

Elementi di Psicometria con Laboratorio di SPSS 1

20-Chi quadro
(v. 1.5, 9 maggio 2018)
versione per stampa

Germano Rossi¹

`germano.rossi@unimib.it`

¹Dipartimento di Psicologia, Università di Milano-Bicocca

9 maggio 2018

Analisi di dati qualitativi

- Finora ci siamo occupati di studiare le analisi dei dati (e le tecniche statistiche) che trattano le variabili **quantitative**.
 - t-test per campioni appaiati (2 quantitative)
 - t-test per campioni indipendenti (1 qualitativa suddivisa in due gruppi)
 - t-test per campione unico (1 quantitativa e una media di riferimento)
 - correlazione di Pearson (2 quantitative)
- Adesso affrontiamo una tecnica di analisi dei dati che utilizza variabili **qualitative**
- Questa e altre tecniche vengono anche chiamate **non parametriche** perché non fanno riferimento ai parametri della popolazione

Dati qualitativi

- Le variabili qualitative possono essere sia ordinali sia nominali
- e possono presentarsi da sole oppure associate con altre
- L'utilizzo più frequente è con le tabelle di contingenza (ad es. vedere se il nostro campione è bilanciato per genere e fasce d'età)
- Altri utilizzi sono
 - una variabile di cui si vuole testare la distribuzione casuale (ad es. se la variabile fasce di età nel nostro campione si distribuisce in modo omogeneo)
 - una distribuzione teorica (ad es. se una variabile del nostro campione rispetta la proporzione presente nella popolazione)
 - tre o più variabili qualitative (analisi loglineare)

Dati qualitativi

- La maggior parte di queste tecniche utilizza variabili qualitative con poche categorie
- Ci sono tecniche di analisi che utilizzano una o più variabili qualitative con molte/moltissime categorie (analisi delle corrispondenze semplice; analisi delle corrispondenze multiple)
- Noi ci occuperemo della tecnica del chi-quadro e del test esatto di Fisher

Chi-quadro (χ^2)

- Il termine *chi-quadro* si usa con tre 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
 - 3 Per indicare la **tecnica di analisi dei dati**
- 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

$$\chi^2 = \sum \frac{(f_o - f_a)^2}{f_a} = \sum \frac{(O - A)^2}{A} = \sum \frac{O^2}{A} - N$$

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

Avviso

Per farvi capire meglio, procederò secondo questo ordine:

- 1 Modello casuale (o di equiprobabilità) [non presente nel libro]
- 2 Modello di indipendenza [presente nel libro]
- 3 Modello teorico [non presente nel libro]

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

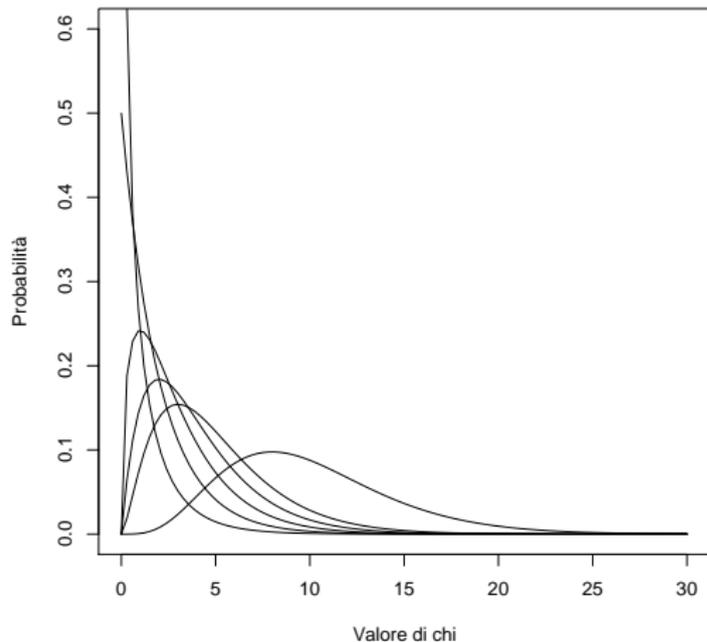
- Se N frequenze si distribuiscono in c celle, noi possiamo mettere un numero arbitrario di valori nelle prime $c - 1$ celle, mentre nell'ultima dobbiamo mettere forzatamente quello che ci avanza:

				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.

