

TEST D'IPOTESI

• Test per l'assenza di associazione tra due caratteri:
Chi-Quadrato

Esempio motivante: Disturbi respiratori all'età di 14 anni e bronchiti prima dei 5 anni (Holland et al. 1978)

OBIETTIVO : valutare se vi è associazione tra bronchite nell'infanzia e disturbi respiratori (tosse).

Segnalazione di disturbi respiratori (tosse) durante il giorno e la notte all'età di 14 anni e di bronchiti prima dei 5 anni (Holland et al. 1978).

I **DATI OSSERVATI** possono essere rappresentati nella tabella di contingenza:

| | Bronchite prima dei 5 anni | | Totale |
|----------|----------------------------|------|--------|
| | si | no | |
| Tosse | 26 | 44 | 70 |
| No tosse | 247 | 1002 | 1249 |
| Totale | 273 | 1046 | 1319 |

Test per l'associazione tra due caratteri

| | Bronchite prima dei 5 anni | | Totale |
|----------|----------------------------|------|--------|
| | si | no | |
| Tosse | 26 | 44 | 70 |
| No tosse | 247 | 1002 | 1249 |
| Totale | 273 | 1046 | 1319 |

Sistema di Ipotesi :

H_0 : NON vi è associazione tra bronchite entro i 5 anni e disturbi respiratori (tosse) a 14 anni

H_1 : Vi è associazione tra bronchite entro i 5 anni e disturbi respiratori (tosse) a 14 anni

I metodi per la verifica di ipotesi visti sino ad ora si basano su

1. una statistica test "sensibile" ad indicare delle deviazioni da H_0 (z o t)
2. la distribuzione di tale statistica sotto H_0 (Gaussiana, t di Student)

In base a queste considerazioni

- si calcola il valore della statistica nel campione
- si decide se è verosimile che tale valore provenga da H_0 confrontandolo con una zona di rifiuto costruita sulla distribuzione al punto 2.

Nella verifica di ipotesi sulla **ASSOCIAZIONE** la statistica test si ottiene **confrontando** le tabelle di contingenza

osservata nel campione ed attesa sotto H_0

Costruzione della tabella attesa sotto Ho

Tabella di contingenza
OSSERVATA

| | Y | N | Totale |
|----------|-----|------|--------|
| Tosse | 26 | 44 | 70 |
| No tosse | 247 | 1002 | 1249 |
| Totale | 273 | 1046 | 1319 |

Tabella di contingenza
ATTESA sotto Ho

Se la probabilità di segnalazione di tosse fosse la stessa:

- quante segnalazioni attendo per chi ha avuto la bronchite?
- quante per chi non la ha avuta?

La stima della probabilità di "tosse" indipendentemente dall'esperienza di bronchite (sotto Ho) è $70/1319 = 0.053$

Sotto Ho attendiamo

$$273 \cdot 0.053 = 14.5 \text{ segnalazioni in "Y"}$$

$$1046 \cdot 0.053 = 55.5 \text{ segnalazioni in "N"}$$

Costruzione della tabella attesa sotto Ho

Tabella di contingenza
OSSERVATA

| | Y | N | Totale |
|----------|-----|------|--------|
| Tosse | 26 | 44 | 70 |
| No tosse | 247 | 1002 | 1249 |
| Totale | 273 | 1046 | 1319 |

Tabella di contingenza
ATTESA sotto Ho

| | Y | N | Totale |
|----------|--------------------|---------------------|--------|
| Tosse | 14.5 | 55.5 | 70 |
| No tosse | 273-14.5 =258.5 | 1046-55.5 =990.5 | 1249 |
| Totale | 273 | 1046 | 1319 |

La frequenza attesa di ciascuna cella si ottiene come
[(tot. riga) · (tot. colonna)] / (tot. generale)

Il confronto tra le freq. nelle tabelle OSSERVATA vs ATTESA sotto Ho, ci dà un'idea di quanto la realtà sia compatibile con Ho

Costruzione della tabella attesa sotto Ho

Tabella di contingenza
OSSERVATA

| | Y | N | Totale |
|----------|-----|------|--------|
| Tosse | 26 | 44 | 70 |
| No tosse | 247 | 1002 | 1249 |
| Totale | 273 | 1046 | 1319 |

Tabella di contingenza
ATTESA sotto Ho

Se tosse e bronchite fossero indipendenti:
quante segnalazioni attendo per chi ha avuto la bronchite?

