II χ^2 (Pearson, 1900)

Relazioni tra variabili: le tabelle di contingenza

"The Physicians' Health Study" è uno studio clinico randomizzato condotto allo scopo di valutare il possibile effetto di riduzione della mortalità cardiovascolare legato ad un uso regolare e continuato di aspirina

Ciascun medico che partecipò allo studio prese a giorni alterni una pasticca 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

In simboli

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

dove

- 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

Calcoliamo 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

Nel nostro studio siamo tuttavia particolarmente interessati a comprendere 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 dell'esito, condizionatamente al tipo di trattamento Adesso i totali di riferimento sono quelli marginali corrispondenti alla numerosità totale del gruppo dei controlli ($n_{1.}$) e dei "trattati" ($n_{2.}$) In generale la distribuzione di frequenza della variabile condizionata $Y|(X = x_i)$ sarà

Modalità di $Y\|(X = x_i)$	y_1	y_2	$\ldots$	y_h	Totale
Frequenze assolute	n_{i1}	n_{i2}	$\ldots$	n_{ih}	$n_{i.}$
Frequenze relative	$n_{i1}/n_{i.}$	$n_{i2}/n_{i.}$	$\ldots$	$n_{ih}/n_{i.}$	1

	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

Esiste una qualche differenza? Come la misuriamo?

È possibile calcolare anche le distribuzioni di frequenze relative del tipo di trattamento, condizionatamente all'esito...

	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

Indipendenza in probabilità

L'ipotesi nulla del nostro test è l'indipendenza delle due variabili Ricordiamo la definizione probabilistica di indipendenza

Prob(A|B)=Prob(A)

Nel nostro esempio l'ipotesi di indipendenza implica

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

...

Ricordiamo che l'ipotesi nulla riguarda la popolazione e non il campione osservato

Al fine di costruire la nostra statistica test dobbiamo stimare le probablità coinvolte nell' ipotesi nulla

Come spesso accade in statistica per stimare delle probabilità utilizziamo le corrisponenti frequenze relative

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

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

Ciò implica che le due distribuzioni condizionate ai diversi trattamenti saranno uguali tra loro

$$\frac{n_{11}}{n_{1.}} = \frac{n_{21}}{n_{2.}}$$

In generale 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

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

Nel nostro caso

	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

Nel caso di indipendenza

	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

Quanto sono "distanti" i dati osservati dalla situazione di indipendenza (date le marginali)?

Il χ^2 (Pearson, 1900)

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

Il χ^2 vale 0 nel caso di indipendenza ma non ha un massimo univoco e dipende dalla numerosità del nostro collettivo

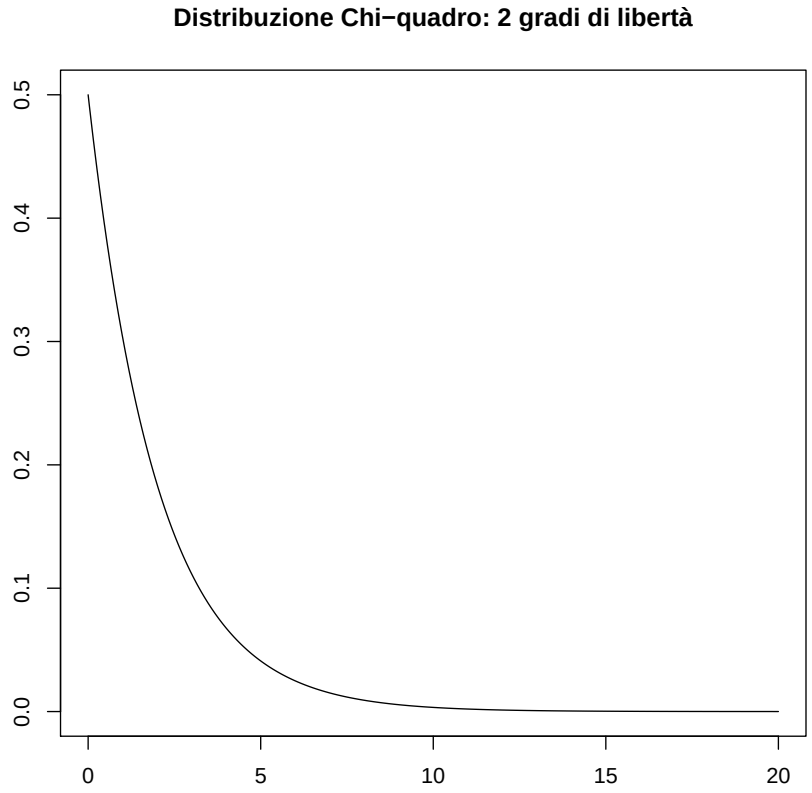
La "connessione" massima (per tabelle quadrate) si verifica quando ad ogni modalità di un carattere corrisponde una ed una sola modalità dell'altro

Allora $max\chi^2 = nmin\,[h-1, k-1]$

Passiamo all'indice relativo $\chi^2/max\chi^2$ oppure usiamo la statistica χ^2 come statistica test.

