

# Statistica Medica

## Mini-corso

Elementi di statistica descrittiva e prime  
nozioni e strumenti sui test di ipotesi  
per le associazioni

## La statistica: capire un fenomeno dai dati

Dati osservati relativi all'esito di un tipo di intervento chirurgico,  
effettuato con due tecniche alternative

|           | vivo | deceduto | tot |
|-----------|------|----------|-----|
| tecnica A | 37   | 13       | 50  |
| tecnica B | 80   | 32       | 112 |
|           | 117  | 45       | 162 |

Domande:

- C'è una differenza fra le due tecniche? → **descrivere, misurare, sintetizzare**
- Se c'è, vuol dire che una tecnica è migliore dell'altra? Oppure ...
  - E' un puro caso, ma in un'altra situazione simile non si osserverebbe la stessa differenza → **generalizzare le conclusioni tratte dai dati osservati**
  - La differenza osservata non dipende dalla tecnica usata ma da qualche altro fattore → **analizzare le relazioni**

## Statistiche basilari: le frequenze

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

(carattere qualitativo non ordinato)

| 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 |

SINTESI

freq. assoluta    freq. percentuale (%)

es. per la seconda modalità:

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

$$1227 : 1861 = 65.9 : 100$$

Numerosità  
 totale del  
 campione

totale  
 =100

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

E' il concetto di **proporzione**

## Statistiche basilari: le frequenze

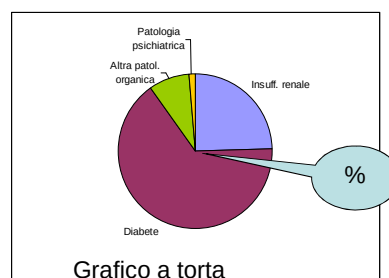
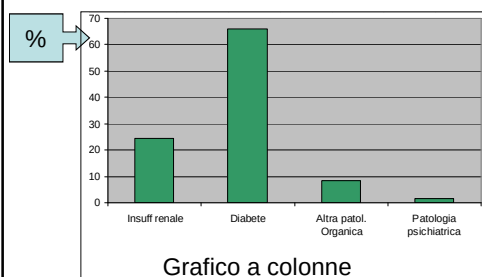
Distribuzione dei pazienti ricoverati sottoposti a regimi dietetici particolari rispetto al TIPO DI MALATTIA

(carattere qualitativo non ordinato)

| 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 |

SINTESI

**Moda:** modalità con la maggiore frequenza



GRAFICI

## Variabili continue: distribuzioni in classi

Distribuzione di 56 pazienti pediatrici per ETA\*

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

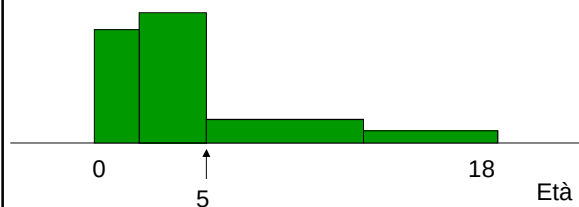
(carattere quantitativo  
continuo)

Età  
media:  
5 anni

Media: modalità  
"centrale"

SINTESI

GRAFICI



## Variabili (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

#### SCONNESSI

sex → M,F  
patologia → ulcera, tumore gastrico,  
tumore intestinale, ...

#### ORDINATI

titolo di studio → nessuno o licenza  
elementare, licenza media, licenza  
superiore, laurea  
stadio malattia → I,II,III, IV

### QUANTITATIVI

#### DISCRETI

numero di componenti (della famiglia) →  
1,2,3,4, ...  
gravidanze precedenti → 0, 1, 2, 3, ...

#### 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!!)

## 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

## Gerarchia dei caratteri

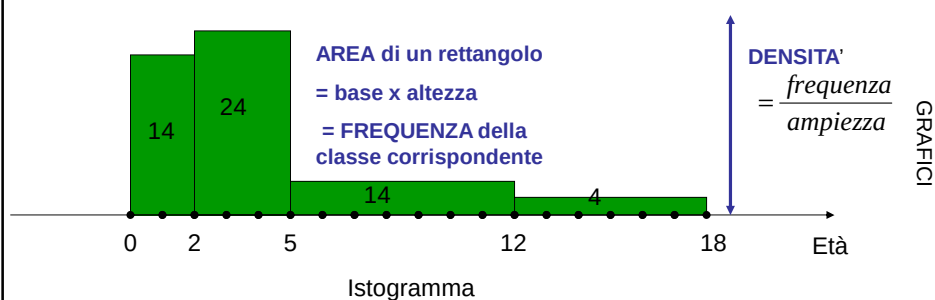
| Carattere             | Operazioni possibili sulle modalità e sintesi statistiche                                                                                                                                                                                                                           |
|-----------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Qualitativo sconnesso | Confronto: Stabilire uguaglianza o diversità ( $=$ o $\neq$ )<br>Manipolazione: accorpamento delle modalità<br>Sintesi: moda                                                                                                                                                        |
| Qualitativo ordinato  | Confronto: Stabilire relazioni di superiorità / inferiorità<br>Manipolazione: accorpamento (mantenendo l'ordinamento)<br>Sintesi: moda e modalità mediana                                                                                                                           |
| Quantitativo          | Confronto: Differenza o rapporto ( $-$ , $/$ )<br>Manipolazione: Suddivisione in classi; applicazione di operazioni matematiche ( $+$ , $-$ , $\cdot$ , $/$ , $\log$ , ...)<br>Sintesi: (classe modale), mediana, media aritmetica, deviazione standard, coefficiente di variazione |