Come calcolo $E(n_{ij})$?

Dalla definizione di indipendenza ($T \perp Y$)

$$E(n_{ij})/n = P(T \cap Y) = P(T)P(Y) = (n_i/n) * (n_j/n)$$

Se moltiplico entrambi i membri per n ottengo $\rightarrow E(n_{ij}) = n_i * n_j / n$

Da cui $n_{ty} = 70 * 273 / 1319 = 14.5$

Confronto tra tabella osservata ed attesa sotto Ho

Statistica Test

$$X^2 = \sum_{\text{celle della tabella di contingenza}} \frac{(\text{freq. oss.} - \text{freq. att.})^2}{\text{freq. att.}} = \sum \frac{(O - E)^2}{E}$$

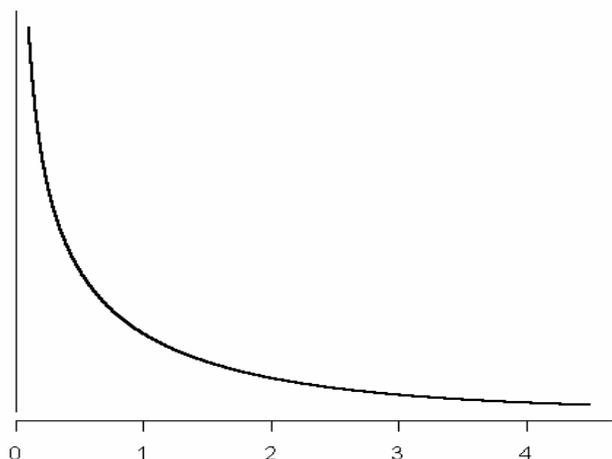
1. La statistica X^2 è "sensibile" ad indicare delle deviazioni da Ho, più è grande meno è verosimile che Ho sia vera
2. Sotto Ho $X^2 \sim$ Chi-Quadrato con 1 grado di libertà ($X^2(1)$)

Il valore della statistica nel campione è

$$X^2 = \frac{(26-14.5)^2}{14.5} + \frac{(44-55.5)^2}{55.5} + \frac{(247-258.5)^2}{258.5} + \frac{(1002-990.5)^2}{990.5} = 12.18$$

12.18 andrà confrontato con la zona di rifiuto unilaterale costruita sulla distribuzione Chi-Quadrato con 1 grado di libertà

DISTRIBUZIONE Chi-quadrato 1 g.d.l



Si dimostra che $X^2(1) \sim Z^2$ dove $Z \sim N(0,1)$

$$x^2_{\alpha}(1) = (z_{\alpha/2})^2 \quad \text{quindi} \quad X^2_{0.05}(1) = (1.96)^2 = 3.84$$

Test Chi-quadrato

Sotto H_0 , X^2 si distribuisce secondo una Chi-Quadrato con 1 grado di libertà

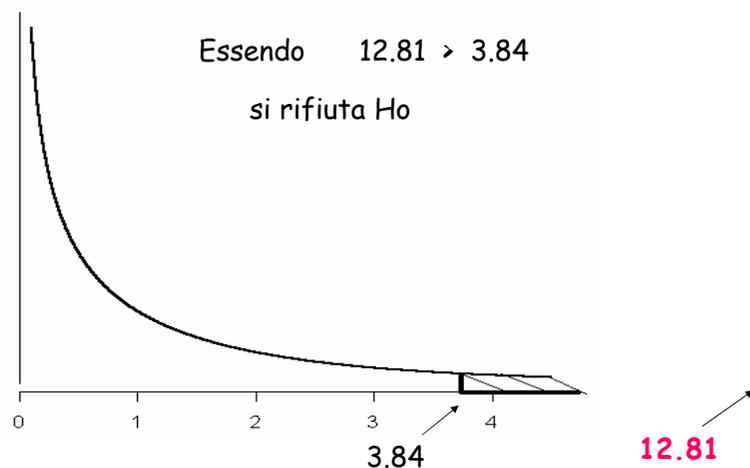
E' verosimile che 12.18 provenga da tale distribuzione ?

