Sui dati qualitativi

Alessandra Nardi

alenardi@mat.uniroma2.it

11 settembre 2019

"The Physicians' Health Study" è uno studio clinico randomizzato, doppio cieco, condotto allo scopo di valutare il possibile effetto di riduzione della mortalità cardiovascolare (infarto del mocardio) legato ad un uso regolare e continuato di piccole dosi di aspirina
Ciascun medico, di età compresa tra 40 e 84 anni, prese a giorni alterni 325 mg di aspirina o un semplice placebo senza essere a conoscenza di quale sostanza stesse realmente assumendo
Riportiamo nella seguente tabella i risultati relativi ad un rapporto preliminare (N.Engl.J.Med., 1988)

	Attacco cardiaco fatale	Attacco cardiaco non fatale	Nessun attacco cardiaco	Totale
Placebo	18	171	10845	11034
Aspirina	5	99	10933	11037
Totale	23	270	21778	22071

In generale una tabella di contingenza descrive la distribuzione congiunta di sue caratteri

$X \downarrow Y \rightarrow$	y_1	y_2	$\dots$	y_j	$\dots$	y_h	Totale
x_1	n_{11}	n_{12}	$\dots$	n_{1j}	$\dots$	n_{1h}	$n_{1.}$
x_2	n_{21}	n_{22}	$\dots$	n_{2j}	$\dots$	n_{2h}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_{i1}	n_{i2}	$\dots$	n_{ij}	$\dots$	n_{ih}	$n_{i.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_{k1}	n_{k2}	$\dots$	n_{kj}	$\dots$	n_{kh}	$n_{k.}$
Totale	$n_{.1}$	$n_{.2}$	$\dots$	$n_{.j}$	$\dots$	$n_{.h}$	$n_{..}$

- n_{ij} è la frequenza assoluta delle osservazioni che presentano contemporaneamente la modalità x_i del carattere X e la modalità y_j del carattere Y
- $n_{i.}$ è la frequenza assoluta marginale delle osservazioni che presentano la modalità x_i del carattere X, quale che sia la modalià del carattere Y
- $n_{.j}$ è la frequenza assoluta marginale delle osservazioni che presentano la modalità y_j del carattere Y, senza tener conto della presenza del carattere X

Il primo passo per interpretare i dati è il passaggio dalle frequenze assolute a quelle relative (perchè ?) Tuttavia, in una tabella doppia, esistono modi diversi di calcolare le frequenze relative
Se dividiamo le frequenze assolute per il totale delle osservazioni $(n = 22071)$, otteniamo le frequenze relative della distribuzione doppia $f_{ij} = \frac{n_{ij}}{n}$ e delle due distribuzioni marginali corrispondenti ai caratteri X $f_{i.} = \frac{n_{i.}}{n}$ e Y $f_{.j} = \frac{n_{.j}}{n}$

	Attacco cardiaco fatale	Attacco cardiaco non fatale	Nessun attacco cardiaco	Totale
Placebo	18	171	10845	11034
Aspirina	5	99	10933	11037
Totale	23	270	21778	22071

	Attacco cardiaco fatale	Attacco cardiaco non fatale	Nessun attacco cardiaco	Totale
Placebo	0.0008	0.008	0.491	0.50
Aspirina	0.0002	0.004	0.496	0.50
Totale	0.001	0.012	0.987	1

Tuttavia nel nostro studio siamo particolarmente interessati a leggere le differenze tra il gruppo di medici che hanno assunto aspirina ed il gruppo di controllo a cui è stato somministrato un semplice placebo

Calcoliamo allora le frequenze relative separatamente per i due gruppi, cioè le distribuzioni di frequenze relative della risposta, condizionatamente al tipo di trattamento. I totali di riferimento sono quelli marginali corrispondenti alla numerosità totale del gruppo dei controlli e dei "trattati".

	Attacco cardiaco fatale	Attacco cardiaco non fatale	Nessun attacco cardiaco	Totale
Placebo	18	171	10845	11034
Aspirina	5	99	10933	11037
Totale	23	270	21778	22071

	Attacco cardiaco fatale	Attacco cardiaco non fatale	Nessun attacco cardiaco	Totale
Placebo	0.0016	0.0155	0.9829	1
Aspirina	0.0005	0.0090	0.9905	1
Totale	0.0010	0.0122	0.9868	1

È possibile anche calcolare anche le distribuzioni di frequenze relative del tipo di trattamento, condizionatamente all'esito (per colonna)

	Attacco cardiaco fatale	Attacco cardiaco non fatale	Nessun attacco cardiaco	Totale
Placebo	18	171	10845	11034
Aspirina	5	99	10933	11037
Totale	23	270	21778	22071

	Attacco cardiaco fatale	Attacco cardiaco non fatale	Nessun attacco cardiaco	Totale
Placebo	0.78	0.63	0.498	0.50
Aspirina	0.22	0.37	0.502	0.50
Totale	1	1	1	1

...anche se nel nostro caso non ha molto senso, trattandosi di uno studio prospettico

Quando avrebbe senso?

