Statistica

Capitolo 13

Test sulla Bontà di Adattamento e
Tabelle di Contingenza

Obiettivi del Capitolo

Dopo aver completato il capitolo, sarete in grado di:

- Usare il test sulla bontà di adattamento per determinare se i dati sono generati da una specifica distribuzione
- Effettuare test di normalità
- Costruire una tabella per l'analisi della contingenza ed effettuare un test chi-quadrato di associazione

Test sulla Bontà di Adattamento

- I dati campionari si adattano ad una distribuzione ipotizzata?
  - Esempi:
    - I risultati campionari si adattano a specifiche probabilità attese?
    - Il numero di chiamate al supporto tecnico è lo stesso per tutti i giorni della settimana? (i.e., le chiamate hanno una distribuzione uniforme?)
    - Le misurazioni relative ad un processo di produzione seguono una distribuzione normale?

Test sulla Bontà di Adattamento

(continuazione)

- Il numero di chiamate al supporto tecnico è lo stesso per tutti i giorni della settimana? (i.e., le chiamate hanno una distribuzione uniforme?)
  - Per ciascun giorno della settimana, raccogliamo i dati campionari relativi a 10 giorni:

	Somma delle chiamate per giorno della settimanai:
Lunedi	**290**
Martedi	**250**
Mercoledi	**238**
Giovedi	**257**
Venerdi	**265**
Sabato	**230**
Domenica	**192**
	Σ = 1722

Logica del Test sulla Bontà di Adattamento

- Se le chiamate **sono** distribuite uniformemente, le 1722 chiamate dovrebbero essere equamente divise fra i 7 giorni:

$$\frac{1722}{7} = 246 \quad \text{chiamate attese per giorno se distribuite uniformemente}$$

- Test chi-quadrato sulla bontà di adattamento: test per vedere se i risultati campionari sono consistenti con i risultati attesi

Frequenze Osservate vs. Attese

	Osservate O_i	Attese E_i
Lunedi	290	246
Martedi	250	246
Mercoledi	238	246
Giovedi	257	246
Venerdi	265	246
Sabato	230	246
Domenica	192	246
TOTALE	1722	1722

Statistica Test Chi-Quadrato

H_0: La distribuzione delle chiamate è uniforme rispetto ai giorni della settimana

H_1: La distribuzione delle chiamate non è uniforme

- La statistica test è

$$\chi^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} \quad (\text{dove g.d.l.} = K - 1)$$

dove:

K = numero di categorie

O_i = frequenza osservata per la categoria i

E_i = frequenza attesa per la categoria i

Regione di Rifiuto

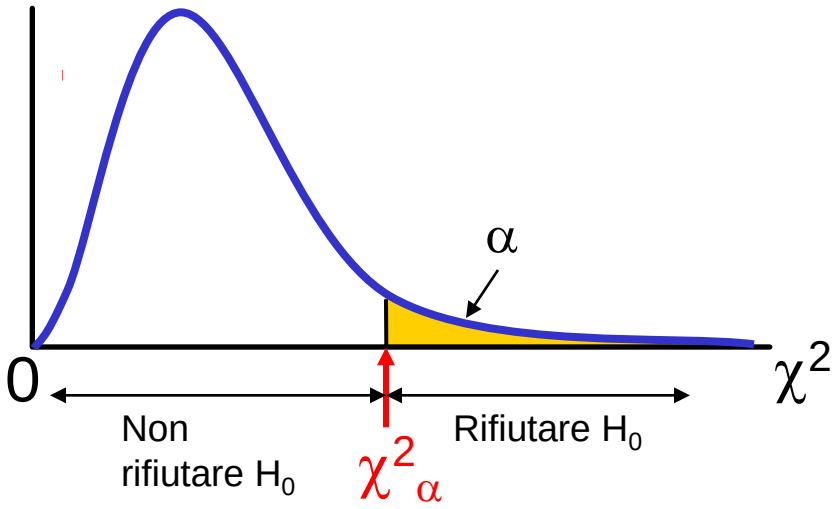
H_0: La distribuzione delle chiamate è uniforme rispetto ai giorni della settimana

