Caso studio 9

Si consideri la seguente tabella che riporta i dati dei Laureati nel 2004 dei tre principali gruppi di corsi di laurea, per condizione occupazionale a tre anni dalla laurea (Fonte: ISTAT, Indagine sulla condizione occupazionale dei laureati, Anno 2008):

Corsi di laurea	Lavorano	Cercano lavoro	Non cercano lavoro	Totale
Economico-statistico	21.156	2.780	2.515	26.451
Giuridico	13.336	6.478	5.491	25.305
Ingegneria	16.526	799	835	18.160
Totale	51.018	10.057	8.841	69.916

Il tipo di corso di laurea influenza la possibilità di trovare lavoro? In che modo? Come si può sintetizzare il grado di dipendenza del carattere «Condizione occupazionale» dal tipo di corso di laurea (un indice sintetico permetterebbe di fare confronti: ad esempio, la crisi economica come ha modificato la situazione)?

Distribuzioni doppie

- Quando vengono considerate congiuntamente due colonne di una matrice di dati si ha una ***distribuzione doppia disaggregata*** (o *unitaria*). Si tratta dell'elencazione delle modalità di due caratteri (*X* e *Y*) osservate per ogni unità statistica del collettivo considerato:

$$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$$

Nome (*i*)	Sesso (x_i)	Regione (y_i)
Verdi M.	M	Marche
Bianchi C.	F	Calabria
Rossi V.	F	Umbria
Gialli L.	M	Piemonte
Grandi A.	F	Marche
Pini B.	F	Umbria

- L'informazione contenuta in una distribuzione doppia disaggregata è solitamente *sintetizzata* tramite una ***distribuzione doppia di frequenza*** che viene rappresentata tramite una tabella a doppia entrata in cui per ogni coppia di modalità dei due caratteri

$$(x_i, y_j), \quad i = 1, \cdots, H, \quad j = 1, \cdots, K$$

viene indicata la corrispondente *frequenza congiunta* (n_{ij}).

X/Y	y_1	y_2	…	y_j	…	y_K
x_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1K}
x_2	n_{21}	n_{22}	…	n_{2j}	…	n_{2K}
…	…	…	…	…	…	…
x_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{iK}
…	…	…	…	…	…	…
x_H	n_{H1}	n_{H2}	…	n_{Hj}	…	n_{HK}

Quando il carattere è quantitativo con molte modalità (tipicamente continuo) possono essere utilizzate le ***classi*** al posto delle modalità.

Esempi

- Si consideri la seguente *distribuzione doppia disaggregata* per due caratteri qualitativi (*Sesso*, *Regione*)

Nome (*i*)	Sesso (x_i)	Regione (y_i)
Verdi M.	M	Marche
Bianchi C.	F	Calabria
Rossi V.	F	Umbria
Gialli L.	M	Piemonte
Grandi A.	F	Marche
Pini B.	F	Umbria

- La corrispondente *distribuzione doppia di frequenza* è

	Regione			
Sesso	Calabria	Marche	Umbria	Piemonte
M	0	1	0	1
F	1	1	2	0

- Esempio di distribuzione in cui il carattere quantitativo è *in classi*

	Reddito annuo (x 1.000€)		
Titolo di studio	0-10	10-30	30-100
Lic. Media	88	142	120
Diploma	9	39	38
Laurea	3	19	42

Distribuzioni marginali

- Sommando le frequenze congiunte *separatamente per ogni riga* si ottengono le ***frequenze marginali*** di *X* (colonna del totale) che corrispondono al numero di soggetti che presentano una certa modalità di questo carattere a prescindere dalla modalità di *Y*:

$$n_{i\circ} = \sum_{j=1}^{K} n_{ij}$$

- Analogamente, le ***frequenze marginali*** di *Y* si ottengono sommando le frequenze congiunte *separatamente per ogni colonna* (riga del totale):

$$n_{\circ j} = \sum_{i=1}^{H} n_{ij}$$

- La somma di tutte frequenze congiunte (o di tutte le frequenze marginali) corrisponde alla ***numerosità del collettivo***

$$n = \sum_{j=1}^{K} \sum_{i=1}^{H} n_{ij} = \sum_{j=1}^{K} n_{\circ j} = \sum_{i=1}^{H} n_{i\circ}$$

