

Capitolo 5. Chi quadro. Goodness-of-fit test. Test di simmetria. Tabelle 2x2. Tabelle mxn. Correzione per la continuità. Test esatto di Fisher. Tabelle 2x2 per dati appaiati: test di McNemar.

Il chi-quadro

Il test del chi-quadro χ^2 serve a saggiare l'ipotesi che una certa discrepanza tra frequenze attese e frequenze osservate sia dovuta:

H0: al caso (campionamento, imprecisione, errore distribuito, ecc.) oppure a

H1: al fatto che il campione provenga da una popolazione diversa da quella da cui deriva la frequenza attesa.

Il test consiste nel rapporto:

$$c^2 = \frac{(\text{frequenze osservate} - \text{frequenze attese})^2}{\text{frequenze attese}}$$

Nota bene: le frequenze sono sempre frequenze assolute. Mai si possono utilizzare frequenze percentuali o relative. Le percentuali quindi vanno sempre ritrasformate in frequenze assolute moltiplicandole per il numero di osservazioni. Con i soli dati percentuali (senza il numero totale di osservazioni da cui ricavare le frequenze assolute) non si può fare alcun test.

Come al solito, esiste una tabella di valori critici. Se il valore del χ^2 supera almeno quello tabulato per $\alpha=0.05$ si accetta l'H1 e si rifiuta l'H0.

Valori critici di χ^2 corrispondenti a $P \leq 0.05$ e $P \leq 0.01$		
	Probabilità di avere un χ^2 maggiore	
Gradi di libertà	$\alpha = 0.05$ o 5%	$\alpha = 0.01$ o 1%
1	3.841	6.635
2	5.991	9.210
3	7.815	11.345
4	9.488	13.277
5	11.070	15.086
6	12.592	16.812
7	14.067	18.475
8	15.507	20.090
9	16.919	21.666
10	18.307	23.209

Si noti che il valore critico di χ^2 con 1 grado di libertà corrisponde al valore critico di z al quadrato:

$3.841 = 1.96^2$ (se non altro questo ci può aiutare a ricordarlo).

Esistono diversi test basati sul χ^2 . Vediamo insieme i più importanti:

Goodness-of-fit (bontà di adattamento)

Per testare l'adattamento dei dati ad una distribuzione basta confrontare le frequenze rappresentate nell'istogramma con le frequenze previste dalla distribuzione per le stesse classi dell'istogramma.

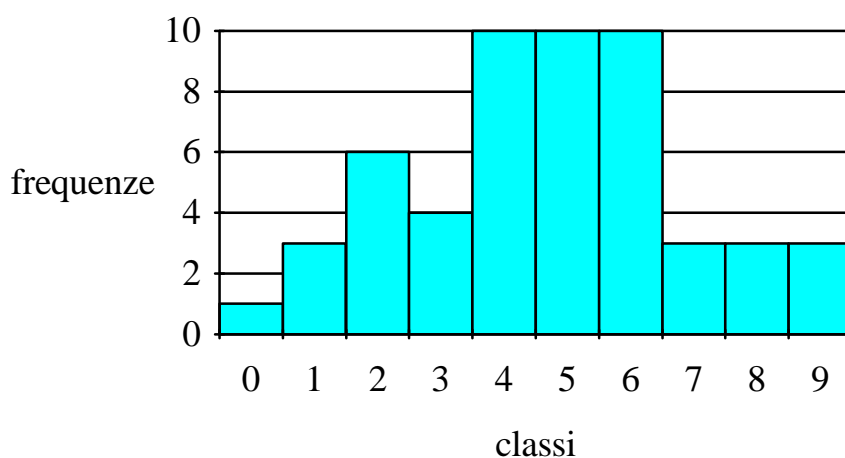
Un esempio. Immaginiamo il seguente campione:

1,7,6,4,5,6,8,9,4,5,6,5,4,2,4,8,2,6,4,9,2,3,4,7,6,5,4,5,2,3,5,6,8,5,3,6,7,0,2,4,3,5,6,
2,5,1,1,6,5,9,4,6,4

con media=4.7, deviazione standard=2.17 e numerosità=53.

