

Psicometria 1

PS 123

Parte 11

Indipendenza e associazione tra variabili qualitative

Corrado Caudlek

Dipartimento di Psicologia, Università di Trieste

Parte 11 – Variabili qualitative

Psicometria 1 - p. 1/119

Sommario

Il materiale coperto è contenuto nel capitolo 8 del testo di Agresti e Finlay (1997).

- Indipendenza e associazione tra variabili qualitative.
- Test χ^2 per due proporzioni.
- Test χ^2 per tavole di contingenza $r \times c$.
- La famiglia di distribuzioni χ^2 .
- Condizioni di validità per il test χ^2 .
- Misure di associazione per variabili categoriali e ordinali.

Sommario

Parte 11 – Variabili qualitative

Psicometria 1 - p. 3/119

Associazione e Indipendenza

- Il problema che ci poniamo è quello di studiare l'**associazione** tra variabili qualitative.
- Diciamo che due variabili qualitative sono associate se, nella popolazione, alcune modalità della prima variabile tendono a presentarsi più spesso in relazione ad alcune modalità della seconda variabile.

Associazione e
Indipendenza
Distribuzioni
marginali
Distribuzioni
marginali
Distribuzioni
condizionate
Distribuzione
 congiunta
Indipendenza e
associazione
Indipendenza e
associazione

Parte 11 – Variabili qualitative

Psicometria 1 - p. 5/119

Sommario

Sommario

Parte 11 – Variabili qualitative

Psicometria 1 - p. 2/119

Associazione e Indipendenza

Associazione e
Indipendenza
Distribuzioni
marginali
Distribuzioni
marginali
Distribuzioni
condizionate
Distribuzione
 congiunta
Indipendenza e
associazione
Indipendenza e
associazione

Parte 11 – Variabili qualitative

Psicometria 1 - p. 4/119

Associazione e Indipendenza

- Per esempio, negli Stati Uniti, le variabili qualitative "affiliazione religiosa" con modalità (protestante, cattolico, altro) e "gruppo etnico" con modalità (caucasico, afro-americano, latino, altro) sono associate nel senso che:
 - ◆ gli appartenenti al gruppo caucasico tendono ad essere protestanti, mentre
 - ◆ gli appartenenti al gruppo latino tendono ad essere cattolici.
- La nozione di indipendenza statistica tra variabili qualitative risulta più chiara se le osservazioni sono rappresentate mediante una tavola di contingenza.

Associazione e
Indipendenza
Distribuzioni
marginali
Distribuzioni
marginali
Distribuzioni
condizionate
Distribuzione
 congiunta
Indipendenza e
associazione
Indipendenza e
associazione

Parte 11 – Variabili qualitative

Psicometria 1 - p. 6/119

Indipendenza e associazione

Per esempio, se la distribuzione dell'identificazione partitica nella popolazione fosse diversa all'interno delle classi dei maschi e delle femmine, allora le variabili "genere" e "identificazione partitica" sarebbero associate o dipendenti.

Il problema che ci poniamo è: le differenze nelle distribuzioni condizionate osservate nel campione possono essere attribuite alla variabilità campionaria, oppure sono così grandi da rendere implausibile l'ipotesi di indipendenza nella popolazione?

Associazione e
indipendenza
Distribuzioni
marginali
Distribuzioni
marginali
Distribuzioni
condizionate
Distribuzioni
condizionate
Distribuzione
 congiunta
Indipendenza e
associazione
associazione

Parte 11 – Variabili qualitative

Psicometria 1 - p. 13/119

Parte 11 – Variabili qualitative

Psicometria 1 - p. 14/119

Differenza di due proporzioni

Differenza di due
proporzioni
ipotesi nulla e
alternativa

Differenza di due proporzioni

■ Un test statistico che consente di rispondere alla domanda precedente, nel caso più semplice di due variabili qualitative, ciascuna con due modalità, è quello sulla differenza di due proporzioni.

■ Possiamo infatti pensare ai dati forniti da due proporzioni come ad una tavola di contingenza in cui la variabile esplicativa e la variabile di risposta hanno due modalità ciascuna.

Differenza di due
proporzioni
ipotesi nulla e
alternativa

Differenza di due proporzioni

Illustrazione. Consideriamo uno studio relativo all'applicazione della pena di morte negli Stati Uniti nei casi in cui la vittima è di razza bianca.

Le unità sono 214 accusati di omicidio in 20 contee della Florida tra il 1976 e il 1977, classificati secondo la razza (modalità: bianco e afro-americano) e il verdetto di condanna a morte (modalità: no e sì).

Accusati	Si	No	Totale
Bianchi	19	132	151
Afro-americani	11	52	63
Totale	30	184	214

Differenza di due
proporzioni
ipotesi nulla e
alternativa

Parte 11 – Variabili qualitative

Psicometria 1 - p. 15/119

Parte 11 – Variabili qualitative

Psicometria 1 - p. 16/119

Illustrazione

■ Dato che il numero di accusati bianchi è 2,4 volte maggiore del numero di accusati afro-americani è difficile fare dei confronti diretti.

■ Per stabilire se l'applicazione della pena di morte varia in funzione della razza dell'accusato, esprimiamo in termini percentuali le modalità della variabile risposta (pena di morte), *all'interno* di ciascuna modalità della variabile esplicativa (razza).

Accusati	Si	No	Numero
Bianchi	12.6	87.4	151
Afro-americani	17.5	82.5	63

Differenza di due
proporzioni
ipotesi nulla e
alternativa

Ipotesi nulla e alternativa

Vogliamo verificare l'ipotesi nulla

$$H_0 : \pi_1 = \pi_2 \quad o \quad H_0 : \pi_1 - \pi_2 = 0$$

la probabilità di un verdetto di condanna a morte è la stessa per gli accusati bianchi e gli accusati afro-americani,

verso l'ipotesi alternativa

$$H_a : \pi_1 < \pi_2 \quad o \quad H_a : \pi_1 - \pi_2 < 0$$

gli accusati bianchi hanno una probabilità minore di ottenere un verdetto di condanna a morte degli accusati afro-americani.

Differenza di due
proporzioni
ipotesi nulla e
alternativa

Parte 11 – Variabili qualitative

Psicometria 1 - p. 17/119

Parte 11 – Variabili qualitative

Psicometria 1 - p. 18/119

Inferenze per le tavole di contingenza

Inferenze per le tavole di contingenza
Test χ^2 per due proporzioni
Terminologia
Statistica χ^2
Test χ^2 con R

Parte 11 – Variabili qualitative

Psicometria 1 - p. 19/119

Inferenze per le tavole di contingenza

Inferenze per le tavole di contingenza
Test χ^2 per due proporzioni
Terminologia
Statistica χ^2
Test χ^2 con R

- La proporzione complessiva di accusati a cui è stata inflitta la pena di morte nel campione è

$$\hat{\pi} = \frac{19 + 11}{151 + 63} = \frac{30}{214} = 0.140$$

- Il test z per la differenza tra proporzioni è

$$\begin{aligned} z &= \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\hat{\pi}(1 - \hat{\pi}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \\ &= \frac{0.126 - 0.175}{\sqrt{0.140(1 - 0.140) \left(\frac{1}{151} + \frac{1}{63} \right)}} = -0.94 \end{aligned}$$

Parte 11 – Variabili qualitative

Psicometria 1 - p. 20/119

Inferenze per le tavole di contingenza

Inferenze per le tavole di contingenza
Test χ^2 per due proporzioni
Terminologia
Statistica χ^2
Test χ^2 con R

- Il p -valore della statistica $z = -0.94$ è uguale a 0.1736.
- Quindi questo particolare campione non fornisce evidenze sufficienti per rifiutare l'ipotesi nulla.

