

CORSI DI LAUREA AFFERENTI AL TRONCO COMUNE III
A.A. 2013/14

Corso integrato Fisica Statistica e Informatica
Statistica Medica

Alessandra Spagnoli

Info

LEZIONI: 28-30 Gennaio 2014

RICEVIMENTO: si consiglia di prendere appuntamento
Contatti: inviare una e-mail a alessandra.spagnoli@uniroma1.it

MATERIALE DIDATTICO

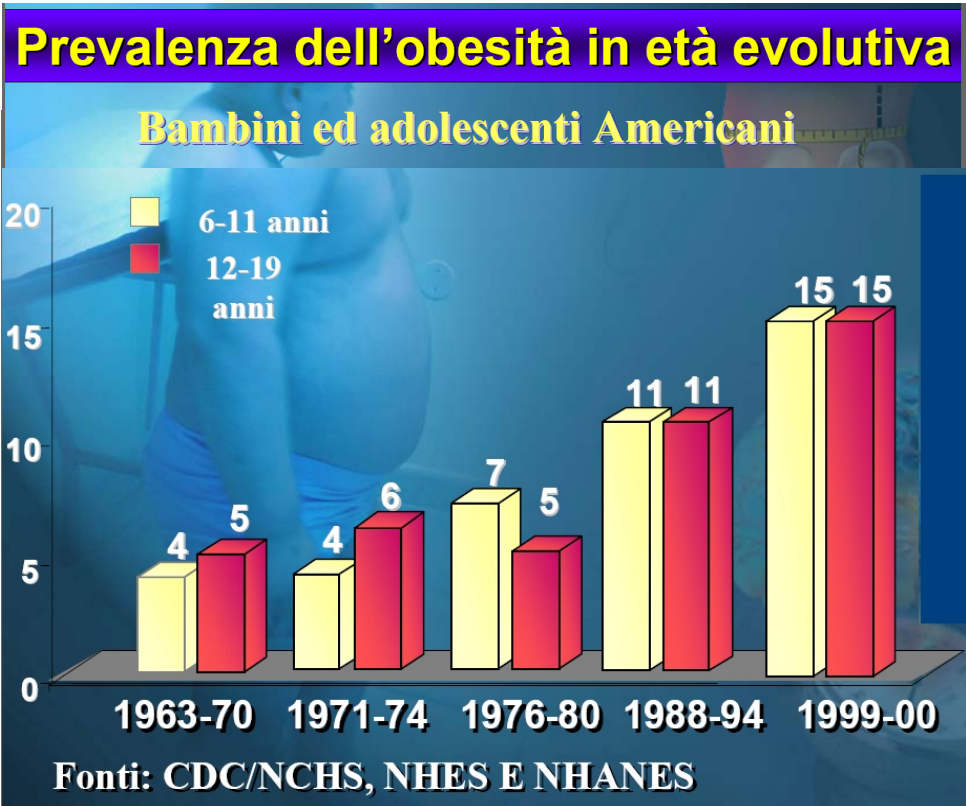
- Un testo di riferimento utile è: Lantieri PB, Risso D, Ravera G: Statistica medica per le professioni sanitarie, II ed. McGraw-Hill (2004)
- Appunti e stampati delle slides a cura della dr. Simona Iacobelli (disponibili in rete: http://www.uniroma2.it/didattica/statistica_tronco3)

MODALITA' D'ESAME

Le prove sono scritte, e comprendono domande a risposta multipla e piccoli esercizi.

Un po' di *statistiche* ...

I dati presentati nelle prossime slides sono tratti da un intervento del prof. Del Giudice (II Università Napoli) al convegno della Società Italiana di Pediatria Preventiva e Sociale (2008) sul tema dell'obesità infantile



EBM: operare secondo l'evidenza scientifica

- Si stima che il 15% degli errori nella pratica clinica sia di tipo cognitivo, ossia imputabile a:
 - a) Cattive informazioni
 - b) Cattivi ragionamenti – derivati dal trascurare o utilizzare male “buone” informazioni, ricorrendo sistematicamente a metodi errati
- L'etica impone di usare al meglio le risorse cognitive
- Oggi in ambito biomedico la conoscenza basata sui dati è sempre più alla base delle decisioni e degli interventi, sia sui singoli individui (pratica clinica) sia per le collettività (politiche sanitarie).
- Per tutti gli operatori in ambito biomedico è necessario:
 - a) Conoscere i metodi statistici per l'elaborazione e la comunicazione delle informazioni
 - b) Imparare a utilizzare correttamente le informazioni (processo di deduzione e interpretazione delle evidenze statistiche)

Programma del corso

Statistica Descrittiva

Terminologia

Strumenti

Elaborazione e
Comunicazione
dei dati (fase
descrittiva)

Elementi di
**Inferenza
Statistica**

Basi di
**Calcolo delle
Probabilità**

Elementi per
l'interpretazione

Elementi per
una corretta
elaborazione /
deduzione

Terminologia iniziale



- Popolazione; Campione; Unità statistiche
- Carattere, modalità
- Classificazione dei caratteri

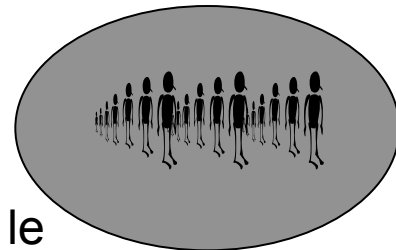
Popolazione

- Considerato un fenomeno di interesse, possiamo immaginare che esista una **POPOLAZIONE** di individui* che, se interamente osservata, ci permette di conoscere ogni aspetto di interesse del fenomeno
- Essa è anche detta **POPOLAZIONE OBIETTIVO**
- Può essere una popolazione **reale**, potenzialmente osservabile interamente (es. sondaggio fra gli italiani), o una popolazione **ideale, fittizia**, non identificabile

Esempio: Interessa studiare gli effetti del virus dell'influenza stagionale

Popolazione Obiettivo: tutti gli individui che sono stati già esposti al contagio, o lo saranno, tutti i pazienti che si sono ammalati, o si ammaleranno; compresi i soggetti esposti o ammalatisi in passato, e deceduti

Rappresentiamo la Popolazione come un insieme



*Gli elementi che costituiscono la popolazione sono le **unità statistiche**

Unità statistiche

A volte il fenomeno non si riferisce a individui umani (o animali), ma a gruppi di individui (es. famiglie) o enti (es. ospedali) o altri organismi (es. cellule). Si usa allora il termine più generale di UNITA' STATISTICA.

L'unità statistica è **l'elemento della popolazione su cui studiamo il fenomeno che ci interessa**, andando ad osservare alcune loro caratteristiche.

<i>unità</i>	<i>caratteristiche</i>
 Pazienti con tumore della mammella	Età, menopausa, stadio del tumore, dimensioni del tumore, ...
 Famiglie assistite dal consultorio	Numero di componenti, titolo di studio del capofamiglia, reddito complessivo, presenza di anziani >65 anni ...
 Ospedali presenti in Regione	Addetti, numero di posti letto, presenza di unità rianimazione ...

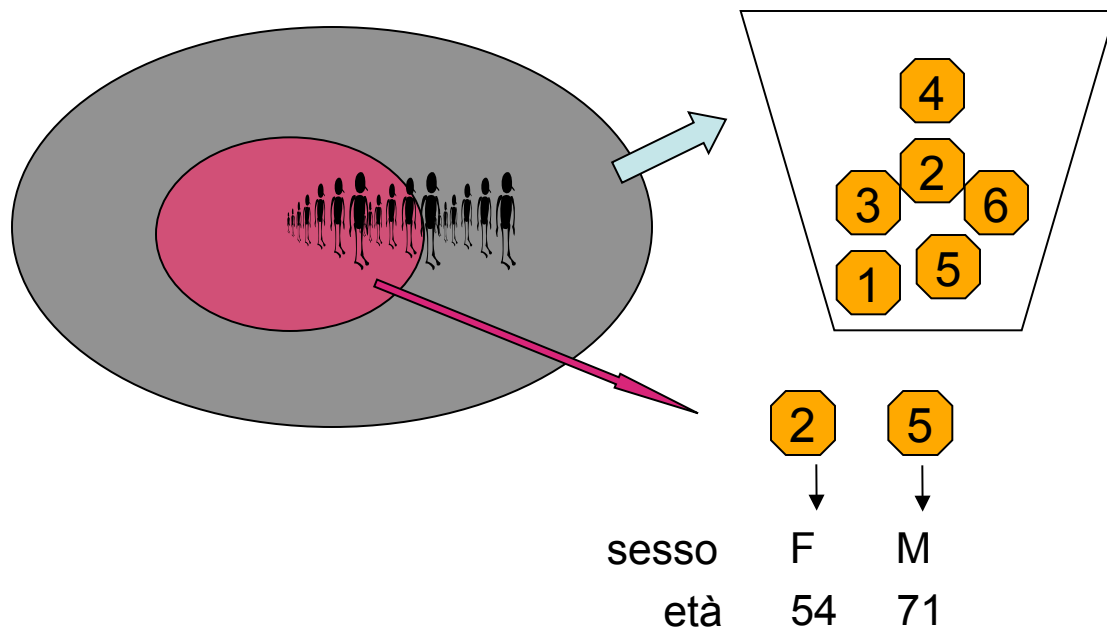
Campione

L'insieme degli individui su cui andiamo effettivamente a osservare il fenomeno è detto collettivo, o popolazione osservata, o CAMPIONE.

Idealmente, questi individui sono stati estratti dalla popolazione obiettivo, come palline estratte da un'urna.

→ Per questo la Statistica utilizza quella parte della Matematica che è il Calcolo delle Probabilità

→ Anche il campione è rappresentato come un insieme, ed essendo una parte della popolazione ("sottoinsieme"), è tutto contenuto nell'altro insieme



Spesso il termine CAMPIONE si riferisce non più alle unità estratte, ma direttamente ai dati osservati su tali unità

→ I dati sono assimilabili a numeri estratti da un'urna

L'elemento essenziale: un insieme di dati

caratteri (variabili)

unità statistiche

paziente	sex	age	BMI	weight*	pathology	diabetes
A	F	54	20.2	normopeso	diabetes	sì
B	M	64	31.2	obesity	dislipidemia	no
C	M	32	17.8	underweight	diabetes	sì
D	F	74	23.1	normopeso	insuff. renale	no
...						

*Classificazione del peso (soggetti adulti) secondo Body Mass Index: basata su classi di peso

< 18,5 sottopeso

18,5 – 24,9 normopeso

25 – 29,9 sovrappeso

> 30 obeso

Caratteri e Modalità

(modalità)

paziente	sex	età	BMI	peso	patologia	diabete
A	F	54	20.2	normopeso	diabete	sì
B	M	64	31.2	obesità	dislipidemia	no
C	M	32	17.8	sottopeso	diabete	sì
D	F	74	23.1	normopeso	insuff. renale	no
...						

- Le caratteristiche di interesse delle unità statistiche sono dette **CARATTERI**, o **VARIABILI**
- I caratteri presentano (si esprimono attraverso) dei **VALORI** o **MODALITA'**
 - Le unità statistiche differiscono fra loro per le modalità che esse presentano: il carattere presenta una **variabilità** che è l'oggetto di studio della statistica

Adozione di una codifica numerica

paziente	sex	age	BMI	weight	pathology	diabetes
A	F 2	54	20.2	normopeso 1	diabete 2	sì 1
B	M 1	64	31.2	obesità 3	dislipidemia 3	no 0
C	M 1	32	17.8	sottopeso 0	diabete 2	sì 1
D	F 2	74	23.1	normopeso 1	insuff. renale 1	no 0
...						

modalità - e loro “etichette” (labels)

peso:

< 18,5 sottopeso 0
 18,5 – 24,9 normopeso 1
 25 – 29,9 sovrappeso 2
 > 30 obeso 3

età, BMI: $\mathbb{R}^+$

sex: 1=M 2=F

diabete: 1=sì 0=no

patologia:

1 = insuff. renale
 2 = diabete
 3 = altro

Classificazione dei caratteri

- La natura del carattere dipende da che modalità esso presenta, e ha una corrispondenza nel tipo di operazione che è possibile fare:
 - Per confrontare due modalità / due unità
 - Per manipolare le sue modalità

QUALITATIVI

SCONNESSI

sex → M,F

patologia → ulcera, tumore gastrico, tumore intestinale, ...

ORDINATI

titolo di studio → nessuno o licenza elementare, licenza media, licenza superiore, laurea

stadio malattia → I,II,III

QUANTITATIVI

DISCRETI

numero di componenti (della famiglia) → 1,2,3,4, ...

gravidanze precedenti → 0, 1, 2, 3, ...

CONTINUI

età (anni compiuti) → 0,1,2,...,24,...,88,...

peso (kg) → 56.4, 78.2, ...

WBC (x 10³/ml) → 3.4, 2.8, ...