È evidente che la chiave di lettura di una tabella doppia è legata al tipo di disegno clinico

Torniamo al nostro studio e proviamo a rispondere ad alcuni quesiti

- Cosa stimano le frequenze relative che abbiamo calcolato?
- Esiste una differenza *effettiva* tra i due trattamenti?
- Come quantifichiamo la differenza nella risposta ai due trattamenti?

Partiamo dalla prima domanda: cosa stimano le frequenze relative che abbiamo calcolato?

Esiste in statistica uno stretto legame tra frequenze relative e probabilità sancito dalla famosa *legge dei grandi numeri* , spesso citata, raramente conosciuta.
In verità il riferimento è quasi sempre al Terorema di Bernoulli. Immaginiamo una successione di prove indipendenti, ripetute nelle medesime condizioni. In ciascuna prova osserviamo (o meno) un evento aleatorio. Allora la frequenza relativa con cui osserviamo l'evento (numero di successi) tende alla probabilità dell'evento stesso al tendere ad infinito del numero di prove.
La versione *empirica* (legge empirica del caso) sancisce il fatto che la frequenza relativa con cui osserviamo un evento stima la probbailità dell'evento e che la precisione di tale stima cresce al crescere del numero delle prove.

Dovrebbe allora chiarirsi che la frequenza relativa 0.0016 stima la probbailità di avere un attacco cardiaco fatale assumendo solo un placebo o se preferite l'incidenza di infarti del miocardio nell'arco nel periodo di follow-up nell'intera popolazione
Analogamente 0.0005 stima la stessa incidenza nell'ipotesi che l'intera popolazione fosse trattata in modo regolare e continuato con dosi ridotte di Aspirina
Accanto ad ogni stima deve sempre comparire l'errore ad essa associato Cosa accadrebbe se ripetessimo l'esperimento nelle medesime condizioni?
Lo *standard error* è una misura della variabilità delle stime che otterremmo in ipotetiche ripetizioni dell'esperimento (principio del campionamento ripetuto)
Nel nostro caso lo stardard error è 0.00038 per il placebo e 0.00020 nel caso dell'aspirina, in entrambi i casi un errore molto basso
È spesso utile riassumere stima ed errore in un intervallo di confidenza (al 95%): (0.0009971129, 0.0026327613) nel caso di placebo e (0.000166796, 0.001123226) nel caso di assunzione di aspirina

Veniamo al secondo quesito: esiste una differenza *effettiva* tra i due trattamenti?
Le probabilità che abbiamo stimato suggeriscono che una differenza esista. La domanda allora può essere rifrasata come: la differenza osservata corrisponde ad un effetto reale o potrebbe essere legata al caso? Ripetendo l'esperimento nelle medesime condizioni comparirebbe ancora? Potrebbe diventare irrilevante o oddirittura cambiare direzione?
È la domanda a cui risponde un test d'ipotesi
L'ipotesi nulla del nostro test è che la differenza osservata sia casuale ovvere che il tipo di trattamento e la risposta siano due variabili *indipendenti*
L'ipotesi alternativa, quella che vorremmo dimostrare, è la presenza di un legame (associazione) tra le due variabili

Ricordiamo la definizione probabilistica di indipendenza: due eventi aleatori A e B sono indipendenti se $Prob(A|B) = Prob(A)$
Nel nostro esempio l'ipotesi di indipendenza implica

$$\begin{aligned} Prob(Att\,card\,fatale|Aspirina) &= Prob(Att\,card\,fatale) \\ Prob(Att\,card\,fatale|Placebo) &= Prob(Att\,card\,fatale) \\ Prob(Att\,card\,fatale|Aspirina) &= Prob(Att\,card\,fatale|Placebo) \end{aligned}$$

...
L'ipotesi alternativa implica che almeno una di queste disuguaglianze non sia vera
Ricordiamo che le ipotesi riguardano la popolazione e non il campione osservato

Un test statistico è una procedura decisionale che, sulla base dei dati osservati, porta a concludere a favore di una delle due ipotesi

A questo fine l'informazione sperimentale deve essere sintetizzata definindo una opportuna *statistica test*
Quella che cerchiamo è una funzione dei soli dati campionari che misuri il grado di associazione tra il trattamento e la risposta. In altri termini una misura della *distanza* dell'evidenza sperimentale dall'ipotesi nulla. Per ottenerla costruiamo, accanto alla tabella osservata, una tabella *attesa* nel caso di indipendenza tra le due variabili

Sotto l'ipotesi nulla ci attendiamo

$$\begin{aligned} Prob(Att\,card\,fatale|Placebo) &= Prob(Att\,card\,fatale) \\ \frac{n_{11}}{n_{1.}} &= \frac{n_{.1}}{n_{..}} \\ Prob(Att\,card\,fatale|Aspirina) &= Prob(Att\,card\,fatale) \\ \frac{n_{21}}{n_{2.}} &= \frac{n_{.1}}{n_{..}} \\ \frac{n_{11}}{n_{1.}} &= \frac{n_{21}}{n_{2.}} \end{aligned}$$