## Il grafico della densità di frequenza

Distribuzione di 56 pazienti pediatrici per ETA'

(carattere quantitativo continuo)

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



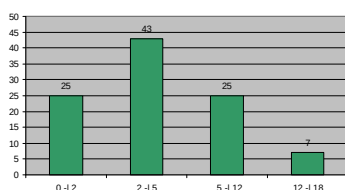
## (Perché usare la densità di frequenza)

Distribuzione di 56 pazienti pediatrici per età

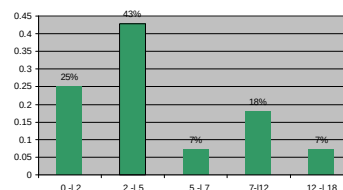
| Età      | freq. | %   |
|----------|-------|-----|
| 0 -  2   | 14    | 25  |
| 2 -  5   | 24    | 43  |
| 5 -  12  | 14    | 25  |
| 12 -  18 | 4     | 7   |
|          | 56    | 100 |

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 !!



## (Perché usare la densità di frequenza)

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

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

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

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

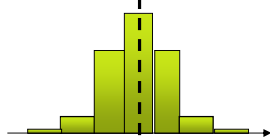
La frequenza va rapportata all'ampiezza della classe, ottenendo la densità di frequenza, un valore che **rappresenta quante unità sono presenti in ogni intervallino di ampiezza 1**

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

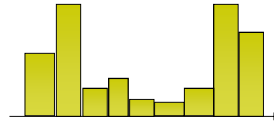
$$\Leftrightarrow \text{frequenza} = \text{ampiezza} \times \text{densità}$$

## Varie forme della distribuzione

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

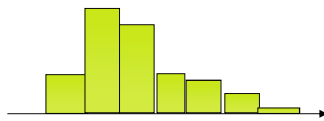


La forma "a campana" è tipica di fenomeni che possano essere ricondotti agli effetti "del caso", come l'altezza degli individui

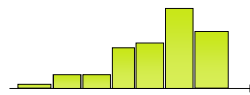


Distribuzione **BIMODALE**, cioè con la densità concentrata in due masse.

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

## Indici che descrivono la forma della distribuzione

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

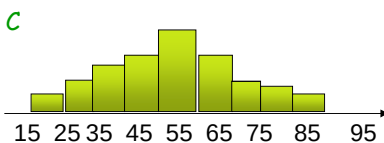
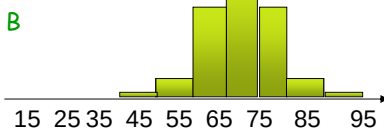
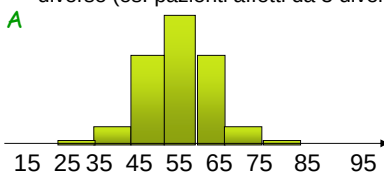


Tabelle e grafici di frequenza forniscono una rappresentazione completa dei dati.

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

I due aspetti essenziali sono:

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

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

## La media aritmetica

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

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

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

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

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

| Voto |    |
|------|----|
| 26   | 24 |
| 18   | 24 |
| 28   | 24 |
| 72   | 72 |

Media = 72 / 3 = 24

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

### In pratica

### Media di un carattere quantitativo discreto da una tabella di frequenze

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

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

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

= 0 · 4 + 1 · 3 + 2 · 1

Media = 5 / 8 = 0.6



Non confondere modalità (Parti) e frequenze!

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

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

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

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

## In pratica

## 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      |

$$\text{Media} = 4180 / 64 = 65.3$$



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

## In pratica

## 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<br>$\bar{X}$ | n.ro<br>casi<br>$n$ | totale      |
|---|--------------------|---------------------|-------------|
| M | 198                | 40                  | 198×40=7920 |
| F | 190                | 16                  | 190×16=3040 |
|   |                    | 56                  | 10960       |

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



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

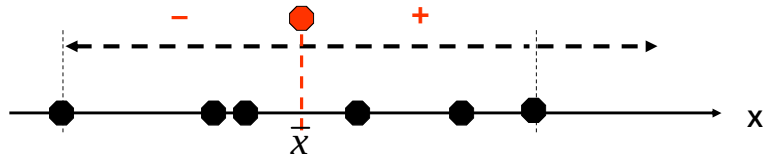
Ricostuiamo il totale di ciascun gruppo, e lo dividiamo per il totale delle unità

Questo ci conduce al concetto di MEDIA PONDERATA

$$\bar{X} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$$

$$\bar{X}_p = \frac{x_1 p_1 + x_2 p_2 + \dots + x_n p_n}{p_1 + p_2 + \dots + p_n}$$