Caratteri Qualitativi

- Presentano modalità che corrispondono a diciture, attributi, caratteristiche descrivibili attraverso “parole” (ovvero, attraverso numeri che però non corrispondono a conteggi o misurazioni, ma esprimono convenzioni)
 - Non ammettono operazioni matematiche!!
- **SCONNESSI**: non si ha un ordinamento naturale o “tipico” (stabilito per convenzione)
 - è possibile solo dire se due unità sono uguali o diverse (se presentano la stessa modalità o modalità diverse)
- **ORDINATI**: esiste un ordinamento naturale o “tipico”
 - è possibile stabilire relazioni di superiorità / inferiorità fra due unità;
 - non è però possibile (o non ha senso) calcolare delle differenze per stabilire la “distanza” fra due unità
 - (Non farsi ingannare dalle codifiche numeriche!!)

Caratteri Dicotomici

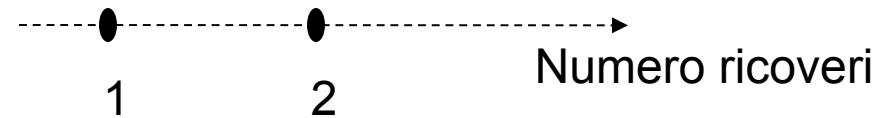
- Un tipo particolare di carattere qualitativo sconnesso è quello BINARIO o DICOTOMICO, cioè che assume 2 sole modalità
- Esso può essere solitamente inteso come indicatore di presenza/assenza di una certa caratteristica
- Corrispondentemente, di solito si usa la codifica numerica 0/1 (0=no=assenza, 1=sì=presenza)

Esempi

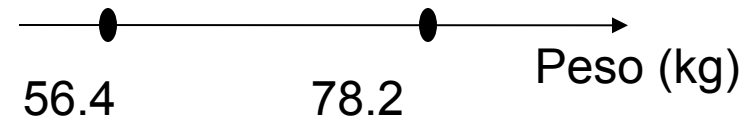
- Fumatore: sì/no
- Rispondente (alla terapia): sì/no
- Sesso = M/F, ovvero:
- Paziente maschio: sì/no

Caratteri Quantitativi

- Presentano modalità effettivamente numeriche, ottenute tramite conteggio o misurazione; sulle modalità è possibile eseguire operazioni matematiche
- **DISCRETI**: le modalità possono essere enumerate; i valori compresi fra due modalità possono NON essere a loro volta delle modalità
 - generalmente ottenuti tramite conteggio



- **CONTINUI**: le modalità NON possono essere enumerate; i valori compresi fra due modalità sono sempre a loro volta delle modalità
 - generalmente ottenuti tramite misurazione

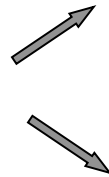


→ L'imprecisione dello strumento di misura determina una APPROSSIMAZIONE o ARROTONDAMENTO, ma la natura del carattere è continua

→ E' assimilabile a un continuo un carattere di natura discreta che assuma un numero molto alto di modalità, es. il numero di abitanti di un comune, o l'età misurata in anni compiuti

Ricodifica delle variabili (1)

PATOLOGIA
a - tumore gastrico
b - ulcera gastrica
c - tumore intestinale



PATOLOGIA
ulcera (b)
tumore (a, c)

PATOLOGIA
gastrica (a, b)
intestinale (c)

STADIO TUMORE
I
II
III
IV



STADIO TUMORE
I - iniziale
II-III – progredito
IV - terminale

Per i caratteri qualitativi si può fare un accorpamento di modalità

Per i qualitativi sconnessi, esso può seguire vari criteri.

Per un qualitativo ordinato, è bene rispettare l'ordinamento delle modalità

Ricodifica delle variabili (2)

Età	
- 25	
25 - 45	$25 < \text{Età} \leq 45 \quad (25, 45]$
45 - 65	
65 -	$\text{Età} > 65 \text{ (classe aperta)}$

I caratteri quantitativi possono essere ridotti in CLASSI, accorpendo le modalità. Vanno così ad assomigliare ai qualitativi ordinati.

WBC		$\ln(\text{WBC})$
2.2		0.788
3.2		1.160
1.8		0.588
2.1		0.742

Le modalità quantitative possono essere trasformate mediante operazioni matematiche.

Scelta della codifica

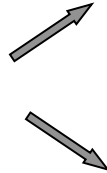
La codifica, e quindi la natura del carattere, possono cambiare a seconda della definizione che gli si dà, e dipendere dagli obiettivi dello studio

Es: Caratteristica di interesse: il fumo di sigaretta

→ Fumo → Numero di sigarette fumate (mediamente) in un giorno: 0, 1, 2, 3, ...20, ...
Carattere quantitativo discreto ma **assimilabile a continuo**

Sigarette
0
1-5
6-10
10-20
> 20

Il carattere
quantitativo in classi
mantiene una natura
quantitativa, ma
perde alcune
caratteristiche ...



Fumo
no = 0 sigarette
si = > 0 sigarette

Dicotomico

Fumatore
no = 0 sigarette
moderato = 1-10 sigarette
forte = 10 sigarette

Qualitativo ordinato

Alternativa:

non fumatore	ex-fumatore	fumatore
--------------	-------------	----------

Qualitativo sconnesso (o ordinato?)

Gerarchia dei caratteri (1)

Carattere	Operazioni possibili sulle modalità
Qualitativo sconnesso	Confronto: Stabilire uguaglianza o diversità ($=$ o $\neq$) Manipolazione: accorpamento, secondo criteri vari
Qualitativo ordinato	Confronto: Stabilire relazioni di superiorità / inferiorità Manipolazione: accorpamento, mantenendo l'ordinamento
Quantitativo	Confronto: Differenza o rapporto ($-$, $/$) Manipolazione: Suddivisione in classi; applicazione di operazioni matematiche ($+$, $-$, $\cdot$, $/$, $\log$, ...)

Descrivere: tabelle, grafici e indici sintetici



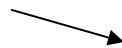
- Tabelle e grafici
 - Frequenze relative e percentuali; frequenze cumulate
 - Concetto di Densità di Frequenza, istogramma
- Indici statistici
 - di posizione: moda, media, mediana, quartili
 - di variabilità: deviazione standard, varianza, coeff. di variazione
- Forma della distribuzione
 - la Normale

Le tabelle di frequenza

unità	SESSO	ETA
1	M	55
2	F	51
3	F	44
4	M	62
5	M	48
6	F	51
7	M	69
8	F	58
9	F	72
10	M	50
11	F	78
12	F	46



SESSO	n
M	5
F	7
tot	12



ETA'	n
- 50	4
50 - 65	5
65 -	3
tot	12

• La prima operazione utile per sintetizzare una serie di dati relativa ad un carattere è il **conteggio**: ad ogni modalità (o classe, intervallo di valori) si associa la **frequenza**, ossia il numero di unità che presentano quella modalità (o cadono in quella classe)

→ Rispetto alla serie originaria, la tabella è una **sintesi**, in cui si è persa una parte di informazione [il riferimento alle singole unità], e si è guadagnata una visione generale e “rapida” del fenomeno

Frequenze relative e percentuali

Distribuzione dei pazienti ricoverati sottoposti a regimi dietetici particolari rispetto al TIPO DI MALATTIA

Patologia	n	f	p (%)
Insuff. renale	454	0.244	24.4
Diabete	1227	0.659	65.9
Altra patol. Organica	153	0.082	8.2
Patologia psichiatrica	27	0.015	1.5
	1861	1.000	100.0

es. per la seconda modalità:

$$\frac{1227}{1861} = 0.659 \quad 0.659 \cdot 100 = 65.9$$

Queste quantità esprimono lo stesso rapporto della parte al tutto (frazione):

E' il concetto di **proporzione**

$$1227 : 1861 = 0.659 : 1 = 65.9 : 100$$

freq.
assoluta

totale delle
osservazioni
nel campione

freq.
relativa

totale
=1

freq percentuale
(%)

totale
=100

Percentuali: interpretazione e uso (1)

Risposta al trattamento	%
No	42.9
Si	57.1
<i>tot</i>	100.0

Risultati di uno studio clinico: RISPOSTA AL TRATTAMENTO

- Le **percentuali** di Risposta forniscono la DISTRIBUZIONE del carattere, e **possono essere interpretate come le probabilità**, per un generico paziente, di rispondere o non rispondere al trattamento
- Dunque, sottoponendo al trattamento 20 (nuovi) pazienti, ci si aspettano circa 11 rispondenti (circa il 60%):
 $0.571 \times 20 = 11.42$

Percentuali: interpretazione e uso (2)

Risposta al trattamento	%
No	42.9
Si	57.1
<i>tot</i>	100.0

freq.
6
8
14

freq.
600
800
1400

Presentiamo 2 scenari in cui le freq. percentuali di Risposta sono le stesse.

L'attendibilità dello studio è la stessa? Quale studio è più "affidabile"?

- Rispetto al conteggio delle frequenze assolute, il passaggio alle frequenze relative è una ulteriore sintesi: si perde l'informazione sulla **numerosità totale**, che è invece **fondamentale per capire l'attendibilità / la precisione dei dati**.

→ In presenza di percentuali, guardiamo e riportiamo sempre la numerosità totale del campione!!

Frequenze cumulate

Un'altra utile elaborazione delle frequenze, ma solo per caratteri ordinati

Numero figli	freq	p (%)
0	25	40%
1	21	33%
2	12	19%
3	4	6%
4 +	1	2%
totale	63	100%

cum	% cum
25	40%
46	73%
58	92%
62	98%
63	100%

Le frequenze cumulate (assolute o percentuali) rappresentano semplicemente le somme parziali delle frequenze fino alla modalità corrente

Ad esempio, guardando l'ultima colonna, posso subito vedere che:

→ 3 donne su 4 (73%) hanno al massimo 1 figlio;

→ il 92% delle donne hanno al massimo 2 figli, e quindi solo l'8% ha più di 2 figli

etc

Una sintesi di tutta la tabella: la Moda

Distribuzione dei pazienti ricoverati sottoposti a regimi dietetici particolari rispetto al TIPO DI MALATTIA

Patologia	n	p (%)
Insuff. renale	454	24.4
Diabete	1227	65.9
Altra patol. Organica	153	8.2
Patologia psichiatrica	27	1.5
	1861	100.0

La modalità più rappresentativa di questo carattere è quella che presenta la frequenza più alta: questo indice viene chiamato **MODA**

Qui, la moda è la modalità “Diabete”.

Possiamo dire che il “tipico” paziente ricoverato che richiede un regime dietetico particolare è affetto da diabete. Ovvero, in un gruppo di pazienti ricoverati sottoposti a regime dietetico particolare, la maggior parte soffre di diabete.

Grafici da tabelle di caratteri qualitativi

Distribuzione dei pazienti ricoverati sottoposti a regimi dietetici particolari rispetto al TIPO DI MALATTIA

Patologia	n	p (%)
Insuff. renale	454	24.4
Diabete	1227	65.9
Altra patol. Organica	153	8.2
Patologia psichiatrica	27	1.5
	1861	100.0

%

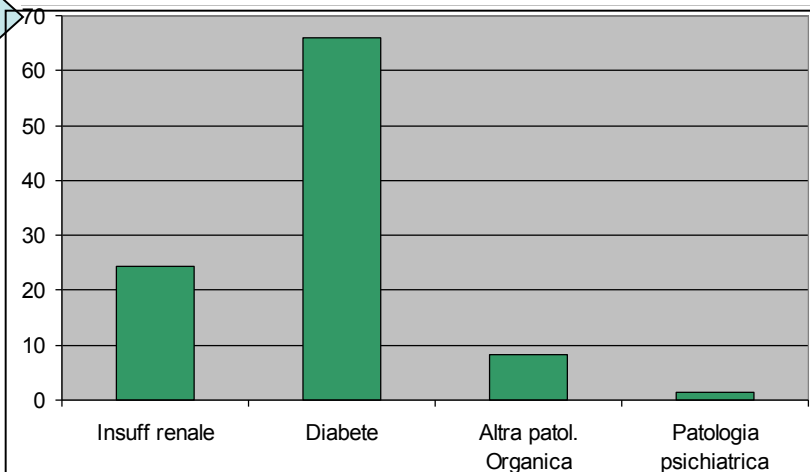


Grafico a colonne

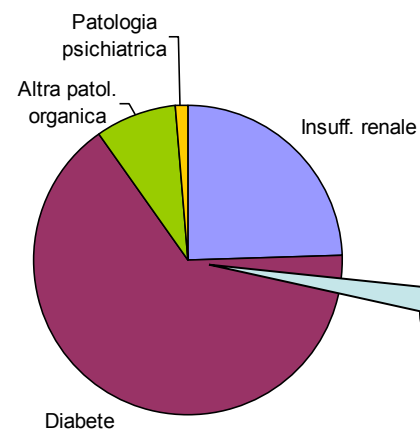


Grafico a torta

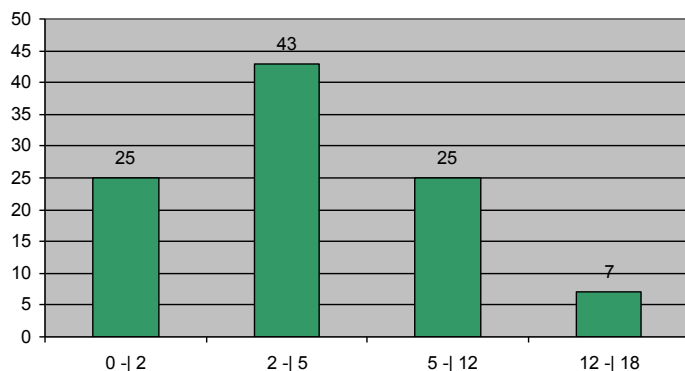
Grafici da tabelle di caratteri continui

Distribuzione di 56 pazienti pediatrici per età

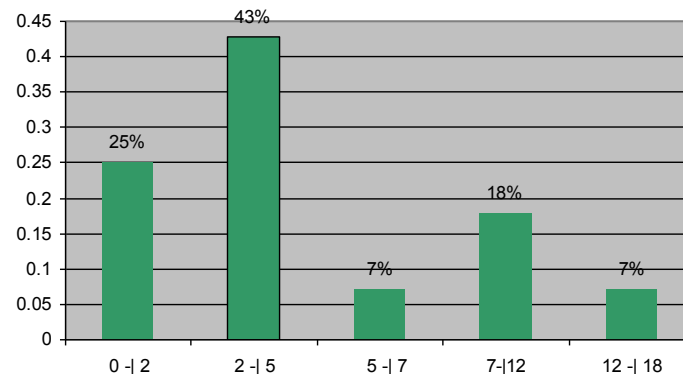
Età	freq.	%
0 - 2	14	25
2 - 5	24	43
5 - 12	14	25
12 - 18	4	7
	56	100

La semplice rappresentazione delle frequenze percentuali delle classi fornisce una **rappresentazione distorta** del fenomeno **se le classi non hanno la stessa ampiezza**

Ad esempio: le classi 0-|2 e 5-|12 hanno la stessa frequenza, e quindi vengono rappresentate come aventi la stessa importanza:



Immaginiamo di suddividere l'intervallo 5-|12 in due classi: con 4 pazienti di età 5-|7 e gli altri 10 di 7-|12: diventano “meno importanti” della classe 0-|2 !!



