

Test di significatività

I dati statistici sono affetti dal caso — il caso, anzi, viene impiegato di proposito, per esempio negli studi clinici randomizzati. Servono quindi metodi per distinguere i fenomeni realmente rilevabili nei dati, dagli effetti del caso. Questi metodi si chiamano *test di significatività* o *test di ipotesi*. Il risultato di un test di significatività è un *livello di significatività* — *p-value* — che misura quanto sia verosimile attribuire i dati osservati alla mera azione del caso: *più piccolo* è il *p-val.* *maggiore* è la *significatività statistica* della rilevazione del fenomeno corrispondente. Abbiamo già visto, studiando la regressione lineare, che, nelle scienze della vita, un *p-value* < 0.05 è frequentemente richiesto per considerare il risultato *statisticamente significativo*, e un *p-value* < 0.01 viene considerato *fortemente significativo*.

Nota. È fondamentale ricordare che i *p-value*

misurano il grado di evidenza statistica che abbiamo a favore di una certa ipotesi, e *non* l'intensità dell'effetto ipotizzato, *né* il rilievo scientifico del risultato.

Consideriamo la seguente tabella, che mostra i risultati di uno

Patients	No. of Attacks of Pain	
	Pronethalol	Placebo
1	7	14
2	16	23
3	29	71
4	25	34
5	15	17
6	1	8
7	0	2
8	2	7
9	0	3
10	348	323
11	65	79
12	41	55

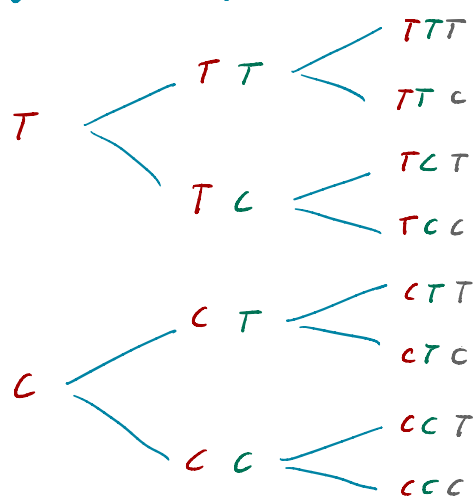
studio clinico sullo uso del pronotalolo per il trattamento dell'angina pectoris (Prichard et al. '63). In questo studio, dodici pazienti sono

stati trattati, a periodi alterni, con pronotalolo o con un placebo. (Nota: le durate dei periodi di trattamento coincidono e la loro disposizione nel tempo è randomizzata.) Per ciascun paziente si riporta il numero di attacchi subiti durante ciascuno dei due trattamenti (pronotalolo e placebo). Si vede bene che 11 dei 12 pazienti hanno avuto meno attacchi mentre si trovavano sotto l'azione del farmaco: possiamo quindi concludere

che questo è efficace? Se il farmaco non facesse nulla, se fosse equivalente a un placebo, ci aspetteremo, nel confronto col placebo, di vedere un miglioramento grosso modo nella metà dei casi, e nell'altra metà un peggioramento: insomma la distribuzione di miglioramenti e peggioramenti, se il farmaco non avesse effetto, sarebbe statisticamente equivalente al lancio di una moneta. Ci chiediamo quindi: dal mero lancio di 12 monete, quanto sarebbe probabile ottenere 11 o più risultati concordi, come è avvenuto nell'esperimento?

Dodici monete possono dare luogo a $2^{12} = 4096$ diverse sequenze di teste e croci:

una moneta due monete tre monete ...



Ogni nuova moneta duplica il numero delle possibili sequenze. Con 12 monete le sequenze possibili sono quindi

$$\underbrace{2 \cdot 2 \cdot 2 \cdot \dots \cdot 2}_{12 \text{ duplicazioni}} = 2^{12} = 4096$$

Quante sequenze presentano 11 o più risultati concordi? Abbiamo due sequenze con 12 risultati concordi: $TTTT...T$ e $CCCC...C$. Abbiamo poi 12 sequenze con 11 teste: $CTTT...T$, $TCTT...T$, $TTCT...T$, ... (sono 12 perché l'unica C può stare in posizione 1, 2, 3 ... 12: che sono 12 possibilità). Infine, allo stesso modo, abbiamo 12 sequenze con 11 croci: gran totale $2+12+12=26$. La probabilità di ottenere 11 o più risultati concordi da 12 monete è quindi:

$$p = \frac{26}{4096} \approx 0.0063$$

Questo numero è il p-value prodotto dal test che abbiamo appena eseguito. Siccome $p < 0.01$ sappiamo già che i risultati dell'esperimento mostrano una forte evidenza statistica a favore dell'efficacia del pronetaleto. Ma perché questo? Quale è il ragionamento? Noi stiamo dicendo che se il farmaco non avesse effetto, ALLORA l'esperimento sarebbe come il lancio ripetuto di 12 monete, quindi un risultato sbilanciato come quello che abbiamo osservato, o più, si verificherebbe solo 26 volte su 4096 o nello 0.63% dei casi. Di conseguenza $\bullet$ il farmaco

ha effetto, oppure si è verificata una coincidenza molto rara: meno di un caso su 100. Più rara è la coincidenza — ossia minore è il p -value — più siamo spinti a credere che questa non si sia, in realtà, verificata, la qual cosa equivale a dire che crediamo che il farmaco abbia effetto.

