

Facoltà di Scienze Politiche Università di Macerata

Corso di

Statistica Sociale

L'analisi bivariata

docente: Cristina Davino

a.a.: 2009-2010

Prof.ssa C. Davino

UNIMC
scienze politiche

Corso di Statistica Sociale

L'analisi bivariata

Distribuzione unitaria multipla

	Genere	Tipo diploma
1	Femmina	ITC
2	Femmina	ITC
3	Femmina	Classico
4	Femmina	ITC
5	Maschio	ITC
6	Femmina	Scientif.
7	Femmina	ITC
8	Femmina	Classico
9	Femmina	ITC
10	Femmina	ITC
11	Maschio	Scientif.
12	Femmina	ITC
13	Femmina	Scientif.
14	Femmina	ITC
15	Femmina	ITC
16	Maschio	Scientif.
17	Maschio	Scientif.
18	Femmina	ITC
19	Femmina	Scientif.
20	Maschio	ITC
21	Maschio	ITC
22	Maschio	ITC
23	Femmina	Scientif.
:	:	:

Distribuzioni di frequenza

Genere	Frequenza	%
Maschio	97	42,7
Femmina	130	57,3
Totale	227	100,0

Tipo diploma	Frequenza	%
Liceo classico	10	4,4
Liceo Scientifico	64	28,2
ITC	141	62,1
Altro	12	5,3
Totale	227	100,0

Tabella di contingenza

		Tipo diploma				
		Liceo classico	Liceo Scientifico	ITC	Altro	Totale
Genere	Maschio	6	30	55	6	97
	Femmina	4	34	86	6	130
Totale		10	64	141	12	227

L'analisi bivariata

- Relazioni statistiche o probabilistiche
- Non si individuano nessi di causa-effetto
- Distinzione tra variabili indipendenti e dipendenti

Variabile indipendente

Nominale *Cardinale*

Variabile
dipendente

Nominale

Cardinale

Tabelle di contingenza	
Analisi della varianza	Regressione e correlazione

Prof.ssa C. Davino

UNIMC
scienze politiche

Corso di Statistica Sociale

Tabelle di contingenza

Esempio Reddito / Consumi

		Consumi p.c.				Totale
		5-10 mila	10000-12500	12500-15000	15-20 mila	
Reddito p.c.	10-15 mila	275	151	14		440
	15-20 mila	28	151	165	14	358
	20-25 mila	14	14	413	96	537
	25-30 mila				83	83
Totale		317	316	592	193	1418

Prof.ssa C. Davino

UNIMC
scienze politiche

Corso di Statistica Sociale

Tabelle di contingenza

Esempio Reddito / Consumi

		Consumi p.c.				Totale
		5-10 mila	10000-12500	12500-15000	15-20 mila	
Reddito p.c.	10-15 mila	275	151	14		440
	15-20 mila	28	151	165	14	358
	20-25 mila	14	14	413	96	537
	25-30 mila				83	83
Totale		317	316	592	193	1418

○ Distribuzioni marginali

Tabelle di contingenza

Esempio Reddito / Consumi

		Consumi p.c.				Totale
		5-10 mila	10000-12500	12500-15000	15-20 mila	
Reddito p.c.	10-15 mila	275	151	14		440
	15-20 mila	28	151	165	14	358
	20-25 mila	14	14	413	96	537
	25-30 mila				83	83
Totale		317	316	592	193	1418

○ Distribuzioni marginali

○ Distribuzioni condizionate

Tabelle di contingenza

Esempio Reddito / Consumi

		Consumi p.c.				Totale
		5-10 mila	10000-12500	12500-15000	15-20 mila	
Reddito p.c.	10-15 mila	275	151	14		440
	15-20 mila	28	151	165	14	358
	20-25 mila	14	14	413	96	537
	25-30 mila				83	83
Totale		317	316	592	193	1418

○ Distribuzioni marginali

○ Distribuzioni condizionate

Analisi delle tabelle di contingenza

Esempio: Pratica religiosa / età

	18-34	35-54	Oltre 54	totale
Praticanti	223	313	182	718
Saltuari	266	317	88	671
Non praticanti	425	504	168	1097
totale	914	1134	438	2486