H_1: La distribuzione delle chiamate non è uniforme

$$\chi^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i}$$

- Rifiutare H_0 se $\chi^2 > \chi^2_\alpha$

(con K – 1 gradi di libertà)



Statistica Test Chi-Quadrato

H_0: La distribuzione delle chiamate è uniforme rispetto ai giorni della settimana

H_1: La distribuzione delle chiamate non è uniforme

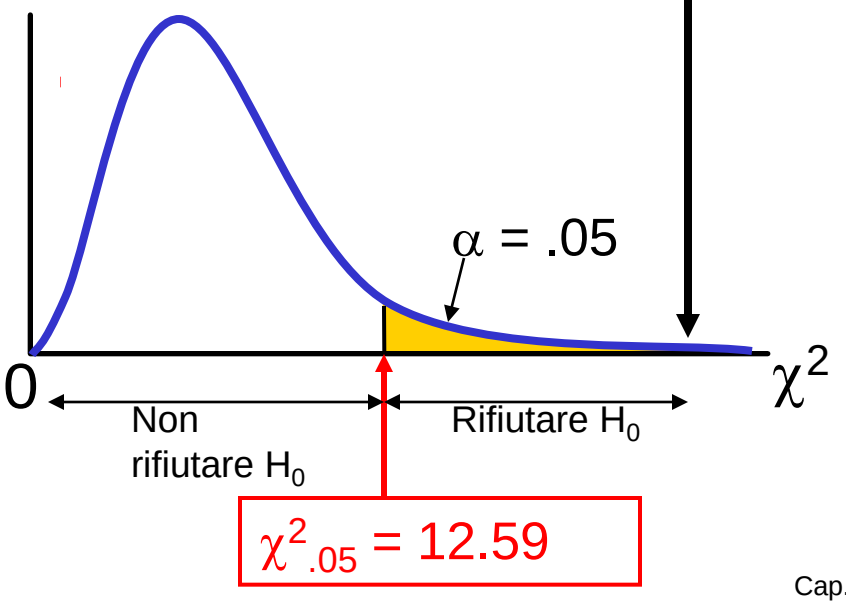
$$\chi^2 = \frac{(290-246)^2}{246} + \frac{(250-246)^2}{246} + \ldots + \frac{(192-246)^2}{246} = 23.05$$

K– 1 = 6 (7 giorni della settimana quindi usiamo 6 gradi di libertà):

$\chi^2_{.05} = 12.59$

Conclusione:

$\chi^2 = 23.05 > \chi^2_{\alpha} = 12.59$ quindi **rifiutiamo H_0** e concludiamo che la distribuzione non è uniforme



Test sulla Bontà di Adattamento, Parametri della Popolazione non Noti

Idea:

- Verificare se i dati hanno una specifica distribuzione (per esempio binomiale, Poisson, o normale) . . .
- . . . senza assumere che i parametri della popolazione siano noti
- Usiamo i dati campionari per stimare i parametri della popolazione che non sono noti

Test sulla Bontà di Adattamento, Parametri della Popolazione non Noti

(continuazione)

- Supponiamo che l'ipotesi nulla specifichi la probabilità per le categorie che dipendono dalla stima (dai dati) di m parametri non noti della popolazione
- Il test sulla bontà di adattamento appropriato coincide con quello fornito precendentemente . . .

$$\chi^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i}$$

- . . . tranne che nel numero di gradi di libertà per la variabile Chi-quadrato che è

$$\text{Gradi di Libertà} = (K - m - 1)$$

- Dove K è il numero di categorie

Test di Normalità

- In statistica è comune l'assunzione che i dati abbiano una distribuzione normale
- La normalità è stata verificata precedentemente
  - Normal probability plot (Capitolo 6)
- Qui sviluppiamo un test chi-quadrato

