

Statistica medica – *parte 2*

Statistica medica - 377ME

cds TeRP

A.A. 2022/23

I° anno – I° semestre

2+3 crediti / 50 ore

Prof. Lucio Torelli

Dipartimento Clinico di Scienze mediche, chirurgiche e della salute

Università degli Studi di Trieste

torelli@units.it

Attenzione: queste slide sono solo alcune note per le lezioni, non sono pertanto
un riassunto del seminario

12/12/2022

Calcolo delle probabilità

- Il problema dell'incertezza
- E' possibile/corretto misurare l'incertezza?
Come siamo soliti misurare l'incertezza?
- Il calcolo delle probabilità, strumento necessario per un uso corretto della statistica
- Saper leggere un risultato, una statistica, un output di un software, in cui viene valutata l'incertezza.

Uno strumento utile: **le tavole di contingenza**

sappiamo “leggere” questa tavola di 2 x 2?

	Dalt.	Non Dalt.	totali
M	60	540	600
F	5	495	500
totali	65	1035	1100

sappiamo “leggere” la probabilità da questi dati?

	Dalt.	Non Dalt.	totali
M	60	540	600
F	5	495	500
totali	65	1035	1100

- $P(\text{Dalt.}) ?$
 - $P(\text{Dalt.} \mid F) ?$
 - $P(\text{Dalt.} \mid F) = P(\text{Dalt.}) ?$
- “Sapendo che”*
- Eventi dipendenti !!

Siamo in grado di **completare**
questa tavola di contingenza 2 x 2?

	Malati	Sani	totali
M			60%
F			40%
totali	10%	90%	100%

Siamo in grado di completare
questa tavola 2 x 2?

	Malati	Sani	totali
M	10%	50%	60%
F	0%	40%	40%
totali	10%	90%	100%

????????????????????

Siamo in grado di completare
questa tavola 2 x 2?

	Malati	Sani	totali
M	6%	54%	60%
F	4%	36%	40%
totali	10%	90%	100%

Se la malattia non dipende dal genere !!!

*Ad es. $P(F \cap \text{Sani}) = P(F) * P(\text{Sani}) = 40\% * 90\% = 36\% \dots$*

Diverse definizioni,
da applicare in maniera “corretta”...



Definizione **classica**: casi favorevoli / casi possibili
se “eventi elementari” sono “equiprobabili”

Definizione **frequentista**: $\cong$ freq. relativa osservata
se abbiamo un numero “suff. grande” di osservazioni

Definizione **soggettivista**: è il prezzo p ($0 \leq p \leq 1$) che sono
disposto a pagare in una scommessa equa ...
il problema della soggettività ...

e quindi, come fare, in particolare in ambito bio-medico?

Probabilità condizionata e test diagnostici

	T+	T-	
M+			
M-			

Valore predittivo pos.: è la probabilità che una persona sia malata, sapendo che è positiva al test

Valore predittivo neg.: è la probabilità che una persona sia sana, sapendo che è negativa al test

$$Vp+ = P(M+ | T+) \cong \frac{T+ \cap M+}{T+} = \frac{\text{veri positivi}}{\text{positivi}}$$

$$Vp- = P(M- | T-) \cong \frac{T- \cap M-}{T-} = \frac{\text{veri negativi}}{\text{negativi}}$$

Probabilità condizionata e test diagnostici

	T+	T-	
M+			
M-			

Sensibilità: è la probabilità che una persona risulti positiva al test, sapendo che è malata

Specificità: è la probabilità che una persona risulti negativa al test, sapendo che è sana

$$Se = P(T+ | M+) \cong \frac{T+ \cap M+}{M+} = \frac{\text{veri positivi}}{\text{malati}}$$

$$Sp = P(T- | M-) \cong \frac{T- \cap M-}{M-} = \frac{\text{veri negativi}}{\text{sani}}$$

Domanda: In un test con sensibilità del 100% non ci possono essere falsi positivi o falsi negativi?

3 giugno 2020



scienzainrete
il gruppo 2003 per la ricerca scientifica



Perché un tampone non basta

L'accuratezza diagnostica del test può aumentare con una raccolta ripetuta dei tamponi. In uno studio condotto su 51 pazienti affetti da COVID-19, la sensibilità del primo tampone respiratorio era del 70%; tale valore raggiungeva il 94% e il 98% con l'esecuzione del secondo e del terzo tampone.

... Sul mercato sono oggi disponibili diversi test sierologici, dai più rapidi e meno specifici (immunocromatografici) per arrivare all'analisi d'elezione effettuata su esame da prelievo endovenoso con tecnologie differenti (ELISA, chemiluminescenza). L'affidabilità di un test sierologico dipende dalle sue caratteristiche di specificità e sensibilità. È fortemente raccomandato l'utilizzo di test del tipo CLIA e/o ELISA e/o ELFA che presentino una specificità non inferiore a 95% e una sensibilità non inferiore al 90%, come indicato in molte linee guida, per consentire di ridurre il numero di risultati falsi positivi e falsi negativi.

<https://www.aboutpharma.com/blog/2020/07/09/test-tamponi-mutazioni-il-punto-su-sars-cov2-con-un-11-microbiologo/> 9 luglio 2020

22 ottobre 2020

Università degli studi di Trieste x Il Piccolo x Coronavirus: a cosa servono i test x +

corriere.it/salute/malattie_infettive/20_ottobre_22/coronavirus-cosa-servono-test-sierologici-cosa-possono-scoprire-af16cb7c-f8e9-11ea-b4b0-f49c5435d3f2.shtml

App Gmail YouTube Maps Traduci Notizie Posta - DELEGATO...