Una volta messi in ordine crescente i dati è più facile costruire l'istogramma di frequenze:

0,1,1,1,2,2,2,2,2,3,3,3,3,4,4,4,4,4,4,4,4,4,5,5,5,5,5,5,5,5,6,6,6,6,6,6,6,6,6,
6,7,7,7,8,8,8,9,9,9



Con gli stessi dati allestiamo la seguente tabella:

Limite inferiore (incluso)	Limite superiore (escluso)	Frequenze osservate	Limite (z) inferiore standardizzato	Limite (z) superiore standardizzato	Frequenze attese percentuali (dall'area compresa tra i confini della classe)	Frequenze attese assolute	χ^2
da....	a...		da...	a...			
$-\infty$	0.5	1	$-\infty$	-1.94	0.02	1	0.00
0.5	1.5	3	-1.94	-1.47	0.04	2	0.50
1.5	2.5	6	-1.47	-1.01	0.09	5	0.20
2.5	3.5	4	-1.01	-0.55	0.13	7	1.29
3.5	4.5	10	-0.55	-0.09	0.17	9	0.11
4.5	5.5	10	-0.09	+0.37	0.19	10	0.00
5.5	6.5	10	+0.37	+0.83	0.15	8	0.50
6.5	7.5	3	+0.83	+1.29	0.11	6	1.50
7.5	8.5	3	+1.29	+1.75	0.06	3	0.00
8.5	$+\infty$	3	+1.75	$+\infty$	0.04	2	0.50
		53			1	53	5.78

Il χ^2 calcolato ha gradi di libertà pari al numero delle classi - 3, cioè $10-3=7$.

Il valore critico del χ^2 (vedi tabella) con 7 GDL al livello di probabilità dello 0.05 è 14.067. Il valore trovato è inferiore. Quindi le discrepanze tra le frequenze osservate e le frequenze attese dalla distribuzione normale non sono significative. Pertanto manteniamo l'ipotesi nulla che vuole che la nostra distribuzione sia compatibile con quella di un campione estratto da una popolazione con distribuzione normale.

P.S.: le prime e ultime classi con frequenza attesa inferiore a 5 vanno accorpate alle classi adiacenti. Nell'esempio ciò è stato omesso per non complicare la dimostrazione.

Test di simmetria

Molto più semplice del precedente, il test di simmetria è basato sul fatto che, se la distribuzione è simmetrica, metà dei valori ($n/2$) saranno inferiori alla media e metà ($n/2$) saranno superiori (escludendo che esistano valori identici alla media, o se ve ne sono li togliamo dal conto). In pratica, tale test corrisponde ad un confronto tra media e mediana. Si tratta di confrontare le frequenze osservate dei valori inferiori e superiori alla media con le frequenze attese ($n/2$ ed $n/2$). Non occorre calcolare nessun altro parametro.

Per esempio, immaginiamo il seguente campione:

4, 5, 8, 3, 4, 8, 4, 5, 6, 3, 4, 2, 3, 4, 9, 5

con media=4.81 e numerosità=16.

I valori inferiori alla media sono 9: 4, 3, 4, 4, 3, 4, 2, 3, 4

I valori superiori alla media sono 7: 5, 8, 8, 5, 6, 9, 5

contro frequenze attese di 8 e 8.

Pertanto il test consisterà nel calcolo:

$$c^2 = \frac{(9-8)^2}{8} + \frac{(7-8)^2}{8} = 0.25$$

Anche questa volta il risultato non è statisticamente significativo. Si tratta comunque di esempi costruiti su campioni molto piccoli per amore di semplicità.

Tale χ^2 ha 1 solo grado di libertà in quanto le frequenze attese sono date dalla media delle due frequenze osservate. Pertanto, se una frequenza osservata varia in un senso rispetto alla frequenza attesa, l'altra varierà in senso opposto. Quindi una sola frequenza osservata è libera di variare mentre l'altra è vincolata a variare in senso opposto per rispettare la media.

Tabella 2'2

La tabella 2x2 serve a valutare l'associazione tra due caratteri o a confrontare due proporzioni. Le due cose sono in realtà due aspetti dello stesso fenomeno. La tabella 2x2 si costruisce come una normale tabella a due entrate. Ogni carattere è scisso in due modalità che devono essere

1. mutualmente esclusive (senza alcuna sovrapposizione)
2. esaustive (comprendono o esauriscono tutte le possibilità)
3. indipendenti tra i soggetti (il fatto di trovare una modalità di un carattere in un soggetto non influisce sulla modalità presente nel soggetto successivamente campionato e in tutti gli altri).