## Limitazioni della media aritmetica

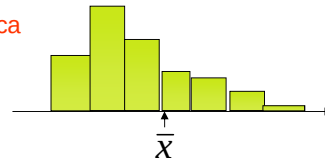


Dovendo BILANCIARE scarti positivi e negativi, e collocarsi nel centro (rispetto ai valori), la media è influenzata dai valori molto alti e dai valori molto bassi →

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:

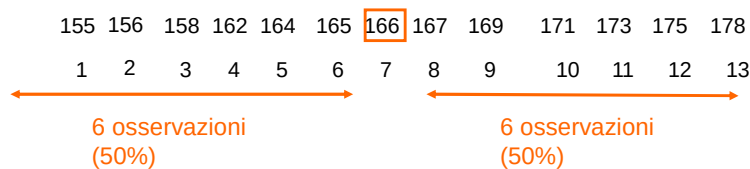
- Quando si hanno uno o più valori estremi molto anomali
- Quando la distribuzione è asimmetrica



## La mediana

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**):



mediana = 166

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

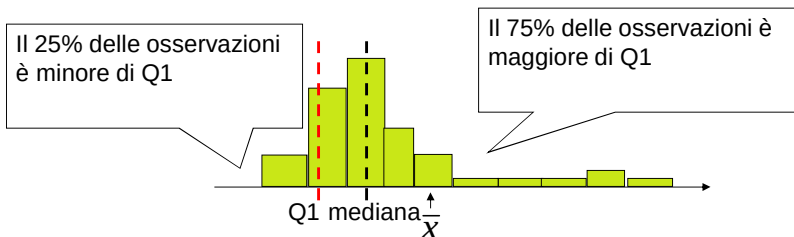
$n$  pari → mediana = modalità intermedia fra quelle di posto  $n/2$  e  $n/2+1$

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



## Generalizzazione della mediana: i 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.



## Interpretazione dei quantili

Es: Per l'età di 70 studenti di un corso di statistica, sappiamo che:

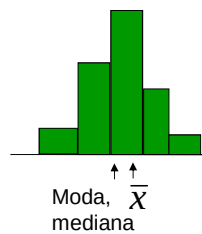
Quantili: Mediana (Q2) = 20.5      Q1 = 20.1      Q3 = 22

P10 (primo decile) = 18.5

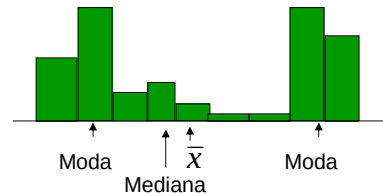
P66 (secondo terzile) = 21.7

- metà studenti avevano meno di 20.5 anni, e metà studenti più di 20.5 anni
- Uno su 4 (25%) aveva meno di 20.1 anni, 1 su 10 meno di 18.5 anni
- Uno su 4 (25%) aveva più di 22 anni. Ovvero  $\frac{3}{4}$  avevano al massimo 22 anni
- Uno su 3 (33%) aveva più di 21.7 anni. Ovvero  $\frac{2}{3}$  avevano al massimo 21.7 anni
- e per differenza, ad es.:
- uno su 4 avevano un'età compresa fra 20.1 e 20.5
- il 15% aveva un'età compresa fra 18.5 e 20.1
- etc

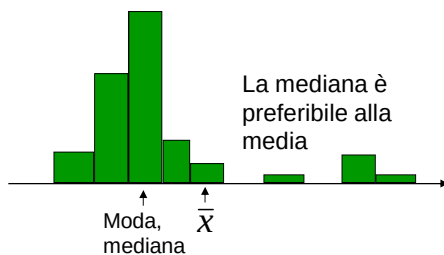
## Appropriatezza degli indici



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



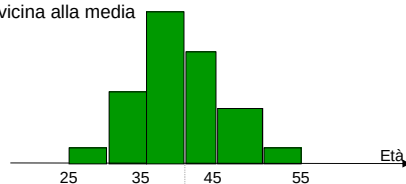
E' opportuno rimarcare la bimodalità: ne' media ne' mediana sono sintesi soddisfacenti



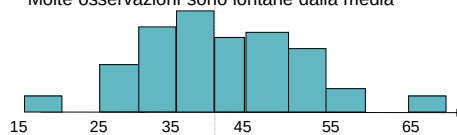
La mediana è preferibile alla media

## Variabilità: deviazione standard

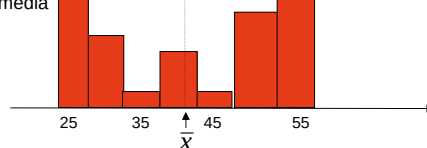
La maggior parte delle osservazioni è vicina alla media



Molte osservazioni sono lontane dalla media



La maggior parte delle osservazioni è lontana dalla media



- La principale misura di variabilità è una sintesi delle distanze delle osservazioni dalla media ( $x_i - \bar{x}$ )

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

- La deviazione standard rappresenta la **distanza media fra tutte le osservazioni e la media** ed è una sorta di "unità di misura rilevante" del fenomeno osservato
  - Es.  $X$  = peso paziente,  $std = 4.5\text{kg}$ : è la "distanza rilevante" fra due pazienti (1kg è irrilevante ai fini della descrizione del carattere)

