

INOLTRE: la media può essere calcolata per variabili quantitative (continue o discrete), ma non per variabili categoriali nominali o ordinali. Un'eccezione a questa regola si ha con le variabili dicotomiche le cui modalità siano codificate 0 e 1. Ad esempio se nello studio ci sono 8 foglie verdi (valore 1) e 5 foglie gialle (valore 0), la media (interpretabile come una proporzione) è

$$m_x = \frac{(1+1+1+\ldots+0+0+\ldots+0)}{13} = \frac{8}{13} = 0.615$$

ossia il 61.5% delle foglie è verde.

La media aritmetica ponderata

Supponiamo di disporre di una tabella di frequenza per una variabile quantitativa (discreta)

X	Freq.
x_1	n_1
x_2	n_2
$\dots$	$\dots$
x_j	n_j
$\dots$	$\dots$
x_k	n_k
Totale	n

La media può allora essere calcolata come

$$m_x = \frac{x_1 n_1 + \ldots + x_k n_k}{n_1 + \ldots + n_k} = \frac{\sum_{j=1}^{k} x_j n_j}{\sum_{j=1}^{k} n_j}$$

NOTA: Poiché $\sum_{j=1}^{k} n_j = n$, si ha

$$m_x = \frac{1}{n} \sum_{j=1}^{k} x_j n_j = \sum_{j=1}^{k} x_j \frac{n_j}{n} = \sum_{j=1}^{k} x_j p_j$$

E se i dati sono raggruppati in classi? Esempio

Con una distribuzione di frequenza per una variabile quantitativa continua con raggruppamento in classi, si usa la stessa formula con x_j valore centrale del j-esimo intervallo, n_j frequenza assoluta del j-esimo intervallo e k numero di intervalli.

Aumento di peso	tratt. A Freq. Ass.	tratt. B Freq. Ass.
$(-15, -10]$	4	0
$(-10, -5]$	5	1
$(-5, 0]$	6	3
$(0, 5]$	5	1
$(5, 10]$	2	5
$(10, 15]$	3	6
$(15, 20]$	1	0
$(20, 25]$	0	1
Totale	26	17

tratt. A: $m_x = \frac{(-12.5)\times 4+(-7.5)\times 5+...+17.5\times 1}{26} = -0.76$ libbre (-0.45 sui dati grezzi)

tratt. B: $m_x = \frac{(-7.5)\times 1+(-2.5)\times 3...+22.5\times 1}{17} = 7.21$ libbre (7.26 sui dati grezzi)

1. PROPRIETÀ DI BARICENTRO: la somma degli scarti dei dati dalla propria media m_x è zero:

$$\sum_{i=1}^{n}(x_i - m_x) = 0 \ ,$$

o anche, introducendo le frequenze,

$$\sum_{j=1}^{k}(x_j - m_x)n_j = 0 \ .$$

Dimostrazione.

$$\begin{aligned}\sum_{i=1}^{n}(x_i - m_x) &= (x_1 - m_x) + \ldots + (x_n - m_x) \\ &= (x_1 + \ldots + x_n) - (m_x + \ldots + m_x) = nm_x - nm_x = 0 \ .\end{aligned}$$

2. MINIMI QUADRATI: la somma degli scarti al quadrato dei dati da un valore C è minima se $C = m_x$:

$$\sum_{i=1}^{n}(x_i - C)^2 \geq \sum_{i=1}^{n}(x_i - m_x)^2 \ ,$$

o anche, introducendo le frequenze,

$$\sum_{j=1}^{k}(x_j - C)^2 n_j \geq \sum_{j=1}^{k}(x_j - m_x)^2 n_j \ .$$

Dimostrazione.

$$\begin{aligned}\sum_{i=1}^{n}(x_i - C)^2 &= \sum_{i=1}^{n}[(x_i - m_x) + (m_x - C)]^2 \\ &= \sum_{i=1}^{n}[(x_i - m_x)^2 + (m_x - C)^2 + 2(x_i - m_x)(m_x - C)]\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{n}(x_i - m_x)^2 + n(m_x - C)^2 + 2(m_x - C)\sum_{i=1}^{n}(x_i - m_x) \\
&= \sum_{i=1}^{n}(x_i - m_x)^2 + n(m_x - C)^2 + 0
\end{aligned}
$$

infatti, per la proprietà di baricentro, $\sum_{i=1}^{n}(x_i - m_x) = 0$.