SEZIONI EDIZIONI LOCALI CORRIERE TV ARCHIVIO TROVOCASA TROVOLAVORO SERVIZI CERCA ABBONATI LOGIN

Tempi di risposta (esito)

La risposta arriva in **circa 15 minuti** in laboratorio. In poche ore all'utente (ma dipende dai laboratori e da quanti anticorpi vengono cercati).

Affidabilità

La **sensibilità** arriva **fino al 96%** quando siano eseguiti dai 21 ai 35 giorni dall'esordio dei sintomi (**più giorni passano** dall'incontro con il virus, più i sierologici sono affidabili). Ma secondo una recente ricerca ci sono notevoli variazioni nell'accuratezza dei test sierologici disponibili in commercio, la cui sensibilità va dal 60,9% all'87,3%.

Risultato

A seconda dei kit, il modello più completo ed esaustivo ci dice se abbiamo

redazione di Salute.

ISCRIVITI



ESERCIZIO – dati Se, Sp, Pr, calcolare i Vp

	T+	T-	
M+			
M-			

Se:	60%
Sp:	90%
Pr:	1%

$$Se = P(T+ | M+) = \frac{\text{veri_positivi}}{\text{malati}}$$

$$Sp = P(T- | M-) = \frac{\text{veri_negativi}}{\text{sani}}$$

Risulta quindi:

$$\text{veri_positivi} = Se * \text{malati} = 60\% * 1\% = 0,6\%$$

$$\text{veri_negativi} = Sp * \text{sani} = 90\% * 99\% = 89,1\%$$

	T+	T-	
M+	0,6%	0,4%	1,0%
M-	9,9%	89,1%	99,0%
	10,5%	89,5%	100%

sens:	60%
spec:	90%
Prev:	1%

Risulta pertanto:

$$Vp+ = P(M+ | T+) = \frac{\text{veri_positivi}}{\text{test_positivi}} = 5,7\%$$

$$Vp- = P(M- | T-) = \frac{\text{veri_negativi}}{\text{test_negativi}} = 99,6\%$$

**Cosa significano
tali valori?**

Cosa succede se passiamo a Se=95%?

	T+	T-	
M+	0,95%	0,05%	1,0%
M-	9,90%	89,10%	99,0%
	10,85%	89,15%	100%

Se:	95%
Sp:	90%
Pr:	1%

Risulta pertanto:

$$Vp+ = P(M+ | T+) = \frac{\text{veri_positivi}}{\text{test_positivi}} = 8,8\%$$

$$Vp- = P(M- | T-) = \frac{\text{veri_negativi}}{\text{test_negativi}} = 99,9\%$$

e se fosse Pr=20%?

	T+	T-	
M+	19%	1%	20%
M-	8%	72%	80%
	27%	73%	100%

Se:	95%
Sp:	90%
Pr:	20%

Vp+=	70,4%
Vp-:	98,6%

Probabilità e misure del rischio

Prevention of Breast Cancer in Postmenopausal Women: Approaches to Estimating and Reducing Risk

Steven R. Cummings, Jeffrey A. Tice, Scott Bauer, Warren S. Browner, Jack Cuzick, Elad Ziv, Victor Vogel, John Shepherd, Celine Vachon, Rebecca Smith-Bindman, Karla Kerlikowske

Endogenous Hormone Levels and Estimation of Breast Cancer Risk

A study (34) that combined data from nine prospective cohort studies of breast cancer found that women in the highest quintile of estradiol or testosterone had a higher relative risk of breast cancer (for estradiol, 2.00-fold, 95% CI = 1.47- to 2.71-fold; for testosterone, 2.22-fold, 95% CI = 1.59- to 3.10-fold) than those in the lowest quintile. We found six subsequent prospective studies (32,35-39), all of which found statistically significant associations between estradiol or testosterone levels and the risk of breast cancer.

Che differenza c'è tra uno studio di coorte o uno studio caso controllo?

Perché alcune volte negli articoli troviamo scritto:

Rischio Relativo (RR) e altre Odds Ratio (OR)?

Injectable and Oral Contraceptive Use and Cancers of the Breast, Cervix, Ovary, and Endometrium in Black South African Women: Case-Control Study

Methods and Findings: We analysed data from a South African hospital-based case-control study of black females aged 18-79 y, comparing self-reported contraceptive use in patients with breast ($n=1,664$), cervical ($n=2,182$), ovarian ($n=182$) and endometrial ($n=182$) cancer, with self-reported contraceptive use in 1,492 control patients diagnosed with cancers with no known relationship to hormonal contraceptive use. We adjusted for potential confounding factors, including age, calendar year of diagnosis, education, smoking, alcohol, parity/age at first birth, and number of sexual partners. Among controls, 26% had used injectable and 20% had used oral contraceptives. For current and more recent users versus never users of oral or injectable contraceptives, the odds ratios (ORs) for breast cancer were significantly increased in users of oral and/or injectable contraceptives (OR 1.66, 95% CI 1.28-2.16, $p<0.001$) and separately among those exclusively using oral (1.57, 1.03-2.40, $p=0.04$) and exclusively using injectable (OR 1.83, 1.31-2.55, $p<0.001$) contraceptives; corresponding OR for cervical cancer were 1.38 (1.08-1.77, $p=0.01$), 1.01 (0.66-1.56, $p=0.96$), and 1.58 (1.16-2.15, $p=0.004$). There was no