Tabella delle percentuali sul totale

	18-34	35-54	Oltre 54	totale
Praticanti	9,0	12,6	7,3	28,9
Saltuari	10,7	12,7	3,5	26,9
Non praticanti	17,1	20,3	6,8	44,2
totale	36,8	45,6	17,6	100

Analisi delle tabelle di contingenza

Esempio: Pratica religiosa / età

	18-34	35-54	Oltre 54	totale
Praticanti	31,1	43,6	25,3	100
Saltuari	39,6	47,3	13,1	100
Non praticanti	38,7	46,0	15,3	100
Totale	36,8	45,6	17,6	100

Tabella delle percentuali di riga

	18-34	35-54	Oltre 54	totale
Praticanti	24,4	27,6	41,5	28,9
Saltuari	29,1	28,0	20,1	26,9
Non praticanti	46,5	44,4	38,4	44,2
totale	100,0	100,0	100,0	100,0

Tabella delle percentuali di colonna

Analisi delle tabelle di contingenza

Un criterio

- Si scelgono i profili colonna quando si vuole analizzare l'influenza che la variabile posta in colonna ha sulla variabile posta in riga
- Si scelgono i profili riga quando si vuole analizzare l'influenza che la variabile posta in riga ha sulla variabile posta in colonna

Analisi delle tabelle di contingenza

Risposte alla domanda "Lei è soddisfatto dell'operato del governo?" per ampiezza del comune di residenza (in migliaia)

	>250	100-250	50-100	10-50	1-10	<1	totale
Per nulla	97	26	29	22	62	10	246
Poco	168	71	99	249	253	8	848
Non so	89	36	49	118	153	1	446
Abbastanza	159	66	92	246	326	5	894
Molto	3	2	2	8	14	1	30
Totale	516	201	271	643	808	25	2464

Tabella delle percentuali di colonna

	>250	100-250	50-100	10-50	<10
Per nulla	18,8	12,8	10,8	3,4	8,6
Poco	32,5	35,3	36,5	38,7	31,3
Non so	17,3	18,0	18,0	18,4	18,5
Abbastanza	30,8	32,9	34,0	38,3	39,8
Molto	0,6	1,0	0,7	1,2	1,8
Totale	100	100	100	100	100
N	516	201	271	643	833

Analisi delle tabelle di contingenza

Tabella delle percentuali di colonna

	>250	100-250	50-100	10-50	<10
Per nulla	18,8	12,8	10,8	3,4	8,6
Poco	32,5	35,3	36,5	38,7	31,3
Non so	17,3	18,0	18,0	18,4	18,5
Abbastanza	30,8	32,9	34,0	38,3	39,8
Molto	0,6	1,0	0,7	1,2	1,8
Totale	100	100	100	100	100
N	516	201	271	643	833

Tabella delle percentuali di colonna

	>250	100-250	50-100	10-50	<10
Per nulla + poco	51,3	48,1	47,3	42,1	39,9
Non so	17,3	18,0	18,0	18,4	18,5
Abbastanza + Molto	31,4	33,9	34,7	39,5	41,6
Totale	100	100	100	100	100
IDP	-19,9	-14,2	-12,6	-2,6	+1,7

Indice di differenza percentuale

Analisi delle tabelle di contingenza

	18-34	35-54	Oltre 54	totale
Praticanti	223	313	182	718
Saltuari	266	317	88	671
Non praticanti	425	504	168	1097
totale	914	1134	438	2486

Indipendenza

$$\frac{n_{ij}}{n_{i.}} = \frac{n_{.j}}{n_{..}} = \frac{n_{ij}}{n_{i.} \cdot n_{..} / n_{..}}$$

	18-34	35-54	Oltre 54	totale
Praticanti	31,1	43,6	25,3	100
Saltuari	39,6	47,3	13,1	100
Non praticanti	38,7	46,0	15,3	100
Totale	36,8	45,6	17,6	100

Le frequenze teoriche:
(in caso di indipendenza)

$$\frac{n_{ij}}{n_{i.}} = \frac{n_{.j}}{n_{..}} \Rightarrow \hat{n}_{ij} = \frac{n_{i.} \times n_{.j}}{n_{..}}$$

