

LAUREA TRIENNALE IN DIETISTICA
A.A. 2010/11

Corso integrato Fisica Statistica e Informatica
Statistica Medica

Simona Iacobelli

2 CFU, 20 ore (?)

Info

LEZIONI: martedì (e giovedì) h 14:00-16:00

RICEVIMENTO: preferibilmente il martedì dopo la lezione; presso il CIBB (Centro Interdipartimentale di Biostatistica e Bioinformatica), Edificio H (Fisica Medica)
Contatti: inviare una e-mail a simona.iacobelli@uniroma2.it

MATERIALE DIDATTICO

- Un testo di riferimento utile è: Lantieri PB, Rizzo D, Ravera G: Statistica medica per le professioni sanitarie, II ed. McGraw-Hill (2004)
- Appunti e stampati delle slides (disponibili in rete)

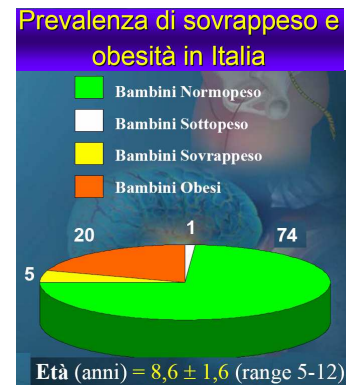
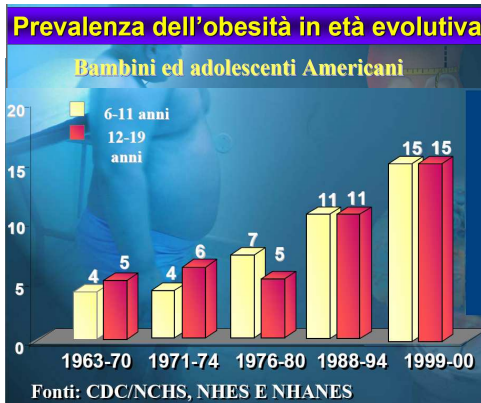
MODALITA' D'ESAME

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

Introduzione

Un po' di statistiche ...

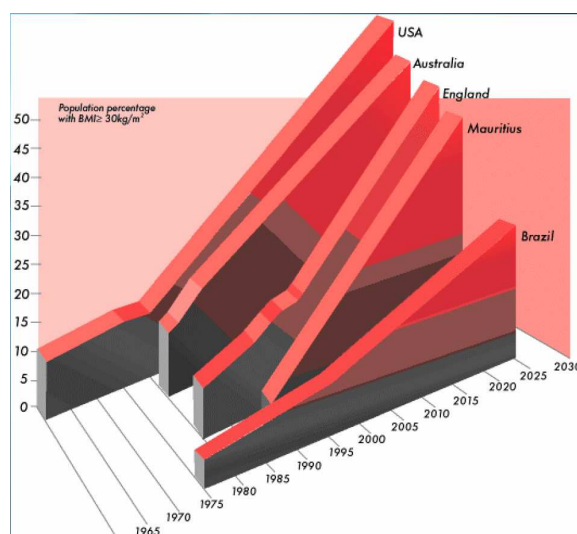
I dati presentati nelle prossime slides sono tratti da un intervento del prof. Del Giudice (II Università Napoli) al convegno della Società Italiana di Pediatria Preventiva e Sociale (2008) sul tema dell'obesità infantile



Introduzione

... Previsioni ...

(o proiezioni? o estrapolazioni?)

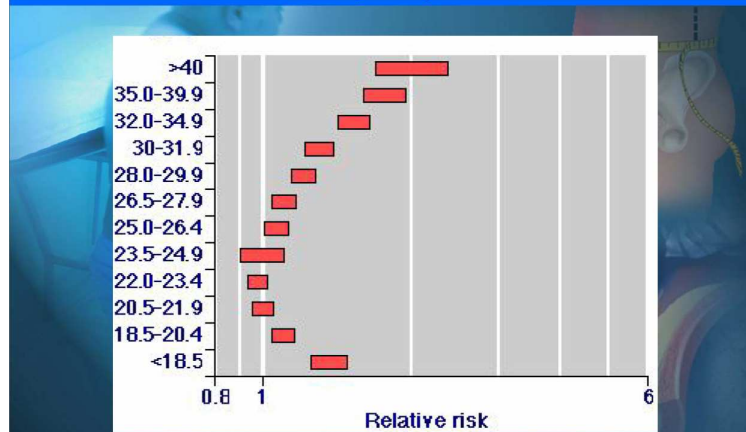


Introduzione

... Oltre le frequenze ...

Quantificazione del rischio di un evento: il Risk Ratio (con l'Intervallo di Confidenza)

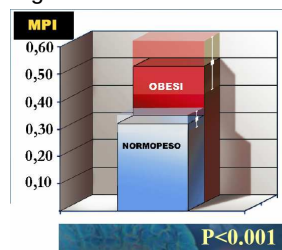
Rischio relativo di mortalità per qualsiasi causa per donne non fumatrici senza alcuna storia di patologia in relazione al BMI



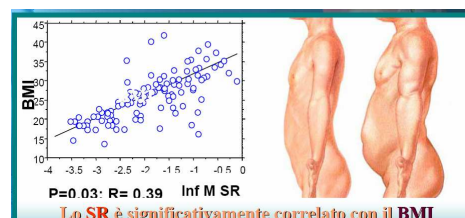
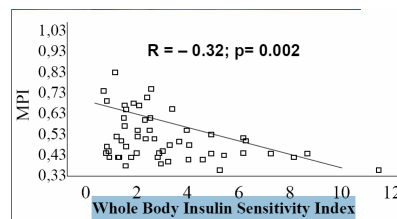
Introduzione

... strumenti per la conoscenza ...

Ancora per lo studio delle relazioni fra fenomeni (qui: fra MPI e obesità; fra MPI e WBISI; fra BMI e SR): modelli di regressione e test di significatività



- MPI: Indice di Performance Miocardica [alto = deterioramento della contrattilità miocardica]
- WBISI: Whole Body Insulin Sensitivity Index [basso = ridotta attività regolatrice dell'insulina]
- SR: Strain Rate, indice di contrattilità miocardica

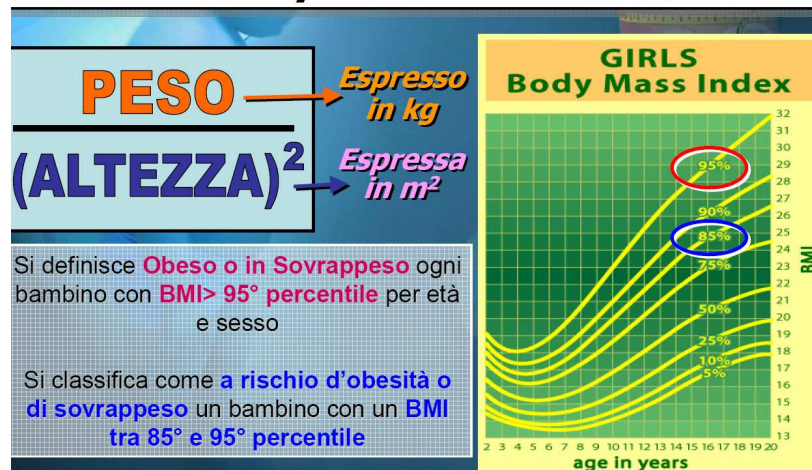


Introduzione

... e per la pratica clinica

Definizione dell'obesità infantile: i quantili

Il Body Mass Index (BMI)



Introduzione

La Statistica

Parole-chiave

- Fenomeni collettivi (fenomeni che presentano variabilità)
- Relazioni fra fenomeni
- Usare dati (osservare)
- Quantificare

Finalità

- Descrivere
- Conoscere / capire
- Prevedere
- Utilizzare / prendere decisioni



Evidence-Based

Medicine / Nursing / Prevention

Fasi di intervento

- Pianificazione degli studi
- Analisi dei dati
- Interpretazione dei risultati
- Comunicazione dei risultati

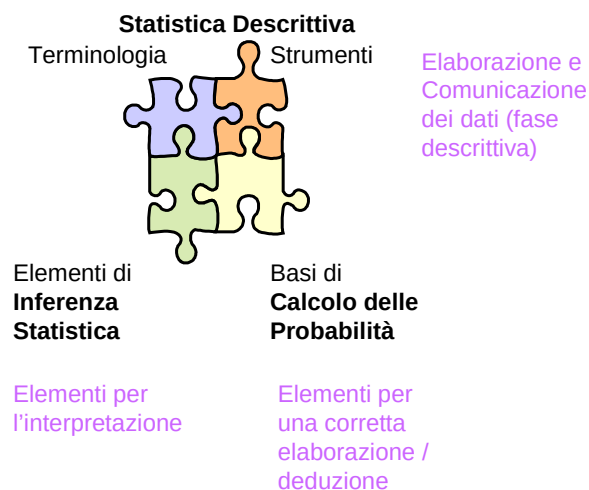
Strumenti

- Ragionamento analitico ("buon senso")
- Matematica (Probabilità)

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



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

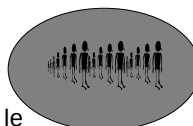
→ Essa è anche detta **POPOLAZIONE OBIETTIVO**

→ 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.

unità	caratteristiche
 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 ...

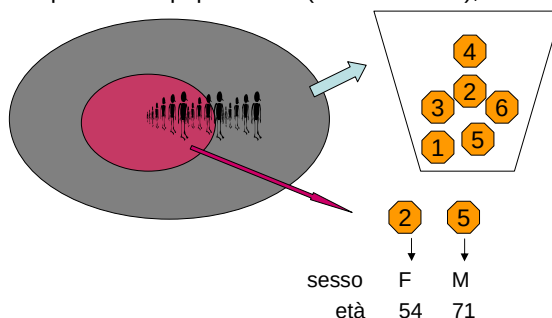
Campione

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	sexso	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
	...						