Fattori di rischio

- Esposizioni, comportamenti o attributi personali che influenzano il rischio di sviluppare una malattia
- Fattori che aumentano la probabilità di comparsa di una malattia o di una determinata condizione
- 4 categorie principali:
 - *Componenti genetiche o ereditarie*
 - *Componenti ambientali*
 - *Fattori nocivi presenti nell'ambiente di lavoro*
 - *Abitudini tipicamente individuali*

Associazione tra “esposizione” e malattia

- l'esistenza della associazione si può dimostrare confrontando la **frequenza della malattia** in due popolazioni costituite rispettivamente da persone **esposti** e **non esposti** ad una presunta causa (o determinante o «fattore di rischio»)
- **Per “esposizione”** si intende la presenza di qualsiasi variabile che, in linea di ipotesi, può causare un certo effetto

DUE APPROCCI PER LA SCOPERTA DI UN'ASSOCIAZIONE

**STUDIO PROSPETTICO
di COORTE**

**STUDIO RETROSPETTIVO
CASO-CONTROLLO**

Disegno di uno Studio Prospettivo

Reclutamento due gruppi di **individui sani**

- uno composto da soggetti esposti al fattore di rischio
- uno composto da soggetti quanto più possibile simili ai primi, salvo per il fatto di non essere esposti

Osservazione dei singoli individui per un periodo di tempo appropriato (follow-up) al fine di **stimare l'incidenza** di malattia

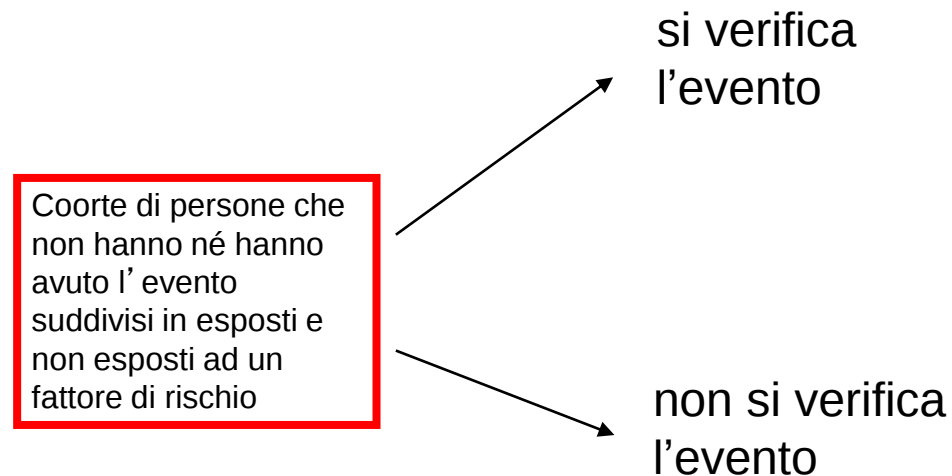
nei due gruppi di esposizione (esposti e non esposti).

Studi prospettici

Passato

Presente

Futuro



In uno studio prospettico, all'inizio l'evento malattia non è ancora avvenuto; trascorso un periodo di tempo si determina quanti individui hanno sviluppato la malattia tra gli esposti o non esposti ad un fattore di rischio

CALCOLO DEL RISCHIO RELATIVO (RR)

	Ammalati (M+)	Non ammalati(M-)	Totale
Esposti	a	b	a+b
Non esposti	c	d	c+d

Numericamente l'associazione tra "fattore di rischio e malattia" è misurata come rapporto (**R**ischio **R**elativo, RR) tra tassi di incidenza, ovvero come rapporto tra il rischi negli esposti e quello nei non esposti .

RR=

Rischio negli esposti: $a/(a+b)$

Rischio nei non esposti: $c/(c+d)$

Troppa carne rossa nella dieta?

Rischio del 32% in più di tumore del seno ...

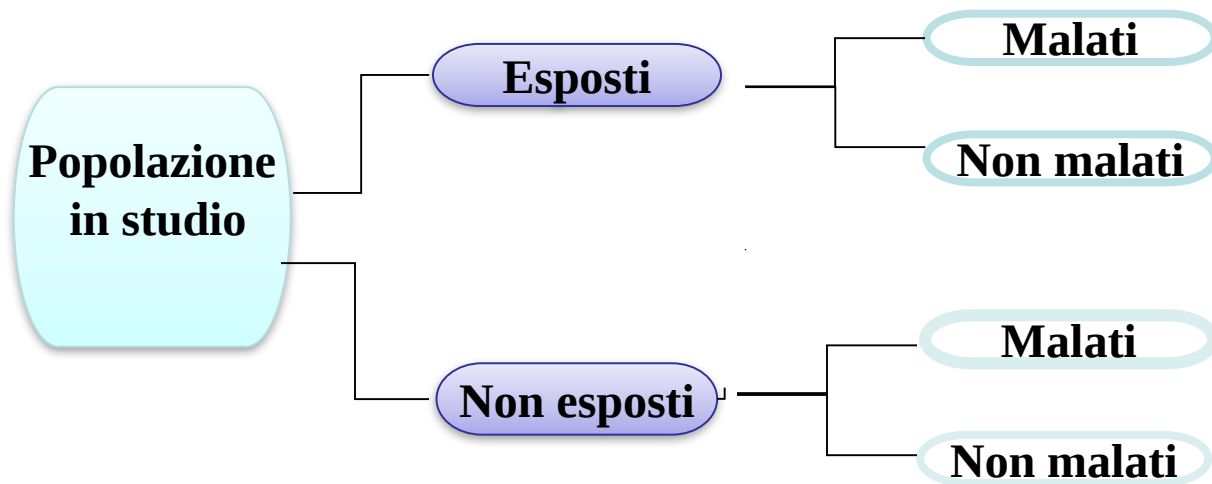
Sembra essere questa la conclusione di uno studio pubblicato dagli *Archives of Internal Medicine*, studio che aveva attirato una grande attenzione.