Concetto di densità di frequenza

Età	freq.	%
0 - 2	14	25.0
2 - 5	24	42.9
5 - 12	14	25.0
12 - 18	4	7.1
	56	100

La stessa frequenza (14 unità) della prima e della terza classe viene “spalmata” su intervalli di ampiezza diversa, rispettivamente di 2 anni (2-0) e di 7 anni (12-5);

Immaginando di passare a intervallini di età di ampiezza 1 (0-1 anno; 1-2 anni; 2-3 anni; etc) si avrebbero:

- dalla classe 0-|2, 14 casi spalmati su 2 anni $\rightarrow$ circa $14 / 2 = 7$ casi per ciascun intervallino
- dalla classe 5-|12, 14 casi spalmati su 7 anni $\rightarrow$ circa $14 / 7 = 2$ casi per ciascun intervallino

La frequenza va rapportata all'ampiezza della classe, ottenendo la densità di frequenza, un valore che **rappresenta quante unità sono presenti in ogni intervallino di ampiezza 1**

$$densità = \frac{frequenza}{ampiezza}$$

$$\Leftrightarrow frequenza = ampiezza \times densità$$

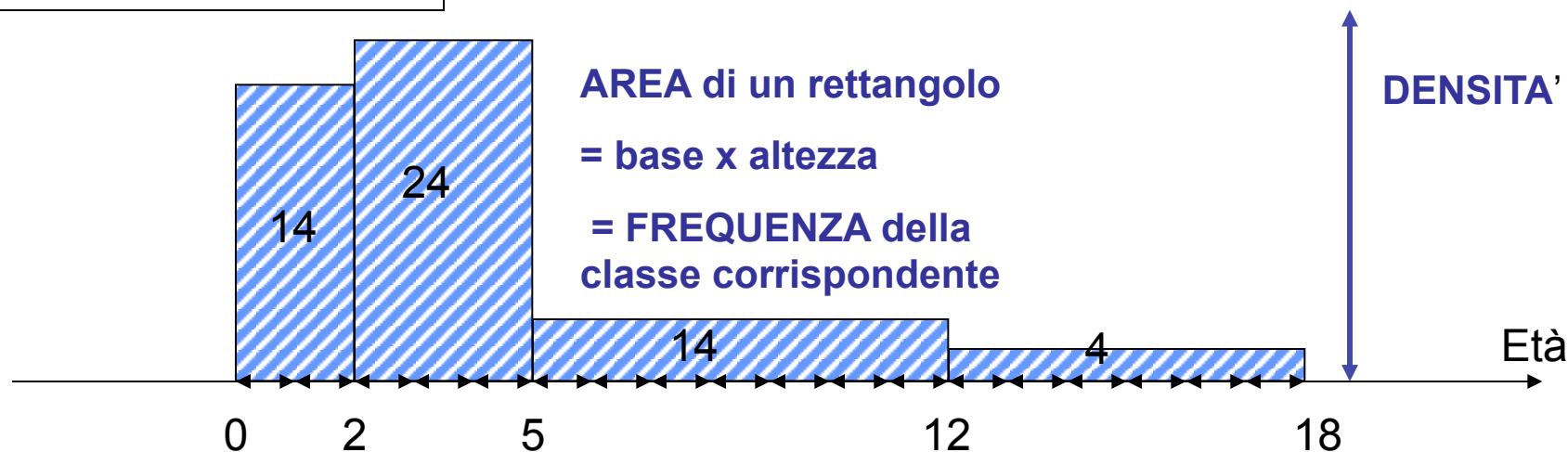
L'istogramma: il grafico della densità

Età	freq.	%
0 - 2	14	25.0
2 - 5	24	42.9
5 - 12	14	25.0
12 - 18	4	7.1
	56	100

Le classi vengono riportate sulle ascisse

ampiezza	densità
$2 - 0 = 2$	$14 / 2 = 7.0$
$5 - 2 = 3$	$24 / 3 = 8.0$
$12 - 5 = 7$	$14 / 7 = 2.0$
$18 - 12 = 6$	$4 / 6 = 0.7$

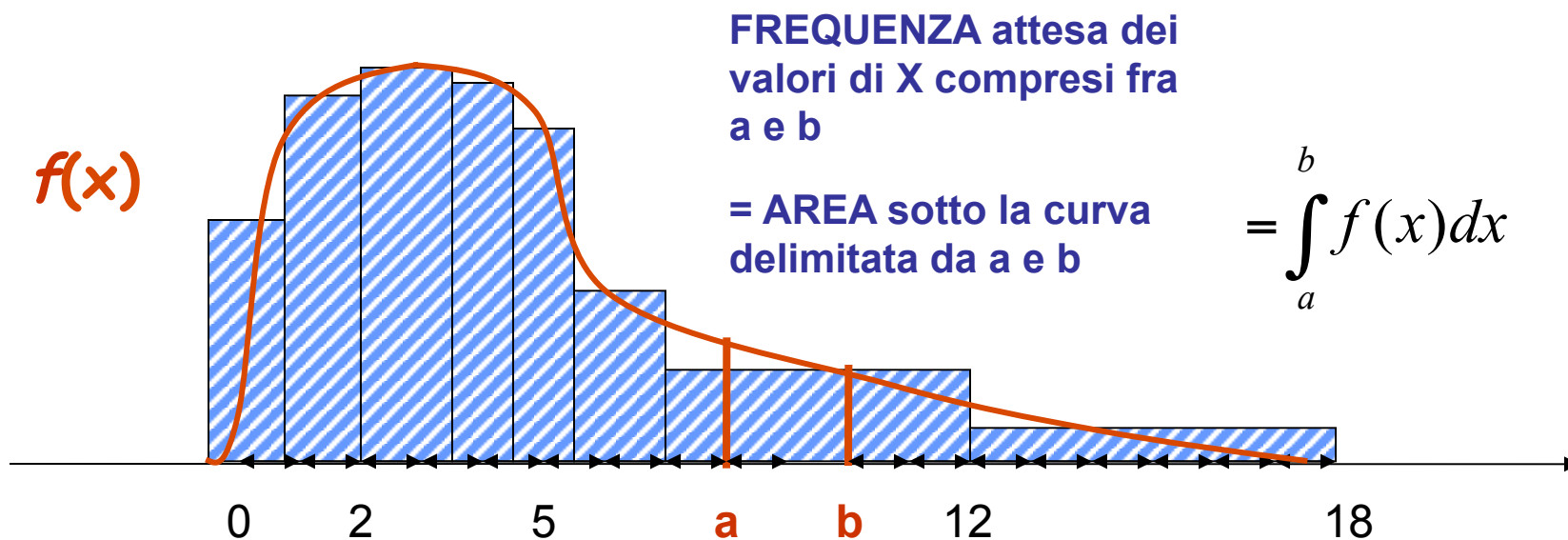
Le densità vengono poste in ordinata



Curve teoriche di densità

Se immaginiamo di fare un istogramma con intervallini piccolissimi, e di unire i punti medi delle colonne, otteniamo un grafico dato da una curva continua.

La matematica fornisce equazioni di curve continue che possono essere interpretate come curve di densità teoriche, corrispondenti a distribuzioni “ideali” di fenomeni quantitativi di interesse X.

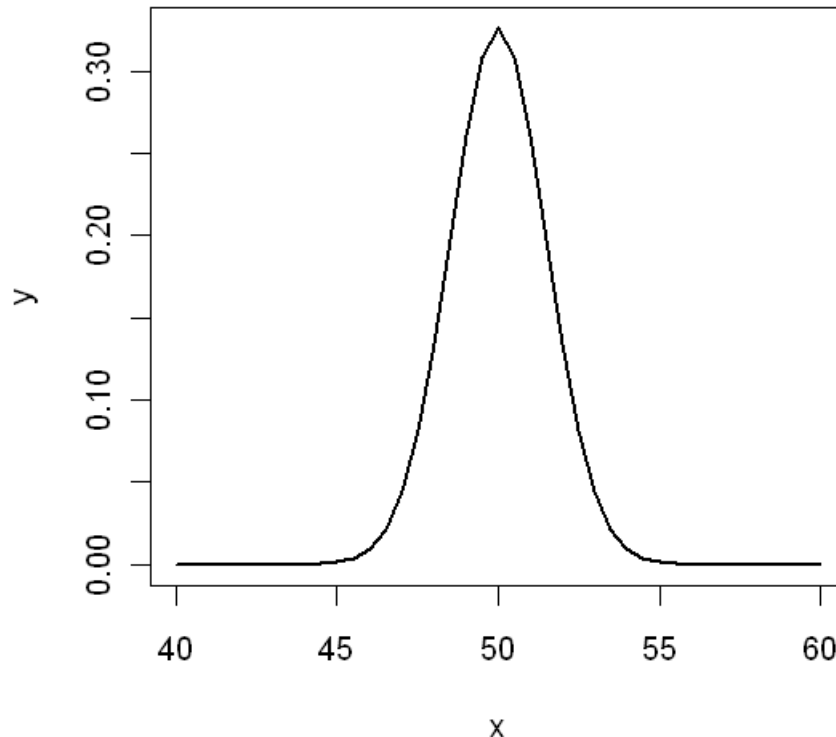


(vd. la curva Normale)

La curva Normale (i)

Un modello per la variabilità biologica / per gli “errori”

La principale **curva di densità teorica** è la Normale (o Gaussiana), che descrive l'andamento di quei fenomeni misurabili come caratteri continui che dipendono “dal caso”, come gli errori di misurazione. E' infatti simmetrica e ha una forma a campana.



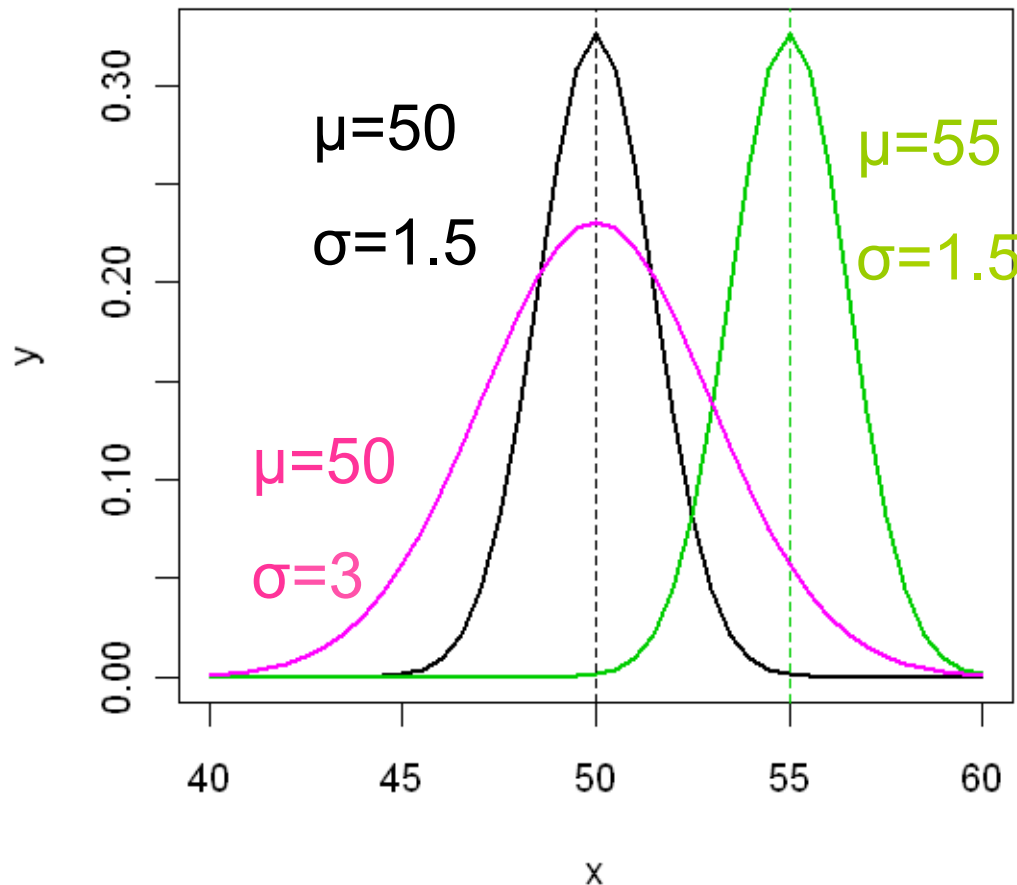
$$y = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x - \mu)^2}{2\sigma^2}\right\}$$