Nota: Prichard et al. dichiarano un p -value, per

Of the 12 patients, 11 had fewer attacks on pronethalol than on placebo. This event in itself is statistically significant ($P=0.003$) and has the advantage over other analyses of being independent of any assumptions.

questo test, di **0.003**,
ossia la metà del nostro.

Chiaramente loro hanno calcolato la probabilità di ottenere 11 o più risultati positivi (miglioramenti) e non concordi (che potrebbero essere positivi o negativi). Il nostro si chiama test a due code, il loro a una coda: il primo considera risultati pari o più estremi di quelli dell'esperimento, il secondo risultati pari o più positivi. I test a due code assegnano una minore significatività statistica (p -value più alto), sono quindi più conservativi. Generalmente, la scelta più conservativa, ossia quella più cauta nell'asserire la significatività dei risultati, è ritenuta preferibile (ma il lavoro citato è del '63, erano altri tempi...).

Ricapitolando, un test di significatività oppone due ipotesi

Ipotesi Nulla: "i dati sono unicamente affetti dal caso"

Ipotesi Alternativa: "i dati esibiscono un fenomeno che non si può spiegare con la mera casualità"

Se stiamo valutando un trattamento, l'I.A. equivale quindi a dire "il trattamento produce un effetto".

Un test di significatività (o di ipotesi) può solo rigettare l'ipotesi nulla, e anche questo solamente a un certo livello di significatività, $p < \alpha$. Soglie comunemente impiegate sono $\alpha = 0.05$ e $\alpha = 0.01$. Più bassa è, numericamente, la soglia, più forte la significatività statistica con la quale rigettiamo l'ipotesi nulla (se effettivamente $p < \alpha$).

In termini precisi dire che un test rifiuta l'I.N. al livello di significatività $p < \alpha$ equivale a dire che SE l'I.N. fosse vera Allora la probabilità che il test rigetti l'I.N. stessa — "falso positivo" o "errore del primo tipo" — sarebbe α .

Nota: Fissare la soglia α determina solo in modo indiretto

le probabilità di ottenere un risultato positivo o negativo nel caso in cui sia l'I.A. ad essere vera. Va da sé che più bassa è la soglia, più il test è severo, maggiore sarà quindi la probabilità, se è vera l'I.A., di un falso negativo, o errore del secondo tipo. Questa probabilità, solitamente denotata con β , non dipende solo da α , e, in questo corso non vedremo come si possa calcolare. Il numero $1-\beta$, che è la probabilità di rigettare la I.M. se questa è falsa, si chiama potenza del test.

Nota: un test non può rigettare l'ipotesi alternativa. Ossia, un risultato di non significatività ci dice che non siamo in grado di rigettare l'ipotesi nulla e non che abbiamo dimostrato l'inefficacia del trattamento.

Nota: nessun livello di significatività statistica equivale alla certezza. Anche una coincidenza che capita solo una volta su un milione, quella volta, capita. E nulla vieta che quella volta sia proprio il primo tentativo. L'unico modo di avvicinarsi alla certezza, in statistica, è ripetendo gli esperimenti (meglio se ad opera di ricercatori diversi, con metodi

diversi, etc.) Nessun esperimento, nessun risultato, in questo senso, può dirsi definitivo.

Per estrarre il p -value 0.0063 dai dati di Pritchard et al. abbiamo impiegato un ragionamento ad hoc. In generale, non è necessario, per ogni singolo caso, fare una piccola invenzione; ci sono dei metodi standard che coprono tutti i casi più frequenti. Nel seguito vedremo due di questi metodi:

- il test t di Student per saggiare le medie di variabili statistiche numeriche
- il test χ^2 per saggiare la distribuzione di variabili statistiche categoriche

Test t di Student

Supponiamo di estrarre un campione di n individui e consideriamo i valori $x_1, \dots, x_n$ assunti dalla variabile statistica X su questi individui.

Intendiamo saggiare l'ipotesi che la media di X (su tutta la popolazione) sia un certo valore μ_0 .

$$IN: \mu_X = \mu_0$$

$$IA: \mu_X \neq \mu_0$$

Perché questo test si applichi è necessario che la distribuzione di X sia grosso modo gaussiana o il numero n degli individui grande.

