LA DIPENDENZA STATISTICA FRA VARIABILI QUALITATIVE

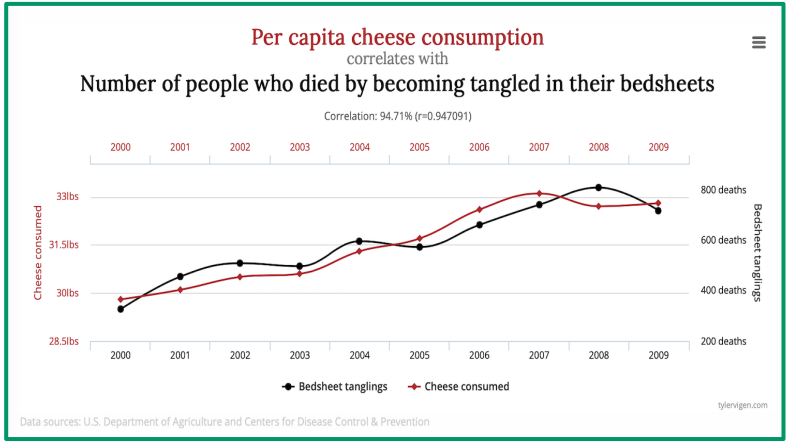
L’analisi della contingenza

PREMESSA

DIPENDENZA STATISTICA

Esistenza di una relazione fra due variabili osservate X e Y: ad esempio al crescere di X, Y aumenta o viceversa. La teoria economica individua una dipendenza fra livello di salari e occupazione, fra prezzo e domanda, ...

Quando fra due variabili esiste una dipendenza statistica, ma non una relazione logica di causa - effetto, si parla di **correlazione spurie**.



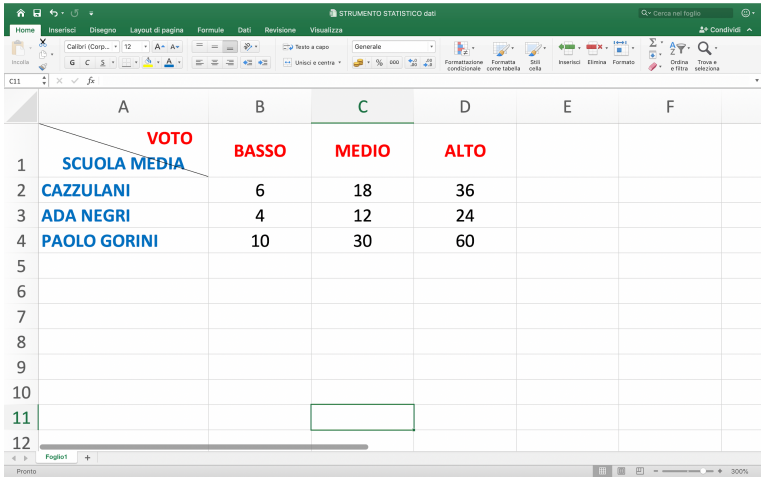
PREMESSA

VARIABILI QUALITATIVE

Variabili che non possono essere espresse in termini numerici (quantitativi); sono dette anche **mutabili**.

Ne sono esempi: colore degli occhi, titolo di studio, squadra per la quale si fa il tifo, mezzo di trasporto utilizzato per raggiungere la scuola, ...

TABELLE A DOPPIA ENTRATA



VOTO / SCUOLA MEDIA	BASSO	MEDIO	ALTO
CAZZULANI	6	18	36
ADA NEGRI	4	12	24
PAOLO GORINI	10	30	60

RISPONDI

Quanti sono in totale gli studenti?

Quanti studenti provengono dalla scuola Paolo Gorini?

Quanti studenti hanno ottenuto un voto di diploma medio?

Quanti studenti che hanno frequentato la Cazzulani hanno ottenuto un voto di diploma alto?

Quanti studenti che hanno ottenuto un voto basso hanno frequentato l'Ada Negri?

