X^2

È un test utile a determinare quanto due variabili nominali siano correlate fra loro in un certo dataset.
Considerando un dataset con due attributi A e B, dove A e B hanno rispettivamente c e r tipi di valori diversi

$$A = \{a_1 \dots a_c\}$$
$$B = \{b_1 \dots b_r\}$$

Se rappresentiamo tutti i valori nella matrice di contingenza:

	b_1	...	b_r	
a_1	$o_{1,1}$	...	$o_{1,c}$	o_1
⋮	⋮		⋮	
a_c	$o_{r,1}$	...	$o_{r,c}$	o_r
	o_1		o_c	TOT

Dove $o_{i,j}$ indica quanti record nel dataset hanno come attributo $A = a_i$ e $B = b_j$.

Esempio:

	NON FUMATORI	FUMATORI	
MASCHIO	200	100	300
FEMMINA	150	50	200
	350	150	500

FORMULA X^2

$$X^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

Che possiamo anche riscrivere come:

$$X^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(O_{ij} - e_{ij})^2}{e_{ij}}$$

Dove i valori O_{ij} vengono presi dalla tab di contingenza mentre e_{ij} vanno stimati seguendo l'assunzione che le due variabili siano indipendenti. Usiamo la formula:

$$e_{ij} = \frac{\text{COUNT}(A=a_i) \times \text{COUNT}(B=b_j)}{n}$$

Se consideriamo $\frac{\text{COUNT}(A=a_i)}{n}$ è la probabilità $P(A=a_i)$ e $\frac{\text{COUNT}(B=b_j)}{n}$ è $P(B=b_i)$, la formula viene fuori dal moltiplicare queste probabilità fra loro e per n:

$$P(A=a_i) \cdot P(B=b_j) \cdot n =$$

$$\frac{\text{COUNT}(A=a_i)}{n} \cdot \frac{\text{COUNT}(B=b_j)}{n} \cdot n =$$

$$\frac{\text{COUNT}(A=a_i) \times \text{COUNT}(B=b_j)}{n}$$

Sotto l'assunzione che A e B siano indipendenti, posso dire che la probabilità di avere $A=a_i$ insieme a $B=b_j$ è il prodotto delle probabilità:

$$P(A=a_i) \cup P(B=b_j) = P(A=a_i) \cdot P(B=b_j) \cdot n$$

Si moltiplica il tutto per n perché ho la necessità di capire quante istanze ci sono nel database con i valori a_i e b_j

A e B sono davvero indipendenti?

Questo lo si capisce perché $o_{ij} \approx e_{ij}$ se A e B sono davvero indipendenti. Così il χ^2 va verso 0.

Esempio

OBSERVED

	B_1 Giocatori di scacchi	B_2 Non gioca a scacchi	
A_1 Persone che amano fiction scientifiche	250	200	450
A_2 Persone che NON amano fiction scientifiche	50	1000	1050
	300	1200	1500

Calcoliamo i valori EXPECTED

$$e_{11} = P(A_1) \cdot P(B_1) \cdot n = \frac{\text{COUNT}(A_1) \cdot \text{COUNT}(B_1)}{n}$$

$$e_{11} = \frac{450 \cdot 300}{1500} = 90$$

Seguendo gli stessi calcoli anche per la altra costruiamo la tabella dei valori EXPECTED

	B_1 Giocatore di scacchi	B_2 Non gioca a scacchi
A_1 Persone che amano fiction scientifiche	90	360
A_2 Persone che NON amano fiction scientifiche	210	840

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(200-360)^2}{360} + \frac{(50-210)^2}{210} + \frac{(1000-840)^2}{840} = 507,93$$

TEST PARAMETRICO

Quando si lavora sotto l'assunzione che i dati siano distribuiti secondo una ben conosciuta distribuzione di probabilità. Dobbiamo solo valutare i parametri della distribuzione (media e varianza per la Gaussiana)

TEST NON-PARAMETRICO

Quando non è necessario o non si può ricondurre a una forma parametrica della distribuzione dei dati

NULL-HYPOTESIS (H_0)

Le due variabili sono indipendenti

ALTERNATIVE HYPOTESIS (H_1)

Variabili strettamente relate

Attenzione!!!

Due variabili correlate NON IMPLICATO CAUSALITÀ. Per esempio, il numero di ospedali in una zona è correlato col numero di furti d'auto perché entrambi una CAUSA del numero di abitanti.

DEGREE OF FREEDOM

Indica il numero di celle nella tabella che considera le variabili in esame che possono variare

$$df = (c-1)(r-1)$$

Esempio

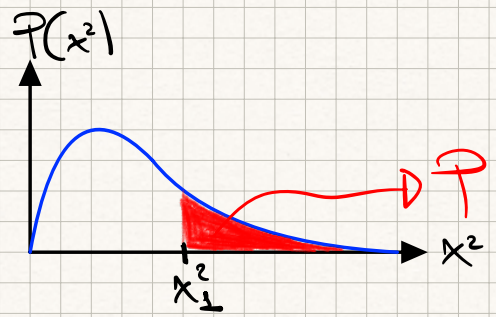
	A		tot
B	3	3	6
	7	8	15
tot	10	11	21

Usando la formula $df = (2-1)(2-1) = 1$
In effetti, per completare la tabella conoscendo i totali ci basta solo fissare il valore di una cella, che quindi può variare entro i limiti.

Come possiamo determinare l'indipendenza?

Il χ^2 è un tipo di distribuzione probabilistica che dipende dai gradi di libertà. La forma della distribuzione del χ^2 cambia a seconda del valore di df.

Diciamo di aver già calcolato un certo df tale per cui la distribuzione del χ^2 è:



Diciamo che il valore del nostro $\chi^2 = \chi^2_1$. Per avere la probabilità che χ^2_1 si presenti dobbiamo calcolare l'area sotto la curva da $\emptyset$ a χ^2_1.

L'area sotto la curva oltre χ^2_1 viene chiamata P-VALUE e ci da indicazioni su con che probabilità si può verificare H_0.

Calcolare l'area sotto la curva può essere difficoltoso per questo, per il nostro scopo, è sufficiente consultare una tabella precompilata che, fissato df e calcolato χ^2 ci da indicazione sul P-VALUE relativo. Per capire se quel P-VALUE va bene dobbiamo basarci sul nostro relativo contesto.

Se la tabella ci dice che per $df = K$ e $P = V$ abbiamo $\chi^2 = X$, allora la cosa funziona anche per valori di $\chi^2 > X$. NON per minori.