# Il test del $\chi^2$

Federico Plazzi

24 Novembre 2015

# A che cosa serve?

# A che cosa serve?

## Condizioni

- **Variabili qualitative**: il test del $\chi^2$ si usa quando si ha a che fare con delle variabili *qualitative*, ossia *categoriche* (sì/no; malato/sano; vivo/morto; ecc.).

# A che cosa serve?

## Condizioni

- **Variabili qualitative**: il test del $\chi^2$ si usa quando si ha a che fare con delle variabili *qualitative*, ossia *categoriche* (sì/no; malato/sano; vivo/morto; ecc.).
- Le variabili qualitative si distinguono da quelle *quantitative* perché i soggetti non *si misurano* con un numero espresso in una qualche unità di misura; i soggetti *vengono invece assegnati* ad una categoria, un gruppo.

# A che cosa serve?

## Condizioni

- **Variabili qualitative**: il test del $\chi^2$ si usa quando si ha a che fare con delle variabili *qualitative*, ossia *categoriche* (sì/no; malato/sano; vivo/morto; ecc.).
- Le variabili qualitative si distinguono da quelle *quantitative* perché i soggetti non *si misurano* con un numero espresso in una qualche unità di misura; i soggetti *vengono invece assegnati* ad una categoria, un gruppo.
- **Il test del $\chi^2$ risente molto della dimensione del campione**: non ha senso utlizzarlo se in una qualche categoria ricadono meno di cinque soggetti. In questo caso, è da preferirsi, se applicabile, il **metodo esatto di Fisher**.

# A che cosa serve?

## Condizioni

- **Variabili qualitative**: il test del $\chi^2$ si usa quando si ha a che fare con delle variabili *qualitative*, ossia *categoriche* (sì/no; malato/sano; vivo/morto; ecc.).
- Le variabili qualitative si distinguono da quelle *quantitative* perché i soggetti non *si misurano* con un numero espresso in una qualche unità di misura; i soggetti *vengono invece assegnati* ad una categoria, un gruppo.
- **Il test del $\chi^2$ risente molto della dimensione del campione**: non ha senso utlizzarlo se in una qualche categoria ricadono meno di cinque soggetti. In questo caso, è da preferirsi, se applicabile, il **metodo esatto di Fisher**.
- Il test del $\chi^2$ confronta la ripartizione dei soggetti nelle varie categorie con l'ipotesi nulla della *distribuzione attesa*.

# Tabelle di contingenza

# Tabelle di contingenza

## Cosa sono?

- Sono semplicemente le tabelle in cui si riportano i soggetti ripartiti per categorie:

Tabella: Ambiti preferiti **osservati**

| Ambito | Numero di studenti |
|---|---|
| Botanica | 12 |
| Ecologia | 14 |
| Geologia | 12 |
| Paleontologia | 8 |
| Zoologia | 54 |

# Come stabilire le ripartizioni attese?

Questo lo decidiamo noi!

# Come stabilire le ripartizioni attese?

## Questo lo decidiamo noi!

- Possiamo decidere che ci aspettiamo che i soggetti siano distribuiti a caso nelle categorie...

Tabella: Ambiti preferiti **attesi** casualmente

| Ambito | Numero di studenti |
|---|---|
| Botanica | 20 |
| Ecologia | 20 |
| Geologia | 20 |
| Paleontologia | 20 |
| Zoologia | 20 |

# Come stabilire le ripartizioni attese?

## Questo lo decidiamo noi!

- Oppure possiamo avere aspettative più complicate...

Tabella: Ambiti preferiti **attesi** in base ai CFU

| Ambito | CFU | Numero di studenti |
|---|---|---|
| Botanica | 18 | 17,48 |
| Ecologia | 14 | 13,59 |
| Geologia | 37 | 35,92 |
| Paleontologia | 6 | 5,83 |
| Zoologia | 28 | 27,18 |

# Calcolo di $\chi^2$

## La logica del test

# Calcolo di $\chi^2$

## La logica del test

- Per ogni categoria, calcoliamo lo scostamento dall'atteso in proporzione:

$$\Delta_{OA} = \frac{O_i - A_i}{A_i} \tag{1}$$

dove $O_i$ è il valore osservato dell'$i$-esima categoria ed $A_i$ è il valore atteso dell'$i$-esima categoria.

# Calcolo di $\chi^2$

## La logica del test

- Per ogni categoria, calcoliamo lo scostamento dall'atteso in proporzione:

$$\Delta_{OA} = \frac{O_i - A_i}{A_i} \tag{1}$$

  dove $O_i$ è il valore osservato dell'$i$-esima categoria ed $A_i$ è il valore atteso dell'$i$-esima categoria.