La quantità $n(m_x - C)^2$ è maggiore o uguale a zero (zero se $C = m_x$).

Quindi, $\sum_{i=1}^{n}(x_i - C)^2$ è uguale a $\sum_{i=1}^{n}(x_i - m_x)^2$ più una quantità non negativa e questo mostra la proprietà.

Interpretazione: Se sostituiamo ai dati un singolo valore C, possiamo attribuire a ogni sostituzione di x_i con C la 'perdita' $(x_i - C)^2$. La perdita totale è dunque $\sum_{i=1}^{n}(x_i - C)^2$ ed è la più piccola possibile prendendo $C = m_x$.
In questo senso, m_x è la 'miglior' sostituzione dei dati $x_1, \ldots, x_n$ con un singolo indice di sintesi.

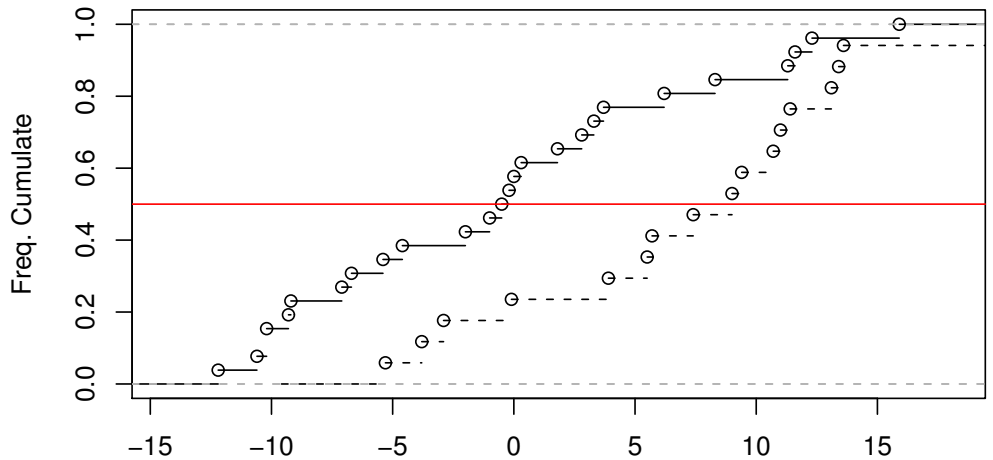
La mediana

La mediana è definita come l'osservazione che occupa la posizione centrale nella sequenza dei valori di una variabile ordinati in senso crescente.

Basta che la variabile sia ordinabile per calcolarla.

La metà, o poco più, dei valori osservati sarà dunque maggiore o uguale alla mediana, mentre metà, o poco più, sarà minore o uguale ad essa.

DEF: Se un insieme di dati contiene n osservazioni (n dispari), la mediana è il valore centrale: la misurazione corrispondente alla posizione $(n+1)/2$. Se n è pari, la mediana è definita come la media dei due valori centrali, l'osservazione di posto $(n/2)$ e quella di posto $(n/2)+1$.



Esempio

Su foglie di $n = 13$ diverse specie di felci è stato misurato il tempo di insorgenza di una particolare muffa (dati già ordinati in senso crescente):

$$0.10, 0.25, 0.50, 4, 12, 12, 24, 24, 31, 36, 42, 55, 96$$

La mediana è:

$$(13+1)/2 = 7 \quad \Rightarrow \quad 7^a \text{ osservazione} \quad \Rightarrow \quad 24 \text{ giorni}$$

Togliendo l'ultima osservazione si trova:

$$12/2 = 6 \quad \Rightarrow \quad 6^a \text{ osservazione} \quad \Rightarrow \quad 12 \text{ giorni}$$
$$12/2 + 1 = 7 \quad \Rightarrow \quad 7^a \text{ osservazione} \quad \Rightarrow \quad 24 \text{ giorni}$$

$$\Rightarrow \frac{12 + 24}{2} = 18 \text{ giorni}$$

La mediana è meno sensibile (più ROBUSTA) della media aritmetica alla presenza di osservazioni anomale.