*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	sexso	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 **MODALITÀ'**
 - 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: $\mathfrak{N}^+$

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

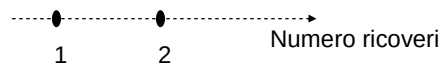
- Un tipo particolare di carattere qualitativo sconnesso è quello BINARIO o DICOTOMICO, cioè che assume 2 sole modalità
- Esso può essere solitamente inteso come indicatore di presenza/assenza di una certa caratteristica
- Corrispondentemente, di solito si usa la codifica numerica 0/1 (0=no=assenza,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

Esempi

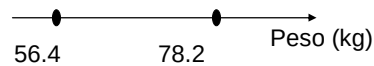
- Fumatore: sì/no
- Rispondente (alla terapia): sì/no
- Sesso = M/F, ovvero:
- Paziente maschio: sì/no

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



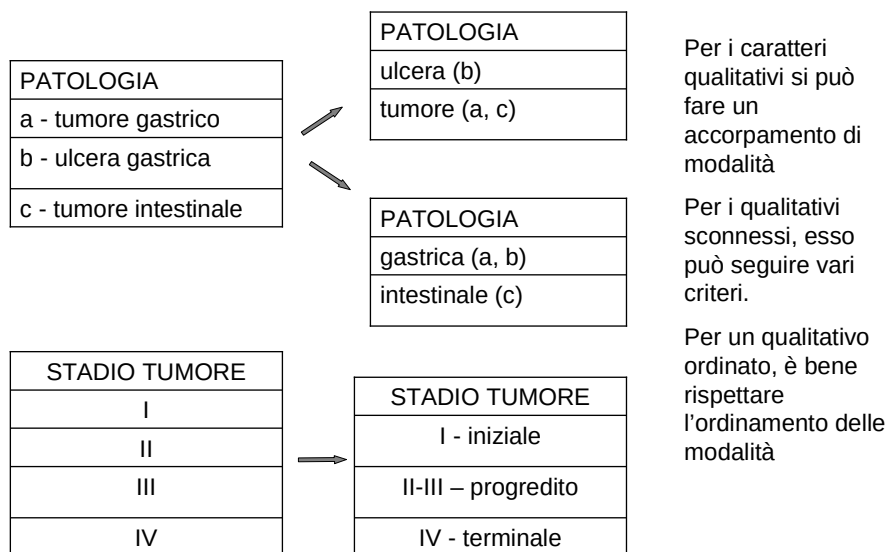
- **CONTINUI**: le modalità NON possono essere enumerate; i valori compresi fra due modalità sono sempre a loro volta delle modalità
 - generalmente ottenuti tramite misurazione



→ 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)



Ricodifica delle variabili (2)

Età
- 25
25 - 45
45 - 65
65 -

$25 < \text{Età} \leq 45$ (25, 45]

Età > 65 (classe aperta)

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

Le modalità quantitative possono essere trasformate mediante operazioni matematiche.

WBC		ln(WBC)
2.2	→	0.788
3.2	→	1.160
1.8	→	0.588
2.1	→	0.742

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:

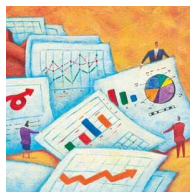
non fumatore	ex-fumatore	fumatore
--------------	-------------	----------

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 freq. relativa freq. percentuale (%)
 totale delle osservazioni nel campione totale = 1 totale = 100

Percentuali: interpretazione e uso (1)

Risposta al trattamento	%
No	42.9
Si	57.1
tot	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
tot	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 (%)	cum	% cum
0	25	40%	25	40%
1	21	33%	46	73%
2	12	19%	58	92%
3	4	6%	62	98%
4 +	1	2%	63	100%
totale	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**

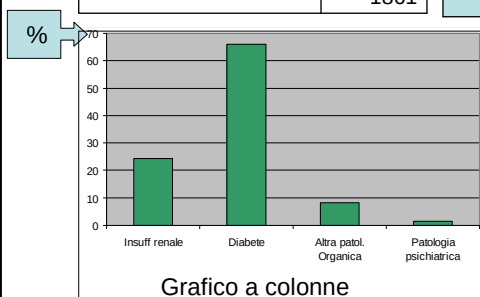
Qui, la moda è la modalità "Diabete".

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



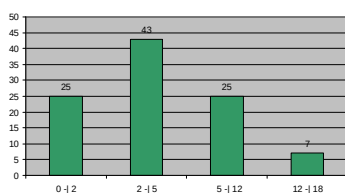
Grafici da tabelle di caratteri continui

Distribuzione di 56 pazienti pediatrici per età

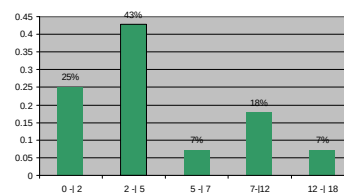
Età	freq.	%
0 - 2	14	25
2 - 5	24	43
5 - 12	14	25
12 - 18	4	7
	56	100

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

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

$$\Leftrightarrow \text{frequenza} = \text{ampiezza} \times \text{densità}$$

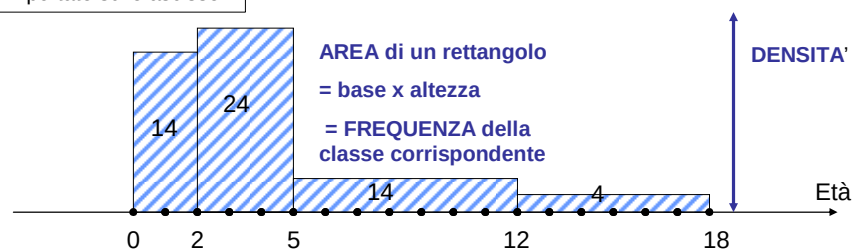
L'istogramma: il grafico della densità

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

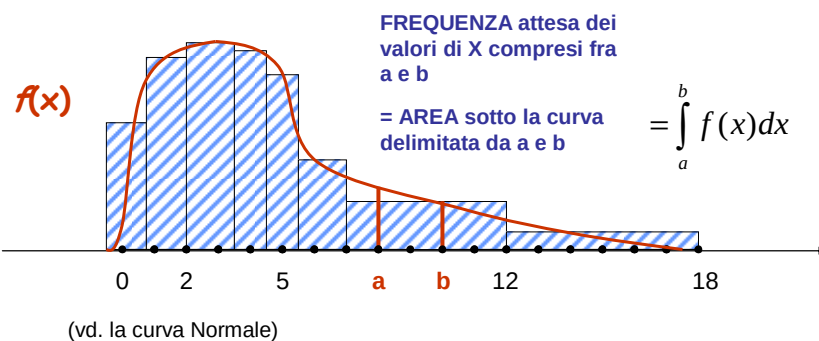
Le densità vengono poste in ordinata



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.

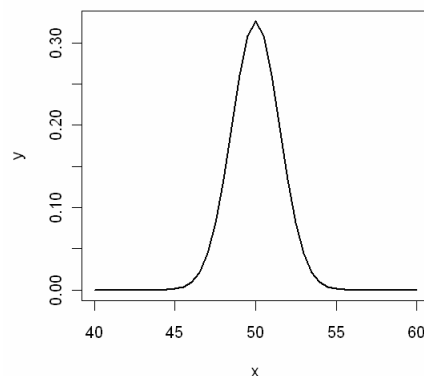
La matematica fornisce equazioni di curve continue che possono essere interpretate come curve di densità teoriche, corrispondenti a distribuzioni "ideali" di fenomeni quantitativi di interesse X.



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.



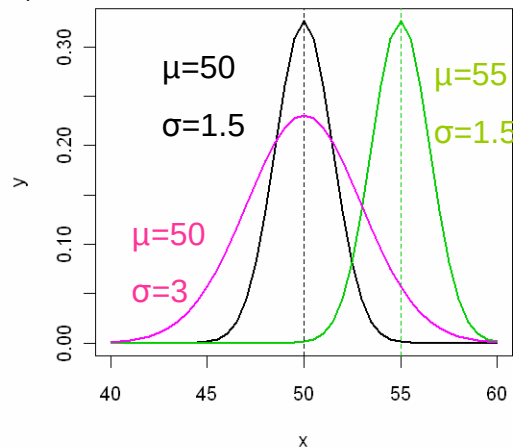
$$y = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

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"

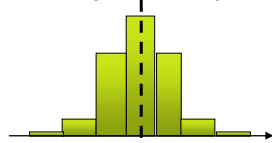
La formula che descrive la curva contiene 2 parametri μ e σ , che determinano rispettivamente dove si posiziona la curva rispetto all'asse x e quanto è ampia la campana



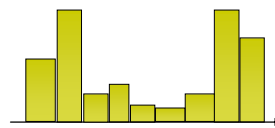
$$y = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

Varie forme della distribuzione

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

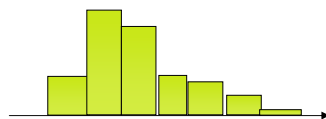


La forma "a campana" è tipica di fenomeni che possano essere ricondotti agli effetti "del caso", come l'altezza degli individui

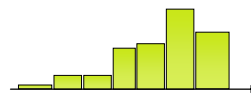


Distribuzione **BIMODALE**, cioè con la densità concentrata in due masse.

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



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



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)

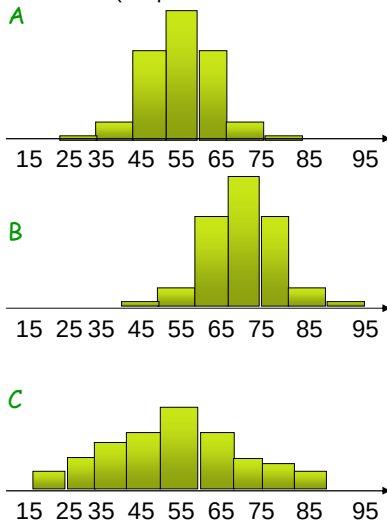


Tabelle e grafici di frequenza forniscono una rappresentazione completa dei dati.

