Elementi di Psicometria

20-Chi quadro
vers. 1.0 (6 dicembre 2011)
versione per stampa

Germano Rossi[1]
germano.rossi@unimib.it

[1] Dipartimento di Psicologia, Università di Milano-Bicocca

2011-2012

Chi-quadro (χ^2)

- Il termine *chi-quadro* si usa con due significati
  1. Per indicare una famiglia di **distribuzioni di probabilità**
  2. Per un indicare una **statistica** il cui risultato si distribuisce approssimativamente come la distribuzione di probabilità omonima
- Come *statistica* è un **indice di discrepanza**
- Si usa con variabili nominali e/o ordinali

Scopo

- La statistica di chi-quadro (χ^2) ha lo scopo di verificare se un determinato valore osservato si discosta (o no) da un valore teorico (l'ipotesi nulla)
- in concreto si applica a:
  1. una variabile nominale si distribuisce casualmente (ipotesi di omogeneità o di Equiprobabilità: ogni cella ha la stessa probabilità di tutte le altre)
  2. due variabili nominali sono fra loro indipendenti (ipotesi di Indipendenza: Il valore atteso di ogni cella dipende dal prodotto delle probabilità)
  3. una o due variabili si distribuiscono in base a un modello predefinito (Verifica di un modello: io stabilisco qual è il valore atteso di ogni cella)
- le differenze dipendono dal modo in cui vengono calcolate le frequenze teoriche

La formula completa

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

- f_o = frequenza osservata
- f_e = frequenza teorica attesa
- La statistica di χ^2 è la sommatoria degli scarti quadratici fra le frequenze osservate (f_o) e quelle teoriche attese (f_e, expected) ponderate sulle attese.
- Il suo valore oscilla da 0 ad ∞ e aumenta all'aumentare degli scarti ($f_o - f_e$)
- L'uso che si può fare, dipende dal modo in cui si calcola il valore atteso

Esempio 1: equiprobabilità

- Alcuni medici-psichiatri hanno notato che la maggior parte degli schizofrenici sono nati in periodo invernale.
- Ci chiediamo se anche i nostri schizofrenici sono nati in prevalenza nel periodo invernale.
- Usando le cartelle cliniche di 636 pazienti (fittizi) costruiamo la nostra tabella:

Primavera	Estate	Autunno	Inverno	Totale
125	130	153	228	636

Verifica di ipotesi e valori teorici

- Se la nascita di schizofrenici non dipende dal periodo, le 4 stagioni hanno la stessa probabilità
- $H_0 : P(p) = P(e) = P(a) = P(i) = 0.25$
- $H_1 : P(p) \neq P(e) \neq P(a) \neq P(i) \neq 0.25$
- H_0 è l'unica ipotesi su cui possiamo lavorare
- In base ad H_0, ci aspettiamo che in ogni stagione nascano $636/4 = 159$ bambini ovvero $636 * 0.25 = 159$

	P	E	A	I	T
O	125	130	153	228	636
T	159	159	159	159	636
d	-34	-29	-6	69	0

Scarti

- Abbiamo il solito problema che la somma degli scarti si annulla. Lo risolviamo nel solito modo, elevando a quadrato gli scarti:

	P	E	A	I	T
d	-34	-29	-6	69	0
d^2	1156	841	36	4761	6794

- Ora abbiamo il problema di valutare quanto effettivamente grandi siano questi scarti. Un modo per “standardizzarli” è quello di dividerli per il valore teorico di ogni cella.
- Così facendo esprimiamo gli scarti al quadrato, come “numero di valori teorici che stanno nello scarto” (qualcosa di simile a quanto si è fatto con i punti z).

