



Alessandra Nardi

Alcuni studi
sperimentali

Iniziamo a
descrivere i dati:
la distribuzione
di frequenze

Rappresentazioni
grafiche

Sintetizziamo i
dati: le medie

Le medie non
basta

Sulla descrizione dei dati sperimentali

Alessandra Nardi

Dipartimento di Matematica

Università di Roma "Tor Vergata"

4 aprile 2022



Outline

Alessandra Nardi

Alcuni studi
sperimentali

Iniziamo a
descrivere i dati:
la distribuzione
di frequenze

Rappresentazioni
grafiche

Sintetizziamo i
dati: le medie

Le media non
basta

- ① Alcuni studi sperimentali
- ② Iniziamo a descrivere i dati: la distribuzione di frequenze
- ③ Rappresentazioni grafiche
- ④ Sintetizziamo i dati: le medie
- ⑤ Le media non basta



Alessandra Nardi

Alcuni studi sperimentali

Iniziamo a descrivere i dati: la distribuzione di frequenze

Rappresentazioni grafiche

Sintetizziamo i dati: le medie

Le media non basta

Statistica descrittiva ed inferenziale: un confine incerto

Vediamo alcuni studi sperimentali che ci accompagneranno nel prosieguo



Alessandra Nardi

Alcuni studi sperimentali

Iniziamo a descrivere i dati: la distribuzione di frequenze

Rappresentazioni grafiche

Sintetizziamo i dati: le medie

Le media non basta

Il primo è uno studio sperimentale che ha coinvolto 67 volontari che non presentavano patologie e non erano sottoposti a terapie in grado di influire sui livelli dell'ormone della crescita.

Per ciascun soggetto sono stati rilevati i livelli di concentrazione dell'ormone nel sangue prima della stimolazione e quindi a distanza di 60, 90 e 120 minuti dalla somministrazione delle sostanze stimolanti. Obiettivo dello studio era stimare i livelli attesi dell'ormone della crescita prima e dopo stimolazione in individui adulti sani.

La definizione di un intervallo di valori *normali* è infatti il primo passo per poter individuare casi di carenza.



Alessandra Nardi

Alcuni studi sperimentali

Iniziamo a descrivere i dati: la distribuzione di frequenze

Rappresentazioni grafiche

Sintetizziamo i dati: le medie

Le media non basta

Per comprendere la struttura del database osserviamo i primi dieci individui del nostro studio:

ID	BMI	SEX	AGE	GH1	GH2	GH3	GH4	GHMAX
E09963	40.2	m	31	0.01	2.55	1.7	1.2	2.55
E09327	24.7	m	22	0.03	11.5	12.2	3.3	12.2
E09670	37.3	m	42	0.03	2.1	2.8	0.9	2.8
E09822	35.4	f	32	0.03	2.3	4	1.6	4
E10388	25.1	m	43	0.03	19	8.1	1.2	19
E08979	30.8	f	33	0.04	19.3	18.6	8.8	19.3
E09355	26.5	m	28	0.04	17.6	14.3	1.3	17.6
E09536	29.5	m	25	0.04	5	6.4	1.2	6.4
E09047	30.8	f	37	0.05	13.7	9.2	1.95	13.7
E09453	34.3	f	52	0.06	2.5	1	0.1	2.5



Alessandra Nardi

Alcuni studi sperimentali

Iniziamo a descrivere i dati: la distribuzione di frequenze

Rappresentazioni grafiche

Sintetizziamo i dati: le medie

Le media non basta

Il secondo è uno studio retrospettivo su pazienti affetti da Leucemia Mieloide Cronica e sottoposti a trapianto allogenico di midollo osseo. Analizziamo la casistica relativa a 181 pazienti sottoposti a trapianto in centri italiani dal marzo 1979 al novembre 1988.

Obiettivi:

- comprendere in generale l'andamento della terapia trapiantologica sia in termini di sopravvivenza che di assenza di recidive della malattia;
- generare delle ipotesi circa possibili fattori di rischio o protettivi.



Alessandra Nardi

Alcuni studi sperimentali

Iniziamo a
descrivere i dati:
la distribuzione
di frequenze

Rappresentazioni
grafiche

Sintetizziamo i
dati: le medie

Le medie non
basta

Il modo in cui descriveremo i nostri dati dipende dalla natura delle variabili che abbiamo osservato.

La distinzione fondamentale è tra variabili *qualitative* (categorical data) e *quantitative*.

Nel primo caso osserviamo l'appartenenza o meno ad una certa categoria (ad esempio maschio o femmina) mentre nel secondo i nostri dati sono numeri reali.

Questo determina il livello di misurabilità delle nostre osservazioni.



Alessandra Nardi

Alcuni studi
sperimentali

Iniziamo a
descrivere i dati:
la distribuzione
di frequenze

Rappresentazioni
grafiche

Sintetizziamo i
dati: le medie

Le media non
basta

Nel caso di caratteri qualitativi parliamo di:

- *scala nominale* quando, date due osservazioni, siamo in grado di dire soltanto se, rispetto al carattere in esame, sono uguali o diverse, si tratta del livello più basso di misurabilità e l'analisi statistica lavorerà prevalentemente sulle frequenze;
- *scala ordinale* quando, date due osservazioni, possiamo stabilire quale precede l'altra siamo cioè in grado di definire un ordinamento.

