Statistica bivariata

Tabella di contingenza

% di riga

I tre tipi di percentuale rimandano a significati differenti

	Si più info	No più info	Totale
Sufficiente	46	567	613
Insufficiente	218	419	637
Totale	264	986	1250

	Si più info	No più info	Totale
Sufficiente	7,5%	92,5%	100,0%
Insufficiente	34,2%	65,8%	100,0%
Totale	21,1%	78,9%	100,0%

Statistica bivariata

Tabella di contingenza

% di colonna

I tre tipi di percentuale rimandano a significati differenti

	Si più info	No più info	Totale
Sufficiente	46	567	613
Insufficiente	218	419	637
Totale	264	986	1250

	Si più info	No più info	Totale
Sufficiente	17,4%	57,5%	49,0%
Insufficiente	82,6%	42,5%	51,0%
Totale	100,0%	100,0%	100,0%

Statistica bivariata

Tabella di contingenza

% di cella

I tre tipi di percentuale rimandano a significati differenti

	Si più info	No più info	Totale
Sufficiente	46	567	613
Insufficiente	218	419	637
Totale	264	986	1250

	Si più info	No più info	Totale
Sufficiente	3,7%	45,4%	49,0%
Insufficiente	17,4%	33,5%	51,0%
Totale	21,1%	78,9%	100,0%

Statistica bivariata

Tabella di contingenza

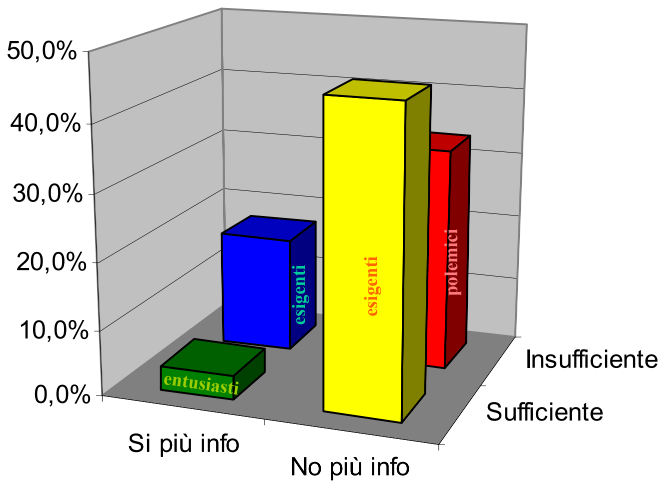
Una tipologia

		Più info	
		Sì	No
Valutazione Informazione	sufficiente	Entusiasti 3,7%	Soddisfatti 45,4%
	insufficiente	Esigenti 17,4%	Polemici 33,5%

Statistica bivariata

Tabella di contingenza

Una tipologia



Statistica bivariata

Operatori di connessione – Misure basate sul Chi quadro

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(n_{ij} - \hat{n}_{ij} \right)^2}{\hat{n}_{ij}}$$

Il valore del Chi quadro è influenzato dalla numerosità campionaria.

Due tabelle con le stesse percentuali di cella ma costruita con N differenti forniscono valori di Chi quadro diversi. In particolare il valore del Chi quadro cresce al crescere del valore di N

Statistica bivariata

Operatori di connessione – Misure basate sul Chi quadro

Una misura di associazione derivata dal Chi quadro ed indipendente dalla numerosità campionaria è Φ^2

$$\Phi^2 = \frac{X^2}{N} = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(f_{ij} - \hat{f}_{ij}\right)^2}{\hat{f}_{ij}}$$

Statistica bivariata

Operatori di connessione – Misure basate sul Chi quadro

Il valore massimo di Φ^2 dipende dal numero di modalità delle variabili. Per questo motivo si utilizza anche la statistica T di Tschuprov che, in ogni caso, ha come valore massimo 1

$$T = \frac{\Phi^2}{\sqrt{(J-1)(I-1)}}$$

Statistica bivariata

Operatori di concordanza – K di Cohen

Si calcola per variabili categoriali che presentano una specifica affinità da un punto di vista logico.