Esempi di tabelle 2x2:

		Colore degli occhi	
		celeste	non-celeste
Colore dei capelli	biondi	24	6
	non-biondi	28	90

per rispondere al quesito: esiste una relazione tra colore degli occhi e colore dei capelli?

		Tipo di scuola	
		liceo 'A'	Liceo 'B'
Esiti scrutini	promossi o rimandati	450	640
	bocciati	110	60

per rispondere al quesito: la proporzione promossi/bocciati è uguale nelle due scuole?

		Recettore K	
		presente	assente
Recettore Q	presente	72	14
	assente	54	73

per rispondere al quesito: l'espressione del recettore K è associata all'espressione del recettore Q

Le tre tabelle rispondono a quesiti solo apparentemente diversi. In tutti i casi ci si chiede se ci sia una certa relazione tra le modalità delle due variabili o caratteri.

Se esiste una certa relazione, allora i rapporti o proporzioni o percentuali in tabella tendono a divergere: il $24/(24+28) = 0.46 = 46\%$ delle persone con occhi celesti sono anche bionde, mentre appena il $6/(6+90) = 0.06 = 6\%$ appena delle persone con occhi non-celesti sono bionde.

Se invece non c'è relazione, i rapporti tendono a coincidere.

Il test del χ^2 permette di verificare l'ipotesi nulla. Occorre a questo punto trovare le frequenze attese (dall'ipotesi nulla) che equalizzano le due proporzioni. Ciò è estremamente semplice. Basta, per ogni casella, calcolare il prodotto (totale riga \times totale colonna) diviso per il totale generale:

Frequenze osservate:

		Colore degli occhi		totali di riga
		celeste	non-celeste	
Colore dei capelli	biondi	24 (46%)	6 (6%)	30
	non-biondi	28 (54%)	90 (94%)	118
	totali di colonna	52	96	148 totale generale

Frequenze attese:

		Colore degli occhi		totali di riga
		celeste	non-celeste	
Colore dei capelli	biondi	$52 \times 30 / 148 = 10.5$ (20%)	$96 \times 30 / 148 = 19.5$ (20%)	30
	non-biondi	$52 \times 118 / 148 = 41.5$ (80%)	$96 \times 118 / 148 = 76.5$ (80%)	118
	totali di colonna	52	96	148 totale generale

Le frequenze attese trovate equalizzano le proporzioni nel senso che $10.5/52 = 19.5/96$ (a parte qualche lieve differenza dovuta all'arrotondamento) lasciando inalterati i totali marginali di riga e di colonna.

$$\begin{aligned}
 \chi^2 &= \frac{(24 - 10.5)^2}{10.5} + \frac{(6 - 19.5)^2}{19.5} + \frac{(28 - 41.5)^2}{41.5} + \frac{(90 - 76.5)^2}{76.5} = \\
 &= \frac{13.5^2}{10.5} + \frac{13.5^2}{19.5} + \frac{13.5^2}{41.5} + \frac{13.5^2}{76.5} = \\
 &= \frac{182.25}{10.5} + \frac{182.25}{19.5} + \frac{182.25}{41.5} + \frac{182.25}{76.5} = \\
 &= 17.35 + 9.35 + 4.39 + 2.38 = 33.47
 \end{aligned}$$

Il valore trovato 33.47 è ben superiore al valore critico di χ^2 con 1 grado di libertà sia per $\alpha=0.05$ che per $\alpha=0.01$. Pertanto si rigetta l'ipotesi nulla di indifferenza, concludendo che esiste una relazione tra colore dei capelli e colore degli occhi.

Si sarà notato come la differenza tra frequenza osservata e frequenza attesa sia pari a 13.5 in tutte le 4 caselle. Infatti, dovendo rispettare i totali marginali, se sottraiamo una certa quantità ad una casella dobbiamo aggiungere la stessa quantità alla casella adiacente, sia in verticale che in orizzontale. Per questo motivo la tabella 2x2, pur sviluppando quattro quozienti, ha 1 solo grado di libertà. In altre parole, una volta calcolata la frequenza attesa di una casella, le frequenze attese delle altre tre sono vincolate dal rispetto dei totali marginali, ed infatti il modo più semplice di calcolarle è per differenza.