Calcoliamo

$$t = \frac{|\hat{M} - \mu_0|}{\frac{s}{\sqrt{n}}}$$

media campionaria (pointing to $\hat{M}$)
varianza campionaria (under s^2)
s (under s)
 $\sqrt{n}$ (under $\sqrt{n}$)

se fosse vera l'ipotesi nulla, questa quantità sarebbe distribuita come una t di Student con $n-1$ gradi di libertà. Quindi:

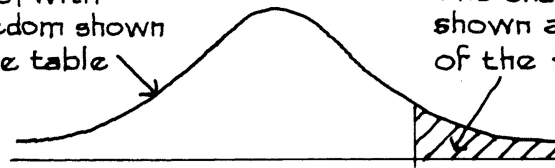
Il test rigetta l'IN con p -value $p < \alpha$ se $t > t_{\alpha/2, n-1}$

Dove la soglia $t_{\alpha/2, n-1}$ si può ricavare da

una tabella delle aree sotto il grafico della t di Student come segue.

Student's curve, with degrees of freedom shown at the left of the table

The shaded area is shown along the top of the table



is shown in the body of the table

$\frac{\alpha}{2}$

Degrees of freedom	25%	10%	5%	2.5%	1%	0.5%
$n-1$ { 1	1.00	3.08	6.31	12.71	31.82	63.66
2	0.82	1.89	2.92	4.30	6.96	9.92
3	0.76	1.64	2.35	3.18	4.54	5.84
4	0.74	1.53	2.13	2.78	3.75	4.60
5	0.73	1.48	2.02	2.57	3.36	4.03

$p < 0.05$ $p < 0.01$

Per esempio, con un campione di 5 individui, perché il p -value sia < 0.01 , la soglia è

$$t_{0.005, 4} \approx 4.60$$

Nota: per valori grandi di n , la distribuzione t di Student è bene approssimata da una gaussiana, per cui non è, solitamente, tabulata per più di ~30 gradi di libertà. In questi casi, le soglie possono essere desunte da una tabella della gaussiana.

Esempio

Ripetiamo l'analisi dei dati di Prichard et al. sull'uso del pronetalolo per l'angina pectoris. Il lavoro ci dice sinteticamente cosa è stato fatto.

The figures have also been analysed using a standard "analysis of variance" technique, after transforming the variable to $\log(\text{number of attacks} + 1)$. This transformation makes the distribution more nearly symmetrical, and the variability more nearly equal in the different groups, as required for the analysis. The difference between the placebo and pronethalol figures is found to be significant at the 1% level. There is no evidence in this small series that the drug is any better for some patients than for others.

Si capisce che, a partire da questi dati:

Patients	No. of Attacks of Pain	
	Pronethalol	Placebo
1	7	14
2	16	23
3	29	71
4	25	34
5	15	17
6	1	8
7	0	2
8	2	7
9	0	3
10	348	323
11	65	79
12	41	55

È stata considerata la variabile statistica

$$X = \log(\text{Placebo} + 1) - \log(\text{Pronethalol} + 1)$$

L'ipotesi nulla che vorremmo rigettare è che il

farmaco non abbia effetto. Se questo avvenisse, le due quantità sottotutte sarebbero, in media, uguali, quindi avremmo

Ipotesi Nulla: $\mu_X = 0$ ← Questo è μ_0

Il lavoro asserisce di poter rigettare questa ipotesi con un livello di significatività $p < 0.01$, quindi, per confermare i calcoli dobbiamo verificare che

$$t = \frac{|\hat{\mu}|}{s/\sqrt{12}} > t_{0.005, 11} \approx 3.11$$

$\hat{\mu} - \mu_0 = \hat{\mu} - 0 = \hat{\mu}$

Dove $\hat{\mu}$ è la media di X e s^2 la varianza campionaria. Prichard et al. non ci dice che base sia stata usata per i logaritmi: usiamo e .

	pronethalol	placebo	X	
1	7	14	0.62860866	$= x_1 = \log_e(14+1) - \log_e(7+1)$
2	16	23	0.34484049	$= x_2 = \log_e(23+1) - \log_e(16+1)$
3	29	71	0.87546874	
4	25	34	0.29725152	
5	15	17	0.11778304	
6	1	8	1.50407740	
7	0	2	1.09861229	
8	2	7	0.98082925	
9	0	3	1.38629436	
10	348	323	-0.07432841	
11	65	79	0.19237189	
12	41	55	0.28768207	$= x_{12} = \log_e(55+1) - \log_e(41+1)$

etc.