Parte 11 – Variabili qualitative

Psicometria 1 - p. 21/119

Test χ^2 per due proporzioni

Inferenze per le tavole di contingenza
Test χ^2 per due proporzioni
Terminologia
Statistica χ^2
Test χ^2 con R

- Supponiamo però di volere confrontare più di due campioni, o che la variabile di risposta abbia più di due modalità. In tali circostanze, un semplice test basato sulla differenza tra due proporzioni non sarà sufficiente.
- Esaminiamo ora un altro metodo eseguire un test sulla differenza tra due proporzioni. Tale metodo può essere esteso al caso più generale in cui la variabile esplicativa e la variabile di risposta sono composte da un numero qualsiasi di modalità.

Parte 11 – Variabili qualitative

Psicometria 1 - p. 22/119

Test χ^2 per due proporzioni

Inferenze per le tavole di contingenza
Test χ^2 per due proporzioni
Terminologia
Statistica χ^2
Test χ^2 con R

- Se le variabili "razza dell'accusato" e "verdetto di condanna a morte" fossero indipendenti, allora la pena di morte verrebbe inflitta nella stessa proporzione di casi agli accusati bianchi e agli accusati afro-americani.
- Questa proporzione può essere stimata nel modo seguente:

$$\hat{\pi} = \frac{19 + 11}{151 + 63} = \frac{30}{214} = 0.14$$

Parte 11 – Variabili qualitative

Psicometria 1 - p. 23/119

Test χ^2 per due proporzioni

Inferenze per le tavole di contingenza
Test χ^2 per due proporzioni
Terminologia
Statistica χ^2
Test χ^2 con R

- Nel caso presente, ci sono $n = 151$ accusati bianchi e la proporzione di "successi" è $\hat{\pi} = 0.14$. Sotto l'ipotesi di indipendenza, dunque, ci aspettiamo un verdetto di condanna a morte per il seguente numero di accusati bianchi nel campione:
- $$151 \times 0.14 = \frac{151 \times 30}{214} = 21.2$$
- Tale frequenza attesa si calcola allo stesso modo della media di una variabile bernoulliana X con parametri $n = 151$ e probabilità di successo $p = 0.14$, ovvero $\mathbb{E}(X) = np$.

Parte 11 – Variabili qualitative

Psicometria 1 - p. 24/119

Test χ^2 per due proporzioni

In maniera corrispondente, ci aspettiamo che il numero seguente di accusati bianchi nel campione non ottenga un verdetto di condanna a morte:

$$151 \times 0,86 = \frac{151 \times 184}{214} = 129,8$$

Inferenze per le tavole di contingenza
Test χ^2 per due proporzioni
Terminologia
Statistica χ^2
Test χ^2 con R

Test χ^2 per due proporzioni

- Sotto l'ipotesi di indipendenza, ci aspettiamo un verdetto di condanna a morte per il seguente numero di accusati afro-americani nel campione:

$$63 \times 0,14 = \frac{63 \times 30}{214} = 8,8$$

- La frequenza attesa di accusati afro-americani nel campione a cui non dovrebbe essere inflitta la pena di morte è

$$63 \times 0,86 = \frac{63 \times 184}{214} = 54,2$$

Inferenze per le tavole di contingenza
Test χ^2 per due proporzioni
Terminologia
Statistica χ^2
Test χ^2 con R

Test χ^2 per due proporzioni

Queste frequenze attese corrispondono a ciò che ci aspettiamo di osservare *in media* se venissero estratti molti campioni composti da 151 accusati bianchi e 63 accusati afro-americani da una popolazione con probabilità di un verdetto di condanna a morte pari a $p = 0,14$ per entrambi.

Accusati	Si	No	Totale
Bianchi	19 (21,2)	132 (129,8)	151
Afro-americani	11 (8,8)	52 (54,2)	63
Totale	30	184	214

Inferenze per le tavole di contingenza
Test χ^2 per due proporzioni
Terminologia
Statistica χ^2
Test χ^2 con R

Test χ^2 per due proporzioni

Se la razza dell'accusato e il verdetto di condanna a morte sono indipendenti, allora le frequenze attese in ciascuna cella della tavola di contingenza si possono calcolare con la seguente formula:

$$\text{freq. attesa} = \frac{\text{totale riga} \times \text{totale colonna}}{\text{totale generale}}$$

dove il totale generale è la dimensione n del campione.

Inferenze per le tavole di contingenza
Test χ^2 per due proporzioni
Terminologia
Statistica χ^2
Test χ^2 con R

Test χ^2 per due proporzioni

Per esempio, la frequenza attesa nella prima cella della tavola di contingenza è

$$\frac{151 \times 30}{214} = 21,2$$

Inferenze per le tavole di contingenza
Test χ^2 per due proporzioni
Terminologia
Statistica χ^2
Test χ^2 con R

Test χ^2 per due proporzioni

La formula precedente può anche essere concepita come l'applicazione della regola del prodotto per eventi indipendenti.

- La proporzione complessiva di accusati a cui viene inflitta la pena di morte è

$$\hat{\pi}_m = \frac{30}{214} = 0,140$$

- La proporzione di accusati bianchi è

$$\hat{\pi}_b = \frac{151}{214} = 0,706$$

Inferenze per le tavole di contingenza
Test χ^2 per due proporzioni
Terminologia
Statistica χ^2
Test χ^2 con R

Test χ^2 per due proporzioni

- Se i due eventi "pena di morte" e "razza dell'accusato" sono indipendenti, allora la proporzione di "accusati bianchi a cui viene inflitta la pena di morte" è

$$\hat{\pi}_{mb} = \hat{\pi}_m \hat{\pi}_b = 0.140 \times 0.706 = 0.0988$$

- da cui deriva la frequenza attesa di accusati bianchi nel campione a cui viene inflitta la pena di morte:

$$n\hat{\pi}_{mb} = 214 \times 0.0988 = 21.2$$

Inferenze per le tavole di contingenza
Test χ^2 per due proporzioni
Terminologia
Statistica χ^2
Test χ^2 con R

Terminologia

Le frequenze attese sotto l'ipotesi di indipendenza dovrebbero essere chiamate **frequenze attese stimate** in quanto sono basate sulle distribuzioni marginali delle due variabili in un campione (e non sulle frequenze marginali della popolazione).

Inferenze per le tavole di contingenza
Test χ^2 per due proporzioni
Terminologia
Statistica χ^2
Test χ^2 con R

Statistica χ^2

La statistica χ^2 confronta le frequenze attese con le frequenze osservate in tutte le celle della tavola di contingenza:

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

Inferenze per le tavole di contingenza
Test χ^2 per due proporzioni
Terminologia
Statistica χ^2
Test χ^2 con R

Statistica χ^2

Illustrazione. Per l'esempio presente, avremo

f. osservate	f. attese	oss. - att.	$\frac{(\text{oss.}-\text{att.})^2}{\text{att.}}$
19	21.2	-2.168	0.222
132	129.8	2.168	0.036
11	8.8	2.168	0.532
52	54.2	-2.168	0.087
214	214.0	0.0	$\chi^2=0.877$

Inferenze per le tavole di contingenza
Test χ^2 per due proporzioni
Terminologia
Statistica χ^2
Test χ^2 con R

Illustrazione

- Il p -valore della statistica $\chi^2 = 0.877$ è $P = 0.349$.
- Il test χ^2 per una tavola di contingenza 2×2 è soltanto un modo un pò più complicato di calcolare il test z relativo alla differenza tra due proporzioni. In questo caso, $\chi^2 = z^2$.
- Per l'esempio presente, infatti $-0.94^2 = 0.88$.
- La sola differenza tra i due test è che il test z può essere usato per sottoporre a verifica un'ipotesi nulla unilaterale mentre il test χ^2 è bilaterale. Per questa ragione, nel caso presente il p -valore del test χ^2 è pari al doppio (dato l'errore di approssimazione) del p -valore del test z unilaterale.

