L'ANALISI BIVARIATA

Relazioni fra variabili

- Studio delle relazioni fra variabili. Cosa significa dire che c'è una relazione fra due o più variabili?

- Significa dire che c'è una «variazione concomitante» fra i loro valori, una co-variazione: per esempio, al variare del titolo di studio varia il reddito

1. Si tratta di relazioni statistiche, cioè relazioni di tipo probabilistico

2. La statistica ci può dire solo che esiste una relazione fra due variabili→Sarà compito e responsabilità del ricercatore di conferire a tale relazione il significato di nesso causale e di attribuire a essa una direzione

TAB. 3.1. Le tecniche di analisi bivariata

		VARIABILE INDIPENDENTE	
		Nominale	*Cardinale*
VARIABILE DIPENDENTE	*Nominale*	Tavole di contingenza	
	Cardinale	Analisi della varianza	Regressione e correlazione

Tipi di relazioni

Nome della relazione	Tipi di variabili messe in relazione
Concordanza	dicotomiche e/o variabili categoriali
Nessun nome	1 variabile cardinale e categoriale (dicotomica)
Covariazione/ controvariazione	variabili cardinali e/o variabili categoriali ordinate
1. Cograduazione	2 variabili categoriali ordinate
2. Correlazione	2 variabili cardinali

- **Le tavole di contingenza**
  - Percentuali di riga
  - Percentuali di colonna
  - Percentuali sul totale

TAB. 3.2. Pratica religiosa per età

	18-34	35-54	Oltre 54	Totale
a) Tabella dei valori assoluti (frequenze) di cella				
Praticanti	223	313	182	718
Saltuari	266	317	88	671
Non praticanti	425	504	168	1.097
Totale	914	1.134	438	2.486
b) Tabella delle percentuali di riga				
Praticanti	31,1	43,6	25,3	100,0
Saltuari	39,6	47,2	13,1	100,0
Non praticanti	38,7	45,9	15,3	100,0
c) Tabella delle percentuali di colonna				
Praticanti	24,4	27,6	41,6	
Saltuari	29,1	28,0	20,1	
Non praticanti	46,5	44,4	38,4	
Totale	100,0	100,0	100,0	
d) Tabella delle percentuali sul totale				
Praticanti	9,0	12,6	7,3	28,9
Saltuari	10,7	12,8	3,5	27,0
Non praticanti	17,1	20,3	6,8	44,1
Totale	36,8	45,6	17,6	100,0

Fonte: Itanes 1996.

- Si sceglie la percentuale ***di colonna*** quando si vuole analizzare l'influenza che la variabile posta in colonna ha sulla variabile posta in riga
- Si sceglie la percentuale ***di riga*** quando si vuole analizzare l'influenza che la variabile posta in riga ha sulla variabile posta in colonna
- Si definisce qual è la variabile indipendente e si percentualizza all'interno delle sue modalità

- Presentazione delle tavole
  - Parsimoniosità
  - Totali
  - Basi delle percentuali
  - Cifre decimali, decimale zero, arrotondamenti, quadratura
  - Intestazione

Gli Atteggiamenti verso la politica dei giovani di Napoli e Provincia

(% colonna - Base N= 996)

	Tot	Genere		Residenza			Età			
		M	F	Napoli	Comuni Prov. Nord	Comuni Prov. Sud	15-17	18-23	24-29	30-34
Mi considero politicamente impegnato	1,7	1,8	1,6	0,6	0,7	4,5	0,0	0,7	4,8	0,0
Mi tengo al corrente della politica, ma non partecipo	**51,7**	**51,4**	**52,0**	**60,4**	**47,6**	**48,1**	**34,9**	**47,6**	**58,9**	**55,9**
Bisogna lasciare la politica a chi ha più competenza di me	28,9	27,7	30,2	25,0	32,9	27,1	34,1	38,4	18,8	28,0
La politica mi disgusta	17,7	19,1	16,2	14,0	18,7	20,3	31,0	13,4	17,5	16,1
TOTALE	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0

Le frequenze attese

Come si è detto in precedenza, ogni cella contiene quattro tipi di informazioni:

- Frequenza osservata (f_o);
- Frequenza minima possibile;
- Frequenza massima possibile;
- Frequenza attesa (f_e).

*Si definisce frequenza **attesa** la frequenza che si otterrebbe in caso di indipendenza tra le variabili*.

- Essa è data dal ***prodotto dei marginali*** (di riga e di colonna) corrispondenti, ***diviso il numero totale*** dei casi: (vedi figura).