Nel caso di caratteri quantitativi parliamo di:

- *scala ad intervallo* quando possiamo confrontare due osservazioni facendo la differenza tra le modalità del carattere che assumono, si tratta di casi molto rari in cui la mancanza di uno zero assoluto non consente di calcolare rapporti;
- *scala di rapporto* quando possiamo confrontare sia per differenza sia utilizzando dei rapporti, è il livello più alto di misurabilità consente di definire una unità di misura e una distanza.



Alessandra Nardi

Alcuni studi
sperimentali

Iniziamo a
descrivere i dati:
la distribuzione
di frequenze

Rappresentazioni
grafiche

Sintetizziamo i
dati: le medie

Le media non
basta

Sintetizziamo i dati osservati al fine di interpretarli ricordando sempre che **ogni sintesi corrisponde una perdita di informazione**.

Tornando allo studio sull'orname della crescita, rispetto alla variabile Sesso, abbiamo osservato sul nostro insieme di 67 individui le seguenti modalità

(m m m f m f m m f f m m f f f m f f f m m m m m f f m m f m f f m m
m f m f m f m m f m f m m m m m m m f m m m m f m m f f f f m m m)

che, sintetizzate in una distribuzione di frequenza, si presentano come

Modalità	Frequenze assolute	Frequenze relative
Maschi	41	0.61
Femmine	26	0.39
Totale	67	1

È questo il caso di una variabile qualitativa nominale ed il passaggio alle frequenze è obbligato.



Alessandra Nardi

Alcuni studi
sperimentali

Iniziamo a
descrivere i dati:
la distribuzione
di frequenze

Rappresentazioni
grafiche

Sintetizziamo i
dati: le medie

Le medie non
basta

Frequenze assolute o relative?

I dati che seguono si riferiscono al numero di pazienti

Covid ricoverati in interapia intensiva in Emilia Romagna il 31 gennaio 2022.

	Freq. ass.	Totale	Freq. rel.	Freq. rel. x 100000
Non vaccinati	88	300000	0.000293	29.3
Vaccinati	61	3700000	0.0000165	1.65

Quando confrontiamo gruppi di numerosità diversa il confronto tra le frequenze assolute è sbagliato (equivale all'errore che si commette scrivendo "o confrontato"). Perchè i numeri siano confrontabili è necessario passare alle frequenze relative.

Il calcolo delle frequenze relative equivale ad una proporzione che riporta ad 1 (o a 100000) il totale in entrambi i gruppi rendendoli comparabili

$$88 : 300000 = 0.000293 : 1 = 29.3 : 100000)$$

Se scegliamo di confrontare i dati per rapporto avremo $29.3/1.65=17.8$ e non $88/61=1.4$.



Alessandra Nardi

Alcuni studi
sperimentali

Iniziamo a
descrivere i dati:
la distribuzione
di frequenze

Rappresentazioni
grafiche

Sintetizziamo i
dati: le medie

Le medie non
basta

A volte viene riportato il dato relativo alle sole terapie intensive da cui risulta che su un totale di 149 pazienti ricoverati il 41% è non vaccinato.

In questo caso si tratta di un peccato di omissione
L'informazione completa è contenuta nella tabella doppia.

	Ricoverati	Non ricoverati	Totale
Non vaccinati	88	299912	300000
Vaccinati	61	3699939	3700000
Totale	149	3999851	4000000

Bisognerebbe aggiungere che tra i soggetti che non sono ricoverati in terapia intensiva il 92.5% è vaccinato o se preferite che soltanto il 7.5 è non vaccinato.



Alessandra Nardi

Alcuni studi
sperimentali

Iniziamo a
descrivere i dati:
la distribuzione
di frequenze

Rappresentazioni
grafiche

Sintetizziamo i
dati: le medie

Le media non
basta

Se il carattere è quantitativo continuo, come costruiamo una distribuzione di frequenze? Riportare valori osservati e relative frequenze non garantisce una reale sintesi poichè i valori osservati raramente coincidono

Diventa allora necessario ricodificare la variabile in classi

Consideriamo ad esempio nello studio sull'ormone della crescita l'età dei soggetti che ricodifichiamo in classi aperte a sinistra e chiuse a destra, della stessa ampiezza (10 anni)

N.B. le classi devono costituire una partizione dell'intervallo che va dal valore minimo al valore massimo osservato. Questo vuol dire che la loro unione deve restituirci tutto l'intervallo per assicurare che tutti i valori osservabili possano essere classificati e non devono esserci intersezioni tra le classi per evitare ambiguità nell'assegnazione.



Alessandra Nardi

Alcuni studi
sperimentali

Iniziamo a
descrivere i dati:
la distribuzione
di frequenze

Rappresentazioni
grafiche

Sintetizziamo i
dati: le medie

Le medie non
basta

Otteniamo

Modalità	Frequenze assolute	Frequenze relative	Frequenze relative cumulate
(10, 20]	10	0.15	0.15
(20, 30]	20	0.30	0.45
(30, 40]	14	0.21	0.66
(40, 50]	13	0.20	0.86
(50, 60]	6	0.09	0.95
(60, 70]	1	0.01	0.96
(70, 80]	3	0.04	1
Totale	67	1	



Alessandra Nardi

Alcuni studi
sperimentali

Iniziamo a
descrivere i dati:
la distribuzione
di frequenze

Rappresentazioni
grafiche

Sintetizziamo i
dati: le medie

Le media non
basta

Classi equi-ampie oppure equi-distribuite?

Nella distribuzione precedente l'informazione non è distribuita uniformemente.

