SDE 2018

Marco Riani
mriani@unipr.it
http://www.riani.it

Tabella di contingenza $r \times c$

gradi di libertà:

$(r\text{-}1) \times (c\text{-}1)$

x \ y	y_1	...	y_c	Tot
x_1	n_{11}		n_{1c}	$n_{1.}$
...				
x_r	n_{r1}		n_{rc}	$n_{r.}$
Tot	$n_{.1}$		$n_{.c}$	n

Variabili indipendenti se:

$$\frac{n_{i1}}{n_{.1}} = \frac{n_{i2}}{n_{.2}} = \ldots = \frac{n_{ij}}{n_{.j}} = \ldots = \frac{n_{ic}}{n_{.c}} = \frac{n_{i.}}{n}$$

Tabella di contingenza $r \times c$

gradi di libertà:

$(r\text{-}1) \times (c\text{-}1)$

Variabili indipendenti se:

$$n_{ij} = \frac{n_{i.} n_{.j}}{n} \quad \forall i = 1,\ldots,r; j = 1,\ldots,c$$

x \ y	y_1	...	y_c	Tot
x_1	n_{11}		n_{1c}	$n_{1.}$
...				
x_r	n_{r1}		n_{rc}	$n_{r.}$
Tot	$n_{.1}$		$n_{.c}$	n

Definizione: Indipendenza

In caso di indipendenza, la modalità assunta da X non influenza le modalità assunte da Y.

Due variabili nominali *X* ed *Y* si dicono indipendenti, con riferimento alle *n* unità statistiche analizzate, se e solo se, per ogni *i* e *j*:

$$n_{ij} = \quad n^*_{ij} = \frac{n_{i.} \cdot n_{.j}}{n}$$

Tabella

- Occorre generalizzare la statistica di Pearson per tabelle 2 x 2 al caso di tabella r x c

$$\chi^2 = \frac{\left(n_{11} - n_{11}^*\right)^2}{n_{11}^*} + \frac{\left(n_{12} - n_{12}^*\right)^2}{n_{12}^*} + \frac{\left(n_{21} - n_{21}^*\right)^2}{n_{21}^*} + \frac{\left(n_{22} - n_{22}^*\right)^2}{n_{22}^*}$$

Indice X^2 (CHI QUADRATO) di Pearson in tabelle *rxc*

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

$$\left(n_{ij} - n_{ij}^*\right)$$

Contingenze

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

Φ compreso tra -1 e +1 in tabella 2x2

(attribuzione convenzionale del segno in base a: $n_{11}\, n_{22} - n_{12}\, n_{21}$)

- non normalizzato in tabella *rxc* (può risultare >1)

Indice di Cramer

$$V = \frac{\phi}{\sqrt{\min[(r-1),(c-1)]}} \quad \begin{cases} 0 \\ 1 \end{cases}$$

V = 1 nell'ipotesi di massima dipendenza tra i caratteri

(per qualsiasi valore di r e c)

$V = \Phi$ se r =2 e/o c =2

INFERENZA SULLA ASSOCIAZIONE

Obiettivo dell'inferenza:

Estendere la conoscenza da un campione di osservazioni ad una popolazione (universo) più ampia

$\Rightarrow$ è un processo che ha a che fare con l'estrazione della conoscenza (in una fase confermativa, non esplorativa)

ASSUNZIONE: le n osservazioni bivariate presentate nella tabella di contingenza costituiscono un campione casuale

DECISIONE

- Calcolo del valore di χ^2 campionario
- H_0: indipendenza; H_1: associazione (significativa)
- g = (r-1)(c-1)
- Scelta di α (livello di significatività = *prob. di errore di prima specie*)
- Lettura sulla tavola del valore χ^2_{α}
- **RIFIUTO H_0 SE:**
- $\chi^2_{camp} > \chi^2_{\alpha}$
- **Con SPSS: RIFIUTO H0 se: *p-value* <α**
- **La probabilità di errore è minore di α**