- Ad esempio, un valore di 0.50 significherebbe che il dato osservato è il 50% più grande di quello atteso.

# Calcolo di $\chi^2$

## La logica del test

- Per ogni categoria, calcoliamo lo scostamento dall'atteso in proporzione:

$$\Delta_{OA} = \frac{O_i - A_i}{A_i} \tag{1}$$

dove $O_i$ è il valore osservato dell'$i$-esima categoria ed $A_i$ è il valore atteso dell'$i$-esima categoria.

- Ad esempio, un valore di 0.50 significherebbe che il dato osservato è il 50% più grande di quello atteso.
- La somma di tutti i $\Delta_{OA}$ farebbe zero, perché i positivi ed i negativi si compensano. Perciò, prima di sommare, passiamo al quadrato:

$$\chi^2 = \sum_{i=1}^{N} \frac{(O_i - A_i)^2}{A_i} \tag{2}$$

dove $N$ è il numero di categorie.

# Il nostro esempio

## Nel primo caso (ripartizione casuale)...

Tabella: Scostamenti quadratici proporzionali

| Ambito | $\frac{(O_i - A_i)^2}{A_i}$ |
|---|---|
| Botanica | 3,20 |
| Ecologia | 1,80 |
| Geologia | 3,20 |
| Paleontologia | 7,20 |
| Zoologia | 57,80 |

$$\chi^2 = 3,20 + 1,80 + 3,20 + 7,20 + 57,80 = 73,20$$

# Il nostro esempio

## Nel secondo caso (ripartizione in base ai CFU)...

Tabella: Scostamenti quadratici proporzionali

| Ambito | $\frac{(O_i - A_i)^2}{A_i}$ |
|---|---|
| Botanica | 1,72 |
| Ecologia | 0,01 |
| Geologia | 15,93 |
| Paleontologia | 0,81 |
| Zoologia | 26,46 |

$$\chi^2 = 1,72 + 0,01 + 15,93 + 0,81 + 26,46 = 44,93$$

# Il test

## La distribuzione di $\chi^2$

- La cosa più semplice è ricavarla in modo empirico, usando un elevato numero di campioni.

# Il test

## La distribuzione di $\chi^2$

- La cosa più semplice è ricavarla in modo empirico, usando un elevato numero di campioni.
- Generiamo 10.000 campioni sotto la nostra ipotesi nulla e calcoliamo 10.000 volte $\chi^2$.

# Il test

## La distribuzione di $\chi^2$

- La cosa più semplice è ricavarla in modo empirico, usando un elevato numero di campioni.
- Generiamo 10.000 campioni sotto la nostra ipotesi nulla e calcoliamo 10.000 volte $\chi^2$.
- Le due distribuzioni sono molto simili, anche se sono generate da due ipotesi nulle diverse! *In realtà, il parametro che governa la distribuzione di $\chi^2$ è il numero di gradi di libertà, pari ad $N - 1$.*

# Il test

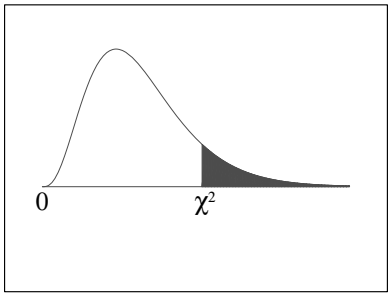
## La distribuzione di $\chi^2$

- La cosa più semplice è ricavarla in modo empirico, usando un elevato numero di campioni.
- Generiamo 10.000 campioni sotto la nostra ipotesi nulla e calcoliamo 10.000 volte $\chi^2$.
- Le due distribuzioni sono molto simili, anche se sono generate da due ipotesi nulle diverse! *In realtà, il parametro che governa la distribuzione di $\chi^2$ è il numero di gradi di libertà, pari ad $N - 1$.*

## Risultato del test

- A questo punto, ci basta sapere quale parte dell'area della distribuzione di $\chi^2$ è sottesa dai valori pari al nostro $\chi^2$ o maggiori: se è meno del 5%, possiamo rigettare l'ipotesi nulla!