Dove:

- c_i = marginale di colonna della cella i
- r_i = marginale di riga della colonna i
- N = numero dei casi

$$F_e = \frac{c_i * r_i}{N}$$

Interpretazione frequenze

		Frequenze Osservate	
		Alte	***Basse***
Frequenze Attese	***Alte***	Affidabili ma poco interessanti	Affidabili e interessanti
	Basse	Non molto affidabili ma interessanti	Poco affidabili e poco interessanti

VALORI ASSOLUTI

OCCUPAZIONE	Conservatori	Centro sinistra	Nazionalisti	Tot. Riga
Pensionati	180	68	24	272
occupati	80	180	44	304
Casalinghe	140	46	10	196
Studenti	50	150	28	228
Tot.colonna	450	444	106	1000

PERCENTUALI TOTALI

OCCUPAZIONE	Conservatori	Centro sinistra	Nazionalisti	Tot. Riga
Pensionati	18%	7%	2%	27%
occupati	8%	18%	4%	30%
Casalinghe	14%	5%	1%	20%
Studenti	5%	15%	3%	23%
Tot.colonna	45%	44%	11%	100%

PERCENTUALE DI COLONNA

OCCUPAZIONE	Conservatori	Centro sinistra	Nazionalisti	Tot. Riga
Pensionati	40%	15%	23%	27%
occupati	18%	41%	42%	30%
Casalinghe	31%	10%	9%	20%
Studenti	11%	34%	26%	23%
Tot.colonna	100%	100%	100%	100%

PERCENTUALE DI RIGA

OCCUPAZIONE	Conservatori	Centro sinistra	Nazionalisti	Tot. Riga
Pensionati	66%	25%	9%	100%
occupati	26%	59%	14%	100%
Casalinghe	71%	23%	5%	100%
Studenti	22%	66%	12%	100%
Tot.colonna	45%	44%	11%	100%

OCCUPAZIONE	Conservatori	Centro sinistra	Nazionalisti	MARG RIGA
Pensionati	180	68	24	272
Occupati	80	180	44	304
Casalinghe	140	46	10	196
Studenti	50	150	28	228
MARG. COLONNA	450	444	106	1000

Frequenze osservate

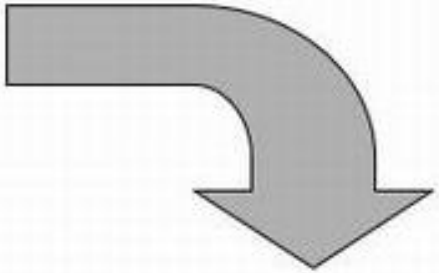
OCCUPAZIONE	Conservatori	Centro sinistra	Nazionalisti	MARG RIGA
Pensionati	122	121	29	272
Occupati	137	135	32	304
Casalinghe	88	87	21	196
Studenti	103	101	24	228
MARG. COLONNA	450	444	106	1000

Frequenze attese

OCCUPAZIONE	Conservatori	Centro sinistra	Nazionalisti	MARG RIGA
Pensionati	-58	53	5	0
Occupati	57	-45	-12	0
Casalinghe	-52	41	11	0
Studenti	53	-49	-4	0
MARG. COLONNA	0	0	0	0

Scarti

Frequenze osservate					
Pensionati	50	.....	.....	.....	261
Semi occupati	33	.....	.....	.....	303
.....	.....	.....	.....	.....	.....
Casalinghe	.....	.....	1	.....	56
Studenti	.....	6	.....	.....	88
MARG. COLONNA	391	22	76	.....	2085
Frequenze attese					
Pensionati	49	...	...	...	261
Semi occupati	56	...	...	...	303
...	...	...	...	...	...
Casalinghe	...	...	2	. ...	56
Studenti	...	0,93	...	...	88
MARG. COLONNA	391	22	76	...	2085



F_o

F_o	Alte	Basse
Alte	Pensionati 49-50 = -1 Affidabili ma poco interessanti	Semi occupati 56 – 33 = 23 Affidabili e interessanti
Basse	Studenti 0,93-6=-5,07 Non molto affidabili ma interessanti	Casalinghe 2 – 1 = +1 Poco affidabili e poco interessanti

Esercizio

		Religioso			
		No	Poco	Molto	
	Giovane	20	15	10	?
Età	Adulto	?	20	20	60
	Anziano	5	5	?	?
		?	?	45	?

1) Completare la tabella di contingenza, inserendo tutto ciò che manca al posto dei punti interrogativi
2) Individuare:
a) La percentuale di anziani che sono molto religiosi
b) La percentuali di poco religiosi in generale
c) La percentuale, sul totale, di giovani non religiosi

Quando si intende analizzare la relazione tra due variabili ***categoriali ordinate*** si impiegano le **misure di cograduazione.**