Poichè lo stesso vale per le altre probabilità in esame, ne deriva che le due distribuzioni (frequenze relative) della risposta condizionate all'assunzione di Placebo e Aspirina saranno uguali tra loro

Nel nostro caso la tabella attesa nel caso di indipendenza sarà

	Attacco cardiaco fatale	Attacco cardiaco non fatale	Nessun attacco cardiaco	Totale
Placebo	11.4 0.001	135 0.012	10887.6 0.987	11034 1
Aspirina	11.5 0.001	135 0.012	10887.5 0.987	11037 1
Totale	23 0.001	270 0.012	21778 0.987	22071 1

Da mettere a confronto con la tabella osservata

	Attacco cardiaco fatale	Attacco cardiaco non fatale	Nessun attacco cardiaco	Totale
Placebo	18 0.002	171 0.015	10845 0.983	11034 1
Aspirina	5 0.0004	99 0.009	10933 0.9906	11037 1
Totale	23 0.001	270 0.012	21778 0.987	22071 1

In generale, nell'ipotesi di indipendenza, le distribuzioni condizionate di frequenze relative di $Y|(X = x_i)$ saranno uguali tra loro ed in particolare uguali alla distribuzione marginale di Y (indipendenza in distribuzione)
In simboli

$$\frac{n_{ij}}{n_{i.}} = \frac{n_{kj}}{n_{k.}} = \frac{n_{.j}}{n_{..}}$$

da cui

$$\tilde{n}_{ij} = \frac{n_{i.}n_{.j}}{n_{..}} \quad o \quad \tilde{f}_{ij} = f_{i.}f_{.j}$$

Quanto sono "distanti" i dati osservati dalla situazione di indipendenza (date le marginali)?
Il χ^2 (Pearson, 1900)

$$\begin{aligned} \chi^2 &= \sum_{i=1}^{k}\sum_{j=1}^{h} \frac{(n_{ij} - \tilde{n_{ij}})^2}{\tilde{n_{ij}}} \\ &= n\sum_{i=1}^{k}\sum_{j=1}^{h} \frac{(f_{ij} - f_{i.}f_{.j})^2}{f_{i.}f_{.j}} \end{aligned}$$

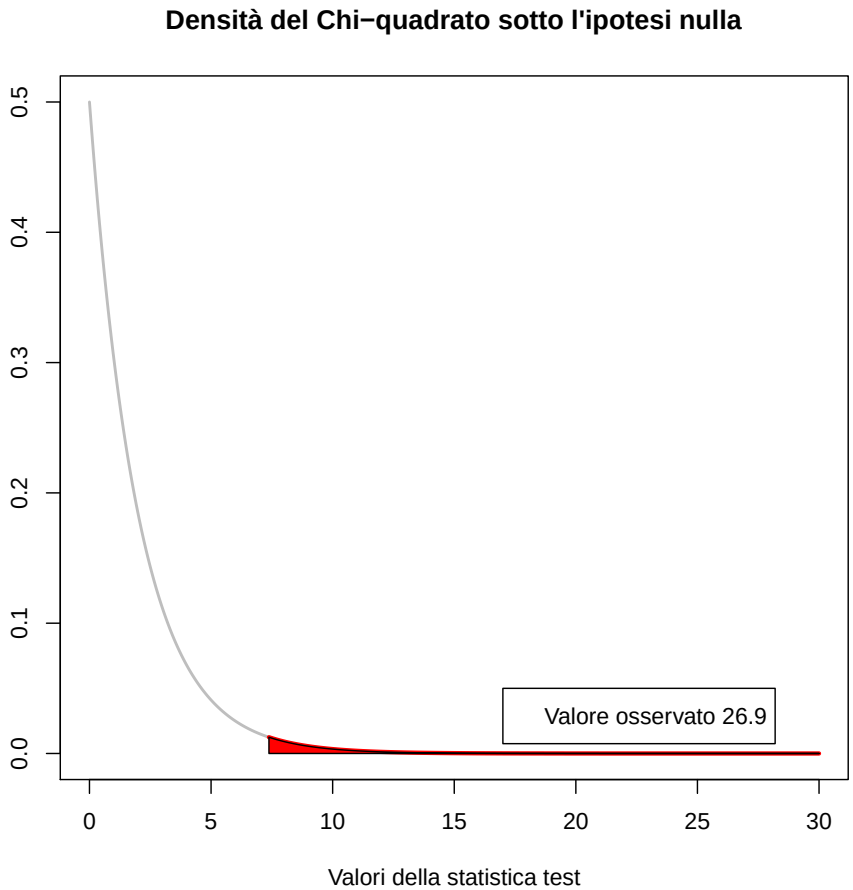
La statistica test χ^2 misura il grado di associazione tra trattamento e risposta come distanza tra la tabella osservata e quella attesa nel caso di indipendenza (nessuna differenza tra i due trattamenti)
Come misura del lengame tra le due variabili ha due difetti: non ha un massimo univoco e dipende dalla numerosità del nostro campione
È invece appropriata per portare a termine il nostro test.