# Chi-Square Distribution Table




The shaded area is equal to $\alpha$ for $\chi^2 = \chi^2_\alpha$.

| *df* | $\chi^2_{.995}$ | $\chi^2_{.990}$ | $\chi^2_{.975}$ | $\chi^2_{.950}$ | $\chi^2_{.900}$ | $\chi^2_{.100}$ | $\chi^2_{.050}$ | $\chi^2_{.025}$ | $\chi^2_{.010}$ | $\chi^2_{.005}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.000 | 0.000 | 0.001 | 0.004 | 0.016 | 2.706 | 3.841 | 5.024 | 6.635 | 7.879 |
| 2 | 0.010 | 0.020 | 0.051 | 0.103 | 0.211 | 4.605 | 5.991 | 7.378 | 9.210 | 10.597 |
| 3 | 0.072 | 0.115 | 0.216 | 0.352 | 0.584 | 6.251 | 7.815 | 9.348 | 11.345 | 12.838 |
| 4 | 0.207 | 0.297 | 0.484 | 0.711 | 1.064 | 7.779 | 9.488 | 11.143 | 13.277 | 14.860 |
| 5 | 0.412 | 0.554 | 0.831 | 1.145 | 1.610 | 9.236 | 11.070 | 12.833 | 15.086 | 16.750 |
| 6 | 0.676 | 0.872 | 1.237 | 1.635 | 2.204 | 10.645 | 12.592 | 14.449 | 16.812 | 18.548 |
| 7 | 0.989 | 1.239 | 1.690 | 2.167 | 2.833 | 12.017 | 14.067 | 16.013 | 18.475 | 20.278 |
| 8 | 1.344 | 1.646 | 2.180 | 2.733 | 3.490 | 13.362 | 15.507 | 17.535 | 20.090 | 21.955 |
| 9 | 1.735 | 2.088 | 2.700 | 3.325 | 4.168 | 14.684 | 16.919 | 19.023 | 21.666 | 23.589 |
| 10 | 2.156 | 2.558 | 3.247 | 3.940 | 4.865 | 15.987 | 18.307 | 20.483 | 23.209 | 25.188 |
| 11 | 2.603 | 3.053 | 3.816 | 4.575 | 5.578 | 17.275 | 19.675 | 21.920 | 24.725 | 26.757 |
| 12 | 3.074 | 3.571 | 4.404 | 5.226 | 6.304 | 18.549 | 21.026 | 23.337 | 26.217 | 28.300 |
| 13 | 3.565 | 4.107 | 5.009 | 5.892 | 7.042 | 19.812 | 22.362 | 24.736 | 27.688 | 29.819 |
| 14 | 4.075 | 4.660 | 5.629 | 6.571 | 7.790 | 21.064 | 23.685 | 26.119 | 29.141 | 31.319 |
| 15 | 4.601 | 5.229 | 6.262 | 7.261 | 8.547 | 22.307 | 24.996 | 27.488 | 30.578 | 32.801 |
| 16 | 5.142 | 5.812 | 6.908 | 7.962 | 9.312 | 23.542 | 26.296 | 28.845 | 32.000 | 34.267 |
| 17 | 5.697 | 6.408 | 7.564 | 8.672 | 10.085 | 24.769 | 27.587 | 30.191 | 33.409 | 35.718 |
| 18 | 6.265 | 7.015 | 8.231 | 9.390 | 10.865 | 25.989 | 28.869 | 31.526 | 34.805 | 37.156 |
| 19 | 6.844 | 7.633 | 8.907 | 10.117 | 11.651 | 27.204 | 30.144 | 32.852 | 36.191 | 38.582 |
| 20 | 7.434 | 8.260 | 9.591 | 10.851 | 12.443 | 28.412 | 31.410 | 34.170 | 37.566 | 39.997 |
| 21 | 8.034 | 8.897 | 10.283 | 11.591 | 13.240 | 29.615 | 32.671 | 35.479 | 38.932 | 41.401 |
| 22 | 8.643 | 9.542 | 10.982 | 12.338 | 14.041 | 30.813 | 33.924 | 36.781 | 40.289 | 42.796 |
| 23 | 9.260 | 10.196 | 11.689 | 13.091 | 14.848 | 32.007 | 35.172 | 38.076 | 41.638 | 44.181 |
| 24 | 9.886 | 10.856 | 12.401 | 13.848 | 15.659 | 33.196 | 36.415 | 39.364 | 42.980 | 45.559 |
| 25 | 10.520 | 11.524 | 13.120 | 14.611 | 16.473 | 34.382 | 37.652 | 40.646 | 44.314 | 46.928 |
| 26 | 11.160 | 12.198 | 13.844 | 15.379 | 17.292 | 35.563 | 38.885 | 41.923 | 45.642 | 48.290 |
| 27 | 11.808 | 12.879 | 14.573 | 16.151 | 18.114 | 36.741 | 40.113 | 43.195 | 46.963 | 49.645 |
| 28 | 12.461 | 13.565 | 15.308 | 16.928 | 18.939 | 37.916 | 41.337 | 44.461 | 48.278 | 50.993 |
| 29 | 13.121 | 14.256 | 16.047 | 17.708 | 19.768 | 39.087 | 42.557 | 45.722 | 49.588 | 52.336 |
| 30 | 13.787 | 14.953 | 16.791 | 18.493 | 20.599 | 40.256 | 43.773 | 46.979 | 50.892 | 53.672 |
| 40 | 20.707 | 22.164 | 24.433 | 26.509 | 29.051 | 51.805 | 55.758 | 59.342 | 63.691 | 66.766 |
| 50 | 27.991 | 29.707 | 32.357 | 34.764 | 37.689 | 63.167 | 67.505 | 71.420 | 76.154 | 79.490 |
| 60 | 35.534 | 37.485 | 40.482 | 43.188 | 46.459 | 74.397 | 79.082 | 83.298 | 88.379 | 91.952 |
| 70 | 43.275 | 45.442 | 48.758 | 51.739 | 55.329 | 85.527 | 90.531 | 95.023 | 100.425 | 104.215 |
| 80 | 51.172 | 53.540 | 57.153 | 60.391 | 64.278 | 96.578 | 101.879 | 106.629 | 112.329 | 116.321 |
| 90 | 59.196 | 61.754 | 65.647 | 69.126 | 73.291 | 107.565 | 113.145 | 118.136 | 124.116 | 128.299 |
| 100 | 67.328 | 70.065 | 74.222 | 77.929 | 82.358 | 118.498 | 124.342 | 129.561 | 135.807 | 140.169 |