Probabilità

- Quindi, sommiamo tutti gli scarti standardizzati:

	P	E	A	I	T
d^2	1156	841	36	4761	6794
f_t	159	159	159	159	636
	7.27	5.29	0.23	29.94	42.72

- Ottenendo un χ^2 di 42.72
- Qual è la probabilità che $\chi^2 = 42.72$ indichi una variazione casuale rispetto a 4 celle?
- Tutti i valori di chi-quadro si distribuiscono secondo una particolare famiglia di distribuzione di probabilità che variano in base ai “gradi di libertà” (o g.l. o gdl o df)

Gradi di libertà

- Se N frequenze si distribuiscono in c celle, noi possiamo mettere un numero arbitrario di valori nelle prime $c-1$ celle, mentre nell'ultima dobbiamo mettere forzatamente quello che ci avanza:

				totale
125	130	153	X	636

- Nel nostro esempio, i g.l. sono $4-1=3$ perché dopo aver distribuito i 636 casi nelle prime 3 celle, nell'ultima devo mettere gli avanzi.

Significatività

- Stabiliamo un livello $\alpha = .05$
- Usiamo le tavole del chi-quadro (Tavola H, p.487) e cerchiamo il valore critico di χ^2 per 3 g.l.
- $\chi^2_c = 7.815$
- Poiché il nostro χ^2 (42.72) è superiore a quello critico (7.815), concludiamo che le nascite non sono state casuali

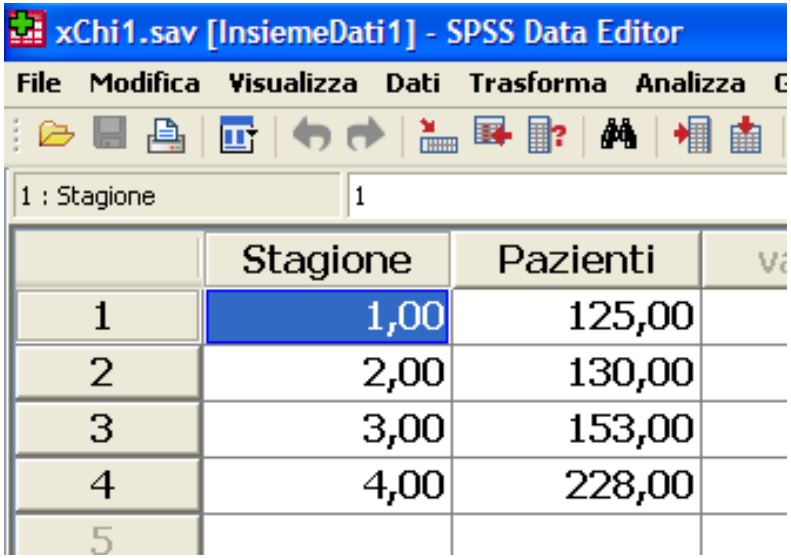
N.B. In Excel si può trovare il χ^2 critico con `=INV.CHI(alfa;gl)`

N.B. In R si può trovare il χ^2 critico con `qchisq(1-alfa,gl)`

dove alfa=α e gl=gradi di libertà

Chi quadro in Spss: equiprobabilità

- `Analizza | Test non parametrici | Chi-quadrato...`
- Mettiamo la variabile qualitativa in `Variabili` `oggetto del test`
- `OK`



`Dati | Pesa casi`, *Pesa casi per* **Pazienti** ovvero

```
WEIGHT BY Pazienti.
```

Test

	Stagione
Chi-quadrato	42,730a
df	3
Sig. Asint.	,000

Per 0 celle (,0%) erano previste frequenze minori di 5. Il valore minimo previsto per la frequenza in una cella è 159,0.