- Si determina una zona di rifiuto unilaterale con una probabilità totale di errore del I tipo pari ad α

La zona di rifiuto deve essere costituita solo dalla coda di destra perché **SOLO** valori elevati della statistica test portano a sospettare che H_0 sia falsa

- Si rifiuta H_0 se $X^2 > X^2_{\alpha}(1)$
Per $\alpha = 0.05$ $X^2_{\alpha}(1) = 3.84$

Test Chi-quadrato



Essendo $12.81 > 3.84$
si rifiuta H_0

I dati suggeriscono che sussiste una relazione tra gli episodi di tosse manifestati e l'aver avuto o meno una storia di bronchite.

Estensione del Test Chi-quadrato

- Sino a questo punto abbiamo considerato il confronto di
- una caratteristica che assume due modalità (tosse/no)
 - in due gruppi (bronchite/no)

mediante una soluzione basata su una analisi della associazione tra due caratteri.

Questa soluzione è anche estendibile anche al confronto di

- una caratteristica che assume due o più modalità
- in due o più gruppi

Esempio motivante: problemi mestruali e pratica sportiva

OBIETTIVO : E' noto che le donne che praticano sport hanno in media un numero più ridotto di cicli mestruali. Ci si chiede se a seconda della **pratica sportiva** vi sia anche una differente insorgenza di **problemi mestruali**

Si osservano :

- pratica sportiva (NO; SI, amatoriale; SI, professionale)
 - manifestazione di problemi mestruali (SI, NO)
- in una coorte di donne

METODOLOGIA

si valuta se vi sia associazione tra pratica sportiva e manifestazione di problemi mestruali

I **DATI OSSERVATI** possono essere rappresentati nella tabella di contingenza

| Problemi mestruali ----- | SI | NO | Totale |
|-----------------------------|-----------|-----------|------------|
| <i>Sport</i> | | | |
| NO | 14 | 40 | 54 |
| SI , amatoriale | 9 | 14 | 23 |
| SI , professionale | 46 | 42 | 88 |
| Totale | 69 | 96 | 165 |

Sistema di Ipotesi :

H_0 : NON vi è associazione tra pratica sportiva e problemi mestruali

H_1 : Vi è associazione tra pratica sportiva e problemi mestruali

Costruzione della tabella attesa sotto H_0

Tabella di contingenza
OSSERVATA

| Problemi mestruali ----- | SI | NO | Totale |
|-----------------------------|-----------|-----------|------------|
| <i>Sport</i> | | | |
| NO | 14 | 40 | 54 |
| SI , amatoriale | 9 | 14 | 23 |
| SI , professionale | 46 | 42 | 88 |
| Totale | 69 | 96 | 165 |

Tabella di contingenza
ATTESA sotto H_0

- Se la probabilità di avere problemi mestruali fosse la stessa,
- quante segnalazioni per chi non pratica sport?
 - e così' via ...