X/Y	y_1	y_2	…	y_j	…	y_K	Totale
x_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1K}	$n_{1\circ}$
x_2	n_{21}	n_{22}	…	n_{2j}	…	n_{2K}	$n_{2\circ}$
…	…	…	…	…	…	…	…
x_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{iK}	$n_{i\circ}$
…	…	…	…	…	…	…	…
x_H	n_{H1}	n_{H2}	…	n_{Hj}	…	n_{HK}	$n_{H\circ}$
Totale	$n_{\circ 1}$	$n_{\circ 2}$	…	$n_{\circ j}$	…	$n_{\circ K}$	n

- Associando a ogni modalità del carattere *X* la corrispondente frequenza marginale, si ottiene la ***distribuzione marginale*** di *X*. E' la stessa distribuzione che avremmo ottenuto osservando il carattere singolarmente.

Modalità (x_i)	Frequenze ($n_{i\circ}$)
x_1	$n_{1\circ}$
x_2	$n_{2\circ}$
…	…
x_i	$n_{i\circ}$
…	…
x_H	$n_{H\circ}$
Totale	n

- In modo analogo si ottiene la ***distribuzione marginale*** di *Y*.
- Entrambe le distribuzioni possono essere *lette* direttamente dalla tabella a doppia entrata, quando sono presenti i totali (***margini***) di riga e di colonna.

Esempi

- Alla prima distribuzione doppia considerata nell'esempio precedente

	Regione				
Sesso	Calabria	Marche	Umbria	Piemonte	Totale
M	0	1	0	1	2
F	1	1	2	0	4
Totale	1	2	2	1	6

si hanno le seguenti distribuzioni marginali di X e Y

Sesso (x_i)	Frequenze ($n_{i\circ}$)
M	2
F	4
Totale	6

Regione (y_j)	Frequenze ($n_{\circ j}$)
Calabria	1
Marche	2
Umbria	2
Piemonte	1
Totale	6

- Per la seconda distribuzione doppia considerata nell'esempio precedente,

	Reddito annuo (x 1.000€)			
Titolo di studio	0-10	10-30	30-100	Totale
Lic. Media	88	142	120	350
Diploma	9	39	38	86
Laurea	3	19	42	64
Totale	100	200	200	500

si hanno le seguenti distribuzioni marginali di X e Y

Titolo di studio (x_i)	Frequenze ($n_{i\circ}$)
Lic. media	350
Diploma	86
Laurea	64
Totale	500

Reddito (y_j)	Frequenze ($n_{\circ j}$)
0-10	100
10-30	200
30-100	200
Totale	500

Distribuzioni condizionate

- La ***distribuzione condizionata di Y dato X = x_i*** è la distribuzione di *Y* limitatamente ai soggetti che presentano la modalità x_i di *X*. Si ottiene associando a ogni modalità y_j di *Y* la frequenza congiunta di (x_i, y_j).

Modalità (y_j)	Frequenze (n_{ij})
y_1	n_{i1}
y_2	n_{i2}
…	…
y_j	n_{ij}
…	…
y_K	n_{iK}
Totale	$n_{i\circ}$

- Ogni riga della tabella a doppia entrata corrisponde a una distribuzione condizionata di *Y* per una certa modalità *X*.
- In modo analogo possono essere ottenute le ***distribuzioni condizionate di X dato Y = y_j*** . Ognuna di queste distribuzioni corrisponde a una diversa colonna della tabella a doppia entrata.

Distr. condizionate relative e percentuali

- Per la distribuzione condizionata di *Y* dato *X* = x_i, le frequenze relative e percentuali possono essere calcolate come

$$f_{j|i} = \frac{n_{ij}}{n_{i\circ}} \quad e \quad p_{j|i} = 100\frac{n_{ij}}{n_{i\circ}} = 100 f_{j|i}$$

- Associando alla distribuzione condizionata di *Y* dato *X* = x_i le corrispondenti frequenze relative (o percentuali) si ottiene la distribuzione condizionata relativa (o percentuale) di *Y* dato *X* = x_i. Questa distribuzione permette di capire come *X* influenza *Y*.