Distribuzione di χ^2 sotto H_0

IPOTESI DI INDIPENDENZA STOCASTICA

Stima di n_{ij} : $$n_{ij}^* = \frac{n_{i.} n_{.j}}{n}$$

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

= variabile aleatoria

Se H_0 è vera $$\chi^2 \to \chi^2((r-1)(c-1)) \quad n \to \infty$$

Variabile aleatoria χ^2

$$Y = (Z_1)^2 + (Z_2)^2 + \ldots + (Z_g)^2$$

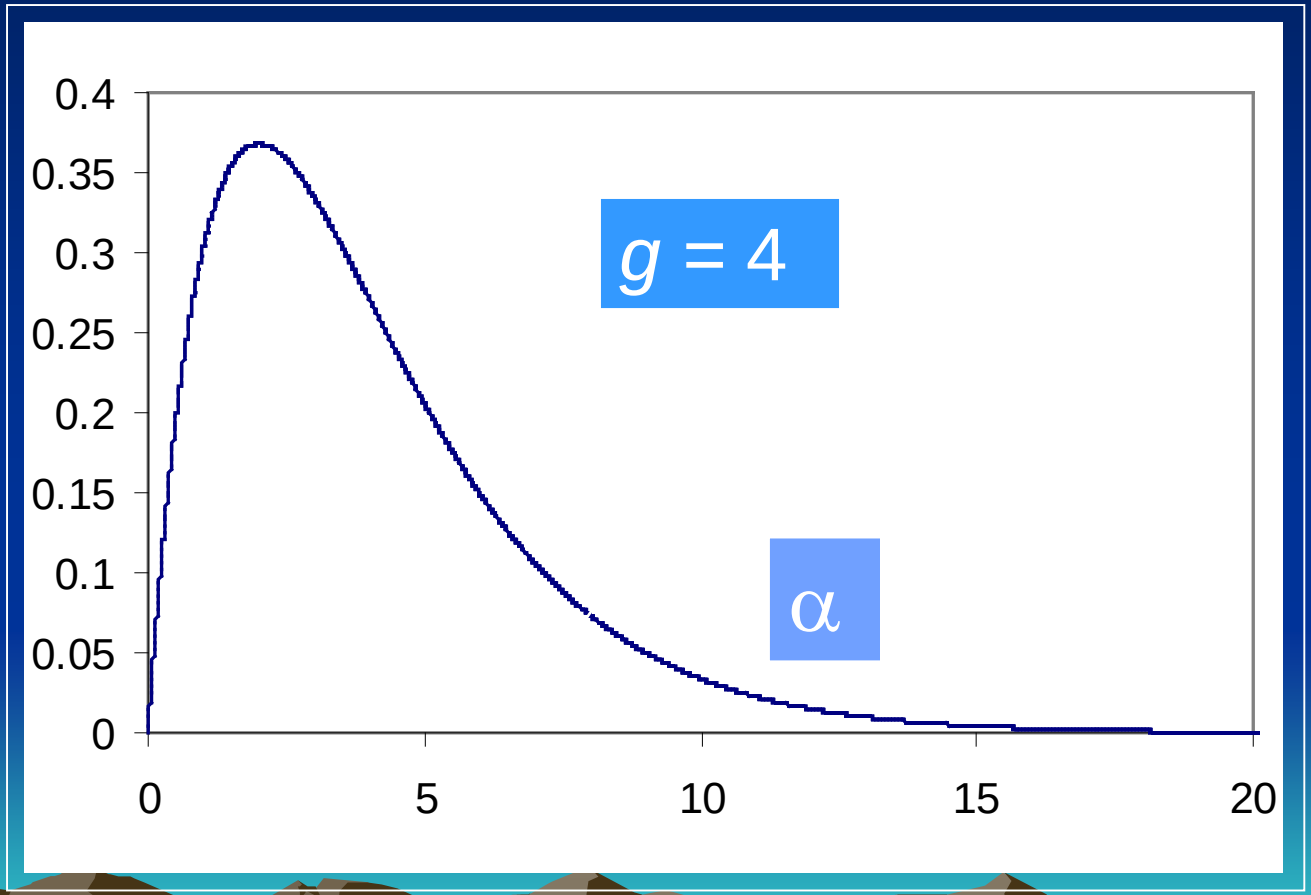
ove $Z_i \sim N(0,1)$ $i = 1, \ldots, g$