Si dimostra infatti che sotto l'ipotesi di indipendenza, al crescere di n, la statistica χ^2 tende a distribuirsi come una variabile aleatoria χ^2 con (k-1)(h-1) gradi di libertà

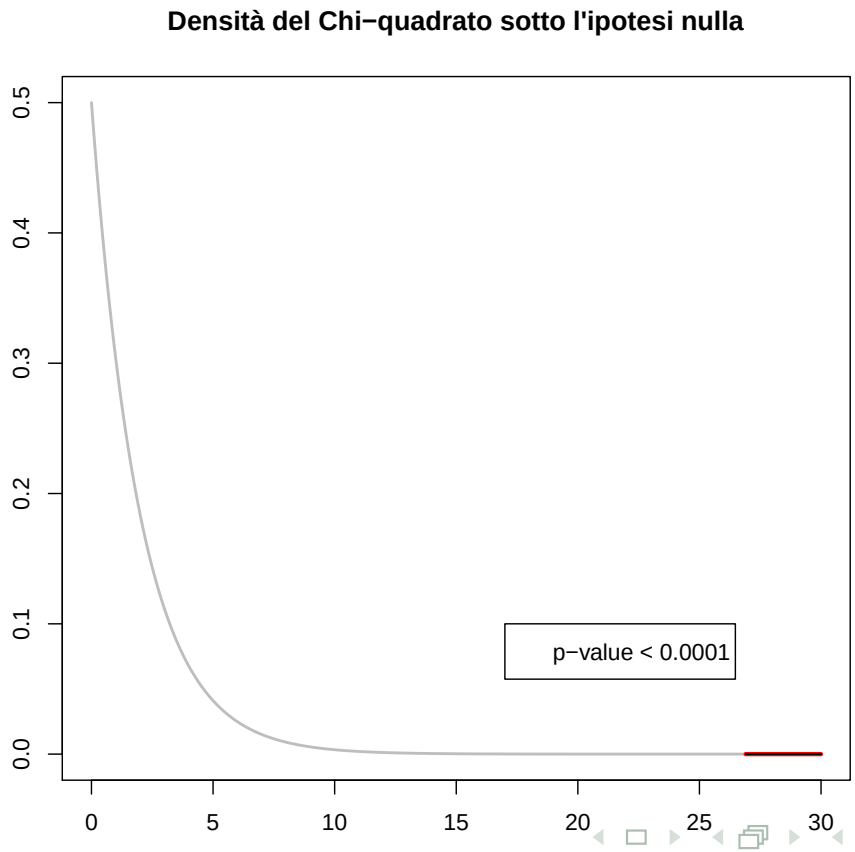
La conoscenza della distribuzione della statistica test sotto l'ipotesi nulla è fondamentale per poter costruire la nostra regione di rifiuto ed arrivare alla decisione finale.

Ai fini di tale decisione non è tanto importante il valore osservato della statistica test quanto la sua probabilità sotto l'ipotesi nulla di indipendenza.

Fissata la dimensione del test a 0.05 (area in rosso) possiamo costruire la regione di rifiuto del test. Nel nostro esempio abbiamo ottenuto $\chi^2 = 26.9\ldots$



. . . che sotto la distribuzione χ^2 con 2 gradi di libertà corrisponde ad un p-value praticamente nullo. È quindi confermata la presenza di un legame tra trattamento e risposta, la differenza osservata corrisponde ad una differenza effettiva nella popolazione, seppure con un margine di errore.



Stabilita la presenza di un legame, dobbiamo entrare nel merito della sua natura
Passiamo a rispondere alla terza domanda: come quantifichiamo la differenza nella risposta ai due trattamenti?
Per semplicità focalizziamo la nostra attenzione sull'evento cardiaco senza distinguere se fatale o meno (collassiamo la tabella).

	Attacco cardiaco	Nessun attacco cardiaco	Totale
Placebo	189	10845	11034
Aspirina	104	10993	11037
Totale	293	21778	22071

Per valutare l'entità dell'effetto dell'aspirina come trattamento di prevenzione dell'infarto del miocardio indichiamo con $\pi_{AC|A}$ e $\pi_{AC|P}$ le probabilità di avere un attacco cardiaco se sottoposti a terapia preventiva rispettivamente a base di aspirina e placebo. Queste probabilità si riferiscono all'intera popolazione (o se volete ad un paziente estratto casualmente) e all'arco temporale dello studio, 8 anni dall'inizio del trattamento.
Non possono invece essere riferite al singolo paziente, con precise caratteristiche.
Le stime che otteniamo sono rispettivamente $104/11037 = 0.0094$ e $189/11037 = 0.0171$.