Modalità (y_j)	Frequenze (n_{ij})	Freq. relative ($f_{j\|i}$)	Freq. percentuali ($p_{j\|i}$)
y_1	n_{i1}	$f_{1\|i}$	$p_{1\|i}$
y_2	n_{i2}	$f_{2\|i}$	$p_{2\|i}$
…	…	…	…
y_j	n_{ij}	$f_{j\|i}$	$p_{j\|i}$
…	…	…	…
y_K	n_{iK}	$f_{K\|i}$	$p_{K\|i}$
Totale	$n_{i\circ}$	1	100

Esempio

- Distribuzioni condizionate del reddito al titolo di studio.

Licenza media

Reddito (y_j)	Frequenze (n_{ij})	Relative ($f_{j|i}$)	Percentuali ($p_{j|i}$)
0-10	88	0,2514	251,4
10-30	142	0,4057	405,7
30-100	120	0,3429	342,9
Totale	350	1	100

Diploma

Reddito (y_j)	Frequenze (n_{ij})	Relative ($f_{j|i}$)	Percentuali ($p_{j|i}$)
0-10	9	0,1047	104,7
10-30	39	0,4535	453,5
30-100	38	0,4419	441,9
Totale	86	1	100

Laurea

Reddito (y_j)	Frequenze (n_{ij})	Relative ($f_{j|i}$)	Percentuali ($p_{j|i}$)
0-10	3	0,0469	46,9
10-30	19	0,2969	296,9
30-100	42	0,6562	656,2
Totale	64	1	100

Esercizio

Si considerino i dati del Caso Studio 9.

Il tipo di corso di laurea influenza la possibilità di trovare lavoro? In che modo? Per rispondere si calcolino le distribuzioni del carattere «Condizione occupazionale» condizionatamente al corso di laurea.

Corsi di laurea	Lavorano	Cercano lavoro	Non cercano lavoro	Totale
Economico-statistico	21156	2780	2515	26451
Giuridico	13336	6478	5491	25305
Ingegneria	16526	799	835	18160
Totale	51018	10057	8841	69916

Corsi di laurea	Lavorano	Cercano lavoro	Non cercano lavoro	Totale
Economico-statistico	0,800	0,105	0,095	1,000
Giuridico	0,527	0,256	0,217	1,000
Ingegneria	0,910	0,044	0,046	1,000
Totale	0,730	0,144	0,126	1,000

Analisi dell'associazione tra caratteri

- Lo scopo principale dell'*analisi di una distribuzione doppia* è usualmente quello di stabilire se tra i due caratteri considerati esiste una relazione e se, in particolare, uno dei due (tipicamente X) ha ***influenza*** sull'altro (Y). Esempi:
  - relazione tra la *provincia di residenza* e *spesa per beni alimentari*;
  - relazione tra *voto di maturità* e *voto a un certo esame universitario*;
  - relazione tra sesso e reddito;
- Se X non ha alcuna influenza su Y, allora si dice che Y è ***indipendente*** da X. In termini statistici questa situazione si ha quando le distribuzioni condizionate di Y sono equivalenti per ogni modalità di X , cioè hanno le stesse frequenze relative:

$$f_{j|1} = f_{j|2} = \cdots = f_{j|H}, \qquad j = 1, \cdots, K$$

- Si può dimostrare che si ha ***indipendenza statistica*** se e solo se le frequenze congiunte osservate corrispondono alle *frequenze teoriche sotto indipendenza*

$$n'_{ij} = \frac{n_{io} n_{oj}}{n}, \qquad i = 1, \cdots, H, \qquad j = 1, \cdots, K$$