$$Y \sim \chi^2(g)$$

1. $E(\chi^2(g)) = g$
2. $VAR(\chi^2(g)) = 2g$
3. $f(\chi^2(g))$ è asimmetrica, ma se $g \rightarrow \infty$ la $\chi^2(g)$ standardizzata $\rightarrow$ *NORMALE*

Verifica di ipotesi sulla χ^2

$f(\chi^2)$



v.c. Chi quadrato	Densità, ripartizione e quantili in una v.c. Chi quadrato	chi2.m chi2.mlx	chi2_out.m chi2_out.mlx	24/10/2018

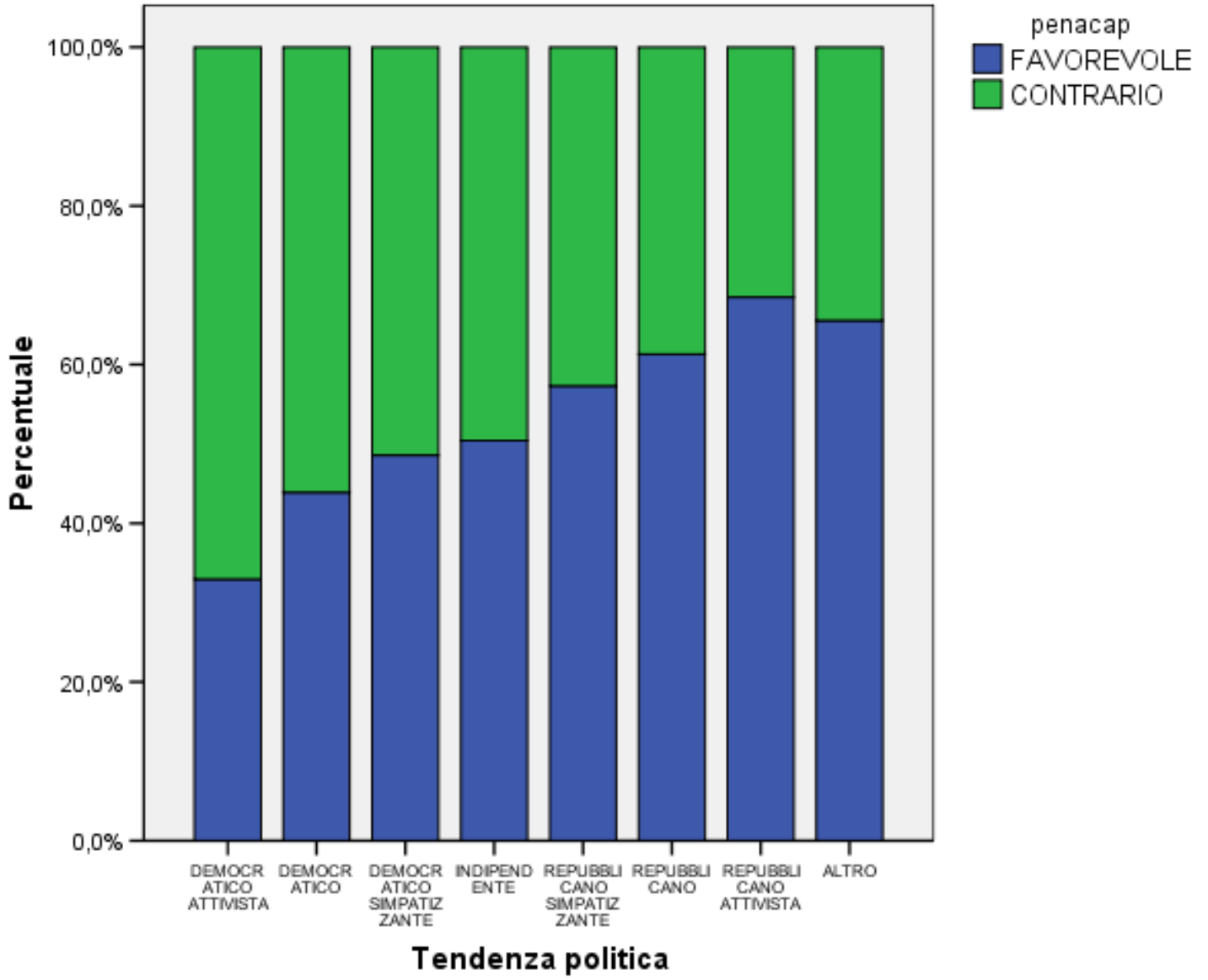
Esercizio

- Il file SONDAUSA.xlsx è una parte di un'indagine multiscopo
- Per ciascun intervistato sono riportate informazioni anagrafiche, sulle tendenze politiche e il pensiero a riguardo di alcune tematiche di stretta attualità. Gli studiosi sono interessati ad indagare su relazioni intercorrenti tra le variabili
- *Obiettivo 1* È sostenibile la tesi secondo la quale la tendenza politica (var. **partito**) influenzi in qualche modo il parere riguardo alla pena capitale (var. **penacap**)?

Esercizio

- Costuire la tabella di contingenza tra le variabili partito e penacap
- Mostrare tramite un grafico a barre sovrapposto (in pila al 100% l'andamento dei favorevoli e contrari al progredire della tendenza politica)
- Calcolare il X^2 per misurare l'associazione tra le due variabili

Barre in pila al 100%



Quarta settimana

Tabelle di contingenza	Test di associazione in tabelle di contingenza rxc	tabellerxc.m tabellerxc.mlx sondUSA.xlsx	tabellerxc_out.m tabellerxc_out.mlx	24/10/2018

Concetto di dipendenza e interdipendenza