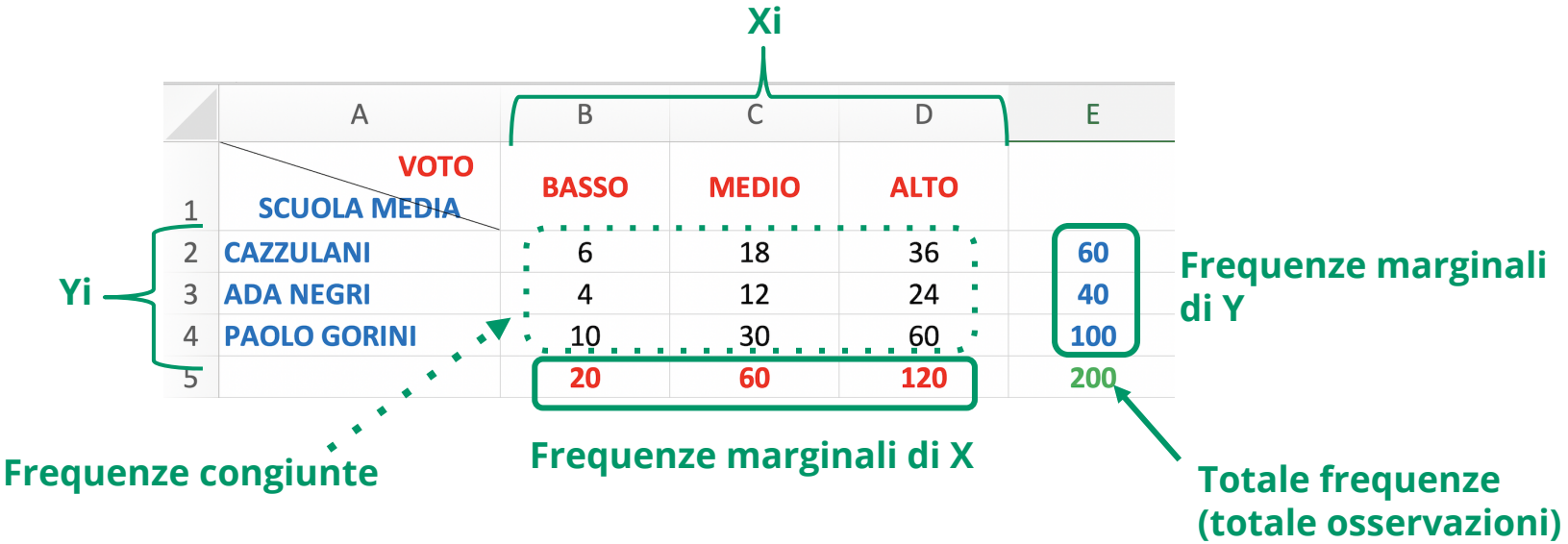
TABELLE A DOPPIA ENTRATA

Aggiungiamo alla precedente tabella a doppia entrata i totali di riga e di colonna.

	A	B	C	D	E
1	VOTO / SCUOLA MEDIA	BASSO	MEDIO	ALTO	
2	CAZZULANI	6	18	36	60
3	ADA NEGRI	4	12	24	40
4	PAOLO GORINI	10	30	60	100
5		20	60	120	200

TABELLE A DOPPIA ENTRATA

Introduciamo la terminologia statistica



	A	B	C	D	E
1	VOTO / SCUOLA MEDIA	BASSO	MEDIO	ALTO	
2	CAZZULANI	6	18	36	60
3	ADA NEGRI	4	12	24	40
4	PAOLO GORINI	10	30	60	100
5		20	60	120	200

TABELLE A DOPPIA ENTRATA

Formalizzando

	x_1	x_2	x_3	.....	.....	Totale
y_1	$f_{1,1}$	$f_{1,2}$	$f_{1,3}$	.....	.....	$f_{1,0}$
y_2	$f_{2,1}$	$f_{2,2}$	$f_{2,3}$	.....	.....	$f_{2,0}$
y_3	$f_{3,1}$	$f_{3,2}$	$f_{3,3}$	.....	.....	$f_{3,0}$
yi	.....	.....	.....	.....	.....	$f_{i,0}$
Totale	$f_{0,1}$	$f_{0,2}$	$f_{0,3}$	.....	$f_{0,j}$	f

TABELLE A DOPPIA ENTRATA

Le tabelle a doppia entrata vengono spesso rappresentate graficamente attraverso il **diagramma a bolle**, dove il **raggio** di ciascuna bolla è proporzionale alla frequenza congiunta.

INDIPENDENZA STATISTICA

Si dimostra che **se** le variabili X e Y sono indipendenti **allora** la frequenza congiunta ($f_{i,j}$) è uguale al prodotto delle corrispondenti frequenze marginali divise per il numero di osservazioni n.

$$f_{i,j} = (f_{i,0} * f_{0,j}) / n$$

Se tale condizione è rispettata le variabili sono **indipendenti**

Se tale condizione non è rispettata le variabili non sono indipendenti ovvero sono **dipendenti**

INDIPENDENZA STATISTICA

	BASSO	MEDIO	ALTO
CAZZULANI	(60*20)/200 = 6	(60*60)/200 = 18	(60*120)/200 = 36
NEGRI	(40*20)/200 = 4	(40*60)/200 = 12	(40*120)/200 = 24
GORINI	(100*20)/200 = 10	(100*60)/200 = 30	(100*120)/200 = 60

	A	B	C	D	E
1	VOTO / SCUOLA MEDIA	BASSO	MEDIO	ALTO	
2	CAZZULANI	6	18	36	60
3	ADA NEGRI	4	12	24	40
4	PAOLO GORINI	10	30	60	100
5		20	60	120	200

Le frequenze congiunte sono sempre uguali al prodotto delle frequenze marginali divise per n.
Vale cioè:

$$f_{i,j} = (f_{i,0} * f_{0,j}) / N$$

Le variabili sono fra loro indipendenti ovvero il voto di diploma non dipende dalla scuola media di provenienza.