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

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

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

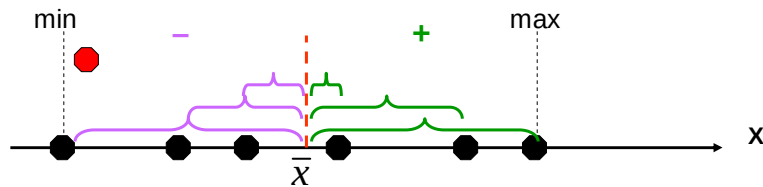
→ La media, sostituita a ciascuna osservazione, ricostituisce la somma totale delle modalità

Voto	
26	24
18	24
28	24
72	72

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:

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

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

→ 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)

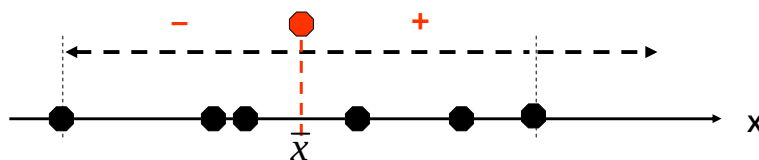
Naturalmente la formula vale anche nel caso di calcolo della media di K medie:

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

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

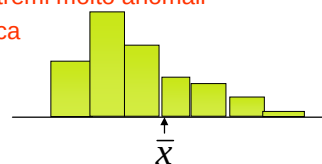


Dovendo BILANCIARE scarti positivi e negativi, e collocarsi nel centro (rispetto ai valori), la media è influenzata dai valori molto alti e dai valori molto bassi →

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

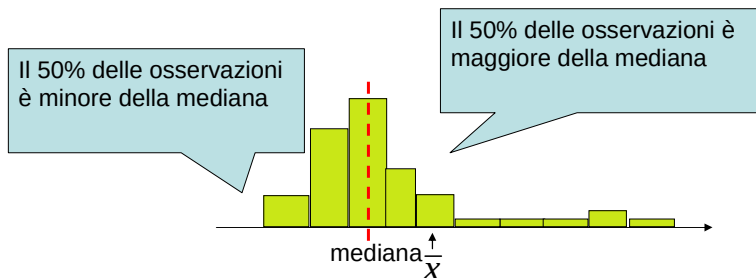
→ La media aritmetica è una sintesi insoddisfacente della distribuzione:

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



La mediana

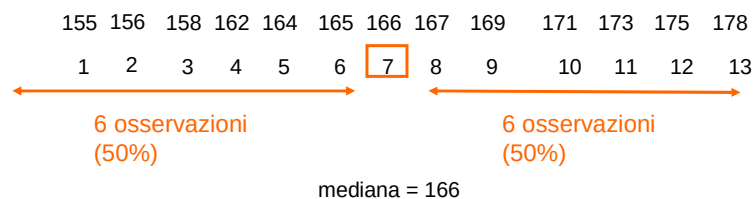
- La media aritmetica è una sintesi insoddisfacente della distribuzione:
 - Quando la distribuzione è (molto) asimmetrica
 - Quando si hanno uno o più valori estremi molto anomali
- In questi casi è più rappresentativa la mediana: il valore x tale che la metà delle osservazioni è $< x$ (e l'altra metà 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 di posto $(n+1)/2$

n dispari → 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°)

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:

173 155 162 165 167 ~~175~~ 171 169 164 ~~178~~ 156 158 166 $\bar{x} = 166.1$

$$189 \qquad 210 \qquad \bar{x} = 169.6$$

155 156 158 162 164 165 166 167 169 171 173 189 210

1 2 3 4 5 6 7 8 9 10 11 12 13

6 osservazioni
(50%)

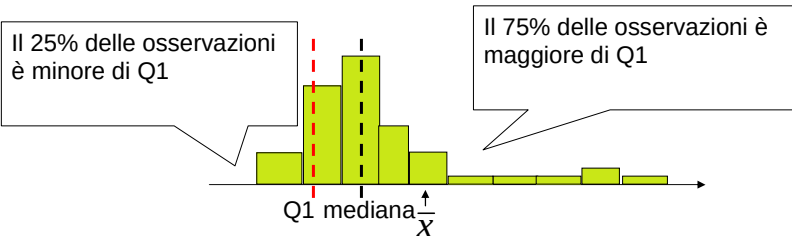
6 osservazioni
(50%)

mediana = 166

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

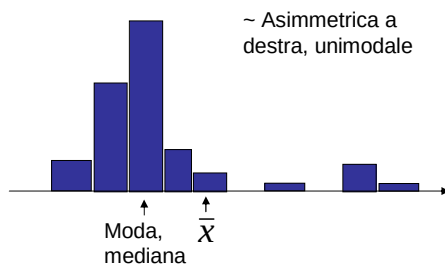
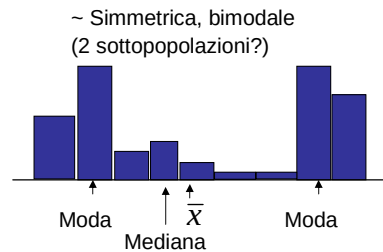
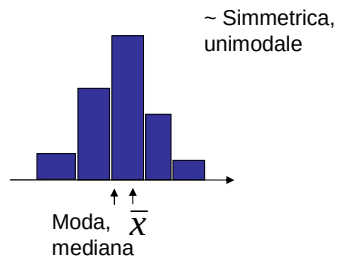
Generalizzazione della mediana: quantili

- 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.



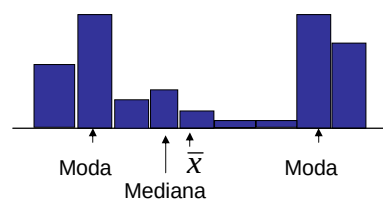
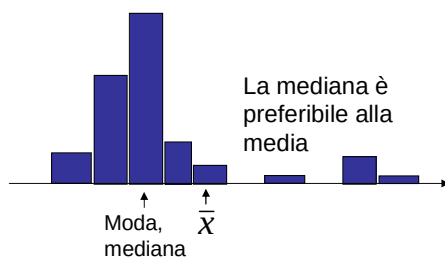
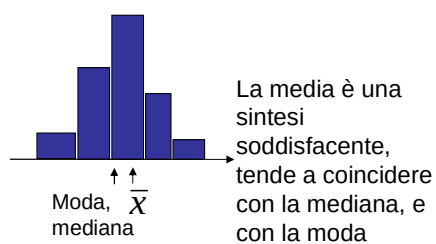
Nota: la mediana e tutti i quantili possono essere calcolati anche per caratteri QUALITATIVI ORDINATI

Forma della distribuzione e indici



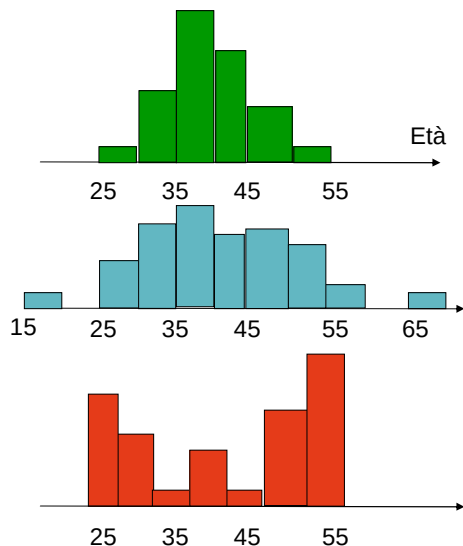
La forma della distribuzione è individuabile (in maniera grossolana) a partire dagli indici sintetici – e viceversa.

Appropriatezza degli indici



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

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*

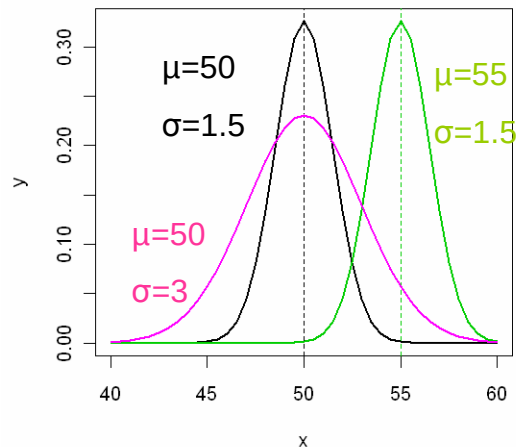
Nota: al denominatore si mette (n-1) anziché per n per motivi legati ad un concetto (distorsione) che affronteremo nella parte di inferenza

- 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

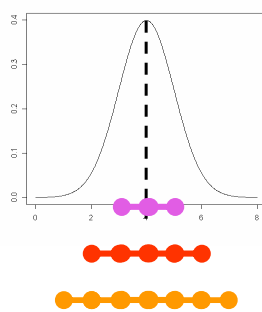
La curva Normale (ii)

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 kg / 3.2 kg) · 100 = 15.6

Altezza: CV = (3.5 cm / 51 cm) = 6.9

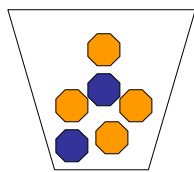
Peso Madre: CV = (4.5 kg / 64 kg) = 7.0

→ 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

Elementi di calcolo delle probabilità, e loro applicazione in medicina

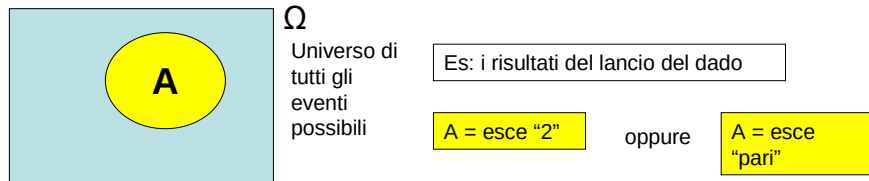


- Gli eventi e la Probabilità: le regole basilari
 - Il concetto di dipendenza probabilistica
- La regola di Bayes e sue implicazioni
 - I test diagnostici
- Le distribuzioni di probabilità per i caratteri continui: es. la Normale

Eventi e Probabilità

- Le nozioni di evento e probabilità sono intuitive e comunemente utilizzate in ogni ambito, anche nella vita quotidiana, e non solamente nei contesti di gioco. Un evento è un fatto che può o meno verificarsi. **La probabilità esprime l'aspettativa nel verificarsi dell'evento, e in genere viene espressa in percentuale.**
- In ambito scientifico, esistono diverse impostazioni "filosofiche" che danno luogo a diverse definizioni. Prescindendo da esse, proponiamo di adottare un approccio intuitivo per cui un **evento** sia **qualsiasi oggetto (fenomeno, avvenimento o caratteristica) che possa essere immaginato come il risultato di una prova – paragonabile all'estrazione da un'urna** – non limitandosi a oggetti che si verificheranno nel futuro.
 - Es. in ambito biomedico sono oggetti di interesse – la probabilità di infezione durante il ricovero, di presentare un'anomalia cromosomica, di essere un fumatore, etc.
- La probabilità esprime il grado di aspettativa, basata su criteri logici, nozioni esistenti e aspettativa "soggettiva", e viene formalizzata nell'ambito del calcolo delle probabilità.

Eventi - Insiemi - e Probabilità



La probabilità dell'evento A è un numero:

$$0 < p(A) < 1$$

Ω è l'evento certo: $p(\Omega)=1$

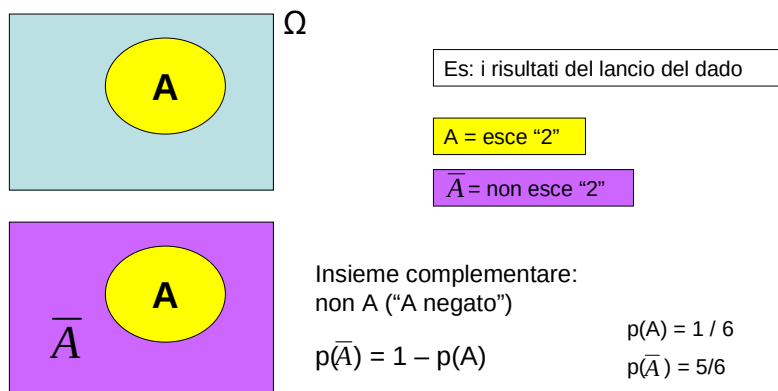
Nei casi più semplici, dove la prova ha un numero finito di possibili esiti, e tutti sono ugualmente probabili, $p(A)$ = numero casi favorevoli / numero casi possibili.