Es: distribuzione dei risultati della misurazione ripetuta del peso di un paziente di 50 kg

La curva Normale (ii)

Un modello per la variabilità biologica / per gli “errori”

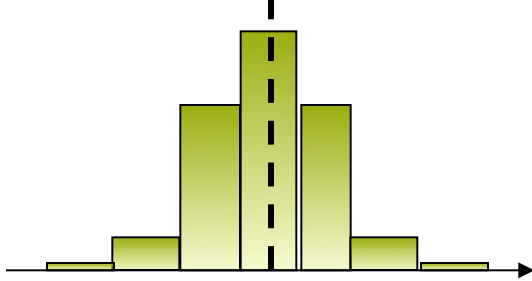
La formula che descrive la curva contiene 2 parametri μ e σ , che determinano rispettivamente dove si posiziona la curva rispetto all'asse x e quanto è ampia la campana



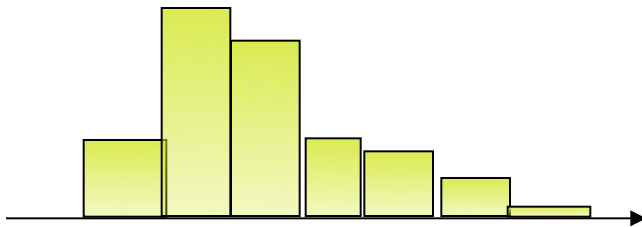
$$y = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x - \mu)^2}{2\sigma^2}\right\}$$

Varie forme della distribuzione

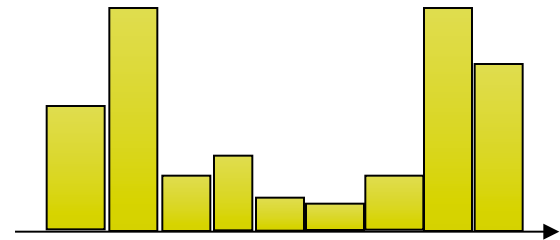
Distribuzioni **SIMMETRICHE**: la massa di densità si dispone in parti “uguali” rispetto ad un immaginario asse (“di simmetria”)



La forma “a campana” è tipica di fenomeni che possano essere ricondotti agli effetti “del caso”, come l’altezza degli individui

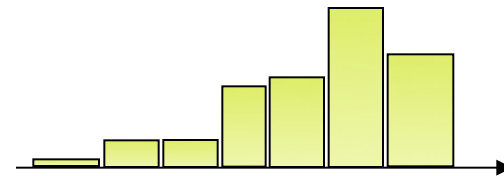


La distribuzione **ASIMMETRICA a destra** è tipica di molti fenomeni biologici, ad es. per i caratteri a valori positivi che possono assumere valori molto alti, ma non molto bassi, come il peso corporeo, il valore dei WBC, etc



Distribuzione **BIMODALE**, cioè con la densità concentrata in due masse.

Spesso è indice fenomeno che è diverso in due sotto-popolazioni, es: altezza delle Femmine e dei Maschi

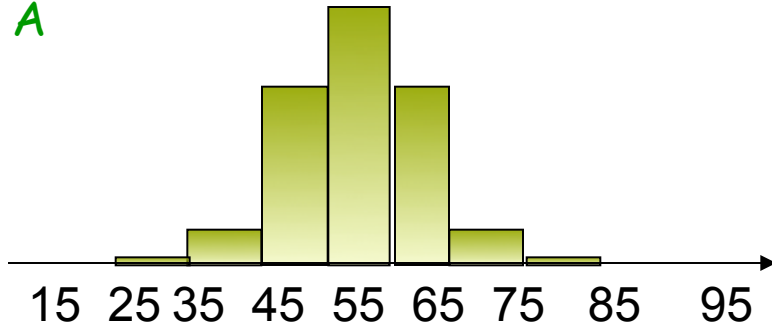


Nella distribuzione Asimmetrica a sinistra, rispetto a un ipotetico asse di simmetria, vi è una massa di densità nella coda sinistra, su valori bassi

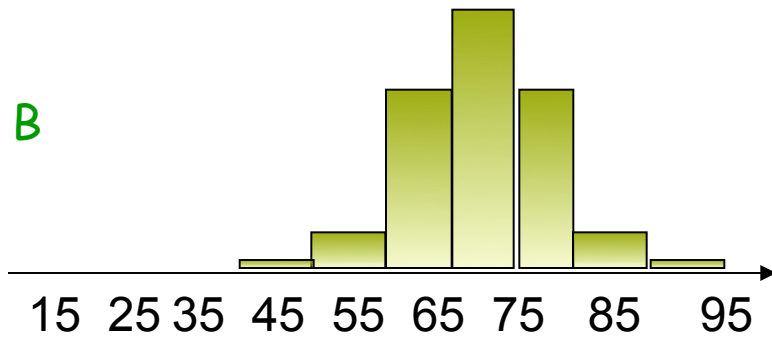
Sintesi di caratteri quantitativi

Distribuzione dell'ETA' ALLA DIAGNOSI in 3 popolazioni diverse (es: pazienti affetti da 3 diverse malattie)

A



B



C

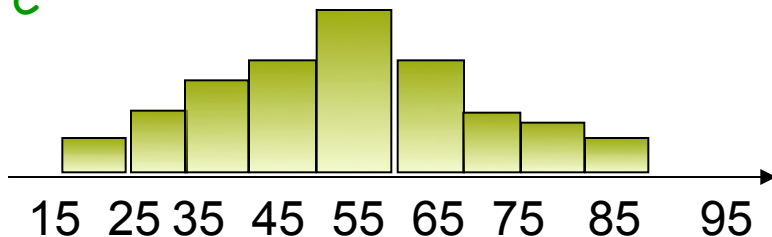


Tabelle e grafici di frequenza forniscono una rappresentazione completa dei dati.

Gli indici statistici servono a fornire delle sintesi di alcuni aspetti delle distribuzioni.

I due aspetti essenziali sono:

→ La **posizione** del carattere sull'asse, eventualmente indicando un valore che sia rappresentativo di tutti gli altri

→ La **variabilità** del carattere, ossia se le osservazioni sono omogenee, simili fra loro, oppure tendono a essere eterogenee, disperse

La media aritmetica

La media aritmetica è una delle **sintesi di posizione** più importanti

La media è l'ammontare totale del carattere (somma di tutte le osservazioni) ripartito in parti uguali

➔ La media, sostituita a ciascuna osservazione, ricostituisce la somma totale delle modalità

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

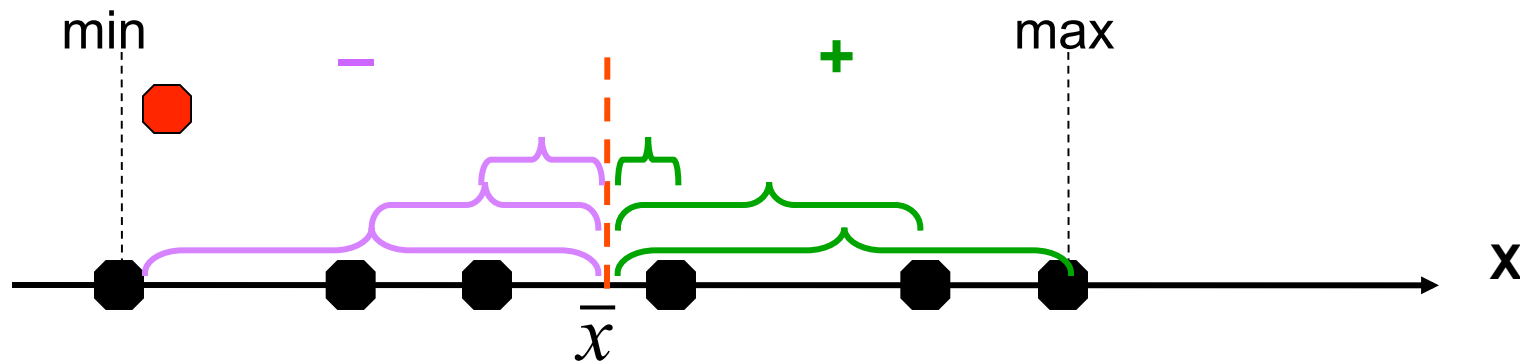
$$\bar{x} = \frac{\sum x_i}{n} \Leftrightarrow \sum x_i = n\bar{x}$$

Voto	
26	24
18	24
28	24
72	72

$$\text{Media} = 72 / 3 = 24$$

➤ Una serie di proprietà illustrano che il comportamento della media aritmetica è quello di un **baricentro**: si colloca al centro delle osservazioni, per questo le “rappresenta”, ne è una sintesi efficace

Principali proprietà della media



$$\min(x_i) \leq \bar{x} \leq \max(x_i)$$

La media è interna al range, ossia, è sempre compresa fra l'osservazione più bassa e quella più alta

$$\sum_{i=1}^n (x_i - \bar{x}) = 0$$

La somma degli **scarti dalla media** è nulla: ossia, la media si colloca “al centro” dei valori osservati, bilanciando scarti positivi e scarti negativi

$$\text{dist} = \sqrt{\sum_{i=1}^n (x_i - C)^2}$$

Se misuriamo la distanza delle osservazioni da un valore C secondo questa misura globale, essa assume il minimo se C è la media aritmetica: ossia, la media aritmetica è il punto “globalmente meno distante” dalle osservazioni

(Altre medie (quadratica; geometrica; armonica) godono di altre proprietà, ma sono meno utili: le trascuriamo)

Media ponderata (1)

Caso particolare: la media di 2 medie

In presenza di 2 gruppi di cui conosciamo numerosità e media aritmetica, possiamo calcolare la media globale:

gruppo	n.ro casi	media
1	n_1	$\bar{x}_1$
2	n_2	$\bar{x}_2$
tot	$n = n_1 + n_2$	

Conosciamo la numerosità totale; ricostituiamo l'ammontare totale dagli ammontari dei due gruppi, usando la relazione fra ammontare e media:

$$\bar{x} = \frac{\sum x_i}{n} \Leftrightarrow \sum x_i = n\bar{x}$$

$$\bar{x} = \frac{\sum_{tutti} x_i}{n} = \frac{\sum_{gr1} x_i + \sum_{gr2} x_i}{n_1 + n_2} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$$

→ La media complessiva non è la media semplice fra le due medie!!

→ Bisogna tener conto delle diverse numerosità, che vanno a fare da “peso” (“ponderazione”)

Media ponderata (2)

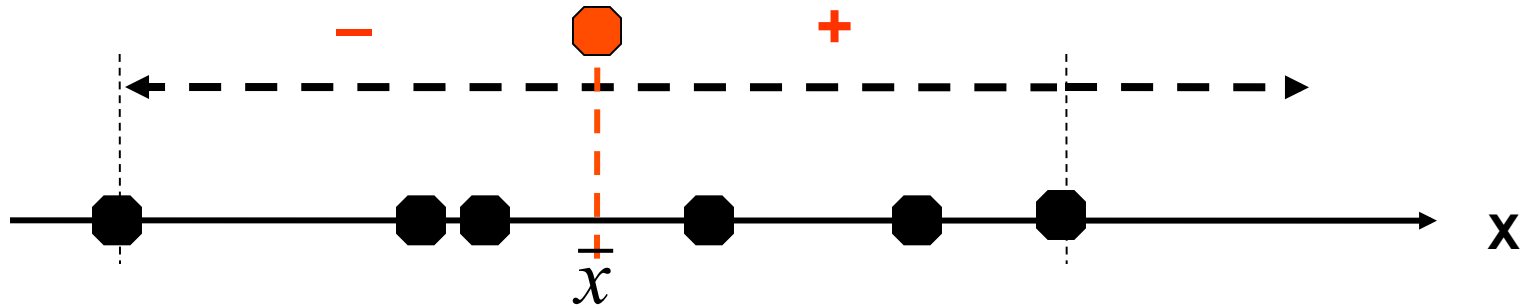
Naturalmente la formula vale anche nel caso di calcolo della media di K medie:

$$\bar{x} = \frac{\sum_{j=1}^K \bar{x}_j n_j}{\sum_{j=1}^K n_j}$$

L'idea si può generalizzare: si può fare la media di K oggetti assegnando a ciascuno un "peso" p_i

$$\bar{x}_P = \frac{\sum_{j=1}^K x_j p_j}{\sum_{j=1}^k p_j}$$

Limitazioni della media aritmetica

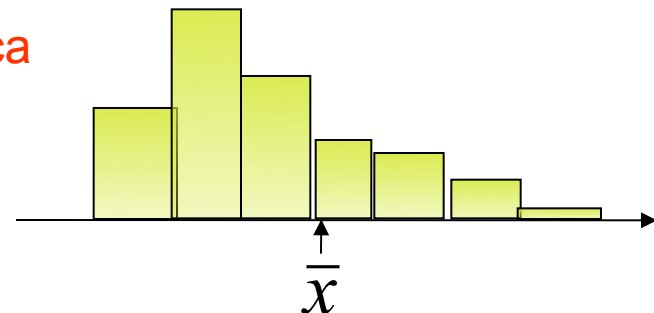


Dovendo BILANCIARE scarti positivi e negativi, e collocarsi nel centro (rispetto ai valori), la media è influenzata dai valori molto alti e dai valori molto bassi ➔