Distribuzione di chi-quadro (χ^2)

Curve di chi quadro per 1,2,3,4,5 e 10 g.l

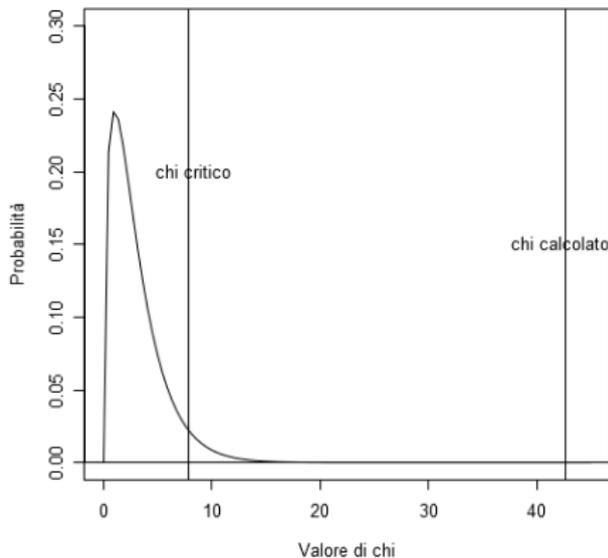


$$f(x) = \frac{2^{-k/2}}{\Gamma(k/2)} x^{(k-2)/2} e^{-x/2}$$

- Con pochi gdl (1 e 2), i valori più probabili sono molto vicini a 0
- già con gdl=3, 0 non è probabile
- all'aumentare dei gdl la curva assomiglia sempre più a una normale (ma non lo è)

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. per un certo livello di alfa
- per $\alpha = .05$ è $\chi_c^2 = 7.815$
- il χ^2 da noi trovato è $\chi_t^2 = 42.72$



Significatività

- Poiché il nostro χ^2 (42.72) è superiore a quello critico (7.815), concludiamo che le nascite non sono state casuali
- Più precisamente, ipotizzando l'equiprobabilità, il chi-quadro significativo ci dice che se rifiutiamo H_0 corriamo un rischio inferiore (di molto inferiore) al 5% di prendere una decisione sbagliata
- siccome il χ_c^2 per $\alpha = .01$ è 11.35, possiamo anche dire che il rischio che stiamo correndo è inferiore al 1%
- siccome il χ_c^2 per $\alpha = .001$ è 16.27, possiamo anche dire che il rischio che stiamo correndo è inferiore al 0.1%

Chi quadro in Spss: equiprobabilità

- Analizza | Test non parametrici | Chi-quadrato...

- Mettiamo la variabile qualitativa in Variabili oggetto del test



Test

Stagione

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

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

Stagione
1
3
4
2
3
1
4
...

Prime 7 osservazioni (in totale 636 righe)
In alternativa possiamo “pesare i casi”

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: $z = 1.4$
- $1.4^2 = 1.96$ $\sqrt{1.96} = 1.4$

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 $df \geq 3$, una può essere =1 se le altre sono almeno 5

Esempio 2: indipendenza

- Ho raccolto un campione di 42 ragazze/i e ho misurato (fra l'altro):
- **Genere:** Maschi (M) e femmine (F)
- **Livello socio-economico:** Basso (B) e alto (A)
- Mi chiedo “Le variabili sono fra loro associate?” ovvero una variabile ha qualche influenza sull'altra?
- H_0 : le variabili sono fra loro indipendenti
- H_1 : le variabili non sono indipendenti

<i>Genere</i>		<i>Livello Educativo</i>		
		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

	Bas	Alt	Tot
Maschi	13	9	22
Femmine	13	7	20
Totale	26	16	42

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

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

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

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

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

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)

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

Valori teorici

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

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

Sesso	Livello Educativo			Totale
	Basso	Alto		
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_c^2)
- Se il nostro χ^2 è inferiore al χ_c^2 , allora accetteremo H_0
- Se il nostro χ^2 è superiore o uguale al χ_c^2 , 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.