In alternativa

Modalità	Frequenze assolute	Frequenze relative
14 ÷ 20	10	0.150
20 ÷ 26	12	0.179
26 ÷ 31	11	0.164
31 ÷ 40	11	0.164
40 ÷ 48	11	0.164
48 ÷ 79	12	0.179
Totale	67	1

Attenzione adesso le frequenze di classi diverse non sono confrontabili.



Alessandra Nardi

Alcuni studi
sperimentali

Iniziamo a
descrivere i dati:
la distribuzione
di frequenze

Rappresentazioni
grafiche

Sintetizziamo i
dati: le medie

Le media non
basta

La scelta dipende dagli obiettivi che ci prefiggiamo. Per scopi puramente descrittivi è preferibile utilizzare classi della stessa ampiezza che, garantendo la confrontabilità, facilitano l'interpretazione dei dati. Se invece le frequenze campionarie hanno un valore inferenziale, ad esempio stiamo *stimando* la distribuzione nella popolazione, allora una uguale numerosità garantisce una simile precisione.

Torneremo sull'argomento quando andremo a confrontare curve di sopravvivenza.



Alessandra Nardi

Alcuni studi
sperimentali

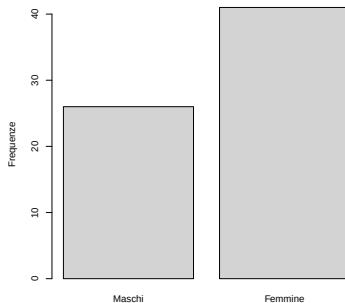
Iniziamo a
descrivere i dati:
la distribuzione
di frequenze

**Rappresentazioni
grafiche**

Sintetizziamo i
dati: le medie

Le media non
basta

Per variabili qualitative possiamo rappresentare la distribuzione di frequenze ponendo ascissa le modalità ed in ordinata le frequenze assolute o relative.



Questo grafico a barre, che non è un istogramma, può essere utilizzato anche nel caso di caratteri quantitativi discreti.



Alessandra Nardi

Alcuni studi
sperimentali

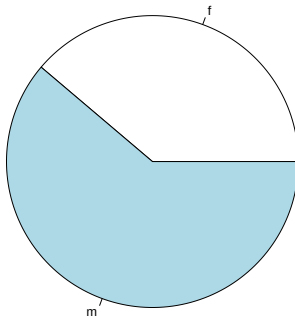
Iniziamo a
descrivere i dati:
la distribuzione
di frequenze

**Rappresentazioni
grafiche**

Sintetizziamo i
dati: le medie

Le media non
basta

In alternativa possiamo utilizzare il diagramma a *torta* che è una rappresentazione areale, è infatti l'area delle sezioni ad essere proporzionale alla frequenza.





Alessandra Nardi

Alcuni studi
sperimentali

Iniziamo a
descrivere i dati:
la distribuzione
di frequenze

Rappresentazioni
grafiche

Sintetizziamo i
dati: le medie

Le media non
basta

Nel caso e solo nel caso di variabili quantitative continue facciamo corrispondere ad ogni classe un rettangolo di area pari (o proporzionale) alla frequenza della classe.

Se consideriamo una generica classe

$$(x_i, x_{i+1})$$

l'ampiezza della base sarà

$$(x_{i+1} - x_i)$$

. Osserviamo come stiamo di fatto calcolando una distanza che non avrebbe senso per variabili qualitative. Perchè l'area del rettangolo corrispondente sia pari alla frequenza (assoluta o relativa) l'altezza dovrà essere pari a

$$\frac{n_i}{x_{i+1} - x_i}$$

La rappresentazione così ottenuta prende il nome di **istogramma**



Alessandra Nardi

Alcuni studi
sperimentali

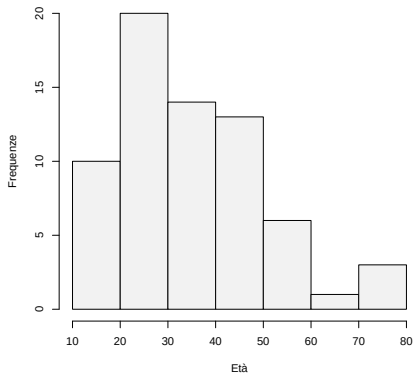
Iniziamo a
descrivere i dati:
la distribuzione
di frequenze

**Rappresentazioni
grafiche**

Sintetizziamo i
dati: le medie

Le media non
basta

Istogramma



In ordinata compaiono le frequenze, come mai?



Alessandra Nardi

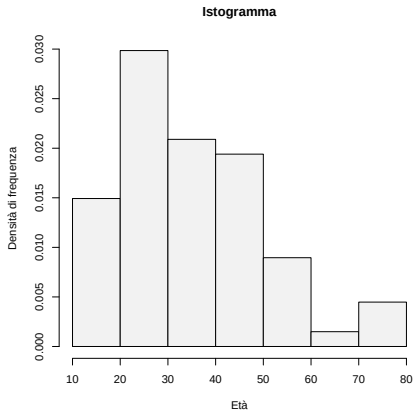
Alcuni studi
sperimentali

Iniziamo a
descrivere i dati:
la distribuzione
di frequenze

**Rappresentazioni
grafiche**

Sintetizziamo i
dati: le medie

Le media non
basta



In questo caso l'istogramma è stato calcolato utilizzando le frequenze relative rappresentate dalle aree dei rettangoli.