Es: lavoro del padre e del figlio

Diagnosi operate da due diversi psicologi

Il K di Cohen considera solamente le frequenze poste sulla diagonale principale, quelle cioè che riguardano categorie affini

Statistica bivariata

Operatori di concordanza – K di Cohen

Es: si confrontano fatte diagnosi di due psicologi che hanno analizzato le cartelle cliniche di 20 pazienti

	Psicosi	Borderline	Nevrosi	**Totale**
Psicosi	4	0	0	**4**
Borderline	1	3	1	**5**
Nevrosi	1	1	9	**11**
TOTALE	**6**	**4**	**10**	**20**

Statistica bivariata

Operatori di concordanza – K di Cohen

L'accordo complessivo atteso è dato dalla somma delle probabilità degli eventi congiunti definiti dalle celle poste sulla diagonale principale

$$\hat{\theta} = \sum_{i=1}^{I} f_{i.} f_{.i}$$

Frequenze relative congiunte teoriche

	Psicosi	Borderline	Nevrosi	**Totale**
Psicosi	0,06			**0,20**
Borderline		0,05		**0,25**
Nevrosi			0,275	**0,55**
TOTALE	**0,30**	**0,20**	**0,50**	**1**

$$\hat{\theta} = 0.06 + 0.05 + 0.275 = 0.385$$

Statistica bivariata

Operatori di concordanza – K di Cohen

L'accordo complessivo osservato è dato dalla somma delle frequenze relative delle celle poste sulla diagonale principale

$$\theta = \sum_{i=1}^{I} f_{ii}$$

Frequenze relative congiunte osservate

	Psicosi	Borderline	Nevrosi	**Totale**
Psicosi	0,2			**0,20**
Borderline		0,15		**0,25**
Nevrosi			0,45	**0,55**
TOTALE	**0,30**	**0,20**	**0,50**	**1**

$$\theta = 0.4 + 0.15 + 0.45 = 0.8$$

Statistica bivariata

Operatori di concordanza – K di Cohen

$$K = \frac{\theta - \hat{\theta}}{1 - \hat{\theta}}$$

$$K = \frac{0.8 - 0.385}{1 - 0.385} = \frac{0.415}{0.615} = 0.675$$

K è abbastanza elevato, possiamo quindi concludere che c'è una relazione tra le diagnosi fatte dai due psicologi

Statistica bivariata

Operatori di concordanza – K di Cohen

Kappa	Livello di accordo
< 0,00	Povero
0,00 – 0,20	Esile
0,21 – 0,40	Discreto
0,41 – 0,60	Moderato
0,61 – 0,80	Sostanziale
0,81 – 1,00	perfetto

Landis, Koch 1977

Statistica bivariata

Operatori di concordanza – K di Cohen

Per capire se il valore osservato è significativo bisogna impostare un test:

$$H_0 = K \leq 0$$
$$H_1 = K > 0$$

$$\alpha = .01 \Rightarrow z_c = 2{,}33$$

$$Z = \frac{K}{\sigma_{k^0}} \Rightarrow \sigma_{k^0} = \sqrt{\frac{\hat{\theta}}{N(1-\hat{\theta})}}$$

Statistica bivariata

Operatori di concordanza – K di Cohen

Nell'esempio:

$$Z = \frac{0{,}675}{\sqrt{\frac{0{,}385}{20(1-0{,}385)}}} \qquad Z = \frac{0{,}675}{\sqrt{\frac{0{,}385}{12{,}3}}} \qquad Z = \frac{0{,}675}{0{,}177} = 3{,}25$$

$$Z_c = 2{,}33$$
$$Z = 3{,}25$$

Il valore di k è diverso maggiore di zero con con un livello di significatività alfa=0.01

Statistica bivariata

Operatori di concordanza – Q di Yule

Valuta la concordanza nel caso di due dicotomie

ES: X= soggetti che si sono licenziati nell'ultimo anno

Y = soggetti che hanno seguito almeno un corso di formazione negli ultimi 5 anni

		Formazione		totale
		si	no	
Licenziati	si	***10***	***20***	30
	no	***30***	***20***	50
totale		40	40	80

Statistica bivariata

Operatori di concordanza – Q di Yule

	Valore	Interpretazione
Minimo	-1	Associazione completa negativa
Centrale	0	Indipendenza
Massimo	+1	Associazione completa positiva

Statistica bivariata

Operatori di concordanza – Q di Yule

$$Q = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21}}$$

		Formazione		totale
		si	no	
Licenziati	si	***10***	***20***	30
	no	***30***	***20***	50
totale		40	40	80

$$Q = \frac{10*20 - 20*30}{10*20 + 20*30} = \frac{200-600}{200+600} = \frac{-400}{800} = -0.5$$

Poiché il valore di Q è abbastanza alto possiamo affermare che esiste una relazione tra il licenziamento e la formazione

Statistica bivariata

Coefficiente di correlazione tetracorico

$$r_{phi} = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{1.}n_{.1}n_{2.}n_{.2}}}$$

	Valore	Interpretazione
Minimo	-1	Associazione completa negativa
Centrale	0	Indipendenza
Massimo	+1	Associazione completa positiva