Test di Normalità

(continuazione)

- Con i dati campionari si possono stimare due parametri della popolazione:

$$\text{Indice di Asimmetria} = \frac{\sum_{i-1}^{n}(x_i - \bar{x})^3}{ns^3}$$

$$\text{Indice di Curtosi} = \frac{\sum_{i-1}^{n}(x_i - \bar{x})^4}{ns^4}$$

- Per una distribuzione normale,

Indice di Asimmetria = 0

Indice di Curtosi = 3

Test di Normalità di Jarque-Bera

- Consideriamo l'ipotesi nulla che la distribuzione della popolazione sia normale
- Il Test di Normalità di Jarque-Bera è basato sulla vicinanza dell'asimmetria campionaria a 0 e della curtosi campionaria a 3
- La statistica test è

$$B = n\left[\frac{(\text{Asimmetri a})^2}{6} + \frac{(\text{Curtosi} - 3)^2}{24}\right]$$

- Quando il numero di osservazioni campionarie cresce, questa statistica assume una distribuzione Chi-quadrato con 2 gradi di libertà
- L'ipotesi nulla viene rifiutata per valori grandi della statistica test

Test di Normalità di Jarque-Bera

(continuazione)

- L'approssimazione alla Chi-quadrato è molto buona solo per campioni veramente grandi
- Se il campione non è grande, la statistica test di Jarque-Bera è confrontata con i valori significativi della tabella 13.7 del libro

Ampiezza campionaria n	Livello di significatività 10%	Livello di significatività 5%	Ampiezza campionaria n	Livello di significatività 10%	Livello di significatività 5%
20	2.13	3.26	200	3.48	4.43
30	2.49	3.71	250	3.54	4.61
40	2.70	3.99	300	3.68	4.60
50	2.90	4.26	400	3.76	4.74
75	3.09	4.27	500	3.91	4.82
100	3.14	4.29	800	4.32	5.46
125	3.31	4.34	∞	4.61	5.99
150	3.43	4.39			

Esempio: Test di Normalità di Jarque-Bera

- È stata registrata, per 200 giorni selezionati a caso, la temperatura media giornaliera. L'asimmetria campionaria è risultata 0.232 e la curtosi campionaria 3.319
- Verificare l'ipotesi nulla che la vera distribuzione sia normale

$$B = n\left[\frac{(\text{Asimmetri a})^2}{6} + \frac{(\text{Curtosi} - 3)^2}{24}\right] = 200\left[\frac{(0.232)^2}{6} + \frac{(3.319-3)^2}{24}\right] = 2.642$$

- Dalla tabella 13.7 il valore critico al 10% per n = 200 è 3.48, quindi non ci sono sufficienti evidenze per rifiutare l'ipotesi che la popolazione sia normale

Tabelle di Contingenza

Tabelle di Contingenza

- Usate per classificare le osservazioni campionarie secondo due caratteristiche
- Anche chiamate tabelle cross-classification o cross-tabulation
- Assumiamo ci siano *r* categorie per la caratteristica A e *c* categorie per la caratteristica B
  - Allora ci sono (*r* x *c*) possibili classificazioni

Tabella di Contingenza *r x c*

	Caratteristica B				
Caratteristica A	**1**	**2**	**. . .**	**C**	**Totali**
1	O_{11}	O_{12}	**...**	O_{1c}	R_1
2	O_{21}	O_{22}	**...**	O_{2c}	R_2
.	**.**	**.**	**...**	**.**	**.**
.	**.**	**.**	**...**	**.**	**.**
.	**.**	**.**	**...**	**.**	**.**
r	O_{r1}	O_{r2}	**...**	O_{rc}	R_r
Totali	C_1	C_2	**...**	C_c	n

