Università degli Studi di Padova
Facoltà di Psicologia,
L4, Psicometria, Modulo B

Verifica di ipotesi sulle frequenze

Dr. Marco Vicentini
marco.vicentini@unipd.it

Anno Accademico 2010 – 2011
Rev 20/04/2011

Sommario

- La distribuzione χ^2
- Tabelle di contingenza
  - Ipotesi di indipendenza
  - Gradi di libertà
- Test del Chi-Quadro
  - Test unidimensionale
  - Test bidimensionale

Dimensionalità dei test statistici

- *Test unidimensionale*, quando si deve decidere se una distribuzione di frequenze osservate sia compatibile con la distribuzione che ci si attende in base ad un certo modello teorico;
- *Test bidimensionale*, quando si studia il legame tra due variabili qualitative.

Distribuzione Chi-Quadro

Distribuzione Chi-Quadro

- La distribuzione χ^2 consente di affrontare i seguenti problemi:
  - confronto tra frequenze osservate e frequenze teoriche;
  - studio della relazione tra due variabili qualitative;
  - calcolo dell'intervallo di confidenza per la varianza.

Distribuzione Chi-Quadro

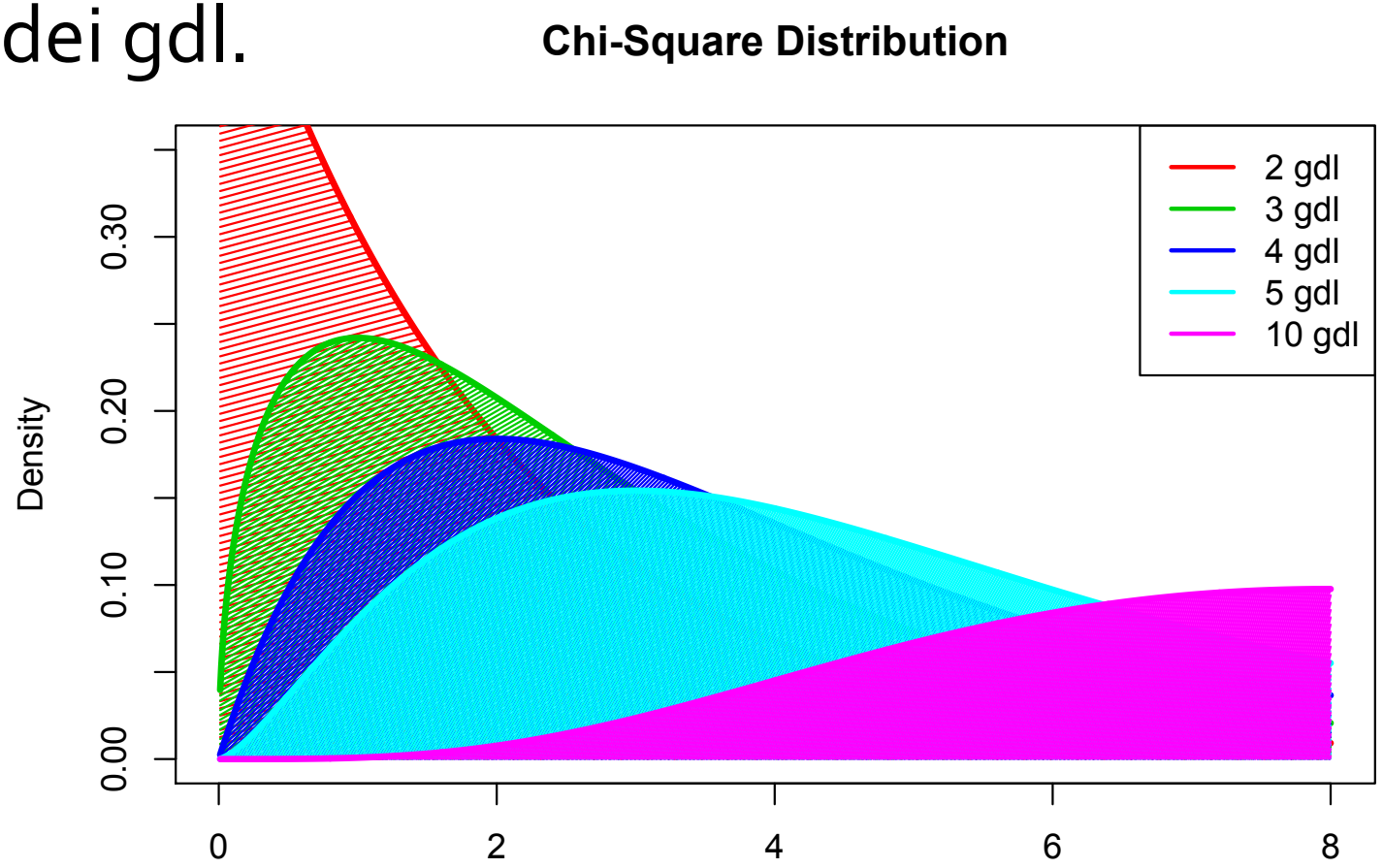
- Si tratta di una distribuzione continua con densità data dalla relazione:

$$f\left(\chi^2\right) = C_{(\nu)}\left(\chi^2\right)^{\left(\frac{\nu}{2}\right)^{-1}} e^{-\left(\frac{\chi^2}{2}\right)}$$

- Dove:
  - ν gradi di libertà della distribuzione
  - $C_{(\nu)}$ costante tale da assicurare che l'area delimitata sotto la curva sia pari a 1

  - * normalmente la funzione di densità della distribuzione chi-quadro viene definita partendo partendo dalla distribuzione Gamma

- La forma della distribuzione varia al variare dei gdl.



Proprietà

- la distribuzione χ^2 è una distribuzione continua, con valori compresi tra 0 e $+\infty$;
- al crescere dei gradi di libertà la curva tende ad assumere la forma della normale
  - gdl > 30-50;
- dato che la v.c. χ^2 è generata da una somma dei quadrati di *n* valori indipendenti di una v.c. normale standardizzata, quando tali valori non sono indipendenti è necessario stabilire le condizioni che li vincolano:
  - sottraendo ad *n* tali vincoli, si ottiene il numero di **gradi di libertà**, cioè il numero di valori, tra loro indipendenti, della v.c. normale standardizzata.

Tavola del Chi-Quadro

- Per ogni particolare gdl, il valore ottenuto di χ^2 è significativo ad un dato livello di significatività se è uguale o maggiore dei valori critici presentati.
- Per esempio, con α=0.05 e gdl=6, il valore critico è 12.59

gdl	a = 0.10	a = 0.05	a = 0.01
1	2,71	3,84	6,63
2	4,61	5,99	9,21
3	6,25	7,81	11,34
4	7,78	9,49	13,28
5	9,24	11,07	15,09
6	10,64	12,59	16,81
7	12,02	14,07	18,48
8	13,36	15,51	20,09
9	14,68	16,92	21,67
10	15,99	18,31	23,21
11	17,28	19,68	24,72
12	18,55	21,03	26,22
13	19,81	22,36	27,69
14	21,06	23,68	29,14
15	22,31	25,00	30,58
16	23,54	26,30	32,00
17	24,77	27,59	33,41
18	25,99	28,87	34,81
19	27,20	30,14	36,19
20	28,41	31,41	37,57

Ipotesi di indipendenza

Indipendenza e indipendenza media

- L'analisi bivariata ha come scopo lo studio della relazione tra due variabili.

- Nel caso di assenza di relazione parliamo di **indipendenza** tra due variabili.
  - In una tabella di contingenza si ha indipendenza tra le due variabili *se ciascuna distribuzione parziale è uguale alla corrispondente distribuzione marginale*.

Indipendenza tra due variabili e tabelle di contingenza

- Nella teoria della probabilità, due eventi A e B sono detti indipendenti se la seguente relazione si verifica:

  - $P(A \cap B) = P(A)\,P(B)$

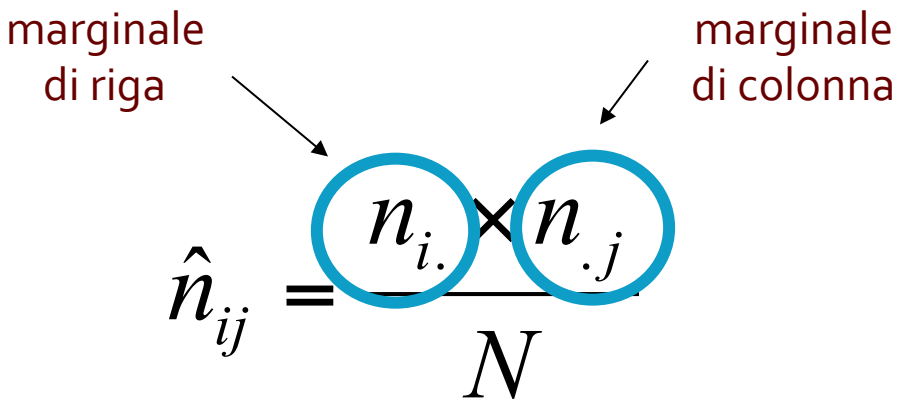
Indipendenza tra due variabili e tabelle di contingenza

