

CdL Medicina, Tor Vergata, 2011-2012

Corso integrato "Fisica e Statistica"

Statistica Medica

Docente: dr. Simona Iacobelli

Informazioni (1)

Ricevimento Studenti

- Orario (indicativo*): Giovedì 15:30-17:30
- Luogo: stanza 99 Edificio H ("Fisica Medica")

(* A causa di impegni fuori sede, è possibile che alcune volte il Ricevimento non possa essere garantito nell'orario prestabilito. Ove possibile la Docente avvertirà gli Studenti in occasione delle lezioni, tuttavia per essere sicuri di poter reperire la Docente, si consiglia agli Studenti di richiedere un appuntamento)

Contatti: e-mail a: simona.iacobelli@uniroma2.it

CdL Medicina, Tor Vergata, 2011-2012

Corso integrato "Fisica e Statistica"

Statistica Medica

Docente: dr. Simona Iacobelli

Informazioni (2)

Libri di testo

- Statistica medica, autore: Bland, Ed. Apogeo ***
- Medical statistics at a glance, autori: Petrie e Sabin, Ed. Wiley-Blackwell ***
- Biostatistica (II ed), autori: Pagano e Gauvreau, Ed. Idelson-Gnocchi *

Materiale didattico (slide delle lezioni, esercizi etc): pagina STATISTICA MEDICA sul sito

<http://www.uniroma2.it/didattica/>

Introduzione - e motivazione



- Perché studiare Statistica in un corso di Medicina
 - Ruolo della Statistica nella ricerca, nella formazione, nei processi decisionali
 - Evidence-Based Medicine / Nursing / Prevention

Fonti per acquisire conoscenza in medicina

- Ricerca di base – in laboratorio
 - Su molecole, cellule ...
 - Su animali
 - Su organi
- Questa non esaurisce le necessità di conoscenza
 - Quando insorge il cancro del colon-retto? quali sono i fattori di rischio?
 - E' vero che i maschi sono più a rischio delle femmine? Di quanto?
 - Quali sono gli esiti della malattia (durata della sopravvivenza, probabilità di metastasi ...) e da cosa dipendono (trattamento, caratteristiche del paziente ...) ?
- Bisogna osservare i fenomeni di interesse sugli esseri umani, sani o malati, raccogliere informazioni, elaborarle e comunicarle in maniera appropriata

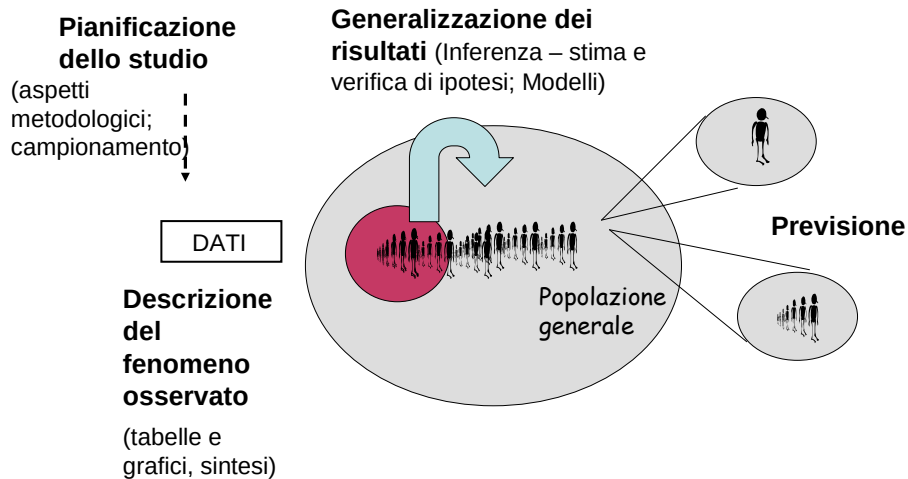
Errori in medicina , etica, EBM

- Si stima che il 15% degli errori nella pratica clinica sia di tipo cognitivo, ossia imputabile a:
 - Cattive informazioni
 - 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
- 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).
 - Evidence-Based Medicine
 - Evidence-Based Nursing
 - Evidence-Based Prevention

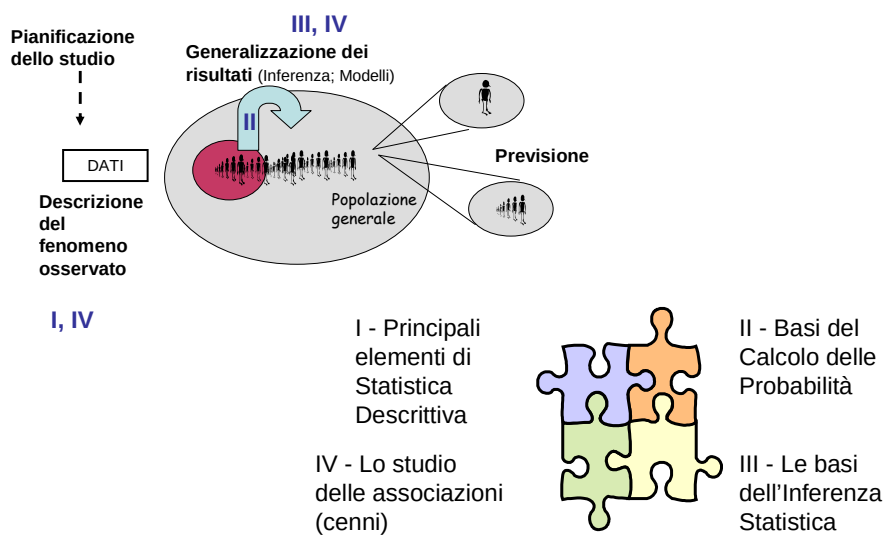
Il ruolo della statistica

- Ci interessano fenomeni che sono **variabili** (da individuo a individuo, da situazione a situazione, nel tempo, etc)
- Larga parte della conoscenza su fenomeni di interesse proviene dall'**OSSERVAZIONE** del fenomeno e da una adeguata **ELABORAZIONE DELLE INFORMAZIONI** raccolte mediante osservazione;
- Si impone la necessità di **MISURARE** i fenomeni (conoscenza di tipo **QUANTITATIVO**) e di metterli in **RELAZIONE** fra loro secondo metodi 'oggettivi'
 - il materiale di partenza è un insieme di **dati**
 - sono necessari **metodi** adeguati per la loro raccolta, elaborazione e interpretazione
- La Statistica è la disciplina che elabora strategie metodologiche e strumenti matematici per produrre conoscenza quantitativa di fenomeni collettivi sulla base di dati osservati, e ne supporta l'utilizzo a fini decisionali e di intervento.

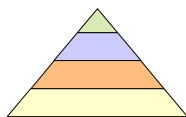
Momenti di intervento della Statistica



Programma del corso



Terminologia basilare e primi concetti



- Popolazione e Campione
- Tipi di dati. Classificazione dei caratteri

Oggetti di interesse

E' oggetto di interesse qualsiasi fenomeno che riguardi una moltitudine di 'individui' e si manifesti con **variabilità** da individuo a individuo.

	Italiani residenti in Italia il 9/10/2011	..., proprietà della casa, connessione a internet, condizione lavorativa, tempo impiegato per andare al lavoro, ...
	Pazienti con tumore della mammella	Età, menopausa, stadio del tumore, dimensioni del tumore, trattamenti fatti, durata della sopravvivenza, ...
	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 ...

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

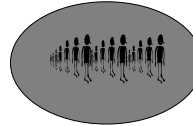
→ Essa è anche detta **POPOLAZIONE OBIETTIVO**

→ Può essere una popolazione **reale**, potenzialmente osservabile interamente (es. sondaggio fra gli italiani), o una popolazione **ideale, fittizia**, non identificabile

Esempio: Interesse studiare gli esiti di una nuova influenza

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



Nota: 'individuo' è usato per '**UNITA' STATISTICA**': il più piccolo elemento costitutivo della popolazione (paziente; famiglia; ospedale; cellula; etc)

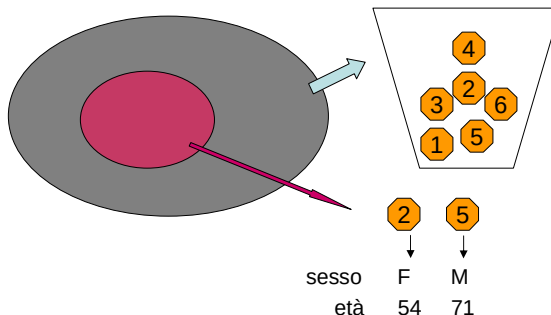
Campione

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

Caratteri

- Le caratteristiche di interesse delle unità statistiche sono dette **CARATTERI**, o **VARIABILI**
- I caratteri presentano (si esprimono attraverso) delle **MODALITÀ**, o determinazioni, o **VALORI**, diversi da unità a unità
- I caratteri vengono classificati secondo la seguente terminologia, che permette di definirne la natura e il tipo di operazioni che è possibile fare sulle sue modalità, per manipolarle, confrontarle e sintetizzarle

QUALITATIVI	QUANTITATIVI
SCONNESSI sesso → M,F patologia → ulcera, tumore gastrico, tumore intestinale, ...	DISCRETI numero di componenti (della famiglia) → 1,2,3,4, ... gravidanze precedenti → 0, 1, 2, 3, ...
ORDINATI titolo di studio → nessuno o licenza elementare, licenza media, licenza superiore, laurea stadio malattia → I,II,III, IV	CONTINUI 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!!)

Manipolazione di caratteri qualitativi

Ricodifica

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

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
- Spesso si usa la codifica numerica 0/1 (0=no=assenza, 1=sì=presenza)

Esempi

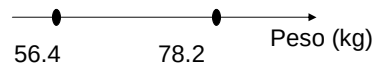
- Fumatore: sì/no
- Rispondente (alla terapia): sì/no
- Sesso = M/F, ovvero:
- Paziente maschio: sì/no

Caratteri Quantitativi

- Presentano modalità effettivamente numeriche, ottenute tramite conteggio o misurazione; sulle modalità è possibile eseguire operazioni matematiche; due modalità sono confrontabili mediante differenza o rapporto
- **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

Durate

- Un tipo particolare di carattere quantitativo è la misura della DURATA di un fenomeno, o meglio del tempo necessario al verificarsi di un evento
Esempi:
 - Tempo dalla diagnosi al decesso per pazienti oncologici
 - Tempo dall'ottenimento di remissione completa alla recidiva
- Quando la durata del periodo di osservazione ("follow-up") non è sufficiente a osservare l'evento per qualche paziente (**dati "censurati"**) si ha un tipo di dato che deve essere analizzato con metodi specifici, di **ANALISI DELLA SOPRAVVIVENZA**
- Quando l'evento non accade con certezza neanche lungo un follow-up infinito, si devono usare metodi per **RISCHI COMPETITIVI**

Questi metodi non possono essere trattati in un corso base di statistica medica

Manipolazione di caratteri quantitativi

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:

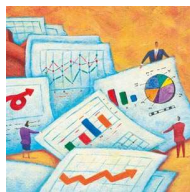
non fumatore	ex-fumatore	fumatore
--------------	-------------	----------

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 ($+$, $-$, $\cdot$, $/$, $\log$, ...)

Descrizione di un carattere: tabelle, grafici e indici sintetici



- Caratteri qualitativi
 - Frequenze relative e percentuali; frequenze cumulate. Moda. Tabelle doppie.
- Caratteri quantitativi
 - Concetto di Densità di Frequenza, istogramma; boxplot. Indici statistici:
 - di posizione: media, mediana, quartili
 - di variabilità: deviazione standard, varianza, coeff. di variazione
 - Curva Normale e altre forme della distribuzione

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

Insieme di DATI originario:

Tutte le unità del campione con le modalità osservate per ciascun carattere

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

freq. assoluta freq. relativa freq percentuale (%)

es. per la seconda modalità:

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

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

Numerosità
totale del
campione

totale
=1

totale
=100

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

E' il concetto di **proporzione**

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:

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

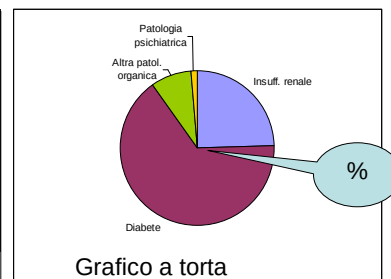
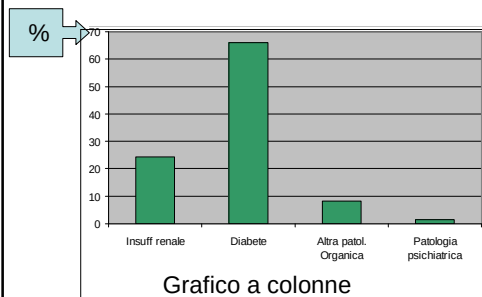
→ il 92% delle donne hanno al massimo 2 figli, e quindi solo l'8% ha più di 2 figli
etc



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



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.