Una volta che si sa come stanno le cose, è possibile utilizzare una formula che ci dà il χ^2 in un solo passaggio. Se chiamiamo le 4 caselle con a, b, c, d; i 4 totali marginali con (a+b) (c+d) (a+c) (b+d); il totale generale con $n=a+b+c+d$:

		Colore degli occhi		totali di riga
		celeste	non-celeste	
Colore dei capelli	biondi	a=24	b= 6	a+b=30
	non-biondi	c=28	d=90	c+d=118
	totali di colonna	a+c=52	b+d=96	n=a+b+c+d=148 totale generale

La formula immediata è:

$$\chi^2 = \frac{(ad - bc)^2 n}{(a+b)(c+d)(a+c)(b+d)}$$

che nel nostro caso darà:

$$\chi^2 = \frac{(24 \cdot 90 - 6 \cdot 28)^2 \cdot 148}{30 \cdot 118 \cdot 52 \cdot 96} = 33.23$$

33.23 corrisponde al 33.47 ottenuto in base al calcolo delle frequenze attese, con la differenza di qualche decimale dovuta agli arrotondamenti.

Tabelle m' n

E' possibile organizzare tabelle con più righe e più colonne (fatta salva la regola della mutua esclusività ed esaustività delle modalità ed indipendenza delle osservazioni). Le frequenze attese si calcolano con la solita formula: per ogni casella, totale di riga moltiplicato totale di colonna diviso totale generale. Non esistono formule semplificate.

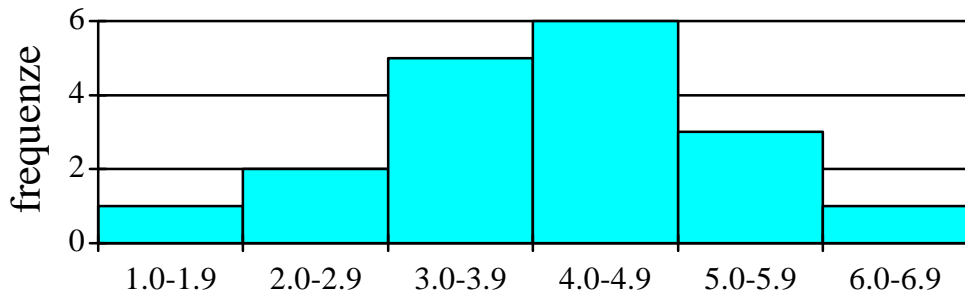
In generale le tabelle 2×2 e le altre tabelle $m \times n$ sono dette tabelle di contingenza. I gradi di libertà del χ^2 saranno $(n^\circ \text{ di righe}-1) \times (n^\circ \text{ di colonne}-1)$. Il ragionamento è lo stesso fatto sopra per la tabella 2×2 : il rispetto dei totali marginali fa sì che in ogni riga ed in ogni colonna un dato sia vincolato dal valore degli altri. Per cui in una riga di n dati, solo $n-1$ saranno liberi di variare. Idem, in una colonna di m dati, solo $m-1$ saranno liberi di variare. In totale i gradi di libertà saranno $(n-1)(m-1)$.

Esempio di tabella 3×3 :

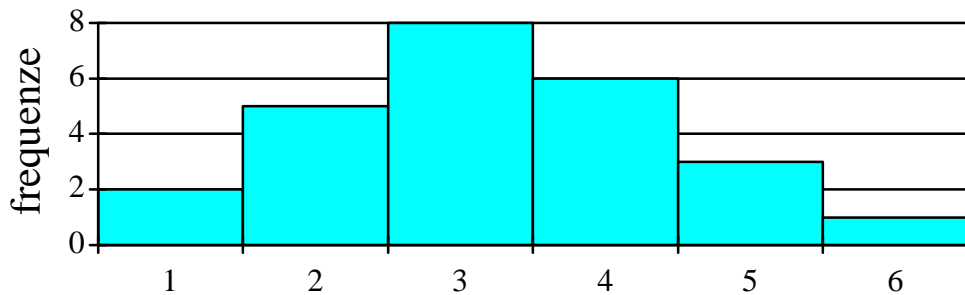
		Severità della malattia		
		lieve	moderata	grave
Quadro istopatologico	normale	13	6	2
	poco alterato	6	18	17
	molto alterato	2	12	24