- **Dipendenza:**
- Esiste un antecedente (logico o temporale) ed un conseguente

- **Interdipendenza**:
- Le due variabili sono sullo stesso piano

Dipendenza ed Interdipendenza

La dipendenza di Y da X (***r*** ≥ ***c***) si definisce massima quando vi è una sola frequenza non nulla per ogni riga

La dipendenza di X da Y (***c*** ≥ ***r***) si definisce massima quando vi è una sola frequenza non nulla per ogni colonna

Si ha massima **interdipendenza** quando

$$n_{ij} = n_{i.} = n_{.j} \qquad (c = r)$$

Massima interdipendenza e dipendenza

	y_1	y_2	Tot
x_1	n_{11}	0	$n_{1.}$
x_2	0	n_{22}	$n_{2.}$
x_3	0	n_{32}	$n_{3.}$
Tot	$n_{.1}$	$n_{.2}$	n

	y_1	y_2	y_3	Tot
x_1	0	n_{12}	0	$n_{1.}$
x_2	$n_{21.}$	0	0	$n_{2.}$
x_3	0	0	n_{33}	$n_{3.}$
Tot	$n_{.1}$	$n_{.2}$	$n_{.3}$	n

Max dipendenza di Y da X

Max interdipendenza

La dipendenza di Y da X (***r*** ≥ ***c***) si definisce massima quando vi è una sola frequenza non nulla per ogni riga

Osservazioni conclusive

- **La verifica della significatività dell'associazione è una condizione preliminare al calcolo degli indici**
- Se non vi è associazione significativa (almeno al 10%) non si può rifiutare l'ipotesi di indipendenza stocastica
- **Quindi la misura dell'associazione è priva di senso**
- Nel caso di un campione piccolo, è opportuno aumentare la numerosità (se possibile!) per verificare se permane la non significatività dell'associazione