L'indice
chi-quadrato

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

Analisi delle tabelle di contingenza

	18-34	35-54	Oltre 54	totale
Praticanti	223	313	182	718
Saltuari	266	317	88	671
Non praticanti	425	504	168	1097
totale	914	1134	438	2486

Le frequenze teoriche:
(in caso di indipendenza)

	18-34	35-54	Oltre 54	totale
Praticanti	264,0	327,5	126,5	718
Saltuari	246,7	306,1	118,2	671
Non praticanti	403,3	500,4	193,3	1097
Totale	914	1134	438	2486

L'indice
chi-quadrato

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}} = \frac{(223 - 264,0)^2}{264,0} + \dots = 45,47$$

Il test del chi-quadrato

L'indice
chi-quadrato

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}} = \frac{(223 - 264,0)^2}{264,0} + \dots = 45,47$$

IL valore dell'indice è significativamente diverso da zero?

Ipotesi

H_0 : indipendenza

H_1 : no indipendenza

Livello di
significatività

$\alpha = 0.05$

Statistica
test

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

Regola di decisione:

Rifiutiamo H_0 , con una probabilità di errore pari a α , se $\chi^2_{oss} > \chi^2_c$
 χ^2_c è il valore critico che si trova nella coda di destra della distribuzione con $(r-1)(c-1)$ gdl



Il test del chi-quadrato

Ipotesi

H_0 : indipendenza

H_1 : no indipendenza

Livello di
significatività

$\alpha = 0.05$

$\chi^2_c = 9,488$

Statistica
test

$\chi^2 = 45,47$

Rifiutiamo H_0

P-value

$$p\text{-value} = P(\chi^2 > 45,47) \approx 0,001$$

Più piccolo è il p-value, più improbabile appare l'ipotesi nulla!



Il test del chi-quadrato

L'effetto della numerosità del campione

	18-34	35-54	Oltre 54	totale
Praticanti	22	31	18	71
Saltuari	27	32	9	68
Non praticanti	43	50	17	110
totale	92	113	44	249

Ipotesi

H_0 : indipendenza

H_1 : no indipendenza

Livello di significatività

$\alpha = 0.05$



$\chi_c^2 = 9,488$



Statistica test

$\chi^2 = 4,547$



Accettiamo H_0

Misure di associazione

L'indice **chi-quadrato**:

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}} \rightarrow \max(\chi^2) = n \times \min[(k-1); (h-1)]$$

L'indice **phi-quadrato**:

$$\Phi^2 = \frac{\chi^2}{n} \rightarrow \max(\Phi^2) = \min[(k-1); (h-1)]$$

L'indice **V di Cramer**:

$$V = \sqrt{\frac{\Phi^2}{\min[(k-1); (h-1)]}} = \sqrt{\frac{\chi^2}{n \cdot \min[(k-1); (h-1)]}}$$

Misure di associazione basate sulla riduzione proporzionale dell'errore

Date due variabili X e Y, la previsione di una modalità di Y può avvenire in due modi:

- Non si ha alcuna informazione aggiuntiva
- Si conosce la corrispondente modalità di X

PRE = $\frac{\text{Probabilità di errore nel caso (a)} - \text{Probabilità di errore nel caso (b)}}{\text{Probabilità di errore nel caso (a)}}$
(Proportional Reduction Error)

λ e τ di Goodman e Kruskal

- 0: la conoscenza della modalità di X non migliora la previsione sulla modalità di Y (Y e X sono indipendenti)
- 1: attraverso la conoscenza della modalità di X è possibile prevedere perfettamente la modalità di Y (Y dipende perfettamente da X)

Misure di cograduazione fra variabili ordinali

- Forza o intensità della relazione
- Segno (positivo o negativo) della relazione

	Basso	Medio	Alto
Alto	20	40	150
Medio	45	250	30
Basso	400	50	15

	Basso	Medio	Alto
Alto	250	75	45
Medio	45	150	60
Basso	30	45	300

Le misure di cograduazione si basano sul confronto fra i valori assunti dalle variabili X e Y su tutte le possibili coppie di casi:

- coppia di casi concordante
- coppia di casi discordante
- coppia di casi appaiata

Misure di cograduazione fra variabili ordinali

Le misure di cograduazione si basano sul confronto fra i valori assunti dalle variabili X e Y su tutte le possibili coppie di casi:

- coppia di casi concordante
- coppia di casi discordante
- coppia di casi appaiata

γ di
Goodman e
Kruskal

$$\gamma = \frac{P - Q}{P + Q}$$

+1: perfetta relazione positiva

0: assenza di relazione

-1: perfetta relazione negativa

Rapporti di probabilità (odds)

		Totale
Atteggiamento verso la pena di morte	Favorevoli	1188
	Contrari	604
	Totale	1792

Proporzione di soggetti favorevoli:

$$p = \frac{1188}{1792} = 0,663$$

Rapporto di probabilità:

$$\varpi = \frac{1188}{604} = 1,97$$

"ci sono quasi due favorevoli per ogni persona contraria"

Rapporto fra la probabilità che un individuo, estratto a caso dalla popolazione, appartenga ad una categoria e la probabilità che non vi appartenga

$$\varpi = \frac{f_1}{f_2} = \frac{p_i}{1 - p_i} \quad [0, +\infty[$$

- $\varpi=1$: le due categorie della variabile hanno lo stesso peso

Rapporti di probabilità e di associazione

		Istruzione		Totale
		Inferiore	Superiore	
Atteggiamento verso la pena di morte	Favorevoli	a 1027	b 161	1188
	Contrari	c 397	d 207	604
	Totale	1424	368	1792

Proporzioni condizionate:

$$p_1 = \frac{1027}{1424} = 0,721$$

$$p_2 = \frac{161}{368} = 0,438$$

Rapporti di probabilità condizionati:

$$\varpi_1 = \frac{1027}{397} = 2,59$$

$$\varpi_2 = \frac{161}{207} = 0,77$$

Rapporti di associazione (odds ratio):

$$\frac{\varpi_1}{\varpi_2} = \frac{a/c}{b/d} = \frac{ad}{bc} = \frac{1027 \cdot 207}{397 \cdot 161} = 3,33$$

Rapporti di probabilità e di associazione

Rapporti di associazione (odds ratio):

$$\frac{\varpi_1}{\varpi_2} = \frac{a/c}{b/d} = \frac{ad}{bc} \quad [0, +\infty[$$

$$\frac{\varpi_1}{\varpi_2} > 1 \Rightarrow$$

Associazione positiva tra le variabili
(i soggetti della categoria X_1 hanno probabilità di collocarsi nella categoria Y_1 maggiore di quanto sia la probabilità dei soggetti nella categoria X_2)

$$\frac{\varpi_1}{\varpi_2} < 1 \Rightarrow$$

Associazione negativa tra le variabili

- L'odds ratio cambia se si inverte l'ordine delle righe o delle colonne
- L'odds ratio resta immutato se si cambia l'orientamento della tavola
- L'odds ratio non risente della dimensione del campione
- L'odds ratio non cambia se le frequenze di una riga o di una colonna sono moltiplicate per una costante
- Gli odds ratio possono essere calcolati solo su tabelle 2@2

L'analisi della varianza (ANOVA)

X Zona geografica → Mutabile indipendente

Y Reddito p.c. → Variabile dipendente

Conteggio		Reddito p.c. (in euro)				Totale
		10-15mila	15-20mila	20-25mila	25-30mila	
Zona geografica	Nord		7	34	5	46
	Centro	1	18	5	1	25
	Sud e Isole	31	1			32
Totale		32	26	39	6	103

Medie condizionate:

$$M(Y|X=\text{Nord}) = \frac{1}{n_{X_1}} \sum_{j=1}^k y_j \cdot n_{1,j}$$

$$= \frac{12,5 \times 0 + 17,5 \times 7 + 22,5 \times 34 + 27,5 \times 5}{46}$$

$$= \frac{1025}{46} = 22,28$$

L'analisi della varianza (ANOVA)

X Zona geografica → Mutabile indipendente

Y Reddito p.c. → Variabile dipendente

Conteggio		Reddito p.c. (in euro)				Totale
		10-15mila	15-20mila	20-25mila	25-30mila	
Zona geografica	Nord		7	34	5	46
	Centro	1	18	5	1	25
	Sud e Isole	31	1			32
Totale		32	26	39	6	103

Y è indipendente in media da X se al variare delle modalità di X le medie condizionate di Y rimangono costanti.