I ricercatori del Department of Medicine del Brigham and Women's Hospital e della **Harvard Medical School** hanno analizzato i dati riguardanti **90.659 donne** in fase pre-menopausale dai 26 ai 46 anni di età.

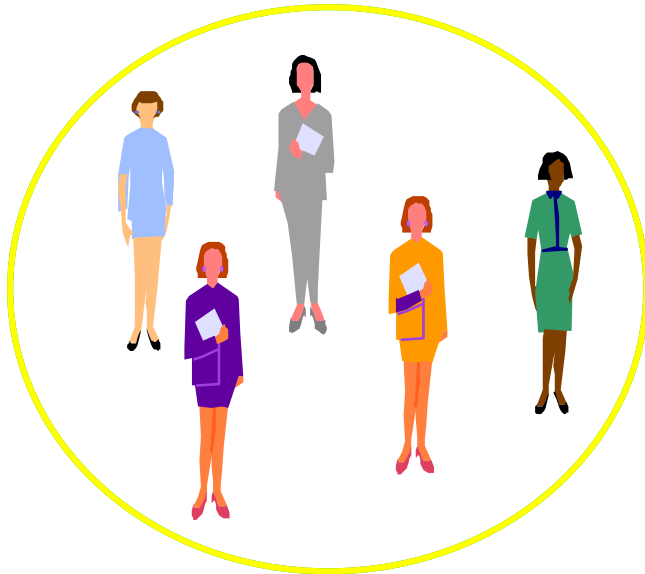
Il **consumo di carne rossa** è stato misurato mediante **questionari** sui regimi alimentari sottoposti alle pazienti ogni 4 anni nell'arco di 12 anni: i **casi** registrati di carcinoma invasivo della mammella sono stati **1021**.

Andiamo a leggere con attenzione quanto riportato nell'articolo scientifico (disegno dello studio, materiali e metodi, ...).

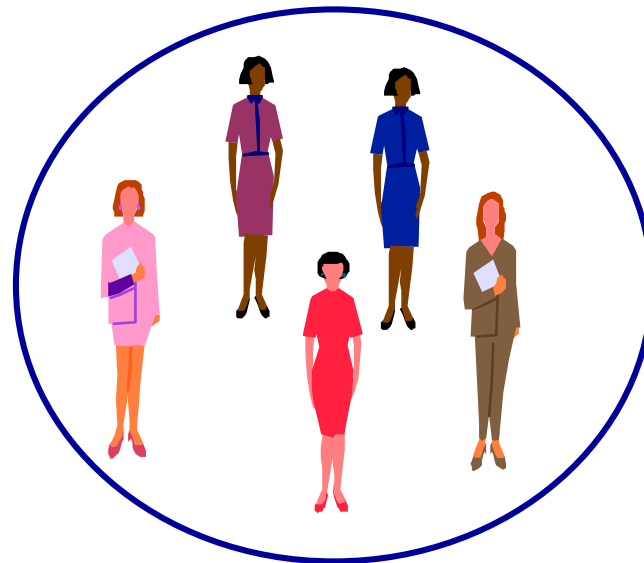


Follow-up di 12 anni

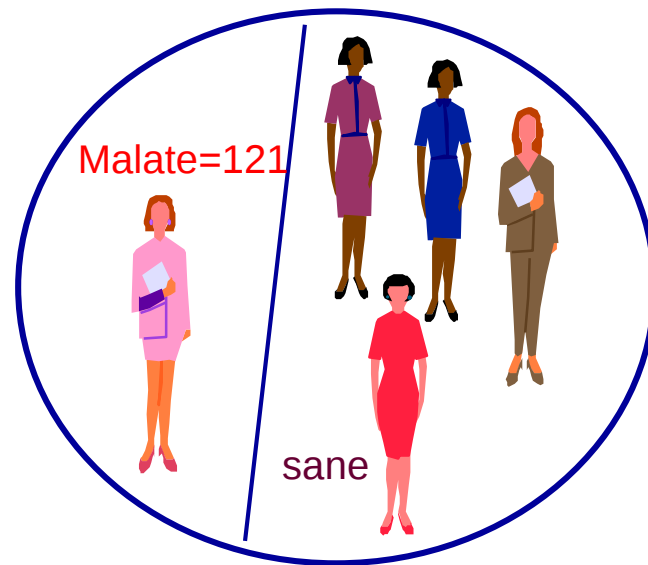
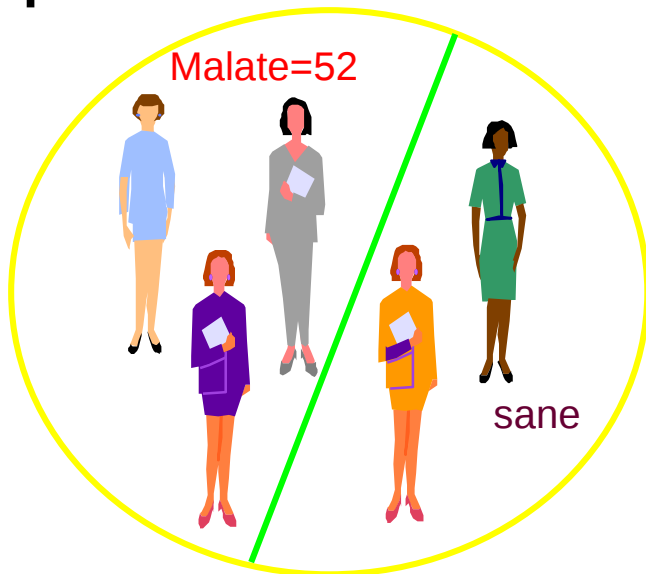
esposte
carne rossa >1,5 al giorno
8649 donne