Alessandra Nardi

Alcuni studi
sperimentali

Iniziamo a
descrivere i dati:
la distribuzione
di frequenze

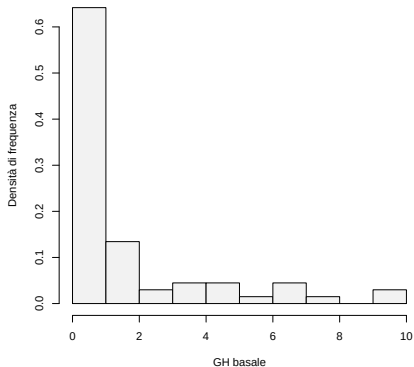
**Rappresentazioni
grafiche**

Sintetizziamo i
dati: le medie

Le media non
basta

Proviamo a descrivere la distribuzione basale dell'ormone della crescita osservata sui nostri dati

Istogramma



Osserviamo una concentrazione dei valori tra 0 e 1 anche se sono presenti anche valori elevati.



Possiamo modulare l'ampiezza delle classi tenendo conto della loro numerosità

Alessandra Nardi

Alcuni studi
sperimentali

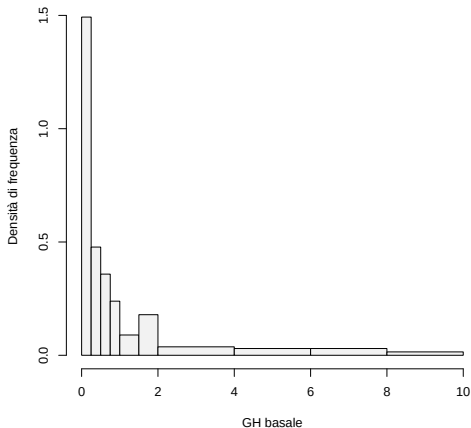
Iniziamo a
descrivere i dati:
la distribuzione
di frequenze

**Rappresentazioni
grafiche**

Sintetizziamo i
dati: le medie

Le media non
basta

Istogramma





Alessandra Nardi

Alcuni studi
sperimentali

Iniziamo a
descrivere i dati:
la distribuzione
di frequenze

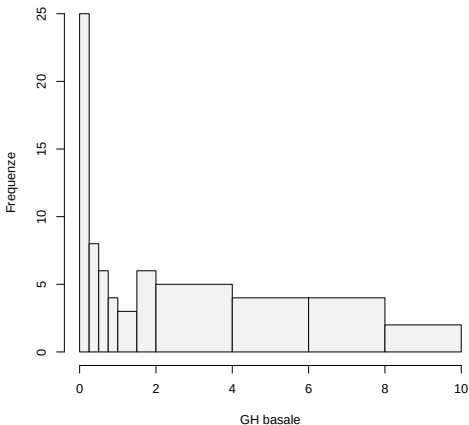
**Rappresentazioni
grafiche**

Sintetizziamo i
dati: le medie

Le media non
basta

Attenzione, se le classi sono diversa ampiezza è sbagliato riportare in ordinata le frequenze

Istogramma





Alessandra Nardi

Alcuni studi
sperimentali

Iniziamo a
descrivere i dati:
la distribuzione
di frequenze

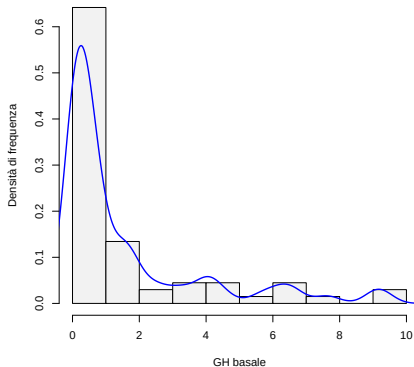
**Rappresentazioni
grafiche**

Sintetizziamo i
dati: le medie

Le media non
basta

Esiste un modo più elegante di descrivere la distribuzione

Istogramma





Alessandra Nardi

Alcuni studi
sperimentali

Iniziamo a
descrivere i dati:
la distribuzione
di frequenze

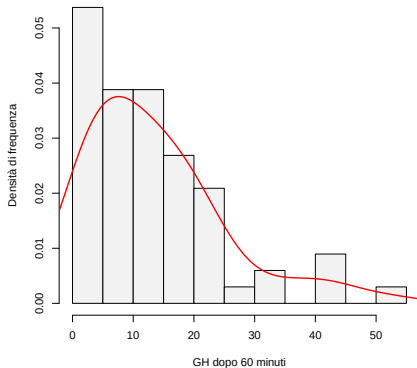
**Rappresentazioni
grafiche**

Sintetizziamo i
dati: le medie

Le media non
basta

Come si modificano i livelli dell'ormone dopo 60 minuti dalla stimolazione?

Istogramma





Alessandra Nardi

Alcuni studi
sperimentali

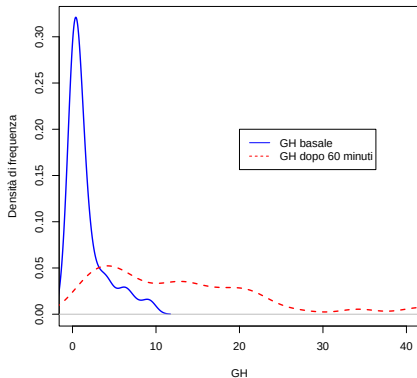
Iniziamo a
descrivere i dati:
la distribuzione
di frequenze