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 le Femmine Fumatrici sono il 22%
=31/142

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

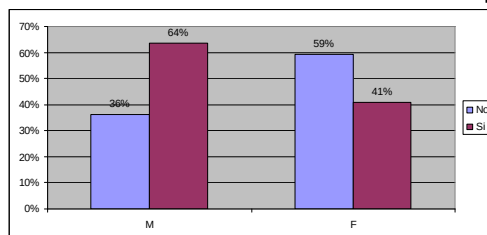
X = Sesso	Y = Fumo		totale
	no	si	
M	24	42	66
F	45	31	76
totale	69	73	142

Sesso	Y = Fumo		totale
	no	si	
M	24/66 = 36%	42/66 = 64%	
F	45/76 = 59%	31/76 = 41%	

C'è relazione fra X e Y? Per capirlo possiamo ricavare le % di riga (o di colonna)

Il 'profilo riga' (distribuzione di Y dato X) indica come si distribuisce Y quando X è fissato [Nota: condizionamento = restrizione]

Associazione o Indipendenza



La distribuzione del Fumo cambia a seconda del Sesso: c'è un certo **grado di associazione statistica** fra i due caratteri

Sesso	Y = Fumo		totale
	no	si	
M	36%	64%	66
F	59%	41%	76
totale	49%	51%	

Se invece fra Fumo e Sesso ci fosse una **indipendenza perfetta**, osserverei sia fra i M che fra le F il 51% di fumatori (vd. l'indice di associazione Chi-quadrato)

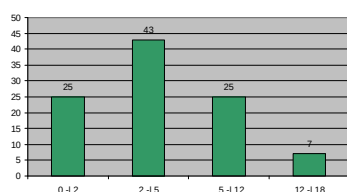
Grafici da tabelle di caratteri continui

Distribuzione di 56 pazienti pediatrici per età

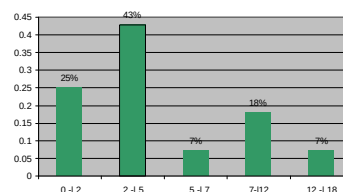
Età	freq.	%
0 - 2	14	25
2 - 5	24	43
5 - 12	14	25
12 - 18	4	7
	56	100

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**

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

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

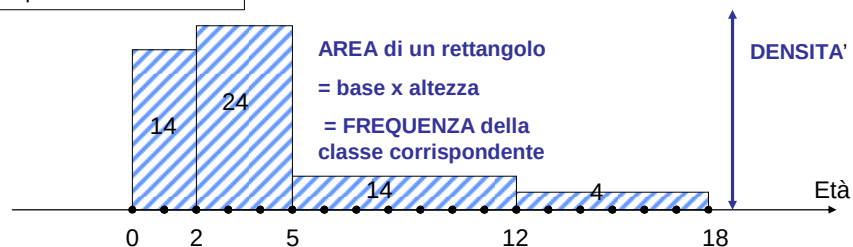
L'istogramma: il grafico della densità

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

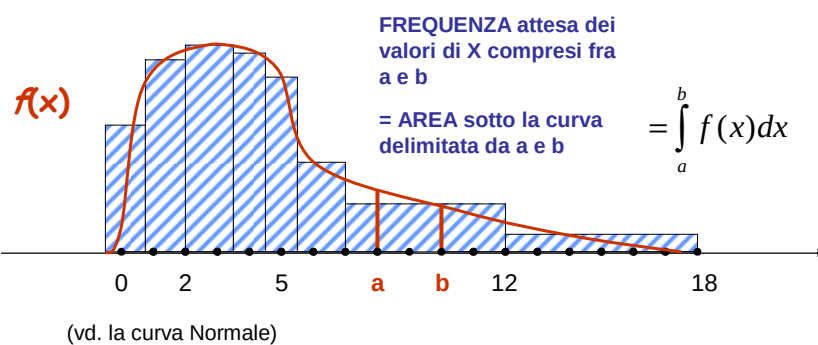
Le densità vengono poste in ordinata



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.

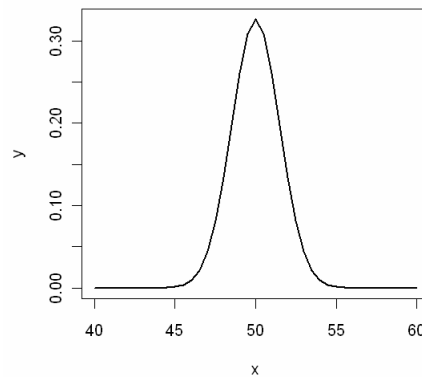
La matematica fornisce equazioni di curve continue che possono essere interpretate come curve di densità teoriche, corrispondenti a distribuzioni "ideali" di fenomeni quantitativi di interesse X.



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.



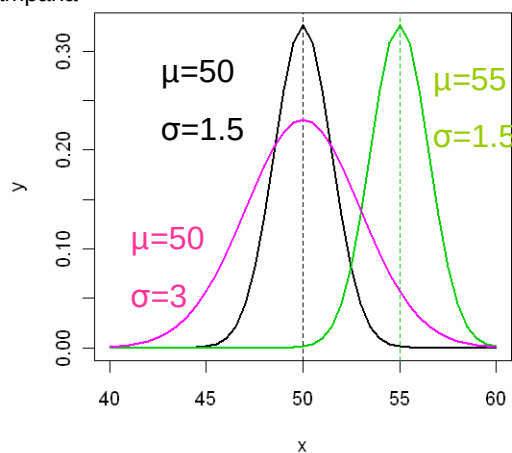
$$y = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

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"

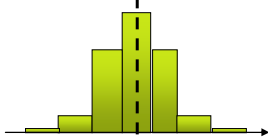
La formula che descrive la curva contiene 2 parametri μ e σ , che determinano rispettivamente dove si posiziona la curva rispetto all'asse x e quanto è ampia la campana



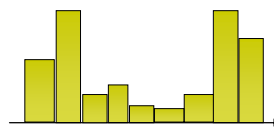
$$y = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

Varie forme della distribuzione

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

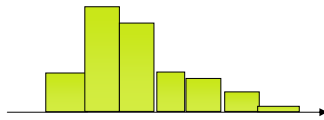


La forma "a campana" è tipica di fenomeni che possano essere ricondotti agli effetti "del caso", come l'altezza degli individui

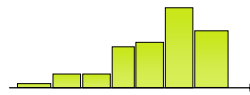


Distribuzione **BIMODALE**, cioè con la densità concentrata in due masse.

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)

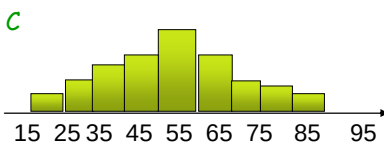
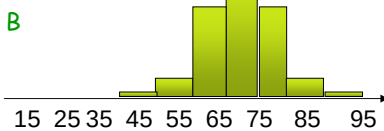
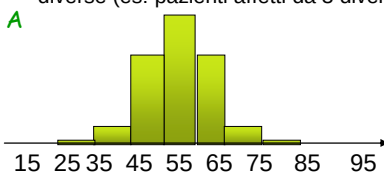


Tabelle e grafici di frequenza forniscono una rappresentazione completa dei dati.

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

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

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

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

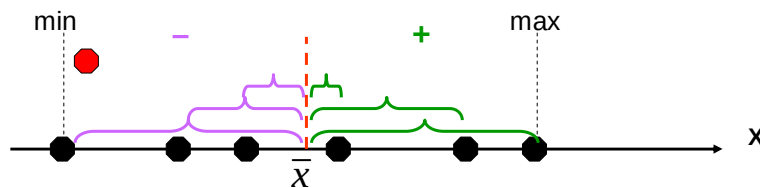
Voto	
26	24
18	24
28	24
72	72

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

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

Le differenze di ciascun valore osservato dalla media sono dette "scarti". 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)

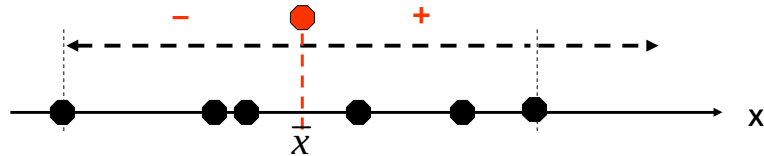
Naturalmente la formula vale anche nel caso di calcolo della media di K medie:

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

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

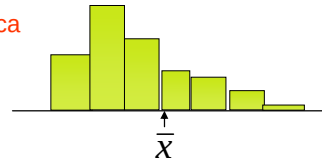


Dovendo BILANCIARE scarti positivi e negativi, e collocarsi nel centro (rispetto ai valori), la media è influenzata dai valori molto alti e dai valori molto bassi →

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

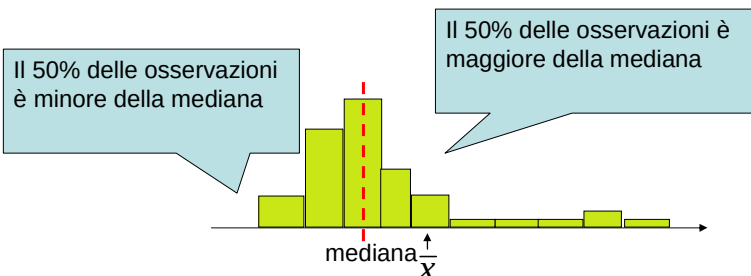
→ La media aritmetica è una sintesi insoddisfacente della distribuzione:

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



La mediana

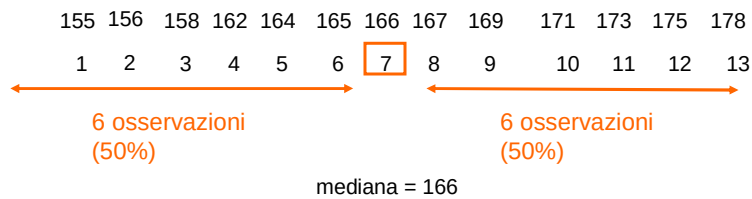
- La media aritmetica è una sintesi insoddisfacente della distribuzione:
 - Quando la distribuzione è (molto) asimmetrica
 - Quando si hanno uno o più valori estremi molto anomali
- In questi casi è più rappresentativa la mediana: il valore x tale che la metà delle osservazioni è $< x$ (e l'altra metà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

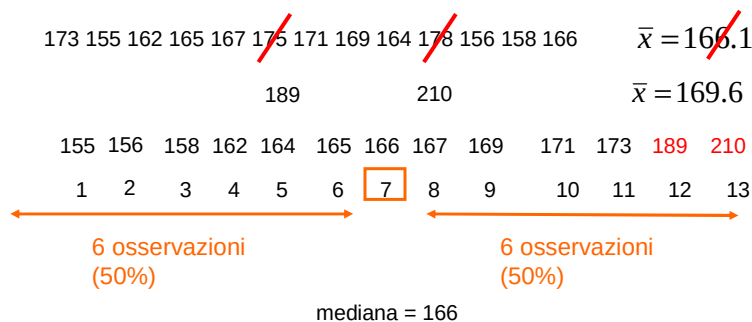
n pari → mediana = modalità intermedia fra quelle di posto $n/2$ e $n/2+1$

(ad esempio, se $n=6$, è la modalità centrale fra la 3° e la 4°)

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:

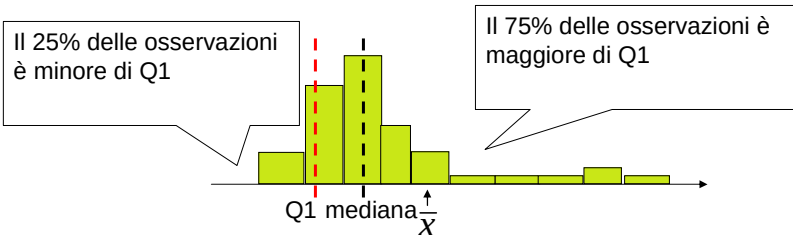


→ La mediana non cambia poichè l'ordinamento delle prime n osservazioni non cambia (invece la media cambia perché l'ammontare totale cambia)



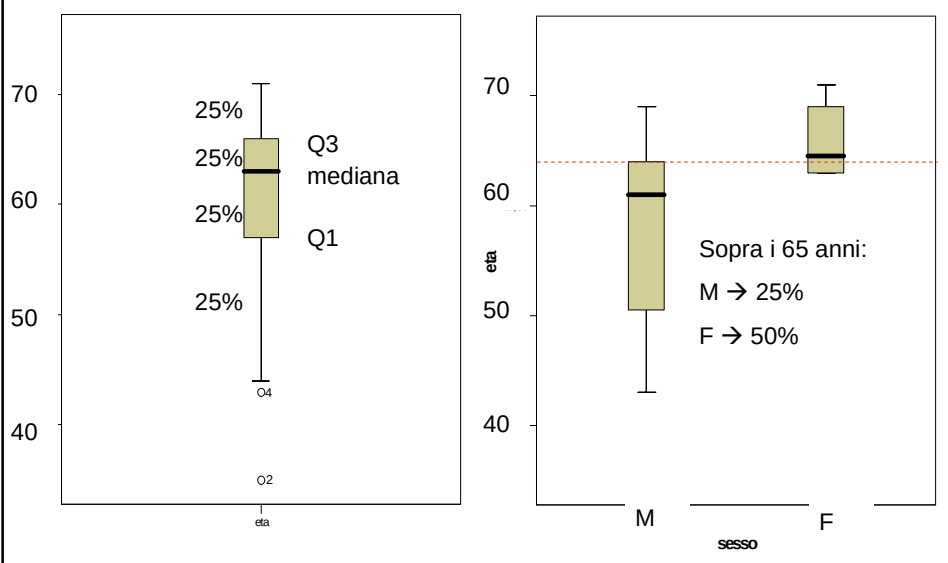
Generalizzazione della mediana: altri quantili

- 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.



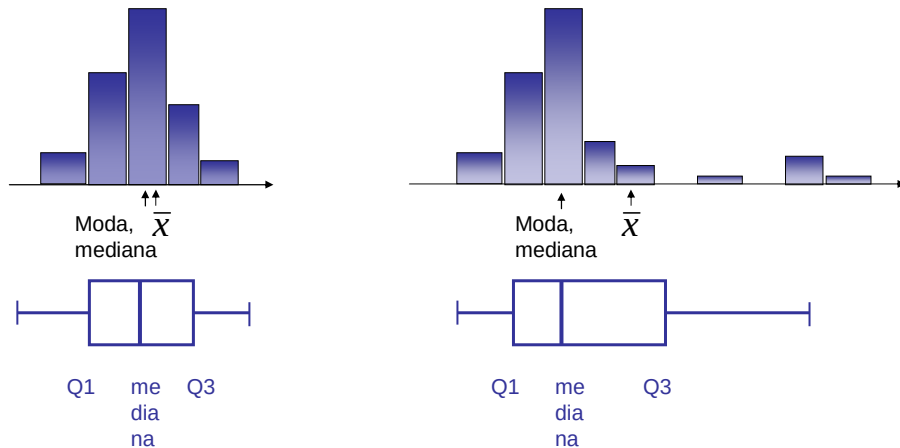
Nota: il concetto di mediana o quantile si applica anche nel caso di caratteri QUALITATIVI ORDINATI

Boxplot: un grafico basato sui quartili

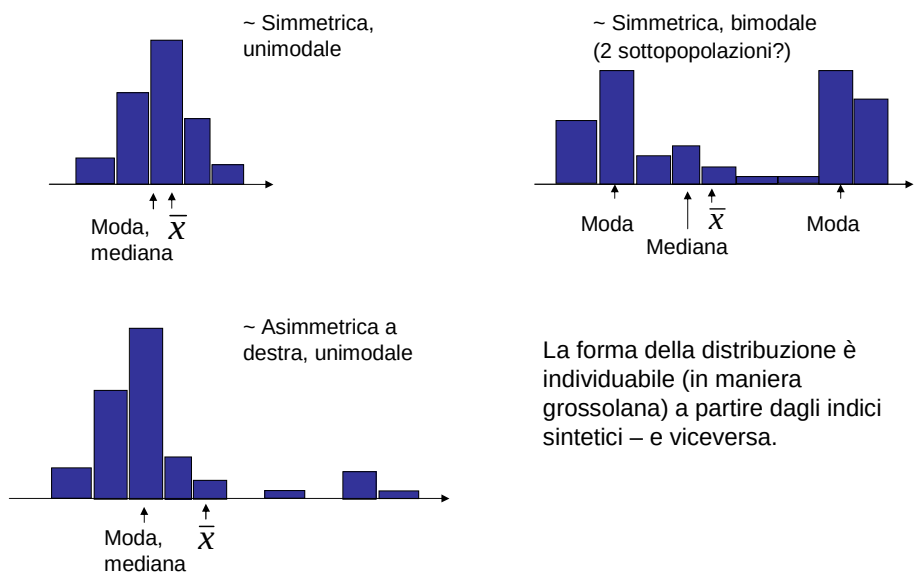


Forma della distribuzione e indici (1)

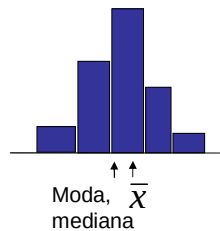
Vediamo alcune rappresentazioni tramite istogramma e boxplot; la forma della distribuzione è individuabile (in maniera grossolana) a partire dagli indici sintetici – e viceversa.



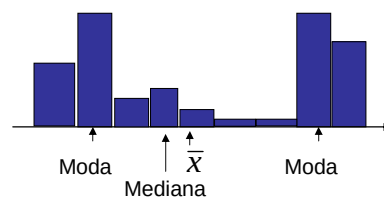
Forma della distribuzione e indici (2)



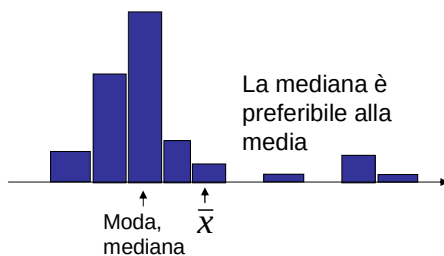
Appropriatezza degli indici



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

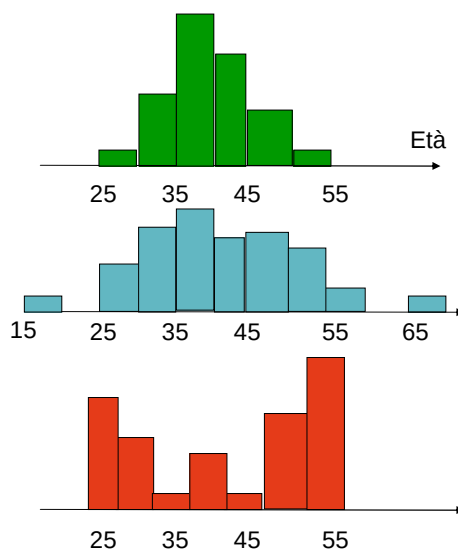


E' opportuno rimarcare la bimodalità: ne' media ne' mediana sono sintesi soddisfacenti



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*

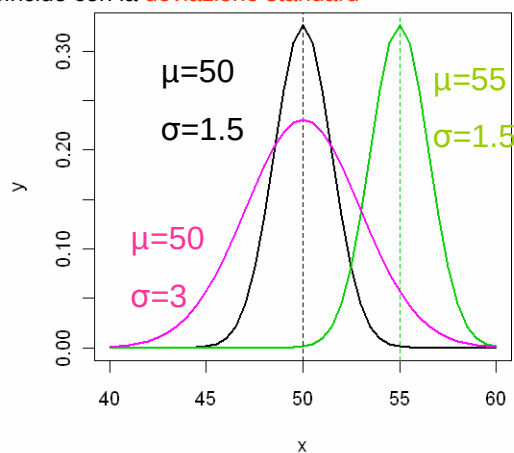
Nota: al denominatore si mette (n-1) anziché per n per motivi legati ad un concetto che affronteremo nella parte di inferenza

- 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

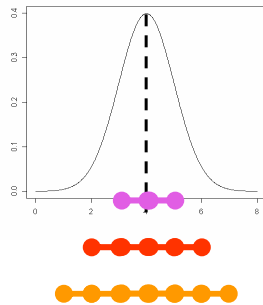
La curva Normale (ii)

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 %)

X = peso neonato: media = 3.2 kg, std = 0.5 kg

Y = peso madre: media = 60 kg, std = 4.5 kg

Z = altezza neonato: media = 51 cm, std = 2.5 cm

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

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

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

Y : CV = $(4.5 \text{ kg} / 60 \text{ kg}) = 7.5$

Z : CV = $(2.5 \text{ cm} / 51 \text{ cm}) = 4.9$

→ I neonati sono più variabili rispetto al peso che all'altezza (circa tre volte tanto) e in termini di peso sono variabili il doppio delle 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

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

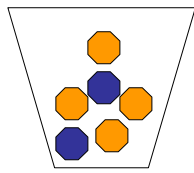
→ 1 su 4 aveva più di ?

41 anni

→ l'età era fortemente variabile?

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

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
 - La Binomiale (e la Poisson)
 - La Normale

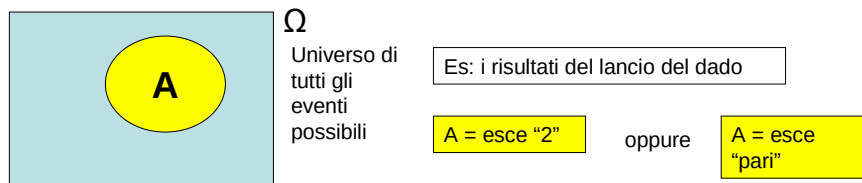
Argomenti

- Eventi e probabilità: concetti intuitivi + formalizzazione matematica.
 - Probabilità e ... scommesse
- Calcolare le probabilità di eventi definiti in base a altri eventi
- Applicazione del calcolo delle probabilità: astrazione: quello che osserviamo sono numeri estratti da un'urna. Il risultato dell'estrazione è descritto da una funzione di distribuzione o di densità di probabilità.
 - Funzioni di densità per caratteri continui. Normale e altre.
 - Distribuzioni utili per caratteri discreti / dicotomici: Binomiale e Poisson
 - Risk Ratio, Odds Ratio, Rate Ratio
- Formula di Bayes. Applicazione ai test diagnostici.

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: $p(\Omega)=1$

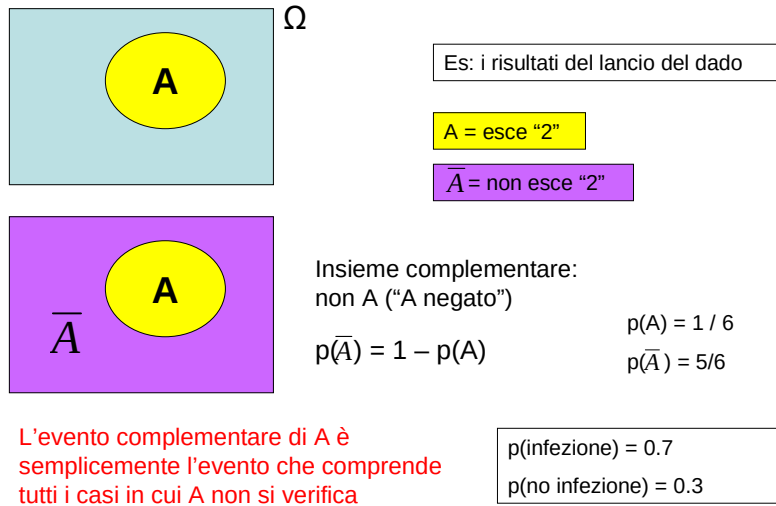
Nei casi più semplici, dove la prova ha un numero finito di possibili esiti, e tutti sono ugualmente probabili, $p(A)$ = numero casi favorevoli / numero casi possibili.