# Il test

## Risultato del test

- Il p-value associato a un valore di $\chi^2$ di 73,20 con 4 gradi di libertà è di $4,77 \cdot 10^{-15}$!

# Il test

## Risultato del test

- Il p-value associato a un valore di $\chi^2$ di 73,20 con 4 gradi di libertà è di $4,77 \cdot 10^{-15}$!
- Il p-value associato a un valore di $\chi^2$ di 44,93 con 4 gradi di libertà è di $4,11 \cdot 10^{-9}$!

# Il test

## Risultato del test

- Il p-value associato a un valore di $\chi^2$ di 73,20 con 4 gradi di libertà è di $4,77 \cdot 10^{-15}$!
- Il p-value associato a un valore di $\chi^2$ di 44,93 con 4 gradi di libertà è di $4,11 \cdot 10^{-9}$!
- In entrambi i casi, quindi, possiamo rifiutare l'ipotesi nulla.

# Il test

## Risultato del test

- Il p-value associato a un valore di $\chi^2$ di 73,20 con 4 gradi di libertà è di $4,77 \cdot 10^{-15}$!
- Il p-value associato a un valore di $\chi^2$ di 44,93 con 4 gradi di libertà è di $4,11 \cdot 10^{-9}$!
- In entrambi i casi, quindi, possiamo rifiutare l'ipotesi nulla.
- I nostri dati si discostano molto significativamente dalle varie ipotesi nulle: gli ambiti non sono scelti né a caso né in base ai CFU del corso.

# $\chi^2$ per categorizzazioni incrociate

## Tabelle a doppia entrata

- Possiamo usare il test del $\chi^2$ per casi anche più complessi; per esempio, possiamo chiederci se le femmine preferiscono significativamente la zoologia alla geologia rispetto ai maschi.

# $\chi^2$ per categorizzazioni incrociate

## Tabelle a doppia entrata

- Possiamo usare il test del $\chi^2$ per casi anche più complessi; per esempio, possiamo chiederci se le femmine preferiscono significativamente la zoologia alla geologia rispetto ai maschi.
- Partiamo costruendo una tabella a doppia entrata:

Tabella: Zoologia e geologia - Valori osservati

| | F | M | Tot. |
|---|---|---|---|
| Geologia | 6 | 6 | 12 |
| Zoologia | 28 | 26 | 54 |
| Tot. | 34 | 32 | 66 |

# $\chi^2$ per categorizzazioni incrociate

- Costruiamo l'ipotesi nulla a partire dai totali di riga e di colonna. Per esempio:

$$A_{ij} : \text{tot}_i = \text{tot}_j : N \tag{3}$$

$$A_{ij} = \frac{\text{tot}_i \cdot \text{tot}_j}{N} \tag{4}$$

dove $A_{ij}$ è il valore atteso della riga $i$ e della colonna $j$, $\text{tot}_i$ è il totale della riga $i$ e $\text{tot}_j$ è il totale della colonna $j$.