**Rappresentazioni
grafiche**

Sintetizziamo i
dati: le medie

Le media non
basta

Mettiamo a confronto le due distribuzioni





Alessandra Nardi

Alcuni studi
sperimentali

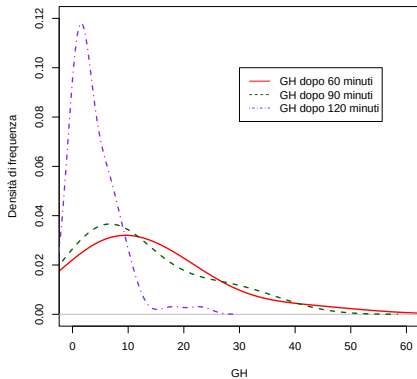
Iniziamo a
descrivere i dati:
la distribuzione
di frequenze

**Rappresentazioni
grafiche**

Sintetizziamo i
dati: le medie

Le media non
basta

Confrontiamo le distribuzioni dell'ormone ai tempi successivi





Alessandra Nardi

Alcuni studi
sperimentali

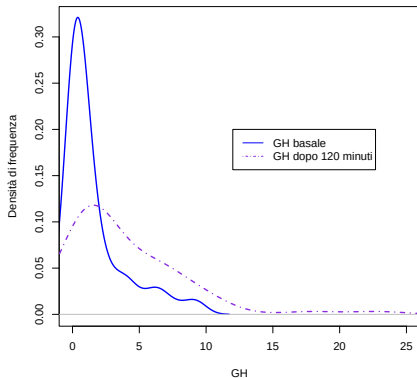
Iniziamo a
descrivere i dati:
la distribuzione
di frequenze

**Rappresentazioni
grafiche**

Sintetizziamo i
dati: le medie

Le medie non
basta

Possiamo chiederci se dopo 120 minuti il livello dell'ormone è tornato ai valori basali





Alessandra Nardi

Alcuni studi
sperimentali

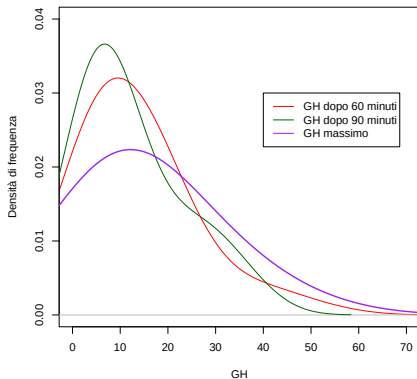
Iniziamo a
descrivere i dati:
la distribuzione
di frequenze

**Rappresentazioni
grafiche**

Sintetizziamo i
dati: le medie

Le media non
basta

Possiamo anche chiederci se il livello massimo raggiunto dell'ormone dopo stimolazione possa essere preferibile





Alessandra Nardi

Alcuni studi
sperimentali

Iniziamo a
descrivere i dati:
la distribuzione
di frequenze

Rappresentazioni
grafiche

Sintetizziamo i
dati: le medie

Le media non
basta

Da una pluralità di dati ad un solo valore a rappresentarli che *posizioni* la loro distribuzione di frequenza e ci dia un'idea del loro ordine di grandezza. Tre sono le *medie* di cui tratteremo:

- la media aritmetica $Ma = \frac{\sum_{i=1}^n x_i}{n}$;
- la mediana, Me, cioè il valore che corrisponde all'osservazione che occupa il posto centrale nella serie ordinata in modo crescente (non decrescente) di tutte le nostre osservazioni;
- la moda, cioè la modalità a cui corrisponde la frequenza più elevata, assoluta o relativa.



Alessandra Nardi

Alcuni studi
sperimentali

Iniziamo a
descrivere i dati:
la distribuzione
di frequenze

Rappresentazioni
grafiche

Sintetizziamo i
dati: le medie

Le media non
basta

La sintesi dei dati è essenziale per la loro interpretazione ma comporta sempre una perdita d'informazione

Passando da n osservazioni ad un solo valore, che indicheremo genericamente con M , è evidente che la nostra sintesi è estrema e la perdita di informazione sarà consistente.

Immaginiamo di sostituire a ciascuna osservazione x_i il valore M ed indichiamo con $d(x_i, M)$ la perdita d'informazione che ne deriva.

Possiamo considerare $d(x_i, M)$ una *distanza* tra l'osservazione e la media M chiamata a rappresentarla.

La perdita d'informazione complessiva sarà

$$\sum_{i=1}^n d(x_i, M)$$



Alessandra Nardi

Alcuni studi
sperimentali

Iniziamo a
descrivere i dati:
la distribuzione
di frequenze

Rappresentazioni
grafiche

Sintetizziamo i
dati: le medie

Le media non
basta

Il principio che ci guida è scegliere M in modo da garantire la minima perdita informativa.

Ora ciascuna delle medie considerate è a suo modo ottimale, cioè garantisce la minima perdita d'informazione in corrispondenza di una particolare scelta della distanza $d(x_i, M)$.

Consideriamo tre diversi modi di valutare la distanza tra x_i e M :

- $d(x_i, M) = (x_i - M)^2$;
- $d(x_i, M) = |x_i - M|$;
- $d(x_i, M) = 0$ se $x_i = M$ e $d(x_i, M) = 1$ se $x_i \neq M$.

In cosa differiscono?



Alessandra Nardi

Alcuni studi
sperimentali

Iniziamo a
descrivere i dati:
la distribuzione
di frequenze

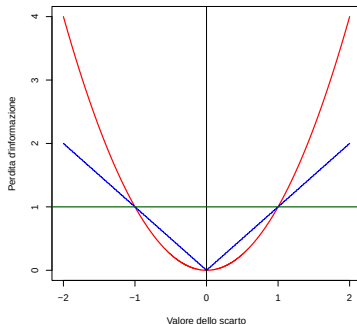
Rappresentazioni
grafiche

Sintetizziamo i
dati: le medie

Le media non
basta

Le tre *distanze* pesano diversamente gli scarti $(x_i - M)$. Nel primo caso lo scarto è elevato al quadrato, nel secondo caso si considera il suo valore assoluto, nel terzo caso la perdita di informazione vale sempre 1 e si annulla soltanto se $x_i = M$.