- La ***tabella di indipendenza*** si ottiene sostituendo a ogni frequenza osservata (n_{ij}) la corrispondente frequenza di indipendenza (n'_{ij}).
- Sotto ***indipendenza*** si hanno le stesse distribuzioni marginali di quelle osservate e la stessa frequenza totale

$$\sum_{j=1}^{K} n'_{ij} = n_{io}, \qquad i = 1, \cdots, H$$

$$\sum_{i=1}^{H} n'_{ij} = n_{oj}, \qquad j = 1, \cdots, K$$

$$\sum_{j=1}^{K} \sum_{i=1}^{H} n'_{ij} = n$$

Corsi di laurea	Lavorano	Cercano lavoro	Non cercano lavoro	Totale
Economico-statistico	21156	2780	2515	26451
Giuridico	13336	6478	5491	25305
Ingegneria	16526	799	835	18160
Totale	51018	10057	8841	69916

Corsi di laurea	Lavorano	Cercano lavoro	Non cercano lavoro	Totale
Economico-statistico	0,800	0,105	0,095	1,000
Giuridico	0,527	0,256	0,217	1,000
Ingegneria	0,910	0,044	0,046	1,000
Totale	0,730	0,144	0,126	1,000

Condizione di indipendenza:

$$\frac{n_{ij}}{n_{i\circ}} = \frac{n_{\circ j}}{n}, \forall i, j$$

Frequenze teoriche di indipendenza:

$$n'_{ij} = \frac{n_{i\circ} \cdot n_{\circ j}}{n}, \forall i, j$$

- Quando *Y non è indipendente* da *X*, *Y* dipende da *X* e quindi i due caratteri si dicono ***connessi***. In pratica, ciò accade ogni volta che la tabella osservata non coincide con quella di indipendenza.
- In particolare, *Y* ***dipende perfettamente*** da *X* quando la modalità di *X* determina automaticamente la modalità di *Y*. *Y* dipende perfettamente da *X* se $K \leq H$ e si ha una sola frequenza positiva in ogni riga della tabella a doppia entrata mentre le altre frequenze sono tutte nulle.

Y ***dipende perfettamente*** da *X*

	y_1	y_2
x_1	25	0
x_2	0	11
x_3	0	31

X ***dipende perfettamente*** da *Y*

	y_1	y_2	y_3
x_1	25	0	31
x_2	0	11	0

X ***dipende perfettamente*** da *Y* e viceversa

	y_1	y_2	y_3
x_1	25	0	0
x_2	0	0	11
x_3	0	31	0

Esempio

- Per la distribuzione doppia dei carattere *titolo di studio* e *reddito* si ha la seguente *distribuzione di indipendenza*

	Reddito annuo (x 1.000€)			
Titolo di studio	0-10	10-30	30-100	Totale
Lic. Media	70	140	140	350
Diploma	17,2	34,4	34,4	86
Laurea	12,8	25,6	25,6	64
Totale	100	200	200	500

- Siccome la tabella osservata non coincide con quella di indipendenza, i due caratteri sono dipendenti e quindi si può ragionevolmente ritenere che il *titolo di studio* influenzi il *reddito*

- Se la distribuzione doppia fosse come la seguente, si avrebbe *perfetta dipendenza* del *reddito* dal *titolo di studio*

	Reddito annuo (x 1.000€)			
Titolo di studio	0-10	10-30	30-100	Totale
Lic. Media	350	0	0	350
Diploma	0	86	0	86
Laurea	0	0	64	64
Totale	100	86	64	500

Esercizio

Si considerino i dati del Caso Studio 9.

Si costruisca la tabella ipotetica di perfetta indipendenza e un'ipotetica tabella di perfetta dipendenza della Condizione lavorativa dal titolo di studio.

Tabella di perfetta dipendenza

Corsi di laurea	Lavorano	Cercano lavoro	Non cercano lavoro	Totale
Economico-statistico	26451	0	0	26451
Giuridico	0	0	25305	25305
Ingegneria	0	18160	0	18160
Totale	26451	18160	25305	69916

Corsi di laurea	Lavorano	Cercano lavoro	Non cercano lavoro	Totale
Economico-statistico	21156	2780	2515	26451
Giuridico	13336	6478	5491	25305
Ingegneria	16526	799	835	18160
Totale	51018	10057	8841	69916