Abbiamo

$$\hat{\mu} = \frac{x_1 + x_2 + \dots + x_{12}}{12} \approx 0.6366243$$

$$s = \sqrt{\frac{(x_1 - \hat{\mu})^2 + \dots + (x_{12} - \hat{\mu})^2}{11}} \approx 0.5220226$$

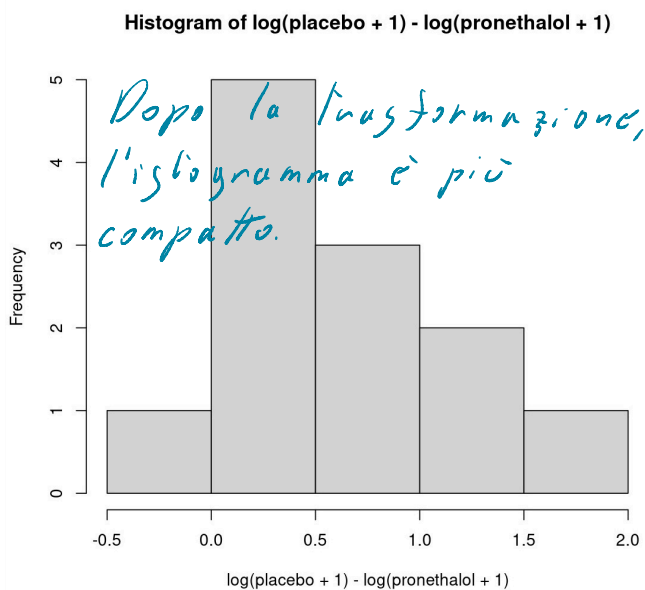
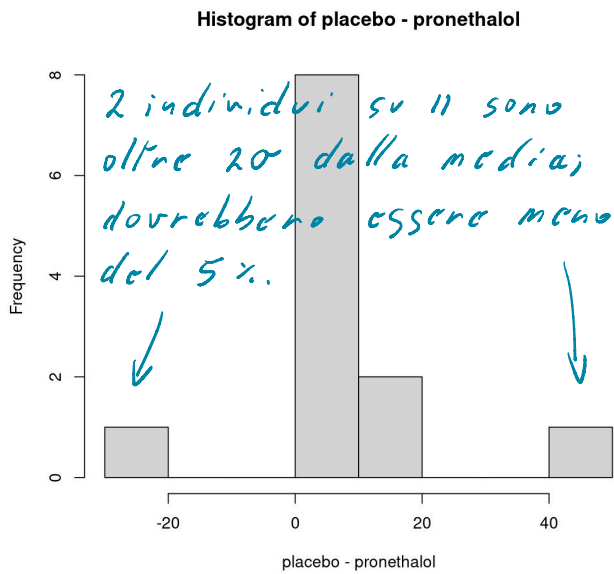
Per cui

$$t = \frac{|\hat{\mu}|}{s/\sqrt{12}} \approx 4.22$$

Siccome $4.22 > t_{0.005, 11} \approx 3.11$, possiamo rigettare l'ipotesi nulla con un livello di significatività statistica dell'1%. (p-value < 0.01), a favore dell'ipotesi alternativa, ossia dell'efficacia del farmaco.

Merita considerazione la scelta di applicare una *trasformazione logaritmica*. A prima vista potrebbe sembrare un tentativo di adulterare i dati, però, in realtà, si tratta di una tecnica standard. Cerca di risolvere il problema che alcune distribuzioni che si incontrano in pratica hanno *coda lunghe*, ossia una proporzione eccessiva di individui molto lontani dalla media, rispetto a quanto *sarebbe*

previsto da una distribuzione gaussiana. Per



esempio, se avessimo usato placebo-pronethalol avremmo incontrato la distribuzione qui accanto. Il guaio è che molti test derivano dalla studio della distribuzione gaussiana, quindi, a distribuzioni del genere, si applicano male. Più profondamente, se noi usassimo placebo-pronethalol, i due individui che si trovano molto lontani dalla media avrebbero un'influenza preponderante sul

risultato del test, e questo non è desiderabile, perché, in ultima analisi, ci interessa sapere se il farmaco sia utile per la generalità dei pazienti, e non se il singolo caso più fortunato di tutti bilanci o no il singolo caso più sfortunato di tutti.

Test t per il confronto di variabili appaiate

Dall'esempio precedente, abbiamo appreso che il test t si può impiegare anche per saggiare l'uguaglianza fra le medie di due variabili statistiche appaiate. Ricapitolando, questa è la situazione. Abbiamo un campione di n individui e due variabili numeriche A e B ; abbiamo quindi i numeri $A_1, A_2, \dots, A_n$ e $B_1, B_2, \dots, B_n$. Nell'esempio abbiamo considerato $A = \log(\text{placobo} + 1)$ e $B = \log(\text{pronetatolo} + 1)$. Vogliamo saggiare l'ipotesi nulla