non esposte
carne rossa $\leq 1,5$ volte alla settimana
26787 donne



esposte



non esposte

	Malate	Non Malate	
Esposte	52	8597	8649
Non Esposte	121	26666	26787
Totale	173	35263	35436

$$RR = \frac{\text{rischio nelle esposte}}{\text{rischio nelle non esposte}} = \frac{52/8649}{121/26787} = 1.32$$

Il valore di RR ottenuto - a patto che sia soddisfatta la condizione iniziale di fondamentale uniformità fra i due gruppi confrontati - ci informa su quanto aumenta la probabilità di contrarre la malattia se ci si espone al fattore di rischio:
come si interpreta il risultato ottenuto???

$$P(\text{malate}|\text{esposte}) / P(\text{malate}|\text{non_esposte})$$

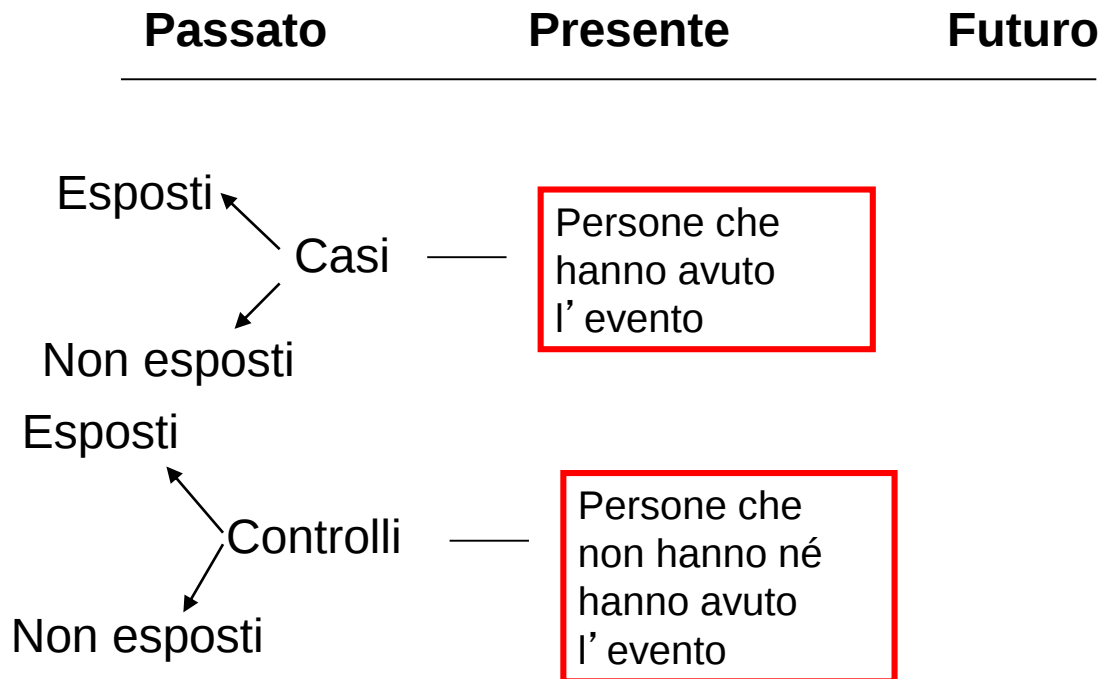
Schema di interpretazione del rischio relativo e dell'odds ratio.



Studi retrospettivi

- Gli studi retrospettivi sono ad esempio basati sempre sul reclutamento di due gruppi di individui da una stessa popolazione ma, questa volta, uno di soggetti che hanno contratto la malattia in studio (**casi**) e uno di individui che non l'hanno contratta (**controlli**) ma con caratteristiche simili ai casi
- Si valuta **retrospettivamente la precedente esposizione** al fattore di rischio , sia nei casi che nei controlli: su questi individui vengono raccolte informazioni circa l'esposizione pregressa al fattore di rischio (questionari, interviste).

Caso-controllo: studio retrospettivo



Studi di coorte (prospettici) e studi caso-controllo



Anche negli studi retrospettivi, i risultati possono essere riassunti in una tabella 2x2 **ma non si può** stimare l'associazione tra malattia e esposizione attraverso il rapporto tra la presenza di malattia negli esposti e nei non esposti, perché il numero di casi e controlli è scelto dal ricercatore

L'associazione tra fattore di rischio e malattia è misurata come rapporto tra **odds** di esposizione per casi e odds di esposizione per i controlli (ODDS RATIO)

ma cosa si intende con odds???