# $\chi^2$ per categorizzazioni incrociate

- Costruiamo l'ipotesi nulla a partire dai totali di riga e di colonna. Per esempio:

$$A_{ij} : \text{tot}_i = \text{tot}_j : N \tag{3}$$

$$A_{ij} = \frac{\text{tot}_i \cdot \text{tot}_j}{N} \tag{4}$$

  dove $A_{ij}$ è il valore atteso della riga $i$ e della colonna $j$, $\text{tot}_i$ è il totale della riga $i$ e $\text{tot}_j$ è il totale della colonna $j$.
- Costruiamo la tabella a doppia entrata delle frequenze attese:

Tabella: Zoologia e geologia - Valori attesi

| | F | M | Tot. |
|---|---|---|---|
| Geologia | 6,18 | 5,82 | 12 |
| Zoologia | 27,82 | 26,18 | 54 |
| Tot. | 34 | 32 | 66 |

# $\chi^2$ per categorizzazioni incrociate

- A questo punto, possiamo calcolare $\chi^2$ come prima, applicando eventualmente una correzione di continuità $(-0.5)$ se la tabella è una tabella $2 \times 2$.

Tabella: $\frac{(|O_{ij} - A_{ij}| - 0.5)^2}{A_{ij}}$

| | F | M |
|---|---|---|
| Geologia | 0,0171 | 0,0171 |
| Zoologia | 0,0037 | 0.0039 |

$$\chi^2 = 0,0171 + 0,0171 + 0,0037 + 0,0039 = 0,0417$$

# $\chi^2$ per categorizzazioni incrociate

- A questo punto, possiamo calcolare $\chi^2$ come prima, applicando eventualmente una correzione di continuità $(-0.5)$ se la tabella è una tabella $2 \times 2$.

Tabella: $\frac{(|O_{ij} - A_{ij}| - 0.5)^2}{A_{ij}}$

| | F | M |
|---|---|---|
| Geologia | 0,0171 | 0,0171 |
| Zoologia | 0,0037 | 0.0039 |

$$\chi^2 = 0,0171 + 0,0171 + 0,0037 + 0,0039 = 0,0417$$

- I gradi di libertà, nel caso delle categorizzazioni incrociate sono $(R-1) \cdot (C-1)$, dove $R$ è il numero delle righe e $C$ il numero delle colonne della tabella a doppia entrata.

# $\chi^2$ per categorizzazioni incrociate

- A questo punto, possiamo calcolare $\chi^2$ come prima, applicando eventualmente una correzione di continuità ($-0.5$) se la tabella è una tabella $2 \times 2$.

Tabella: $\frac{(|O_{ij} - A_{ij}| - 0.5)^2}{A_{ij}}$

| | F | M |
|---|---|---|
| Geologia | 0,0171 | 0,0171 |
| Zoologia | 0,0037 | 0.0039 |

$$\chi^2 = 0,0171 + 0,0171 + 0,0037 + 0,0039 = 0,0417$$

- I gradi di libertà, nel caso delle categorizzazioni incrociate sono $(R-1) \cdot (C-1)$, dove $R$ è il numero delle righe e $C$ il numero delle colonne della tabella a doppia entrata.
- Il p-value associato a un valore di $\chi^2$ di 0,0417 con 1 grado di libertà è di 0,84: non possiamo rifiutare l'ipotesi nulla.

# $\chi^2$ per categorizzazioni incrociate

- A questo punto, possiamo calcolare $\chi^2$ come prima, applicando eventualmente una correzione di continuità ($-0.5$) se la tabella è una tabella $2 \times 2$.

Tabella: $\frac{(|O_{ij} - A_{ij}| - 0.5)^2}{A_{ij}}$

| | F | M |
|---|---|---|
| Geologia | 0,0171 | 0,0171 |
| Zoologia | 0,0037 | 0.0039 |

$$\chi^2 = 0,0171 + 0,0171 + 0,0037 + 0,0039 = 0,0417$$

- I gradi di libertà, nel caso delle categorizzazioni incrociate sono $(R-1) \cdot (C-1)$, dove $R$ è il numero delle righe e $C$ il numero delle colonne della tabella a doppia entrata.
- Il p-value associato a un valore di $\chi^2$ di 0,0417 con 1 grado di libertà è di 0,84: non possiamo rifiutare l'ipotesi nulla.
- **Il fatto però che non possiamo rifiutare l'ipotesi nulla non significa che sia corretta!**