Chi quadro e binomiale

- Se volessimo usare l'approccio dell'equiprobabilità con una variabile di sole due categorie (come il genere) scopriremmo che il valore di chi quadro trovato corrisponderebbe al quadrato dello z calcolato con la binomiale e $p = .05$

Esempio

- File di Sara: verifichiamo se la distribuzione del genere è al 50%
- Usando il chi quadro, troviamo $\chi^2 = 1.96$
- Usando la binomiale: $z = 1.4$
- $1.4^2 = 1.96 \qquad \sqrt{1.96} = 1.4$

Precauzioni nell'uso del chi quadro

- i dati devono essere indipendenti fra loro
- un caso statistico deve stare in una sola cella
- le frequenze attese non devono essere troppo piccole
  - Per df=1, le frequenze attese devono essere almeno 5
  - Per df=2, devono essere almeno 2
  - Per df>=3, una può essere =1 se le altre sono almeno 5

Esempio 2: indipendenza

- **Genere**: Maschi (M) e femmine (F)
- **Livello socio-economico**: Basso (B) e alto (A)
- H_0: le variabili sono fra loro indipendenti
- H_1: le variabili non sono indipendenti

		Livello Educativo		
Sesso		Basso	Alto	Totale
	Maschi	13	9	22
	Femmine	13	7	20
Totale		26	16	42

Valori teorici

- In questo caso non possiamo dividere N per il numero di celle perché avremmo alcuni problemi
- 42/4 = 10.5
- Avremmo quindi 10.5 + 10.5 = 21 *maschi* anziché 22;
- 21 *femmine* anziché 20
- Avremmo anche 21 *Basso* e 21 *Alto*
- I valori teorici devono quindi essere calcolati diversamente
- Devono tener conto del totale dei maschi e delle femmine, ma contemporaneamente dei livelli socio-economici
- Calcoliamo i valori teorici **sulla base della probabilità di 2 eventi indipendenti**

Valori teorici

- La probabilità indipendente di essere *Maschio* di *Basso* livello economico è data dal prodotto delle singole probabilità

$$p(M) = \frac{22}{42} = .52 \qquad p(B) = \frac{26}{42} = .62$$

$$p(MB) = p(M)p(B) = \frac{22}{42} \times \frac{26}{42}$$

- La probabilità ottenuta dovrà essere moltiplicata per il totale per avere la frequenza attesa

$$f_e(MB) = p(M)p(B)N = \frac{22}{42} \times \frac{26}{42} \times 42 = \frac{22 \times 26}{42}$$

Valori teorici

- Dall'applicazione della regola dell'indipendenza degli eventi, si ricava una "regoletta" per il calcolo dei valori teorici:
- La frequenza attesa di una cella è uguale al totale di riga (T_r) per il totale di colonna (T_c) diviso il totale generale (T_t o N)

$$f_e = \frac{T_r \times T_c}{T_t}$$

Valori teorici

- Applicando la regola ad ogni cella della tabella, avremo:

		Freq.	Freq. teorica		
Maschi	Basso	13	22 x 26 / 42	=	13.62
	Alto	9	22 x 16 / 42	=	8.38
Femmine	Basso	13	20 x 26 / 42	=	12.38
	Alto	7	20 x 16 / 42	=	7.62

		Livello Educativo		
Sesso		Basso	Alto	Totale
	Maschi	13 (13.62)	9 (8.38)	22
	Femmine	13 (12.38)	7 (7.62)	20
Totale		26	16	42

- Le frequenze teoriche danno gli stessi totali (di riga, di colonna e generale) delle frequenze osservate

Calcolo del chi-quadro

- Applicando la formula del chi-quadro avremo:

$$\chi^2 = \frac{(13 - 13.62)^2}{13.62} + \frac{(9 - 8.38)^2}{8.38} + \frac{(13 - 12.38)^2}{12.38} +$$

$$+ \frac{(7 - 7.62)^2}{7.62} = 0.0282 + 0.0459 + 0.0311 + 0.0504 = 0.1556$$

- che dovremo confrontare con il chi-quadro critico (χ^2_c)
- Se il nostro χ^2 è inferiore al χ^2_c, allora accetteremo H_0
- Se il nostro χ^2 è superiore o uguale al χ^2_c, allora rifiuteremo H_0

Gradi di libertà

- Per i gradi di libertà, consideriamo che corrispondono al numero di celle necessarie per completare la tabella con i resti, dal momento che i totali (di riga, di colonna e generale) non possono cambiare.