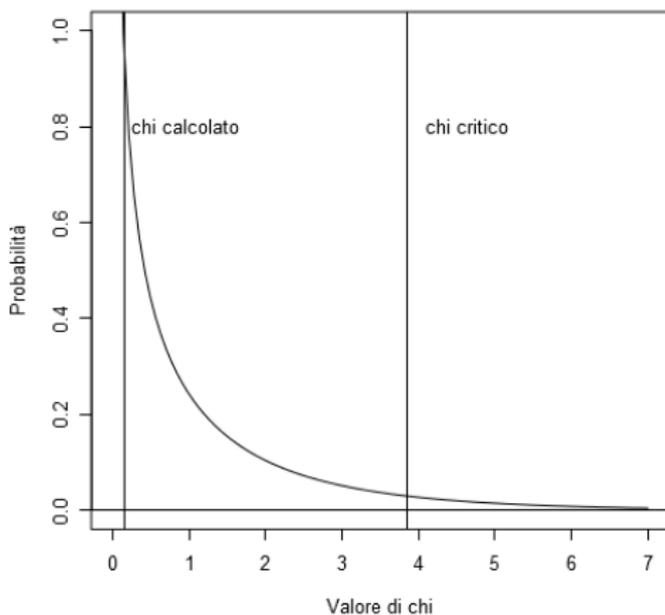
Sesso	Livello Educativo		Totale
	Basso	Alto	
Maschi	13	X	22
Femmine	X	X	20
Totale	26	16	42

- 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 $\alpha = .01$; infatti 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**
- , attiva
Chi-quadrato
- e poi

Genere	LivEdu
1	1
2	1
2	2
1	2
1	2
2	1
...	...

Prime 7 osservazioni (in totale 42 righe)
In alternativa possiamo “pesare i casi”

Chi quadro in Spss: indipendenza

Tavola di contingenza Genere * LivEdu

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

Chi-quadro

	Valore	df	Sig. asint. (2 vie)	Sig. esatta (2 vie)	Sig. esatta (1 via)
Chi-quadro di Pearson	,155 ^a	1	,694		
Correzione di continuità ^b	,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
%	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_c^2 per 3 gdl è ancora 7.815
- Se il nostro χ^2 è inferiore al χ_c^2 , allora accetteremo H_0
- Se il nostro χ^2 è superiore o uguale al χ_c^2 , allora rifiuteremo H_0
- L'ipotesi nulla è il nostro **modello teorico**,
- **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)
-



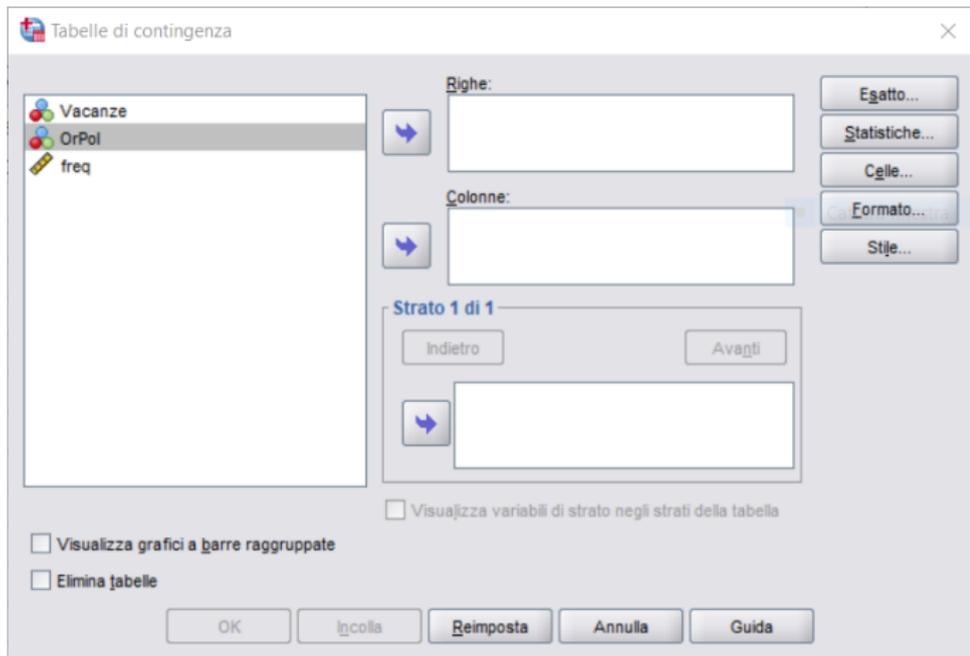
Stagione	
Chi-quadrato	1,482 ^a
df	3
Sig. Asint.	0,686321