Rispettivamente nei due esempi:

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

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

Evento complementare



Probabilità e Odds

- Nell'ambito delle scommesse sportive le probabilità relative agli esiti di una gara sono implicitamente espresse nelle quote
 - Cosa vuol dire che la vittoria di una squadra è quotata 5 a 1?

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

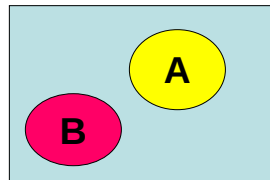
→ Rapporto fra Casi Favorevoli e Casi Possibili

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

→ Rapporto fra Casi Favorevoli e Casi Contrari

- Vincita quotata 5 a 1 vuol dire odds(vittoria)=1/5 cioè pr(vittoria)=1/6
- Gli Odds sono una misura della probabilità che viene utilizzata in Statistica Medica ed Epidemiologia

Intersezione → “e” (And)



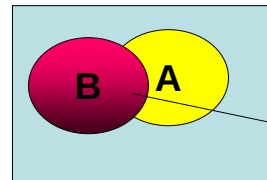
A, B
insiemi
disgiunti

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”
 $A \cap B$

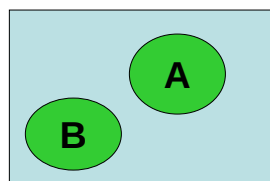
A = esce “pari”

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” (Or) (1)



A, B
insiemi
disgiunti

A = esce “2”

B = esce “3”

$A \cup B$ = esce
“2” oppure “3”

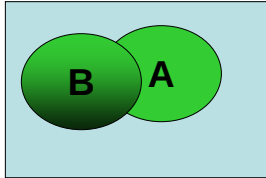
$$p(A \cup B) = p(A) + p(B)$$

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” (Or) (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 $\cup$ 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:

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

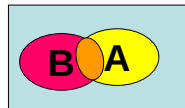
L'evento B non modifica l'evento A in modo materiale, concreto; quello che si modifica è la probabilità

Formule per prob. condizionata e intersezioni

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

Prob. di A condizionata a B

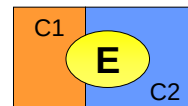
$$\begin{aligned} p(A \cap B) &= p(A | B) \cdot p(B) \\ &= p(B | A) \cdot p(A) \end{aligned}$$



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

Caso particolare per eventi A e B indipendenti

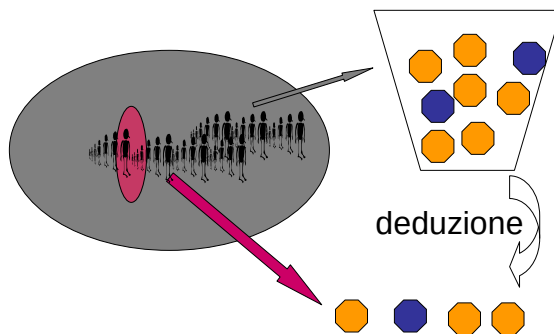
$$p(C1 | E) = \frac{p(E | C1) \cdot p(C1)}{p(E | C1) \cdot p(C1) + p(E | C2) \cdot p(C2)}$$



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

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

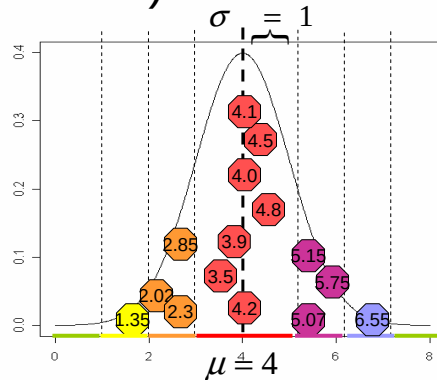
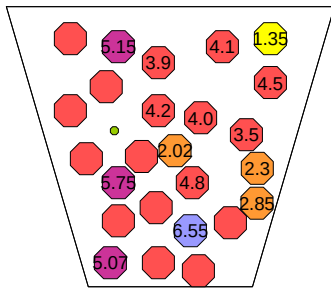
Quesito: $P(A \text{ o } B)$

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

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

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

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:

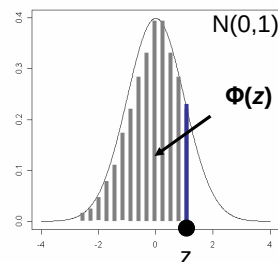
- Molto "densi" attorno a 4; il 68% distanti meno di 1, in eccesso o in difetto, ossia fra 3 e 5
- Un po' meno densi fra 2 e 3 (circa il 14%) o fra 5 e 6 (ancora 14%)
- Soltanto il 2% fra 1 e 2, o fra 6 e 7
- Praticamente nessuno <1 oppure >7 : in tutto, 3 su 1000

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:

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

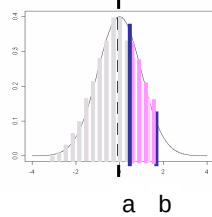
$$\Phi(-z) = 1 - \Phi(z)$$

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

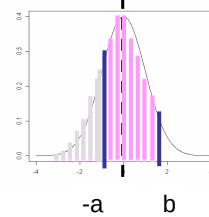
Calcolare probabilità per la Normale

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

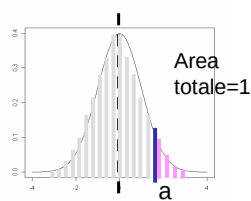
$$\text{Area}(a, b) = \Phi(b) - \Phi(a)$$



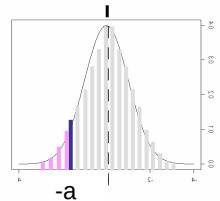
$$\text{Area}(-a, b) = \Phi(b) - (1 - \Phi(a))$$



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



$$= \Phi(-a)$$



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

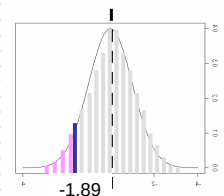
Informazioni: $\mu=23$ $\sigma^2=7$

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

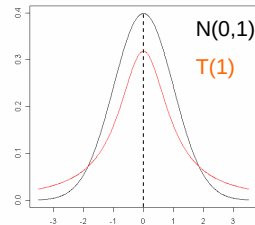


Altre distribuzioni continue

DISTRIBUZIONE T di Student $T(g)$

La curva di densità è a campana e simmetrica attorno all'asse $\mu=0$, con code più alte della $N(0,1)$

(se g grande (>30) è molto simile alla $N(0,1)$)

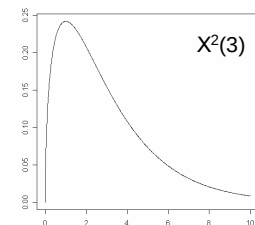


DISTRIBUZIONE Chi Quadrato $\chi^2(g)$

$X \geq 0$

La densità $f(x)$ può avere forme diverse, a seconda di g

Entrambe sono identificate da un unico parametro g detto "gradi di libertà"



Distribuzioni di probabilità discrete



Es: 3 lanci consecutivi di una moneta, X = numero di "Testa"
 X variabile aleatoria

Risultati possibili:

C	C	C	C	C	T	C	T	T	T	T	T
			C	T	C	T	C	T			
			T	C	C	T	T	C			

Distribuzione di probabilità di X

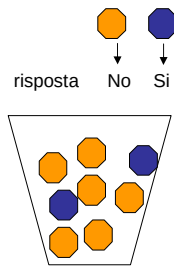
$X = 0$	$X = 1$	$X = 2$	$X = 3$
$p(X=0)$ $= 1/8$	$p(X=1)$ $= 3/8$	$p(X=2)$ $= 3/8$	$p(X=3)$ $= 1/8$

Nota: la somma delle probabilità di tutti i risultati possibili per X vale 1

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

$\pi = 0.25$, N estrazioni



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

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

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

$$0! = 1$$

Quanti pazienti rispondono fra 20 trattati?

$$N \cdot \pi = 20 \cdot 0.25 = 5$$

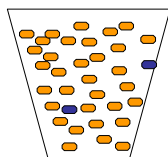
Qual è la prob. che rispondano 10 pazienti?

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

La legge di Poisson

Nel contesto di un fenomeno assimilabile all'osservazione di un evento RARO (π molto piccola) in una serie INFINITA di prove o lungo un INTERVALLO di tempo (o su una superficie, e simili), per il numero totale (X) di eventi osservati si possono usare le seguenti formule

Il parametro λ (detto TASSO) rappresenta il numero di eventi atteso in infinite prove ovvero da la prob. di evento nell'intervallo di osservaz.



- Il numero medio "atteso" di eventi è λ
- La probabilità di osservare esattamente x eventi è data da:

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

Dove: $e^{-\lambda} = \frac{1}{e^{\lambda}}$

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

$$k^0 = 1$$

Se dobbiamo contare il numero X di eventi in N prove con $\pi = P(\text{Evento}) \approx 0$ e $N \gg 0$ usiamo la Poisson con $\lambda = N \cdot \pi$

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

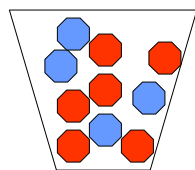
Usiamo la Poisson, con $\mu = N \cdot \pi = 50 \cdot 0.001 = 0.05$

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

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

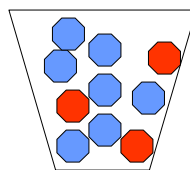
$$P(X>0) = 1 - 0.95 = 0.05 = 5\%$$

Confronto di probabilità: il Risk Ratio



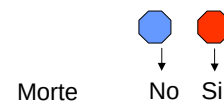
Popolazione (urna) dei soggetti ESPOSTI ad un fattore di interesse, ad es. sottoposti a tecnica chirurgica A

$$\Pr(\text{Morte} | A) = 0.60$$



Popolazione (urna) dei soggetti esposti NON ESPOSTI, ad es. sottoposti a tecnica chirurgica B

$$\Pr(\text{Morte} | B) = 0.30$$



Molto spesso in medicina ed epidemiologia si vogliono confrontare le probabilità che si verifichi un evento fra due gruppi. Quasi mai il confronto si fa calcolando la differenza fra le due probabilità.

Confronto di probabilità: il Risk Ratio

$$\Pr(\text{Morte} | A) = 0.60$$

$$\Pr(\text{Morte} | B) = 0.30$$

$$\text{RISK RATIO} \quad RR = \frac{\pi_E}{\pi_{NE}} = \frac{0.60}{0.30} = 2$$

= 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

= 1.6 : Esposizione aumenta il rischio di evento del 60%

= 2: aumento del 100%

= 0.7: Esposizione diminuisce il rischio di evento del 30% (lo riduce al 70% del rischio dei Non Esposti)

Confronto di probabilità: Odds Ratio

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

→ 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

- Anche il rapporto fra gli odds è una misura di confronto di rischi, che in certi casi (es. studi caso-controllo) è necessario valutare in alternativa al RR:

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

= 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

Tassi e Rate Ratio



Il concetto di TASSO visto nella Poisson è valido in generale per descrivere la “forza” del verificarsi dell’evento non rispetto a un “numero di prove”, che è da ritenersi infinito, ma lungo un intervallo continuo

Es. contagi di TBC in un ospedale lungo un arco di tempo

$\lambda \approx \text{Prob}(E)$ durante l’intervallo d

→ Viene stimato come rapporto fra Casi osservati e “ampiezza dell’intervallo” – quest’ultima definita in alcuni casi come durata complessiva del periodo di osservazione espressa in anni-persona (il concetto viene approfondito in Epidemiologia).

$$\text{RATE RATIO} \quad RR = \frac{\lambda_E}{\lambda_{NE}}$$

(interpretabile come il risk Ratio, fornisce una quantificazione della differenza del rischio fra Esposti e Non Esposti)

Importanza delle probabilità a priori

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

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: $P(C1|E)$. Formula di Bayes:

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

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

$p(C1|E)$: 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

Utilizzare la formula di Bayes

Questa formula trova applicazione in 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)

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

Confronto fra le probabilità delle cause:

A priori: $\frac{p(C1)}{p(C2)}$

A posteriori:
cioè dopo aver
acquisito dei DATI
(aver osservato che
si verifica l'evento
E)

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

RR

L'aggiornamento delle probabilità a priori sulla base del dato osservati avviene tramite il Risk Ratio per il verificarsi di E in presenza dell'una o dell'altra causa

Da questi concetti si origina la Statistica Bayesiana

...