Tabella di perfetta indipendenza

Corsi di laurea	Lavorano	Cercano lavoro	Non cercano lavoro	Totale
Economico-statistico	19301,41	3804,819	3344,775	26451
Giuridico	18465,17	3639,973	3199,861	25305
Ingegneria	13251,43	2612,208	2296,364	18160
Totale	51018	10057	8841	69916

Misura della connessione tra *Y* e *X*

- Il livello di connessione tra i due caratteri è tanto più elevato quanto più la tabella osservata si discosta da quella di indipendenza. Per misurare il livello di connessione si fa quindi uso delle ***contingenze*** (***assolute***)

$$c_{ij} = n_{ij} - n'_{ij} = n_{ij} - \frac{n_{io} n_{oj}}{n}, \qquad i = 1, \cdots, H, \qquad j = 1, \cdots, K$$

- La ***tabella delle contingenze*** si ottiene indicando in una tabella a doppia entrata la contingenza (c_{ij}) corrispondente a ogni coppia di modalità (x_i , y_j). Un'importante proprietà di questa tabella è che la somma delle celle in ogni riga o colonna della tabella è nulla

$$\sum_{j=1}^{K} c_{ij} = 0, \qquad i = 1, \cdots, H \quad \text{e} \quad \sum_{i=1}^{H} c_{ij} = 0, \qquad j = 1, \cdots, K$$

- Dividendo le *contingenze assolute* per le corrispondenti frequenze sotto indipendenza si ottengono le ***contingenze relative***

$$\frac{c_{ij}}{n'_{ij}} = \frac{n_{ij} - n'_{ij}}{n'_{ij}}, \qquad i = 1, \cdots, H, \qquad j = 1, \cdots, K$$

- Un indice sintetico di connessione è l'***indice chi-quadro*** che è una *somma ponderata* delle contingenze relative al quadrato

$$\chi^2 = \sum_{i=1}^{H}\sum_{j=1}^{K}\left(\frac{c_{ij}}{n'_{ij}}\right)^2 n'_{ij} = \sum_{i=1}^{H}\sum_{j=1}^{K}\frac{c_{ij}^2}{n'_{ij}}$$

- Una *formula alternativa*, che non richiede il calcolo delle contingenze, per l'indice *chi-quadro* è

$$\chi^2 = n\left(\sum_{i=1}^{H}\sum_{j=1}^{K}\frac{n_{ij}^2}{n_{i\circ}n_{\circ j}} - 1\right)$$

25

- L'indice ***chi-quadro*** assume valori tra 0 (nel caso di indipendenza) e

$$n \cdot \min[(H-1),(K-1)]$$

nel caso di perfetta dipendenza di *Y* da *X* o di *X* da *Y*. Quindi è un indice di connessione bilaterale.

- E' quindi possibile costruire un indice *relativo* (che varia tra 0 e 1), chiamato ***indice di connessione di Cramér***, come

$$C = \sqrt{\frac{\chi^2}{n \cdot \min[(H-1),(K-1)]}}$$

26

Esempio

- Per la distribuzione doppia dei carattere *titolo di studio* e *reddito* si ha la seguente *tabella delle contingenze assolute*

	Reddito annuo (x 1.000€)			
Titolo di studio	0-10	10-30	30-100	Totale
Lic. Media	18	2	-20	0
Diploma	-8,2	4,6	3,6	0
Laurea	-9,8	-6,6	16,4	0
Totale	0	0	0	0

- Il numero di soggetti con *reddito elevato* e *licenza media* è minore di quello atteso sotto indipendenza (c_{13} = -20) mentre quello dei soggetti con *laurea* è superiore (c_{33} = 16,4).
- L'indice di connessione è pari a

$$\chi^2 = 32{,}13$$

- Il massimo dell'indice χ^2 è 2×500 = 1000 e quindi l'indice di Cramér è

$$C = \sqrt{32{,}13/1000} = 0{,}179$$

27

Esercizio

Si considerino i dati del Caso Studio 9.

Si misuri il grado di dipendenza della Condizione occupazionale dal Corso di laurea attraverso l'indice di connessione di Cramer.