$$H_0: \mu(A) = \mu(B)$$

Per farlo calcoliamo la variabile $X = A - B$, per cui

$$X_1 = A_1 - B_1, \quad X_2 = A_2 - B_2, \quad \dots, \quad X_n = A_n - B_n$$

Ora possiamo rigettare l'ipotesi nulla col livello di significatività statistica $p < \alpha$ se

$$t = \frac{|\hat{\mu}_x|}{s_x / \sqrt{n}} > t_{\alpha/2, n-1}$$

Test t per il confronto di campioni indipendenti

Negli studi clinici avviene spesso di avere due campioni indipendenti: un **gruppo di trattamento** e un **gruppo di controllo**. Una variante del test t permette di confrontare la media di due variabili statistiche sotto sull'uno e sull'altro gruppo, per saggiare la ipotesi che siano diverse. Questa è la situazione: abbiamo due campioni, di m e n individui rispettivamente. Una certa variabile statistica X assume sul primo i valori $X_1, \dots, X_m$, mentre una seconda variabile Y assume sul secondo $Y_1, \dots, Y_n$. Possiamo rigettare l'ipotesi nulla

$$H_0 : \mu_X = \mu_Y$$

con livello di confidenza $p < \alpha$ se

$$t = \frac{|\bar{M}_X - \bar{M}_Y|}{\sqrt{\frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2} \left(\frac{1}{m} + \frac{1}{n} \right)}} > t_{\alpha/2, m+n-2}$$

Test χ^2 ← "chi quadro"

Consideriamo una variabile categorica X che può assumere m valori diversi $x_1, x_2, \dots, x_m$. Per esempio il sesso è una variabile di questo tipo: in questo caso $m=2$, $x_1 = \text{"femmina"}$, $x_2 = \text{"maschio"}$. Un altro esempio potrebbe essere lo stato di occupazione, con $m=3$ e i tre valori "disoccupato", "occupato" e "in pensione". Dati m numeri $f_1, f_2, \dots, f_m$, possiamo chiederci se questi siano le frequenze relative dei valori $x_1, \dots, x_m$ rispettivamente, su tutta la popolazione. Estraiamo un campione casuale di n individui, e indichiamo con $n_1, n_2, \dots, n_m$ le frequenze degli m valori nel campione.

Promemoria: Frequenza = numero di individui

$$\text{Frequenza relativa} = \frac{\text{numero di individui}}{\text{totale}}$$

Per cui $n_1 + n_2 + \dots + n_m = n$ e $f_1 + f_2 + \dots + f_m = 1$

Chiaramente, se le frequenze relative su tutta la popolazione fossero davvero $f_1, \dots, f_m$, sul nostro campione ci aspetteremmo le frequenze teoriche $\bar{n}_1 = n \cdot f_1$, $\bar{n}_2 = n \cdot f_2, \dots$. Calcoliamo la quantità

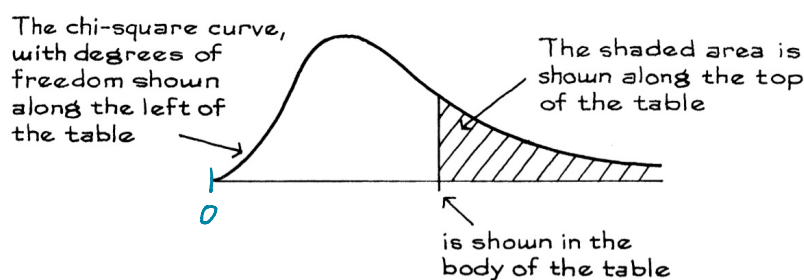
$$\chi^2 = \frac{(n_1 - \bar{n}_1)^2}{\bar{n}_1} + \frac{(n_2 - \bar{n}_2)^2}{\bar{n}_2} + \dots + \frac{(n_m - \bar{n}_m)^2}{\bar{n}_m}$$

Questo numero, chiaramente, misura la distanza fra le frequenze osservate $n_1, \dots, n_m$ e le frequenze teoriche $\bar{n}_1, \dots, \bar{n}_m$. Il test χ^2 rigetta l'ipotesi nulla

I.N.: Le frequenze relative di $x_1, \dots, x_m$ sono $f_1, \dots, f_m$

se $\chi^2 > \chi^2_{\alpha, m-1}$. Le soglie $\chi^2_{\alpha, m-1}$ sono tabulate, similmente a come accade per la t di Student.