Graficamente





Alessandra Nardi

Alcuni studi
sperimentali

Iniziamo a
descrivere i dati:
la distribuzione
di frequenze

Rappresentazioni
grafiche

Sintetizziamo i
dati: le medie

Le media non
basta

Lo scarto al quadrato assegna perdite molto elevate a scarti elevati, enfatizzando il loro peso mentre poco rilievo è dato agli scarti piccoli. Se scegliamo questa distanza la perdita totale di informazione sarà

$$\sum_{i=1}^n (x_i - M)^2$$

Tra tutte le possibili medie M quella che rende minima questa funzione è la media aritmetica.

$$\min_M \left(\sum_{i=1}^n (x_i - M)^2 \right) = \sum_{i=1}^n (x_i - M_a)^2$$

Forse avete riconosciuto nella funzione di perdita totale, quando ad M sostituiamo M_a , il numeratore nella formula della varianza. Si annuncia fin da subito un legame, quello tra media aritmetica e varianza, che sarà inscindibile



Alessandra Nardi

Alcuni studi
sperimentali

Iniziamo a
descrivere i dati:
la distribuzione
di frequenze

Rappresentazioni
grafiche

**Sintetizziamo i
dati: le medie**

Le medie non
basta

La media aritmetica per sua natura sintetizza i dati cercando di non allontanarsi troppo proprio dai valori più lontani. Ne deriva inevitabilmente che è estremamente sensibile ai dati estremi.

Immaginiamo che i valori a seguire corrispondano ai livelli basali dell'ormone della crescita osservati su 5 volontari.

$$[0.1, 0.3, 0.6, 0.2, 0.5]$$

La media aritmetica è 0.34, come vedete in grado di ben rappresentare le nostre osservazioni.

Ora proviamo a modificare l'ultimo valore portandolo a 15

$$[0.1, 0.3, 0.6, 0.2, 15]$$

La media aritmetica si sposta e diventa 3.24.

In termini di capacità rappresentativa dei dati non va benissimo.



Alessandra Nardi

Alcuni studi
sperimentali

Iniziamo a
descrivere i dati:
la distribuzione
di frequenze

Rappresentazioni
grafiche

**Sintetizziamo i
dati: le medie**

Le media non
basta

Vediamo alcune sue proprietà:

- è sempre compresa tra il minimo ed il massimo dei valori osservati;
- è un baricentro, cioè bilancia esattamente scarti positivi e negativi ($\sum_{i=1}^n (x_i - Ma) = 0$);
- se si aggiunge una costante ai dati, la media aritmetica sarà modificata dello stesso ammontare;
- se si moltiplicano i dati per una costante, la media aritmetica risulterà moltiplicata per la stessa quantità;
- è associativa nel senso che la media di dati osservata in più gruppi si ottiene come media delle medie dei singoli gruppi, tenuto opportunamente conto della differente numerosità;
- se sostituita alle singole osservazioni, lascia inalterato l'ammontare complessivo del carattere (equidistribuisce).



Alessandra Nardi

Alcuni studi
sperimentali

Iniziamo a
descrivere i dati:
la distribuzione
di frequenze

Rappresentazioni
grafiche

**Sintetizziamo i
dati: le medie**

Le media non
basta

Lo scarto in valore assoluto si comporta invece in maniera equa, assegnando perdite che corrispondono ai valori degli scarti (con il segno positivo).

Se scegliamo questa distanza la perdita totale di informazione sarà

$$\sum_{i=1}^n |x_i - M|$$

Tra tutte le possibili medie M quella che rende minima questa funzione è la mediana cioè quel valore che occupa il posto centrale nella serie delle osservazioni ordinata in modo non decrescente.



Alessandra Nardi

Alcuni studi
sperimentali

Iniziamo a
descrivere i dati:
la distribuzione
di frequenze

Rappresentazioni
grafiche

**Sintetizziamo i
dati: le medie**

Le media non
basta

Torniamo all'esempio precedente

$[0.1, 0.3, 0.6, 0.2, 0.5]$

Ordiniamo i dati

$[0.1, 0.2, 0.3, 0.5, 0.6]$

Le mediana corrisponde al valore 0.3 che occupa il posto centrale
Se le osservazioni sono in numero pari sarà semplicemente la media
dei 2 valori centrali.



Alessandra Nardi

Alcuni studi
sperimentali

Iniziamo a
descrivere i dati:
la distribuzione
di frequenze

Rappresentazioni
grafiche

Sintetizziamo i
dati: le medie

Le media non
basta

Ora proviamo di nuovo a modificare l'ultimo valore portandolo a 15

$$[0.1, 0.3, 0.6, 0.2, 15]$$

Ordiniamo di nuovo i dati

$$[0.1, 0.2, 0.3, 0.6, 15]$$

e ci accorgiamo che la mediana è ancora 0.3.

A differenza della media aritmetica, la mediana resta rappresentativa dei nostri dati anche in presenza di valori estremi. In altri termini è più robusta rispetto alla presenza di valori anomali o comunque molto distanti dalla massa dei dati.