Come confrontiamo le due probabilità stimate?
Tre sono le misure possibili

- La differenza

  $\pi_{AC|P} - \pi_{AC|A}$ $0.0171 - 0.0094 = 0.0077$

- Il rapporto

  $\frac{\pi_{AC|P}}{\pi_{AC|A}}$ $\frac{0.0171}{0.0094} = 1.82$

- L'odds ratio

  $\frac{\pi_{AC|P}/(1-\pi_{AC|P})}{\pi_{AC|A}/(1-\pi_{AC|A})}$ $\frac{0.0171/0.9829}{0.0094/0.9906} = 1.83$

Sebbene meno intuitiva, la misura più utilizzata nell'analisi statistica è la terza, l'*odds ratio*
Tra le ragioni, alcune delle quali tecniche, c'è il fatto che, a certe condizioni, può essere considerato un'approssimazione del rischio relativo, misura cara ai clinici.
Prima di preseguire dobbiamo però chiarire cosa intendiamo per *rischio*

Il concetto di rischio è strettamento legato al tempo
Per darne una definizione formale iniziamo col chiarire cosa intendiamo per tempo di sopravvivenza specificando

- un'origine in cui inizia l'osservazione del paziente
- l'evento di interesse
- una unità di misura

Defiamo poi come

$$T : \text{tempo di sopravvivenza}$$

l'intervallo tra l'origine e il momento aleatorio in cui si verifica l'evento
Ne deriva che T è una variabile aleatoria a valori positivi.

T è caratterizzata da alcune funzioni del tempo
La densità di probabilità

$$f(t) = \lim_{\Delta \to 0^+} \frac{Pr\{t \leq T < t + \Delta\}}{\Delta}$$

da cui

$$Pr\{t \leq T < t + \Delta\} \simeq f(t)dt$$

La funzione di rischio

$$\lambda(t) = \lim_{\Delta \to 0} \frac{Pr\{t \leq T < t + \Delta | T \geq t\}}{\Delta}$$

$$Pr\{t \leq T < t + \Delta | T \geq t\} \simeq \lambda(t)dt$$

La funzione di sopravvivenza

$$S(t) = Pr\{T \geq t\}$$

da cui

$$\lambda(t)dt = \frac{f(t)dt}{S(t)}$$

o in altri termini

$$Pr\{t \leq T < t + \Delta | T \geq t\} = \frac{Pr\{t \leq T < t + \Delta\}}{Pr\{T \geq t\}}$$

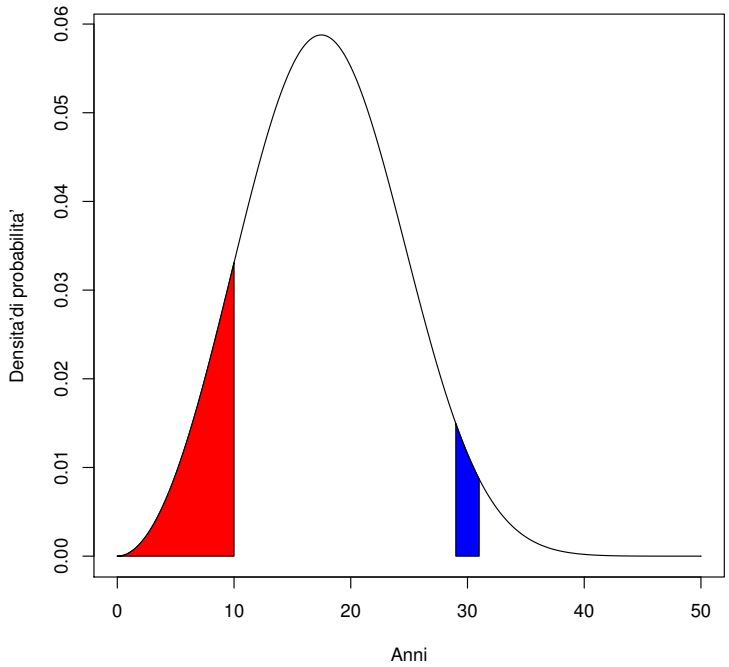
Questo implica che probabilità e rischio sono simili solo se $S(t) \simeq 1$

Apriamo una piccola parentesi....cos'è una densità di probabilità ?
Se i valori osservabili come risultato del nostro esperimento sono nel discreto (pochi) potremo assegnare ad ognuno una probabilità di essere osservato, facendo attenzione a garantire che la loro somma sia 1.
Se il risultato del nostro esperimento o della nostra osservazione è una variabile continua
non è più possibile assegnare una probabilità ad ogni singolo valore osservabile (sono davvero troppi!)