A CHI-SQUARE TABLE



Degrees of freedom	99%	95%	90%	70%	50%	30%	10%	5%	1%
$m-1$ { 1	0.00016	0.0039	0.016	0.15	0.46	1.07	2.71	3.84	6.64
2	0.020	0.10	0.21	0.71	1.39	2.41	4.60	5.99	9.21
3	0.12	0.35	0.58	1.42	2.37	3.67	6.25	7.82	11.34
4	0.30	0.71	1.06	2.20	3.36	4.88	7.78	9.49	13.28
5	0.55	1.14	1.61	3.00	4.35	6.06	9.24	11.07	15.09

Nota: Il test χ^2 è generalmente eseguito a una coda. Infatti abbiamo osservato che il valore di χ^2 misura la distanza fra le frequenze osservate e quelle teoriche, ossia previste dall'ipotesi nulla. È chiaro quindi che solo un valore grosso di χ^2 fornisce una ragione per rigettare quest'ultima.

Affinché il test χ^2 produca risultati validi è generalmente richiesto che sia soddisfatta la seguente condizione dovuta a Cochran.

Almeno l'80% delle frequenze teoriche deve essere > 5 e tutte le frequenze teoriche devono essere > 1 .

Se questa condizione non è soddisfatta, allora la tecnica non si applica. In questa situazione è possibile ricondursi ad un caso trattabile combinando alcuni dei valori fra loro. Per esempio, se abbiamo una variabile "classe di reddito" suddivisa in dieci categorie: I, II, ..., X, possiamo accorpate queste categorie a due a due ottenendo una nuova variabile che può assumere i cinque valori I II, II III, III IV, etc. Va da sé che, sommandosi le frequenze, sarà più facile che la condizione di Cochran sia soddisfatta.

Esempio

Introducendo i test di significatività abbiamo analizzato i dati di Prichard et al. osservando

Patients	No. of Attacks of Pain		X
	Pronethalol	Placebo	
1	7	14	+
2	16	23	+
3	29	71	+
4	25	34	+
5	15	17	+
6	1	8	+
7	0	2	+
8	2	7	+
9	0	3	+
10	348	323	-
11	65	79	-
12	41	55	-

che 11 dei 12 pazienti sono apparentemente migliorati durante il trattamento. Se questo non avesse effetto, la miglioramenti e peggioramenti si verifi-

cherebbero con uguale frequenza (50% e 50%). Con un calcolo ad hoc abbiamo sfruttato questo fatto per determinare il p-value 0.0063 : la probabilità che, lanciate 12 monete, vengano 11 o più risultati uguali. Ripetiamo la medesima analisi usando il test χ^2 .

Consideriamo la variabile categorica X che associa ad ogni paziente il valore $x_1 = "+"$ se questi è migliorato e $x_2 = "-"$ se è peggiorato. Quindi $m = 2$. L'ipotesi nulla è che miglioramenti e peggioramenti si verifichino con uguale frequenza

I.N.: Le frequenze relative di "+" e "-" sono $f_1 = 0.5$ e $f_2 = 0.5$.

Il numero dei pazienti è $n=12$, per cui le frequenze teoriche sono

$$\bar{n}_1 = n f_1 = 6 \quad \bar{n}_2 = n f_2 = 6$$

e queste vanno confrontate con le frequenze osservate $n_1=11$ e $n_2=1$. Quindi

$$\chi^2 = \frac{(11-6)^2}{6} + \frac{(1-6)^2}{6} \approx 8.33$$

Se ci interessa un p-value < 0.01 la soglia con cui confrontare questo numero è $\chi^2_{0.01,1} = 6.64$.

Siccome $8.33 > 6.64$ possiamo rigettare l'ipotesi nulla con p-value < 0.01 .

Avendo a disposizione un computer, possiamo anche calcolare esattamente a che p-value corrisponda $\chi^2 = 8.33$: il risultato è $p = 0.0033$. Il valore corretto, però, è quello che abbiamo già calcolato:

$p = 0.0063$. La differenza si spiega per una combinazione delle approssimazioni implicite nel test e il fatto che le frequenze teoriche sono appena accettabili per il criterio di Cochran.

Test χ^2 di indipendenza

Un test χ^2 può essere anche impiegato per valutare la correlazione fra due variabili categoriche. Per esempio, in uno studio clinico, si vuole capire se trattamento/controllo e l'esito della cura siano variabili, almeno in parte, correlati.

Consideriamo questa situazione:

abbiamo due variabili categoriche X e Y , che possono assumere i valori $x_1, \dots, x_a$ e $y_1, \dots, y_b$ rispettivamente. Estraiamo un campione di n individui e indichiamo con n_{ij}^i , al variare dei due indici $i=1, \dots, a$ e $j=1, \dots, b$, le frequenze delle coppie x_i, y_j . Con n_i^i indichiamo la frequenza di x_i e con n_j^j la frequenza di y_j . Sotto forma di tabella a due vie possiamo rappresentare la situazione.