Inferenze per le tavole di contingenza
Test χ^2 per due proporzioni
Terminologia
Statistica χ^2
Test χ^2 con R

Test χ^2 con R

Il test χ^2 può essere eseguito con la funzione `chisq.test` di R:

```
> deathpenalty <-
+ matrix(c(19, 132, 11, 52),
+        byrow=T, ncol=2,
+        dimnames = list("Razza" = c("Bianca", "Nera"),
+        "Pena di morte" = c("Si", "No")))
> deathpenalty
      Pena di morte
Razza  Si No
Bianca 19 132
Nera   11  52
```

Inferenze per le tavole di contingenza
Test χ^2 per due proporzioni
Terminologia
Statistica χ^2
Test χ^2 con R

Test χ^2 con R

```
> chisq.test(deathPenalty, correct=FALSE)

Pearson's Chi-squared test

data: deathPenalty
X-squared = 0.8774, df = 1, p-value = 0.3489
```

Inferenza per le
tavole di
contingenza
Test χ^2 per due
proporzioni
Terminologia
Statistica χ^2
Test χ^2 con R

Test χ^2 con R

Come in precedenza, l'argomento `correct=FALSE` fa in modo che non venga applicata la correzione per la continuità.

```
> chisq.test(deathPenalty)

Pearson's Chi-squared test with Yates' continuity correction

data: deathPenalty
X-squared = 0.5194, df = 1, p-value = 0.4711
```

Tavole di contingenza $r \times c$

Tavole di
contingenza $r \times c$
Distribuzioni χ^2
Distribuzioni χ^2 in
R
Distribuzioni χ^2
Test χ^2
Gradi di libertà

Tavole di contingenza $r \times c$

- I dati della tavola seguente provengono da un'indagine condotta nel 1989 sulla popolazione canadese e sono riportati da Sniderman, Northrup, Felcher, Russell e Tellock (1993).
- L'indagine riguardava gli atteggiamenti dei canadesi nei confronti dei diritti civili.
- Tra gli altri termini trattati vi era l'antisemitismo e, a questo proposito, il questionario conteneva la seguente domanda.

Tavole di
contingenza $r \times c$
Distribuzioni χ^2
Distribuzioni χ^2 in
R
Distribuzioni χ^2
Test χ^2
Gradi di libertà

Tavole di contingenza $r \times c$

We realize, NO STATEMENT is true of all people in a group, but GENERALLY SPEAKING, please tell me whether you agree or disagree with the following statements. The first one is: Most Jews don't care what happens to people who aren't Jewish. Would you say that you agree strongly, agree somewhat, disagree somewhat, or disagree strongly with that statement?

Tavole di
contingenza $r \times c$
Distribuzioni χ^2
Distribuzioni χ^2 in
R
Distribuzioni χ^2
Test χ^2
Gradi di libertà

Tavole di contingenza $r \times c$

I dati possono essere presentati in una tavola di contingenza 4×2 avente come righe le modalità della risposta e come colonne le modalità della variabile esplicativa (gruppo linguistico).

	Lingua			Totale
	Inglese	Francese	Totale	
<i>Jews don't care</i>				
<i>Agree</i>	261	171	432	
<i>Disagree Somewhat</i>	576	159	735	
<i>Disagree Strongly</i>	561	73	634	
<i>Don't know or Refused</i>	134	117	251	
Totale	1,532	520	2,052	

Tavole di
contingenza $r \times c$
Distribuzioni χ^2
Distribuzioni χ^2 in
R
Distribuzioni χ^2
Test χ^2
Gradi di libertà

Tavole di contingenza $r \times c$

Per interpretare più facilmente la tavola di contingenza calcoliamo le **percentuali di colonna** – dato che la variabile esplicitiva è riportata in colonna.

	Lingua	
	Inglese	Francese
<i>Jews don't care</i>		
<i>Agree</i>	17.1	33.0
<i>Disagree Somewhat</i>	37.6	30.5
<i>Disagree Strongly</i>	36.6	14.1
<i>Don't Know or Refused</i>	8.7	22.5
Totale	100.0	100.1
Numero di casi	1532	520

Tavole di contingenza $r \times c$
Distribuzioni χ^2 in **R**
Distribuzioni χ^2 in **Testi** χ^2
Gradi di libertà

Tavole di contingenza $r \times c$

- Non è possibile riassumere la distribuzione di risposte per gli intervistati di lingua inglese e francese con una singola percentuale dato che la variabile di risposta (*Jews don't care*) ha più di due modalità.
- Di conseguenza, il test z per la differenza di due proporzioni non può essere applicato.
- Possiamo però calcolare le frequenze attese assumendo che vi sia indipendenza tra la variabile "gruppo linguistico" e la variabile "anti-semitismo", e poi usare il test χ^2 per confrontare le frequenze attese e le frequenze osservate.

Tavole di contingenza $r \times c$
Distribuzioni χ^2 in **R**
Distribuzioni χ^2 in **Testi** χ^2
Gradi di libertà

Tavole di contingenza $r \times c$

La frequenza attesa per gli intervistati di lingua inglese che concordano con l'affermazione *Jews don't care*, per esempio, è

$$\text{freq. attesa} = \frac{\text{totale riga} \times \text{totale colonna}}{\text{totale generale}} = \frac{432 \times 1532}{2052} = 322.5$$

Tavole di contingenza $r \times c$
Distribuzioni χ^2 in **R**
Distribuzioni χ^2 in **Testi** χ^2
Gradi di libertà

Tavole di contingenza $r \times c$

Le frequenze attese in ciascuna cella della tavola di contingenza sono le seguenti:

	Lingua		
	Inglese	Francese	Totale
<i>Jews don't care</i>			
<i>Agree</i>	322.5	109.5	432
<i>Disagree Somewhat</i>	548.7	186.3	735
<i>Disagree Strongly</i>	473.3	160.7	634
<i>Don't Know or Refused</i>	187.4	63.6	251
Totale	1,531.9	520.1	2,052

Tavole di contingenza $r \times c$
Distribuzioni χ^2 in **R**
Distribuzioni χ^2 in **Testi** χ^2
Gradi di libertà

Si noti che, tenendo conto dell'errore di arrotondamento, la somma delle frequenze attese è uguale ai totali di riga e di colonna.

Tavole di contingenza $r \times c$

- Il test χ^2 viene calcolato nel modo seguente:

$$\chi^2 = \sum \frac{(\text{frequenze osservate} - \text{frequenze attese})^2}{\text{frequenze attese}} = \frac{(261 - 322.5)^2}{322.5} + \frac{(171 - 109.5)^2}{109.5} + \dots + \frac{(117 - 63.6)^2}{63.6} = 175.8$$

- Prima di trovare il p -valore della statistica test dobbiamo esaminare le proprietà delle distribuzioni χ^2 .

Tavole di contingenza $r \times c$

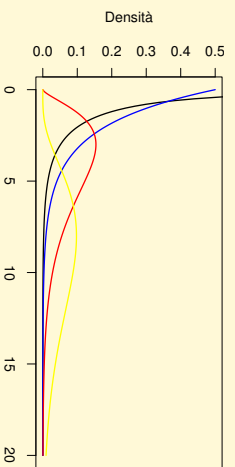
- Come nel caso delle distribuzioni t di Student, anche le distribuzioni χ^2 formano una famiglia di funzioni di densità parametrizzate dai gradi di libertà.
- ◆ C'è una diversa distribuzione χ^2 a seconda dei numeri di grado di libertà 1, 2, 3, ...
- ◆ Diversamente dalle distribuzioni t di Student e dalla distribuzione normale, la distribuzione χ^2 possiede un'asimmetria positiva.
- L'asimmetria della distribuzione diminuisce al crescere dei gradi di libertà.
- La media della distribuzione χ^2 è uguale al numero ν di gradi di libertà: $\mathbb{E}(\chi^2) = \nu$.
- La varianza della distribuzione χ^2 è uguale a due volte il numero di gradi di libertà: $\mathbb{V}(\chi^2) = 2\nu$.