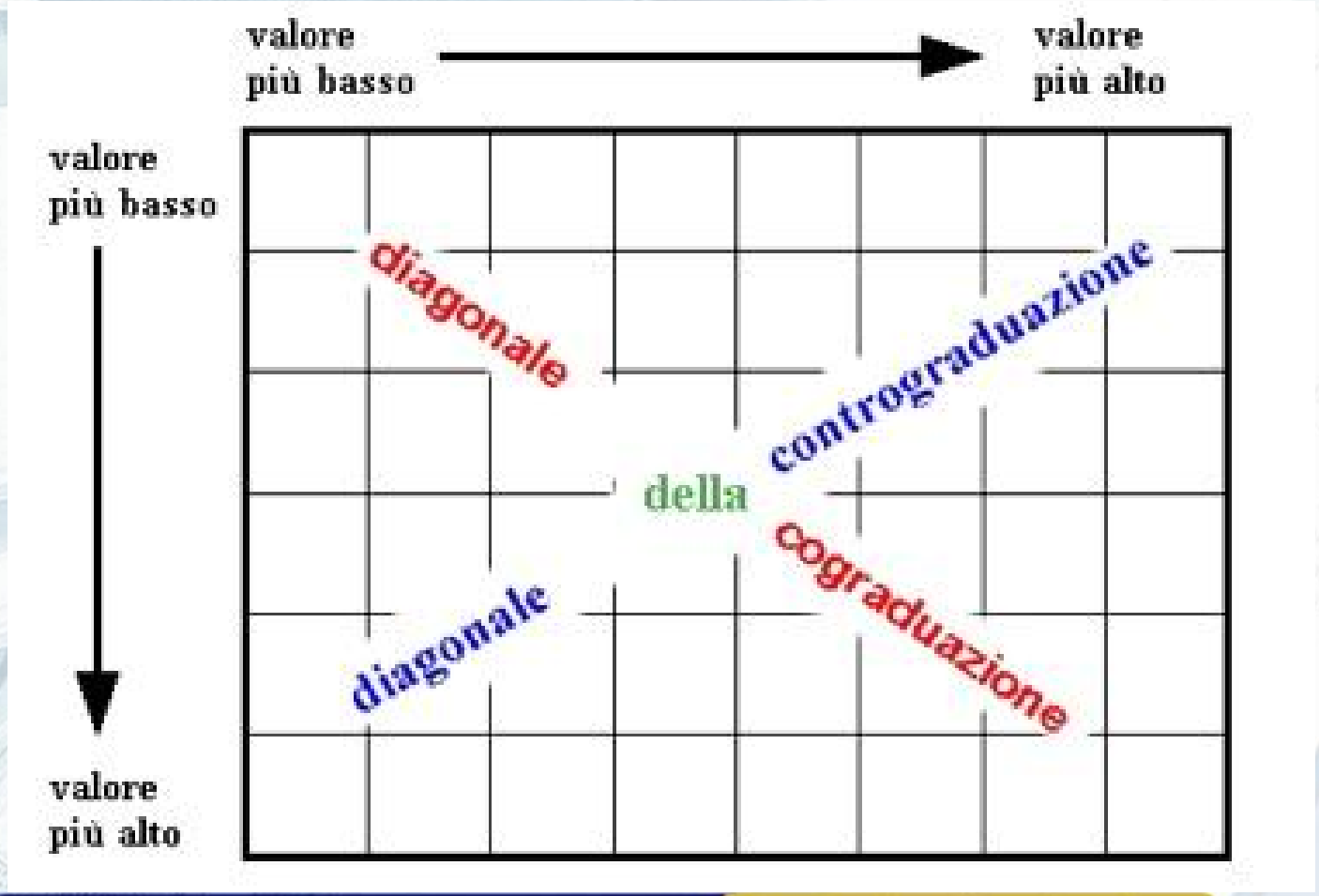
- **Co-variazione**: si ha quando a valori alti di una variabile corrispondono valori alti anche dell'altra. Ciò accade quando i valori sono direttamente proporzionali.
  Esempio: se all'aumentare del grado di fiducia nelle istituzioni aumenta anche il grado di soddisfazione nei confronti del loro operato.
- **Contro-variazione**: si ha quando a valori alti di una variabile corrispondono valori bassi dell'altra. In tal caso i valori delle due variabili sono inversamente proporzionali.
  Esempio: se all'aumentare del grado di fruizione nei confronti di un servizio diminuisce il grado di soddisfazione nei confronti della gestione dello stesso.

Numero di abitanti per comune (in classi)	Titolo di studio: nessuno	Lic. Elemen.	Lic. Media inf.	Lic. media sup.	Laurea	Totale
meno di 5.000	21	29	32	15	3	100
5.001-10.000	17	32	37	12	2	100
10.001-15.000	9	24	41	20	6	100
15.001-20.000	6	14	44	25	11	100
20.001-25.000	4	14	39	31	12	100
25.001-30.000	4	16	34	30	16	100
più di 30.000	5	15	30	29	21	100
Marg. di colonna	*8*	*19*	*35*	*26*	*12*	

Per stabilire se c'è covariazione o contro variazione è necessario prestare attenzione a come i dati si distribuiscono lungo le ***diagonali*** della tabella di contingenza .