	x_1	x_2	x_3	...	x_a	Totali righe
y_1	n_{11}^1	n_{12}^2	n_{13}^3		n_{1a}^a	$n_{1\cdot}$
y_2	n_{21}^1	n_{22}^2	n_{23}^3		n_{2a}^a	$n_{2\cdot}$
$\vdots$						
y_b	n_{b1}^1	n_{b2}^2	n_{b3}^3		n_{ba}^a	$n_{b\cdot}$
Totali colonne	$n_{\cdot 1}^1$	$n_{\cdot 2}^2$	$n_{\cdot 3}^3$		$n_{\cdot a}^a$	n

Vogliamo saggiare l'ipotesi nulla

I.N.: Le variabili X e Y sono indipendenti:
(ossia non c'è correlazione fra X e Y)

Per farlo, calcoliamo la frequenza teorica

$$\bar{n}_j^i = \frac{n^i \cdot n_j}{n} \quad \text{per es.} \quad \bar{n}_2^2 = \frac{n^2 \cdot n_2}{n}$$

Nota: il ragionamento è che, se le variabili fossero indipendenti la proporzione di y_2 nella seconda colonna della tabella sarebbe uguale alla proporzione di y_2 sul totale quindi

n. di y_2 nella col. x_2 : tot. col. x_2 = n. di y_2 totali : n. individui
teorico (se vale I.N.)

$$\begin{array}{ccccccc} \parallel & & \parallel & & \parallel & & \parallel \\ \bar{n}_2^2 & : & n^2 & = & n_2 & : & n \end{array}$$

da cui, risolvendo la proporzione, $\bar{n}_2^2 = \frac{n^2 \cdot n_2}{n}$

Ora calcoliamo χ^2 di tutte le frequenze n_j^i e le corrispondenti frequenze teoriche $\bar{n}_j^i$.

$$\chi^2 = \frac{(n_1^1 - \bar{n}_1^1)^2}{\bar{n}_1^1} + \frac{(n_2^1 - \bar{n}_2^1)^2}{\bar{n}_2^1} + \dots + \frac{(n_5^3 - \bar{n}_5^3)^2}{\bar{n}_5^3} + \dots + \frac{(n_6^a - \bar{n}_6^a)^2}{\bar{n}_6^a}$$

Rigettiamo l'I.N. con $p < \alpha$ se $\chi^2 > \chi_{\alpha, (a-1) \cdot (b-1)}^2$

Esempio

La tabella seguente proviene dallo studio sulle efficacia della streptomicina che abbiamo già esaminato.

TABLE II.—Assessment of Radiological Appearance at Six Months as Compared with Appearance on Admission

Radiological Assessment	Streptomycin Group x_1		Control Group x_2	
x_1 Considerable improvement ..	28	51%	4	8%
x_2 Moderate or slight improvement	10	18%	13	25%
x_3 No material change	2	4%	3	6%
x_4 Moderate or slight deterioration	5	9%	12	23%
x_5 Considerable deterioration ..	6	11%	6	11%
x_6 Deaths	4	7%	14	27%
Total	55	100%	52	100%

Vogliamo usare un test χ^2 per saggiare l'ipotesi
 Ipotesi Nulla: Non c'è correlazione fra il trattamento
 con streptomicina e l'esito a sei mesi.

Chiamiamo X la variabile trattamento che può assumere due valori: x_1 = "streptomicina" e x_2 = "controllo". Chiamiamo Y la variabile esito, e denotiamo con $y_1 \dots y_6$ i sei possibili esiti considerati. Quindi a = numero degli x_i = 2 e b = numero dei y_j = 6. La tabella a due vie è come qui accanto.

	x_1	x_2	tot.
y_1	28	4	32
y_2	10	13	23
y_3	2	3	5
y_4	5	12	17
y_5	6	6	12
y_6	4	14	18
tot.	55	52	107

Da questa possiamo dedurre la tabella delle frequenze teoriche. $\rightarrow$

	x_1	x_2
y_1	16.45	15.55
y_2	11.82	11.18
y_3	2.57	2.43
y_4	8.74	8.26
y_5	6.17	5.83
y_6	9.25	8.75