Medie condizionate:

$$M(Y|X=\text{Nord}) = 22,3$$

$$M(Y|X=\text{Centro}) = 18,7$$

$$M(Y|X=\text{Sud}) = 12,7$$

Media generale:

$$M(Y) = \frac{1}{n} \sum_{j=1}^h y_j \cdot n_{.j}$$

$$= \frac{(12,5 \times 32 + 17,5 \times 26 + 22,5 \times 39 + 27,5 \times 6)}{103}$$

$$= 18,4$$

$$M(Y) = 22,3 \times \frac{46}{103} + 18,7 \times \frac{25}{103} + 12,7 \times \frac{32}{103}$$

L'analisi della varianza (ANOVA)

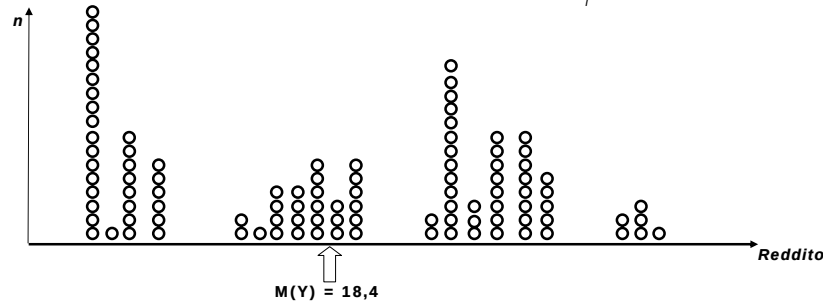
X Zona geografica → Mutabile indipendente

Y Reddito p.c. → Variabile dipendente

Conteggio		Reddito p.c. (in euro)				Totale
		10-15mila	15-20mila	20-25mila	25-30mila	
Zona geografica	Nord					46
	Centro					25
	Sud e Isole					32
Totale		32	26	39	6	103

$$M(Y) = 18,4$$

$$Dev(Y) = \sum_i (y_i - \bar{y})^2 \cdot n_i$$



L'analisi della varianza (ANOVA)

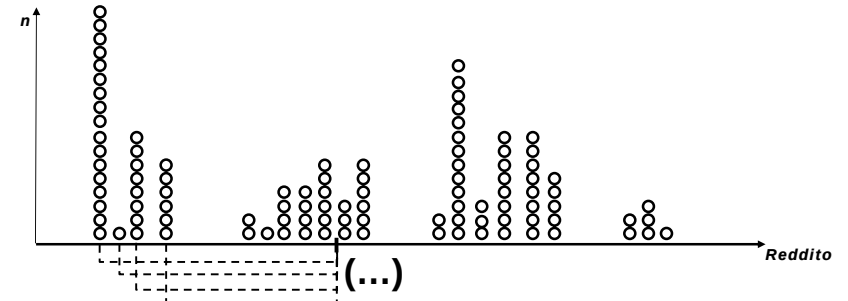
X Zona geografica → Mutabile indipendente

Y Reddito p.c. → Variabile dipendente

Conteggio		Reddito p.c. (in euro)				Totale
		10-15mila	15-20mila	20-25mila	25-30mila	
Zona geografica	Nord					46
	Centro					25
	Sud e Isole					32
Totale		32	26	39	6	103

$$M(Y) = 18,4$$

$$Dev(Y) = \sum_i (y_i - \bar{y})^2 \cdot n_i$$



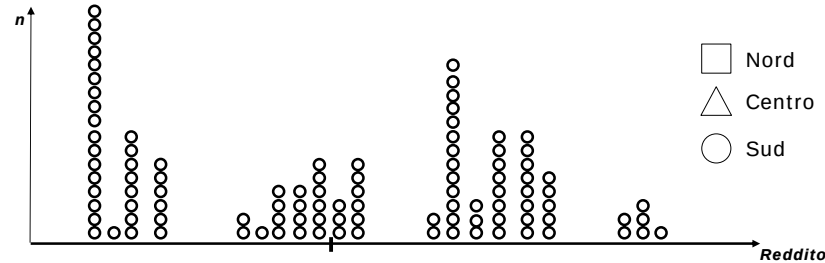
L'analisi della varianza (ANOVA)