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

SPSS: Chi quadro indipendenza

Analizza | Statistiche descrittive | Tabella di contingenza...

- Inserire almeno una variabile qualitativa in Righe e almeno una in Colonne
- Usare Celle... per i dati da visualizzare
- Usare Statistiche... per i test da calcolare



SPSS: Chi quadro indipendenza, Celle

Celle...

- Serve per selezionare i contenuti delle celle
- **Osservati** è attivo per default
- **Previsti** visualizza le frequenze attese (utile per freq. attese <5)
- **Non standardizzato**, **Standardizzato** non si usano
- **Standardizzato adattato** per avere i residui trasformati in punti z

Tabelle di contingenza: Visualizzazione cella

Conteggi

Osservati

Previsti

Nascondi conteggi piccoli

Minore di

test z

Confronta proporzioni di colonna

Adatta i valori P (metodo di Bonferroni)

Percentuali

Riga

Colonna

Totale

Residui

Non standardizzato

Standardizzato

Standardizzato adattato

Pesi non interi

Arrotonda conteggi delle celle

Arrotonda pesi del caso

Tronca conteggi delle celle

Tronca pesi del caso

Nessun adattamento

SPSS: Chi quadro indipendenza, Statistiche

Statistiche...

- **Chi quadrato** per calcolare il χ^2
- **Coefficiente di contingenza, Phi e V di Cramer** per stampare l'*effect size*

Tabelle di contingenza: Statistiche

Chi-quadrato

Correlazioni

Nominale

Coefficiente di contingenza

Phi e V di Cramer

Lambda

Coefficiente di incertezza

Ordinale

Gamma

D di Somers

Tau-b di Kendall

Tau-c di Kendall

Nominale per intervallo

Eta

Kappa

Rischio

McNemar

Statistiche di Cochran e Mantel-Haenszel

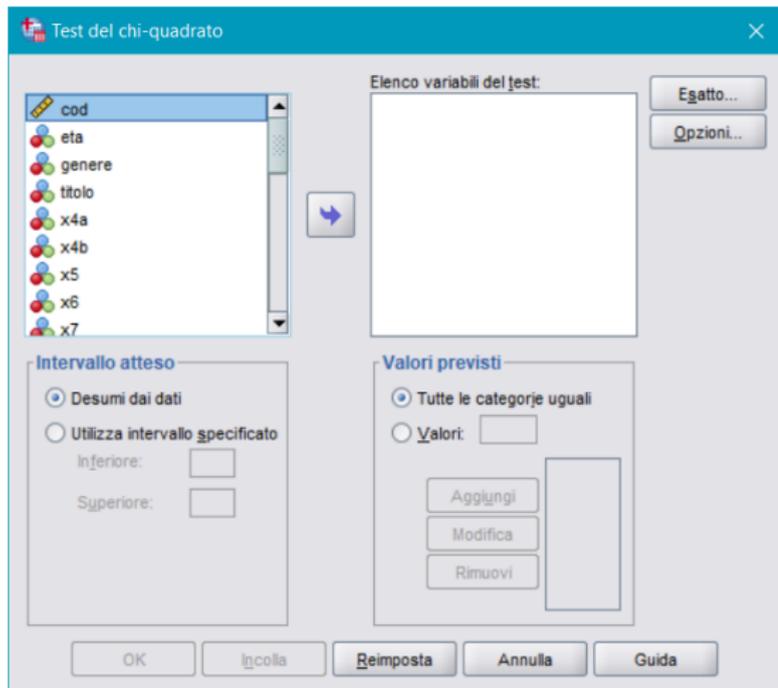
Test di uguaglianza rapporto odd comune: 1