Rispettivamente nei due esempi:

$$p(A) = 1 / 6$$

$$p(A) = 3 / 6 = 1 / 2$$

Evento complementare

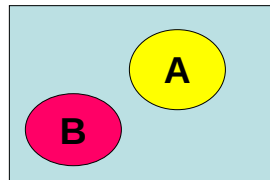


L'evento complementare di A è semplicemente l'evento che comprende tutti i casi in cui A non si verifica

$$p(\text{infezione}) = 0.7$$

$$p(\text{no infezione}) = 0.3$$

Intersezione → “e”



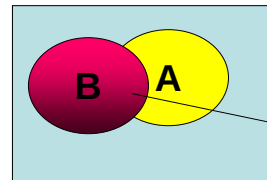
A, B
insiemi
disgiunti

A = esce “2”

B = esce “3”

$A \cap B = \emptyset$

Insieme vuoto =
complementare di Ω =
evento impossibile



A, B insieme che
si intersecano

Intersezione:
“A & B”
 $A \cap B$

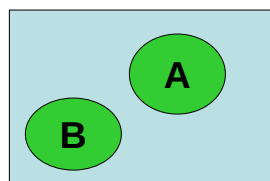
A = esce “pari”

B = esce un numero ≤ 3

$A \cap B$ = esce “2”

L'intersezione di due eventi A e B comprende tutti i casi in cui si verificano sia A che B: può essere vuota, ossia “impossibile”

Unione → “oppure” (1)



A, B
insiemi
disgiunti

A = esce “2”

B = esce “3”

$A \cup B$ = esce
“2” oppure “3”

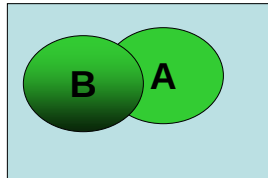
$$p(A \cup B) = p(A) + p(B)$$

Es: i risultati del lancio del dado

“pari” = “2 o 4 o 6”

$$\begin{aligned} p(\text{pari}) &= p(2) + p(4) + p(6) \\ &= 1/6 + 1/6 + 1/6 = 3/6 = 1/2 \end{aligned}$$

Unione → “oppure” (2) “o anche”



A, B insieme che si intersecano

L'unione di A e B comprende tutti i casi in cui si verifica A oppure B (compresi gli eventuali casi in cui si verificano entrambi - intersezione)

A = esce “pari”

B = esce un numero ≤ 3

A & B = esce “2”

A U B = esce “1” oppure “2” oppure “3” oppure “4” oppure “6”

$$p(A \cup B) = p(A) + p(B) - p(A \& B)$$

Nota: questo caso generale comprende quello particolare con eventi disgiunti.

Anche la formula è in realtà la stessa, poiché se gli eventi sono disgiunti $p(A \& B) = 0$

Probabilità condizionata

Spesso, la probabilità di un evento cambia a seconda dell'informazione che abbiamo

Es: $p(\text{esce } 2) = 1/6$

Ma: se so che “esce pari” → la prob. che esca 2 sale a $1/3$

Introduciamo quindi il concetto di probabilità condizionata:

$p(A|B)$ = prob. di A condizionata a B

- “dato B”
- “se si verifica B”
- “sapendo che si verifica B”
- “restringendosi ai casi in cui si verifica B”

Es: Nella popolazione generale, la prob. di decesso per infarto è 5%; fra gli obesi, è 10%.

$p(\text{decesso per infarto}) = 0.05$

$p(\text{decesso per infarto}|\text{obeso}) = 0.10$

Eventi dipendenti e indipendenti

Quando la probabilità di un evento NON cambia in presenza di condizionamento ad un altro evento, essi si dicono **indipendenti**

$$p(A|B) = p(A)$$

→ Il condizionamento non agisce!

→ L'aspettativa di A non si modifica sapendo che si verifica B

Nota: non è una indipendenza materiale, logica, causale delle prove. E' una indipendenza "della probabilità".

Analogamente, A e B si dicono **dipendenti** se:

$$p(A|B) \neq p(A)$$

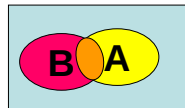
L'evento B non modifica l'evento A in modo materiale, concreto; quello che si modifica è la probabilità

Formule per prob. condizionata e intersezioni

$$p(A | B) = \frac{p(A \cap B)}{p(B)}$$

Prob. di A condizionata a B

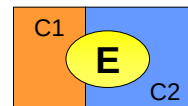
$$\begin{aligned} p(A \cap B) &= p(A | B) \cdot p(B) \\ &= p(B | A) \cdot p(A) \end{aligned}$$



$$p(A \cap B) = p(A) \cdot p(B)$$

Caso particolare per eventi A e B indipendenti

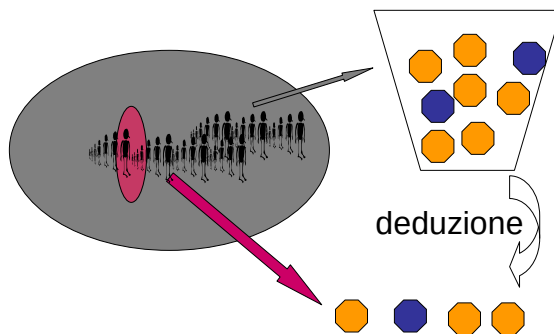
$$p(C1 | E) = \frac{p(E | C1) \cdot p(C1)}{p(E | C1) \cdot p(C1) + p(E | C2) \cdot p(C2)}$$



Formula di Bayes: per calcolare la probabilità a posteriori di C1 dato E: vd. applicazioni

Utilizzare la probabilità in medicina

Ci interessa un fenomeno "in generale", in una Popolazione obiettivo. Usiamo un modello matematico per rappresentare il suo "andamento teorico". **Se conosciamo i parametri che descrivono come è composta l'urna, possiamo elaborare ulteriormente le nostre informazioni.**



Es: X Risposta al trattamento

Supponiamo di conoscere la composizione dell'urna = sappiamo che nella popolazione / in generale, il trattamento è efficace nel 25% dei casi:

$$P(\text{Risposta})=0.25$$

Allora in un campione di 4 pazienti, mi aspetto di osservare una risposta.

Esempio: dal quesito al problema di probabilità

In una certa popolazione:

- Il 20% dei parti va incontro a complicazioni;
- La metà di questi richiede un taglio cesareo;
- In generale, il 30% dei parti è cesareo.

Qual è la prob. di avere un parto cesareo o con complicazioni?

Eventi: A = complicazioni; B = cesareo

Informazioni: $p(A)=0.2$ $p(B) = 0.3$ $P(B|A) = 0.5$

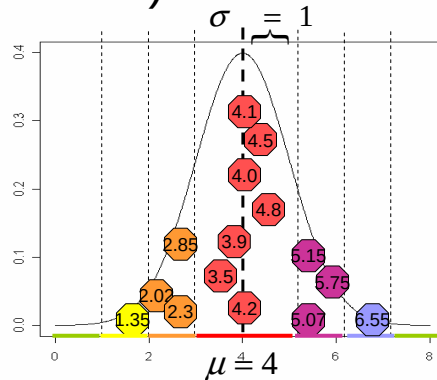
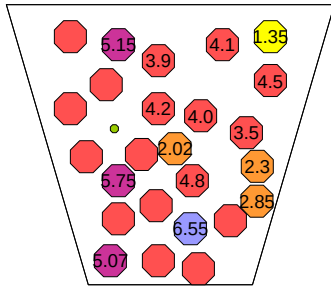
Quesito: $P(A \text{ o } B)$

$$P(A \text{ o } B) = p(A) + p(B) - p(A \text{ e } B)$$

$$p(A \text{ e } B) = p(B|A) \cdot p(A) = 0.5 \cdot 0.2 = 0.1 \text{ (prob. cesareo con complicazioni)}$$

$$P(A \text{ o } B) = 0.2 + 0.3 - 0.1 = 0.4$$

Caratteri continui: una curva di densità teorica (es. la Normale) descrive l'urna



Il carattere X con densità Normale ($\mu=4$, $\sigma=1$) assume valori:

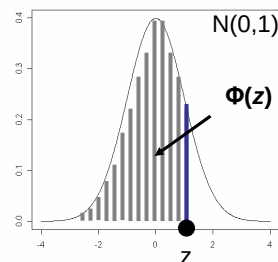
- Molto "densi" attorno a 4; il 68% distanti meno di 1, in eccesso o in difetto, ossia fra 3 e 5
- Un po' meno densi fra 2 e 3 (circa il 14%) o fra 5 e 6 (ancora 14%)
- Soltanto il 2% fra 1 e 2, o fra 6 e 7
- Praticamente nessuno <1 oppure >7 : in tutto, 3 su 1000

Calcolare probabilità per la Normale

Per la Normale(0,1) (detta Standard) calcolatori o tavole forniscono i valori dell'area sotto la curva, fino a z : indichiamola con $\Phi(z)$, per ogni z .

Per qualsiasi altra Normale(μ, σ), per avere l'area fino a x , basta calcolare Φ sul valore trasformato:

$$z = \frac{x - \mu}{\sigma} \quad (\text{Standardizzazione})$$



Per calcolare aree con altra forma, basta comporla o scomporla in pezzi del tipo di $\Phi(z)$, ricordando che vale la simmetria attorno all'asse μ , per cui:

$$\text{Area}(Z < -z) = \text{Area}(Z > z)$$

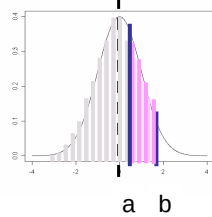
$$\Phi(-z) = 1 - \Phi(z)$$

$$\Phi(0) = 0.5 \quad \Phi(+\infty) = 1$$

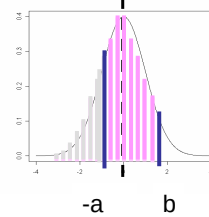
Calcolare probabilità per la Normale

Utilizzando tavole che forniscono $\Phi(z) = \text{Area}(-\infty, z)$ per $z > 0$:

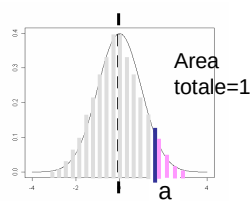
$$\text{Area}(a, b) = \Phi(b) - \Phi(a)$$



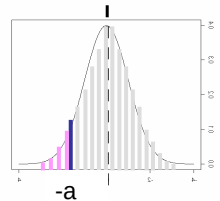
$$\text{Area}(-a, b) = \Phi(b) - (1 - \Phi(a))$$



$$\text{Area}(a, +\infty) = 1 - \Phi(a)$$



$$= \Phi(-a)$$



Due valori di Φ da ricordare:

$$\Pr(Z > 1.96) = \Pr(Z < -1.96) = 2.5\%$$

$$\Pr(Z > 1.64) = \Pr(Z < -1.64) = 5\%$$

Esempio: Normale

In una popolazione di ragazze adolescenti, il Body Mass Index (BMI) si distribuisce secondo una Normale con media 23 e varianza 7. Se definiamo "sottopeso" le ragazze con BMI inferiore a 18, qual è la probabilità di essere sottopeso? Quante ragazze risulteranno sottopeso in un gruppo di 60?