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)

		Risultato del test diagnostico	
		Test +	Test –
Situazione (incognita) del soggetto	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 - SENSITIVITA'

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

=1 - SPECIFICITA'

$\Pr(\text{Errore}) = \Pr(\text{Errore} \& \text{Malato}) + \Pr(\text{Errore} \& \text{Non Malato}) =$

$= \Pr(\text{Errore} | \text{Malato}) \Pr(\text{Malato}) + \Pr(\text{Errore} | \text{Non Malato}) \Pr(\text{Non Malato}) =$

$= \Pr(\text{Test} - | \text{Malato}) \Pr(\text{Malato}) + \Pr(\text{Test} + | \text{Non Malato}) (1 - \Pr(\text{Malato})) =$

$= (1 - \text{sensitività}) \Pr(\text{Malato}) + (1 - \text{specificità}) (1 - \Pr(\text{Malato}))$

Le probabilità nel test diagnostico (3)

Situazione (incognita) del soggetto	Risultato del test diagnostico		SENSITIVITA': $p(\text{Test} + \text{Malato})$ SPECIFICITA': $p(\text{Test} - \text{Non Malato})$ FN: $p(\text{Test} - \text{Malato}) = 1 - \text{SENS}$ FP: $p(\text{Test} + \text{Non Malato}) = 1 - \text{SPEC}$
	Test +	Test -	
	ok	Falso negativo	
Malato	ok	Falso negativo	
Non Malato	Falso positivo	ok	

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" → 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)

Situazione (incognita) del soggetto	Risultato del test diagnostico		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 = $P(\text{Malato})$
	Test +	Test -	
	ok	Falso negativo	
Malato	ok	Falso negativo	
Non Malato	Falso positivo	ok	

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

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

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

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

Quesito: $P(M | TP)$

Si applica la formula di Bayes:

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

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

Stima di Sensitività e Specificità

Si vuole stimare la sensibilità 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

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

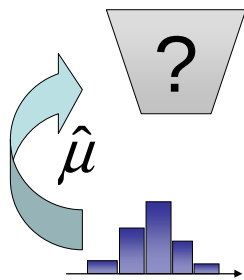
$$\text{specificità} \approx \frac{vn}{n - m} = \frac{vn}{vn + fp}$$

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)

L'inferenza statistica: risalire dal campione alla popolazione. Principi.

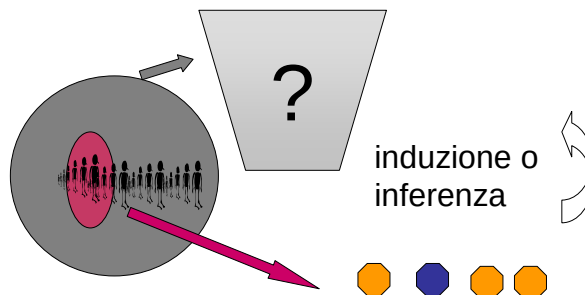


- Stima puntuale e intervallare
 - Stimatori non distorti
 - Intervalli di confidenza
- La verifica delle ipotesi
 - Significatività
 - Elementi formali
- Rapporto tra IC e Test

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.**

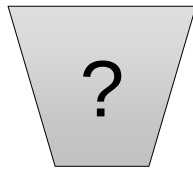
Guardiamo i **DATI** in un Campione. Essi sono come un insieme di **palline estratte dall'urna.** Allora, **i dati del campione ci danno informazione su com'è fatta l'urna.**



La stima statistica

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

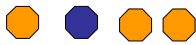
Dobbiamo dare una valutazione ai parametri della distribuzione.



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

Diversi principi per fare Inferenza

- La teoria statistica propone diversi approcci per risalire dal campione alla popolazione, ossia fare Inferenza statistica. I principali sono:

- l'approccio frequentista
- l'inferenza basata sulla verosimiglianza
- l'approccio Bayesiano

- Questo corso si limita ai principali elementi dell'approccio **frequentista**, basato sul cosiddetto principio del campionamento ripetuto. In sostanza:

- Si individuano delle "buone regole" per fare inferenza, a prescindere dai dati osservati
- Una "regola" è "buona" se ha un **buon comportamento con riferimento alla possibilità di ripetere l'estrazione del campione (e quindi la procedura inferenziale) infinite volte**

Distribuzione della media campionaria

“Comportamento”, ossia: che valori assume la media aritmetica di infiniti campioni estratti dalla stessa popolazione

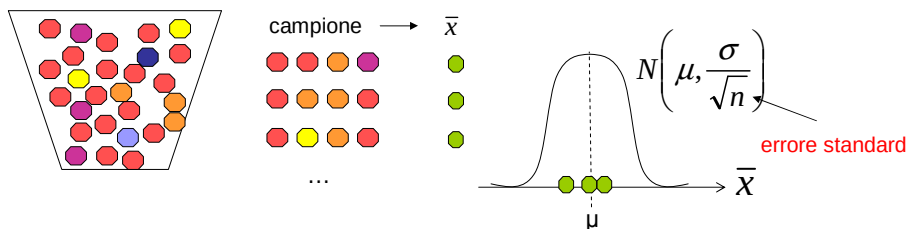
L'urna rappresenta una popolazione con media μ e varianza σ^2 .

Estraiano ripetutamente campioni, ciascuno contenente n osservazioni, e ne calcoliamo la media aritmetica. Per ogni campione, essa assume un valore diverso.

Come si distribuiscono questi valori?

Attenzione: questo risultato vale **se** sussiste almeno una delle seguenti due condizioni:

- La popolazione è Normale
- Il campione (n) è grande



Proprietà della media campionaria

La media calcolata sul campione può esserci utile per conoscere la media della popolazione?

Giudichiamo la “bontà” del suo “comportamento” ripetendo il campionamento all’infinito:

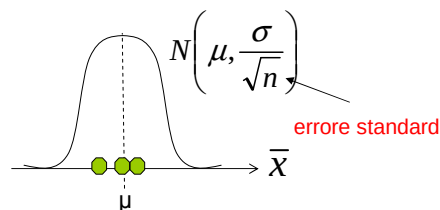
- La distribuzione di $\bar{X}$ è simmetrica e centrata su μ ; quindi estraendo ripetutamente campioni, di volta in volta avremo un valore più alto (sovrastima) o più basso (sottostima) di μ , in maniera bilanciata.

-Nel 95% dei campioni, la media dei dati cade nell’intervallo $\mu \pm 1.96 \cdot \sigma/\sqrt{n}$

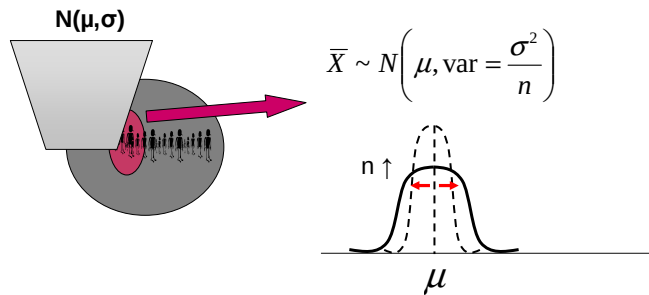
-Se l’errore standard è piccolo, ossia la Normale è “stretta”, concentrata, la media del campione cade vicino alla media μ della popolazione:

-Questo succede se la varianza σ^2 della popolazione è piccola, e se la numerosità del campione è grande:

- tanto più il campione è grande, tanto più vicina la media dei dati è vicina a μ .



Variabilità campionaria



In generale, e non solo per la media campionaria, la varianza della distribuzione di una statistica calcolata nel campione può diminuire se:

- la varianza della popolazione σ^2 diminuisce → condizioni sperimentali controllate, omogeneità dei soggetti inclusi nello studio
- la numerosità del campione n aumenta

Stimatori

Nella stima puntuale, la “regola” da applicare per produrre una stima è una funzione che indica come calcolarla dai dati. Questa funzione è chiamata **STIMATORE**.

Le principali proprietà che indicano che è una “buona regola” sono:

→ 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

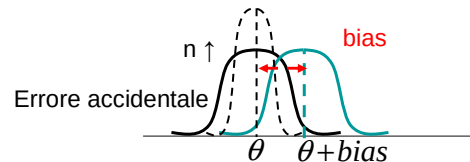
Fra due bilance è non-distorte, preferisco quella che non dà molti valori “estremi”

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)



Stimatori non distorti

Per media, varianza e frequenza relativa, le stime calcolate nel campione sono valide poichè si tratta di stimatori non distorti.

Per un carattere continuo, con media μ e varianza σ^2 :

$$\hat{\mu} = \bar{x} = \frac{\sum_{i=1}^n x_i}{n} \quad \hat{\sigma}^2 = s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

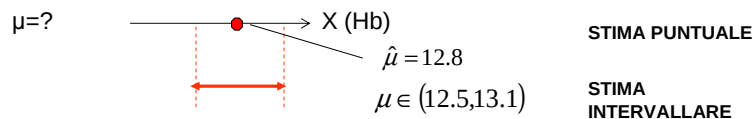
(Se in s^2 dividiamo per n anziché $(n-1)$ lo stimatore è distorto)

Per un carattere dicotomico, con probabilità π di “successo” $X=1$ (ad es. il verificarsi di un evento), si prende semplicemente la frequenza relativa (percentuale) (o: proporzione) di successi osservata nel campione:

$$\hat{\pi} = f = \frac{\text{n.ro casi con } (x = 1)}{n}$$

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.



- 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

- Il metodo più diffuso per calcolare "buoni" intervalli di stima è quello dell'intervallo di confidenza.
- Il metodo è caratterizzato da un livello α , detto "**livello di confidenza**". Valori tipici per α sono 0.95 o 0.90
- Immaginando di disporre di infiniti campioni, costruire l'**intervallo di confidenza di 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.
- In generale, con livello di confidenza $1-\alpha$: l'IC contiene il parametro da stimare tranne che con lo $\alpha\%$ dei possibili campioni

Intervalli di Confidenza al 95%

Per un **carattere continuo**, con **media μ** e **varianza σ^2** nota, e se:

- Ha distribuzione Normale nella popolazione

oppure

- Il campione è grande

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

(la formula si complica un po' se la varianza non è nota: vd dispensa)

Per un **carattere dicotomico**, con **probabilità π** di "successo" $X=1$, solo se campione è grande:

$$\left(\hat{\pi} - 1.96 \cdot \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}, \hat{\pi} + 1.96 \cdot \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \right)$$

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%

$$\bar{x} = 160$$

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

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

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

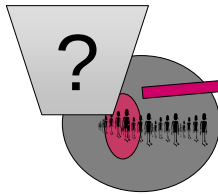
Verifica di ipotesi sulla Popolazione

Oltre che la stima dei parametri che caratterizzano la Popolazione, ci interessa **rispondere a una domanda precisa**, “dimostrando” una certa **ipotesi** su un carattere o sulla relazione fra due caratteri. Es:

- a) Possiamo affermare che i soggetti che prendono un certo farmaco sono soggetti ad anemia? [confronto del carattere con un valore di riferimento]
- b) Il farmaco A è più efficace di B per l'ottenimento della Risposta? [confronto del carattere fra due gruppi]

I dati osservati possono mostrare differenze a supporto di queste ipotesi.

Problema: queste differenze possono essere frutto del puro caso.



Es:

- a) Nei pazienti trattati, l'emoglobina media era pari a 9.8
- b) Nei due gruppi, i trattati con A hanno risposto nel 77% dei casi, nel gruppo B solo nel 54%

Impostazione del test statistico

Le ipotesi

Si formulano due ipotesi sulla Popolazione = sulla composizione dell'urna:

- **H0 [ipotesi di base, o “nulla”]** : è l'ipotesi di riferimento, corrispondente a uno stato “neutro” di conoscenze
- **H1 [ipotesi alternativa]** : rappresenta l'ipotesi sperimentale, la “nuova conoscenza”, il risultato atteso della ricerca.