Continua Annulla Guida

SPSS: Chi quadro equiprobabilità

Analizza | Test non parametrici | Finestre di dialogo legacy | Chi-quadrato...

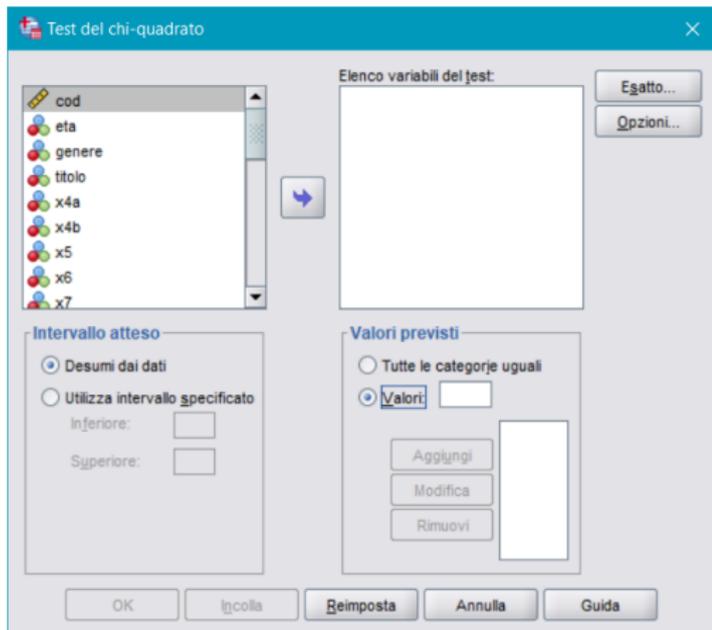
- Selezionare almeno una variabile
- Non cambiare nulla, per usare tutti i valori della variabile originale
- Scegliere *Utilizza intervallo specificato*, se si vuole limitare il numero di categorie (che devono essere però consecutive)
- In tal caso inserire il valore minimo da usare e quello massimo



SPSS: Chi quadro modello teorico

Analizza | Test non parametrici | Finestre di dialogo legacy | Chi-quadrato...

- Selezionare almeno una variabile
- Nel riquadro Valori previsti selezionare Valori
- Inserire di fianco il valore teorico della prima categoria e poi premere Aggiungi
- Ripetere per tutte le categorie
- I valori inseriti devono sommare a 1 (proporzioni), 100 (percentuali) o N (frequenze)
- Altrimenti vengono sommati e ricalcolate le proporzioni



Studio dei residui

- Il chi-quadro fornisce un'informazione complessiva sull'intera tabella
- Ovvero, c'è ($p \leq \alpha$) o non c'è ($p > \alpha$) associazione fra frequenze osservate e frequenze attese? (sì/no)
- Se c'è associazione ($p \leq \alpha$), possiamo capire quali celle sono in qualche modo responsabili della significatività?
- Dal momento che il χ^2 è la somma dei χ^2 di tutte le celle, le celle che producono χ^2 più elevati siano quelle che contribuiscono maggiormente ad innalzare il valore del χ^2 totale.
- Però, questi χ^2 parziali sono confrontabili fra loro, ma non sappiamo quanto devono essere "grandi" per poter dare un contributo maggiore