Se questi si spostano ancora più verso “l'esterno”, la media li segue: è attratta dai VALORI ESTREMI

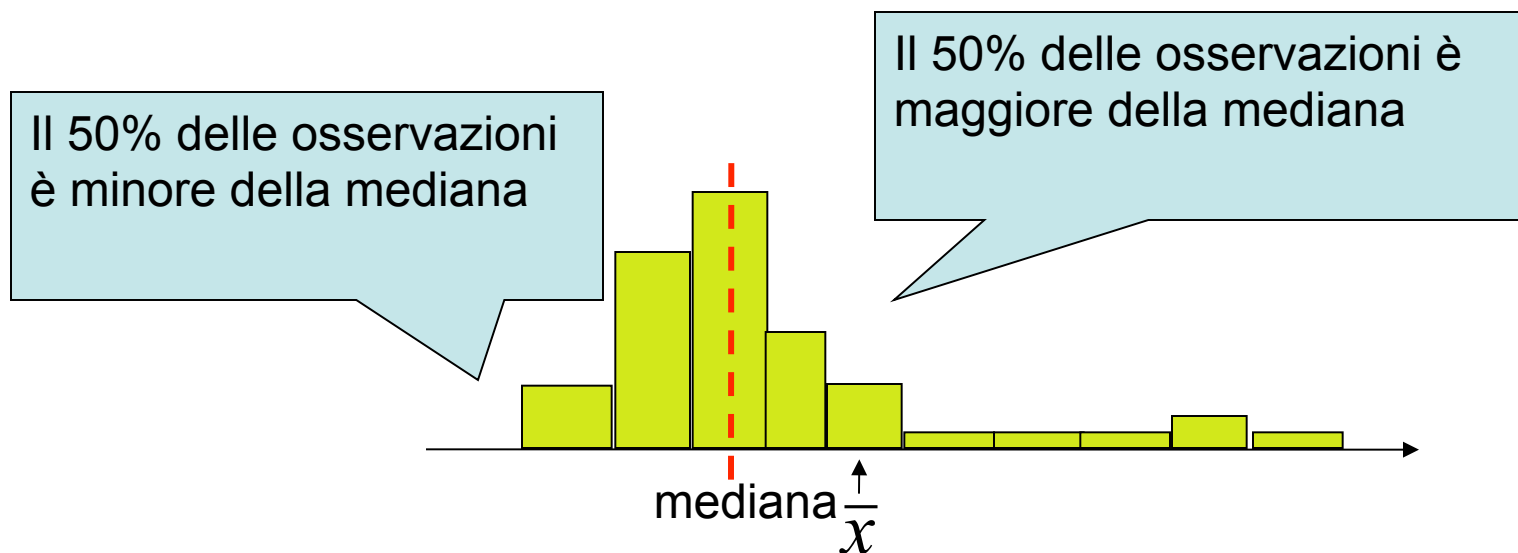
➔ La media aritmetica è una sintesi insoddisfacente della distribuzione:

- Quando si hanno uno o più valori estremi molto anomali
- Quando la distribuzione è asimmetrica



La mediana

- La media aritmetica è una sintesi insoddisfacente della distribuzione:
 - Quando la distribuzione è (molto) asimmetrica
 - Quando si hanno uno o più valori estremi molto anomali
- In questi casi è più rappresentativa la mediana: il valore x tale che la metà delle osservazioni è $< x$ (e l'altra metà è $> x$)

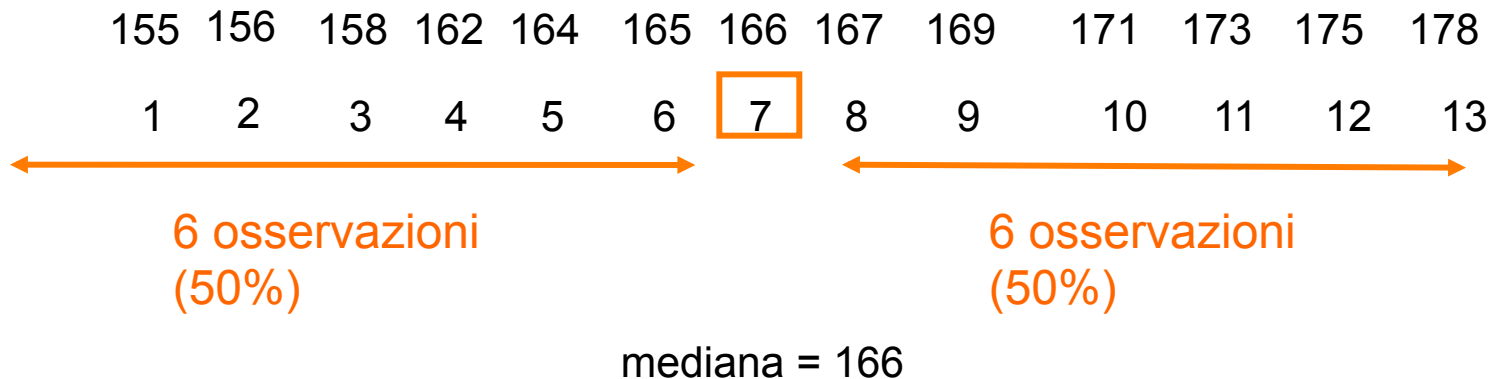


La mediana

Esempio: In un campione di 13 soggetti viene osservato il carattere Altezza (cm):

173 155 162 165 167 175 171 169 164 178 156 158 166

→ Ordiniamo in senso crescente le osservazioni, attribuendogli la posizione in graduatoria (**RANGO**):



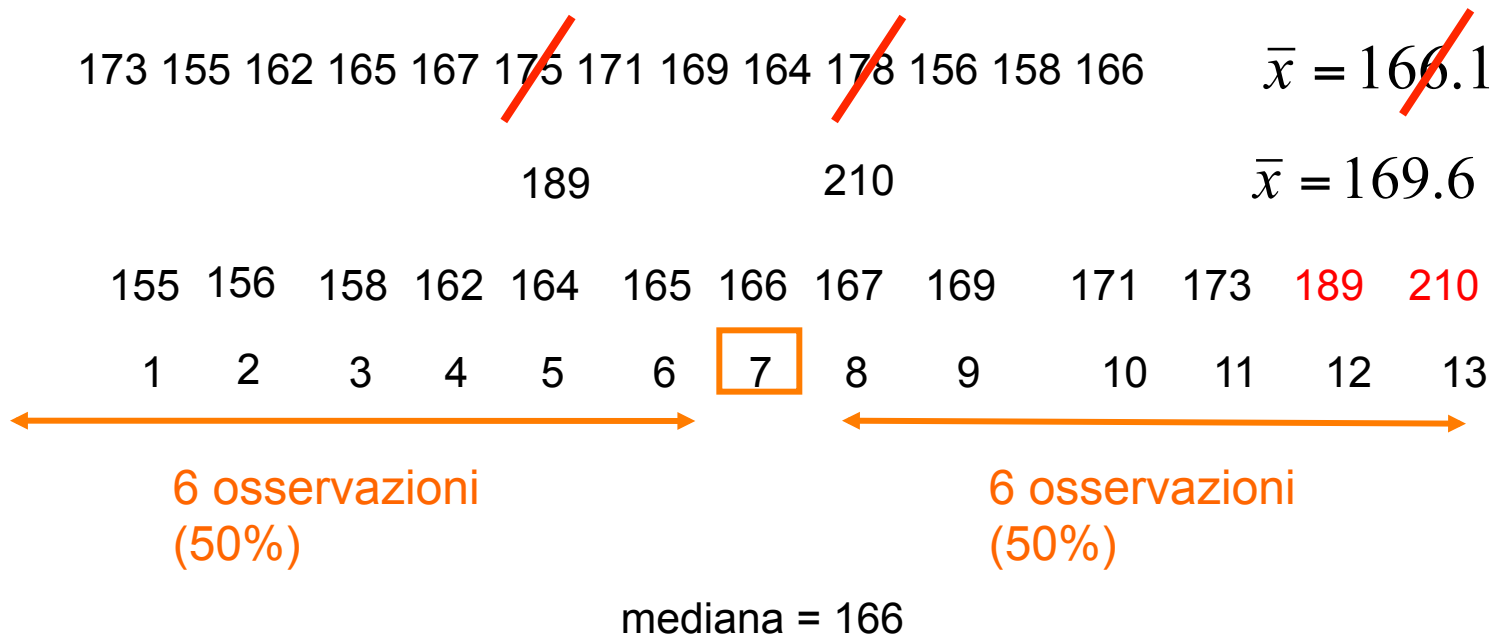
n pari → mediana = modalità intermedia fra quelle di posto $n/2$ e $n/2+1$ (ad esempio, se $n=6$, è la modalità centrale fra la 3° e la 4°)

n dispari → mediana = modalità di posto $(n+1)/2$

Robustezza della mediana

La mediana non cambia o cambia di poco (è “robusta”) in presenza di alcuni dati molto estremi (ad es. con alcuni valori molto alti rispetto agli altri)

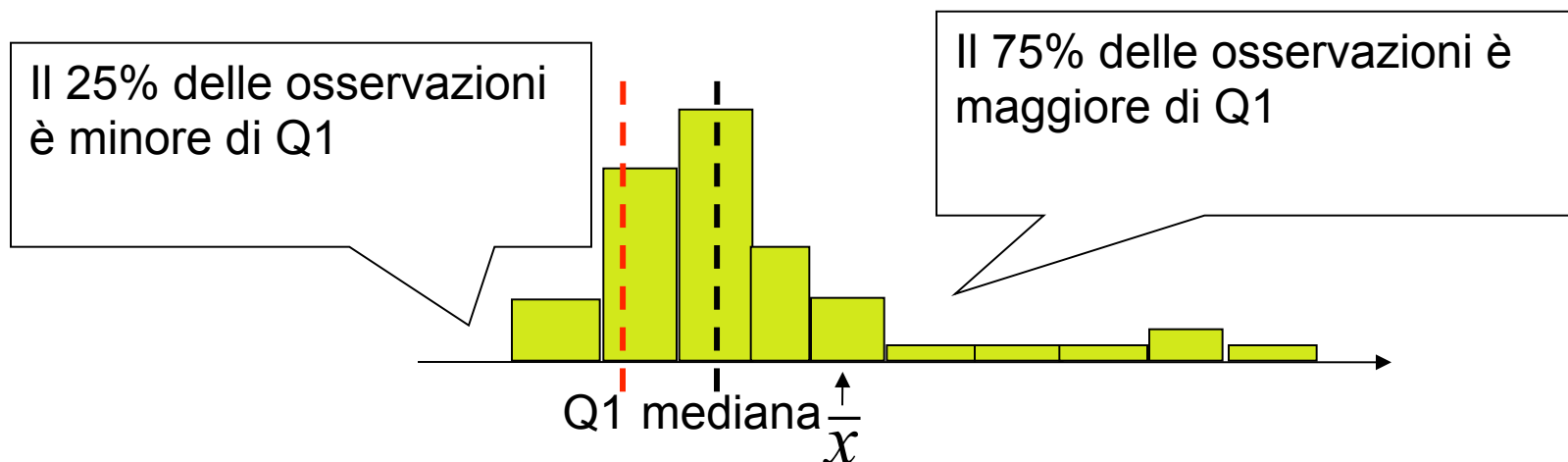
Vediamo per esempio che succede se nel campione precedente i due soggetti più alti sono ancora più alti:



→ La mediana non cambia poichè l'ordinamento delle prime n osservazioni non cambia (invece la media cambia perché l'ammontare totale cambia)

Generalizzazione della mediana: quantili

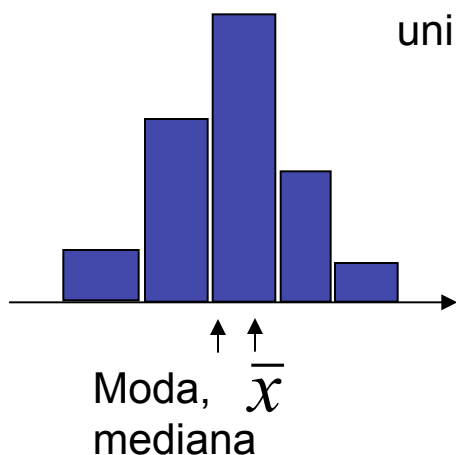
- La mediana separa la distribuzione in due parti, ognuna comprendente il 50% delle osservazioni
- I quantili separano la distribuzione ad altre frazioni percentuali, ad esempio:
 - Il 1° quartile (Q1) separa il primo 25% dal restante 75%
 - Il 3° quartile (Q3) separa il primo 75% dal restante 25%
 - Il 1° decile separa il primo 10% dal restante 90%
 - Il 95° percentile è tale che solo il 5% ha un valore superiore a esso
 - etc.