È importante però tener presente che:

- non sempre la tabella è quadrata e quindi in tali casi è improprio parlare di "diagonali" o intendere con tale termine le diagonali vere e proprie;
- le diagonali della covariazione e della contro variazione si possono invertire qualora i valori non siano collocati in ordine crescente come in figura (dal più basso al più alto) ma in ordine decrescente.
- tra due variabili ordinali c'è cograduazione se le frequenze della tabella di contingenza si addensano o sulla diagonale della cograduazione o intorno ad essa. Lo stesso accade sulla o intorno alla diagonale della contrograduazione quando c'è contrograduazione.



- **Rappresentazioni grafiche della relazione fra due variabili nominali**
- Si utilizzano gli strumenti già visti per le distribuzioni di frequenza, e cioè sostanzialmente i *diagrammi a barre* oppure quelli a *linee spezzate* che congiungono i punti di interesse.

X^2

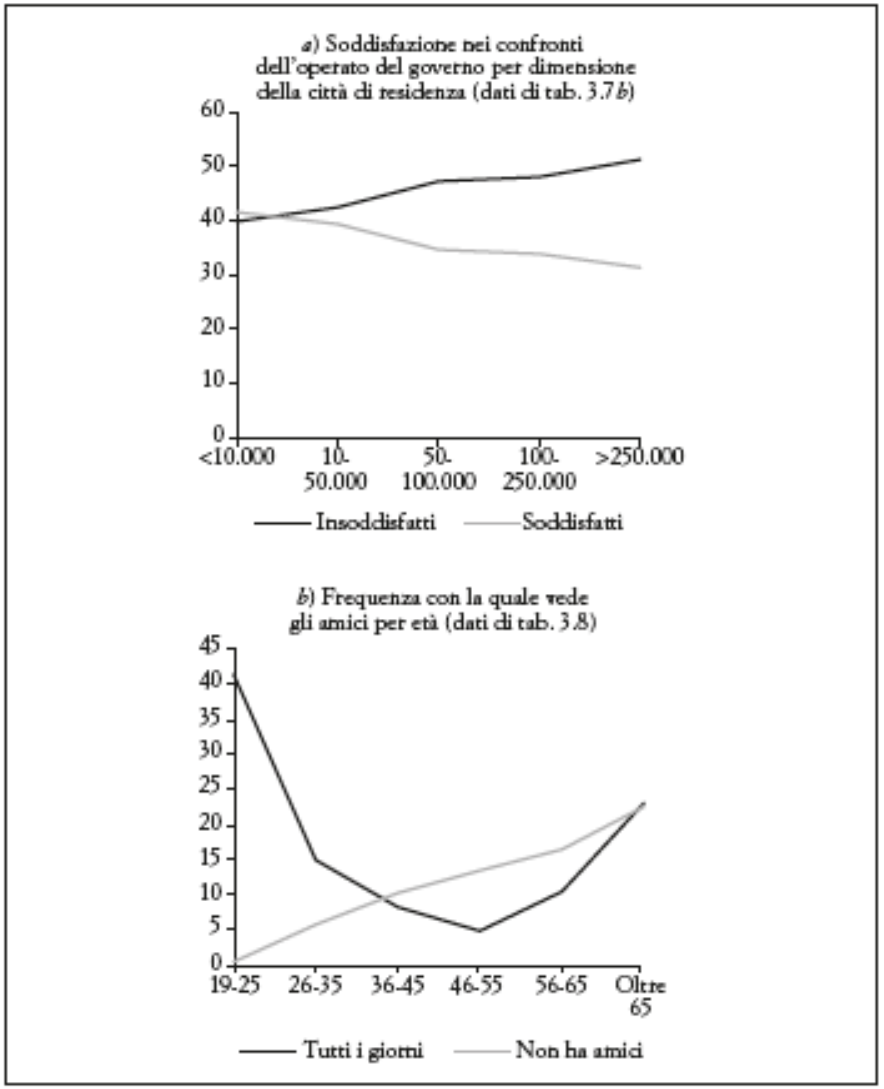


fig. 3.2. *Rappresentazioni grafiche di tavole di contingenza: spezzata.*

- Aprire il file **Gym**
- Incrociare **D4** con **D5** (tabella pivot)
- Commentare **l'eventuale presenza di co-(o contro-)variazione**
- Creare un grafico adeguato di presentazione

- **Significatività della relazione fra due variabili nominali: il test statistico del chi-quadrato (χ^2)**

- Test statistico di verifica delle ipotesi applicato al caso della relazione fra due variabili: formulare l'ipotesi nulla H_0 secondo la quale nella popolazione non esiste relazione fra le due variabili e dimostrare, dati alla mano, che essa è falsa: cioè che questa ipotesi non è compatibile (= è assai improbabile) con i dati di cui disponiamo.
- Se l'ipotesi nulla H_0 di assenza di relazione viene respinta, automaticamente resta accettata la sua alternativa, l'ipotesi di ricerca H_1 che sostiene l'esistenza della relazione.