Residui aggiustati standardizzati

- Possiamo allora considerare i residui semplici ($O - A$) ma dipendono dall'ampiezza delle frequenze della cella
- Questi residui semplici si possono standardizzare, ma diventano confrontabili ma non sappiamo se sono sufficientemente grandi statisticamente.
- Si possono allora trasformare in punti z
- I residui trasformati in punto Z si chiamano **residui aggiustati standardizzati**
- se il valore di un *residuo aggiustato standardizzato* di una cella è significativo, vuol dire che in quella cella la differenza fra osservate e attese non dipende da fluttuazioni casuali

Residui aggiustati standardizzati

- La formula per il calcolo è:

$$z_r = \frac{O - A}{\sqrt{T_r * T_c * \frac{\frac{1 - T_r}{N} * \frac{1 - T_c}{N}}{N}}}$$

- se z_r è positivo e significativo, ci sono più Osservate di quanto previsto
- se z_r è negativo e significativo, ci sono meno Osservate di quanto previsto
- quando la differenza fra O e A è statisticamente significativa?

Residui aggiustati standardizzati

- Nella curva normale,
 - con ipotesi bidirezionale
 - e $\alpha = .5$,
- gli z critici che corrispondono al 5% sulle due code (2.5% sul lato negativo e 2.5% su quello positivo) sono **-1.96 e +1.96**
- I residui aggiustati standardizzati (valore assoluto) superiori a 1.96 sono statisticamente significativi

Residui aggiustati in SPSS

- **Vacanze** 'Vai in vacanza prevalentemente nello stesso posto?' 1=sì 2=no
- **OrPol** 'Qual è il tuo orientamento politico?' 1=Nessuno 2=Sinistra 3=Centro 4=Destra
- $\chi^2(3) = 34.735, p < .001, V = .150$

Tavola di contingenza Vacanze * OrPol

			OrPol				Totale
			1	2	3	4	
Vacanze	1	Conteggio	127	347	286	290	1050
		Conteggio previsto	149,6	301,9	303,3	295,2	1050,0
		Residuo adattato	-3,5	5,4	-2,1	-,6	
2	2	Conteggio	94	99	162	146	501
		Conteggio previsto	71,4	144,1	144,7	140,8	501,0
		Residuo adattato	3,5	-5,4	2,1	,6	
Totale		Conteggio	221	446	448	436	1551
		Conteggio previsto	221,0	446,0	448,0	436,0	1551,0

Residui aggiustati in SPSS

- La somma dei residui aggiustati (per riga e per colonna) deve essere 0
- $-3.5 + (-2.1) + (-0.6) = 5.4$
- Con 2 celle, uno dei residui è positivo, l'altro negativo
- -3.5 vs. 3.5

1	2	3	4
127	347	286	290
149,6	301,9	303,3	295,2
-3,5	5,4	-2,1	-,6
94	99	162	146
71,4	144,1	144,7	140,8
3,5	-5,4	2,1	,6
221	446	448	436
221,0	446,0	448,0	436,0

Indice di associazione/Effect size

- Abbiamo già visto che la statistica χ^2 (per tabelle di contingenza) ci dice se Osservate e Attese sono discrepanti o meno
- Se c'è discrepanza, allora non c'è indipendenza fra le variabili
- Se non c'è indipendenza le variabili sono fra loro associate
- È possibile calcolare degli indici di ampiezza dell'effetto che ci dicono "quanto sono associate"
- questi indici sono il Coefficiente di contingenza, la ϕ e la V di Cramer

Coefficiente phi

- 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
- ϕ ci dice **quanto** le due variabili sono legate fra loro
- Il coefficiente **phi** è calcolato con due distinte formule

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

- La ϕ oscilla fra -1 e +1 ed è a tutti gli effetti una correlazione.