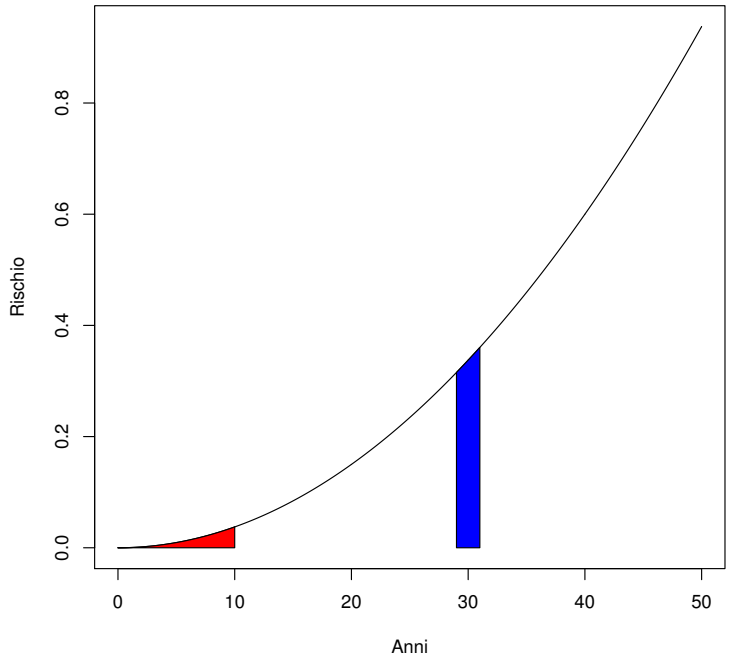
Passiamo a considerare invece che singoli valori degli intervalli (per quanto piccoli ognuno di essi conterrà sempre un numero infinito di punti) e definiamo

$$P(x \leq X < x + dx) = \int_{x}^{x+dx} f(t)dt \approx f(x)dx$$

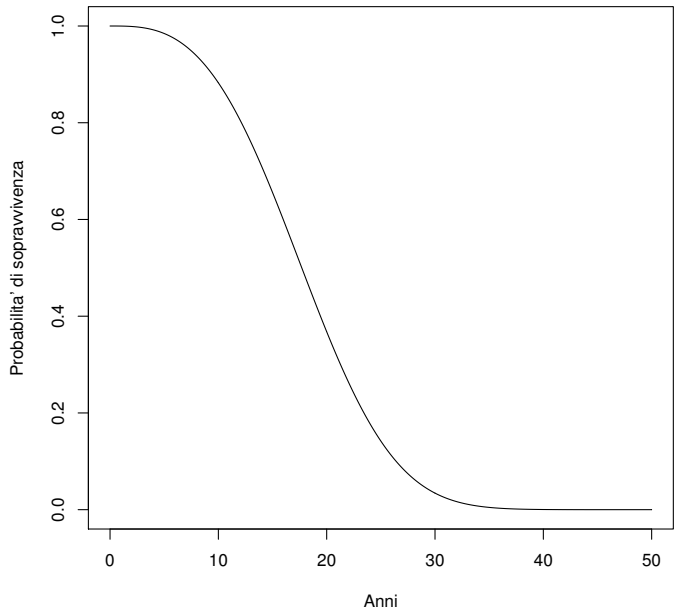
La funzione $f(x)$ prende il nome di densità e distribuisce la massa di probabilità unitaria sull'isieme dei valori osservabili.
La probabilità di osservare valori compresi in un dato intervallo è l'area corrispondente sotto questa curva