Distribuzioni χ^2

Tavole di contingenza $r \times c$
Distribuzioni χ^2 in **R**
Distribuzioni χ^2 in **Testi** χ^2
Gradi di libertà

Distribuzioni χ^2 in R

```
x <- seq(0, 20, 0.1)
chisq1 <- dchisq(x[-1], df = 1)
plot(x[-1], chisq1, type = "l", xlab = "Densità(\`a)",
      ylim=c(0, .5) )
chisq2 <- dchisq(x, df = 2) lines(x, chisq2, col = "blue")
chisq5 <- dchisq(x, df = 5) lines(x, chisq5, col = "red")
chisq10 <- dchisq(x, df = 10) lines(x, chisq10, col = "yellow")
```



Parte 11 – Variabili qualitative

Psicometria 1 - p. 49/119

Distribuzioni χ^2

- La variabile χ^2 è definita nell'intervallo $[0, \infty]$.
- La statistica χ^2 dipende dal quadrato delle differenze tra frequenze osservate e attese e, quindi, non può assumere valori negativi.
- Se le frequenze osservate e le frequenze attese sono uguali, la statistica χ^2 ha valore 0.
- Tanto maggiore è la differenza tra frequenze osservate e attese, tanto più grande sarà il valore della statistica χ^2 .

Tabelle di
contingenza $r \times c$
Distribuzioni χ^2
Distribuzioni χ^2 in
R
Distribuzioni χ^2
Test χ^2
Gradi di libertà

Parte 11 – Variabili qualitative

Psicometria 1 - p. 50/119

Gradi di libertà

- I gradi di libertà per il test χ^2 in tabelle di contingenza $r \times c$ sono

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

dove r è il numero di righe e c è il numero di colonne.

- Nel primo esempio, $r = 2, c = 2, gdl = (2 - 1)(2 - 1) = 1$.
- Nel secondo esempio, $r = 4, c = 2, gdl = (4 - 1)(2 - 1) = 3$.

Tabelle di
contingenza $r \times c$
Distribuzioni χ^2
Distribuzioni χ^2 in
R
Distribuzioni χ^2
Test χ^2
Gradi di libertà

Parte 11 – Variabili qualitative

Psicometria 1 - p. 52/119

Test χ^2

- Il test χ^2 è unilaterale:
 - l'ipotesi nulla di indipendenza tra le due variabili rappresentate in una tavola di contingenza viene rifiutata quando la statistica χ^2 è sufficientemente grande.
- L'ipotesi alternativa, invece, è non-direzionale:
 - una qualunque forma di associazione tra le due variabili.

Tabelle di
contingenza $r \times c$
Distribuzioni χ^2
Distribuzioni χ^2 in
R
Distribuzioni χ^2
Test χ^2
Gradi di libertà

Parte 11 – Variabili qualitative

Psicometria 1 - p. 51/119

Gradi di libertà

- I gradi di libertà per il test χ^2 in una tavola di contingenza obbediscono al seguente vincolo:
 - i totali marginali attesi devono essere uguali ai totali marginali osservati.*
- Quindi, se calcoliamo le frequenze attese per tutte le celle della tavola in una riga o colonna *tranne l'ultima*, il valore rimanente può essere calcolato per sottrazione dai valori già considerati.

Tabelle di
contingenza $r \times c$
Distribuzioni χ^2
Distribuzioni χ^2 in
R
Distribuzioni χ^2
Test χ^2
Gradi di libertà

Parte 11 – Variabili qualitative

Psicometria 1 - p. 53/119

Distribuzioni χ^2

Illustrazione. Nel caso dell'esempio presente, all'interno di ciascuna colonna è sufficiente specificare soltanto 3 frequenze attese (l'ultima frequenza attesa è determinata dalle altre, dato che il totale marginale è fisso).

	Lingua			Totale
	Inglese	Francese	Totale	
<i>Jews don't care</i>				
Agree	322.5	?	432	
Disagree Somewhat	548.7	?	735	
Disagree Strongly	473.3	?	634	
<i>Don't Know or Refused</i>	?	?	251	
Totale	1,531.9	520.1	2,052	

Tabelle di
contingenza $r \times c$
Distribuzioni χ^2
Distribuzioni χ^2 in
R
Distribuzioni χ^2
Test χ^2
Gradi di libertà

Parte 11 – Variabili qualitative

Psicometria 1 - p. 54/119

Test χ^2

Tabelle di contingenza $r \times c$
Distribuzioni χ^2
Distribuzioni χ^2 in R

Test χ^2
Gradi di libertà

- Il p -valore della statistica χ^2 può essere trovato con R oppure consultando una tavola sinottica che riporta i valori critici della distribuzione χ^2 .
- La tavola C nel testo di Agresti e Finlay è costruita come la tavola della distribuzione t di Student e riporta i gradi di libertà (df) sulle righe e la probabilità nella coda destra della distribuzione χ^2 sulle colonne.
- Primo esempio: $\chi^2_1 = 0,877 \rightarrow P > 0,25$.
- Secondo esempio: $\chi^2_3 = 175,8 \rightarrow P < 0,001$.

```
> 1 - pchisq(0.877, 1)
[1] 0.3490247
> 1 - pchisq(175.8, 3)
[1] 0
```

Parte 11 – Variabili qualitative

Psicometria 1 - p. 55/119

Test χ^2 in R

Il test χ^2 per l'esempio presente può essere eseguito con R nel modo seguente.

```
> antiSemitism <-
+ matrix(c(261, 171, 576, 159, 561, 73, 134, 117),
+        byrow=T, ncol=2,
+        dimnames = list("Jews Don't Care" =
+ c("Agree", "Disagree Somewhat", "Disagree Strongly",
+ "Don't Know or Refused"),
+ "Language" = c("English", "French")))
>
> antiSemitism
      Language
Jews Don't Care  English French
Agree           261      171
Disagree Somewhat  576      159
Disagree Strongly  561      73
Don't Know or Refused 134     117
```

Parte 11 – Variabili qualitative

Psicometria 1 - p. 56/119

Test χ^2 in R

```
> chisq.test(antiSemitism, correct=FALSE)
```

```
    Pearson's Chi-squared test
```

```
data:  antiSemitism
X-squared = 175.7601, df = 3, p-value = < 2.2e-16
```

Tabelle di contingenza $r \times c$
Distribuzioni χ^2
Distribuzioni χ^2 in R
Distribuzioni χ^2
Test χ^2
Gradi di libertà

Parte 11 – Variabili qualitative

Psicometria 1 - p. 57/119

Condizioni di validità del test χ^2

Condizioni di validità del test χ^2
Residui di Pearson
Residui di Pearson
e R

Parte 11 – Variabili qualitative

Psicometria 1 - p. 58/119

Condizioni di validità del test χ^2

Il test χ^2 può essere usato per verificare l'esistenza di un'associazione in una tavola di contingenza.

- È necessario avere dei c.c.i. di osservazioni per ciascuna modalità della variabile esplicativa.
- Il p -valore della statistica χ^2 non è accurato se il campione rappresenta una frazione considerevole della popolazione ($> 10\%$).
- Il test χ^2 può essere utilizzato anche quando i soggetti vengono assegnati in maniera casuale a diversi trattamenti sperimentali e la variabile di risposta è categoriale.

Condizioni di validità del test χ^2
Residui di Pearson
Residui di Pearson
e R

Parte 11 – Variabili qualitative

Psicometria 1 - p. 59/119

Condizioni di validità del test χ^2

- La distribuzione χ^2 fornisce soltanto un'approssimazione alla distribuzione esatta della statistica χ^2 .
- L'approssimazione è adeguata se
 - ◆ il campione è abbastanza grande ($n > 100$ oppure più restrittivamente $n > 200$), e
 - ◆ entro ogni casella solo poche frequenze attese (non oltre il 20%) sono inferiori a 5 e nessuna frequenza attesa è minore di 1.

Condizioni di validità del test χ^2
Residui di Pearson
Residui di Pearson
e R

Parte 11 – Variabili qualitative

Psicometria 1 - p. 60/119

Residui di Pearson

- Il test χ^2 , così come gli altri test di significatività, fornisce poche informazioni. Se il p -valore del test è molto piccolo, i dati consentono di rifiutare l'ipotesi nulla di indipendenza tra i due criteri di classificazione. Questo significa che le variabili sono associate.
- Il test χ^2 , però, non dice nulla a proposito della natura di tale associazione.
- L'analisi dei residui dovrebbe essere fatta seguire al rifiuto dell'ipotesi nulla in quanto il confronto tra le frequenze osservate e le frequenze attese può contribuire a rivelare la fonte dell'associazione tra le variabili.