Test di Associazione

- Considera *n* osservazioni tabulate in una tabella di contingenza *r x c*
- Denotiamo con O_{ij} il numero di osservazioni nella cella che corrisponde alla i^{ma} riga e j^{ma} colonna
- L'ipotesi nulla è

H_0 : Assenza di associazione fra le due caratteristiche nella popolazione

- L'appropriato test è un test chi-quadrato con (*r*-1)(*c*-1) gradi di libertà

Test di Associazione

(continuazione)

- Siano R_i e C_j i totali per riga e per colonna
- Il numero atteso di osservazioni nella cella che corrisponde alla riga i e alla colonna j, se H_0 è vera, è

$$E_{ij} = \frac{R_i C_j}{n}$$

- Un test di associazione a livello di significatività α è basato sulla distribuzione Chi-quadrato e sulla seguente regola di decisione

$$\text{Rifiutare } H_0 \text{ se } \chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} > \chi^2_{(r-1)c-1),\alpha}$$

Esempio: Tabella di Contingenza

Essere mancini vs. Sesso

- Mano Dominante: Sinistra vs. Destra
- Sesso: Maschio vs. Femmina

H_0: Assenza di associazione tra mano dominante e sesso

H_1: La mano dominante non è independente dal sesso

Esempio: Tabella di Contingenza

(continuazione)

Risultati campionari organizzati in una tabella di contingenza:

Dimensione campione $n = 300$:

120 Femmine, 12 erano mancine

180 Maschi, 24 erano mancini

Sesso	Mano dominante		
	Sinistra	Destra	
Femmina	12	108	120
Maschio	24	156	180
	36	264	300

Logica del Test

H_0: Assenza di associazione tra mano dominante e sesso

H_1: La mano dominante non è independente dal sesso

- Se H_0 è vera, allora la proporzione di donne mancine dovrebbe coincidere con la proporzione di uomini mancini
- Le due proporzioni precedenti dovrebbero coincidere con la proporzione generale di persone mancine

Calcolo delle Frequenze Attese

120 Femmine, 12 erano mancine

180 Maschi, 24 erano mancine

In generale:

P(mancino) = 36/300 = .12

Se non c'è associazione, allora

P(Mancino | Femmina) = P(Mancino | Maschio) = .12

Quindi ci aspetteremmo che il 12% delle 120 femmine e il 12% dei 180 maschi siano mancini…

i.e., ci aspetteremmo (120)(.12) = 14.4 femmine mancine
(180)(.12) = 21.6 maschi mancini

Calcolo delle Frequenze Attese

(continuazione)

- Frequenza attesa delle celle:

$$E_{ij} = \frac{R_i C_j}{n} = \frac{(\text{totale } i^{ma} \text{ Riga})(\text{totale } j^{ma} \text{ Colonna})}{\text{Dimensione Totale del campione}}$$

Esempio:

$$E_{11} = \frac{(120)(36)}{300} = 14.4$$

Frequenze
Osservate vs. Attese

Frequenze osservate vs. frequenze attese:

Sesso	Mano dominante		
	Sinistra	Destra	
Femmina	Osservate = 12 Attese = 14.4	Osservate = 108 Attese = 105.6	120
Maschio	Osservate = 24 Attese = 21.6	Osservate = 156 Attese = 158.4	180
	36	264	300

Statistica Test Chi-Quadrato

La statistica test chi-quadrato è:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad \text{con} \quad g.d.l. = (r-1)(c-1)$$

dove:

O_{ij} = frequenza osservata nella cella (i, j)