Correzione di Yates per la continuità

A differenza delle classi di frequenza delle variabili continue, riferite a intervalli di scala, le classi di frequenza dei conteggi hanno come riferimento valori discreti centrali:

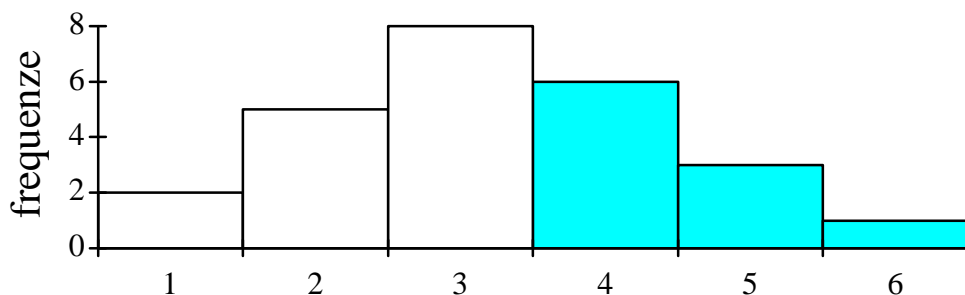


classi di frequenza riferite a intervalli di scala continua (variabile = misura)



classi di frequenza riferite a valori discreti centrali (variabile = conteggio)

Supponiamo ora di dover calcolare l'area di una certa porzione dell'istogramma. Ad esempio, l'area dell'istogramma comprendente le classi 4, 5 e 6 sarà:



Il confine tra le classi considerate e la restante parte di istogramma è dato dal valore nominale 4 meno $0.5 = 3.5$. Questa semplice operazione è detta correzione per la continuità ed è di fondamentale importanza quando l'area in gioco è piccola, come al livello delle code critiche della distribuzione. Per le tabelle 2×2 la cosiddetta correzione di Yates per la continuità consiste nella modifica:

$$c_c^2 = \sum \frac{(|\text{frequenza osservata} - \text{frequenza attesa}| - 0.5)^2}{\text{frequenza attesa}}$$

La lettera c sotto χ^2 sta appunto ad indicare la correzione per la continuità. Il valore del χ^2 con la correzione per la continuità è minore di quello non corretto. La correzione per la continuità è obbligatoria quando il campione è piccolo, ma la si può comunque applicare anche quando il campione è grande, nel qual caso il suo effetto sarà molto leggero. E' quindi buona abitudine utilizzarla sempre per le tabelle 2x2. La correzione per la continuità non va invece applicata alle tabelle m x n o comunque quando il χ^2 ha più di 1 grado di libertà. La formula semplificata del χ^2 con la correzione per la continuità per tabelle 2x2 è la seguente:

$$c_c^2 = \frac{(|ad - bc| - n/2)^2 n}{(a + b)(c + d)(a + c)(b + d)}$$

Test esatto di Fisher

Quando n è più piccolo di 40 e vi sono frequenze attese minori di 5 bisogna ricorrere ad un metodo più preciso, basato sul calcolo combinatorio, detto *metodo esatto di Fisher*. Tale metodo stima direttamente la probabilità di osservare la situazione rappresentata nella tabella più tutte quelle ancora più estreme. Per questo si parla anche di *probabilità esatta di Fisher*. Il procedimento è un po' complesso per essere spiegato in queste pagine, comunque ci proviamo. Esso è basato sul seguente ragionamento: se aumentiamo o diminuiamo la frequenza di una casella (qualsiasi) di una tabella 2×2 , tutte le altre frequenze dovranno insieme variare per mantenere gli stessi totali marginali. Come conseguenza la disproporzione tenderà in un senso a crescere e nell'altro senso a livellarsi. Supponiamo:

3	8
7	5

in cui $\mathbf{a/c}$ ($3/7=0.43$) è minore di $\mathbf{b/d}$ ($8/5=1.60$)

Se aggiungiamo 1 alla casella **a**, per rispettare i totali marginali, siamo costretti anche a togliere 1 alle caselle **b** e **c** e ad aggiungere 1 a **d**:

In pratica,

3+1	8-1
7-1	5+1

=

4	7
6	6

In tal modo i rapporti $\mathbf{a/c}$ e $\mathbf{b/d}$ tendono ad equalizzarsi. Infatti aggiungendo nuovamente 1 ad **a**, ecc. ecc., si raggiunge la condizione più bilanciata possibile, e cioè:

5	6
5	7

in cui $\mathbf{a/c}$ ($5/5=1.00$) è circa uguale a $\mathbf{b/d}$ ($6/7=0.86$) [di meglio non si può].