Condizioni di validità del test χ^2
Residui di Pearson e R

Parte 11 – Variabili qualitative

Psicometria 1 - p. 61/119

Residui di Pearson

La differenza

frequenza osservata - frequenza attesa

è chiamata **residuo**.

- Gli scarti tra le frequenze osservate e le frequenze attese non risentono solamente delle differenze realmente esistenti tra le due distribuzioni a confronto, ma anche delle variazioni casuali.
- Quanto grandi devono essere i residui affinché si possa escludere che siano dovuti agli errori di campionamento?

Condizioni di validità del test χ^2
Residui di Pearson e R

Parte 11 – Variabili qualitative

Psicometria 1 - p. 62/119

Residui di Pearson

Pearson ha dimostrato che, sotto H_0 , il rapporto

$$z = \frac{\text{frequenze osservate} - \text{frequenze attese}}{\sqrt{\text{frequenze attese}}} \sim \mathcal{N}(0, 1)$$

segue la distribuzione normale standardizzata, ovvero $z \sim \mathcal{N}(0, 1)$.

Un grande residuo standardizzato (diciamo > 3) può quindi essere preso come evidenza che, nella corrispondente cella della tavola di contingenza, la frequenza osservata si discosta in maniera significativa dalla frequenza attesa in base all'ipotesi di indipendenza.

Condizioni di validità del test χ^2
Residui di Pearson e R

- Utilizzando **R**, le frequenze attese si ottengono nel modo seguente:

```
> out <- chisq.test(tantisemitism, correct=FALSE)
> out$expected
```

	Language
Jews Don't Care	English French
Agree	322.5263 109.47368
Disagree Somewhat	548.7427 186.25731
Disagree Strongly	473.3372 160.66277
Don't Know or Refused	187.3938 63.60624

- Calcoliamo il residuo standardizzato per la prima cella della tavola di contingenza:

```
[1] -3.425928
```

Parte 11 – Variabili qualitative

Psicometria 1 - p. 63/119

Parte 11 – Variabili qualitative

Psicometria 1 - p. 64/119

Residui di Pearson e R

I residui standardizzati in tutte le celle sono:

```
> out$residuals

Language
Jews Don't Care      -3.425929    5.880389
Agree                 1.163586   -1.997222
Disagree Somewhat    4.029302   -6.916041
Disagree Strongly    -3.900434    6.694847
Don't Know or Refused
```

Condizioni di validità del test χ^2
Residui di Pearson e R

- Nel gruppo di lingua inglese, troviamo una frequenza significativamente minore di quella attesa per la categoria "Agree" mentre, per la stessa categoria, il gruppo francese esibisce una frequenza significativamente maggiore a quella attesa.

- Nella cella "Disagree Strongly" troviamo una frequenza significativamente maggiore di quella attesa per il gruppo inglese e significativamente minore di quella attesa per il gruppo francese.

Parte 11 – Variabili qualitative

Psicometria 1 - p. 65/119

Parte 11 – Variabili qualitative

Psicometria 1 - p. 66/119

Residui di Pearson e R

Condizioni di validità del test χ^2
Residui di Pearson e R

- Nella cella "*Don't Know or Refused*", infine, troviamo una frequenza minore di quella attesa in base all'ipotesi di indipendenza per il gruppo inglese e maggiore di quella attesa per il gruppo francese.
- In conclusione, l'esame dei residui indica che l'associazione tra anti-semitismo e gruppo linguistico è dovuta al fatto che, in questo campione, atteggiamenti anti-semiti sono presenti in misura maggiore all'interno della comunità francese.

Parte 11 – Variabili qualitative

Psicometria 1 - p. 67/119

Residui di Pearson e R

Condizioni di validità del test χ^2
Residui di Pearson e R

I residui nella categoria "*Don't Know or Refused*" per i due gruppi potrebbero significare, però, che i membri della comunità francese hanno opinioni meno forti su tale questione degli intervistati di lingua inglese.

Parte 11 – Variabili qualitative

Psicometria 1 - p. 68/119

Misure di associazione

Misure di associazione

Parte 11 – Variabili qualitative

Psicometria 1 - p. 69/119

Misure di associazione

Misure di associazione

- Nella tavola precedente, le opinioni a favore sono il 60% del totale, in entrambi i gruppi. La variabile *opinione* non è associata alla variabile *razza*.
- Nella tavola seguente, invece, l'associazione è perfetta: tutti i bianchi esprimono un'opinione favorevole, tutti i neri un'opinione contraria. Per questi dati, la variabile *opinione* è completamente determinata dalla variabile *razza*.

Opinione		
Razza	Favorevole	Contraria
Bianca	600	0
Nera	0	400
Totale	600	400

Parte 11 – Variabili qualitative

Psicometria 1 - p. 71/119

Misure di associazione

Misure di associazione

- Una misura di associazione è una statistica che rivela la forza della dipendenza statistica tra due variabili.
- Nella tavola 8.12 l'opinione nei confronti della legalizzazione dell'aborto (favorevole verso contraria) è indipendente dalla razza degli intervistati (bianca verso nera).

Opinione			
Razza	Favorevole	Contraria	Totale
Bianca	360	240	600
Nera	240	160	400
Totale	600	400	1000

Parte 11 – Variabili qualitative

Psicometria 1 - p. 70/119

Misure di associazione

Misure di associazione

- Le misure di associazione tra variabili categoriali rivelano quanto una tavola di contingenza sia simile a uno o l'altro degli estremi che rappresentano l'assenza di associazione o la perfetta associazione.
- Il test χ^2 verifica l'ipotesi nulla di indipendenza. Se l'ipotesi di indipendenza viene rifiutata, però, il test χ^2 non consente di valutare l'intensità dell'associazione tra due variabili.

Parte 11 – Variabili qualitative

Psicometria 1 - p. 72/119

Odds

Odds

Parte 11 – Variabili qualitative

Psicometria 1 - p. 73/119

Odds

Odds

- **Lodds ratio** (*ratio of odds*) è la più importante misura della forza di una associazione tra due variabili categoriali.
- Per comprendere questa misura, occorre introdurre il concetto di "odds" (*quote di scommessa*).
- Supponiamo di avere una variabile risposta binaria B con modalità 0 (*insuccesso*) e 1 (*successo*).
- Oltre alle probabilità di successo spesso si considerano anche **le quote di scommessa** (gli odds).

Parte 11 – Variabili qualitative

Psicometria 1 - p. 74/119

Odds

Odds

- Gli odds di successo ω sono per definizione il rapporto tra la probabilità di successo (π) e la probabilità di insuccesso ($1 - \pi$):

$$\omega = \frac{\pi}{1 - \pi}$$

- La quota di scommessa è un indice non negativo che misura quanti successi ci si attendono per ogni insuccesso.

Parte 11 – Variabili qualitative

Psicometria 1 - p. 75/119

Odds

Odds

Si noti che

- $\omega = 1$: la probabilità di successo è uguale alla probabilità di insuccesso;
- $\omega < 1$: la probabilità di successo è minore della probabilità di insuccesso;
- $\omega > 1$: la probabilità di successo è maggiore della probabilità di insuccesso.

Parte 11 – Variabili qualitative

Psicometria 1 - p. 76/119

Odds

Odds

- Se $\pi = 0.75$ la quota è $\omega = 0.75/0.25 = 3$. Scommettendo su $B = 0$ si riceve 3 volte la posta.
- Se l'odds è minore di 1, possiamo calcolare il reciproco ($1/\omega$) e interpretare il risultato come riferito all'evento complementare.
- ◆ Per esempio, se ω è 0.3333, ci aspettiamo 0.3333 successi per ogni insuccesso ossia $1/0.3333 = 3$ insuccessi per ogni successo.