La stima della probabilità di avere problemi mestruali (sotto H_0) è $69/165 = 0.4182$

Costruzione della tabella attesa sotto H_0

Tabella di contingenza
OSSERVATA

| <i>Sport</i> | Pr.Mes. SI | Pr.Mes. NO | Totale |
|---------------|------------|------------|------------|
| NO | 14 | 40 | 54 |
| SI amatoriale | 9 | 14 | 23 |
| SI profess. | 46 | 42 | 88 |
| Totale | 69 | 96 | 165 |

Tabella di contingenza
ATTESA sotto H_0

| <i>Sport</i> | Pr. Mes. SI | Pr. Mes. NO | Totale |
|---------------|--------------------------|----------------------|------------|
| NO | $0.418 \cdot 54 = 22.57$ | $54 - 22.57 = 31.43$ | 54 |
| SI amatoriale | $0.418 \cdot 23 = 9.61$ | $23 - 9.61 = 13.39$ | 23 |
| SI profess. | $0.418 \cdot 88 = 36.78$ | $88 - 36.78 = 51.22$ | 88 |
| Totale | 69 | 96 | 165 |

In alternativa, $(69 \cdot 54)/165$, $(96 \cdot 54)/165$ e così' via

Confronto tra tabella osservata ed attesa sotto Ho

Statistica Test

$$X^2 = \sum_{\text{celle della tabella di contingenza}} \frac{(\text{freq.oss.} - \text{freq.att.})^2}{\text{freq.att.}} = \sum \frac{(O-E)^2}{E}$$

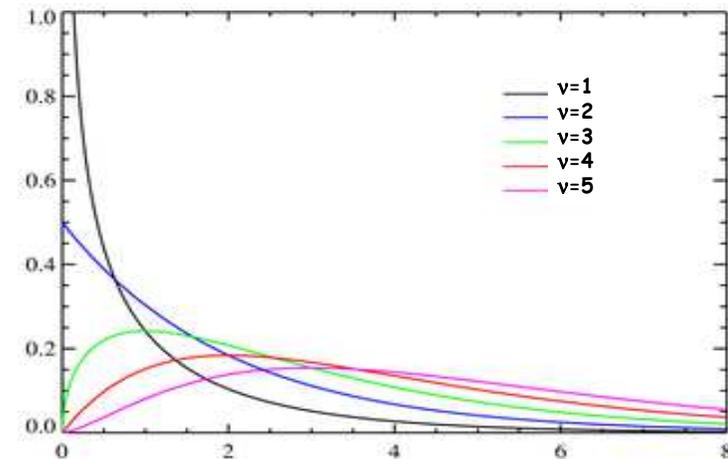
1. Sotto Ho $X^2 \sim \text{Chi}((n-1) \cdot (m-1))$ dove m ed n sono il numero di modalità dei 2 caratteri

Il valore della statistica nel campione è

$$X^2 = \frac{(14 - 22.57)^2}{22.57} + \dots + \frac{(42 - 51.22)^2}{51.22} = 9.63$$

9.63 andrà confrontato con la zona di rifiuto unilaterale costruita sulla distribuzione Chi-Quadrato con $(2-1)(3-1)$ g.d.l.

DISTRIBUZIONE Chi-Quadrato k gradi di libertà



Si noti che all'aumentare dei gradi di libertà aumenta la dispersione della distribuzione

Chi-Quadrato k gradi di libertà - FRATTILI

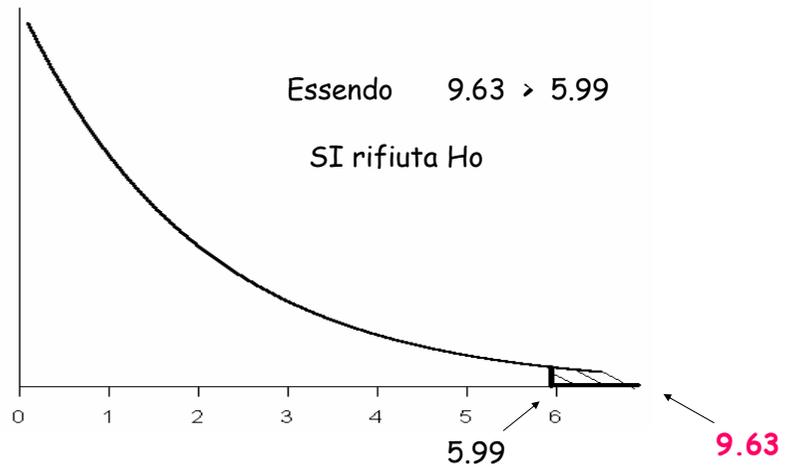
| | $\alpha = 0.05$ | $\alpha = 0.01$ | $\alpha = 0.001$ |
|---------|-----------------|-----------------|------------------|
| $v = 1$ | 3.84 | 6.63 | 10.83 |
| $v = 2$ | 5.99 | 9.21 | 13.82 |
| $v = 3$ | 7.81 | 11.34 | 16.27 |
| $v = 4$ | 9.49 | 13.28 | 18.47 |
| $v = 5$ | 11.07 | 15.09 | 20.52 |