- X** Zona geografica → Mutabile indipendente
Y Reddito p.c. → Variabile dipendente

Conteggio		Reddito p.c. (in euro)				
		10-15mila	15-20mila	20-25mila	25-30mila	Totale
Zona geografica	Nord		7	34	5	46
	Centro	1	18	5	1	25
	Sud e Isole	31	1			32
Totale		32	26	39	6	103

$$M(Y) = 18,4$$

$$Dev(Y) = \sum_i (y_i - \bar{y})^2 \cdot n_i$$



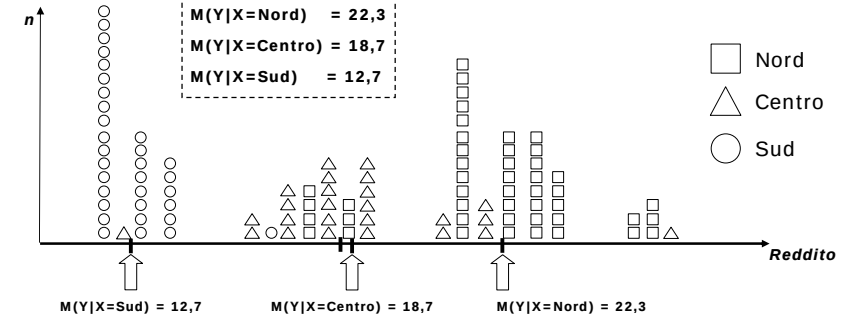
L'analisi della varianza (ANOVA)

- X** Zona geografica → Mutabile indipendente
Y Reddito p.c. → Variabile dipendente

Conteggio		Reddito p.c. (in euro)				
		10-15mila	15-20mila	20-25mila	25-30mila	Totale
Zona geografica	Nord		7	34	5	46
	Centro	1	18	5	1	25
	Sud e Isole	31	1			32
Totale		32	26	39	6	103

$$M(Y) = 18,4$$

$$Dev(Y) = \sum_i (y_i - \bar{y})^2 \cdot n_i$$



L'analisi della varianza (ANOVA)

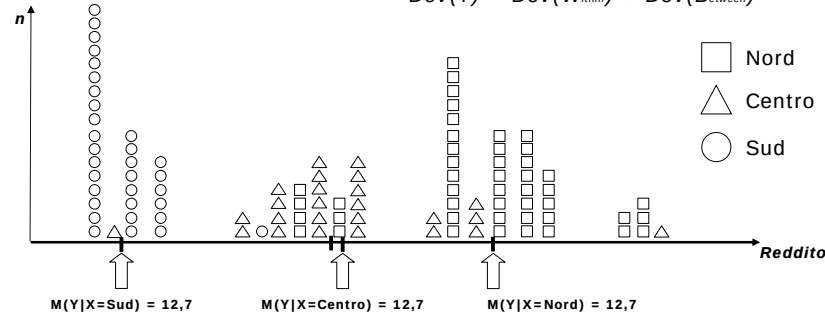
La decomposizione della devianza

Conteggio		Reddito p.c. (in euro)				
		10-15mila	15-20mila	20-25mila	25-30mila	Totale
Zona geografica	Nord		7	34	5	46
	Centro	1	18	5	1	25
	Sud e Isole	31	1			32
Totale		32	26	39	6	103

$$Dev(Y) = \sum_i (y_i - \bar{y})^2 \cdot n_i$$

$$Dev(Y) = Dev(W) + Dev(B)$$

$$Dev(Y) = Dev(W_{within}) + Dev(B_{between})$$



L'analisi della varianza (ANOVA)

- X** Zona geografica → Mutabile indipendente
Y Reddito p.c. → Variabile dipendente