- Prima di procedere all'analisi bivariata tra due variabili categoriali, è sempre opportuno analizzare le distribuzioni monovariate, per verificare che queste non siano eccessivamente squilibrate. Si deve sempre considerare anche l'opportunità di svolgere analisi bivariate in relazione al numero di casi.

Chi Quadrato: limiti

- Se le frequenze attese delle celle sono *inferiori a 5*, il valore del coefficiente diventa *inaffidabile*
- Per campioni di grandi dimensioni (sopra i *2000* casi) il valore del Chi quadrato sembra essere *sempre* significativo e quindi il test evidenzia spesso delle associazioni che vanno verificate
- Il Chi quadrato non mette in luce la forza della relazione tra due variabili, per tale scopo si usano i coefficienti di associazione statistica (bidirezionali e unidirezionali)
- Questi ultimi sono in genere normalizzati nell'intervallo 0-1 per le variabili categoriali non ordinate e (-1;+1) per tutte le altre.

TAB. 3.15. Frequenze osservate e frequenze attese sotto l'ipotesi nulla H_0 di indipendenza

	18-34		35-54		Oltre 54		Totale	
	v.a.	%	v.a.	%	v.a.	%	v.a.	%
a) Frequenze osservate								
Praticanti	223	24,4	313	27,6	182	41,6	718	28,9
Saltuari	166	29,1	317	28,0	88	20,1	671	27,0
Non praticanti	425	46,5	504	44,4	168	38,4	1.097	44,1
Totale	914	100,0	1.134	100,0	438	100,0	2.486	100,0
b) Frequenze attese sotto l'ipotesi di indipendenza								
Praticanti	264,0	28,9	327,5	28,9	126,5	28,9	718	28,9
Saltuari	246,7	27,0	306,1	27,0	118,2	27,0	671	27,0
Non praticanti	403,3	44,1	500,4	44,1	193,3	44,1	1.097	44,1
Totale	914	100,0	1.134	100,0	438	100,0	2.486	100,0

Calcolo della frequenza attesa per la cella (1,1): $f_e = 914 \cdot 718/2.486 = 264$

$$\chi^2 = \Sigma \frac{(f_o - f_e)^2}{f_e} = \frac{(223-264,0)^2}{264,0} + \frac{(313-327,5)^2}{327,5} + \frac{(182-126,5)^2}{126,5} + \ldots + \frac{(168-193,3)^2}{193,3} = 45,47$$

$p < 0,001$ (χ^2 significativo al livello dello 0,001);
$\Phi = 0,14$;
$V = 0,10$.

Fonte: Itanes.

$$\chi^2 = \Sigma \frac{(f_o - f_e)^2}{f_e}$$

- Per convenzione, noi respingiamo l'ipotesi nulla di indipendenza se $p \leq 0{,}05$, cioè se il valore del chi-quadrato è così grande da avere solo il 5% (o meno) di probabilità di essere dovuto al caso (cioè a errori casuali pur derivando da una popolazione dove c'è effettiva indipendenza)
- Tavola di distribuzione del χ^2 → In questa tavola abbiamo tante righe, cioè distribuzioni del χ^2, quanti sono i gradi di libertà della tabella.
- I gradi di libertà della tabella si determinano nel seguente modo:

gradi di libertà gl = (n. righe – 1) (n. colonne – 1)