## In pratica

## Calcolo della deviazione standard

Età per un campione di 7 pazienti

| ETA' $x_i$ | $x_i - m$ | $(x_i - m)^2$ |
|------------|-----------|---------------|
| 65         | 9.3       | 86.49         |
| 35         | -20.71    | 428.49        |
| 44         | -11.70    | 136.89        |
| 43         | -12.70    | 161.29        |
| 71         | 15.30     | 234.09        |
| 63         | 7.30      | 53.29         |
| 69         | 13.30     | 176.89        |
|            | 0         | 1277.43       |

media  $m = 55.7$ 

Ad esempio alla seconda riga:

$$(35 - 55.7) = -20.7 \quad ; \quad (-20.7)^2 = 428.49$$

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

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

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



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

## In pratica

## 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)

## 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 kg / 3.2 kg) · 100 = 15.6

Y : CV = (4.5 kg / 60 kg) = 7.5

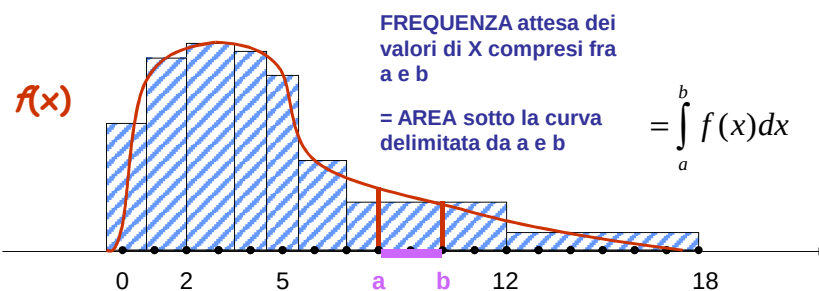
Z : CV = (2.5 cm / 51 cm) = 4.9

→ 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

## Curve teoriche di densità

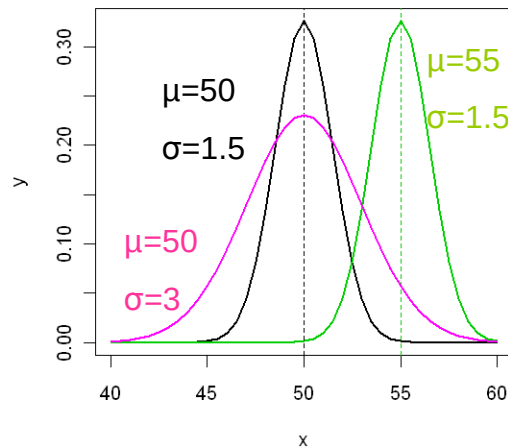
Se immaginiamo di fare un istogramma con intervallini piccolissimi, e di unire i punti medi del lato superiore delle colonne, otteniamo un grafico dato da una curva continua.

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 in statistica

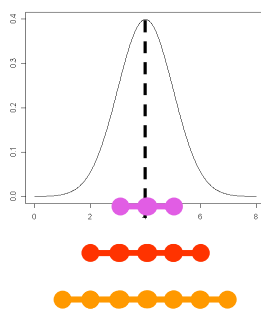
I 2 parametri  $\mu$  e  $\sigma$ , che determinano rispettivamente dove si posiziona l'asse di simmetria della curva rispetto all'asse x e quanto è ampia la campana, esprimono due indici statistici della distribuzione:



→  $\mu$  = media aritmetica (= mediana)

→  $\sigma$  = 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%

RANGE

intervallo

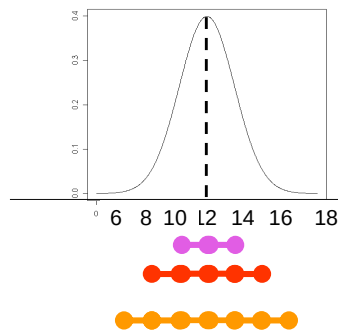
Area corrispondente

I due quartili  $Q_1$  e  $Q_3$  si trovano a distanza  $0.67\sigma$  dalla media:

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

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

## Esempio di utilizzo della Normale



Siamo interessati al numero di pazienti ( $X$ ) che necessitano di ricovero in un certo reparto durante il mese di dicembre. Supponiamo che in base a dati raccolti in passato possiamo assumere che questa variabile abbia distribuzione Normale, con media  $\mu=12$  e deviazione standard  $\sigma=2$ .

Pertanto ad esempio ci aspettiamo che:

- Non si presenteranno meno di 6 pazienti, ma non più di 18 ( $\mu - 3\sigma, \mu + 3\sigma$ )  $\approx 99.7\%$

- C'è solo un 2.5% di probabilità che avremo più di 16 pazienti ( $\mu - 2\sigma, \mu + 2\sigma$ )  $\approx 95\%$

- C'è una probabilità del 25% che avremo al massimo 11 pazienti

$$Q_1 = \mu - 0.67 \cdot \sigma \approx 10.66$$

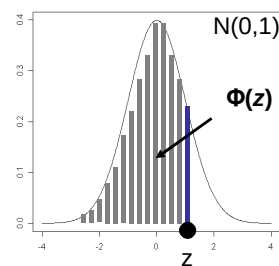
Ripasso!