Esempio a) Anemia?

Parametro: media dell'emoglobina, μ

H0: L'emoglobina media nella popolazione è pari a 12 [valore soglia di riferimento. Fonte: dato convenzionale; letteratura]

H1: L'emoglobina media nella popolazione è inferiore a 12

H0: $\mu=12$

H1: $\mu<12$ [test “unilaterale”, one-sided. Test “bilaterale” se H1: $\mu\neq12$]

Impostazione del test statistico

Le ipotesi

Si formulano due ipotesi sulla Popolazione = sulla composizione dell'urna:

- **H0 [ipotesi di base, o "nulla"]** : è l'ipotesi di riferimento, corrispondente a uno stato "neutro" di conoscenze;
- **H1 [ipotesi alternativa]** : rappresenta l'ipotesi sperimentale, la "nuova conoscenza", il risultato atteso della ricerca.

Esempio b) Trattamento A più efficace di B?

parametro: $\Pr(\text{Risposta})$

H0: non c'è differenza fra la prob. di Risposta fra farmaco A e farmaco B

H1: c'è differenza, e A è migliore di B.

H0: $\pi_A = \pi_B$

H1: $\pi_A > \pi_B$ [test "unilaterale", one-sided. Test "bilaterale" se H1: $\pi_A \neq \pi_B$]

Logica del test statistico

Uso dei dati osservati

Nei nostri esempi, i dati osservati nel campione sono:

- a) Il valore campionario, da confrontare col valore di riferimento
- b) La differenza fra i valori campionari nei due gruppi

Sotto H0, nel campione non ci aspettiamo differenze. "E' però molto probabile osservare "un po'" di differenza per la variabilità casuale.

Se osservo una differenza nel campione, possono essere successe due cose:

- H0 è vera, il caso ha originato la differenza osservata
- H0 è falsa, la popolazione è piuttosto come descritto da H1

Un test statistico stabilisce un criterio per scegliere fra queste due affermazioni, sulla base dei dati (e del modello probabilistico utilizzato per descrivere come essi vengono generati)

Logica del test statistico

Criterio probabilistico

Quanto è probabile che l'urna H_0 generi i dati che sono stati osservati o dati anche meno attesi, e più verosimili sotto H_1 ?

→ Estremamente improbabile. Non crediamo all'azione del caso, H_0 è falsa. Va rigettata in favore di H_1 .

I dati CONFUTANO H_0 e supportano H_1

I dati vengono detti **SIGNIFICATIVI**. H_0 è **RIGETTATA** a favore dell'ipotesi H_1 .

→ Se NON è molto improbabile, non posso affermare che i dati confutano H_0 . Non posso rigettare H_0 .

Non ho evidenza che la differenza osservata nel campione sia presente nella popolazione, e quindi sia riproducibile in un altro campione estratto.

Interpretazione del test statistico

P-value

La probabilità che assumendo come vera l'ipotesi H_0 si osservino dati come quelli osservati o dati anche più "lontani" da H_0 e "vicini" invece ad H_1 è chiamata **P-VALUE**.

P-value molto piccoli (es. 0.001, 0.01) sono un buon supporto al rigetto dell'ipotesi nulla H_0 .

P-value grandi, es. 0.4 o superiori, sono indice di sostanziale assenza di evidenza a supporto di H_1 : ci sono pochi dubbi sull'opportunità di mantenere l'ipotesi di riferimento H_0

Nei casi intermedi, il giudizio va affidato all'esperienza e dipende dal contesto.

Spesso tuttavia si utilizza un criterio rigido: si fissa un "**livello di significatività**" α , per esempio 0.05

- $p < 0.05$ "significativo" (nessuna distinzione fra $p = 0.003$ e 0.045 ???)

Test statistico in pratica

P-value in pratica

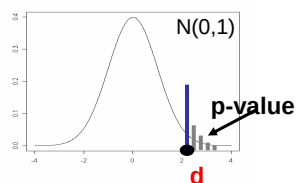
Il p-value viene calcolato dai software statistici, ma vi sono anche delle tavole per i calcoli più comuni.

Esso è l'area sotto la curva di densità dei risultati attesi sotto l'ipotesi H_0 , corrispondente alla zona delimitata dal valore osservato della statistica test, dalla parte indicata dall'ipotesi H_1

Esempio b) $H_0: \pi_A - \pi_B = 0$ $H_1: \pi_A - \pi_B > 0$

La differenza calcolata nel campione, opportunamente standardizzata, segue (sotto certe condizioni) una distribuzione Normale (0,1).

La differenza osservata è pari a d.



Con H_1 bilaterale anche la coda sinistra è una zona 'critica': il p-value è il doppio dell'area in figura

Test statistico in pratica

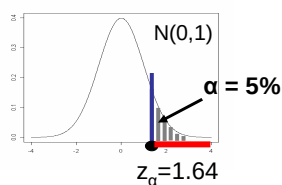
Metodo della regione di rifiuto

Assumiamo un criterio rigido per valutare il p-value: se $p < 0.05$ rigettiamo H_0 .

Troviamo quindi una zona, dalla parte indicata da H_1 , a cui corrisponde, sotto la curva di densità dei risultati attesi sotto l'ipotesi H_0 , un'area pari a 0.05

Se la statistica test cade in questa zona, possiamo rigettare H_0 al livello del 5%

Esempio b) $H_0: \pi_A - \pi_B = 0$ $H_1: \pi_A - \pi_B > 0$



Zona oltre z_α è regione di rifiuto

Con H_1 bilaterale avremmo avuto una metà della zona di rifiuto sulla coda sinistra: a ogni coda corrisponderebbe non area α bensì area $\alpha/2$. $Z_{0.025} = 1.96$

Es: test su una media – p-value

Vd. dispensa

Assumendo che in una popolazione di soggetti sani il valore della pressione sanguigna si distribuisca secondo una distribuzione Normale con media 150 e deviazione standard pari a 15, e osservando un campione di 20 soggetti con media campionaria pari a 160, verificare l'ipotesi che tali soggetti siano stati estratti da quella popolazione, contro l'ipotesi che siano soggetti ipertesi.

Ipotesi e tipo di Test: $H_0 : \mu = 150$ vs $H_1 : \mu > 150$; T-test per la media

Ipotesi del test: la distribuzione è Normale e la varianza è nota: soddisfatte.

Calcolo della statistica test:
$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{160 - 150}{3.35} = 2.98$$

Metodo del p-value: $\Phi(2.98) = 0.999$ (sulle tavole)
 $p\text{-value} = (1 - 0.999) = 0.001$

Conclusione: Il valore osservato è significativamente diverso da 150: possiamo concludere che i soggetti osservati sono ipertesi rispetto alla norma.

Es: test su una media – regione di rifiuto

Assumendo che in una popolazione di soggetti sani il valore della pressione sanguigna si distribuisca secondo una distribuzione Normale con media 150 e deviazione standard pari a 15, e osservando un campione di 20 soggetti con media campionaria pari a 160, verificare l'ipotesi che tali soggetti siano stati estratti da quella popolazione, contro l'ipotesi che siano soggetti ipertesi.

Ipotesi e tipo di Test: $H_0 : \mu = 150$ vs $H_1 : \mu > 150$; T-test per la media

Ipotesi del test: la distribuzione è Normale e la varianza è nota: soddisfatte.

Calcolo della statistica test:
$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{160 - 150}{3.35} = 2.98$$

Metodo delle regioni di rifiuto:

Fissiamo $\alpha=0.05$; trattandosi di un test a una coda, la zona è la coda destra, ed è delimitata dal valore 1.64 (proprietà della Normale: $\Phi(1.64) = 0.95$)

La statistica test vale 2.98: cade nella zona di rifiuto

Conclusione: ... (come prima)

Interpretazione del test statistico

Elementi per la valutazione della significatività

Gran parte della ricerca clinica procede sulla base di risultati valutati come “significativi”. E' fondamentale capire il senso della significatività.

Elementi:

- Significato e implicazioni degli errori di I e II tipo nel test statistico. Gli errori quando si utilizzano soglie di significatività fissate
- La significatività AUMENTA (cioè, il p-value diminuisce) all'aumentare dell'ampiezza campionaria
- Significatività non corrisponde a rilevanza clinica
 - Relazione fra Intervallo di Confidenza e test di ipotesi

Gli errori di I e II tipo

	Rifiuto H_0	Accetto H_0
H_0 vera	Errore I tipo	ok
H_0 falsa	ok	Errore II tipo

- L'errore di I tipo è “più grave” perché corrisponde a modificare erroneamente le nostre conoscenze di riferimento sulla popolazione, acquisendo “conoscenze false”.
- L'errore di II tipo è “meno grave” perché possiamo ricorrere a nuove prove statistiche per falsificare l'ipotesi H_0 .
- Tuttavia, è importante anche esso, poichè vanifica la ricerca

$p(\text{errore I tipo}) = \alpha$ **livello di significatività**

$p(\text{errore II tipo}) = \beta$

$p(\text{rifiuto } H_0 \mid H_0 \text{ falsa}) = 1 - \beta$ **potenza**

Gli errori di I e II tipo

	Rifiuto H_0	Accetto H_0
H_0 vera	Errore I tipo	ok
H_0 falsa	ok	Errore II tipo

$p(\text{errore I tipo}) = \alpha$
livello di significatività
 $p(\text{errore II tipo}) = \beta$
 $p(\text{rifiuto } H_0 \mid H_0 \text{ falsa}) = 1 - \beta$ **potenza**

- Quando scegliamo il metodo della soglia fissa per valutare il p-value, es. 0.05, stiamo fissando la probabilità di commettere Errore di I tipo
 - Circostanze diverse possono farci desiderare di essere più o meno cauti nel trarre conclusioni! Quali sono le conseguenze pratiche del nostro errore?
 - Possiamo permetterci di sbagliare nel 5% dei casi?
- Un piano di ricerca dovrebbe sempre avere una potenza sufficiente a dimostrare la fallacia di H_0 quando essa è falsa.
 - Nei trial clinici, tipicamente si richiede una potenza pari a 80% o 90%
- Si aumenta la potenza riducendo la variabilità casuale dell'esperimento e utilizzando campioni molto numerosi
 - Attenzione: con n grande, tutti i p-value diventano piccolissimi! Come si giudica allora una differenza?

Relazione fra intervallo di confidenza e test

La Popolazione è descritta da un parametro θ

- Facciamo una stima di θ mediante IC 95%
- Facciamo un test di ipotesi bilaterale al livello di significatività $\alpha=5\%$ per rigettare $H_0: \theta=\theta_0$

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

→ se l'IC non contiene θ_0 vuol dire che il test rifiuta H_0

e viceversa:

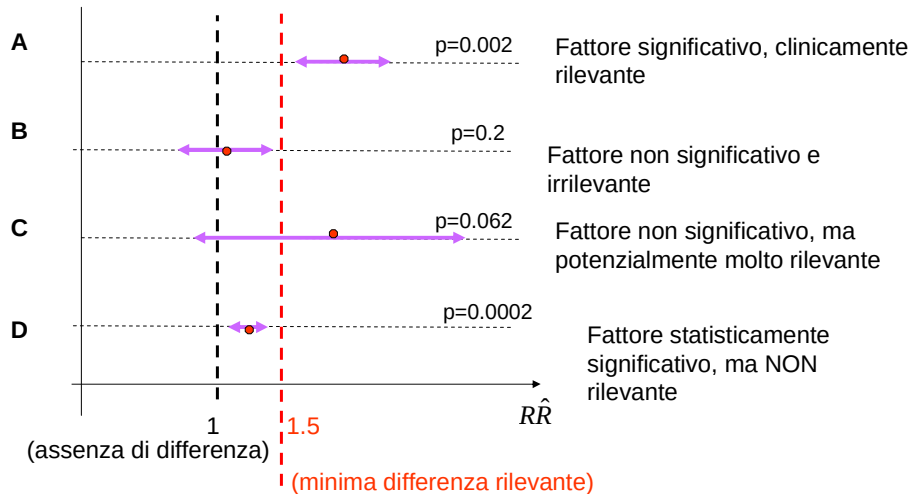
→ se il test accetta H_0 , vuol dire che θ_0 è dentro l'IC

→ Se il test rifiuta H_0 vuol dire che θ_0 non è incluso nell'IC



Rilevanza vs. Significatività (1)

Valutiamo l'impatto di 4 fattori di esposizione sul rischio di un evento (malattia) calcolando per ciascuno il Risk Ratio con un IC 95%. Viene considerato "importante" un fattore che aumenta almeno del 50% il rischio di evento



Rilevanza vs. Significatività (2)

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 al 95%, abbiamo informazione anche sulla significatività al livello 5%, e inoltre possiamo valutare più a fondo la rilevanza:

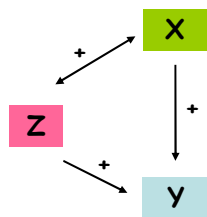
$$RR_A \rightarrow 95\%CI = (1.15, 1.25)$$

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

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.