Più sensibile allo squilibrio dei marginali di riga e di colonna

Statistica bivariata

Coefficiente di correlazione di Spearman

Detto anche coefficiente di correlazione per ranghi, misura la concordanza tra due variabili espresse a livello di scala ordinale

$$r_s = 1 - 6\sum_{i=1}^{N} \frac{\left(x_i - y_i\right)^2}{N\left(N^2 - 1\right)}$$

Statistica bivariata

Coefficiente di correlazione di Spearman

Esempio: uno psicologo ha sottoposto 5 soggetti a due test. Si chiede se esiste una relazione tra le prestazioni nei due test

id	Graduatoria TEST A	Graduatoria TEST B
1	1°	3°
2	2°	4°
3	3°	1°
4	4°	5°
5	5°	2°

Statistica bivariata

Coefficiente di correlazione di Spearman

Esempio: per ciascun soggetto possiamo calcolare la differenza tra il rango del test A e quello del test B ed elevare queste differenze al quadrato

id	Graduatoria TEST A	Graduatoria TEST B	D	D^2
1	**1**	**3**	-2	4
2	**2**	**4**	-2	4
3	**3**	**1**	2	4
4	**4**	**5**	-1	1
5	**5**	**2**	3	9
			somma=	**22**

Statistica bivariata

Coefficiente di correlazione di Spearman

Esempio: riprendendo la formula $r_s = 1 - 6\sum_{i=1}^{N} \frac{(x_i - y_i)^2}{N(N^2 - 1)}$

$$r_s = 1 - 6\frac{22}{5(25-1)} = 1 - 6\frac{22}{120} = 1 - 6(0{,}183) = -0{,}1$$

	Valore	Interpretazione
Minimo	-1	Associazione completa negativa
Centrale	0	Indipendenza
Massimo	+1	Associazione completa positiva

Statistica bivariata

La covarianza

La covarianza è la media dei prodotti degli scarti dalla media di due variabili

$$s_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

Momento centrale misto di ordine 2

- Sx Sy | 0 | + Sx Sy

Max discordanza

Assenza di relazione

Max concordanza

Statistica bivariata

La covarianza

$$s_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

Esempio

ident	altezza	peso	Scarti altezza	Scarti peso	prodotto	
1	170	68	-3	-5	15	
2	164	73	-9	0	0	
3	178	70	5	-3	-15	
4	168	53	-5	-20	100	
5	174	86	1	13	13	
6	184	88	11	15	165	
				somma=	278	n=6
media	173	73		cov=	46,33333	

$S_{alt}=7.24$ $S_{pes}=12.87$

Statistica bivariata

La correlazione

La correlazione r di Bravais e Pearson è l'indice di concordanza maggiormente utilizzato per variabili di tipo cardinale

$$r_{xy} = \frac{s_{xy}}{|s_x s_y|}$$

Esempio

$$r_{altezza-peso} = \frac{46,33}{7,24 * 12,87} = \frac{46,33}{93,18} = 0,497$$

Statistica bivariata

La correlazione

La correlazione r di Bravais e Pearson è la covarianza tra le due variabili standardizzate

$$r_{xy} = s_{z_x z_y}$$

	Valore	Interpretazione
Minimo	-1	Associazione completa negativa
Centrale	0	Indipendenza
Massimo	+1	Associazione completa positiva

Statistica bivariata

La correlazione

La significatività r di Bravais e Pearson è data da

$$f = \frac{r^2}{1-r^2}(N-2)$$

Distribuzione F con 1 e n-2 gdl

$$t = \frac{r}{1-r^2}\sqrt{(N-2)}$$

Distribuzione t con n-2 gdl

Statistica bivariata

La correlazione

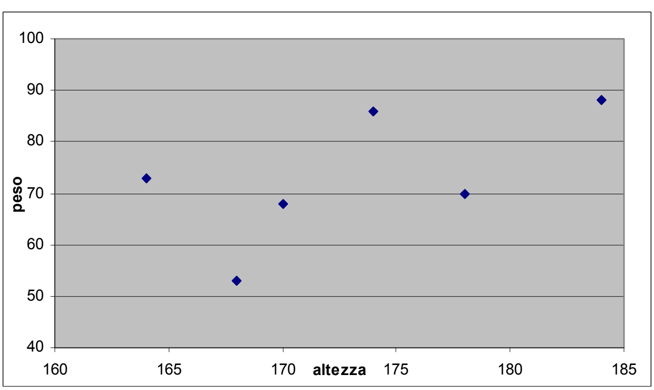


Grafico a dispersione
Scatterplot

La relazione tra due variabili può essere rappresentata per mezzo di un grafico a dispersione, assegnando ad ogni oggetto coordinate definite dai valori assunti sulle due variabili