## 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( $\mu, \sigma$ ), per avere l'area fino a  $x$ , basta calcolare  $\Phi$  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  $\mu$ , per cui:

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

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

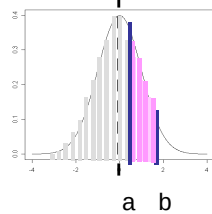
$$\Phi(0) = 0.5 \quad \Phi(+\infty) = 1$$

Ripasso!

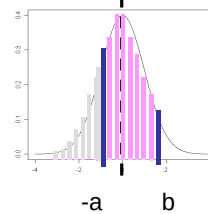
## Calcolare probabilità per la Normale

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

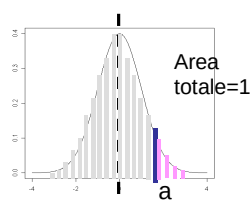
$$\text{Area}(a, b) = \Phi(b) - \Phi(a)$$



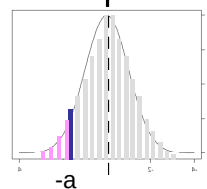
$$\text{Area}(-a, b) = \Phi(b) - (1 - \Phi(a))$$



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



$$= \Phi(-a)$$



Due valori di  $\Phi$  da ricordare:

$$\begin{aligned} \Pr(Z > 1.96) &= \Pr(Z < -1.96) = 2.5\% \\ \Pr(Z > 1.64) &= \Pr(Z < -1.64) = 5\% \end{aligned}$$

Ripasso!

## 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:**  $\mu=23$        $\sigma^2=7$

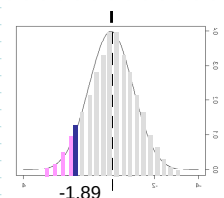
**Quesito:**  $P(X < 18)$

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



## 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 =  $7/36 \cdot 100 = 19.4$ )

## Tabelle doppie per descrivere relazioni

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

46%

54%

$69/142=49\%$

$73/142=51\%$

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

Es i Fumatori sono il 51% del totale

Le celle centrali presentano le frequenze delle combinazioni dei 2 caratteri

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



Distribuzioni marginali



## Distribuzioni condizionate

(profili riga e profili colonna)

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

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

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

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

## Confronto di probabilità: il Risk Ratio

I fumatori fra i M e fra le F sono rispettivamente il 64% e il 41%.  
Esprimiamo la differenza fra questi due gruppi (fra queste due percentuali, o probabilità) facendone il rapporto:

Risk Ratio (M vs. F) = 1.56

Ripasso!

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

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

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

= 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)

## Verifica di ipotesi

H0: **Ipotesi "nulla"**. Esprime uno stato di conoscenze attuale, di riferimento; l'obiettivo è superarlo, rigettarlo

I fumatori fra i M e fra le F sono rispettivamente il 64% e il 41% ( $RR=1.6$ ). Come interpretare questa differenza osservata in un campione?

- Possiamo **generalizzare** questo risultato osservato **alla popolazione** da cui il campione è stato estratto, e assumere che i maschi sono più spesso fumatori delle femmine? Abbiamo cioè una nuova "conoscenza" sulla popolazione?

Andiamo quindi a confrontare due ipotesi (H0 e H1):

- H0: no, nella popolazione non c'è differenza; questa differenza osservata nel campione è frutto del caso; non ci aspettiamo di osservarla in altri campioni
- H1: sì, nella popolazione c'è differenza, la differenza osservata nel campione è "riproducibile" ogni volta che osserviamo un campione estratto da quella popolazione

H1: **Ipotesi "alternativa"**. Esprime uno stato di conoscenze nuovo, innovativo, un obiettivo da dimostrare

## Significatività (concetto e terminologia)

I fumatori fra i M e fra le F sono rispettivamente il 64% e il 41% ( $RR=1.6$ ). Confronto di ipotesi: useremo i dati per concludere alternativamente che:

- vale H0: nella popolazione non c'è differenza, e la differenza osservata nel campione è frutto del caso; non ci aspettiamo di osservarla in altri campioni
- vale H1: nella popolazione c'è differenza, la differenza osservata è "riproducibile" estraendo altri campioni. **Diremo che la differenza osservata nel campione è "significativa"**

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

Significativo = incompatibile con H0, nel senso di improbabile sotto H0, tanto da indurci a rigettare H0.

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

## Significatività: come valutarla (i)

Usando metodi matematici (calcolo delle probabilità) possiamo calcolare il cosiddetto **p-value**:

La probabilità che **SE è vera l'ipotesi di base H1, otteniamo i dati** (la differenza) **effettivamente osservati, o dati ancora più a supporto di H1**.

Se questa probabilità è molto bassa (es.  $< 5\%$ ), concludiamo che i dati sono significativi in favore dell'ipotesi alternativa H1, e rigettiamo l'ipotesi di base H0.

Se è alta, manteniamo lo stato di conoscenze attuali, cioè H0. Non abbiamo evidenza che H0 sia falsa.

Il valore soglia per il p-value che usiamo per discriminare fra accettazione e rifiuto di H0 è detto **livello di significatività** e indicato con  $\alpha$

$\alpha$  è una misura di quanto ci sentiamo di rischiare di commettere un grave errore (detto "**errore di I tipo**"): rigettare H0 quando H0 è vera

Un altro errore (detto "**errore di II tipo**") è accettare H0 quando H0 è falsa. Indichiamo con  $\beta$  la probabilità di commetterlo.