.....

Chi-square Distribution Table

| d.f. | .995 | .99 | .975 | .95 | .9 | .1 | .05 | .025 | .01 |
|------|------|------|------|-------|-------|-------|-------|-------|-------|
| 1 | 0.00 | 0.00 | 0.00 | 0.00 | 0.02 | 2.71 | 3.84 | 5.02 | 6.63 |
| 2 | 0.01 | 0.02 | 0.05 | 0.10 | 0.21 | 4.61 | 5.99 | 7.38 | 9.21 |
| 3 | 0.07 | 0.11 | 0.22 | 0.35 | 0.58 | 6.25 | 7.81 | 9.35 | 11.34 |
| 4 | 0.21 | 0.30 | 0.48 | 0.71 | 1.06 | 7.78 | 9.49 | 11.14 | 13.28 |
| 5 | 0.41 | 0.55 | 0.83 | 1.15 | 1.61 | 9.24 | 11.07 | 12.83 | 15.09 |
| 6 | 0.68 | 0.87 | 1.24 | 1.64 | 2.20 | 10.64 | 12.59 | 14.45 | 16.81 |
| 7 | 0.99 | 1.24 | 1.69 | 2.17 | 2.83 | 12.02 | 14.07 | 16.01 | 18.48 |
| 8 | 1.34 | 1.65 | 2.18 | 2.73 | 3.49 | 13.36 | 15.51 | 17.53 | 20.09 |
| 9 | 1.73 | 2.09 | 2.70 | 3.33 | 4.17 | 14.68 | 16.92 | 19.02 | 21.67 |
| 10 | 2.16 | 2.56 | 3.25 | 3.94 | 4.87 | 15.99 | 18.31 | 20.48 | 23.21 |
| 11 | 2.60 | 3.05 | 3.82 | 4.57 | 5.58 | 17.28 | 19.68 | 21.92 | 24.72 |
| 12 | 3.07 | 3.57 | 4.40 | 5.23 | 6.30 | 18.55 | 21.03 | 23.34 | 26.22 |
| 13 | 3.57 | 4.11 | 5.01 | 5.89 | 7.04 | 19.81 | 22.36 | 24.74 | 27.69 |
| 14 | 4.07 | 4.66 | 5.63 | 6.57 | 7.79 | 21.06 | 23.68 | 26.12 | 29.14 |
| 15 | 4.60 | 5.23 | 6.26 | 7.26 | 8.55 | 22.31 | 25.00 | 27.49 | 30.58 |
| 16 | 5.14 | 5.81 | 6.91 | 7.96 | 9.31 | 23.54 | 26.30 | 28.85 | 32.00 |
| 17 | 5.70 | 6.41 | 7.56 | 8.67 | 10.09 | 24.77 | 27.59 | 30.19 | 33.41 |
| 18 | 6.26 | 7.01 | 8.23 | 9.39 | 10.86 | 25.99 | 28.87 | 31.53 | 34.81 |
| 19 | 6.84 | 7.63 | 8.91 | 10.12 | 11.65 | 27.20 | 30.14 | 32.85 | 36.19 |
| 20 | 7.43 | 8.26 | 9.59 | 10.85 | 12.44 | 28.41 | 31.41 | 34.17 | 37.57 |

Test Chi-quadrato



I dati suggeriscono che vi è associazione tra pratica sportiva e manifestazione di problemi mestruali