Densità di probabilità



Funzione di rischio

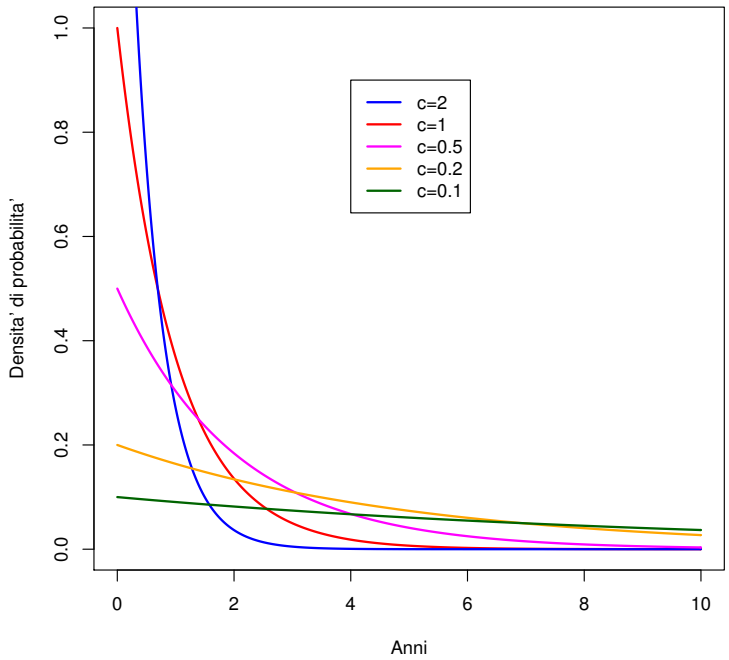


Funzione di sopravvivenza

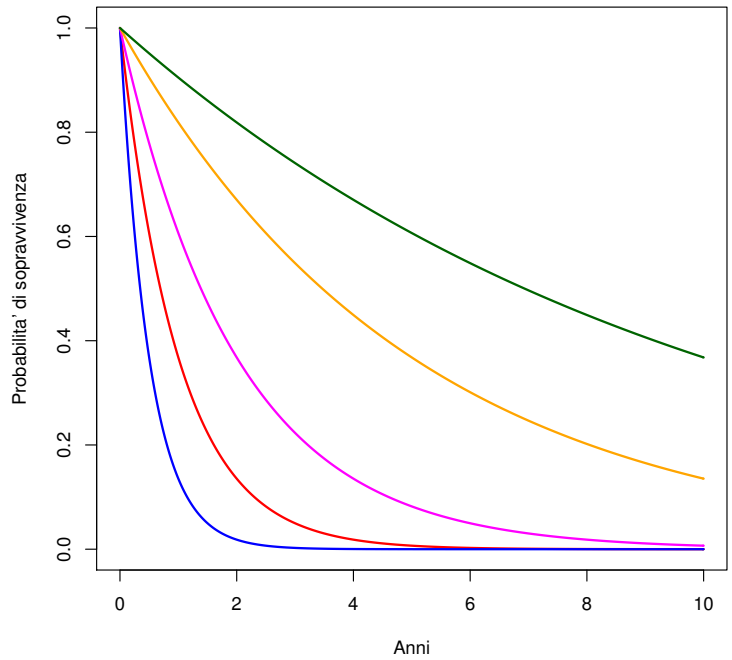
Compreso che densità e rischio sono funzioni del tempo diventa chiaro che quando ignoriamo t stiamo implicitamnete assumendo che il rischio sia costante almeno nell'intervallo di tempo che stiamo considerando.
Questa assunzione, di fatto molto restrittiva e spesso irrealistica se l'intervallo non è breve, è di fatto equivalente a quella di ipotizzare per T un modello esponenziale caratterizzato proprio da

$$\lambda(t) = c$$

dove c è una costante positiva. Vediamo alcuni esempi di distribuzioni esponenziali



Modello esponenziale: densità di probabilità



Modello esponenziale: funzioni di sopravvivenza

La stima della densità come quella della funzione di rischio sono molto complesse e spesso associate ad una bassa precisione.
Ne segue che si preferisce in genere stimare la funzione di sopravvivenza, probabilità valutate su intervalli di tempo oppure misure relative, prima fra tutte il rischio relativo.
Torniamo quindi al nostro studio.

Il rapporto $\frac{\pi_{AC|P}}{\pi_{AC|A}}$ può essere considerato un rischio relativo?
Ricordiamo che

$$\lambda(t)dt = \frac{f(t)dt}{S(t)}$$

Ne deriva che probabilità e rischio sono simili se l'evento è raro:
$S(t) \simeq 1$ e quindi $\lambda(t)dt \simeq f(t)dt$,
Attenzione tuttavia al fatto che le nostre probabilità fanno riferimento ad un intervallo di 8 anni mentre il rischio è riferito ad un intervallo infinitesimo.
Quindi il rapporto $\frac{\pi_{AC|P}}{\pi_{AC|A}}$ può essere assimilato ad un rischio relativo solo sotto l'ulteriore assunzione di un rischio constante nel tempo,
$\frac{\lambda_{AC|P} \; x \; 8anni}{\lambda_{AC|A} \; x \; 8anni}$

Torniamo sul significato di odds per osservare che l'odds di un attacco cardiaco assumendo aspirina è

$$\frac{\pi_{AC|A}}{(1-\pi_{AC|A})} = \frac{104/11037}{10933/11037} = \frac{104}{10933}$$

Mentre la probabilità viene stimata dal rapporto tra chi ha avuto un attacco cardiaco e il totale dei pazienti, la stima dell'odds è basata sul rapporto tra chi ha avuto e chi non ha avuto un attacco cardiaco. Si tratta di un linguaggio comune tra ... scommetitori

Se l'evento è raro, $1 - \pi_{AC|A} \simeq 1$, e il rischio costante, odds ratio e rischio relativo tendono a coincidere.

L'odds ratio diventa essenziale se invece di uno studio prospettico randomizzato analizziamo dati che provengono da uno studio caso-controllo
Doll e Hill nel 1952 dimostrarono per la prima volta una relazione significativa tra fumo e cancro polmonare. I dati si riferiscono ad uno studio retrospettivo caso-controllo condotto in Inghilterra

Numero medio giornaliero di sigarette	Cancro polmonare	Controlli	Totale
Nessuna	7	61	68
< 5	55	129	184
$5 - 14$	489	570	1059
$15 - 24$	475	431	906
$25 - 49$	293	154	447
≥ 50	38	12	50
Totale	1357	1357	2714

Lo studio caso-controllo parte dal reclutamento di un campione di casi, nel nostro caso di pazienti effetti da cancro polmonare.
A questo campione viene affiancato un campione di individui *sani* e su entrambi vengono raccolte retrospettivamente le informazioni relative ai fattori rischio in esame, nel nostro caso il fumo.
Notate come la logica causa-effetto sia invertita, adesso partiamo conoscendo l'evento (effetto) ed è la causa che lo ha generato ad essere aleatoria.
Come interpretare i risultati?