Studio delle relazioni fra fenomeni



- Associazioni causali, associazioni statistiche, confondimento
- Strumenti per lo studio delle associazioni
 - 2 Caratteri qualitativi: Chi-Quadrato sulla tabella doppia
 - 2 Caratteri quantitativi: correlazione e regressione

Studio delle associazioni



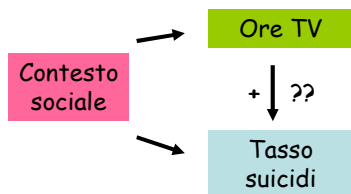
Fattore di esposizione (caratteristica clinica, trattamento) Variabile risposta

- La ricerca clinica ha fra i suoi obiettivi principali lo studio delle associazioni fra caratteri. Esempi di quesiti sull'associazione fra due fenomeni:
 - Qualitativo & qualitativo: Tipo di farmaco (A,B,C) e Risposta (Sì/No)
 - Qualitativo & quantitativo: Tipo di farmaco (A,B,C) e Diminuzione Colesterolo
 - Quantitativo & quantitativo: Dose di farmaco e Diminuzione Colesterolo
- In Medicina e Epidemiologia interessano soprattutto le **relazioni causali**, intese in senso non stretto
 - L'esposizione non causa con certezza una certa risposta, ma ne aumenta la probabilità
- Relazioni non causali possono interessare a fini di stima e previsione

Associazioni non causali fra X e Y

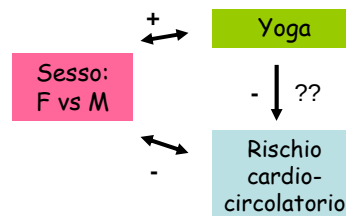
➤ Il terzo fattore Z è causale per X e Y; il legame fra questi ultimi è puramente statistico ("relazione spuria")

Studio sull'incidenza di suicidi fra i bambini giapponesi in diversi periodi



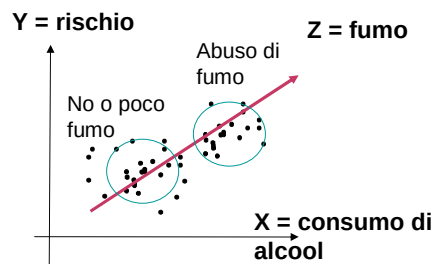
➤ Il terzo fattore Z è associato non causalmente con X e Y: anche in questo caso "appare" una relazione puramente statistica fra X e Y

Osservazione su soggetti giovani del rischio cardiocircolatorio



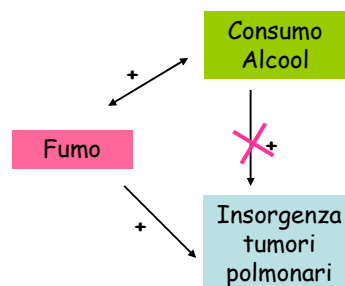
Confondimento (es. 1)

Si osserva una relazione crescente tra consumo di alcool e rischio di tumori polmonari.



Per motivi culturali-sociali, chi abusa di alcool di solito abusa anche di fumo
Il fumo è un fattore causale accertato dei tumori polmonari.

Questa relazione è solo apparente, e "scompare" se la si osserva all'interno del gruppo dei forti fumatori e dei non fumatori.

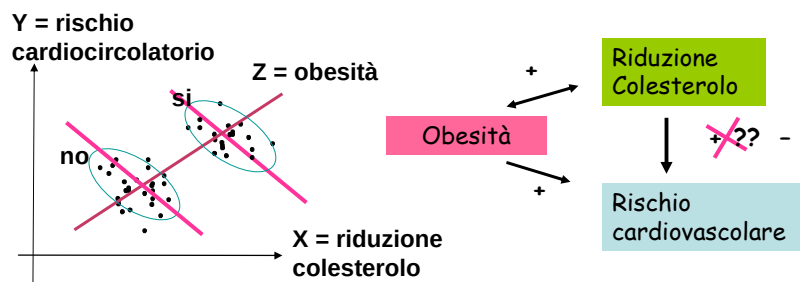


➤ Il terzo **fattore Z** si dice "confondente" se è **associato non causalmente con X** e **causalmente con Y**

Confondimento (es. 2)

Ricerca: se - mediante dieta e uso di farmaci specifici – si ottiene una riduzione del livello di colesterolo, questa provoca una riduzione del rischio cardiovascolare?

I soggetti obesi avevano tendenzialmente livelli iniziali di colesterolo più alti, e dunque hanno registrato riduzioni più forti; l'obesità è un fattore di rischio cardiovascolare



La relazione diretta fra X e Y era solo apparente, e SI INVERTE se la si osserva all'interno del gruppo di obesi e non obesi.

Misura e significatività delle associazioni

Fattore X



Fenomeno Y

Fattore di esposizione (caratteristica clinica, trattamento)

Variabile risposta

- Strumenti statistici molto utilizzati per lo studio delle associazioni fra due caratteri sono:
 - Qualitativo & qualitativo: indice e test del chi quadrato per tabelle doppie.
 - Qualitativo (k gruppi) & quantitativo: confronto fra medie mediante T-test (k=2) o ANOVA (k>2) o corrispondenti test nonparametrici: vd. dispensa di Inferenza
 - Quantitativo & quantitativo: coefficiente di correlazione e retta di regressione

Misura dell'associazione sulla tabella doppia

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

M	36%	64%	66
F	59%	41%	76
totale	49%	51%	

Se fra Fumo e Sesso ci fosse una **indipendenza perfetta**, osserverei sia fra i M che fra le F il 51% di fumatori

Tabella di indipendenza perfetta

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

Ad es: il numero atteso (E, "expected") di Maschi Fumatori sarebbe:

$$E : \frac{E}{\text{maschi}(66)} = \frac{\text{fumatori}(73)}{\text{tutti}(142)}$$

$$\Leftrightarrow E = \frac{\text{maschi} \cdot \text{fumatori}}{\text{tutti}} = \frac{66 \cdot 73}{142}$$

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

La frequenza attesa in caso di indipendenza rispetta la proporzione osservata sulla popolazione complessiva (sulla marginale)

Formula generale delle frequenze attese sotto l'ipotesi H0: i due caratteri sono indipendenti



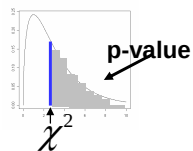
Misura e test sul grado di associazione: χ^2

Il grado di associazione fra 2 caratteri qualitativi può essere misurato sulla tabella doppia valutando la "distanza" dei dati osservati da quelli che si avrebbero nella situazione di indipendenza perfetta

→ L'indice Chi-Quadrato si basa sulle differenze fra le frequenze osservate e le frequenze "attese" che si avrebbero nel caso di indipendenza perfetta

$$\chi^2 = \sum \frac{(\text{Osservate} - \text{Attese})^2}{\text{Attese}}$$

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



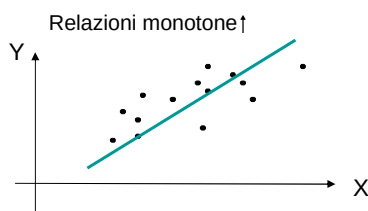
Si può poi valutare se la distanza da 0 del χ^2 è casuale o significativa: attraverso l'omonimo test:

$H_0: \chi^2=0$: indipendenza
 $H_1: \chi^2>0$: associazione

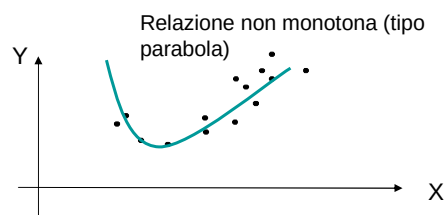
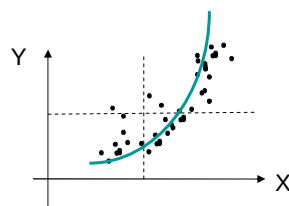


Associazione fra 2 caratteri continui

- Due caratteri continui X e Y mostrano un grado di associazione se, nella nuvola dei punti, è possibile riconoscere una tendenza delle osservazioni a distribuirsi secondo una relazione "regolare", che potrebbe essere rappresentata da una funzione: $Y=f(X)$

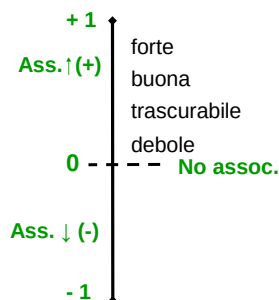
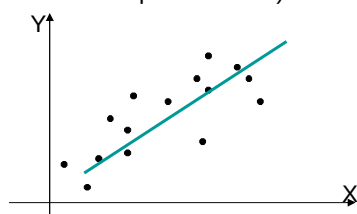


La **retta** è la più semplice relazione funzionale che può rappresentare il modello di associazione fra Y e X



Misura del grado di associazione lineare: r

Il grado di relazione lineare (cioè: con punti disposti attorno ad una ipotetica retta) viene misurato dal **coefficiente di correlazione lineare**, che varia fra -1 e +1, limiti che raggiunge in caso di perfetta associazione lineare (decreciente / crescente rispettivamente)



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

Dove al numeratore c'è la **Covarianza**, che misura la stessa cosa, ma non ha limiti superiore e inferiore

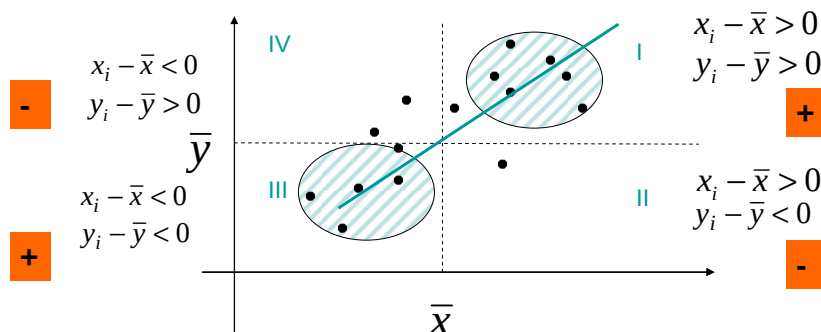
$$\text{COV}_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n}$$