		Livello Educativo		
Sesso		Basso	Alto	Totale
	Maschi	13	X	22
	Femmine	X	X	20
Totale		26	16	42

- In questi caso $gdl = 1$
- Per tabelle di contingenza (incrocio di 2 variabili) la formula generica è quindi:

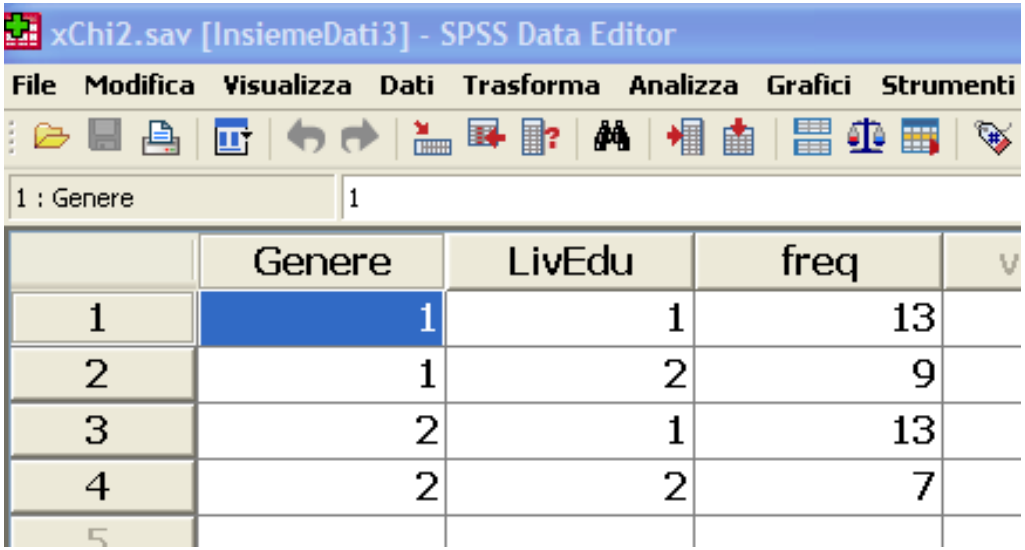
$$gdl = (r - 1)(c - 1)$$

Verifica d'ipotesi

- Il nostro chi quadro ($\chi^2 = 0.1556$) dev'essere confrontato con quello critico
- stabiliamo il livello $\alpha = .05$ e cerchiamo sulla tavola il chi-quadro critico per 1 gdl: $chi_c^2 = 3.841$
- siccome $0.1556 < 3.841$ accettiamo l'ipotesi nulla
- Essendo non significativo per $\alpha = .05$ lo sarà anche per $\alpha = .01$; infatto il chi critico è $chi_c^2 = 6.63$

Chi quadro in Spss: indipendenza

- `Analizza | Statistiche descrittive | Tavole di contingenza...`
- Mettiamo una variabile in `Righe` e una in `Colonne`
- `Statistiche...`, attiva `Chi-quadrato`
- `Continua` e poi `OK`



	Genere	LivEdu	freq
1	1	1	13
2	1	2	9
3	2	1	13
4	2	2	7

`Dati | Pesa casi`, *Pesa casi per* **freq** ovvero

```
WEIGHT BY freq.
```

Chi quadro in Spss: indipendenza

Tavola di contingenza Genere * LivEdu

Conteggio				
		LivEdu		
		1	2	Totale
Genere	1	13	9	22
	2	13	7	20
	Totale	26	16	42

Chi-quadrato

	Valore	df	Sig. asint. (2 vie)	Sig. esatta (2 vie)	Sig. esatta (1 via)
Chi-quadrato di Pearson	,155a	1	,694		
Correzione di continuitàb	,006	1	,940		
Rapporto di verosimiglianza	,155	1	,693		
Test esatto di Fisher				,758	,470
Associazione lineare-lineare	,151	1	,697		
N. di casi validi	42				

a. 0 celle (,0%) hanno un conteggio atteso inferiore a 5.
Il conteggio atteso minimo è 7,62.
b. Calcolato solo per una tabella 2x2