Coefficiente C di contingenza

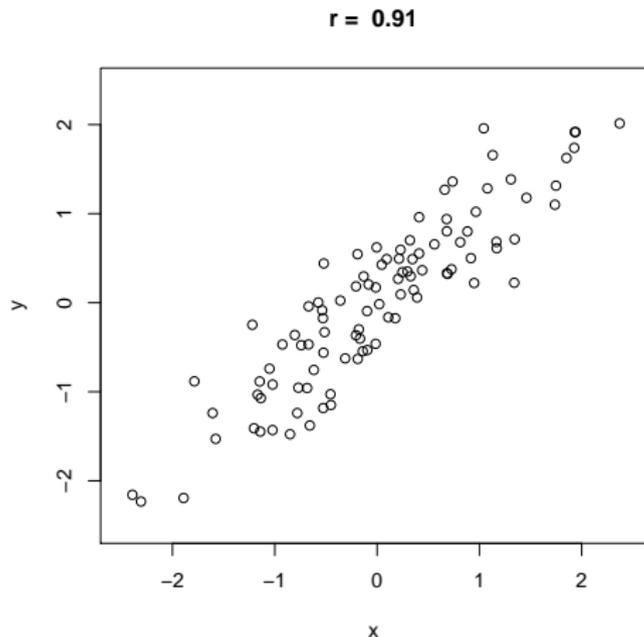
- Anche il coefficiente di contingenza può essere usato come effect size

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}}$$

- Inoltre può essere usato per verificare se la numerosità è responsabile della significatività del χ^2
- L'indice C di contingenza oscilla fra -1 e +1

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



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

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

- 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 ($A < 5$)
 - 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

Sesso	Livello Educativo			Sesso	Livello Educativo		
	Basso	Alto	Totale		Basso	Alto	Totale
Maschi	13	9	22	Maschi	130	90	220
Femmine	13	7	20	Femmine	130	70	200
Totale	26	16	42	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 15.51 (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 **coefficiente phi** (se $g=1$) o **V di Cramer** (se $g>1$)

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

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

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
- *Accorpare* significa unire tra loro due o più categorie di una variabile
- Se la variabile è ordinale, bisogna fare attenzione a cosa si accorpa (la condizione economica “medio-bassa” può essere accorpata a quella “bassa” ma non a quella “medio-alta”!)
- Per variabili nominali è più semplice perché si può creare una categoria “altro”

Frequenze attese < 5

- Le categorie che non si possono accorpate, possono essere “eliminate” dichiarandole come “mancanti definiti dall’utente”
- È possibile usare il test esatto di Fischer (per tabelle 2x2 o 2x3, ma SPSS lo calcola solo se 2x2), $n!$ indica il fattoriale

$$\frac{(a+b)!(c+d)!(a+c)!(b+d)!}{N!a!b!c!d!}$$

a	b
c	d

- Altri ancora suggeriscono di usare il *log chi-quadro* ovvero il corrispondente loglineare del chi-quadro (che in Spss è chiamato *Rapporto di verosimiglianza*)

$$G^2 = 2 \sum (O) \ln \left(\frac{O}{A} \right)$$

Frequenze attese < 5

- Il limite di 5 (o 3) per le frequenze attese deriva da 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

Suddividere chi-quadro

- Un chi-quadro significativo indica che le due variabili sono in qualche modo legate
- Con 6 o più celle non sempre è facile capire il modo in cui le variabili sono legate
- Ci sono alcune tecniche che ci possono aiutare:
 - La partizione del chi-quadro
 - I residui standardizzati corretti

Partizione del chi-quadro

- la partizione implica suddividere la tabella in tante sotto-tabelle di 2 righe e due colonne
- e applicare il chi-quadro a ciascuna delle tabelle
- ogni tabella avrà 1 gdl (essendo 2x2)
- però bisogna aggiustare la significatività tramite il criterio di Bonferroni (α / numero di confronti)
- Se ho una tabella 2x3 posso unire le categoria A1 con A2 e B1 con B2, poi A2 con A3 e B2 con B3...

A1	A2	A3
B1	B2	B3

A1+A2	A3
B1+B2	B3

A1	A2+A3
B1	B2+B3

Riepilogo: equiprobabilità

- *Ipotesi* di equiprobabilità
- *Usando*: 1 variabile qualitativa
- *Valori attesi* calcolati come $A = N/celle$
- *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 $A = p(T_r)p(T_c)N = (T_r \times T_c)/T_t$
- *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]