Ma se continuando ad aggiungere 1 ad **a**, ecc. ecc., si ottiene una tabella in cui i rapporti si distanziano nuovamente, questa volta in senso inverso.

6	5
4	8

in cui $\mathbf{a/c}$ ($6/4=1.50$) è maggiore di $\mathbf{b/d}$ ($5/8=0.63$).

La situazione estrema sarà quella in cui avremo zero in una casella:

10	1
0	12

Qui termina il gioco di aggiungere alla casella **a**. Non possiamo continuare in questo senso perché non si possono avere frequenze negative nella casella **c**!

FINE PRIMO TEMPO

Se invece nella tabella iniziale sottraiamo 1 ad **a** ed a **d**, e sommiamo 1 a **b** e **c**, si ottengono disproporzioni sempre maggiori:

3	8
7	5

2	9
8	4

1	10
9	3

0	11
10	2

La disproporzione così ottenuta è la più estrema possibile.

FINE SECONDO TEMPO

Ora, il metodo di Fisher calcola la probabilità di ottenere la tabella iniziale e tutte le tabelle via via più estreme nel senso opposto rispetto a quello che conduce alla condizione attesa dell'ipotesi nulla (**a/c = b/d**), utilizzando, per ciascuna di queste tabelle, la formula:

$$P = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!} \frac{1}{a!b!c!d!}$$

La somma di tali probabilità è la risposta del test. Si tratta di un calcolo un po' pesante che richiede l'uso di un calcolatore. Ma ha il grosso vantaggio di offrirci direttamente la probabilità, senza passare attraverso altre distribuzioni e tabelle.

Per questo il test si chiama test esatto di Fisher.

Nota: i fattoriali dei numeri interi (> 20) sono numeri giganteschi.

Es. $50! = 3.0414... \times 10^{64}$. Per questo, nel calcolo del test esatto di Fisher ci si avvale della trasformazione logaritmica che trasforma un fattoriale in una semplice sommatoria. Infatti

$$\text{Log}(4!) = \text{Log}(1 \times 2 \times 3 \times 4) = \text{Log}(1) + \text{Log}(2) + \text{Log}(3) + \text{Log}(4)$$

Tabella 2x2 per campioni appaiati (test di Mc Nemar)

Quando i 2 caratteri sono contemporaneamente presenti negli individui, nel senso che si manifestano come diversi comportamenti o risposte a diverse situazioni, allora possiamo configurare una classificazione in cui ogni soggetto si esprime per entrambe le condizioni:

		condizione-A1	
		condizione-B1	condizione-B2
condizione-A2	condizione-B1		
	condizione-B2		

Supponiamo che un certo numero di volontari abbiano in programma di bere, prima di andare a letto, una tazza di caffè o di tè in giorni successivi e che poi esprimano la condizione del prendere sonno associata all'assunzione di ciascuna bevanda. I dati di tale sperimentazione si riassumono nella seguente tabella 2x2 per dati appaiati:

		assunzione di caffè	
		sonno facile	sonno difficile
assunzione di tè	sonno facile		
	sonno difficile		

Notare la differenza rispetto alla tabella 2x2 tradizionale - per dati non appaiati - che sarebbe stata:

		assunzione di	
		caffè	tè
sonno	facile		
	difficile		

In quest'ultima tabella, ogni soggetto si può esprimere per una sola esperienza (o caffè o tè, con risposta di aver avuto sonno facile o difficile). In tal caso vi sarebbe una notevole perdita di informazione.

Torniamo alla tabella per dati appaiati. Occorre stabilire se c'è un diverso effetto tra caffè e tè sul prendere sonno. Le caselle **a** e **d** sono indifferenti. Tutto dipende

dalle altre due caselle: **c** e **b**. Infatti il χ^2 corretto per tabelle 2x2 di dati appaiati è dato da:

$$c_c^2 = \frac{(|b-c|-1)^2}{b+c}$$

Per i più curiosi, il test di McNemar verte sulla differenza tra le frequenze osservate **b** e **c**. L'ipotesi nulla, che sostiene l'indifferenza, vuole che le frequenze attese **b** e **c** siano uguali e precisamente pari a $(b+c)/2$. Per cui si può calcolare:

$$\chi_c^2 = \frac{\left(\left|b - \frac{b+c}{2}\right| - 0.5\right)^2}{\frac{b+c}{2}} + \frac{\left(\left|c - \frac{b+c}{2}\right| - 0.5\right)^2}{\frac{b+c}{2}} = \frac{\left(\left|\frac{b-c}{2}\right| - 0.5\right)^2 + \left(\left|\frac{c-b}{2}\right| - 0.5\right)^2}{\frac{b+c}{2}}$$