...l'odds...

La parola odds deriva dal mondo delle scommesse (classicamente anglosassone...) e non è facilmente traducibile in italiano.

Dal vocabolario

ODDS = probabilità, quotazione



L'odds, in termini matematici, rappresenta pertanto il rapporto fra eventi positivi ed eventi negativi.

L'odds ratio (OR) è un rapporto tra odds

CALCOLO DELL'ODDS RATIO (OR)

	Ammalati (Casi)	Non ammalati (Controlli)	Totale
Esposti	a	b	a+b
Non esposti	c	d	c+d
Totale	a+c	b+d	

Odds di esposizione nei casi = a/c

Odds di esposizione nei controlli = b/d

$$OR = \frac{\text{odds di esposizione nei casi}}{\text{odds di esposizione nei controlli}} = \frac{a/c}{b/d} = \frac{ad}{bc}$$

Esempio

Viene condotto uno studio riguardante l'associazione tra obesità (esposizione) e urolitiasi (patologia): sono stati selezionati 1014 casi e 1487 controlli; si va a valutare in entrambi i gruppi il fattore obesità.

	Urolitiasi si Casi	Urolitiasi no Controlli	Totale
Obesità si	383	322	705
Obesità no	631	1165	1796
Totale	1014	1487	2501

Essendo uno studio caso controllo andiamo a calcolare l'odds ratio

$$\text{Odds casi} = a/c = 383/631 = 0,607$$

$$\text{Odds controlli} = b/d = 322/1165 = 0,277$$

$$\text{Odds ratio } 0,607 / 0,277 = 2,19$$

Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study – March 11, 2020 –*The Lancet*

191 patients (135 from Jinyintan Hospital and 56 from Wuhan Pulmonary Hospital) were included in this study, of whom 137 were discharged and **54 died** in hospital. 91 (48%) patients had a comorbidity, with hypertension being the most common (58 [30%] patients), followed by diabetes (36 [19%] patients) and coronary heart disease (15 [8%] patients). Multivariable regression showed increasing odds of in-hospital death associated with older age (**odds ratio 1·10, 95% CI 1·03–1·17**, per year increase), higher Sequential Organ Failure Assessment (SOFA) score (**5·65, 2·61–12·23**; $p < 0·0001$), and d-dimer greater than 1 µg/mL (**18·42, 2·64–128·55**; $p = 0·0033$) on admission. ...

Current smoker (vs non-smoker): **2·23, 0·65–7·63** ...

Female sex (vs male): **0·61, 0·31–1·20** ...

Naufragio del Titanic

	morti	sopravvissuti	
U	1364	367	
D	126	344	

RR = 2,93

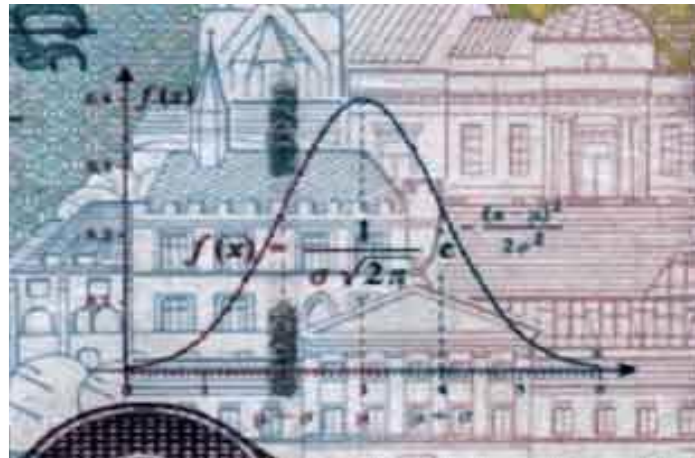
OR = 10,15

???

Gauss e ... la gaussiana / la normale

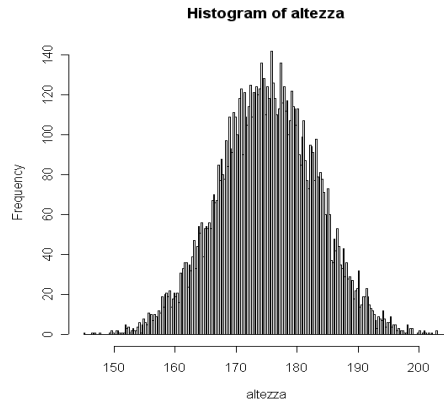


$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$



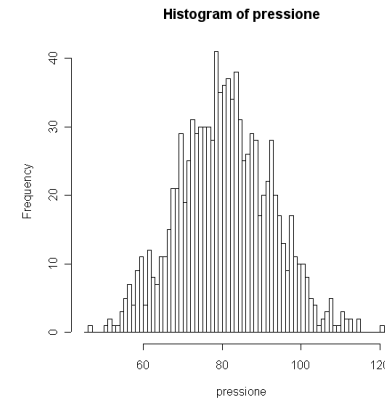
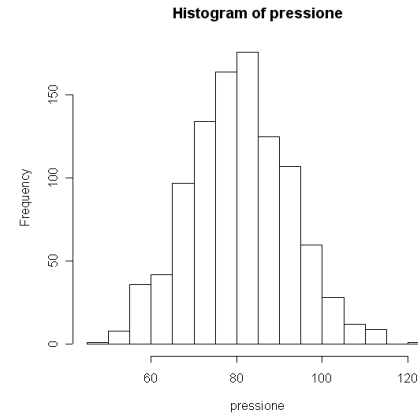
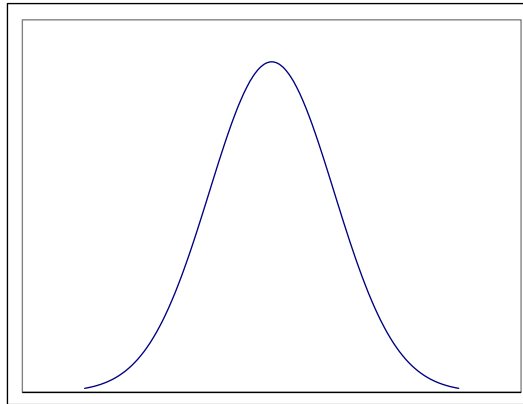
esempio: statura