VALORI PERCENTILI (χ^2_P) PER LA DISTRIBUZIONE CHI-QUADRATO CON ν GRADI DI LIBERTÀ

	α	α	α	α	α	α	α	α	α	α	α	α	α	α
	0.995	0.99	0.975	0.95	0.9	0.75	0.5	0.25	0.1	0.05	0.025	0.01	0.005	0.001
ν	χ^2_P	χ^2_P	χ^2_P	χ^2_P	χ^2_P	χ^2_P	χ^2_P	χ^2_P	χ^2_P	χ^2_P	χ^2_P	χ^2_P	χ^2_P	χ^2_P
1	0.0000	0.0002	0.0010	0.0039	0.0158	0.102	0.455	1.323	2.706	3.841	5.024	6.635	7.879	10.827
2	0.0100	0.0201	0.0506	0.1026	0.211	0.575	1.386	2.773	4.605	5.991	7.378	9.210	10.597	13.815
3	0.0717	0.1148	0.2158	0.352	0.584	1.213	2.366	4.108	6.251	7.815	9.348	11.345	12.838	16.266
4	0.207	0.297	0.484	0.711	1.064	1.923	3.357	5.385	7.779	9.488	11.143	13.277	14.860	18.466
5	0.412	0.554	0.831	1.145	1.610	2.675	4.351	6.626	9.236	11.070	12.832	15.086	16.750	20.515
6	0.676	0.872	1.237	1.635	2.204	3.455	5.348	7.841	10.645	12.592	14.449	16.812	18.548	22.457
7	0.989	1.239	1.690	2.167	2.833	4.255	6.346	9.037	12.017	14.067	16.013	18.475	20.278	24.321
8	1.344	1.647	2.180	2.733	3.490	5.071	7.344	10.219	13.362	15.507	17.535	20.090	21.955	26.124
9	1.735	2.088	2.700	3.325	4.168	5.899	8.343	11.389	14.684	16.919	19.023	21.666	23.589	27.877
10	2.156	2.558	3.247	3.940	4.865	6.737	9.342	12.549	15.987	18.307	20.483	23.209	25.188	29.588
11	2.603	3.053	3.816	4.575	5.578	7.584	10.341	13.701	17.275	19.675	21.920	24.725	26.757	31.264
12	3.074	3.571	4.404	5.226	6.304	8.438	11.340	14.845	18.549	21.026	23.337	26.217	28.300	32.909
13	3.565	4.107	5.009	5.892	7.041	9.299	12.340	15.984	19.812	22.362	24.736	27.688	29.819	34.527
14	4.075	4.660	5.629	6.571	7.790	10.165	13.339	17.117	21.064	23.685	26.119	29.141	31.319	36.124
15	4.601	5.229	6.262	7.261	8.547	11.037	14.339	18.245	22.307	24.996	27.488	30.578	32.801	37.698
16	5.142	5.812	6.908	7.962	9.312	11.912	15.338	19.369	23.542	26.296	28.845	32.000	34.267	39.252
17	5.697	6.408	7.564	8.672	10.085	12.792	16.338	20.489	24.769	27.587	30.191	33.409	35.718	40.791
18	6.265	7.015	8.231	9.390	10.865	13.675	17.338	21.605	25.989	28.869	31.526	34.805	37.156	42.312
19	6.844	7.633	8.907	10.117	11.651	14.562	18.338	22.718	27.204	30.144	32.852	36.191	38.582	43.819
20	7.434	8.260	9.591	10.851	12.443	15.452	19.337	23.828	28.412	31.410	34.170	37.566	39.997	45.314

$$\chi^2 = \Sigma \frac{(f_o - f_e)^2}{f_e}$$

Misure della forza della relazione fra variabili nominali e ordinali

- Misure di associazione fra variabili nominali
  - Misure di associazione basate sul chi-quadrato
  - Misure di associazione basate sulla riduzione proporzionale dell'errore
- Misure di cograduazione fra variabili ordinali

- **Rapporti di probabilità (odds)**: rapporto fra la frequenza di una categoria e la frequenza della categoria alternativa (nel caso di variabili dicotomiche)
- lo indichiamo con la lettera greca omega (ω).
- è anche definibile come il rapporto fra la probabilità che un individuo, estratto a caso dall'universo, appartenga a una categoria della variabile considerata e la probabilità che non vi appartenga (da cui il suo nome italiano di «rapporto di probabilità»)

Rapporto di probabilità (*odds*):

$$\omega = \frac{f_1}{f_2} = \frac{p_i}{1 - p_i}$$

Relazione tra una **dicotomia** e una variabile categoriale ordinale con più di due modalità

Per studiare la relazione tra una variabile categoriale con più di due modalità ed una variabile dicotomica (tabella K*2) è necessario calcolare la probabilità (*odds – rapporto di probabilità*) che ogni caso della popolazione/campione appartenente ad una delle modalità della variabile categoriale ha di appartenere anche ad una delle due modalità della variabile dicotomica.

Si ricorre agli *odds* perché:

- non può essere attribuito segno alla relazione;
- non si può ricorrere a nessuno dei coefficienti delle dicotomie.

Vantaggi

Gli *odds* consentono di confrontare le categorie di risposta della variabile categoriale in relazione alla variabile dicotomica senza ricorrere alle %.