Parte 11 – Variabili qualitative

Psicometria 1 - p. 77/119

Odds

Odds

La relazione inversa tra probabilità di successo e odds è

$$\pi = \frac{\omega}{1 + \omega}$$

- Per esempio, se $\omega = 3$, la probabilità di successo sarà uguale a 0.75.

Parte 11 – Variabili qualitative

Psicometria 1 - p. 78/119

Odds

Odds

- Consideriamo ora il rapporto degli odds in una tavola di contingenza.
- In una tavola di contingenza, poniamo la variabile esplicativa sulle righe e la variabile di risposta sulle colonne.
- Per la ciascuna modalità della variabile esplicativa,
 - ◆ la probabilità di successo è uguale al rapporto tra il numero di successi e il totale di riga;
 - ◆ la probabilità di insuccesso è uguale al rapporto tra il numero di insuccessi e il totale di riga.

Parte 11 – Variabili qualitative

Psicometria 1 - p. 79/119

Odds

Odds

Consideriamo nuovamente i dati relativi alla pena di morte.

Accusati	Sì	No	Totale
Bianchi	19	132	151
Afro-americani	11	52	63
Totale	30	184	214

Parte 11 – Variabili qualitative

Psicometria 1 - p. 80/119

Odds

Odds

- Per gli accusati bianchi, la probabilità di un verdetto di condanna a morte è $19/151 = 0.1258$ e la probabilità dell'evento complementare è $132/151 = 0.8742$.
- Per gli accusati afro-americani, la probabilità di un verdetto di condanna a morte è $11/63 = 0.1746$ e la probabilità dell'evento complementare è $52/63 = 0.8254$.

Parte 11 – Variabili qualitative

Psicometria 1 - p. 81/119

Odds

Odds

- Per gli accusati bianchi l'odds della pena di morte è
- $$\omega_b = \frac{0.1258}{0.8742} = 0.1439$$
- Per gli accusati afro-americani, l'odds della pena di morte è

$$\omega_a = \frac{0.1746}{0.8254} = 0.2115$$

Parte 11 – Variabili qualitative

Psicometria 1 - p. 82/119

Odds ratio

Odds ratio

Interpretazione

L'**odds ratio** (*ratio of odds*) è il rapporto tra gli odds di due eventi.

Per una tavola di contingenza, il rapporto delle quote (*odds-ratio*) detto anche *rapporto incrociato* (*cross-product ratio*) è il rapporto tra gli odds di due righe:

$$\theta = \frac{\omega_i}{\omega_j}$$

Parte 11 – Variabili qualitative

Psicometria 1 - p. 83/119

Parte 11 – Variabili qualitative

Psicometria 1 - p. 84/119

Odds ratio

Illustrazione. Per l'esempio presente, l'odds ratio tra gli accusati bianchi e gli accusati afro-americani è

$$\theta = \frac{0,1439}{0,2115} = 0,6804$$

uguale a 0.6804.

Odds ratio
Interpretazione

Interpretazione

- Come si interpreta il rapporto degli odds?
- Nel caso presente diremo che, per gli accusati bianchi, l'odds della pena di morte è 0.6804 volte più piccolo che per gli accusati afro-americani.
- In maniera equivalente, possiamo calcolare il reciproco di θ e dire che, per gli accusati afro-americani, l'odds della pena di morte è $1/0.6804 = 1.4698$ più grande che per gli accusati bianchi.

Odds ratio
Interpretazione

Parte 11 – Variabili qualitative

Psicometria 1 - p. 85/119

Parte 11 – Variabili qualitative

Psicometria 1 - p. 86/119

Interpretazione

- Questo NON significa che la **probabilità** di un verdetto di condanna a morte sia 1.47 volte maggiore per gli accusati afro-americani che per gli accusati bianchi.
- L'odds ratio è un rapporto tra odds, non un rapporto tra probabilità.

Odds ratio
Interpretazione

Odds ratio

- Quando $\theta = 1$, le probabilità di successo sono uguali in entrambe le righe della tavola di contingenza. Le variabili riga e colonna sono dunque indipendenti.
- Quando $\theta > 1$, le probabilità di successo sono maggiori nella prima riga della tavola: $\pi_1 > \pi_2$.
- Quando $\theta < 1$, le probabilità di successo sono minori nella prima riga della tavola: $\pi_1 < \pi_2$.
- Valori di θ diversi da 0, nelle due direzioni, rappresentano dunque la forza dell'associazione tra le variabili riga e colonna.

Odds ratio
Interpretazione

Parte 11 – Variabili qualitative

Psicometria 1 - p. 87/119

Parte 11 – Variabili qualitative

Psicometria 1 - p. 88/119

Tavola di contingenza 2×2

- Nel caso di una tavola di contingenza 2×2 , l'**odds ratio** si può calcolare direttamente dalle frequenze nella tavola, senza calcolare gli odds.
- L'odds ratio è il rapporto tra il prodotto delle frequenze sulla diagonale principale e il prodotto delle frequenze sulla diagonale opposta. Per questa ragione, l'odds ratio è anche detto **rapporto incrociato**.

a	b
c	d

$$\theta = ad/bc$$

Odds ratio
Interpretazione

Tavola di contingenza 2×2

Illustrazione. Per l'esempio presente avremo

$$\theta = \frac{19 \times 52}{132 \times 11} = 0,6804$$

Odds ratio
Interpretazione

Parte 11 – Variabili qualitative

Psicometria 1 - p. 89/119

Parte 11 – Variabili qualitative

Psicometria 1 - p. 90/119

Rischio relativo

Rischio relativo
Relazione tra RR
e θ
Calcolo dell'odds
ratio con R

Parte 11 – Variabili qualitative

Psicometria 1 - p. 91/119

Rischio relativo

Rischio relativo
Relazione tra RR
e θ
Calcolo dell'odds
ratio con R

Il rischio relativo è il rapporto tra la probabilità che un evento si verifichi dato uno stato del mondo e la probabilità che tale evento si verifichi dato un altro stato del mondo.

Illustrazione. Per l'esempio presente,

$$RR = \frac{P(\text{pena di morte} \mid \text{afro-americano})}{P(\text{pena di morte} \mid \text{bianco})} = \frac{0.1746}{0.1258} = 1.3879$$

che non è uguale all'odds ratio ($\theta = 1.4698$).

Parte 11 – Variabili qualitative

Psicometria 1 - p. 92/119

Relazione tra RR e θ

Rischio relativo
Relazione tra RR
e θ
Calcolo dell'odds
ratio con R

- Anche se θ non è uguale al rischio relativo, questi due indici sono simili se la probabilità di successo ≈ 0 . In tali circostanze, il rapporto degli odds si può interpretare come un rischio relativo.
- Che cosa ci dice θ del rischio relativo quando le probabilità π_1 e π_2 non sono prossime allo zero?
 - ♦ $\theta = 1$ implica che i criteri di classificazione R e C sono indipendenti;
 - ♦ $\theta > 1$ implica $RR > 1$;
 - ♦ $\theta < 1$ implica $RR < 1$;
 - ♦ RR è sempre più simile a 1 di θ .

Parte 11 – Variabili qualitative

Psicometria 1 - p. 93/119

Calcolo dell'odds ratio con R

Rischio relativo
Relazione tra RR
e θ
Calcolo dell'odds
ratio con R

- E' semplice calcolare l'odds ratio. Tuttavia, se vogliamo usare R procediamo nel modo seguente.

```
> library(vcd)
> oddsratio(deadpenalty, log=FALSE)
[1] 0.6718622
> (19.5*52.5)/(132.5*11.5)
[1] 0.6718622
```
- ♦ Si noti che ai dati è stata applicata la correzione per la continuità (0.5 è stato sommato a ciascuna cella della tavola).