Variabile aleatoria: X = valore del BMI

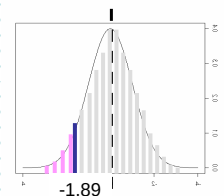
Informazioni: $\mu=23$ $\sigma^2=7$

Quesito: $P(X < 18)$

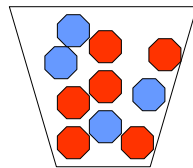
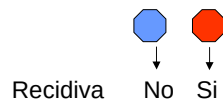
Standardizziamo il valore $x=18$: $z = \frac{18 - 23}{\sqrt{7}} = -1.89$
(è negativo!) $\Phi(-1.89) = 1 - \Phi(1.89)$

$= 1 - 0.971 = 0.029 \approx 3\%$

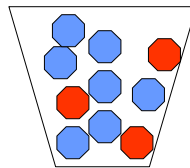
Su 60 ragazze, circa il 3%, pari a $0.029 \cdot 60 = 1.74$, dunque circa 2 risulteranno in sovrappeso



Confronto di probabilità: il Risk Ratio*



Popolazione (urna) dei
soggetti ESPOSTI ad un
fattore di interesse, ad es.
Trattati con farmaco A
 $\Pr(\text{Rec} | A) = 0.60$



Popolazione (urna)
dei soggetti esposti
NON ESPOSTI, ad
es. Trattati con
farmaco B
 $\Pr(\text{Rec} | B) = 0.30$

Per quantificare la
differenza che si verifichi
un evento fra due gruppi,
si calcola il rapporto delle
probabilità: **Risk Ratio**

$$RR = \frac{\pi_E}{\pi_{NE}} = \frac{0.60}{0.30} = 2$$

= 1 : non c'è relazione
tra Esposizione e
verificarsi dell'evento
> 1 : Esposizione fattore
di rischio per l'evento
Tra 0 e 1 : Esposizione
fattore protettivo per
l'evento

Probabilità e Odds**

$\pi = \text{Prob}(E)$

→ Rapporto fra Casi
Favorevoli e Casi Possibili

$$\text{Odds: } \Omega = \frac{\pi}{1-\pi} = \frac{\Pr(E)}{\Pr(\bar{E})}$$

→ Rapporto fra Casi Favorevoli e Casi
Contrari

Gli odds sono una quantificazione alternativa dell'aspettativa
dell'evento

- Sono usati dagli scommettitori: vincita quotata 5 a 1 vuol dire
odds(vittoria)=1/5 cioè $\Pr(\text{vittoria})=1/6$
- Sono usati in Statistica ed Epidemiologia, e dunque in Medicina!
- Infatti, il rapporto fra gli odds è una misura di confronto di rischi, che in certi
casi è necessario valutare in alternativa al RR:

$$\text{ODDS RATIO} \quad OR = \frac{\frac{\pi_E}{1-\pi_E}}{\frac{\pi_{NE}}{1-\pi_{NE}}} = \frac{\pi_E}{\pi_{NE}} \frac{1-\pi_{NE}}{1-\pi_E} = RR \frac{1-\pi_{NE}}{1-\pi_E}$$

Utilizzare la formula di Bayes

Questa formula trova applicazione in quei contesti simili al problema della diagnosi: stabilire la probabilità di una causa (o malattia, o ipotesi; C1) sapendo che si verifica un suo effetto (o sintomo, o conseguenza; E) che può essere altrimenti determinato da altre cause (C2)

$$p(C1 | E) = \frac{p(E | C1) \cdot p(C1)}{p(E | C1) \cdot p(C1) + p(E | C2) \cdot p(C2)}$$

Nella formula, hanno un ruolo:

- le prob. della causa C1 e delle cause alternative C2 (prob. a priori)
- le prob. di osservare l'effetto E sapendo quale causa agisce

Sapendo che si verifica l'effetto E, è più probabile la causa C1 o le cause C2?

$$\frac{p(C1 | E)}{p(C2 | E)} = \underbrace{\frac{p(E | C1)}{p(E | C2)}}_{RR} \cdot \frac{p(C1)}{p(C2)}$$

L'effetto E è più probabile con la causa C1 o con le cause C2?

È più probabile la causa C1 o le cause C2?

La formula di Bayes e la diagnosi (1)

CON CHE PROBABILITA' E' INFARTO??

Un paziente si presenta dal medico per un dolore al braccio, temendo di avere un infarto in corso.

Il medico fa il seguente ragionamento:

- Se c'è un infarto, la probabilità di avere questo tipo di dolore è del 80%;
- D'altra parte, un'inflammation provocherebbe questo dolore nel 30% dei casi;

Il medico prosegue il ragionamento:

- Quest'uomo è giovane, magro, fa attività fisica ... la prob. di infarto in questi casi è bassa, 5%
- Invece, con lo sport che pratica, la prob. di inflammation è 40%

E = dolore
C1 = infarto
C2 = inflammation

$$p(E|C1) = 0.8$$

$$p(E|C2) = 0.3$$

VEROSIMIGLIANZE
delle ipotesi C1 e C2
dato E

$$p(C1) = 0.05$$

$$p(C2) = 0.4$$

Prob. a priori delle
ipotesi C1 e C2

La formula di Bayes e la diagnosi (2)

CON CHE PROBABILITA' E' INFARTO??

Per fare una diagnosi, il medico deve valutare tutti questi elementi, e valutare la probabilità che stia agendo la causa infarto avendo l'evidenza di un suo sintomo.

Quesito: $P(C1|E)$

$$= \frac{p(E | C1) \cdot p(C1)}{p(E | C1) \cdot p(C1) + p(E | C2) \cdot p(C2)}$$

$$= \frac{0.8 \cdot 0.05}{0.8 \cdot 0.05 + 0.3 \cdot 0.4} = 0.25$$

$p(C1|E)$: Prob. a posteriori dell'ipotesi C1

E = dolore

C1 = infarto

C2 = infiammazione

$$p(E|C1) = 0.8$$

$$p(E|C2) = 0.3$$

VEROSIMIGLIANZE
delle ipotesi C1 e C2
dato E

$$p(C1) = 0.05$$

$$p(C2) = 0.4$$

Prob. a priori delle
ipotesi C1 e C2

I test diagnostici

- Il test diagnostico è uno strumento per la diagnosi della presenza di una certa condizione, ad es. un'anomalia genetica, o più semplicemente una malattia, utilizzabile in clinica e negli screening. Test "positivo" indica presenza di quella caratteristica (es malattia).
- Il test diagnostico solitamente non dà risultati sicuri: non tutti i soggetti malati vengono individuati, e viceversa alcuni soggetti sani vengono erroneamente classificati come malati.
 - Si hanno cioè, rispettivamente, i cosiddetti **FALSI NEGATIVI e FALSI POSITIVI**
 - Questi test trovano la loro utilità quando effettuare una diagnosi più accurata sia troppo costoso – invasivo – pericoloso – etc

Le caratteristiche di un test diagnostico vengono sintetizzate da due parametri:

SENSITIVITA': la capacità di individuare i soggetti malati

SPECIFICITA': la capacità di riconoscere i soggetti sani

Capire i test diagnostici per la pratica clinica

Esempio: si stima che il 10% delle persone appartenenti ad una certa categoria di rischio sia affetta dal virus dell'HIV (per semplicità, diciamo "malata").

Supponiamo di dover sottoporre a test diagnostico un individuo di quella categoria; il test utilizzato ha sensibilità = 90% e specificità = 80%.

- Le domande che si può porre l'operatore sono:
 - Per quanti soggetti malati mancheremo la diagnosi?
 - Quanti soggetti non malati sottoporremo inutilmente a ulteriori accertamenti?
 - Quanti errori diagnostici commetteremo in tutto?
- Le domande che il soggetto sottoposto al test può porre sono ad esempio:
 - Il test dà un risultato sicuro?
 - Se sono malato uscirà test positivo?
 - Se il test viene positivo, vuol dire che sono malato?

Le probabilità nel test diagnostico (1)

		Risultato del test diagnostico	
		Test +	Test –
Situazione (incognita) del soggetto	Malato	ok	Falso negativo
	Non Malato	Falso positivo	ok

Caratteristiche del test – sensibilità e specificità:

SENSITIVITA': $p(\text{Test} + | \text{Malato})$

SPECIFICITA': $p(\text{Test} - | \text{Non Malato})$

Le caratteristiche di un test diagnostico vengono sintetizzate da due parametri:

SENSITIVITA': la capacità di individuare i soggetti malati (fornendo risultato positivo)

SPECIFICITA': la capacità di riconoscere i soggetti sani (fornendo risultato negativo)

Le probabilità nel test diagnostico (2)

Situazione (incognita) del soggetto	Risultato del test diagnostico	
	Test +	Test -
Malato	ok	Falso negativo
Non Malato	Falso positivo	ok

SENSITIVITA':
 $p(\text{Test} + | \text{Malato})$
 SPECIFICITA':
 $p(\text{Test} - | \text{Non Malato})$

Errori:

Falso negativo $\rightarrow p(\text{Test} - | \text{Malato})$

$= 1 - \text{SENSITIVITA'}$

Falso positivo $\rightarrow p(\text{Test} + | \text{Non Malato})$

$= 1 - \text{SPECIFICITA'}$

$$\begin{aligned}
 \Pr(\text{Errore}) &= \Pr(\text{Errore} \& \text{Malato}) + \Pr(\text{Errore} \& \text{Non Malato}) = \\
 &= \Pr(\text{Errore} | \text{Malato}) \Pr(\text{Malato}) + \Pr(\text{Errore} | \text{Non Malato}) \Pr(\text{Non Malato}) = \\
 &= \Pr(\text{Test} - | \text{Malato}) \Pr(\text{Malato}) + \Pr(\text{Test} + | \text{Non Malato}) (1 - \Pr(\text{Malato})) = \\
 &= (1 - \text{sensitività}) \Pr(\text{Malato}) + (1 - \text{specificità}) (1 - \Pr(\text{Malato}))
 \end{aligned}$$

Le probabilità nel test diagnostico (3)

Situazione (incognita) del soggetto	Risultato del test diagnostico	
	Test +	Test -
Malato	ok	Falso negativo
Non Malato	Falso positivo	ok

SENSITIVITA':
 $p(\text{Test} + | \text{Malato})$
 SPECIFICITA':
 $p(\text{Test} - | \text{Non Malato})$

FN:
 $p(\text{Test} - | \text{Malato}) = 1 - \text{SENS}$
 FP:
 $p(\text{Test} + | \text{Non Malato}) = 1 - \text{SPEC}$

Se il test viene positivo, l'individuo è malato?