Alessandra Nardi

Alcuni studi
sperimentali

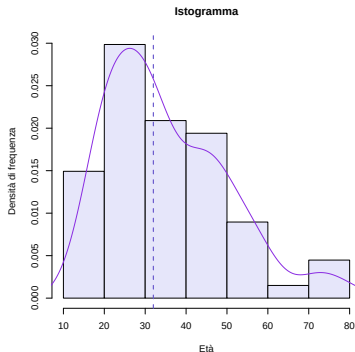
Iniziamo a
descrivere i dati:
la distribuzione
di frequenze

Rappresentazioni
grafiche

**Sintetizziamo i
dati: le medie**

Le media non
basta

Se i nostri dati sono descritti attraverso un istogramma di frequenze allora la mediana è il valore che divide l'area complessiva in due parti uguali, lasciando alla sua sinistra il 50% delle osservazioni.





Alessandra Nardi

Alcuni studi
sperimentali

Iniziamo a
descrivere i dati:
la distribuzione
di frequenze

Rappresentazioni
grafiche

Sintetizziamo i
dati: le medie

Le medie non
basta

Infine, se assumiamo come misura di perdita individuale $d(x_i, M)$ la più semplice possibile, cioè quella che vale 0 se $x_i = M$ ed 1 se $x_i \neq M$, allora il valore che rende minima la perdita complessiva

$$\sum_{i=1}^n I_{(x_i \neq M)}$$

è la Moda, la modalità a cui corrisponde la frequenza più elevata.

Assumiamo come valore rappresentativo di ciascuna osservazione quello che abbiamo osservato più spesso, i.e. la modalità espressa dalla *maggioranza relativa*.

La moda risponde all'idea di *minimizzare gli scontenti* e quindi è utile in quelle situazioni dove il consenso ha significato per la decisione (la moda è un indice per governare).

Si tratta di un indice che ha senso soltanto per variabili qualitative.



Alessandra Nardi

Alcuni studi
sperimentali

Iniziamo a
descrivere i dati:
la distribuzione
di frequenze

Rappresentazioni
grafiche

Sintetizziamo i
dati: le medie

Le media non
basta

La media offre una sintesi *estrema* dei dati ma nasconde molte cose, prima tra tutte la variabilità delle osservazioni. Immagiamo che i dati a seguire si riferiscano ai voti conseguiti da tre studenti, tutti con media e mediana pari a 25 ...

$A : (22, 22, 23, 23, 24, 25, 26, 27, 27, 28, 28)$

$B : (22, 22, 22, 22, 22, 25, 28, 28, 28, 28, 28)$

$C : (25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25)$

... ma molto diversi tra di loro

Come misuriamo la variabilità?



Alessandra Nardi

Alcuni studi
sperimentali

Iniziamo a
descrivere i dati:
la distribuzione
di frequenze

Rappresentazioni
grafiche

Sintetizziamo i
dati: le medie

Le media non
basta

Se tra le medie abbiamo scelto la media aritmetica allora è naturale misurare la variabilità attraverso la varianza

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

Infatti abbiamo visto che la media aritmetica è quel valore che minimizza la perdita di informazione se scegliamo come distanza lo scarto al quadrato $d(x_i, M) = (x_i - M)^2$.

Ricordiamo che la radice quadrata della varianza, σ , è nota come deviazione standard ed è spesso preferita alla varianza perchè ci riporta alla scala di misura originaria.



Alessandra Nardi

Alcuni studi
sperimentali

Iniziamo a
descrivere i dati:
la distribuzione
di frequenze

Rappresentazioni
grafiche

Sintetizziamo i
dati: le medie

Le media non
basta

Se invece tra le medie abbiamo scelto la mediana sarebbe naturale valutare la variabilità utilizzando la media degli scarti in valori assoluto.

$$\frac{\sum_{i=1}^n |x_i - Me|}{n}$$

A questo indice tuttavia è in genere preferita la distanza o range inter-quartilico. Ricordiamo che i quartili $Q_1, Q_2 \equiv Me, Q_3$, corrispondono a valori che dividono in 4 parti uguali la serie ordinata in modo crescente (non decrescente) delle nostre osservazioni.



Alessandra Nardi

Alcuni studi
sperimentali

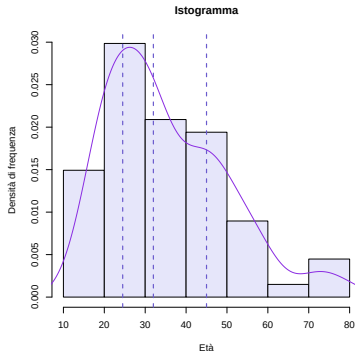
Iniziamo a
descrivere i dati:
la distribuzione
di frequenze

Rappresentazioni
grafiche

Sintetizziamo i
dati: le medie

Le media non
basta

Se i nostri dati sono descritti attraverso un istogramma di frequenze (relative) allora i quartili dividono l'area complessiva in quattro parti uguali ciascuna pari a 0.25.



dove $Q_1 = 24.5, Q_2 \equiv Me = 32, Q_3 = 45$. Il range interquartilico sarà $IQR = Q_3 - Q_1 = 20.5$. Una piccola distanza interquartilica indica che il 50% delle nostre osservazioni è concentrato in un piccolo intervallo, suggerendo una bassa variabilità. Attenzione però a cosa accade nelle code



A qualcuno piace il *boxplot* o *box and whiskers plot*

Alessandra Nardi

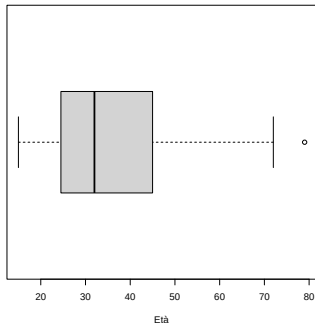
Alcuni studi
sperimentali

Iniziamo a
descrivere i dati:
la distribuzione
di frequenze

Rappresentazioni
grafiche

Sintetizziamo i
dati: le medie

**Le media non
basta**



dove il rettangolo (box) è delimitato da Q_1 e Q_3 mentre il segmento centrale corrisponde alla Mediana. I baffi (whiskers) sono delimitati da massimo e minimo a meno che non compaiano degli outlier descritti da singoli punti esterni ai baffi. In questo caso vengono definiti outlier i valori inferiori a $Q_1 - 1.5 \cdot IQR$ oppure superiori a $Q_3 + 1.5 \cdot IQR$.