Parte 11 – Variabili qualitative

Psicometria 1 - p. 94/119

Associazione tra variabili ordinali

Associazione tra
variabili ordinali
di Goodman e
Kruskal
Calcolo di gamma
con R
Inferenze per
variabili ordinali
intervallo di
confidenza

Associazione tra variabili ordinali

Associazione tra
variabili ordinali
di Goodman e
Kruskal
Calcolo di gamma
con R
Inferenze per
variabili ordinali
intervallo di
confidenza

- Le misure di associazione tra variabili ordinali tengono in considerazione l'ordinamento tra le categorie delle due variabili. Agresti e Finlay discutono l'indice **gamma**.
- Nella tavola 8.17, 104 unità vengono classificate secondo la soddisfazione lavorativa (modalità: *dissatisfied*, *moderately satisfied*, *very satisfied*) e il reddito (livelli: < 5000, 5000 – 25000, > 25000).
- Possiamo ipotizzare che la soddisfazione lavorativa cresca al crescere del reddito, e viceversa. Per quantificare l'associazione tra le due variabili ordinali, Agresti e Finlay usano un indice di **concordanza**.

Parte 11 – Variabili qualitative

Psicometria 1 - p. 95/119

Parte 11 – Variabili qualitative

Psicometria 1 - p. 96/119

Associazione tra variabili ordinali

```
> Jobsatisfaction <-
+ matrix(c(6, 13, 3, 9, 37, 12, 3, 13, 8),
+        byrow=T, ncol=3,
+        dimnames = list("Job Satisfaction" =
+ c("Dissatisfied", "Moderately Satisfied", "Very Satisfied"),
+ "Income" = c("< 5000", "5000-25000", "> 25000")))
>
>
> Jobsatisfaction
      Income
Job Satisfaction < 5000 5000-25000 > 25000
Dissatisfied      6          13          3
Moderately Satisfied 9          37         12
Very Satisfied    3          13          8
```

Parte 11 – Variabili qualitative

Psicometria 1 - p. 97/119

Associazione tra variabili ordinali

- Una coppia di unità nel campione si dice **concordante** se l'unità che riceve il punteggio più alto sulla prima variabile Y riceve anche il punteggio più alto sulla seconda variabile X .
- La coppia si dice **discordante** se l'unità che riceve un punteggio più alto sulla variabile Y ha un punteggio più basso su X .
- La coppia forma un **pareggio** se le due unità sono classificate nello stesso modo su entrambe le variabili Y e X .
- Se il numero di coppie concordanti C eccede il numero di coppie discordanti D , allora vi sono evidenze di una relazione monotonica positiva.

Associazione tra variabili ordinali
 γ di Goodman e Kruskal
 Calcolo di gamma con **R**
 Inferenze per variabili ordinali
 Intervallo di confidenza

Parte 11 – Variabili qualitative

Psicometria 1 - p. 98/119

γ di Goodman e Kruskal

L'indice di associazione ordinale **gamma di Goodman e Kruskal** è basato sulla differenza tra il numero di coppie concordanti e il numero di coppie discordanti:

$$\hat{\gamma} = \frac{C - D}{C + D}$$

Associazione tra variabili ordinali
 γ di Goodman e Kruskal
 Calcolo di gamma con **R**
 Inferenze per variabili ordinali
 Intervallo di confidenza

Parte 11 – Variabili qualitative

Psicometria 1 - p. 99/119

γ di Goodman e Kruskal

- $\hat{\gamma}$ misura il grado di associazione monotonica tra due variabili, e varia nell'intervallo $-1 \leq \hat{\gamma} \leq 1$. $\hat{\gamma}$ assume valori positivi o negativi a seconda che l'associazione tra le variabili sia di tipo positivo o negativo.
- Una perfetta monotonicità nel campione ($\hat{\gamma} = 1$) si verifica quando non ci sono coppie discordanti ($D = 0$, $\hat{\gamma} = 1$) o non ci sono coppie concordanti ($C = 0$, $\hat{\gamma} = -1$).

Associazione tra variabili ordinali
 γ di Goodman e Kruskal
 Calcolo di gamma con **R**
 Inferenze per variabili ordinali
 Intervallo di confidenza

Parte 11 – Variabili qualitative

Psicometria 1 - p. 100/119

γ di Goodman e Kruskal

- La statistica $\hat{\gamma}$ può essere interpretata come una stima della differenza tra la probabilità di concordanza e la probabilità di discordanza:

$$\hat{\gamma} = \frac{C - D}{C + D} = \frac{C}{C + D} - \frac{D}{C + D}$$

- Uno dei limiti di questa misura di associazione ordinale è che $\hat{\gamma}$ tende a sovrastimare l'entità di associazione.
- Altre misure di associazione ordinate sono τ_b e τ_c di Kendall, p_b e p_c di Spearman e d di Somers.

Associazione tra variabili ordinali
 γ di Goodman e Kruskal
 Calcolo di gamma con **R**
 Inferenze per variabili ordinali
 Intervallo di confidenza

Parte 11 – Variabili qualitative

Psicometria 1 - p. 101/119

γ di Goodman e Kruskal

Illustrazione. Consideriamo ora una coppia di soggetti, uno classificato (< \$ 5000, dissatisfied) e l'altro classificato (\$ 5000 - 25000, moderately satisfied).

- Il primo soggetto è uno dei 6 classificati nella cella 1, 1 della tavola di contingenza; il secondo è uno dei 37 classificati nella cella 2, 2.
- Questa coppia è concordante dato che il secondo soggetto ha valori più alti del primo su entrambi i fattori.
- Complessivamente, per queste due celle, ci sono $6 \times 37 = 222$ coppie concordanti.

Associazione tra variabili ordinali
 γ di Goodman e Kruskal
 Calcolo di gamma con **R**
 Inferenze per variabili ordinali
 Intervallo di confidenza

Parte 11 – Variabili qualitative

Psicometria 1 - p. 102/119

Illustrazione

- Calcoliamo il numero di coppie concordanti per tutti i dati della tavola 8.17.

6		13		
37	12		9	
13	8	8	13	8

- Per la prima cella: $C = 6(37 + 12 + 13 + 8)$.
 - Per la seconda cella: $C = 13(12 + 8)$.
 - Per la terza cella: $C = 9(13 + 8)$.
 - Per la quarta cella: $C = 37(8)$.
- Il numero totale di coppie concordanti è $C = 1105$.

Parte 11 – Variabili qualitative

Psicometria 1 - p. 103/119

Illustrazione

- Calcoliamo il numero di coppie discordanti per i dati della tavola 8.17.

	3			13	
9	37		12	9	
3	13	3	13	3	37

- Per la prima cella: $C = 3(9 + 37 + 3 + 13)$.
 - Per la seconda cella: $C = 12(3 + 13)$.
 - Per la terza cella: $C = 13(9 + 3)$.
 - Per la quarta cella: $C = 37(3)$.
- Il numero totale di coppie discordanti è $D = 645$.

Parte 11 – Variabili qualitative

Psicometria 1 - p. 104/119

Illustrazione

- La tavola 8.17 contiene $C = 1165$ coppie concordanti e $D = 645$ coppie discordanti.

- La statistica $\hat{\gamma}$ diventa

$$\hat{\gamma} = \frac{C - D}{C + D} = \frac{1165 - 645}{1165 + 645} = 0.287$$

- In conclusione possiamo dire che i dati della tavola 8.17 esibiscono una moderata correlazione positiva tra soddisfazione lavorativa e reddito.

Associazione tra variabili ordinali γ di Goodman e Kruskal
 Calcolo di gamma
 con R
 Inferenza per variabili ordinali
 Intervallo di confidenza

Parte 11 – Variabili qualitative

Psicometria 1 - p. 105/119

Calcolo di gamma con R

```
gamma2 <- function(x) {
  # x is a matrix of counts.
  f <- list(apply(x, MARGIN=2, FUN=function(x) {
    (colSums(x)) <- NULL;
    a <- matrix("a", nrow=nrow(x), ncol=ncol(x))
    b <- row(x)
    m <- col(x)
    n <- sum(m)
    inferenza <- matrix(0, nrow=nrow(m), ncol=ncol(m))
    for(i in 1:nrow))
      for(j in 1:ncol))
        pi <- (a[i,j] + sum(m[rows(x) < col(x[j])], na.rm=T)) *
          sum(b[rows(x) < col(x[j])], na.rm=T)) /
          (sum(m[rows(x) < col(x[j])], na.rm=T) *
            sum(b[rows(x) < col(x[j])], na.rm=T))
    inferenza[i,j] <- (pi - D / (C + D)) / 2
    sigma2 <- sum(k^2 * pi) / 2 - sum(x * pi) / 2 *
      Agresti, p. 76
    list(gamma = gamma, C = C, D = D, sigma=sqrt(sigma2))
  })
}
```

Associazione tra variabili ordinali γ di Goodman e Kruskal
 Calcolo di gamma
 con R
 Inferenza per variabili ordinali
 Intervallo di confidenza

Parte 11 – Variabili qualitative

Psicometria 1 - p. 106/119

Calcolo di gamma con R

- R non dispone di una funzione per il calcolo di $\hat{\gamma}$. Possiamo però utilizzare la funzione scritta da Laura A. Thompson.