Valori predittivi del test:

$p(\text{Malato} | \text{Test} +)$

$p(\text{Sano} | \text{Test} -)$

Si tratta di prob. "a posteriori" $\rightarrow$ dobbiamo la conoscere la prob. "a priori", non condizionata, di avere la malattia

Dobbiamo avere il dato sulla PREVALENZA della malattia

$P(\text{Malato})$

Le probabilità nel test diagnostico (4)

		Risultato del test diagnostico	
		Test +	Test –
Situazione (incognita) del soggetto	Malato	ok	Falso negativo
	Non Malato	Falso positivo	ok

Se il test viene positivo, l'individuo è malato?

$$p(M | T+) = \frac{p(T+ | M) \cdot p(M)}{p(T+ | M) \cdot p(M) + \underbrace{p(T+ | \text{non } M)}_{1-p(T- | \text{non } M)} \cdot \underbrace{p(\text{non } M)}_{1-p(M)}}$$

$$= \frac{\text{sens} \cdot \text{prev}}{\text{sens} \cdot \text{prev} + (1 - \text{spec}) \cdot (1 - \text{prev})}$$

SENSITIVITA':
 $p(\text{Test} + | \text{Malato})$
SPECIFICITA':
 $p(\text{Test} - | \text{Non Malato})$
FN:
 $p(\text{Test} - | \text{Malato}) = 1 - \text{SENS}$
FP:
 $p(\text{Test} + | \text{Non Malato}) = 1 - \text{SPEC}$
Prevalenza =
 $P(\text{Malato})$

Esempio: test diagnostico

Si stima che una patologia colpisca 1 individuo su 50. L'accertamento della presenza di questa patologia è invasivo. Un test basato su un prelievo di sangue permette di identificare i soggetti affetti. Il test ha sensibilità = 70% e specificità = 90%. Si vuole calcolare la probabilità che un soggetto con Test positivo sia malato.

Eventi: TP= test positivo; M = malattia

Informazioni: $p(\text{TP} | M) = 0.7$ $p(\text{non TP} | \text{non } M) = 0.9$ $P(M) = 1/50 = 0.02$

Quesito: $P(M | \text{TP})$

Si applica la formula di Bayes:

$$= \frac{p(\text{TP} | M) \cdot p(M)}{p(\text{TP} | M) \cdot p(M) + \underbrace{p(\text{TP} | \text{non } M)}_{1-p(\text{non TP} | \text{non } M)} \cdot \underbrace{p(\text{non } M)}_{1-p(M)}}$$

$$= \frac{0.7 \cdot 0.02}{0.7 \cdot 0.02 + 0.1 \cdot 0.98} = 0.125$$

Stima di Sensitività e Specificità*

Si vuole stimare la sensitività e la specificità di una nuova tecnica diagnostica per immagini, alternativa ad una con risultato certo, ma meno invasiva / costosa. Si prende quindi un campione di n soggetti che, sottoposti alla "vecchia" tecnica, vengono classificati in "malati" e "non malati"; li si sottopongono poi alla "nuova" diagnostica, ottenendo i seguenti risultati:

	Test +	Test -	
Malati	vp	fn	m
Non Malati	fp	vn	n-m
	vp+fp	fn+vn	n

n soggetti di cui m malati, gli altri $(n-m)$ non malati

Dei malati, vp hanno Test+ e fn hanno Test- (vp sono i "veri positivi", fn sono i "falsi negativi")

etc

$$\text{sensitività} \approx \frac{vp}{m} = \frac{vp}{vp + fn}$$

$$\text{specificità} \approx \frac{vn}{n-m} = \frac{vn}{vn + fp}$$

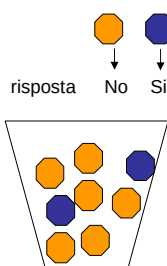
Attenzione! Il valore predittivo si può calcolare solo se conosciamo la **prevalenza** della malattia. **Solo se possiamo pensare di stimarla dal campione, (=m/n)** allora si ha:

$$p(M | T+) \approx \frac{vp}{vp + fn}$$

(stesso risultato con la formula di Bayes)

La legge Binomiale*

In ogni contesto assimilabile all'osservazione di un evento ("successo") che ha probabilità π di verificarsi, in N casi, o soggetti, o "prove", in cui interessa il numero totale (X) di successi, si possono usare le seguenti formule



$$\pi = P(\text{Risposta}) = 0.25$$

Quanti pazienti rispondono fra 20 trattati?

$$20 \cdot 0.25 = 5$$

Qual è la prob. che rispondano 10 pazienti?

$$p(X = 10) = \binom{20}{10} 0.25^{10} (1-0.25)^{20-10} = 0.0099$$

- Il numero medio "atteso" di eventi è $N \cdot \pi$
- La probabilità di osservare esattamente x eventi è data da:

$$p(X = x) = \binom{N}{x} \pi^x (1-\pi)^{N-x}$$

Dove:

$$\binom{N}{k} = \frac{N!}{(N-k)!k!}$$

$$k! = k \cdot (k-1) \cdot (k-2) \cdot \dots \cdot 2 \cdot 1$$

$$5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1$$

$$0! = 1$$

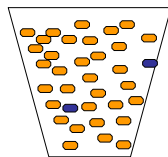
La legge di Poisson**

- Sebbene la sua utilità si espliciti soprattutto in contesti in cui si faccia riferimento al tasso di incidenza di un evento (che noi tralasciamo), questa legge serve anche a estendere la legge Binomiale:
- Nello **stesso contesto della Binomiale**, quando π è molto piccola (evento raro) e N è molto grande, si calcolano le probabilità con la formula:

$$p(X = x) = \frac{e^{-\mu} \mu^x}{x!}$$

in cui $\mu = N \cdot \pi$

Ricordiamo che: $e^{-\mu} = \frac{1}{e^{\mu}}$
 $k^0 = 1$



Se dobbiamo contare il numero X di eventi in N prove con:

$\pi = P(\text{Evento}) \approx 0$

$N \gg 0$

usiamo la Poisson!!

Esempio: Poisson**

Una certa malattia colpisce appena 1 neonato su 1000. Qual è la probabilità che in un campione di 50 neonati si osservi almeno 1 bimbo affetto dalla malattia?

Variabile aleatoria: X = numero di bambini affetti

Informazioni: $p(\text{affetto}) = \pi = 0.001$ numero di prove: $N = 50$

Quesito: $P(X=1 \text{ o } X=2 \text{ o } \dots) = P(X>0)$

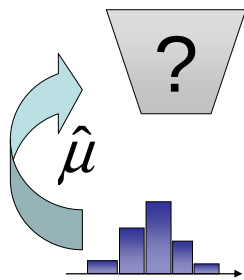
Usiamo la Poisson, con $\mu = N \cdot \pi = 50 \cdot 0.001 = 0.05$

$P(X>0) = 1 - P(X=0)$

$$p(X = 0) = \frac{e^{-0.05} 0.05^0}{0!} = e^{-0.05} = 0.95$$

$$P(X>0) = 1 - 0.95 = 0.05 = 5\%$$

Inferenza statistica: risalire dal campione alla popolazione



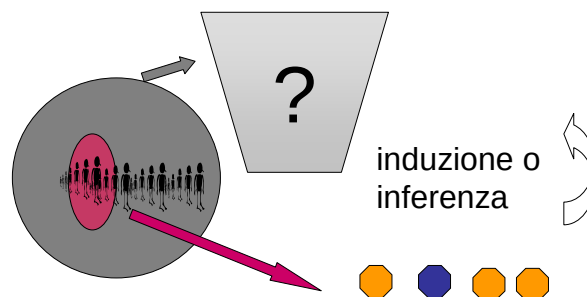
Elementi MOLTO GENERALI
relativi a:

- Stima puntuale e intervallare
 - Stimatori non distorti
 - Intervalli di confidenza
- La verifica delle ipotesi
 - Significatività
- Lo studio delle relazioni
 - Es. di indici e i test

Risalire dal Campione alla Popolazione

Ci interessa un fenomeno “in generale”, in una Popolazione obiettivo. Usiamo un modello matematico per rappresentare il suo “andamento teorico”. Ma **non conosciamo i parametri che descrivono come è composta l'urna.**

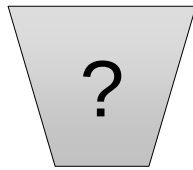
Guardiamo i **DATI** in un Campione. Essi sono come un insieme di **palline estratte dall'urna.** Allora, **i dati del campione ci danno informazione su com'è fatta l'urna.**



La stima statistica

Com'è fatta la popolazione (l'urna)? Ossia, com'è distribuito il carattere X nella popolazione?

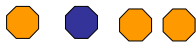
Dobbiamo dare una valutazione ai parametri della distribuzione.



Es: X1 Risposta ad un nuovo trattamento:

$\pi = P(\text{Risposta}) = ?$

X2 Età alla diagnosi. Se è distribuita come una Normale, quanto valgono μ e σ ?



$$\text{freq}(\text{Risposta}) = \frac{1}{4} = 0.25 \Rightarrow \hat{\pi} = 0.25$$

$$\text{media}(\text{Età}) = \bar{x} = 67 \Rightarrow \hat{\mu} = 67$$

$$\text{dev.st.}(\text{Età}) = s = 3.2 \Rightarrow \hat{\sigma} = 3.2$$

Gli indici calcolati nel campione vengono assunti come **STIME** dei parametri nella popolazione

Stimatori

La funzione che calcola la stima di un parametro a partire dai dati del campione è chiamata **STIMATORE**. Come si sceglie un "buon" stimatore?

Immaginando di avere a disposizione un numero infinito di campioni, e di applicare sempre lo stesso stimatore, si richiede che esso soddisfi alcune **proprietà**.