Esempio 3: modello teorico

- Torniamo sull'esempio degli schizofrenici
- Ci possiamo chiedere se nascono più schizofrenici in inverno, perché in inverno nascono più persone
- Per cui, nascendo più persone. è più probabile che nascano anche più schizofrenici
- Per verificare questa ipotesi, devo conoscere la frequenza delle nascite per ogni stagione
- Supponiamo che le percentuali siano:

	Primavera	Estate	Autunno	Inverno	Totale
%	18	20	25	37	

Frequenze teoriche

- Usando le percentuali della popolazione, calcoliamo i nuovi valori teorici ($636 \times 0.18 = 114.48$)

	Primavera	Estate	Autunno	Inverno	Totale
freq. oss.	125	130	153	228	636
% di rif.	18	20	25	37	
freq. att.	114,48	127,2	159	235,32	636

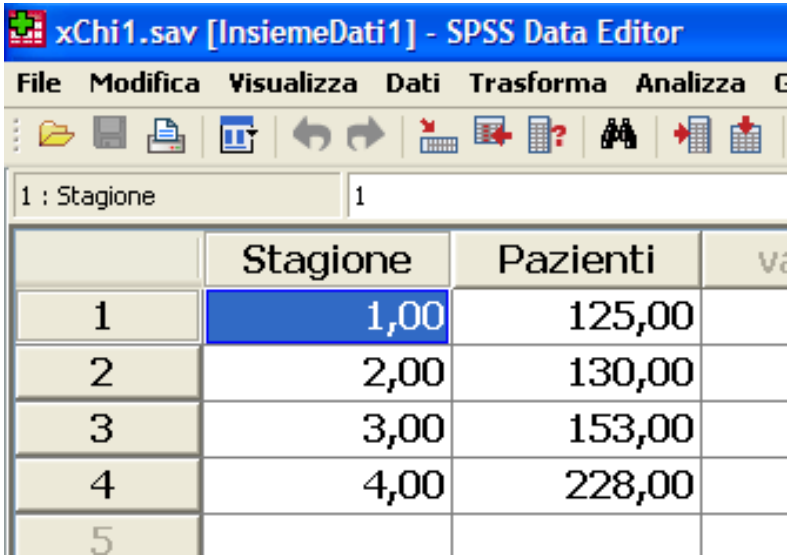
Calcolo del chi-quadro

$$\chi^2 = \frac{(125 - 114.48)^2}{114.48} + \frac{(130 - 127.2)^2}{127.2} + \frac{(153 - 159)^2}{159} +$$
$$+ \frac{(228 - 235.32)^2}{235.32} = 0.9667 + 0.0616 + 0.2264 + 0.2277 = 1.278$$

- χ^2_c per 3 gdl è ancora 7.815
- Se il nostro χ^2 è inferiore al χ^2_c, allora accetteremo H_0
- Se il nostro χ^2 è superiore o uguale al χ^2_c, allora rifiuteremo H_0
- L'ipotesi nulla è ancora l'equiprobabilità
- ma...
- adesso noi vogliamo che il χ^2 sia piccolo perché significa che abbiamo ragione!