Alessandra Nardi

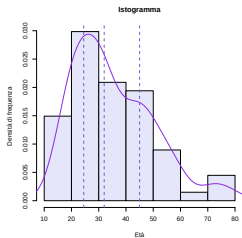
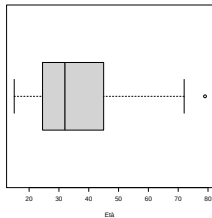
Alcuni studi
sperimentali

Iniziamo a
descrivere i dati:
la distribuzione
di frequenze

Rappresentazioni
grafiche

Sintetizziamo i
dati: le medie

**Le media non
basta**





Alessandra Nardi

Alcuni studi
sperimentali

Iniziamo a
descrivere i dati:
la distribuzione
di frequenze

Rappresentazioni
grafiche

Sintetizziamo i
dati: le medie

**Le media non
basta**

Se il nostro carattere è qualitativo e decidiamo di riportare la modalità più frequente, cioè la Moda, allora è bene indicare la frequenza di coloro che hanno una modalità diversa (coloro che non la pensano come la maggioranza).

Ricordiamo infatti che la Moda è quella Media che rende minima la perdita di informazione

$$\sum_{i=1}^n I_{(x_i \neq M)}$$



Alessandra Nardi

Alcuni studi
sperimentali

Iniziamo a
descrivere i dati:
la distribuzione
di frequenze

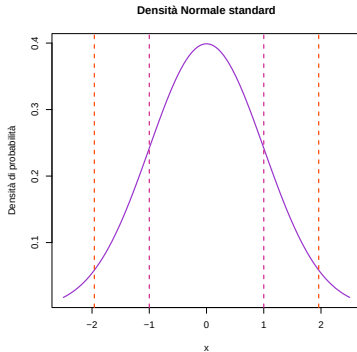
Rappresentazioni
grafiche

Sintetizziamo i
dati: le medie

**Le media non
basta**

La scelta di descrivere i nostri dati attraverso la coppia media e varianza oppure utilizzando i quartili è strettamente legata alla scelta o meno di assumere un modello Normale.

Media e varianza infatti, seppure con un diverso significato, identificano la densità Normale per la quale la probabilità di osservare valori nell'intervallo $\mu \pm \sigma$ è pari a 0.68 e sale a 0.95 se l'intervallo è $\mu \pm 1.96 \cdot \sigma$





Alessandra Nardi

Alcuni studi
sperimentali

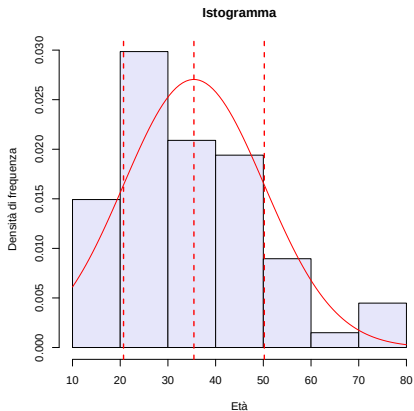
Iniziamo a
descrivere i dati:
la distribuzione
di frequenze

Rappresentazioni
grafiche

Sintetizziamo i
dati: le medie

**Le media non
basta**

Se la distribuzione dei nostri dati è simile ad una densità Normale, media e varianza possono darne una sintesi appropriata anche se tendono a nascondere una eventuale asimmetria (Età: 35.45 ± 14.75)





Alessandra Nardi

Alcuni studi
sperimentali

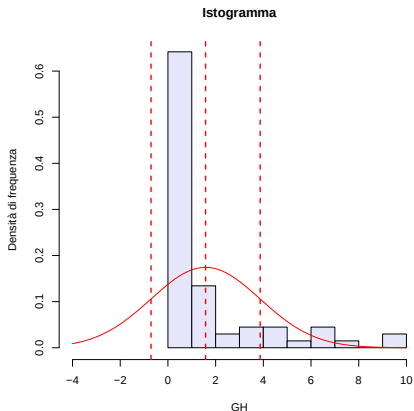
Iniziamo a
descrivere i dati:
la distribuzione
di frequenze

Rappresentazioni
grafiche

Sintetizziamo i
dati: le medie

**Le media non
basta**

Invece questa scelta può essere fuorviante se la forma della distribuzione è lontana da una densità Normale (GH: 1.58 ± 2.29)





Alessandra Nardi

Alcuni studi
sperimentali

Iniziamo a
descrivere i dati:
la distribuzione
di frequenze

Rappresentazioni
grafiche

Sintetizziamo i
dati: le medie

**Le media non
basta**

In questo caso è preferibile riportare i quartili seguendo una logica *non parametrico* (GH: 0.6, 0.17 – 1.70)

Istogramma