La **potenza** del test  $1-\beta$  è la probabilità di rifiutare H0 quando H0 è falsa.

## Significatività: come valutarla (ii)

Alternativamente al calcolo del p-value, sempre usando gli stessi metodi matematici (calcolo delle probabilità), possiamo calcolare i limiti della **regione di rifiuto**:

Scegliendo il livello di significatività  $\alpha$ , la regione di rifiuto è un insieme di valori fra quelli possibili per il dato osservato (la differenza), che sembrano indicare che sia vera H1, a cui corrisponde una probabilità pari a  $\alpha$  sotto l'ipotesi nulla.

**Se il dato osservato appartiene alla regione di rifiuto, rigettiamo l'ipotesi di base H0 a favore dell'ipotesi alternativa H1, avendo dati significativi al livello  $\alpha$ .**

Se il dato osservato non appartiene alla regione di rifiuto, manteniamo lo stato di conoscenze attuali, cioè H0. Non abbiamo evidenza che H0 sia falsa.

## Associazione fra due variabili su una tabella doppia

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

|   |     |     |  |
|---|-----|-----|--|
| M | 36% | 64% |  |
| F | 59% | 41% |  |

Siccome i fumatori fra i M e fra le F sono rispettivamente il 64% e il 41% (RR=1.6 ≠ 1) sembra esserci una associazione fra la variabile Fumo e la variabile Sesso. Andiamo quindi a **misurare** questa **associazione**, e poi a verificare se vi è sufficiente evidenza per generalizzare alla popolazione

## La tabella SE non c'è associazione (indipendenza perfetta)

| X = Sesso | Y = Fumo |       | totale |
|-----------|----------|-------|--------|
|           | no       | si    |        |
| M         | 32.07    | 33.93 | 66     |
| F         | 36.93    | 39.07 | 76     |
| totale    | 69       | 73    | 142    |

Le frequenze attese nel caso di perfetta indipendenza sono tali che c'è la stessa proporzione di fumatori in ciascuna riga condizionata e nella riga marginale

Per esempio il numero atteso E dei Maschi Fumatori dovrebbe essere tale che:

$$\frac{E}{\text{maschi}(66)} = \frac{\text{fumatori}(73)}{\text{tutti}(142)} = 51\% \Leftrightarrow E = \frac{\text{maschi} \cdot \text{fumatori}}{\text{tutti}} = \frac{66 \cdot 73}{142} = 33.93$$

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

Formula generale delle frequenze attese in ciascuna cella sotto l'ipotesi nulla (H0) che non c'è differenza nei gruppi, ossia che le due variabili X e Y sono indipendenti

## In pratica

Calcolo delle frequenze attese in una tabella doppia sotto l'ipotesi H0 di indipendenza (per  $\chi^2$ )

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

Frequenze  
(assolute)  
osservatePasso 1: calcolo  
delle frequenze  
attese (una per  
ciascuna cella  
interna della tabella)

|       | dolore percepito |       |         |     |
|-------|------------------|-------|---------|-----|
| sessu | 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  |

Frequenze  
(assolute)  
attese

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

$$= 12 \cdot 9 / 22$$

Misura e test sul grado di associazione:  $\chi^2$ 

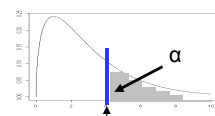
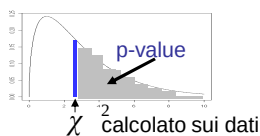
Il grado di associazione fra 2 caratteri qualitativi può essere misurato sulla tabella doppia valutando la "distanza" dei dati osservati da quelli che si avrebbero nella situazione di indipendenza perfetta

→ 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,  $\chi^2$  vale 0; cresce al crescere del grado di associazione.

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



valore soglia      Regione di rifiuto

Si può quindi valutare se l'associazione è significativa (rifiuto  $H_0$ ) o no (accetto  $H_0$ ) calcolando alternativamente:

- Il **p-value**: l'area sotto la curva nella coda delimitata dal valore  $\chi^2$  calcolato sui dati
- Il valore soglia della coda che ha un'area pari al livello di significatività scelto. I valori nella coda sono la **regione di rifiuto** di  $H_0$ .

## In pratica

Calcolo dell'indice di associazione  $\chi^2$ 

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

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

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

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

|      |      |      |
|------|------|------|
| 0.44 | 0.20 | 0.03 |
| 0.37 | 0.17 | 0.02 |

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

Passo 3: sommare tutte le distanze

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

## In pratica

Test  $\chi^2$ 

$H_0$ : indipendenza vs  $H_1$ : associazione

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

$r=2$

$c=3$

Con la tabella fornita nel formulario, possiamo calcolare il valore soglia della regione di rifiuto, in corrispondenza del livello di significatività  $\alpha$  scelto e di un numero  $g$  di gradi di libertà, determinato come

$$g = (r-1) \cdot (c-1)$$

$$\chi^2 = 1.225$$

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

Cioè, solo valori del Chi-Quadrato  $> 5.99$  sono considerati evidenza sufficiente per rifiutare  $H_0$