	HA FIGLI?	SI	NO	TOT	SI
OCCUPAZIONE	Dirigenti e funz. Pubblici	28	12	40	28/12 *2,3*
	Dirigenti e funz. Privati	34	13	47	34/13 *2,6*
	Liberi professionisti	30	16	46	30/16 *1,8*
	Autonomi in comm.; industriali	47	11	51	47/11 *4,2*
	Magistrati, professori	15	7	22	15/7 *2,1*
	Artisti; pubblicitari; pr	13	16	29	13/16 *0,8*
	Impiegati privati	71	28	99	71/28 *2,5*
	Impiegati pubblici	87	22	109	87/22 *3,9*
	Operai	106	29	135	106/29 *3,6*
	Coldiretti; mezzadri	41	5	4	41/5 *8,2*
	TOTALE	472	159	631	472/159 *2,9*

←***ODDS MEDIO***

Esercitazione bivariata

- Calcolare le frequenze attese per questa tabella e evidenziare le differenze rispetto alle fr. effettive;
- Cercare nella tabella chi-quadrato la significatività rispetto ad un ipotetico valore di 1610,5
- Commentare rispetto ai diversi livelli di significatività

https://www.socscistatistics.com/tests/chisquare/

	Non consumo	Consumo	M.Riga
Maschi	1755	6892	8647
Femmine	4576	4787	9363
M.Colonna	6331	11.679	18.010

Esercitazione odds

- Con il file Gym, incrociare il genere (d1) con la motivazione (d3);
- Calcolare i rapporti di probabilità delle diverse categorie;
- Commentare.

- **Analisi della varianza**
- (detta anche Anova) serve per studiare la relazione fra *una variabile nominale* e una *cardinale*
- Anche in questo caso si può stabilire la significatività della relazione (col rapporto F) e misurarne la forza (con l'eta-quadrato)

- **Teorema fondamentale della varianza**

$$\sum_i \sum_j (Y_{ij} - \overline{Y}_{..})^2 = \sum_i \sum_j (Y_{ij} - \overline{Y}_{.j})^2 + \sum_i \sum_j (\overline{Y}_{.j} - \overline{Y}_{..})^2$$

Somma totale dei quadrati	=	Somma interna dei quadrati (devianza non spiegata)	+	Somma esterna dei quadrati (devianza spiegata)

- Devianza (o somma dei quadrati, SQ)
- Abbiamo così scomposto la devianza della variabile cardinale dipendente in due componenti:
  a) la somma dei quadrati degli scarti dei singoli valori dalla rispettiva media di gruppo; essa viene chiamata somma interna dei quadrati («interna» in quanto è interna al gruppo);
  b) la somma dei quadrati degli scarti delle medie di gruppo dalla media generale, che viene chiamata somma esterna dei quadrati.

- La prima somma è una misura della variabilità del fenomeno entro i gruppi
- La seconda una misura della variabilità del fenomeno studiato fra i gruppi.
- La somma interna dei quadrati viene anche chiamata devianza non spiegata
- La somma esterna viene chiamata devianza spiegata. Spiegata da che cosa?
- Spiegata dalla variabile nominale: è quella parte di variabilità della variabile dipendente che è attribuibile alla variabile indipendente.

$$SQ_{totale} = \underbrace{SQ_{interna}}_{=0 \text{ in caso di relazione perfetta}} + \underbrace{SQ_{esterna}}_{=0 \text{ in caso di assenza di relazione}}$$

Significatività della relazione

- Sottoporre a verifica l'ipotesi nulla secondo la quale le medie di gruppo Y provengono tutte da una stessa popolazione e quindi i dati nella popolazione (ipotetica o effettiva) dalla quale derivano sono uguali fra loro

Dividendo la devianza per i gradi di libertà si ottiene la stima della varianza della popolazione dalla quale derivano i dati del campione studiato. I gradi di libertà sono i seguenti:

$$\underset{\text{gradi di libertà totali}}{N-1} = \underset{\text{gradi di libertà interni}}{(N-k)} + \underset{\text{gradi di libertà esterni}}{(k-1)}$$

Le stime della varianza (dette anche «quadrati medi», *mean squares*) sono pertanto:

$$\text{Stima entro i gruppi o stima interna} = \frac{\sum_i \sum_j (Y_{ij} - \overline{Y}_{.j})^2}{N-k}$$

→Se l'ipotesi nulla è vera
le due stime sono uguali;
se l'ipotesi nulla è falsa
la seconda stima
è maggiore della prima

TAB. 3.25. Tabella riassuntiva dell'analisi della varianza

	SQ: somma dei quadrati	*GL*: gradi di libertà	Stima della varianza (quadrati medi)	*F*
Totale	2.901,84	$N-1=27$		
Esterna (fra i gruppi, spiegata)	1.979,41	$k-1=3$	659,80	
Interna (entro i gruppi, non spiegata)	922,93	$N-k=24$	38,46	
				17,16

Calcoli:
Stime della varianza: esterna = 1.979,41/3 = 659,80
interna = 922,23/24 = 38,46
Rapporto *F*: = 659,80/38,46 = 17,16

$$\text{Stima fra i gruppi o stima esterna} = \frac{\sum_i \sum_j (\overline{Y}_{.j} - \overline{Y}_{..})^2}{k-1}$$

- *Forza* della relazione
- Eta-quadrato o η^2: rapporto fra la somma dei quadrati esterna (spiegata) e la somma dei quadrati totale (devianza totale):

$$\eta^2 = \frac{SQ_{esterna}}{SQ_{totale}} = \frac{SQ_{spiegata}}{SQ_{totale}}$$