Conteggio		Reddito p.c. (in euro)				
		10-15mila	15-20mila	20-25mila	25-30mila	Totale
Zona geografica	Nord		7	34	5	46
	Centro	1	18	5	1	25
	Sud e Isole	31	1			32
Totale		32	26	39	6	103

$$Dev(Y) = Dev(W) + Dev(B)$$

$$Dev(Y) = \sum_i (y_i - \bar{y})^2 \cdot n_i \Rightarrow \text{E' la **devianza totale**. Esprime la dispersione della nube di punti attorno alla media generale.}$$

$$Dev(W) = \sum_i \sum_j (y_{ij} - \bar{y}_i)^2 \cdot n_{ij} \Rightarrow \text{E' la **devianza Within**, o **entro i gruppi**. Esprime la dispersione dei k gruppi attorno alle rispettive medie. Si ottiene sommando le k devianze interne ai k gruppi.}$$

$$Dev(B) = \sum_i (\bar{y}_i - \bar{y})^2 \cdot n_i \Rightarrow \text{E' la **devianza Between** o **fra i gruppi**. Esprime la dispersione delle medie dei k gruppi attorno alla media generale.}$$

Una misura della forza della relazione

- X** Zona geografica → Mutabile indipendente
Y Reddito p.c. → Variabile dipendente

Conteggio

		Reddito p.c. (in euro)				
		10-15mila	15-20mila	20-25mila	25-30mila	Totale
Zona geografica	Nord		7	34	5	46
	Centro	1	18	5	1	25
	Sud e Isole	31	1			32
Totale		32	26	39	6	103

Il rapporto di correlazione: $\eta^2 = \frac{Dev(B)}{Dev(Y)}$
 (di Pearson)

M(Y X=N)	22,283
M(Y X=C)	18,700
M(Y X=S)	12,656
M(Y)	18,422
Dev(Y X=N)	297,83
Dev(Y X=C)	214,00
Dev(Y X=S)	24,22
Dev(B)	1751,33
Dev(TOT)	2287,38
Eta quadro	0,77

$$\eta^2 = \frac{Dev(B)}{Dev(Y)} = \frac{\sum_{i=1}^k (\bar{y}_i - \bar{y})^2 \cdot n_i}{\sum_{j=1}^h (y_j - \bar{y})^2 \cdot n_j}$$

Rapporto di correlazione (di Pearson): proporzione della devianza generale "spiegata" dalla variabile indipendente

Una misura della forza della relazione

- X** Zona geografica → Mutabile indipendente
Y Reddito p.c. → Variabile dipendente

Conteggio

		Reddito p.c. (in euro)				
		10-15mila	15-20mila	20-25mila	25-30mila	Totale
Zona geografica	Nord		7	34	5	46
	Centro	1	18	5	1	25
	Sud e Isole	31	1			32
Totale		32	26	39	6	103

Il rapporto di correlazione: $\eta^2 = \frac{Dev(B)}{Dev(Y)}$
 (di Pearson)

M(Y X=N)	22,283
M(Y X=C)	18,700
M(Y X=S)	12,656
M(Y)	18,422
Dev(Y X=N)	297,83
Dev(Y X=C)	214,00
Dev(Y X=S)	24,22
Dev(B)	1751,33
Dev(TOT)	2287,38
Eta quadro	0,77

$$\eta^2 = \frac{Dev(B)}{Dev(Y)} = \frac{\sum_{i=1}^k (\bar{y}_i - \bar{y})^2 \cdot n_i}{\sum_{j=1}^h (y_j - \bar{y})^2 \cdot n_j}$$

$$0 \leq \eta^2 \leq 1$$

◦ $\eta^2 = 0$: nessuna relazione tra le due variabili

◦ $\eta^2 = 1$: relazione perfetta

La significatività della relazione

• ANOVA

Verifica l'uguaglianza tra più valori medi attraverso stime diverse della variabilità il cui rapporto, sotto l'ipotesi H_0 di nessuna differenza tra i valori medi, si distribuisce come una v.c. di Fisher, con opportuni gradi di libertà.

• ANOVA

a 1, 2, ..., k fattori

La distinzione riguarda il numero di condizioni sperimentali, o trattamenti, che variano.

Un fattore può essere una caratteristica, una variabile, una qualità, una circostanza, .., di cui si vuole controllare l'effetto.