- 10000 reclute

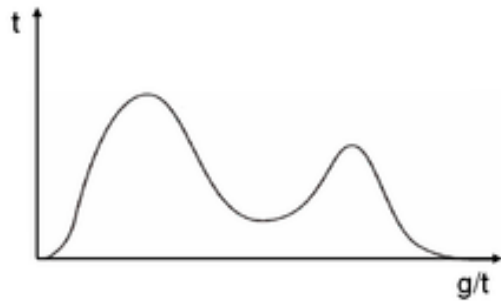
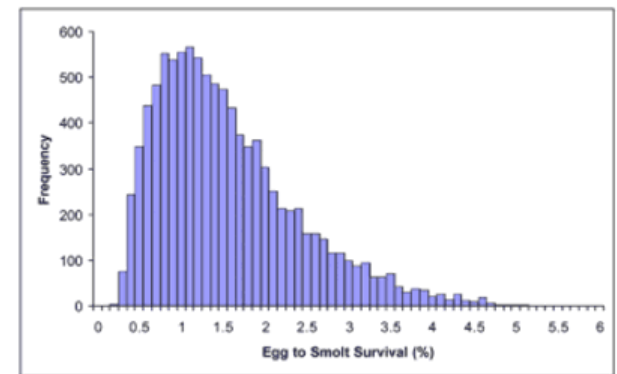
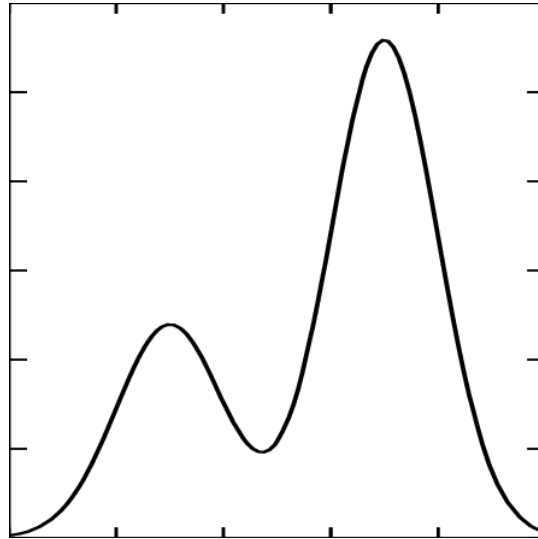


esempio: pressione

- 2000 donatori di sangue, età 35-44.