Regressione e correlazione

- Relazione fra due variabili cardinali
- Diagramma di dispersione

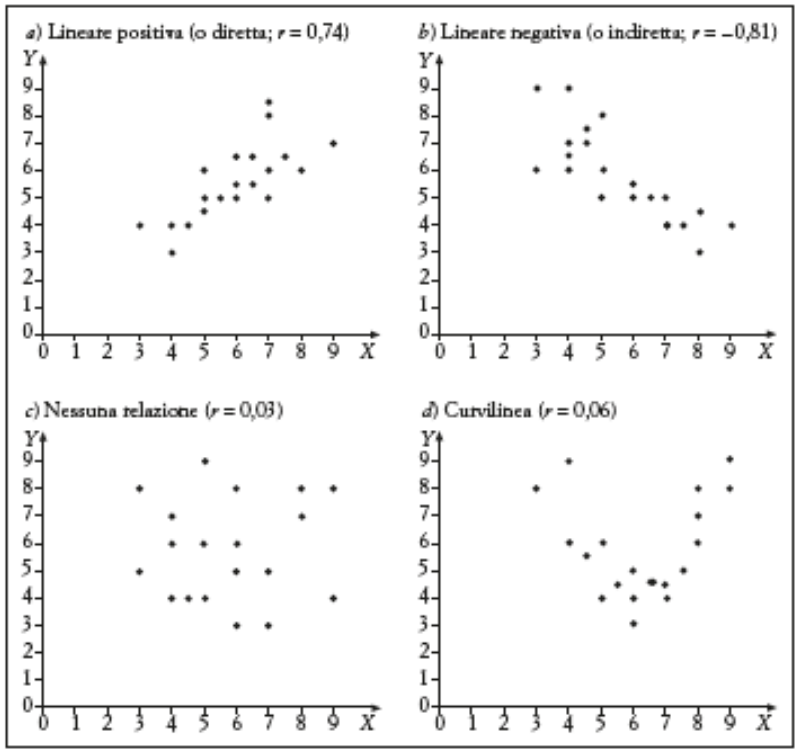


fig. 3.3. *Diagrammi di dispersione raffiguranti quattro tipi di relazioni fra due variabili.*

- Retta di regressione

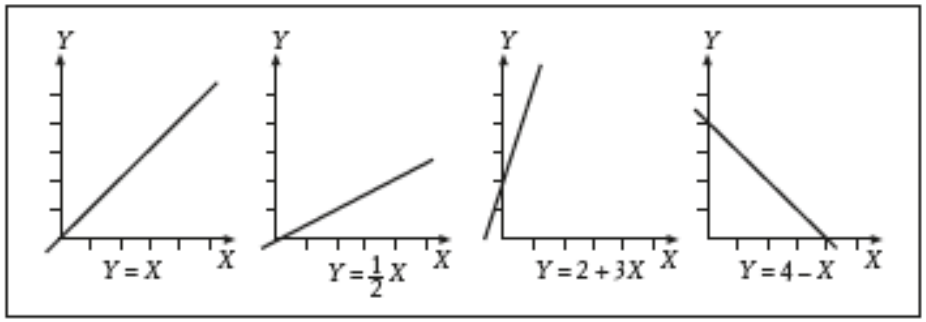


fig. 3.4. *Alcuni esempi di rette e loro equazioni.*

- $Y = a + bX$
- dove a è l'intercetta della retta sull'asse delle Y (cioè l'ordinata della retta quando l'ascissa è 0) e b è l'inclinazione della retta (cioè la variazione dell'ordinata quando l'ascissa varia di un'unità)

Coefficiente di correlazione

- Coefficiente di correlazione di Pearson

$$r = \frac{\Sigma(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\Sigma(X_i - \overline{X})^2 \, \Sigma(Y_i - \overline{Y})^2}}$$

- r assume valore +1 in caso di relazione perfetta positiva, –1 in caso di relazione perfetta negativa, e 0 in caso di assenza di relazione
- r è un numero puro, nel senso che non risente dell'unità di misura delle due variabili

- La relazione fra due variabili X e Y può risultare dai dati, ma tuttavia può non essere dovuta a un affetto di causazione (X causa Y), in quanto può esistere una terza variabile Z che influenza entrambe, producendo una correlazione fra X e Y che non deve essere interpretata in termini di causazione fra X e Y.
- Per il rischio di questo errore, anche nel caso di analisi bivariate **è sempre opportuno introdurre nell'analisi terze variabili**, allo scopo di purificare e chiarire la relazione fra le due variabili iniziali X e Y.
- A seconda del modo di interagire della terza variabile con le prime due, la relazione fra X e Y può risultare:
  - **spuria,**
  - **Indiretta**
  - **condizionata**