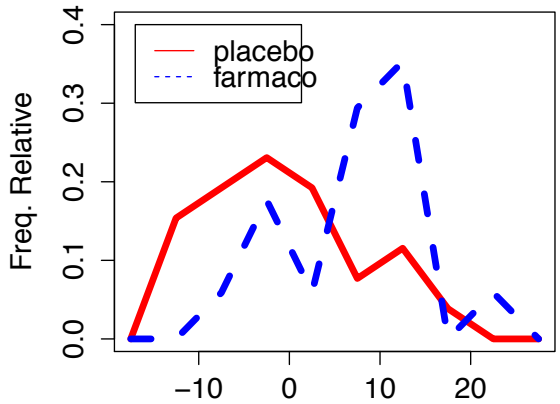
La moda

La moda è definita come la modalità con la frequenza più alta, ossia è l'osservazione che si verifica con maggiore frequenza.

Il concetto di moda è quindi molto semplice ed è analogo a quello usato correntemente nel linguaggio di tutti i giorni.

Proprietà: La moda può essere calcolata per qualsiasi distribuzione di frequenza, ossia sia per variabili qualitative che quantitative.

Può capitare che una distribuzione di frequenza NON presenti una sola moda, ma più di una. In questo caso la distribuzione è bimodale o multimodale.



Quantili (percentili)

I quantili generalizzano la mediana:

DEF: Un quantile-p, con $p \in [0, 1]$, è un valore la lascia alla sua sinistra almeno il $100p\%$ dei dati osservati e alla sua destra almeno il restante $100(1 - p)\%$.

Ad esempio, il quantile-0.1 è quel valore che lascia a sinistra almeno il 10% delle osservazioni e a destra almeno il 90%.

I quantili con p uguale a 0.25, 0.50 e 0.75 vengono chiamati rispettivamente il primo, il secondo e il terzo quartile. Essi dividono la popolazione in quattro parti uguali. Il secondo quartile coincide con la mediana.

Esempio

Su $n = 13$ foglie di differenti felci è stato misurato il tempo di insorgenza di una particolare muffa:

$$0.10, 0.25, 0.50, 4, 12, 12, 24, 24, 31, 36, 42, 55, 96$$

La tabella delle frequenze relative cumulate è

0.1	0.25	0.5	4	12	24	31	36	42	55	96
0.08	0.15	0.23	0.31	0.46	0. 62	0.69	0.77	0.85	0.92	1.00

Il valore 4 è il quantile 0.25

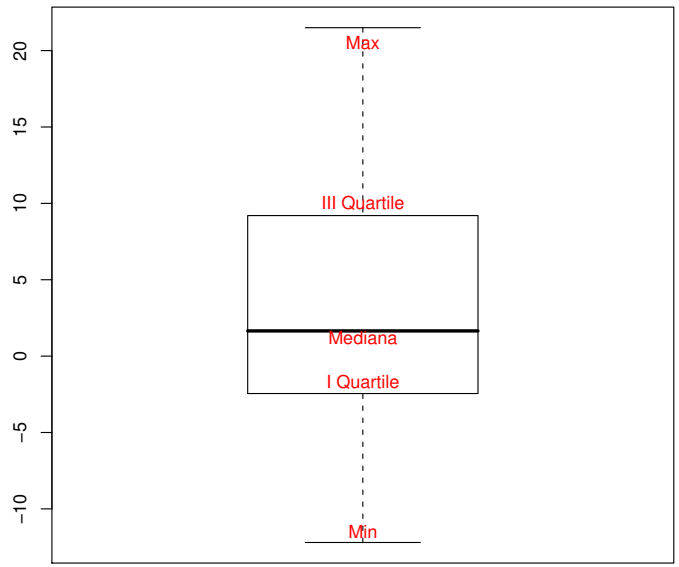
e il valore 36 è il quantile 0.75.

La mediana è 24 giorni.