- Per i dati della tavola 8.17, avremo

```
> Gamma2.F(obsat$atisfaction)
$gamma
[1] 0.2872928
```

```
$C
[1] 1165
```

```
$D
[1] 645
```

```
$sigma
[1] 0.1505677
```

Associazione tra variabili ordinali γ di Goodman e Kruskal
 Calcolo di gamma
 con R
 Inferenza per variabili ordinali
 Intervallo di confidenza

Parte 11 – Variabili qualitative

Psicometria 1 - p. 107/119

Inferenze per variabili ordinali

- Lipotesi nulla di indipendenza tra variabili ordinali può essere formulata come $H_0 : \gamma = 0$.
- Lipotesi alternativa può essere bilaterale, $H_0 : \gamma \neq 0$, o unilaterale, $H_0 : \gamma < 0$ o $H_0 : \gamma > 0$.
- La statistica test si calcola come

$$z = \frac{\hat{\gamma} - 0}{\hat{\sigma}_{\hat{\gamma}}}$$

- sotto H_0 , si distribuisce come una variabile normale standardizzata, $z \sim N(0, 1)$.
- Per l'esempio sulla soddisfazione lavorativa avremo

$$z = \frac{0.287 - 0}{0.1506} = 1.9081$$

Associazione tra variabili ordinali γ di Goodman e Kruskal
 Calcolo di gamma
 con R
 Inferenza per variabili ordinali
 Intervallo di confidenza

Parte 11 – Variabili qualitative

Psicometria 1 - p. 108/119

Inferenze per variabili ordinali

Illustrazione. Nel caso presente l'ipotesi alternativa è direzionale (la soddisfazione lavorativa aumenta all'aumentare del salario) e quindi calcoliamo il p -valore della statistica test corrispondente alla sola coda destra della distribuzione normale standardizzata:

```
> 1 - pnorm(1.908064)
[1] 0.02819147
```

Dato che il p -valore della statistica test è minore di 0.05 possiamo concludere che vi è una significativa associazione positiva tra soddisfazione lavorativa e salario.

Associazione tra variabili ordinali χ^2 di Goodman e Kruskal
Calcolo di gamma con R
Inferenze per variabili ordinali
Intervallo di confidenza

Parte 11 – Variabili qualitative

Psicometria 1 - p. 109/119

Illustrazione

Agresti e Finlay fanno notare che, se avessimo usato il test χ^2 che non fa uso dell'informazione relativa all'ordinamento tra le categorie di ciascuna variabile, tale effetto non sarebbe risultato significativo.

```
> chisq.test(jobssatisfaction)
Pearson's Chi-squared test
```

```
data: jobssatisfaction
X-squared = 4.0942, df = 4, p-value = 0.3934
```

```
Warning message:
Chi-squared approximation may be incorrect in:
chisq.test(jobssatisfaction)
```

Associazione tra variabili ordinali χ^2 di Goodman e Kruskal
Calcolo di gamma con R
Inferenze per variabili ordinali
Intervallo di confidenza

Parte 11 – Variabili qualitative

Psicometria 1 - p. 110/119

Illustrazione

■ Dato che le frequenze in alcune celle sono piccole, il test esatto di Fisher risulta più appropriato

```
> fisher.test(jobssatisfaction, alternative = "greater")
```

```
Fisher's Exact Test for Count Data
```

```
data: jobssatisfaction
p-value = 0.4265
alternative hypothesis: greater
```

■ Entrambe le analisi che considerano i dati a livello nominale, però, non raggiungono la significatività.

Parte 11 – Variabili qualitative

Psicometria 1 - p. 111/119

Intervallo di confidenza

Gli intervalli di confidenza per $\hat{\gamma}$ hanno la forma usuale della stima puntuale più o meno z volte l'errore standard:

$$\hat{\gamma} \pm z\hat{\sigma}_{\hat{\gamma}}$$

Associazione tra variabili ordinali χ^2 di Goodman e Kruskal
Calcolo di gamma con R
Inferenze per variabili ordinali
Intervallo di confidenza

Parte 11 – Variabili qualitative

Psicometria 1 - p. 112/119

Intervallo di confidenza

Illustrazione. Nel caso presente, l'intervallo di confidenza al 95% è

```
0.2872928±1.96×0.1505677 = 0.2872928±0.2951127
ovvero (-0.0078, 0.5824).
```

Dato che l'intervallo di confidenza include lo 0, però, dobbiamo concludere che le evidenze a favore di un'associazione positiva non sono del tutto convincenti.

Per giungere a conclusioni più certe è necessario disporre di un campione più grande.

Associazione tra variabili ordinali χ^2 di Goodman e Kruskal
Calcolo di gamma con R
Inferenze per variabili ordinali
Intervallo di confidenza

Parte 11 – Variabili qualitative

Psicometria 1 - p. 113/119

Conclusioni

Conclusioni

Parte 11 – Variabili qualitative

Psicometria 1 - p. 114/119

Conclusioni

Conclusioni

- Il test χ^2 viene usato per verificare l'ipotesi nulla di indipendenza tra una variabile esplicativa posta sulle righe e una variabile di risposta posta sulle colonne di una tavola di contingenza.
- Sotto l'ipotesi di indipendenza, le frequenze attese in ciascuna cella della tavola di contingenza $r \times c$ si calcolano come
frequenza attesa = $\frac{\text{totale riga} \times \text{totale colonna}}{\text{totale generale}}$

Parte 11 – Variabili qualitative

Psicometria 1 - p. 115/119

Conclusioni

La statistica test

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

è approssimativamente distribuita come χ^2 con $(r-1)(c-1)$ gradi di libertà.

Parte 11 – Variabili qualitative

Psicometria 1 - p. 116/119

Conclusioni

Conclusioni

- Affinché il test χ^2 sia accurato,
 - entro ogni casella solo poche frequenze attese (non oltre il 20%) possono essere inferiori a 5 e nessuna frequenza attesa deve essere minore di 1;
 - le dimensioni del campione devono essere grandi;
 - le frequenze osservate in ciascuna cella della tavola di contingenza devono essere un c.c.i. estratto dalla popolazione.

Parte 11 – Variabili qualitative

Psicometria 1 - p. 117/119

Conclusioni

Conclusioni

- Una volta stabilito che due variabili categoriali sono tra loro associate, dobbiamo stabilire
 - ◆ da che cosa dipende tale associazione, e
 - ◆ quale è la forza dell'associazione.
- L'esame dei residui standardizzati consente di individuare le celle di una tavola di contingenza che si discostano in maniera significativa dalle frequenze attese in base all'ipotesi di indipendenza.

Parte 11 – Variabili qualitative

Psicometria 1 - p. 118/119

Conclusioni

Conclusioni

- L'*odds ratio* consente di confrontare gli *odds* di successo per le diverse modalità della variabile esplicativa.
- Se le probabilità di successo per le modalità della variabile esplicativa sono prossime allo zero (successo = evento raro), allora l'*odds ratio* si può interpretare come un *rischio relativo*.
- Se le variabili sono a livello di scala ordinale, l'associazione tra le variabili può essere misurata con la statistica $\hat{\gamma}$ di Goodman e Kruskal.

Parte 11 – Variabili qualitative

Psicometria 1 - p. 119/119