Chi quadro in Spss: teoria

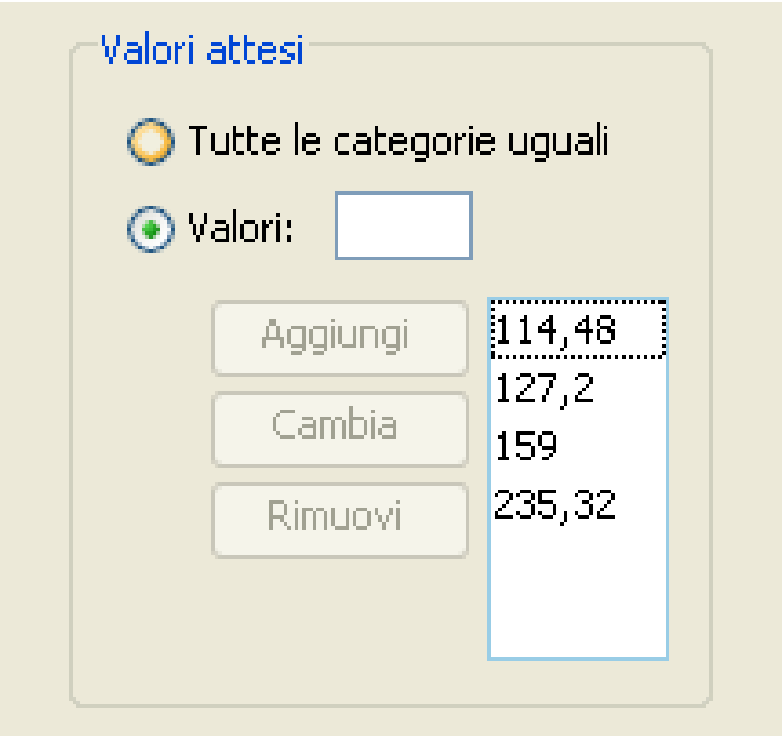
- `Analizza | Test non parametrici | Chi-quadrato...`
- Mettiamo la variabile qualitativa in `Variabili oggetto del test`
- Nell'area `Valori attesi`, scegliere `Valori` e inserire i valori teorici uno alla volta (con `Aggiungi`)
- `OK`



`Dati | Pesa casi`, *Pesa casi per* **Pazienti** ovvero

```
WEIGHT BY Pazienti.
```

Chi quadro in Spss: teoria



Test	
	Stagione
Chi-quadrato	1,482a
df	3
Sig. Asint.	0,686321

a. Per 0 celle (,0%) erano previste frequenze minori di 5. Il valore minimo previsto per la frequenza in una cella è 114,5.

Correzione di continuità

- In certe condizioni, il valore della statistica χ^2 non si approssima bene alla distribuzione di χ^2
- In questi casi si usa la correzione di continuità di Yates

$$\chi^2 = \sum \frac{(|f_o - f_e| - .5)^2}{f_e}$$

- Le condizioni in cui usarlo non sono sempre chiare
  - Quando la tabella è 2x2 (la scelta di Spss)
  - Quando gl=1 e almeno una cella ha una frequenza attesa minore di 5 ($f_e < 5$)
  - Quando gl=2 e almeno una cella ha una freq. attesa minore di 3
  - Quando più del 20% delle celle ha una frequenza attesa minore di 5
  - Sempre perché la distribuzione χ^2 è continua e i dati sono discreti

Problemi di numerosità

- Il chi-quadro è sensibile alla numerosità.
- Riprendiamo l'esempio 2, ma moltiplichiamo tutte le celle per 10

	Livello Educativo		
Sesso	Basso	Alto	Totale
Maschi	13	9	22
Femmine	13	7	20
Totale	26	16	42

	Livello Educativo		
Sesso	Basso	Alto	Totale
Maschi	130	90	220
Femmine	130	70	200
Totale	260	160	420

- Anche il chi quadro risulterà moltiplicato per 10 ($\chi = 1.551$)
- E ancora una volta non è significativo perché inferiore al valore critico (3.815) che non cambia perché dipende dai gdl
- Ma se avessi 4200 valori (tutto moltiplicato per 100)?
- il χ^2 sarebbe 15.5 (significativo!)