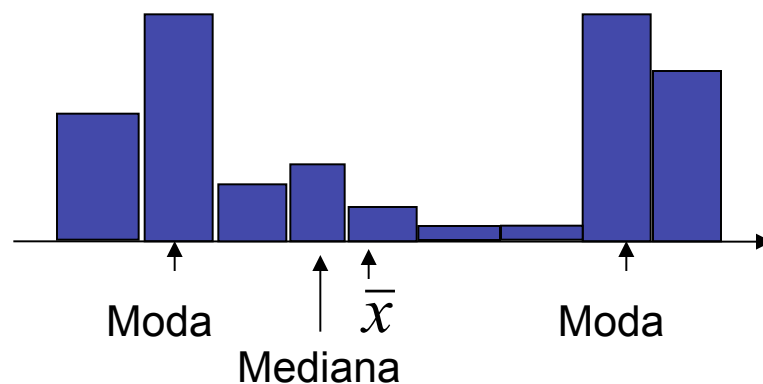
Nota: la mediana e tutti i quantili possono essere calcolati anche per caratteri QUALITATIVI ORDINATI

Forma della distribuzione e indici

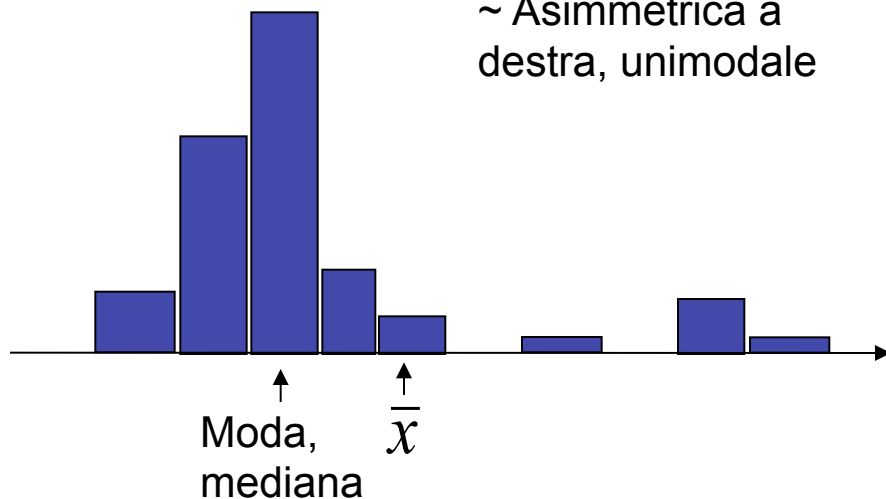
~ Simmetrica,
unimodale



~ Simmetrica, bimodale
(2 sottopopolazioni?)

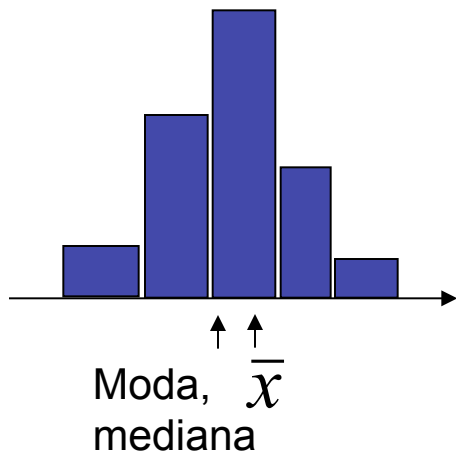


~ Asimmetrica a
destra, unimodale

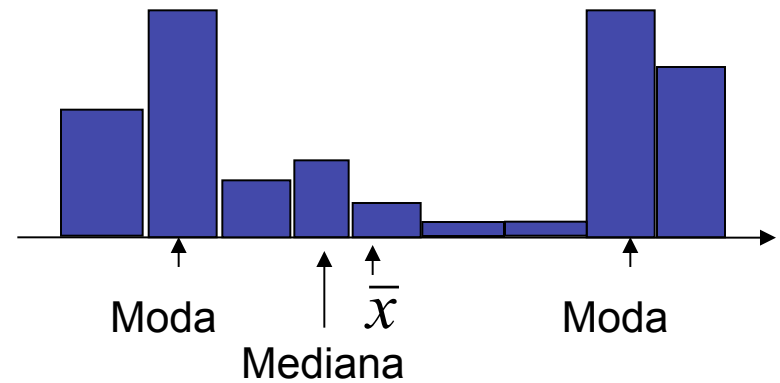


La forma della distribuzione è individuabile (in maniera grossolana) a partire dagli indici sintetici – e viceversa.

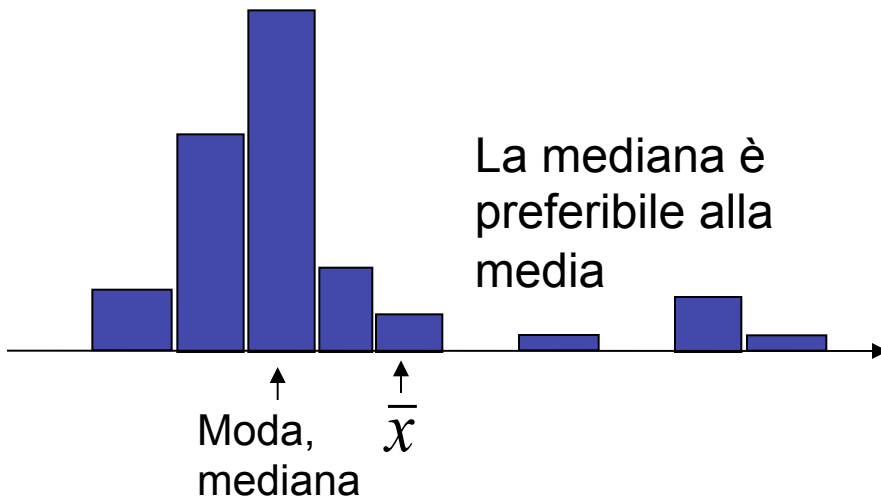
Appropriatezza degli indici



La media è una sintesi soddisfacente, tende a coincidere con la mediana, e con la moda

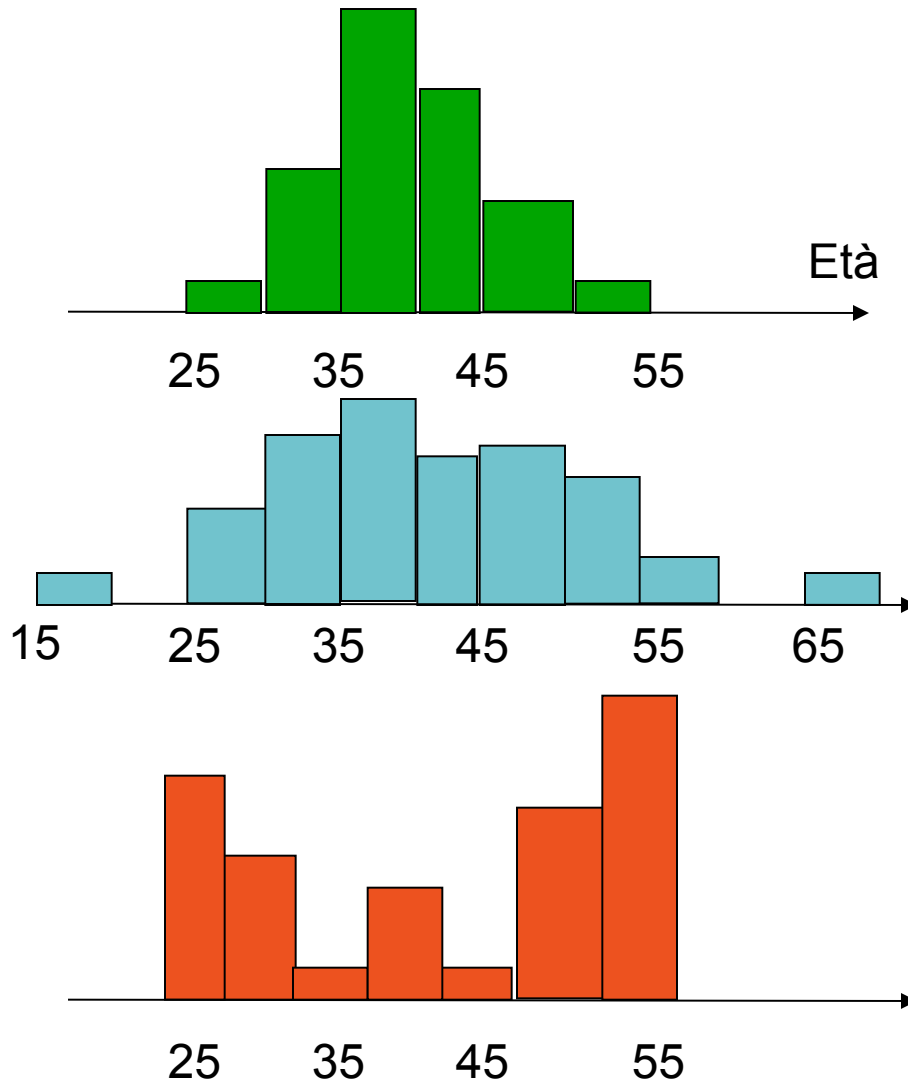


E' opportuno rimarcare la bimodalità: ne' media ne' mediana sono sintesi soddisfacenti



La mediana è preferibile alla media

Misurare la variabilità dalle distanze dalla media $(x_i - \bar{x})$



Queste 3 distribuzioni sono simmetriche, hanno la stessa media aritmetica = mediana = 38 anni

Qui, la maggior parte delle osservazioni è vicina alla media, ci sono pochi ventenni e non ci sono anziani

Qui ci sono tanti soggetti in ciascuna classe, anche alcuni molto giovani o molto anziani: molte osservazioni sono lontane dalla media

Qui ci sono pochi soggetti nelle classi centrali, e molti nelle classi dei giovani e degli anziani: la maggior parte delle osservazioni è lontana dalla media

La Deviazione Standard

(detta anche Scarto o Scostamento Quadratico Medio)

- La deviazione standard rappresenta la **distanza media fra tutte le osservazioni e la media**

$$std = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

Prese le distanze fra ogni osservazione e la media (“scarti”), se ne fa una media non aritmetica - *quadratica*

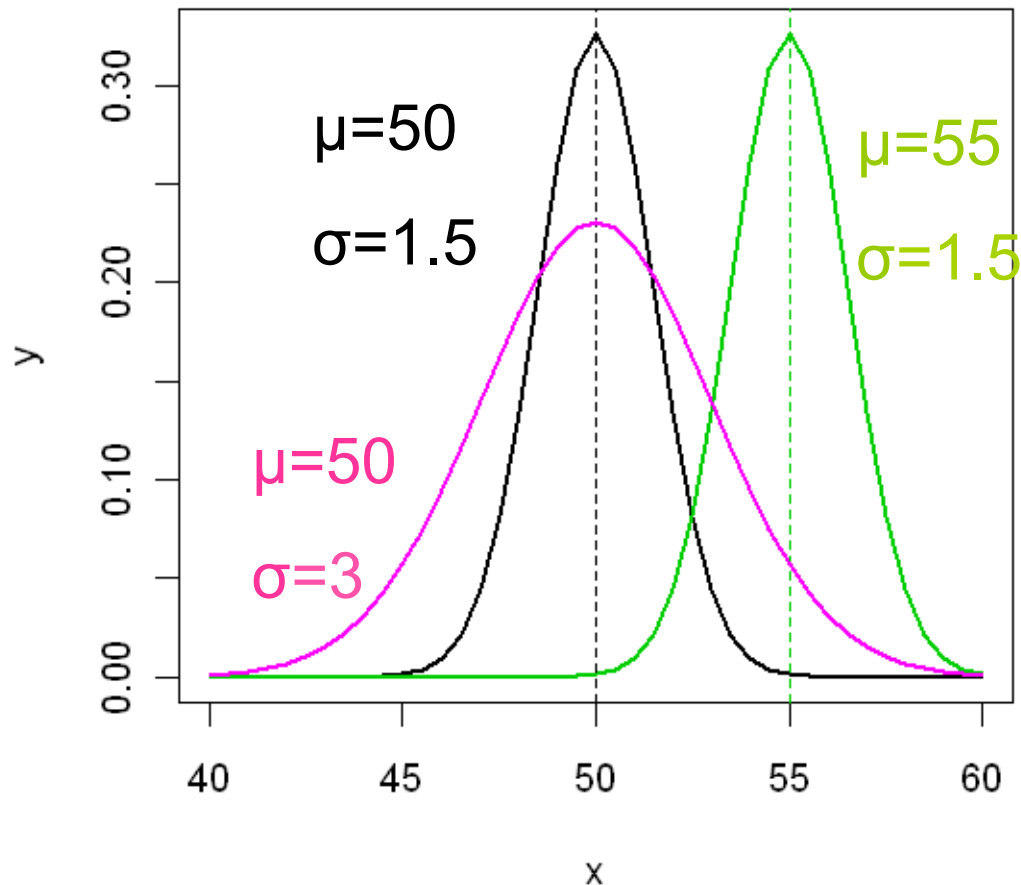
Nota: al denominatore si mette (n-1) anziché per n per motivi legati ad un concetto (distorsione) che affronteremo nella parte di inferenza