→ La non-distorsione: lo stimatore è **non distorto** ("corretto", "unbiased") se **non produce sistematicamente sovrastima o sottostima**

→ L'efficienza: in sostanza, la sua **precisione**, il fatto che i diversi valori di stima ottenuti negli infiniti campioni siano simili fra loro

→ La consistenza: la proprietà secondo cui, più il **campione è numeroso**, più la **stima viene vicina al parametro**

Parallelo: stimatore = bilancia:

E' una bilancia ben tarata: nella singola prova (misura) si può commettere un errore di valutazione, ma ripetendo le prove e facendo la media dei risultati ottengo il valore esatto del peso: l'errore **NON** è sistematico

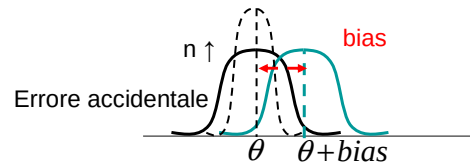
Fra due bilance è non-distorte, preferisco quella che non dà molti valori "estremi"

Se faccio molte misure, voglio avere un valore sempre più vicino a quello che devo stimare

Il concetto di Bias*

- Bias vuol dire “distorsione”, ovvero ERRORE SISTEMATICO.
- Può essere riferito ad un metodo statistico, ad es. ad uno stimatore
 - Es. Se calcoliamo la varianza del campione dividendo per n anziché per $n-1$, abbiamo uno stimatore distorto della varianza della popolazione
- Può essere riferito allo studio (all'esperimento)
- In questo secondo caso, c'è un errore di impostazione, dovuto ad esempio a come viene selezionato il campione o a come vengono valutate le variabili
 - Es. In un trial clinico per confrontare i farmaci A e B, si valuta la differenza dell'incidenza di effetti collaterali. Spesso accade che il personale medico valuti con più meticolosità i pazienti sottoposti al trattamento sperimentale, dunque l'incidenza di effetti collaterali può essere sovrastimata per tale farmaco
- A differenza dell'errore accidentale, NON tende a ridursi all'aumentare dell'ampiezza del campione

Rappresentazione figurata degli effetti dell'errore accidentale e sistematico nella stima di θ , immaginando di ripetere l'esperimento infinite volte (su infiniti campioni)



Stima intervallare

Anzichè considerare come stima di un parametro un valore singolo (stima puntuale) si considera un **intervallo di valori**. Questo tiene implicitamente conto dell'imprecisione insita nella procedura di campionamento e stima, quindi è una valutazione più prudente, e perciò affidabile rispetto al singolo valore.

$$\hat{\mu} = 67 \quad IC95\% : (63.9, 70.1)$$

Un “buon” intervallo di stima deve:

- garantire (in qualche misura) di includere il valore vero del parametro che si vuole stimare,
- fornire una stima abbastanza precisa = rimanere “non troppo ampio”

Intervalli di Confidenza

(in generale: di livello α ; qui: al livello del 95%)

Immaginando di disporre di infiniti campioni, il metodo dell'**intervallo di confidenza al livello 95%** garantisce che per 95 campioni su 100 l'intervallo ottenuto contiene il parametro che si vuole stimare ed è "il più piccolo possibile".

- Il "**livello di confidenza**" dunque esprime un grado di fiducia nella regola di costruzione.
- Per stimare la media di una popolazione usando i dati di un campione di numerosità "grande" (>30), calcolare:

$$\left(\bar{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}} \right)$$

Esempio: IC 95% per la media

Si vuole stimare il valore medio della pressione sanguigna fra i soggetti sottoposti ad un certo trattamento farmacologico. Si dispone di un campione di 130 soggetti, con media campionaria pari a 160 e deviazione standard pari a 25.

Data l'ampiezza del campione, è possibile applicare la formula dell'intervallo di confidenza; scegliamo il livello 95%

$$\bar{x} = 160$$

$$s = 25, n = 130 \Rightarrow s.e.(\bar{X}) = 25 / \sqrt{130} = 2.19$$

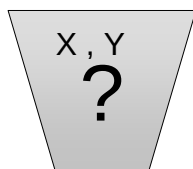
$$z_{\frac{\alpha}{2}} = 1.96$$

$$\begin{aligned} 95\%CI &= \left(\bar{x} - 1.96 \cdot \sigma / \sqrt{n}, \bar{x} + 1.96 \cdot \sigma / \sqrt{n} \right) \\ &= (160 - 1.96 \cdot 4.30, 160 + 1.96 \cdot 4.30) \\ &= (155.7, 164.3) \end{aligned}$$

Quesiti (clinici) sulle relazioni fra caratteri

Nella ricerca clinica, moltissimi quesiti riguardano le relazioni (in generale, nella Popolazione) fra due fenomeni o caratteri. Es:

- Il nuovo farmaco è più efficace di quello in uso per l'ottenimento della Risposta? Ossia, c'è differenza fra le prob. di Risposta con il nuovo farmaco e con il vecchio farmaco?
- Il sesso influenza l'età a cui insorge la malattia? Ossia, c'è differenza fra l'età alla diagnosi fra Maschi e Femmine?
- La pressione si modifica al crescere dell'età? Ossia, c'è qualche relazione fra pressione ed età?



Per studiare l'esistenza delle relazioni fra caratteri dobbiamo:

- Descrivere le relazioni osservate nel campione
- Dire se la relazione osservata sembra o no valere in generale, nella popolazione → Fare un test statistico

Verifica delle ipotesi (o test statistico)

Si hanno due ipotesi sulla composizione dell'urna:

- H_0 [ipotesi di base, o "nulla"] è l'ipotesi di riferimento, corrispondente a uno stato "neutro" di conoscenze;
- H_1 [ipotesi alternativa] contiene una "nuova conoscenza", il risultato atteso della ricerca.

Esempio:

H_0 : non c'è differenza fra la prob. di Risposta fra nuovo farmaco e vecchio farmaco

H_1 : c'è differenza.

Si vanno a vedere i DATI. Sono "compatibili" con l'ipotesi di base? o indicano che essa non è valida, mentre potrebbe essere vera l'ipotesi alternativa?

La "compatibilità" è in termini probabilistici: **SE è vera l'ipotesi di base, con che probabilità ottengo i dati effettivamente osservati, o dati ancora più a supporto di H_1 ?**

Se questa probabilità (**p-value**) è molto bassa (es. < 5%), concludiamo che possiamo rigettare l'ipotesi di base.

Altrimenti, accettiamo H_0 , ossia manteniamo lo stato di conoscenze attuale.

Il valore soglia che usiamo per discriminare fra accettazione e rifiuto di H_0 è detto **livello di significatività**

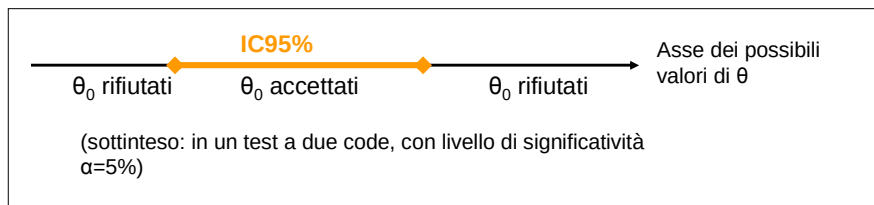
Relazione fra intervallo di confidenza e test*

Investighiamo un parametro θ che descrive una Popolazione:

- facendo una stima mediante IC al livello del 95%
- impostando un test per $H_0: \theta = \theta_0$ vs $H_1: \theta \neq \theta_0$ al livello di significatività del 5%

→ se l'IC contiene θ_0 , vuol dire che il test accetta H_0

→ se θ_0 non è contenuto nell'IC, H_0 viene rigettata



Relazione fra 2 caratteri qualitativi

(overview)

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

Impostiamo un test per confrontare due ipotesi:

- H_0 [ipotesi di base, o "nulla"]: Non c'è differenza; H_1 [ipotesi alternativa]: c'è differenza.

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**).

Facciamo il **test**, calcolando un p-value, per verificare se tale indice ha assunto un valore compatibile con l'ipotesi di base (probabile sotto H_0) o no.

→ Se sì [p-value grande], la differenza osservata fra M e F è attribuibile al caso.

→ Se no [p-value molto piccolo], concludiamo che c'è una **differenza "significativa" e rigettiamo H_0** .

Tabelle doppie

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

Sesso	Y = Fumo		totale
	no	si	
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%.

Significatività

I fumatori fra i M e fra le F sono rispettivamente il 64% e il 41% (RR=1.6).
Può darsi che:

- vale H_0 : nella popolazione non c'è differenza; questa differenza osservata nel campione è frutto del caso
- vale H_1 : nella popolazione c'è differenza. Diremo che la differenza osservata nel campione è "significativa"

Nota: **questo termine non si riferisce all'ammontare della differenza** (grande / piccola; clinicamente "rilevante"; "importante"; è rilevante un RR pari a 1.6?). E' un termine tecnico che sta per "non casuale".

Significativo = incompatibile con H_0 , nel senso di improbabile sotto H_0 , tanto da indurci a rigettare H_0 .

Importante: **anche una differenza molto piccola risulta molto significativa se il campione è molto grande, viceversa una differenza grande può essere non significativa se il campione è piccolo.**

Rilevanza vs. Significatività* (1)

Es: Due fattori di esposizione, A e B, ugualmente diffusi nella popolazione. Il loro effetto sul rischio di malattia stimato è:

$$RR_A = 1.2 \quad (p\text{-value} = 0.0002)$$

$$RR_B = 3.0 \quad (p\text{-value} = 0.020)$$

Quale dei due fattori è più importante prevenire?

Sappiamo che il p-value fornisce una misura della SIGNIFICATIVITA'; entrambi gli effetti osservati sono "significativi", non casuali. A è "più significativo" ... vuole semplicemente dire che abbiamo meno dubbi sull'eventualità che questo effetto sia dovuto al caso.

La RILEVANZA è fornita dalla stima. Se prendiamo gli Intervalli di Confidenza, abbiamo informazione anche sulla significatività (*), e inoltre possiamo valutare più a fondo la rilevanza:

$$RR_A \rightarrow 95\%CI = (1.15, 1.25)$$

$$RR_B \rightarrow 95\%CI = (2.05, 3.95)$$

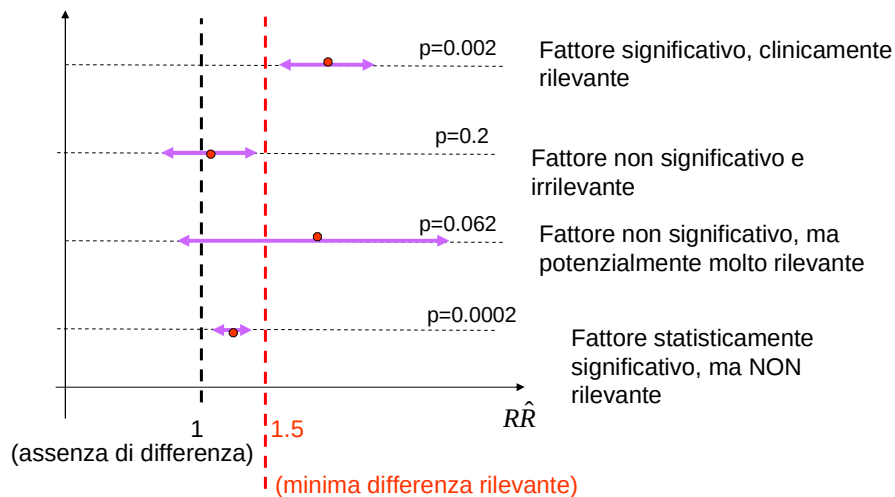
Il fattore più importante è B, perché come minimo raddoppia il rischio (e al massimo lo quadruplica).