Problemi di numerosità

- Dal momento che il chi-quadro tende ad aumentare all'aumentare del totale delle frequenze, e quindi a diventare significativo, si può ragionevolmente dubitare che la significatività trovata sia effettivamente vera
- Una possibile soluzione è il o **coefficiente phi**

$$\phi = \sqrt{\frac{\chi^2}{N}} \qquad \text{(Cramer)}\phi = \sqrt{\frac{\chi^2}{N(k-1)}} \qquad k = min(r, c)$$

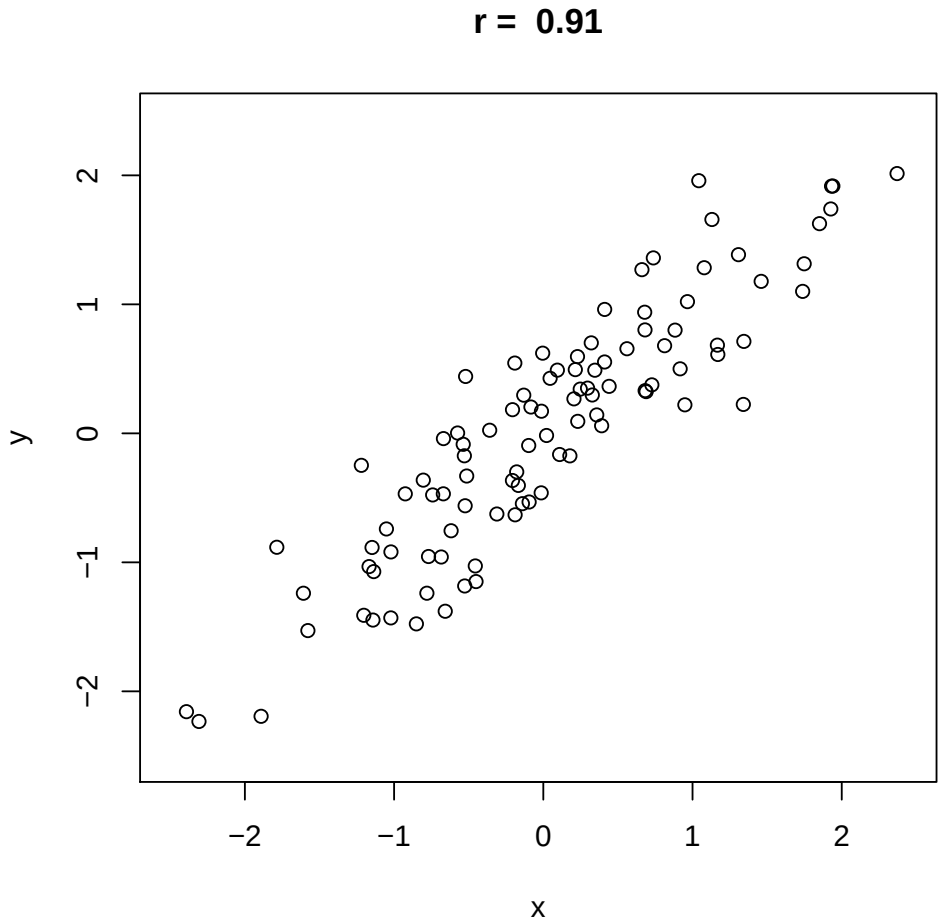
  che (in spss) si può chiedere tramite il pulsante `Statistiche`
- Se tale coefficiente si avvicina a 0, allora il chi-quadro era elevato per colpa della numerosità

Indice di associazione/Effect size

- Il **coefficiente phi** è anche un indice di associazione fra le due variabili
- Un indice di associazione misura la "forza" con cui le due variabili sono legate fra loro
- Per questo motivo, ϕ misura anche l'ampiezza dell'effetto
- Il χ^2 ci dice che le due variabili sono fra di loro dipendenti o indipendenti ed effettua un test probabilistico.
- Rifiutando l'ipotesi nulla, stiamo solo dicendo che, probabilmente, c'è un legame fra le variabili all'interno della popolazione da cui abbiamo estratto il campione
- ϕ ci dice invece **quanto** le due variabili sono legate fra loro

Indice di associazione/Effect size

- ϕ in una tabella 2x2 corrisponde ad una r di Pearson (che è l'indice di associazione per variabili quantitative)
- il segno indica la direzione
- il valore indica l'intensità