- La deviazione standard è una sorta di “**unità di misura rilevante**” del fenomeno osservato
 - Es. X = peso paziente, std = 4.5kg: è la “distanza rilevante” fra due pazienti (1kg è irrilevante ai fini della descrizione del carattere)
- La quantità sotto radice (ossia, il valore elevato al quadrato) è detta **VARIANZA** ed è anch’essa una misura di variabilità

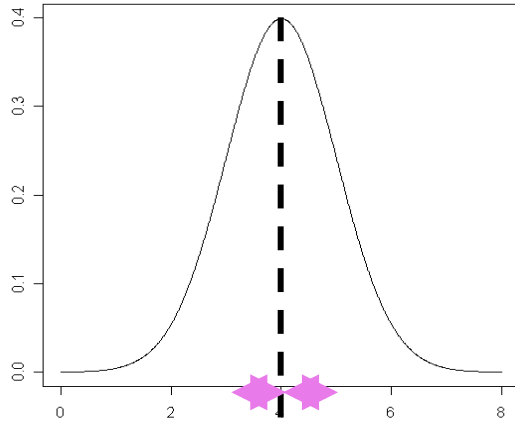
La curva Normale (ii)

I parametri μ e σ

→ μ , che posiziona l'asse di simmetria, ed è interpretabile come **valore medio** →
 σ , che determina l'ampiezza della campana, ossia la **dispersione** di X, e
coincide con la **deviazione standard**



Proprietà della Normale



L'**area** compresa sotto la curva nei seguenti intervalli = la **frequenza** dei valori di X compresi in quegli intervalli è circa(*):

$$(\mu - \sigma, \mu + \sigma)$$

68%

$$(\mu - 2\sigma, \mu + 2\sigma)$$

95%

$$(\mu - 3\sigma, \mu + 3\sigma)$$

99.7%

Mediana=Media= μ . I due quartili Q1 e Q3 si trovano a distanza 0.67σ dalla media:

$$Q_1 = \mu - 0.67 \cdot \sigma$$

$$Q_3 = \mu + 0.67 \cdot \sigma$$

(*) vd. la parte di Probabilità

Coefficiente di variazione

- Il CV è una misura relativa di variabilità: **esprime la variabilità in proporzione alla dimensione media del carattere**; inoltre, è un numero senza unità di misura
- è quindi una misura adatta a **confrontare** la variabilità fra popolazioni diverse, e anche fra caratteri diversi

$$CV = \frac{std}{\bar{x}} \cdot 100$$

Rapporto fra deviazione standard e media aritmetica (espresso in %)

Peso neonato: media = 3.2 kg, std = 0.5 kg

Altezza neonato: media = 51 cm, std = 3.5 cm

Peso Madre: media = 64 kg, std = 4.5 kg

→ I neonati sono più variabili rispetto al peso o all'altezza?

→ Il peso è più variabile nei neonati o nelle madri?

Peso: $CV = (0.5 \text{ kg} / 3.2 \text{ kg}) \cdot 100 = 15.6$

Altezza: $CV = (3.5 \text{ cm} / 51 \text{ cm}) = 6.9$

Peso Madre: $CV = (4.5 \text{ kg} / 64 \text{ kg}) = 7.0$

→ I neonati sono più variabili rispetto al peso che all'altezza (circa il doppio) e in termini di peso sono variabili del doppio anche rispetto alle madri

Gerarchia dei caratteri (2)

Carattere	Sintesi possibili
Qualitativo sconnesso	Moda
Qualitativo ordinato	Moda Mediana
Quantitativo	Se in classi: Classe Modale e Classe Mediana Mediana (e altri quantili) Media aritmetica (e altre medie) Deviazione standard e Coefficiente di Variazione

Relazione fra 2 caratteri qualitativi

(overview)

Es: c'è relazione fra Sesso e Fumo? I Maschi fumano quanto le Femmine?

Raccogliamo dei DATI e organizziamoli in una **tabella di frequenza doppia**.

Usiamo i “profili riga” e “profili colonna” (distribuzioni condizionate) per **capire com'è la relazione** (es. i M fumano di più delle F: 64% vs. 41%). Questo indica che nel campione il Fumo dipende dal (è statisticamente associato al) Sesso.

Calcoliamo una **misura sintetica del grado di associazione osservato** (indice **Chi-Quadrato**).

Tabelle doppie

X = Sesso	Y = Fumo			
	no	si	totale	
M	24	42	66	46%
F	45	31	76	54%
totale	69	73	142	

$$69/142=49\%$$

$$73/142=51\%$$

L'ultima riga rappresenta la distribuzione del carattere Y, senza tener conto di X

Es I
Fumatori
sono il
51% del
totale

Le celle centrali presentano le frequenze delle combinazioni dei 2 caratteri

L'ultima colonna rappresenta la distribuzione del carattere X, senza tener conto di Y



Distribuzioni marginali



Distribuzioni condizionate

(profili riga e profili colonna)

X = Sesso	Y = Fumo		
	no	si	totale
M	24	42	66
F	45	31	76
totale	69	73	142

Sesso	Y = Fumo		
	no	si	totale

M	$24/66 = 36\%$	$42/66 = 64\%$	
F	$45/76 = 59\%$	$31/76 = 41\%$	

Le % di riga sono quelle calcolate rispetto al totale della riga [restringendosi alle sole unità della riga = condizionandosi ad una delle modalità del carattere X].

Esse indicano come si distribuisce il Fumo rispetto al Sesso. I fumatori fra i M e fra le F sono rispettivamente il 64% e il 41%.

χ^2 : distanza dall'indipendenza perfetta

→ Indice complessivo di distanza fra freq. osservate e freq. attese:

$$\chi^2 = \sum \frac{(Osservata - Attesa)^2}{Attesa}$$

Nel caso di indipendenza perfetta, χ^2 vale 0; cresce al crescere del grado di associazione (dipende dalla numerosità del campione).

Difficile da interpretare, dipende dalla numerosità del campione e quindi può assumere valori molto grandi

→ Indice medio di contingenza:

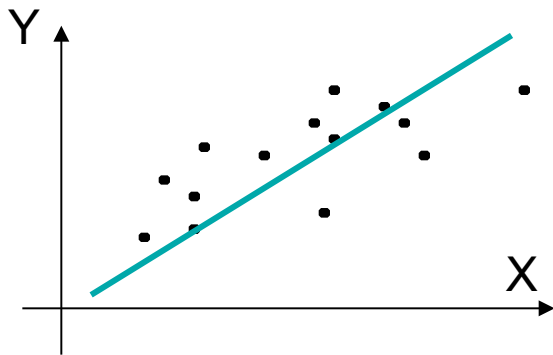
$$\varphi_2 = \chi^2 / [n(num.righe - 1, num.colonne - 1)]$$

Nel caso di indipendenza perfetta, vale 0;

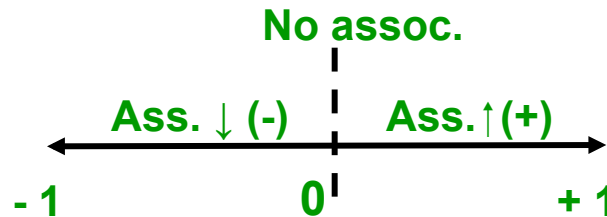
Se il numero di colonne=numero di righe e l'indice vale 1, allora dipendenza perfetta

Associazione fra 2 caratteri continui

- Due caratteri continui X e Y mostrano un grado di associazione se, nella nuvola dei punti che si ottiene su un grafico cartesiano, è possibile riconoscere una tendenza delle osservazioni a distribuirsi secondo una relazione “regolare”, che potrebbe essere rappresentata da una funzione: $Y=f(X)$
- La **retta** è la più semplice relazione funzionale che può rappresentare il modello di associazione fra Y e X. Il grado di associazione LINEARE è misurato dal **coefficiente di correlazione**:



$$r_{xy} = \frac{\text{COV}_{xy}}{\text{std}_x \cdot \text{std}_y}$$

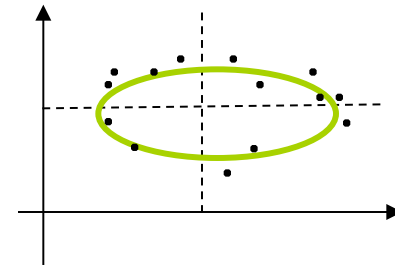
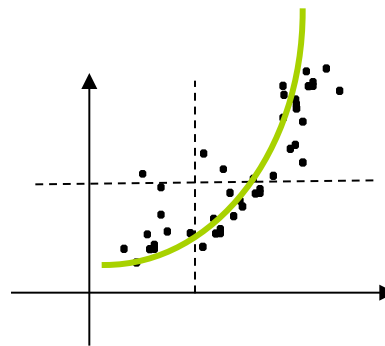
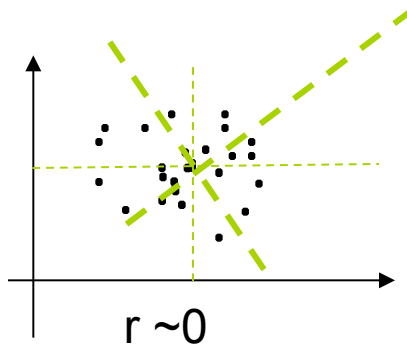
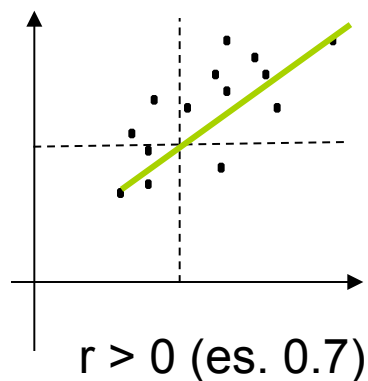


dove al numeratore
c'è la covarianza =

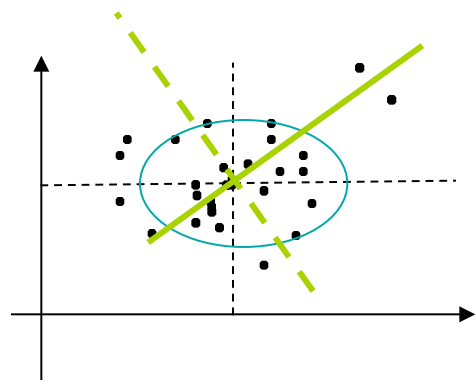
$$\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n}$$

- I valori delle due medie, delle due std e di r determinano anche l'equazione di una **retta di regressione** che passa nella nuvola di punti, fornendo un modello matematico per rappresentare l'effetto di X su Y

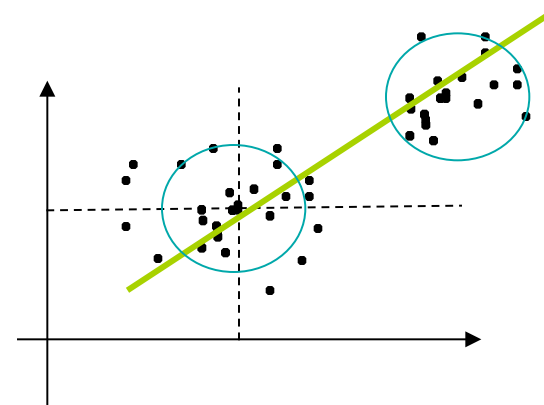
Interpretazione del coeff. di correlazione



r non coglie associazioni non lineari
→ incorrelazione non implica indipendenza



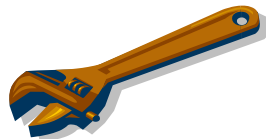
r è affetto da valori estremi, che possono orientare l'ipotetica retta, rendendo apparente una relazione lineare



anche nel caso di presenza di sottopopolazioni r può “leggere” nei dati la presenza di una relazione lineare che non sussiste (vd confondimento)

Appendice

Practicals



- Esercizi elementari di calcolo:
 - Frequenze
 - Media e Mediana
 - Deviazione standard
 - Coeff. di variazione
 - Indice Chi-Quadrato
 - Covarianza (necessaria per calcolare il coefficiente di correlazione lineare)

Regole per l'arrotondamento

- Arrotondare un numero significa ridurre il numero di cifre decimali (quelle dopo “la virgola”, che qui, adottando la convenzione internazionale, rappresentiamo con un punto).

valore originario	→1 decimale	→2 decimali
12.422	12.4	12.42
11.237	11.2	11.24
10.251	10.2	10.25
10.257	10.3	10.26
14.0	14.0	14.00

Se la cifra decimale successiva a quella a cui ci vogliamo fermare è:

<5 → troncare il numero

>5 → aumentare di 1 unità l'ultimo decimale

=5 → guardare alla cifra ancora successiva, e seguire lo stesso criterio



La sommatoria

$$\overbrace{a + a + \dots + a}^{k \text{ volte}} = k \cdot a$$

Somma di k termini tutti uguali fra loro

$$a_1 + a_2 + a_3 + \dots a_i + \dots + a_k = \sum_{i=1}^k a_i$$

Si legge: “sommatoria (o somma) degli a con i per i che va da 1 a k”

Somma di k termini anche diversi fra loro: si usa il simbolo di SOMMATORIA

$$= (a_1 + a_2 + a_3) + (a_4 + \dots + a_k) = \sum_{i=1}^3 a_i + \sum_{i=3}^k a_i$$



Prime sintesi delle tabelle di frequenze

Numero di ricoveri precedenti	n_i
0	4
1	5
2	3
tot	12

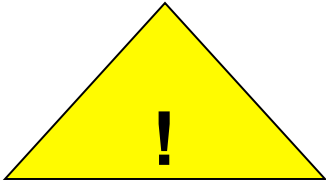
Moda = ?