La significatività della relazione

A Fattore sperimentale di cui vogliamo verificare l'efficacia → **Trattamento**

Y Risultato misurato sull'unità statistica → **Risposta**

Trattamento A_1 : $y_{11}, y_{12}, \dots, y_{1n_1}$

Trattamento A_2 : $y_{21}, y_{22}, \dots, y_{2n_2}$

:

Trattamento A_k : $y_{k1}, y_{k2}, \dots, y_{kn_k}$

μ : valore medio della risposta per l'intero esperimento

μ_i : valore medio della risposta con il trattamento A_i

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k = \mu$$

La significatività della relazione

$$Dev(Y) = Dev(W) + Dev(B)$$

Fonte di variabilità	Devianza	Gradi di libertà	Varianza stimata	F- calcolata
Esterna	D_B	$k - 1$	$S_B^2 = \frac{D_B}{k - 1}$	$F = \frac{S_B^2}{S_W^2}$
Interna	D_W	$n - k$	$S_W^2 = \frac{D_W}{n - k}$	
Totale	D_T	$n - 1$		

La significatività della relazione

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k = \mu$$

$$F = \frac{S_B^2}{S_W^2}$$

Se H_0 è **falsa**: $\Rightarrow$ Il numeratore F sarà maggiore del denominatore il rapporto sarà > 1

Livello di Significatività: $\alpha = 0.05$



Regola di decisione:

Rifiutiamo H_0 , con una probabilità di errore pari a α , se $F > F_c$

Una misura della forza della relazione

X Zona geografica $\rightarrow$ Mutabile indipendente

Y Reddito p.c. $\rightarrow$ Variabile dipendente

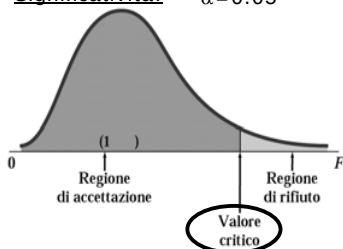
$$F = \frac{S_B^2}{S_W^2}$$

Conteggio		Reddito p.c. (in euro)				Totale
		10-15mila	15-20mila	20-25mila	25-30mila	
Zona geografica	Nord		7	34	5	46
	Centro	1	18	5	1	25
	Sud e Isole	31	1			32
Totale		32	26	39	6	103

		gdl	F
Dev(B)	1751.334	3-1	163.35
Dev(W)	536.050	103-3	
Dev(TOT)	2287.379		

Livello di Significatività: $\alpha = 0.05$

$$F_c = 3.10$$



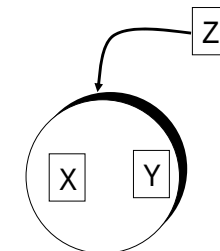
Regola di decisione:

Rifiutiamo H_0 , con una probabilità di errore pari a α , perché $F > F_c$

Introduzione di una terza variabile

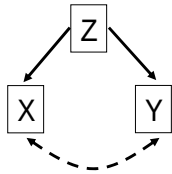
Molte relazioni bivariate sono ingannevoli: statisticamente sono inattaccabili ma la covarianza può essere dovuta all'azione di una terza variabile

- Relazione spuria
- Relazione indiretta
- Relazione condizionata



Relazione spuria

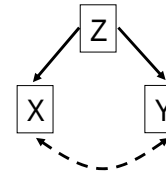
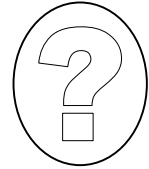
1. La successione oraria dei prezzi dei cavoletti di Bruxelles venduti nel mercato di Londra e la corrispondente altezza del Tamigi
2. Il numero di turisti canadesi che giungono a Roma e il numero di gelati venduti nella capitale
3. Il numero di case con nidi di cicogna sul tetto ed il numero di case in cui nascono più bambini



La relazione tra X e Y è provocata da una terza variabile Z che agisce causalmente sia su X che su Y

Relazione spuria

1. La successione oraria dei prezzi dei cavoletti di Bruxelles venduti nel mercato di Londra e la corrispondente altezza del Tamigi (Z=ora del giorno)
2. Il numero di turisti canadesi che giungono a Roma e il numero di gelati venduti nella capitale (Z=temperatura)
3. Il numero di case con nidi di cicogna sul tetto ed il numero di case in cui nascono più bambini (Z=collocazione urbano-rurale delle case)



- Z è la variabile di controllo
- Bisogna analizzare la relazione tra X e Y in sub-campioni nei quali Z sia costante