E_{ij} = frequenza attesa nella cella (i, j)

r = numero di righe

c = numero di colonne

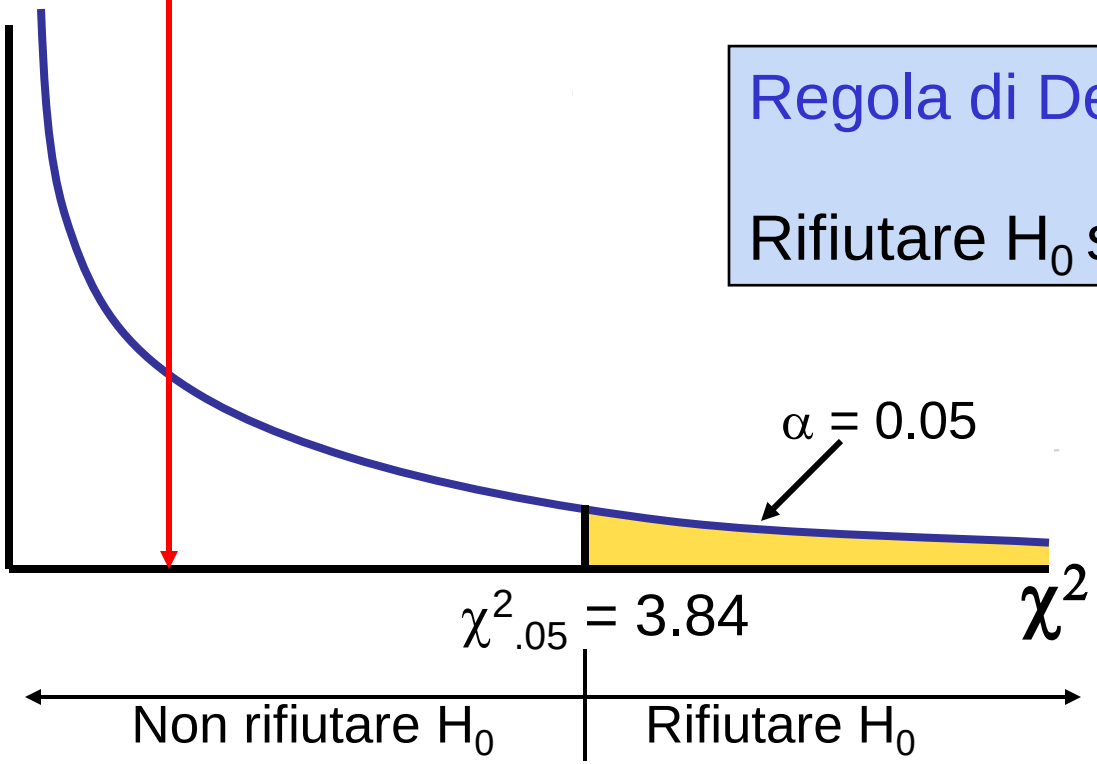
Frequenze Osservate vs. Attese

Sesso	Mano dominante		
	Sinistra	Destra	
Femmina	Osservate = 12 Attese = 14.4	Osservate = 108 Attese = 105.6	120
Maschio	Osservate = 24 Attese = 21.6	Osservate = 156 Attese = 158.4	180
	36	264	300

$$\chi^2 = \frac{(12-14.4)^2}{14.4} + \frac{(108-105.6)^2}{105.6} + \frac{(24-21.6)^2}{21.6} + \frac{(156-158.4)^2}{158.4} = 0.7576$$

Test di Associazione

$\chi^2 = 0.7576$ con g.d.l. $= (r-1)(c-1) = (1)(1) = 1$



Regola di Decisione:

Rifiutare H_0 se $\chi^2 > 3.84$

Poiché $\chi^2 = 0.7576 < 3.84$ non rifiutiamo H_0 e concludiamo che sesso e mano dominante non sono associati

Riepilogo del Capitolo

- Usato il test chi-quadrato sulla bontà di adattamento per determinare se i dati campionari si adattano a specifiche distribuzioni di probabilità
- Effettuati test sulla bontà di adattamento quando i parametri della popolazione non sono noti
- Verificata la normalità usando il test di Jarque-Bera
- Usate tabelle di contingenza per effettuare un test chi-quadrato di associazione
  - Per ogni cella confrontate le frequenze osservate con le frequenze attese