28

Corsi di laurea	Lavorano	Cercano lavoro	Non cercano lavoro	Totale
Economico-statistico	21156	2780	2515	26451
Giuridico	13336	6478	5491	25305
Ingegneria	16526	799	835	18160
Totale	51018	10057	8841	69916

Tabella di perfetta indipendenza

Corsi di laurea	Lavorano	Cercano lavoro	Non cercano lavoro	Totale
Economico-statistico	19301,41	3804,819	3344,775	26451
Giuridico	18465,17	3639,973	3199,861	25305
Ingegneria	13251,43	2612,208	2296,364	18160
Totale	51018	10057	8841	69916

$$\chi^2 = \frac{(21156 - 19301{,}41)^2}{19301{,}41} + \frac{(2780 - 3804{,}819)^2}{3804{,}819} + \frac{(2515 - 3344{,}775)^2}{3344{,}775} + \cdots + \frac{(835 - 2296{,}364)^2}{2296{,}364} = 8935{,}85$$

$$C = \sqrt{\frac{\chi^2}{n \cdot \min[(H-1),(K-1)]}} = \sqrt{\frac{8935{,}85}{69916 \cdot \min[(3-1),(3-1)]}} = 0{,}253$$

29

Esercizio

La tabella che riporta i dati dei Laureati nel 2014 dei tre principali gruppi di corsi di laurea, per condizione occupazionale a tre anni dalla laurea (Fonte: AlmaLaurea, Indagine sulla condizione occupazionale dei laureati, Anno 2015):

Corsi di laurea	Lavorano	Cercano lavoro	Non cercano lavoro	Totale
Economico-statistico	8.184	1.593	570	10.347
Giuridico	4.360	3.598	2.064	10.022
Ingegneria	8.643	859	849	10.351
Totale	21.187	6.050	3.483	30.720

Utilizzando l'indice di Cramer, si confronti l'associazione tra i due caratteri osservata nel 2014 con quella osservata nel 2008.

30

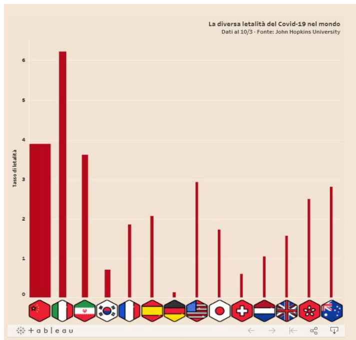
Associazione spuria

- Può accadere che due caratteri logicamente indipendenti mostrino invece un'associazione statistica. Questo può dipendere, ad esempio, dal fatto che entrambi i caratteri siano associati ad un terzo carattere.
  - Esempio: Gli annegamenti aumentano quando aumentano le vendite di gelato. In realtà l'aumento delle temperature può indurre più persone a nuotare, oltre a comprare più gelato.
  - Esempio: I paesi europei con più cicogne sono anche quelli in cui nascono più bambini.

31

- Può accedere anche che, sulla base dei dati osservati, una relazione tra due fenomeni appaia modificata, o perfino invertita, a causa di altri fenomeni non presi in considerazione nell'analisi (**paradosso di Simpson**).

Esempio: La letalità del Covid 19 in Italia



l'Istituto Superiore di Sanità ha spiegato che «Per tutte le fasce d'età il tasso di letalità da Covid-19 in Italia è inferiore a quello che si registra attualmente in Cina». E riporta i dati: quelli italiani al 4 marzo, quelli cinesi al 24 febbraio (ultimo dato disponibile al momento dell'analisi). Tra gli zero e i 69 anni, il tasso era dello 0,5% in Italia e l'1,3% in Cina; tra i 70 e i 79 anni, rispettivamente del 5,3% e dell'8%; per gli over 80, del 10,9% tra gli italiani e del 14,8% tra i cinesi. Se però si fa riferimento al tasso generale che comprende tutte le età, alla data dello studio il tasso di mortalità era rovesciato: in Italia risultava del 3,5% e in Cina del 2,3%.

La ragione di questa apparente contraddizione è che l'Italia ha un'età media superiore a quella della Cina, 44,3 anni contro i 37,4 della Cina; ha un maggior numero di anziani con più di 80 anni e in quella fascia si concentra il maggior numero di eventi letali.