Diagramma a scatola e baffi

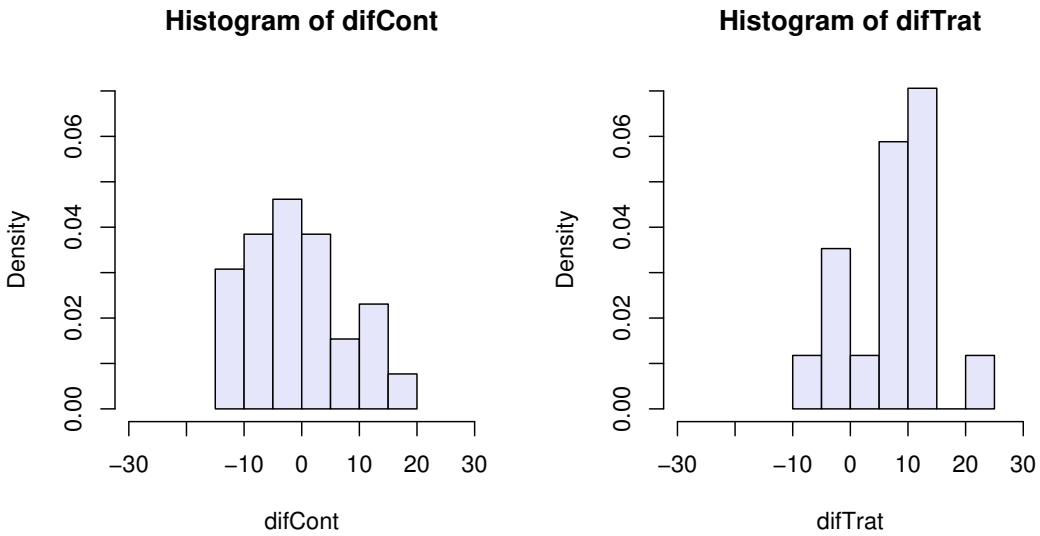
Il nome deriva dall'inglese (*box and whiskers plot*). Anche in italiano è spesso abbreviato in *boxplot*.

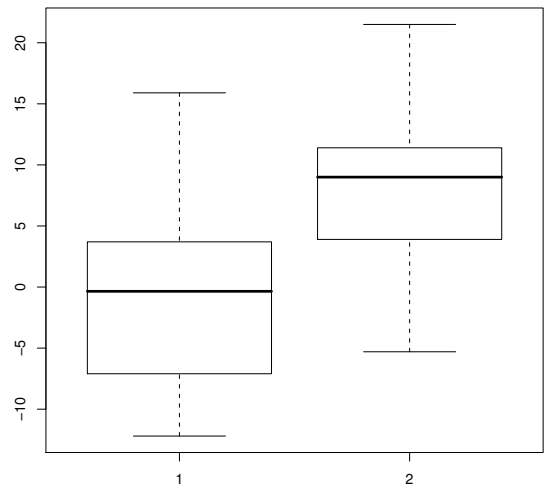
Esso fornisce un'idea schematica di un insieme di dati basata sui quantili. È costituito, come dice il nome, da una scatola e da due baffi costruiti come segue:



Gli estremi dei baffi sono il minimo e il massimo.
Gli estremi della scatola sono il primo e il terzo quartile.
La linea interna alla scatola è la mediana.

Esempio


Histogram of difCont
Density
0.00
0.02
0.04
0.06
−30
−10
0
10
20
30
difCont
Histogram of difTrat
Density
0.00
0.02
0.04
0.06
−30
−10
0
10
20
30
difTrat


−10
−5
0
5
10
15
20
1
2

Esercizi

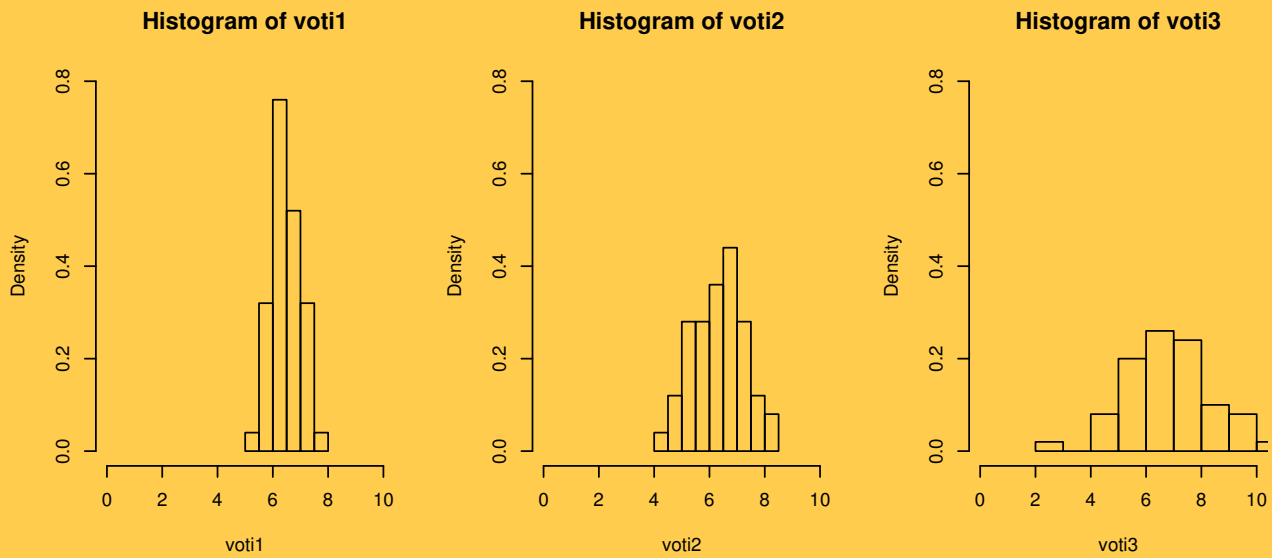
- L'ampiezza dell'apertura alare di 8 farfalle (in mm) è pari a: 8, 17, 10, 11, 13, 15, 15, 5 .
  – Si dia una rappresentazione grafica dei dati tramite un diagramma con scatola e baffi.
  – Si calcoli una misura di posizione dell'ampiezza dell'apertura alare.
- Per 200 laureati in Scienze Naturali, 100 che avevano svolto un tirocinio presso una struttura privata e 100 presso una struttura pubblica, è stato rilevato il tempo dalla laurea al primo lavoro. I risultati sono stati raccolti nella seguente tabella:

	dalla laurea al primo lavoro		
tirocinio	meno di 1 mese	tra 1 e 6 mesi	più di 6 mesi
privato	53	45	2
pubblico	43	51	6

  – Una rappresentazione grafica dei dati opportuna può essere basata su un istogramma? una torta? un diagramma a barre?
  – Calcolare, se possibile, il tempo medio dalla laurea al primo lavoro.
  – Si calcoli la mediana del tempo dalla laurea al primo lavoro per i due gruppi corrispondenti ai due tipi di tirocinio.

Una misura di posizione da sola può essere sufficiente per descrivere la distribuzione di un insieme di dati? Certamente no!

ESEMPIO: Confrontiamo le distribuzioni dei voti di tre diversi professori di matematica, il cui voto medio è però lo stesso (6.5).



Misure di variabilità

DEF: La variabilità di una distribuzione esprime la tendenza delle unità ad assumere modalità diverse del carattere.

Si è detto fin dall'inizio che la variabilità costituisce la ragion d'essere della Statistica.

Per misurare la variabilità si usano INDICI che sintetizzano la diversità tra ogni modalità osservata ed un indice di posizione, oppure tra due particolari valori della distribuzione.

Un indice di variabilità è un valore non negativo che:
– vale zero se e solo se tutte le unità presentano uguale modalità del carattere;
– aumenta all'aumentare della 'diversità' tra le modalità assunte dalle varie unità.

Alcuni indici di variabilità per variabili quantitative sono:

- Il campo di variazione
- Lo scarto interquartile
- La varianza (e lo scarto quadratico medio)

Campo di variazione – *Range*

DEF: Il campo di variazione (range) di un insieme di dati è definito come la differenza tra l'osservazione più grande e quella più piccola:

$$\text{Campo di variazione} = \max(x_1, \ldots, x_n) - \min(x_1, \ldots, x_n)$$

È veloce da calcolare, ma la sua utilità è limitata. Considera solo i valori estremi dei dati e non tutte le osservazioni $\rightarrow$ è molto sensibile ai possibili valori anomali.

ESEMPIO: Per le variabili aumento di peso nei due gruppi di pazienti si ha:

$$\text{Campo di variazione Tratt. A} = 15.90 - (-12.20) = 28.1 \text{ (libbre)}$$

$$\text{Campo di variazione Tratt. B} = 21.50 - (-5.3) = 26.8 \text{ (libbre)}$$

I due valori indicano una variabilità simile nei due gruppi di pazienti.

Scarto interquartile

DEF: Lo scarto interquartile di un insieme di misurazioni è calcolato sottraendo il quantile-0.25 dal quantile-0.75 (ossia il primo quartile dal terzo quartile) e comprende, pertanto, il 50% delle osservazioni:

$$\text{Scarto interquartile} = \text{quantile-0.75 - quantile-0.25}$$

È più robusto in presenza di osservazioni estreme. Per questo viene usato quando si sospetta la presenza di osservazioni anomale. Nel boxplot corrisponde all'ampiezza della scatola.