Invece A al massimo aumenta il rischio del 25%. La maggiore significatività corrisponde solo ad una maggiore precisione della stima del RR.

(*) RR significativo al livello 5% se l'IC95% non contiene il valore 1

Rilevanza vs. Significatività* (2)

Valutiamo 4 fattori prognostici per il verificarsi di un evento; viene considerato "importante" un fattore che aumenta almeno del 50% il rischio di evento



χ^2 : Associazione vs. Indipendenza

Per 2 caratteri qualitativi, vi sono vari test per valutare la significatività della relazione osservata. Un test che si può applicare anche con caratteri con più di 2 modalità ciascuno è basato su un indice del grado di associazione osservato.

Esso è **misurato come differenza fra la situazione osservata e quella che si osserverebbe nel caso di indipendenza perfetta**: se vale H_0 , ossia se il Fumo ha una distribuzione indipendente dal Sesso, dovrei osservare sia fra i M che fra le F il 51% di fumatori:

Su 66 maschi, dovrei osservare $0.51 \cdot 66 = 34$ Fumatori; su 76 femmine, $0.51 \cdot 76 = 39$ Fumatrici.

Vediamo meglio:

$$M \text{ fumatori attesi} = \frac{\text{fumatori tot.}(73)}{\text{tutti}(142)} \cdot \text{maschi tot.}(66) = \frac{\text{tot.col.} \cdot \text{tot.riga}}{\text{tot}}$$

→ Formula generale delle frequenze attese (E, "expected") sotto l'ipotesi H_0 che i due caratteri siano indipendenti:

$$E = \frac{\text{tot riga} \cdot \text{tot colonna}}{\text{tot generale}}$$

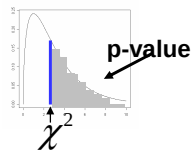
χ^2 : distanza dall'indipendenza perfetta

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

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

Nel caso di indipendenza perfetta, χ^2 vale 0; cresce al crescere del grado di associazione.

Dunque la sua distribuzione sotto H_0 è descritta dalla seguente curva di densità:



Con il computer o usando delle tavole si può calcolare il p-value: la probabilità di osservare quella certa distanza da 0 o una distanza ancora superiore anche se nella popolazione la distanza è 0.

In termini MOLTO generali, per una tabella 2x2, Chi-Quadro > 3.841 è un valore "statisticamente significativo", poiché corrisponde ad un p-value < 0.05.

Differenza di un carattere quantitativo in 2 gruppi (overview)

Es: c'è relazione fra Pressione e Trattamento (A o B)? Assumiamo[*] che la pressione segua una legge Normale, che A e B abbiano la stessa varianza. Impostiamo il test: le **ipotesi da confrontare** sono:

– H_0 : Non c'è differenza, $\mu_A = \mu_B$; H_1 : c'è differenza $\mu_A \neq \mu_B$.

Raccogliamo dei DATI, e descriviamo la pressione separatamente per A e B. In particolare, guardiamo alle due medie della pressione nel campione trattato con A e nel campione trattato con B. Ad es. risultano rispettivamente pari a 96 e 92.

Siamo interessati a valutare la differenza fra le due medie (es. =4): è una differenza dovuta al caso, avendo A e B la stessa distribuzione, oppure è significativa? Calcoliamo una **misura opportuna della differenza, ricorrendo alle formule del T-test**, e procediamo al calcolo del p-value.

[*] Nota: questo è un test detto "parametrico", basato su ipotesi piuttosto stringenti. Le ipotesi devono essere verificate a loro volta usando altri test. La Normalità della popolazione può non valere, purché il campione sia "grande" ($n > 30$).

Esistono test di confronto di un carattere quantitativo fra due gruppi che valgono in casi più generali (test "nonparametrici", es. Mann-Whitney)

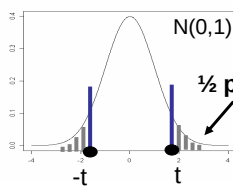
T-test: distanza della differenza osservata da 0

→ Misura della differenza fra le due medie osservate:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$\text{dove } s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Nel caso H0 di uguaglianza delle distribuzioni (delle popolazioni / delle urne) dovremmo avere nei due campioni delle medie molto simili, e quindi una differenza molto prossima allo zero. Si dimostra che, facendo infiniti esperimenti, la differenza osservata t va a distribuirsi secondo una curva a campana (T di Student). Se il campione è grande, sotto H0 t segue una Normale N(0,1):

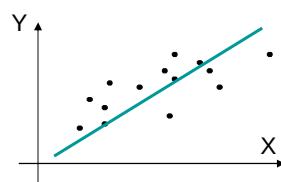


Con il computer o usando delle tavole si può calcolare il p-value: la probabilità di osservare quella certa differenza da 0 o una differenza ancora maggiore anche se nella popolazione la differenza è 0.

Nota: qui H1 considera una differenza sia positiva che negativa, quindi il p-value è da calcolarsi con riferimento a **due code**. Con campioni grandi, una $t > 1.96$ o < -1.96 è "statisticamente significativa" rispetto all'usuale livello 5%. Usando una sola coda, il valore soglia è 1.64.

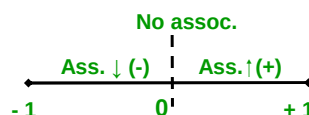
Associazione fra 2 caratteri continui

- 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:



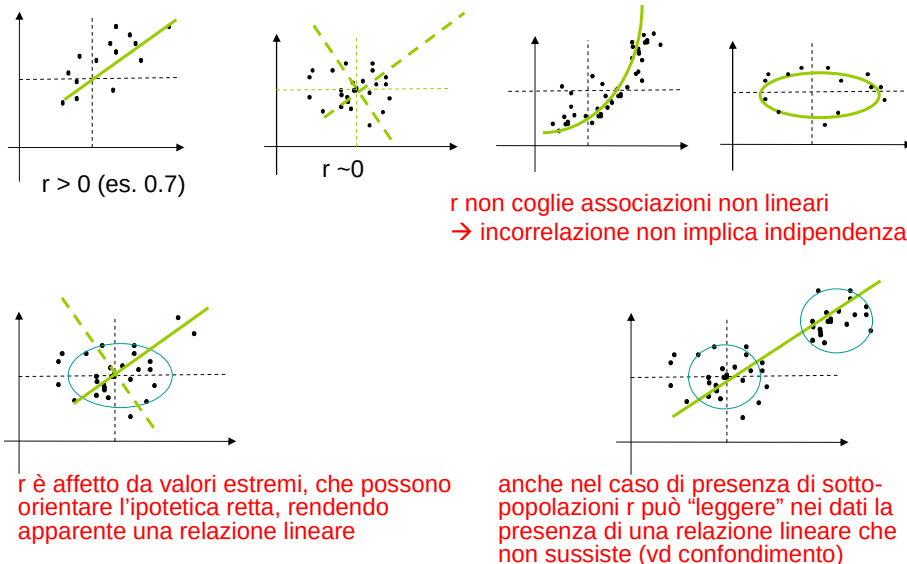
$$r_{xy} = \frac{\text{COV}_{xy}}{\text{std}_x \cdot \text{std}_y}$$

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



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

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

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

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

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



Prime sintesi delle tabelle di frequenza

Numero di ricoveri precedenti	n_i	%
0	4	$4/12 \cdot 100 = 33.3$
1	5	$5/12 \cdot 100 = 41.7$
2	3	$3/12 \cdot 100 = 25.0$
tot	12	

Moda = ?

Numero di ricoveri precedenti	n_i	N (cumulate)
0	4	4 (33.3%)
1	5	9 (75.0%)
2	3	12 (100%)
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



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 $\uparrow$

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

Mediana $\rightarrow$ n pari

Individuare le unità di rango $n/2$ e $n/2+1$

Mediana = somma delle loro modalità / 2

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

$$\text{mediana} = (58+65)/2 = 61.5$$



Practicals

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**

$$n/2=3 \quad n/2+1=4 \\ \text{mediana}=(58+65)/2 = 61.5$$



Practicals

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
tot	8	5

$$\text{Totale Numero di parti} = (0+0+0+0)+(1+1+1)+(2)$$

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

$$\text{Media} = 5 / 8 = 0.6$$



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 ricostruire 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	198×40=7920
F	190	16	190×16=3040
		56	10960

media = 10960 / 56 = 195.7



media ≠ (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	N_i
0	6	6
1	5	11
2	3	14
3	3	17
4	1	18
tot	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
	0	1277.43

media $m=55.7$

Ad esempio alla seconda riga:

$$(35-55.7) = -20.71 ; (-20.71)^2 = 161.65$$

$$\text{Varianza} = 1277.43 / 6 = 212.90$$

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

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



Attenzione a svolgere le operazioni in ordine:

Prima si calcolano gli scarti, $x_i - \text{media}$;

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	$(x_i)^2$	$(x_i-m)^2$
65	4225	86.22
35	1225	429.08
44	1936	137.22
43	1849	161.65
71	5041	233.65
63	3969	53.08
69	4761	176.51
	23006	1277.43

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?

simmetrica, attorno al valore medio centrale di 36 anni

→ Poteva avere una forma a campana?

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

→ Se sì, quale era il range dell'età?

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

→ La metà dei pazienti aveva meno di ?

35 anni

→ 1 su 4 aveva meno di ?

29 anni

→ 1 su 4 aveva più di ?

41 anni

→ l'età era fortemente variabile?

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

Calcolo dell'indice χ^2 (i)

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

Freq.
osservate

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

	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

Freq. attese

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

$$= 12 \cdot 9 / 22$$



Calcolo dell'indice χ^2 (ii)

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

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

$$= [(4 - 4.91)^2] / 4.91$$

	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

Passo 3: sommo tutte le distanze:

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



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

Medie di x e di y

17.6 · 43.3

180 – 162.4
= 17.6

COV_{xy}

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

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