Si tratta di un disegno sperimentale altamente informativo per la disponibilità in tempi brevi di un elevato numero di casi e lapossibilià di indagare simultaneamente l'effetto di diversi fattori di rischio.
Ha tuttavia due punti deboli: ci consente di stimare solo misure relative ed è estremamente suscettibile a possibili distorsioni nel reclutamento (selection bias).
Vediamo aluni punti importanti su cui riflettere

Circa la scelta dei casi:

- dal momento che cerchiamo una relazione causa-effetto è preferibile individuare patologie omogenee per eziologia;
- casi prevalenti o incidenti? La prima soluzione potrebbe portare a selezionare i pazienti sopravvissuti da una serie più ampia di casi incidenti con il rischio di confondere fattori che influenzano lo sviluppo e la prognosi della patologia. L'antigene HLA A2 sembrava inizialmente legato allo sviluppo della leucemia acuta mentre si è poi rilevato un fattore prognostico per la sopravvivenza del paziente (Rogentine et al.1972,1973);
- nei casi prevalenti cause ed affetti della patologia possono sovrapporsi. Il paziente potrebbe cambiare stile di vita a seguito della diagnosi. Necessità di una attenta ricostruzione retrospettiva non sempre facile.

Circa scelta dei controlli

- su base ospedaliera oppure selezioniamo un campione casuale dalla popolazione?
- la scelta dei controlli è strettamente legata alle modalità con cui abbiamo scelto i casi. L'obiettivo primario è garantire la comparabilità dei gruppi che richiede ad esempio omogeneità nei test diagnostici;
- attenzione al rischio di selezionare i casi (o i controlli) in base a caratteristiche legate ai fattori di rischio in esame (maggiore attenzione diagnostica in presenza di sintomi non esclusivamente legati alla patologia in esame);
- bisognerebbe evitare che i controlli presentino patologie la cui eziologia è simile alla malattia in studio o che interessano lo stesso organo (cancro polmonare,broncopatie croniche e fumo);
- l'idea che casi e controlli debbano essere il più possibile simili (a parte la patologia oggetto di studio) che appartiene agli studi propsettici randomizzati, è completamente errata e inapplicabile in uno studio caso-controllo (*overmatching*).

Torniamo allo studio di Doll e Hill

Numero medio giornaliero di sigarette	Cancro polmonare	Controlli	Totale
< 5	62	190	252
≥ 5	1295	1167	2462
$Totale$	1357	1357	2714

Cosa siamo in grado di stimare sulla base dei dati osservati?
I rapporti $1357/2714 = 0.5$, $252/2714 = 0.093$, $62/252 = 0.25$
$1295/2462 = 0.53$ hanno un valore inferenziale?

Il campione osservato non è stato estratto casualmente dalla popolazione e non consente la stima delle caratteristiche della popolazione stessa.
Siamo in grado di stimare solo quantità relative
Obiettivo dello studio è stimare il rapporto $\frac{\pi_{CP|F}}{\pi_{CP|NF}}$ ma in questo disegno sperimentale l'elemento aleatorio è l'esposizione o menoal fumo in pazienti di cui è nota la presenza o mano della malattia. In altri termini potremmo stimare il rapporto $\frac{\pi_{F|CP}}{\pi_{F|NCP}}$ diverso dal precedente.

Le proprietà dell'odds ratio ci vengono in aiuto

$$\frac{\frac{\pi_{F|CP}}{(1-\pi_{F|CP})}}{\frac{\pi_{F|NCP}}{(1-\pi_{F|NCP})}} = \frac{\frac{1295/1357}{62/1357}}{\frac{1167/1357}{190/1357}} = \frac{1295 \times 190}{62 \times 1167} = \frac{\frac{1295/2462}{1167/2462}}{\frac{62/252}{190/252}} = \frac{\frac{\pi_{CP|F}}{(1-\pi_{CP|F})}}{\frac{\pi_{CP|NF}}{(1-\pi_{CP|NF})}}$$

L'elegante simmetria dell'odds ratio implica che numericamente le due quantità coincidono pur mantenendo un significato diverso.

Questo implica che pur stimando $\frac{\frac{\pi_{F|CP}}{(1-\pi_{F|CP})}}{\frac{\pi_{F|NCP}}{(1-\pi_{F|NCP})}}$ noi otteniamo indirettamente una stima della qunatità di interesse $\frac{\frac{\pi_{CP|F}}{(1-\pi_{CP|F})}}{\frac{\pi_{CP|NF}}{(1-\pi_{CP|NF})}}$ pari a $\frac{246050}{72354} = 3.40$

Quando l'odds ratio può essere considerato una approssimazione del rischio relativo?
Se $\pi_{CP|F} \simeq 1$ e $\pi_{CP|NF} \simeq 1$ cioe' se l'evento è raro

$$\frac{\frac{\pi_{CP|F}}{(1-\pi_{CP|F})}}{\frac{\pi_{CP|NF}}{(1-\pi_{CP|NF})}} \simeq \frac{\pi_{CP|F}}{\pi_{CP|NF}}$$

Se inoltre il rischio di sviluppare un tumore può essere considerato costante

$$\frac{\pi_{CP|F}}{\pi_{CP|NF}} \simeq \frac{\lambda_{CP|F} \times \Delta t}{\lambda_{CP|NF} \times \Delta t}$$

Realistica?