32

Esempio: L'effetto della razza sulle condanne a morte

Uno studio del 1991 sull'effetto della razza sulle condanne a morte per i colpevoli di omicidio in Florida ha evidenziato che erano stati condannati a morte 53 imputati caucasici su 483 e 15 imputati afroamericani su 191;[9] ossia, come mostrato nella Tabella 4(a), la percentuale delle condanne a morte era più alta tra i caucasici (11,0 per cento) che non tra gli afroamericani (7,9).

Ora, però, se teniamo conto anche della razza della *vittima* oltre che di quella dell'imputato, ne emerge un quadro differente che, anche in questo caso, ci lascia perplessi.

IMPUTATO	
Caucasico	*Afroamericano*
53/483 = 11,0%	15/191 = 7,9%

(b)

VITTIMA	IMPUTATO	
	Caucasico	*Afroamericano*
Caucasica	53/467 = 11,3%	11/48 = 22,9%
Afroamericana	0/16 = 0,0%	4/143 = 2,8%

33

Esempio: Discriminazione di genere nell'ammissione a Berkley?

Nel 1973, presso la sede dell'Università della California di Berkeley, è stato effettuato uno studio sulla discriminazione di genere nelle ammissioni ai corsi universitari. È stato considerato un campione di 8442 ragazzi e 4321 ragazze che avevano fatto domanda di ammissione per i master post-laurea. Circa il 44% dei richiedenti di sesso maschile aveva ottenuto l'ammissione, contro il 35% delle donne richiedenti.

Ammissioni a Berkeley (1973)

	Ammessi	Rifiutati
Ragazzi	3738 (44.27%)	4704 (55.73%)
Ragazze	1494 (34.57%)	2827 (65.43%)

La differenza tra le percentuali di ammissione di uomini e donne sembrava essere una forte prova empirica della preferenza, da parte dell'università, per i candidati maschi. Per appurare ciò vennero analizzati singolarmente i differenti dipartimenti dell'Università.

La tabella di contingenza mostra che, per ciascuna divisione, la percentuale di candidature ammesse per gli uomini e per le donne è circa uguale. La variabile nascosta non presa in considerazione consiste nella scelta per la specializzazione. Le percentuali mostrano che era più facile essere accettati nei dipartimenti A e B (per i quali hanno applicato soprattutto maschi), rispetto a quelli C, D, E ed F (per i quali hanno applicato la maggioranza delle femmine).

Data on admissions to post-graduate courses in the six main departments (branches of specialization) of the University of California in Berkeley.[17]

Department[18]	Men		Women	
	Number of Applications	% of those admitted	Number of Applications	% of those admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%

Graduate Division - University of California - Berkeley

34

Esempio: I sopravvissuti del Titanic

Nel salvataggio, i passeggeri di terza classe hanno subito discriminazione?

Classe	Sesso	Età	Morti	Sopravvissuti
1^a	Uomini	Bambini	0	5
		Adulti	118	57
	Donne	Bambini	0	1
		Adulti	4	140
2^a	Uomini	Bambini	0	11
		Adulti	154	14
	Donne	Bambini	0	13
		Adulti	13	80
3^a	Uomini	Bambini	35	13
		Adulti	387	75
	Donne	Bambini	17	14
		Adulti	89	76
Equipaggio	Uomini	Bambini	0	0
		Adulti	670	192
	Donne	Bambini	0	0
		Adulti	3	20

Class	Saved	Lost	Total	Survival Rate
Third	151	476	627	24.08%
Crew	212	673	885	23.95%

Table 1: Numbers of adult crew members and third class passengers saved following the sinking of the RMS Titanic

	Men				Women			
Class	Saved	Lost	Total	Survival Rate	Saved	Lost	Total	Survival Rate
Third	75	387	462	16.23%	76	89	165	46.06%
Crew	192	670	862	22.27%	20	3	23	86.96%

35

Table 2: Numbers of adult crew members and third class passengers saved following the sinking of the RMS Titanic, by gender