1.225 appartiene alla regione di accettazione: non è un valore statisticamente significativo

$$(1.225 \rightarrow P\text{-value} = 0.542)$$

## Associazione variabile continua - discreta

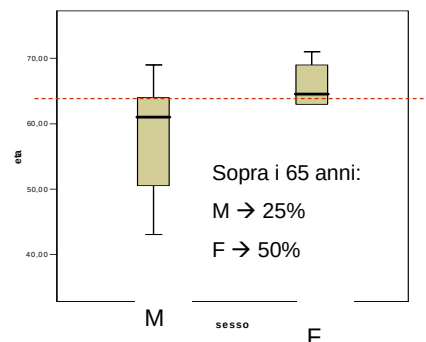
ovvero: valutare una variabile continua in 2 gruppi

| X = Sesso | n  | Y = Età |     |                          |
|-----------|----|---------|-----|--------------------------|
|           |    | media   | std | quantili                 |
| M         | 40 | 55.7    | 42  | Mediana=60.5 Q1=51 Q3=65 |
| F         | 55 | 66.8    | 31  | Mediana=65.5 Q1=63 Q3=69 |

Questi indici sono calcolati per Y=Età condizionatamente a X=Sesso.

E' utile la rappresentazione grafica mediante "boxplot" (basata sui quantili)

Possiamo poi valutare la significatività della differenza fra le medie dei M e delle F: **T-test**



## T-test per confrontare 2 medie

$H_0: \mu_1 = \mu_2$  cioè no associazione VS  $H_1: \mu_1 \neq \mu_2$  presenza di associazione (differenza)

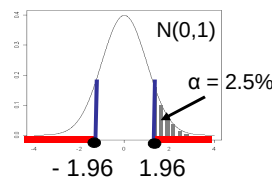
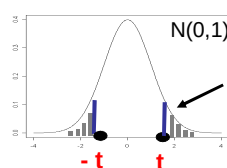
Nota: queste formule si possono applicare se i due gruppi provengono da due popolazioni rispettivamente con media  $\mu_1$  e  $\mu_2$  (incognite) ma uguale varianza  $\sigma^2$  (incognita), ed entrambi i campioni sono grandi ( $n_1, n_2 > 30$ )

Sostanzialmente, si tratta di standardizzare la differenza fra le medie dei due gruppi, e usare la Normale per calcolare il p-value ovvero per determinare la regione di rifiuto (qui, per  $\alpha=5\%$ ).

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

dove:

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



**Zona di rifiuto:**  
per  $\alpha=5\%$ :  
 $t < -1.96$  o  
 $t > 1.96$

## Es: T-test per il confronto fra medie

Due gruppi di pazienti ricevono rispettivamente il trattamento A o B. Si misura il valore della pressione arteriosa, assumendo che la varianza sia uguale nei due gruppi. Dati in tabella. Verificare l'ipotesi che non vi sia differenza vs. l'ipotesi che la pressione dipenda dal trattamento.

|           | Trattamento |    |
|-----------|-------------|----|
|           | A           | B  |
| n         | 32          | 36 |
| $\bar{y}$ | 94          | 92 |
| $s^2$     | 18          | 16 |

**Calcolo della statistica test:**

$$s = \sqrt{\frac{31 \cdot 18 + 35 \cdot 16}{32 + 36 - 2}} = 4.16 \quad t = \frac{94 - 92}{4.16 \sqrt{\frac{1}{32} + \frac{1}{36}}} = 1.98$$

**Metodo della regione di rifiuto:** Con  $\alpha=0.05$  si rigetta l'ipotesi se  $t < -1.96$  o  $t > 1.96$ . La statistica test è pari a 1.98, dunque rigettiamo  $H_0$ .

Concludiamo che il Trattamento (A o B) e la Pressione presentano una associazione statistica significativa al livello del 5%.

**Metodo del p-value:** Sulla tavola di  $N(0,1)$  a  $t=1.98$  corrisponde area=0.976 →  
Area coda =  $1 - 0.976 = 0.024$  →  
 $P=0.048$

## Lo studio di una relazione – un esempio

Dati osservati relativi all'esito di un tipo di intervento chirurgico, effettuato con due tecniche alternative

|           | vivo | deceduto | tot |
|-----------|------|----------|-----|
| tecnica A | 37   | 13       | 50  |
| tecnica B | 80   | 32       | 112 |
|           | 117  | 45       | 162 |

- C'è una differenza fra le due tecniche? → **descrivere, misurare, sintetizzare**

- Se c'è, vuol dire che una tecnica è migliore dell'altra? Oppure ...