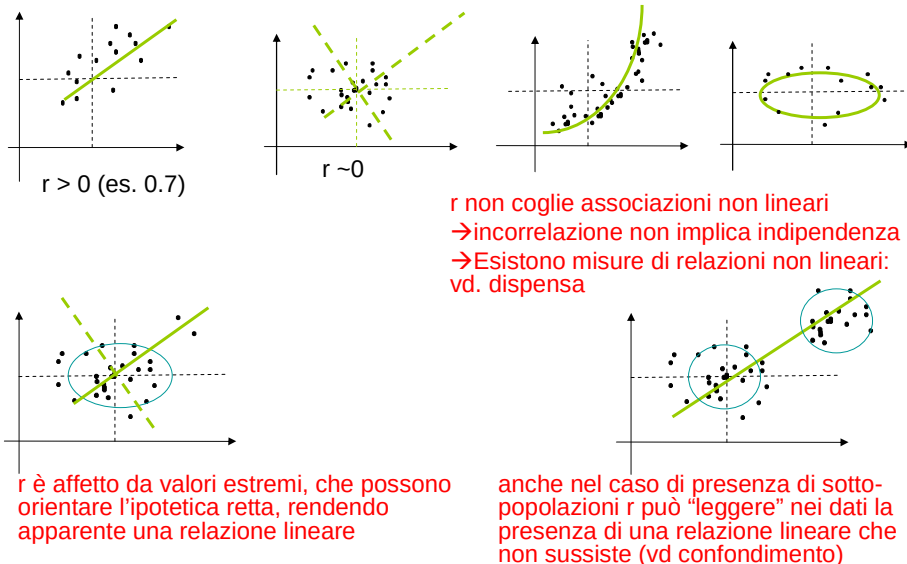
Razionale

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

- Suddividendo il piano cartesiano in 4 quadranti, determinati dagli assi passanti per il baricentro della nuvola dei punti $(\bar{x}, \bar{y})$, quando è presente una relazione lineare crescente (decrecente) tendono a prevalere i punti nel I e III quadrante (II e IV), che hanno scarti concordi (discordi), e quindi con prodotti positivi (negativi)

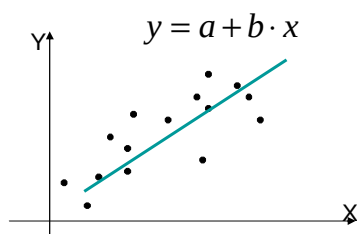


Interpretazione del coeff. di correlazione



Misura della rilevanza della relazione lineare

E' possibile determinare l'equazione della retta che passa nella nuvola di punti, fornendo una **quantificazione dell'effetto di X su Y**



b effetto di un aumento unitario di X su Y

Se $b=0$: retta parallela all'asse delle ascisse:
X non ha effetto su Y

→ La pendenza della retta è stimata attraverso il **coefficiente di regressione**, che viene calcolato in base al coeff. di correlazione lineare r

$$\hat{b} = r_{xy} \frac{std_y}{std_x}$$

→ L'intercetta della retta è stimata come:

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$



Modelli di regressione

- La retta di regressione $y = a + bx$ rappresenta il più semplice modello teorico per lo studio di una “variabile risposta” sulla base di una “variabile esplicativa” X.
- Essa può essere estesa a valutare l'influenza simultanea di più variabili esplicative (“analisi multivariata”)
- Un modello:
 - Fornisce una quantificazione della relazione, ovvero dell'effetto di X su Y
 - È utile nella pratica per stime e previsioni

Es: relazione (non causale!) fra circonferenza addominale (X) e indice di rischio cardiocircolatorio (Y)

 - Utile per individuare soggetti potenzialmente a rischio su cui procedere a ulteriori accertamenti
- Vi sono diversi modelli di regressione per studiare gli effetti su variabili risposta di vario tipo (probabilità; durate; ...) di $k > 1$ variabili esplicative.
Non sono però oggetto di questo corso!!

Appendice I

Prerequisiti di Matematica



- Nozioni elementari o poco più:
 - L'arrotondamento
 - Regole di calcolo
 - La sommatoria
 - La funzione logaritmo
 - La retta e sue generalizzazioni

Prerequisiti



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



Prerequisiti



Il confronto fra numeri

Un confronto fra due numeri a e b si può fare tramite differenza o rapporto

Es. confronto tra le età di due soggetti:

- A ha 10 anni meno di B ($54-64=-10$)

- B ha il doppio dell'età di C ($64/32=2$)

	DIFFERENZA	RAPPORTO
misura di confronto	$d = a - b$	$r = a / b$
valore di uguaglianza	$d = 0$	$r = 1$
ricostruiamo a da b attraverso	somma: $a = b + d$	prodotto: $a = b \cdot r$
progressione	aritmetica	geometrica

Spesso si considera il rapporto in termini percentuali: $r = a / b \times 100$

$r=120$ vuol dire che a è pari a b più il 20% di b

$r=80$ vuol dire che a è pari all'80% di b, ossia pari a b meno il 20% di b





Nozioni minime di matematica ...

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

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

$$k \cdot (a + b) = k \cdot a + k \cdot b \quad -(a + b) = -a - b$$

$$\frac{1}{k} \cdot (a + b) = \frac{1}{k} \cdot a + \frac{1}{k} \cdot b$$

$$a^2 = b \Leftrightarrow a = \sqrt{b}$$

$$\left(\frac{a}{b}\right)^2 = \frac{a^2}{b^2}$$

$$a = b + c \Leftrightarrow a - b = c \Leftrightarrow a - b - c = 0 \Leftrightarrow -a + b + c = 0$$

$$\frac{a}{b} = c \Leftrightarrow a = b \cdot c \Leftrightarrow b = \frac{a}{c}$$

$$\log_a b = x \Leftrightarrow a^x = b$$



... e poco di più

Numeri fattoriali

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

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

$$0! = 1$$

Coefficiente
binomiale
("N sopra k")

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



Prerequisiti



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"

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



Prerequisiti



Es: notazioni e simbologia per la tabella di frequenze

modalità	freq. assoluta (n_i)	freq. relativa (f_i)	freq. percent. (p_i)	freq. cumul. (N_i)
x_1	n_1	f_1	p_1	N_1
x_2	n_2	f_2	p_2	N_2
$x_3 (x_k)$	n_3	f_3	p_3	N_3
tot	n	1	100	

$$n = \sum_{i=1}^k n_i = n_1 + n_2 + \dots + n_i + \dots + n_k$$

Il pedice i indica la generica modalità, assume i valori da 1 a k , il numero totale di modalità del carattere.

$$f_i = \frac{n_i}{n} \quad \sum_{i=1}^k f_i = 1$$

$$p_i = \frac{n_i}{n} \cdot 100 \quad \sum_{i=1}^k p_i = 100$$

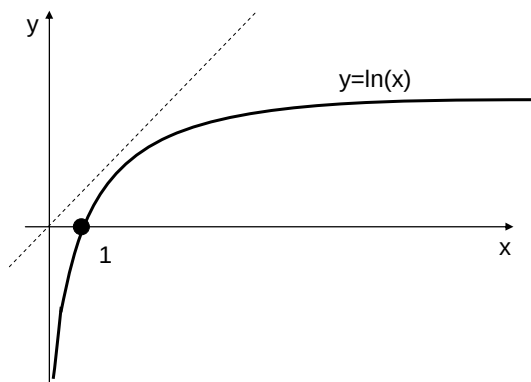
n_i è semplicemente la frequenza della i -esima modalità, cioè quella posta sulla i -esima riga della tabella

$$N_i = \sum_{j=1}^i n_j = n_1 + n_2 + \dots + n_i$$

La funzione logaritmo

$$\log_a b = x \Leftrightarrow a^x = b$$

$$\ln b = \log_e b = x \Leftrightarrow e^x = b \quad (\text{dove } e = 2.718\dots)$$



Logaritmo naturale
(ha in base il numero di Nepero, $e=2.718\dots$)

Trasformare X nel suo logaritmo, $Y=\ln X$, implica "schiacciarne" i valori molto alti

$$\ln(1)=0$$

$$\ln(x)<0 \text{ per } 0<x<1$$

$$\ln(x) \text{ non definito per } x\leq 0$$



La retta (1)

La retta è una curva descritta dall'equazione:

$$y=a+b\cdot x$$

$a \rightarrow$ intercetta

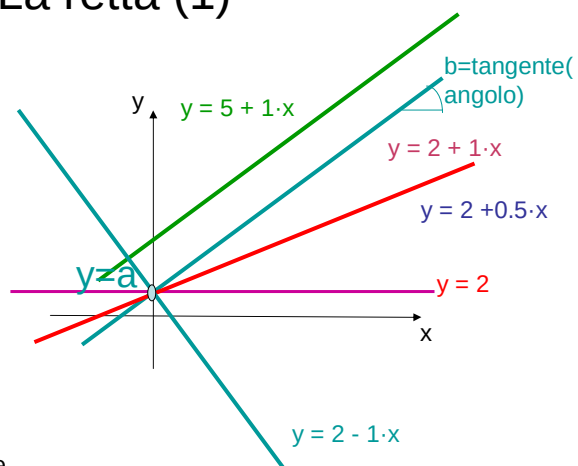
$b \rightarrow$ pendenza

$b > 0$: retta crescente

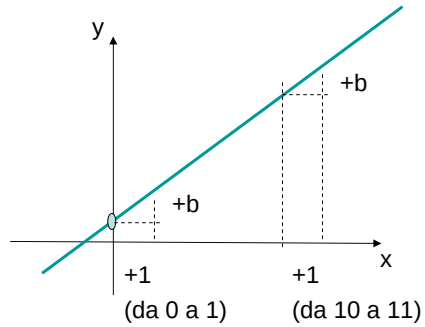
$b < 0$: retta decrescente

$b = 0$: retta parallela all'asse delle ascisse (asse x)

b misura la variazione di Y quando X aumenta di 1; la retta è l'unica curva in cui tale variazione è sempre la stessa, qualunque sia il valore di partenza di x



La retta (2)



x	y
0	a
1	a+b
10	a+10b
11	a+11b

Δx	Δy
1	b
1	b



Appendice II

Practicals



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

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 di un carattere quantitativo discreto da una tabella di frequenza

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

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 $\times$ 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	198×40=7920
F	190	16	190×16=3040
		56	10960

media = 10960 / 56 = 195.7



media ≠ (198+190)/2=194

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

Bisogna fare riferimento al concetto di MEDIA PONDERATA

Media e Mediana di un carattere quantitativo

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
	$\Sigma = 371$	

Ordinamento ↑

Rango	1	2	3	4	5	6
Peso	54	55	58	65	68	71

**Media =
somma / n**

$$= 371 / 6 = 61.8$$

Mediana → n pari

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

Mediana = somma delle loro modalità / 2

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

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



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 → 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

→ 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

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	$(x_i)^2$
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)



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:

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

Practicals

Test del χ^2

H_0 : indipendenza vs H_1 : associazione

Abbiamo calcolato l'indice Chi-Quadrato su una tabella doppia per 2 caratteri qualitativi, con r e c modalità. Per valutare la significatività del Chi-Quadrato [statistica test], facciamo riferimento alla Distribuzione Chi-Quadrato con g gradi di libertà, dove $g = (r-1) \cdot (c-1)$.

Secondo i due approcci al test, possiamo:

- Usare un computer e calcolare il p-value (l'area nella coda), valutando se è "abbastanza piccolo" da confutare l'ipotesi di indipendenza
- Usare le tabelle e vedere quale è il valore soglia corrispondente a α (lo fissiamo ad es $=0.05$) – e a g gradi di libertà

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

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

In una tabella 2x3 $g=2$, e il valore soglia per $\alpha=5\%$ è 5.99

1.225 cade nella zona di accettazione

Il χ^2 non è significativo: i dati non confutano l'ipotesi di indipendenza

(1.225 $\rightarrow$ P-value = 0.542)

Practicals

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

$$17.6 \cdot 43.3$$

$$180 - 162.4 = 17.6$$

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

$150 \cdot 1790$

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



Calcolo di r e della retta di regressione

Bisogna calcolare anche le due deviazioni standard. Risultano:

-X: 13.9

-Y: 137.2

$$r_{xy} = \frac{\text{COV}_{xy}}{\text{std}_x \cdot \text{std}_y} = \frac{480.4}{13.9 \cdot 137.2} = 0.283$$

id	Pressione (x)	kcal/die (y)
1	145	1850
2	180	1960
3	150	1790
4	155	1680
5	148	2090
6	160	2000
7	175	2100
8	180	1900
9	169	1880

$$\hat{b} = r_{xy} \frac{\text{std}_y}{\text{std}_x} = 2.79$$

O direttamente: $\hat{b} = \frac{\text{COV}_{xy}}{(\text{std}_x)^2} = \frac{\text{COV}_{xy}}{\text{var}_x}$

Infine:

$$\hat{a} = \bar{y} - \hat{b}\bar{x} = 1916.7 - 2.79 \cdot 162.4 = 1462.5$$

Es a pressione 150 corrisponde un valore teorico di calorie pari a $1462.5 + 2.79 \times 150 = 1881.9$