Numero di ricoveri precedenti	n_i
0	4
1	5
2	3
tot	12

Che percentuale di pazienti ha già avuto almeno un ricovero?

Il 33% non è mai stato ricoverato; quindi il restante 77% ha avuto almeno un ricovero

%
$4/12 \cdot 100 = 33.3$
$5/12 \cdot 100 = 41.7$
$3/12 \cdot 100 = 25.0$



Non confondere le modalità (Ricoveri = 0, 1, 2) con le frequenze
La MODA è “1”, non “5”!

La somma delle percentuali deve fare 100

Calcolare le cumulate SOLO se il carattere è ORDINATO



Media e Mediana di un carattere quantitativo, dati disponibili unità per unità

n = 6 unità

id	Peso (kg)	Rango
p1	54	1
p2	65	4
p3	71	6
p4	55	2
p5	58	3
p6	68	5
	$\Sigma = 371$	

Media =
somma / n

$$= 371 / 6 = 61.8$$

Ordinamento ↑

Rango	1	2	3	4	5	6
Peso	54	55	58	65	68	71
id	p1	p4	p5	p2	p6	p3

Mediana → n pari
Individuare le unità di rango n/2 e n/2+1
Mediana = somma delle loro modalità /2

$$n=6 \rightarrow n/2=3 \quad n/2+1=4$$

$$mediana=(58+65)/2 = 61.5$$



Media e Mediana di un carattere quantitativo, dati disponibili unità per unità (segue)

n = 7 unità

id	Peso (kg)	Rango
p1	54	1
p2	65	4
p3	71	6
p4	55	2
p5	58	3
p6	68	5
p7	80	7
	$\Sigma = 451$	

Media =
somma / n

= 451 / 7 = 64.4

Mediana → n dispari
Individuare l' unità di rango (n+1)/2
Mediana = la sua modalità

n/2=3 n/2+1=4
mediana=(58+65)/2 = 61.5



Media di un carattere quantitativo discreto, dati raggruppati in una tabella di frequenze

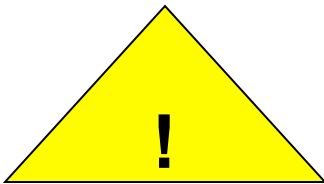
Campione di 8 partorienti, distribuzione del Numero di parti precedenti:

Parti (x_i)	freq. (n_i)	$x_i \cdot n_i$
0	4	0
1	3	3
2	1	2
<i>tot</i>	8	5

Totale Numero di parti = $(0+0+0+0)+(1+1+1)+(2)$

$= 0 \cdot 4 + 1 \cdot 3 + 2 \cdot 1$

Media = $5 / 8 = 0.6$



Non confondere modalità (Parti) e frequenze!

Le unità sono $n=8$, mentre le modalità sono 3.

Occorre ricostruire l'ammontare totale del carattere, e poi dividerlo numero di unità

L'ammontare del carattere corrispondente ad ogni modalità è dato dal prodotto modalità x frequenza

$$\bar{x} = \frac{\sum_{i=1}^k x_i n_i}{n}$$

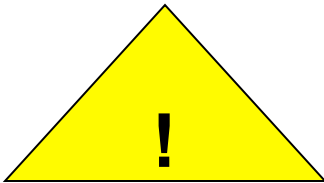


Media di un carattere quantitativo continuo, dati raggruppati in classi

Es: peso corporeo per un campione di 64 pazienti

peso (kg)	freq. (n _i)	x _i	x _i n _i
- 50	4	45	180
50 - 60	17	55	935
60 - 70	24	65	1560
70 - 80	11	75	825
80 -	8	85	680
	64		4180

Media = 4180 / 64 = 65.3



Il principio è sempre quello di ricostituire l’ammontare totale del carattere, e dividerlo per il numero di unità. Il problema è che le modalità sono intervalli di valori del carattere.

Soluzione: assegnare a ciascuna classe un valore rappresentativo – solitamente, il valore centrale

$$x_i = \frac{l_{i-1} + l_i}{2}$$

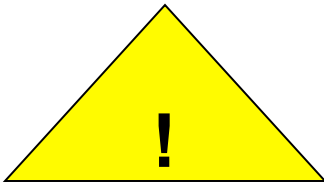
Per le classi aperte si sceglie un valore rappresentativo “plausibile”; la stima della media può cambiare per scelte diverse

Media di due gruppi

Es: Un articolo riporta che il valore medio del colesterolo in un gruppo di 40 uomini è pari a 198 mg/dl, mentre in un gruppo di 16 donne è di 190 mg/dl. Quanto vale la media nella popolazione totale??

	media $\bar{x}$	n.ro casi n	totale
M	198	40	$198 \times 40 = 7920$
F	190	16	$190 \times 16 = 3040$
		56	10960

$$\text{media} = 10960 / 56 = 195.7$$



$$\text{media} \neq (198 + 190) / 2 = 194$$

Ricostituiamo il totale di ciascun gruppo, e lo dividiamo per il totale delle unità

Bisogna fare riferimento al concetto di MEDIA PONDERATA



Mediana / classe mediana per un carattere quantitativo, dati raggruppati

Campione di 18 donne, Numero di gravidanze precedenti

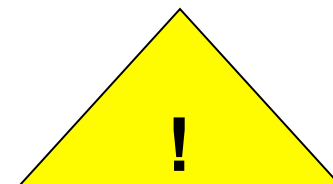
Numero di gravidanze precedenti	n_i
0	6
1	5
2	3
3	3
4	1
tot	18

N_i
6
11
14
17
18

$n/2=9 \rightarrow$ la 9ª unità presenta la modalità “1”

Infatti, con la modalità “0” raggiungiamo solo le prime 6 pazienti, passando a “1” raggiungiamo l’undicesima, e quindi abbiamo già incluso la nona

$\rightarrow$ Mediana=1



Il principio è sempre quello di individuare la modalità di rango $n/2$.

Quindi, occorre calcolare le frequenze cumulate

Nel caso di carattere continuo, si individua la “classe mediana” (allo stesso modo, si parlerà di “classe modale”).

C’è anche una formula per individuare un singolo valore per la mediana: la tralasciamo.



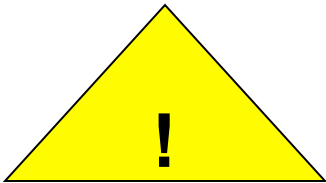
Calcolo della deviazione standard, carattere quantitativo con dati disponibili unità per unità

Età per un campione di 7 pazienti

ETA' x _i	x _i -m	(x _i -m) ²
65	9.29	86.22
35	-20.71	429.08
44	-11.71	137.22
43	-12.71	161.65
71	15.29	233.65
63	7.29	53.08
69	13.29	176.51
media m=55.7	0	1277.43

Ad esempio alla seconda riga:
 (35-55.7) = -20.71 ; (-20.71)² = 161.65
 Varianza = 1277.43 / 6 = 212.90
 std = √212.90 = 14.59

$$\sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$



Attenzione a svolgere le operazioni in ordine:
 Prima si calcolano gli scarti, xi – media;
 Poi ogni scarto viene elevato al quadrato;
 Poi si sommano i quadrati;
 Si divide per (n-1), ottenendo la **VARIANZA**;
 Si estrae la radice quadrata



Calcolo della deviazione standard: formula più rapida

Età per un campione di 7 pazienti

ETA' x_i	$(x_i)^2$
65	4225
35	1225
44	1936
43	1849
71	5041
63	3969
69	4761
	23006

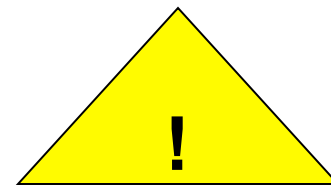
media $m=55.7$

Per il calcolo della varianza:

$$23006 / 7 = 3286.57 \quad 3286.57 - (55.7)^2 = 182.49$$

$$\text{Varianza} = 182.49 \times 7/6 = 212.90$$

$$\text{std} = \sqrt{212.90} = 14.59$$



La VARIANZA si ottiene più rapidamente applicando la seguente formula:

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

Ricordarsi di estrarre la radice quadrata!!

(eventuali discrepanze possono essere dovute all'arrotondamento)



Interpretazione di indici

Es: Per un gruppo di pazienti alla diagnosi di sclerosi multipla:

Media = 36 Dev. St. = 7

Q1=29 Mediana=35 Q3=41

→ La distribuzione era simmetrica o asimmetrica?

→ Poteva avere una forma a campana?

→ Se sì, quale era il range dell'età?

→ La metà dei pazienti aveva meno di ?

→ 1 su 4 aveva meno di ?

→ 1 su 4 aveva più di ?

→ l'età era fortemente variabile?

simmetrica, attorno al valore medio centrale di 36 anni

Sì: nella Normale Q1 e Q3 sono a $36 \pm 0.67 \cdot 7 \approx 31$ e 41

Range: $36 \pm 3 \cdot 7 = (15, 57)$

35 anni

29 anni

41 anni

no: la deviazione standard era il 20% della media (coefficiente di variazione = $7/36 \cdot 100 = 19.4$)

Calcolo dell'indice $\chi^2(i)$

Freq.
osservate

Passo 1: calcolo delle frequenze attese (una per ciascuna cella interna della tabella)

	dolore percepito			
sex	mod.	forte	insopp.	tot
M	2	5	3	10
F	5	4	3	12
tot	7	9	6	22

Freq. attese

$$\tilde{n}_{ij} = \frac{n_{i.} \cdot n_{.j}}{n_{..}}$$

= 12 * 9 / 22

	dolore percepito			
sex	mod.	forte	insopp.	tot
M	3.18	4.09	2.73	10
F	3.82	4.91	3.27	12
tot	7	9	6	22



Calcolo dell'indice $\chi^2_{(ii)}$

Passo 2: calcolo delle distanze fra frequenza osservata e frequenza attesa per ciascuna cella interna della tabella:

$$\frac{(n_{ij} - \tilde{n}_{ij})^2}{\tilde{n}_{ij}}$$

0.44	0.20	0.03
0.37	0.17	0.02

= [(4 – 4.91)²] / 4.91

Passo 3: sommo tutte le distanze:

$\chi^2 = 0.44 + 0.20 + ... = 1.225$

	dolore percepito			
sex	mod.	forte	insopp.	tot
M	2	5	3	10
F	5	4	3	12
tot	7	9	6	22

	dolore percepito			
sex	mod.	forte	insopp.	tot
M	3.18	4.09	2.73	10
F	3.82	4.91	3.27	12
tot	7	9	6	22



Calcolo della covarianza

Formula
generale

$$\text{COV}_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n}$$

1 – calcolo gli
scarti

2 – faccio tutti i
prodotti

3 – li sommo

4 – divido per n

id	Pressione (x)	kcal/die (y)	x-m(x)	y-m(y)	prodotti
1	145	1850	-17.4	-66.7	1163
2	180	1960	17.6	43.3	760.74
3	150	1790	-12.4	-126.7	1576.3
4	155	1680	-7.4	-236.7	1761.9
5	148	2090	-14.4	173.3	-2504
6	160	2000	-2.4	83.3	-203.7
7	175	2100	12.6	183.3	2301.9
8	180	1900	17.6	-16.7	-292.6
9	169	1880	6.6	-36.7	-240.4
somma	1462.0	17250.0	0.0	0.0	4323.3
somma/n	162.4	1916.7	0.0	0.0	480.4

17.6 · 43.3

180 – 162.4
= 17.6

COV_{xy}

Medie di x e di y



Calcolo della covarianza : formula più rapida

Formula
"breve"

$$\text{COV}_{xy} = \frac{\sum_{i=1}^n x_i y_i}{n} - \bar{x} \cdot \bar{y}$$

- 1 – calcolo i prodotti
- 2 – li sommo
- 3 – divido per n
- 4 – sottraggo il prodotto delle medie

id	Pressione (x)	kcal/die (y)	xy
1	145	1850	268250
2	180	1960	352800
3	150	1790	268500
4	155	1680	260400
5	148	2090	309320
6	160	2000	320000
7	175	2100	367500
8	180	1900	342000
9	169	1880	317720
somma	1462.0	17250.0	2806490.0
somma/n	162.4	1916.7	311832.2

Medie di x e di y

150 · 1790

$$\text{COV}_{xy} = 311832.2 - 162.4 \cdot 1916.7$$