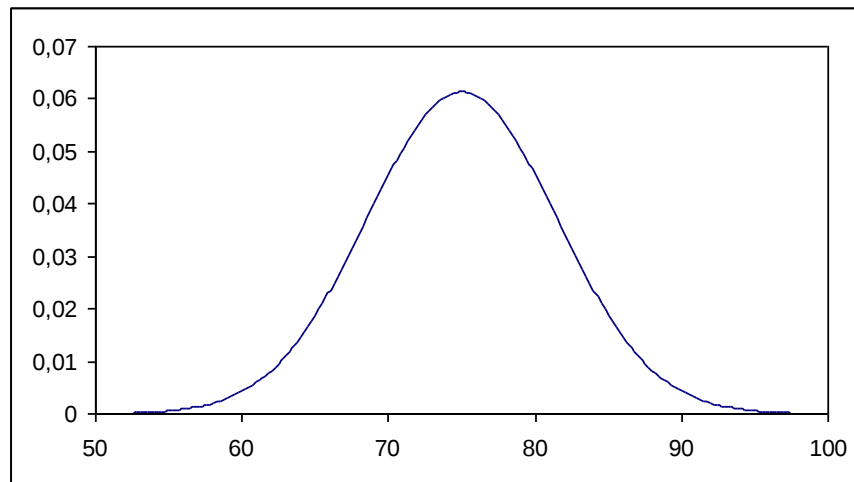


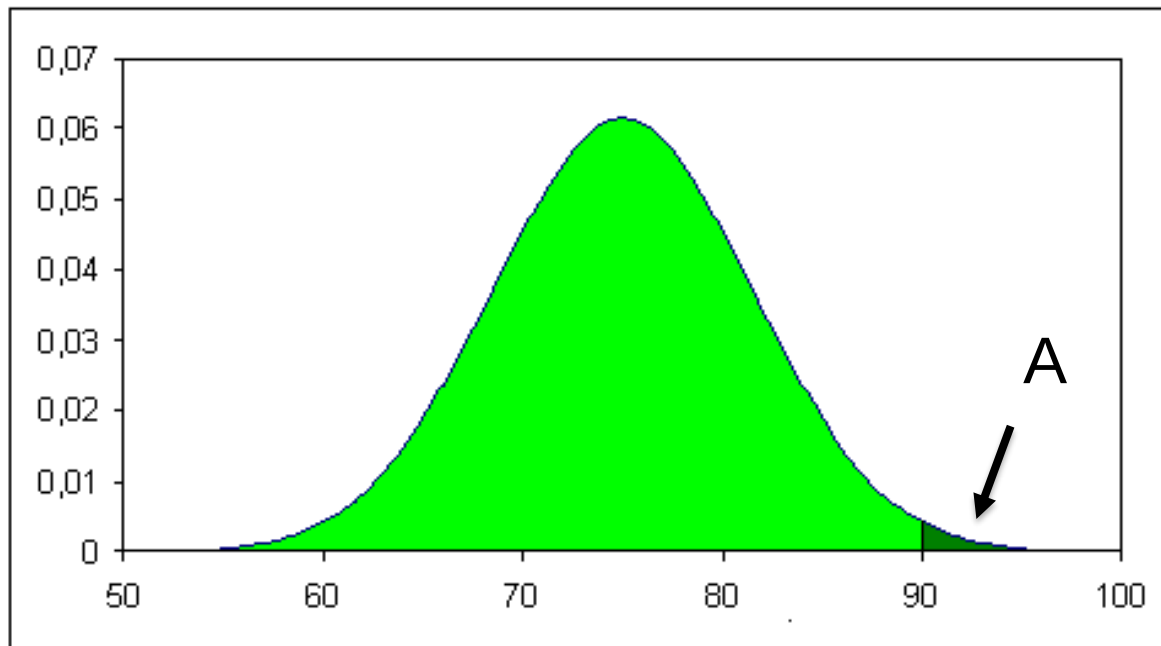
Ma... non tutti i dati si distribuiscono come una gaussiana!!



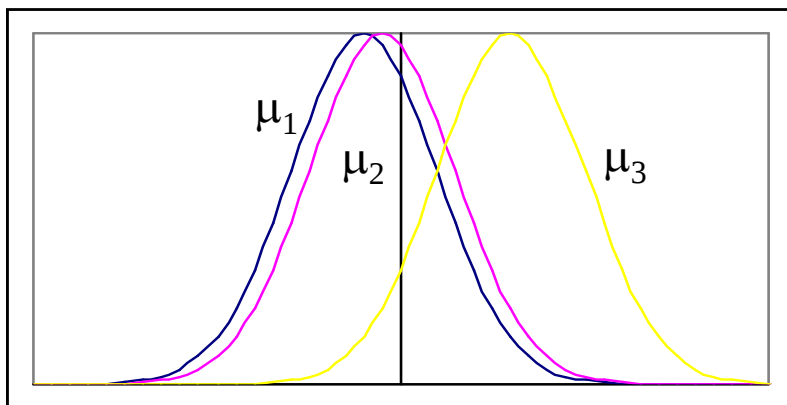
PROBLEMA: Si suppone che il peso della popolazione maschile di tesserati di una certa Federazione sportiva italiana di età maggiore di 18 anni si distribuisca secondo una **gaussiana di media 75 Kg e deviazione standard 6,5 Kg**. E' noto pure che per tale sport un atleta con peso maggiore di 90 Kg è da considerarsi 'sovrappeso'.

Con le informazioni in nostro possesso è possibile determinare la percentuale di popolazione sovrappeso?



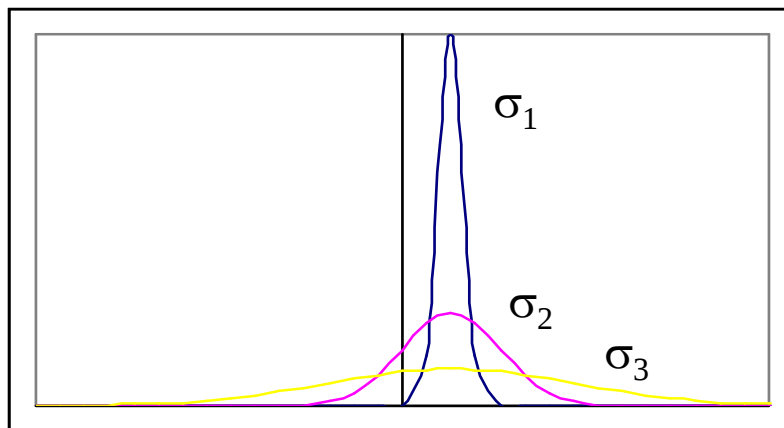


ad es. con Excel – DISTRIB.NORM.N : $A \cong 0,0105$



$$\mu_1 < \mu_2 < \mu_3$$

$$\sigma_1 = \sigma_2 = \sigma_3$$



$$\sigma_1 < \sigma_2 < \sigma_3$$

$$\mu_1 = \mu_2 = \mu_3$$

Principali proprietà della gaussiana $N(\mu;\sigma)$

- Simmetria rispetto a μ
- media = moda = mediana
- $\mu - \sigma$ e $\mu + \sigma$ punti di flesso
- Tra $\mu - \sigma$ e $\mu + \sigma$ sta circa il 68% dei dati
- Tra $\mu - 2\sigma$ e $\mu + 2\sigma$ sta circa il 95% dei dati
- Tra $\mu - 3\sigma$ e $\mu + 3\sigma$ sta circa il 99,7% dei dati
- $N(0;1)$ è la cosiddetta “normale standard”