Frequenze attese < 5

- Quando più del 20% di celle ha una frequenza attesa inferiore a 5, la statistica chi-quadro non si approssima alla sua distribuzione di probabilità
- Alcuni autori suggeriscono di usare la correzione di continuità di Yates
- Altri autori suggeriscono di accorpare qualche categoria di una delle due variabili (o di entrambe) per avere totali di riga (o di colonna) più elevati
- Altri ancora suggeriscono di usare il *log chi-quadro* ovvero il corrispondente loglineare del chi-quadro (che in Spss è chiamato *Rapporto di verosimiglianza*)

$$G^2 = 2 \sum (f_o) ln \left(\frac{f_o}{f_e} \right)$$

Frequenze attese < 5

- Il limite di 5 (o 3) per le frequenze attese deriva uno studio di Lewis e Burke (1949).
- Successivamente, diverse ricerche sono giunte a conclusioni diverse (sintetizzate in Delucchi, 1983)
- Il chi-quadro non è molto sensibile alle frequenze attese piccole o alle celle con poche frequenze se l'N totale è almeno superiore a $r \cdot c \cdot 5$
- Tuttavia questa possibilità incide solo sull'errore α, mentre resta sconosciuto l'effetto sull'errore β

Residui standardizzati

- Il chi-quadro fornisce un'informazione complessiva sull'intera tabella
- È possibile considerare il singolo valore di chi-quadro di una cella per giudicare se questa cella è responsabile (o contribuisce in modo significativo) alla significatività totale.
- È più semplice considerare i *residui aggiustati standardizzati* che si distribuiscono come punti z
- se il valore di un *residuo aggiustato standardizzato* di una cella è positivo e significativo, vuol dire che in quella cella ci sono più frequenze di quante previste dalla teoria
- se è negativo e significativo, che ce ne sono di meno di quanto previsto

Riepilogo: equiprobabilità

- *Ipotesi* di equiprobabilità
- *Usando*: 1 variabile qualitativa
- *Valori attesi* calcolati come $f_e = N/celle$
- *Gdl*: (celle-1)
- *Ipotesi*: $H_0 : \chi^2 = 0$ e $H_1 : \chi^2 \neq 0$
- *Ipotesi da falsificare*: H_0
- *Risultato cercato*: **significatività**, rifiuto di H_0, χ^2 sig.

Riepilogo: indipendenza

- *Ipotesi* di indipendenza
- *Usando*: una tabella di contingenza (2 variabili qualitative)
- *Valori attesi* calcolati come $f_e = p(T_r)p(T_c)N = (T_r \times T_c)/T_t$
- *Gdl*: (celle-1)(righe-1)
- *Ipotesi*: $H_0 : \chi^2 = 0$ e $H_1 : \chi^2 \neq 0$
- *Ipotesi da falsificare*: H_0
- *Risultato cercato*: **significatività**, rifiuto di H_0, χ^2 sig.

Riepilogo: verifica modello

- *Ipotesi* basata su un modello
- *Usando*: indifferente (1 o più variabili)
- *Valori attesi* calcolati in base ad una teoria
- *Gdl*: dipende dal modello
- *Ipotesi*: $H_0 : \chi^2 = 0$ e $H_1 : \chi^2 \neq 0$
- *Ipotesi da verificare*: H_0
- *Risultato cercato*: **non significatività**, accettazione di H_0, χ^2 **non** sig.

Riepilogo

- Chi-quadrato
  - **Equiprobabilità**: una variabile qualitativa viene analizzata per vedere se le categorie sono fra loro equiprobabili
  - **Indipendenza**: due variabili qualitative vengono incrociate (tabella di contingenza) per vedere se sono fra loro indipendenti
  - **Modello teorico**: una variabile qualitativa viene confrontata con un modello teorico per vedere se le categorie si distribuiscono in base a dei valori attesi indicati dalla teoria
  - **Modello generico**: una qualunque tabella di dati osservati viene confrontata con valori attesi calcolati in base ad una teoria (o modello teorico) [non disponibile in SPSS]