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

Nel nostro esempio abbiamo ottenuto $\chi^2 = 26.9$ che sotto la distribuzione χ^2 con 2 gradi di libertà corrisponde ad un p-value praticamente nullo



n.b.: il valore 26.9 non misura la forza del legame tra trattamento ed esito finale ma piuttosto l'evidenza fornita dai dati a favore dell'ipotesi di dipendenza

Focalizziamo la nostra attenzione sul rischio di un evento cardiaco fatale.

	Attacco cardiaco fatale	Nessun attacco cardiaco ...	Totale
Placebo	18	11016	11034
Aspirina	5	11032	11037
Totale	23	22048	22071

Per valutare l'entità dell'effetto dell'aspirina come trattamento di prevenzione indichiamo con $\pi_{AF|A}$ e $\pi_{AF|P}$ le probabilità di avere un attacco cardiaco fatale se sottoposti a terapia preventiva rispettivamente a base di aspirina e beta-carotene.

Stimiamo il rapporto

$$RR = \frac{\pi_{AF|A}}{\pi_{AF|P}}$$

approssimazione del rischio relativo di un attacco cardiaco fatale, utilizzando le corrispondenti frequenze relative

$$\widehat{RR} = \frac{n_{21}/n_{2.}}{n_{11}/n_{1.}} = \frac{5/11037}{18/11034} = 0.31$$

Oppure stimiamo l'*odds ratio*

$$OR = \frac{\pi_{AF|A}/1 - \pi_{AF|A}}{\pi_{AF|P}/1 - \pi_{AF|P}}$$

ancora una volta sulla base delle frequenze relative osservate

$$\begin{aligned} \widehat{OR} &= \frac{n_{21}/n_{2.}/1 - n_{21}/n_{2.}}{n_{11}/n_{1.}/1 - n_{11}/n_{1.}} \\ &= \frac{n_{21} \times n_{12}}{n_{12} \times n_{22}} \\ &= \frac{511016}{1811032} = 0.31 \end{aligned}$$

L'odds di un attacco cardiaco fatale si riduce del 69% utilizzando l'aspirina o, analogamente, è circa 3 volte ($1/0.31 = 3.23$) più alto per coloro che hanno assunto beta-carotene rispetto a coloro che hanno assunto aspirina

Poichè la probabilità di un attacco cardiaco fatale è prossima a zero, le due quantità RR e OR sono molto simili tra loro La situazione di indipendenza corrisponde a OR=1

È utile a volte esprimere l'odds ratio su scala logaritmica:

$$OR = 1 \Rightarrow ln(OR) = 0$$

$$OR = 0.31 \Rightarrow ln(OR) = -1.17$$

$$OR = 3.23 \Rightarrow ln(OR) = 1.17$$

Poichè la nostra è in realtà una tabella 2×3, possiamo descrivere l'associazione tra terapia ed esito calcolando due odds ratio locali che utilizzano le 2 parti separate di informazione di cui disponiamo. I 2 odds ratio corrispondono ai due gradi di libertà del test χ^2

$$OR_1 = \frac{\pi_{AF|A}/\pi_{NA|A}}{\pi_{AF|P}/\pi_{NA|P}}$$

$$\hat{OR}_1 = \frac{n_{21} \times n_{13}}{n_{11} \times n_{23}} = \frac{5 \times 10845}{18 \times 10933} = 0.27$$

$$OR_2 = \frac{\pi_{ANF|A}/\pi_{NA|A}}{\pi_{ANF|P}/\pi_{NA|P}}$$

$$\hat{OR}_2 = \frac{n_{22} \times n_{13}}{n_{12} \times n_{23}} = \frac{99 \times 10845}{171 \times 10933} = 0.57$$

Esiste in realtà un terzo odds ratio che mette a confronto la probabilità di un attacco fatale con quella di un attacco non fatale

$$OR_3 = \frac{\pi_{AF|A}/\pi_{ANF|A}}{\pi_{AF|P}/\pi_{ANF|P}}$$

$$\hat{OR}_3 = \frac{n_{21} \times n_{12}}{n_{11} \times n_{21}} = \frac{5171}{1899} = 0.48$$

ma $OR_1 = OR_2 \times OR_3$

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

- Dimostrare l'esistenza di una associazione significativa

- Collassare la tabella considerando soltanto due livelli per il numero medio giornaliero di sigarette (minore di 5, almeno 5) e calcolare il relativo odds-ratio

- Descrivere la natura dell'associazione calcolando gli odds ratio per ciascuno dei 6 livelli. Esiste un chiaro trend?