Per esempio questo valore 8.26 si ottiene calcolando

$$\frac{n^2 \cdot n_4}{n} = \frac{52 \cdot 17}{107} \approx 8.26$$

Ora calcoliamo

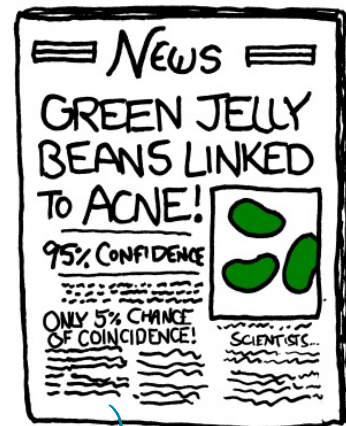
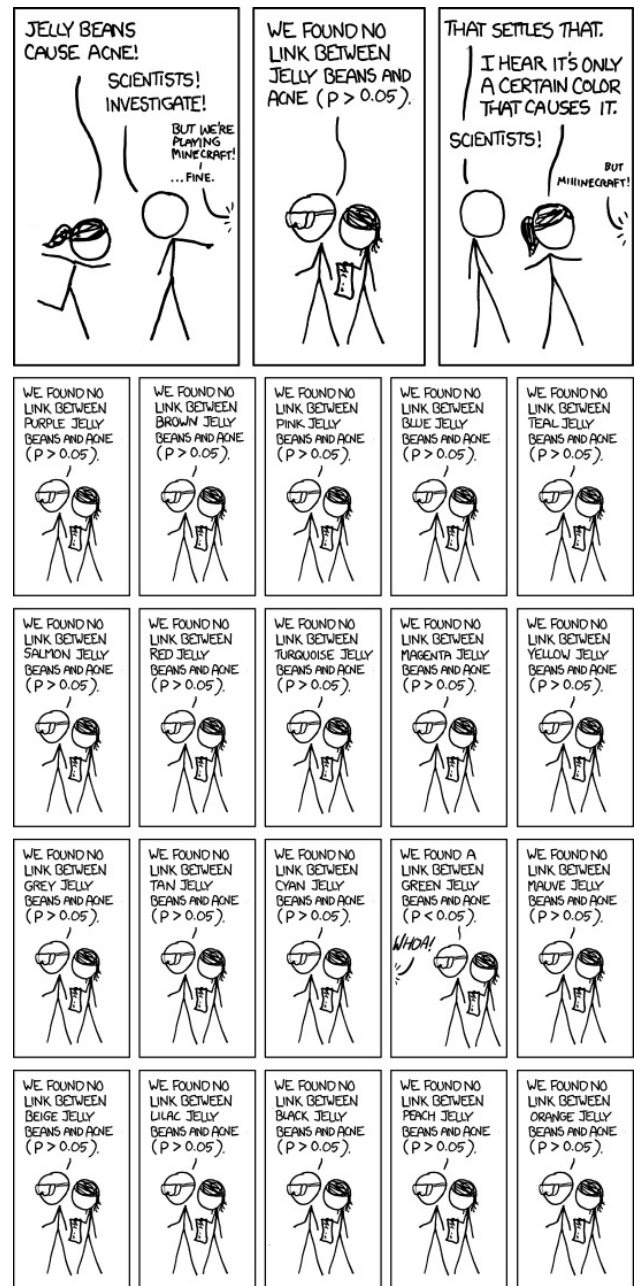
$$\chi^2 = \frac{(28 - 16.45)^2}{16.45} + \frac{(4 - 15.55)^2}{15.55} + \frac{(10 - 11.82)^2}{11.82} + \dots \approx 26.97$$

Siccome $26.97 > \chi^2_{0.01, 5} \approx 15.09$ possiamo rigettare l'ipotesi nulla con livello di significatività $p < 0.01$, a favore dell'ipotesi alternativa, ossia dell'efficacia della streptomicina.

Usando un computer possiamo calcolare che p-value corrisponde a $\chi^2 = 26.97$: si ottiene $5.8 \cdot 10^{-5}$. Questo valore però non è affidabile, perché le approssimazioni implicite nel test falliscono a questo livello di dettaglio. Il valore esatto è, in realtà $1.8 \cdot 10^{-5}$. Cioè, la determinazione p-value < 0.01 rimane valida.

La correzione di Bonferroni

Questo fumetto illustra una ingidia implicita nell'uso di test di significatività statistica: fissata una soglia $p < \alpha$, se l'ipotesi nulla è vera, c'è una probabilità α di ottenere un falso positivo (ossia di rigettare l'I.N. benché sia vera). Se, per esempio, consideriamo la soglia $p < 0.05$, una volta su 20 (in media) rigatteremo un'I.N. vera. Di conseguenza, se uno studio richiede di eseguire n test, e una determinazione positiva seguirebbe dalla riuscita di ciascuno di essi, allora per ogni singolo test è opportuno usare la soglia $p < \frac{\alpha}{n}$ (correzione di Bonferroni)



Da xkcd.com

Riepilogo

Se vogliamo saggiare

Usiamo

la media di una v. numerica

il test t di Student

la distribuzione di una v. categorica

il test χ^2

Se vogliamo saggiare l'indipendenza
(assenza di correlazione) di

Usiamo

due variabili numeriche

la regressione lineare⁺

una v. numerica e una v. binaria

il test t di Student*

(una v. numerica e una v. categorica

ANOVA)^{← non in programma}

due v. categoriche

il test χ^2 di indep.

⁺ se sospettiamo la presenza
di una relazione di tipo
lineare ($Y = aX + b$). Ci
sono tecniche analoghe
per relazioni diverge.

^{*} per campioni indipendenti
o appaiati secondo il caso.