- La probabilità teorica, P(A), può essere approssimata dalla probabilità empirica corrispondente al rapporto tra la frequenza con cui si verifica l'evento A e la grandezza dello spazio campione Ω.

  - $P(A) = |A| / |\Omega|$

Indipendenza tra due variabili e tabelle di contingenza

- Conoscendo le frequenze marginali, possiamo calcolare i valori di cella che dovremmo aspettarci nel caso di completa indipendenza tra le due variabili.
- Per una generica cella n_{ij} la frequenza teorica in base all'ipotesi di indipendenza si calcola come:



Indipendenza tra due variabili e tabelle di contingenza

- Questo stesso risultato si ottiene:

$$\hat{n}_{ij} = P\left(A_i \cap B_j\right) \times N = P\left(A_i\right) \times P\left(B_j\right) \times N$$

$$= \frac{n_{i.}}{N} \times \frac{n_{.j}}{N} \times N = \frac{n_{i.} \times n_{.j}}{N}$$

Gradi di libertà

Gradi di libertà

- Data una tabella di frequenze di dimensioni *n*, è possibile riconoscere come i gradi di libertà della tabella siano espressi dalla formula:

  - $df = (n - 1)$

Gradi di libertà

- Prendiamo una tabella di frequenze con 3 dati:

10	15	20	**45**

- Data la somma, si può osservare come i gradi di libertà siano:
  - (3 – 1) = 2

			45

- Infatti la tabella risulta univocamente determinata in presenza di due valore determinati presenti in qualsiasi cella.

Gradi di libertà

- Data una tabella di contingenza di dimensioni r x c, è possibile riconoscere come i gradi di libertà della tabella siano espressi dalla formula:

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

Gradi di libertà

- Prendiamo una tabella di dimensione 2x2 con i seguenti dati fittizi:

10	15	25
20	25	45
30	40	70

- Dati i marginali di riga e di colonna si può osservare come i gradi di libertà siano:
  - $(2-1)(2-1) = 1$

		25
		45
30	40	70

- Infatti la tabella risulta univocamente determinata in presenza di un valore determinato in una qualsiasi cella.

Test con un solo campione

Test del Chi-Quadro Unidimensionale

- Permette un confronto tra le *frequenze osservate* e le *frequenze attese* con la formula

$$\chi^2_{(n-1)} = \sum_{i=1}^{n} \frac{\left(f_i - F_i\right)^2}{F_i}$$

- Dove:

χ^2
- è il numero di classi della variabile osservata

n
- è la numerosità delle classi, con *n - 1* gdl

f_i
- sono le frequenze osservate della classe *i*-ma

F_i
- sono le frequenze attese della classe *i*-ma

Esempio 1

- Gli appassionati di corse di cavalli ritengono che, in una corsa disputata su una pista circolare, i cavalli disposti nelle posizioni interne siano più avvantaggiati.
- In una corsa di 8 cavalli, la posizione 1 è la più vicina alla staccionata, la 8 la più esterna.
  Vengono osservati 18 giorni di corse, per un totale di 144 gare e, in tabella si riportano i vincitori secondo il numero di gara.

corsie	1	2	3	4	5	6	7	8	**TOTALE**
n.ro di vittorie	29	19	18	25	17	10	15	11	144

- *tratto da: Siegel (1956)*

Esempio 1

- formulazione delle ipotesi
  - H_0: il numero dei vincitori è ripartito equamente tra le 8 posizioni di partenza, non vi è una posizione avvantaggiata;
  - H_1: le posizioni di partenza non danno le stesse possibilità di vittoria.
- In pratica, se è vera H_0 dovremmo osservare nelle varie celle lo stesso numero di vincitori e cioè avere per ciascuna cella la frequenza attesa:

$$\frac{144}{8} = 18$$

Esempio 1

- calcolo del valore χ^2

corsie	*f*	F	*(f - F)*	*(f - F)²*	*(f - F)²/F*
1	29	18	11	121	6,72
2	19	18	1	1	0,06
3	18	18	0	0	0,00
4	25	18	7	49	2,72
5	17	18	-1	1	0,06
6	10	18	-8	64	3,56
7	15	18	-3	9	0,50
8	11	18	-7	49	2,72
TOT	144	144	0	294	16,33

$$\chi^2_{(7)} = 16.33$$

Esempio 1

- Determinazione dei valori critici
  - Fissato un livello di significatività $\alpha = 0.01$, con 7 gdl si ottiene il valore di 18.48.
- Decisione
  - Poiché $\chi^2_{cal} = 16.33$ è minore del valore critico $\chi^2_c = 18.48$, non possiamo rigettare l'ipotesi H_0.
- Conclusione
  - Con un livello di probabilità del 1%, non possiamo affermare che chi parte nelle posizioni interne sia avvantaggiato.

Test con due variabili

Test del Chi-Quadro Bidimensionale

- Permette un confronto tra le *frequenze osservate* e le *frequenze attese* con la formula

$$\chi^2_{(r-1)(c-1)} = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(f_{ij} - F_{ij}\right)^2}{F_{ij}}$$

- Dove:

χ^2
- è il numero di classi della variabile osservata

r, c
- è il numero di righe e colonne della tabella, con (r-1)(c-1) gdl

f_{ij}
- sono le frequenze osservate della cella *i,j*

F_{ij}
- sono le frequenze attese della cella *i,j*

Esempio 2

- A 520 studenti degli ultimi tre anni del corso di laurea in Psicologia viene richiesto di esprimere la propria opinione in merito ad una proposta di cambiamento del corso di laurea.
- Le risposte vengono classificate in tre categorie:
  - A) favorevoli;
  - B) sfavorevoli;
  - C) incerti o senza opinione.
- Si vuole sapere se l'atteggiamento verso la proposta sia legato alla permanenza nel corso di laurea.

	atteggiamento			
anno	*A*	*B*	*C*	**TOTALI**
III	36	72	12	120
IV	98	64	14	176
V	126	90	8	224
TOTALI	260	226	34	520

Esempio 2

- Formulazione delle ipotesi
  - H_0: non vi è relazione tra atteggiamento verso il cambiamento e l'anno di corso;
  - H_1: i soggetti dei vari anni di corso hanno un atteggiamento diverso nei confronti del cambiamento.

  - Formalmente

$$H_0 : F_{1.} = F_{2.} = F_{3.}$$

Esempio 2

- calcolo del valore χ^2
  - frequenze osservate:

	atteggiamento			
anno	*A*	*B*	*C*	**TOTALI**
III	36	72	12	120
IV	98	64	14	176
V	126	90	8	224
TOTALI	260	226	34	520

  - frequenze attese:

	atteggiamento			
anno	*A*	*B*	*C*	**TOTALI**
III	60	52,2	7,8	120
IV	88	76,5	11,5	176
V	112	97,4	14,6	224
TOTALI	260	226	34	520

  - La tabella è calcolata utilizzando la formula:

$$F_{1,1} = \frac{120 \cdot 260}{520} = 60$$

Esempio 2

- calcolo del valore χ^2

$$\chi^2 = \frac{(36-60)^2}{60} + \frac{(72-52.2)^2}{52.2} + \frac{(12-7.8)^2}{7.8} +$$
$$+ \frac{(98-88)^2}{88} + \frac{(64-76.5)^2}{76.5} + \frac{(14-11.5)^2}{11.5} +$$
$$+ \frac{(126-112)^2}{112} + \frac{(90-97.4)^2}{97.4} + \frac{(8-14.6)^2}{14.6} = 28.39$$

$$\chi^2_{(4)} = 28.39$$

Esempio 2

- Determinazione dei valori critici
  - Con un livello di significatività α = 0,05, e gradi di libertà (3-1)(3-1) = 4, il valore critico che si osserva nelle tabelle è χ^2=9.49.
- Decisione
  - Poiché χ^2_{cal} = 28.39 è maggiore del valore critico, dobbiamo rigettare l'ipotesi H_0.
- Conclusione
  - L'atteggiamento degli studenti verso la proposta di cambiamento è legato all'anno di corso.