ESEMPIO: Per le variabili aumento di peso nei due gruppi di pazienti si ha:

$$\text{Scarto interquartile Tratt. A} = 3.60 - (-7) = 10.6$$

$$\text{Scarto interquartile Tratt. B} = 11.40 - (3.9) = 7.5$$

I due valori indicano una variabilità leggermente inferiore nel gruppo delle pazienti trattate con il farmaco.

Varianza e Scarto quadratico medio

DEF: La misura di variabilità più usata è la VARIANZA:

$$S_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m_x)^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - m_x^2$$

ossia "varianza = media dei quadrati degli scarti dalla media", oppure "varianza = media dei quadrati - quadrato della media".

La varianza misura di quanto i dati sono "distanti" dalla media aritmetica (la distanza è valutata usando i quadrati delle differenze).
Si osservi che l'unità di misura della varianza è pari al quadrato dell'unità di misura dei dati originari $\rightarrow$ la radice quadrata della varianza è chiamata scarto quadratico medio (o deviazione standard): $S_x = \sqrt{(S_x^2)}$. La sua unità di misura coincide con l'unità di misura dei dati.

ESEMPIO: Per le variabili aumento di peso nei due gruppi di pazienti (lbs=libbre):

Varianza Tratt. A $= 61.36$ lbs^2 Scarto quadratico medio Tratt. A $= 7.83$ lbs

Varianza Tratt. B $= 48.21$ lbs^2 Scarto quadratico medio Tratt. B $= 6.94$ lbs

Esempi di calcolo

- **Dati:** Numero di incidenti stradali in 4 tratti di autostrada: 1, 3, 2, 5
  Media: $m_x = (1 + 3 + 2 + 5)/4 = 2.75$
  Media dei quadrati: $(12 + 32 + 22 + 52)/4 = 9.75$
  Varianza: $S_x^2 = [(1 - 2.75)^2 + (3 - 2.75)^2 + (2 - 2.75)^2 + (5 - 2.75)^2]/4 = 2.19$
  Varianza: $S_x^2 = 9.75 - 2.75^2 = 2.19$
- **Dati:** Numero di imprese (in migliaia) nel 1991 in 5 regioni italiane: 268, 106, 76, 238, 88
  Media: $m_x = (268 + 106 + 76 + 238 + 88)/5 = 155.2$
  Media dei quadrati: $(2682 + 1062 + 762 + 2382 + 882)/5 = 30644.8$
  Varianza: $S_x^2 = 30644.8 - 155.22 = 6557.76$
  Scarto quadratico medio: $S_x = 6557.76^{1/2} = 80.98$

Esempio

Studio condotto per esaminare il tempo di insorgenza (in giorni) di una particolare muffa su foglie di $n = 13$ diverse piante di felce:

$$0.10, 0.25, 0.50, 4, 12, 12, 24, 24, 31, 36, 42, 55, 96$$

Riassunto della variabile:

Min.	1st Qu.	Mediana	Media	3rd Qu.	Max.
0.10	4.00	24.00	25.91	36.00	96.00

Il campo di variazione è: max - min = 96-0.10 = 95.9.
Dipende molto dal valore massimo.

Lo scarto interquartile è: quantile-0.75 - quantile-0.25 = 36 - 4 = 32.

La varianza è: $S_x^2 = 691.54$.

Lo scarto quadratico medio è: $S_x = 26.29$.

Standardizzazione dei dati

A volte è utile trasformare un insieme di dati $x_1, \ldots, x_n$ in maniera tale che i dati trasformati, detti $z_1, \ldots, z_n$, abbiano media nulla e varianza unitaria.

DEF: La trasformazione appropriata è:

$$z_i = \frac{x_i - m_x}{S_x} \qquad i = 1, \ldots, n$$

I dati così trasformati sono detti standardizzati e il nuovo insieme di dati è tale che:

- media$(z_1, \ldots, z_n) = m_z = 0$
- varianza$(z_1, \ldots, z_n) = S_z^2 = 1$

Dallo studio di dati standardizzati possono emergere altre caratteristiche delle distribuzioni.