- E' un puro caso, ma in un'altra situazione simile non si osserverebbe la stessa differenza → **generalizzare le conclusioni tratte dai dati osservati**

- La differenza osservata non dipende dalla tecnica usata ma da qualche altro fattore → **analizzare le relazioni**

→ Percentuali di riga e Risk Ratio

→ Test di significatività

→ Studio del confondimento

## Lo studio di una relazione – un esempio

Dati osservati relativi all'esito di un tipo di intervento chirurgico, effettuato con due tecniche alternative

|           | vivo | decaduto | tot |
|-----------|------|----------|-----|
| tecnica A | 37   | 13       | 50  |
| tecnica B | 80   | 32       | 112 |
|           | 117  | 45       | 162 |

"Tasso" di Mortalità (prob) con tecnica A:  $13/50 = 0.26 = 26\%$

"Tasso" di Mortalità (prob) con tecnica B:  $32/112 = 0.29 = 29\%$

$P=0.736$

(calcolare il Chi-Quadrato e verificare che cade nella regione di accettazione)

→ Percentuali di riga e Risk Ratio

→ Test di significatività

→ Studio del confondimento

## Il confondimento: concetto ed esempio

|           | vivo | decaduto | tot |
|-----------|------|----------|-----|
| tecnica A | 37   | 13       | 50  |
| tecnica B | 80   | 32       | 112 |
|           | 117  | 45       | 162 |

A: 26%

B: 29%

Questo campione è costituito da 52 giovani e 110 anziani

### Pazienti "giovani"

|           | vivo | decaduto | tot |
|-----------|------|----------|-----|
| tecnica A | 32   | 8        | 40  |
| tecnica B | 10   | 2        | 12  |
|           | 42   | 10       | 52  |

A:  $8/40=20\%$

B:  $2/12=17\%$

complessivamente, fra i giovani:  
pr(decesso)=  
 $10/52=19\%$

### Pazienti "anziani"

|           | vivo | decaduto | tot |
|-----------|------|----------|-----|
| tecnica A | 5    | 5        | 10  |
| tecnica B | 70   | 30       | 100 |
|           | 75   | 35       | 110 |

A:  $5/10=50\%$

B:  $30/100=30\%$

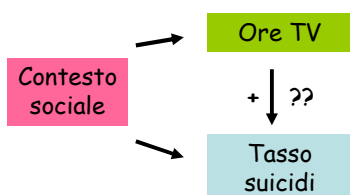
fra gli anziani:  
 $35/110=32\%$

## Il confondimento nell'esempio

- La mortalità sembra dipendere dalla Tecnica ( $B > A$ )
- In realtà, la mortalità dipende dall'Età del paziente: gli anziani hanno maggiore mortalità (circa il 32%) dei giovani ( $\approx 19\%$ )
- Il punto è che con la tecnica B sono stati trattati molti più anziani che con la tecnica A. Ecco perché il numero relativo di deceduti era maggiore con B che con A
- Ma all'interno del gruppo con la stessa età, B ha mortalità minore di A!
- Si dice che la relazione fra Tecnica e Esito è confusa dall'Età
- Se ne deriva quindi che **quando studiamo una relazione fra due variabili X e Y, dobbiamo "controllare" l'effetto di altre variabili**
- (Esistono diversi metodi statistici per controllare il confondimento)
- Usando la statistica, possiamo misurare un'associazione, verificarne la significatività, verificare la presenza di fattori confondenti. E' sempre tuttavia necessario discutere la natura (causale o no) della relazione

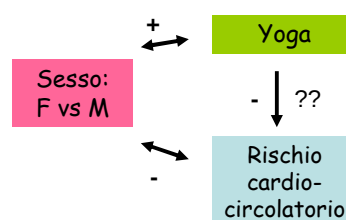
## Associazioni non causali fra X e Y

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



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

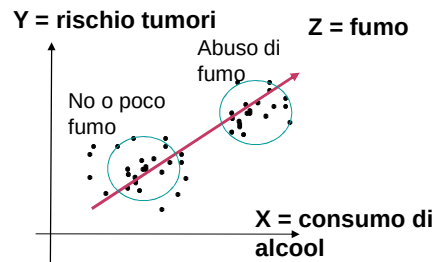
Osservazione su soggetti giovani del rischio cardiocircolatorio



➤ 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

## Un (altro) esempio di confondimento (i)

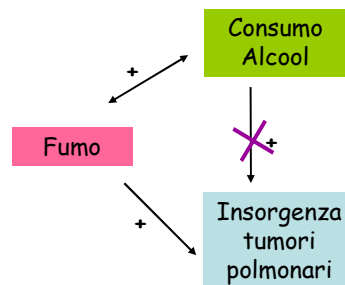
Si osserva una relazione crescente tra consumo di alcool e rischio di tumori polmonari.



Per motivi culturali-sociali, chi abusa di alcool di solito abusa anche di fumo

Il fumo è un fattore causale accertato dei tumori polmonari.

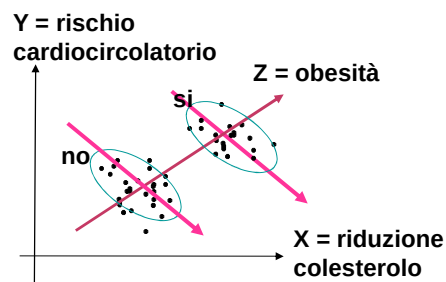
Questa relazione è solo apparente, e "scompare" se la si osserva all'interno del gruppo dei forti fumatori e dei non fumatori.



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

## Un (altro) esempio di confondimento (ii)

Ricerca: somministrando una dieta e farmaci specifici si ottiene una riduzione del livello di colesterolo; questa provoca una riduzione del rischio cardiovascolare? Risultati: opposti??!



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 è solo apparente, essa **SI INVERTE** (come atteso) se la si osserva all'interno del gruppo di obesi e non obesi