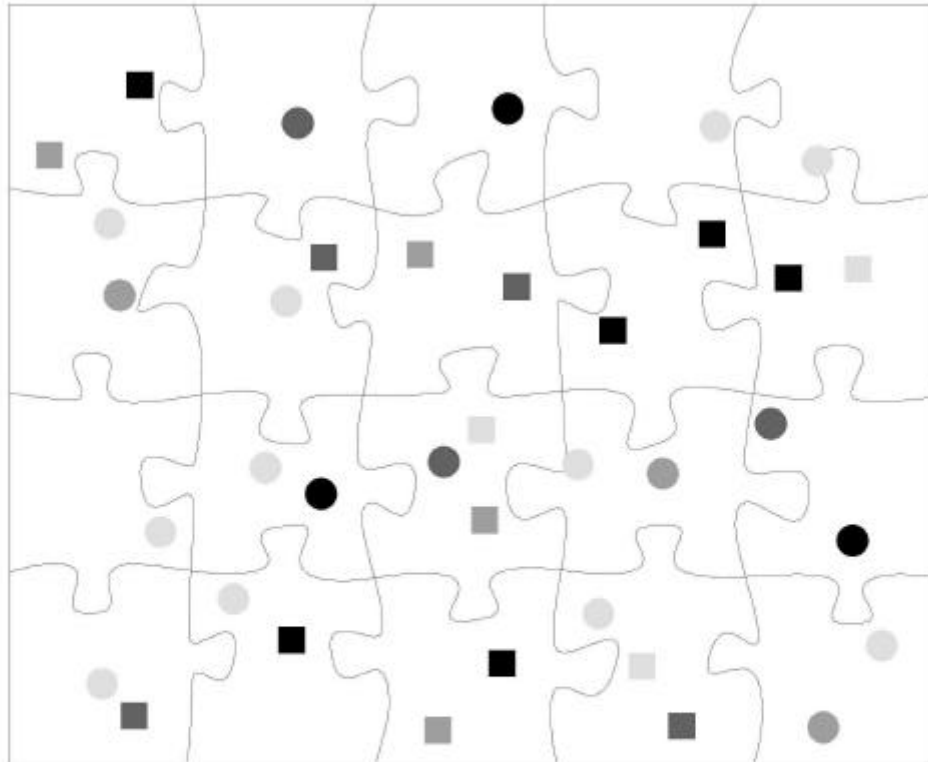
$$\text{ma poichè } \left|\frac{b-c}{2}\right| = \left|\frac{c-b}{2}\right|$$

possiamo semplificare

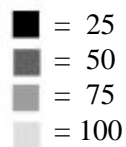
$$\chi_c^2 = \frac{2\left(\left|\frac{b-c}{2}\right| - 0.5\right)^2}{\frac{b+c}{2}} = \frac{2\left(\left|\frac{b-c}{2}\right| - \frac{1}{2}\right)^2}{\frac{b+c}{2}} = \frac{2\left[\frac{1}{2}(|b-c|-1)\right]^2}{\frac{b+c}{2}} = \frac{\frac{2}{4}(|b-c|-1)^2}{\frac{1}{2}(b+c)} = \frac{(|b-c|-1)^2}{(b+c)}$$



Esercizio



SCALA DEI TONI DI GRIGIO



Quante è la frequenza assoluta degli oggetti circolari?	
Quante è la frequenza percentuale degli oggetti circolari?	
Quale è la frequenza media di oggetti circolari per area?	
Quale è la frequenza assoluta delle aree contenenti solo oggetti circolari?	
Quale è la frequenza percentuale delle aree non contenenti alcun oggetto?	
Quale è la frequenza relativa delle aree contenenti solo oggetti quadrati?	
Quale il valore di grigio medio degli oggetti circolari?	