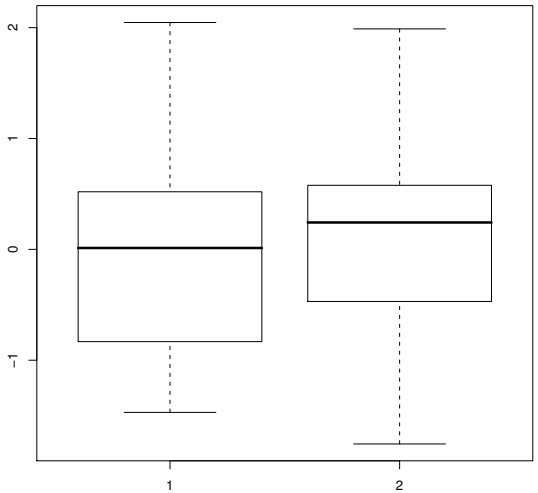
Esempio

Consideriamo due insieme di dati standardizzati, perciò almeno approssimativamente omogenei per posizione e variabilità.

Nonostante la stessa media e varianza, le due distribuzioni sono diverse.

La seconda è più o meno simmetrica rispetto allo zero. La prima invece ha la coda verso i valori alti molto più lunga della coda verso i valori bassi. Si parla in questo caso di asimmetria positiva (nel caso opposto si parla di asimmetrica negativa).

Per una distribuzione di frequenza unimodale e simmetrica si ha:
Media $\cong$ Mediana $\cong$ Moda



Esercizi

(1) L'ampiezza dell'apertura alare di 8 farfalle (in mm) è pari a: 8, 17, 10, 11, 13, 15, 15, 5. Una misura di variabilità dell'ampiezza dell'apertura alare è 12 mm? 4.02 mm? 16.21 mm^2?

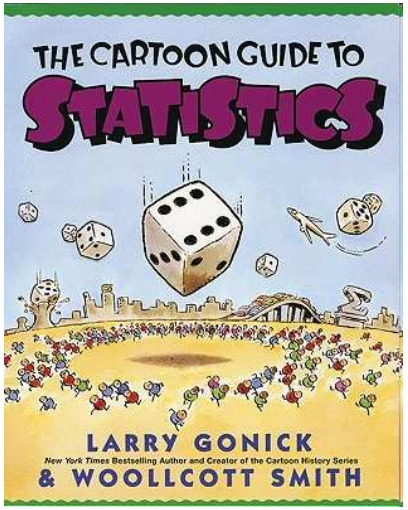
(2) Sono stati rilevati i pesi (in kg) di 10 maschi (M) e 10 femmine (F) di una specie di pesce:
M: 1.2, 3, 5.2, 4, 3.5, 4.3, 3.3, 4.8, 3.8, 3.2
F: 1.3, 2.2, 1.5, 2.3, 1.8, 1.7, 2.1, 2, 1.9, 2.1
Si calcolino per i maschi e le femmine lo scarto quadratico medio, lo scarto interquartile e il campo di variazione. Si costruiscano i due boxplot e commentando i risultati.

(3) In una foresta la foglia di un certo tipo di felce ha una lunghezza media pari a 13 cm e media dei quadrati pari a 173 cm^2. Calcolare lo scarto quadratico medio della lunghezza della foglia. Si sa poi che un altro tipo di felce ha le foglie tre volte più lunghe: si calcolino media e scarto quadratico medio per questo tipo di felce.

Alcuni riferimenti bibliografici

- Agresti, A., Finlay, B. (2009). *Statistica per le scienze sociali*. Pearson.
- Agresti, A., Franklin, C. (2016). *Statistica: l'arte e la scienza d'imparare dai dati*. Pearson.
- Diamond, I., Jefferies, J. (2001). *Introduzione alla statistica per le scienze sociali*. McGraw-Hill.
- Rosenthal, J.S. (2005). *Le Regole del Caso: Istruzioni per l'Uso*. Longanesi.
- Ventura, L., Racugno, W. (2017). *Biostatistica. Casi di Studio in* R, Egea, Milano.

Oppure...


THE CARTOON GUIDE TO
STATISTICS
LARRY GONICK
New York Times Bestselling Author and Creator of the Cartoon History Series
& WOOLLCOTT SMITH


I MANGA DELLE SCIENZE
STATISTICA
SHIN TAKAHASHI
TREND-PRO, CO., LTD.
la Repubblica Le Scienze
I MANGA DELLE SCIENZE
REGRESSIONE
ANALISI STATISTICA DEI DATI
SHIN TAKAHASHI
IROHA INOUE
TREND-PRO CO., LTD.
la Repubblica