CONTINGENZA

La **contingenza** permette di calcolare il grado di dipendenza statistica fra due mutabili.

ESEMPIO

Si consideri la tabella a doppia entrata assegnata e si dica se la scelta post diploma (Università o Lavoro) è dipendente dalla sezione (Sez A, Sez B, Sez C) frequentata dallo studente.

	A	B	C	D
1		UNIVERSITA'	LAVORO	
2	SEZ A	15	10	25
3	SEZ B	7	8	15
4	SEZ C	8	2	10
5		30	20	50

CONTINGENZA

Si costruisce la tabella delle frequenze marginali teoriche calcolate la formula nota:

$$f_{i,j} = (f_{i,0} * f_{0,j}) / N$$

	UNIVERSITA’	LAVORO
SEZ A	30*25/50 = 15	20*25/50 = 10
SEZ B	30*15/50 = 9	20*15/50 = 6
SEZ C	30*10/60 = 6	20*10/50 = 4

TABELLA FREQUENZE CONGIUNTE TEORICHE

	A	B	C	D
1		UNIVERSITA'	LAVORO	
2	SEZ A	15	10	25
3	SEZ B	7	8	15
4	SEZ C	8	2	10
5		30	20	50

TABELLA FREQUENZE CONGIUNTE OSSERVATE

CONTINGENZA

	UNIVERSITA’	LAVORO
SEZ A	30*25/50 = 15	20*25/50 = 10
SEZ B	30*15/50 = 9	20*15/50 = 6
SEZ C	30*10/60 = 6	20*10/50 = 4

TABELLA FREQUENZE CONGIUNTE TEORICHE

≠

	A	B	C	D
1		UNIVERSITA'	LAVORO	
2	SEZ A	15	10	25
3	SEZ B	7	8	15
4	SEZ C	8	2	10
5		30	20	50

TABELLA FREQUENZE CONGIUNTE OSSERVATE

La tabella delle frequenze congiunte è diversa da quella delle frequenze osservate

↓

Le due variabili non sono statisticamente indipendenti, ovvero la scelta post diploma (Università, Lavoro) non è indipendente dalla sezione frequentata (Sez A, Sez B, Sez C)

CONTINGENZA

Definisco contingenza la differenza fra una frequenza congiunta osservata e la frequenza congiunta teorica.

$$C(x_i;y_j) = f(x_i,y_j) \text{ osservata} - f(x_i,y_j) \text{ teorica}$$

Nel caso di **indipendenza statistica le contingenze sono tutte nulle**, mentre cresceranno in valore assoluto, al crescere del grado di dipendenza tra i caratteri.

CONTINGENZA

L'indice che misura il grado di connessione fra due mutabili è l'indice di Pearson noto anche come Chi^2:

$$\mathbf{Chi^2 = \sum C^2(x_i,y_j) \,/\, f(x_i,y_j)\ teoriche}$$

$Chi^2 = 0$ allora le variabili sono indipendenti

Maggiore è il valore di Chi^2, maggiore è il grado di connessione fra le due variabili

CONTINGENZA

	UNIVERSITA’	LAVORO
SEZ A	30*25/50 = 15	20*25/50 = 10
SEZ B	30*15/50 = 9	20*15/50 = 6
SEZ C	30*10/60 = 6	20*10/50 = 4

	A	B	C	D
1		UNIVERSITA'	LAVORO	
2	SEZ A	15	10	25
3	SEZ B	7	8	15
4	SEZ C	8	2	10
5		30	20	50

$Chi^2 = 0^2/15 + 0^2/10 + 2^2/9 + 2^2/6 + 2^2/6 + 2^2/4 = 2,8$

Essendo diverso Chi^2 da 0, le due mutabili non sono indipendenti.

CONTINGENZA

Spesso si calcola il coefficiente di contingenza:

$$\textbf{Coefficiente contingenza} = \sqrt{Chi^2/(Chi^2+n)}$$

Il coefficiente di contingenza assume valori compresi fra 0 e 1.
0 = indipendenza statistica
1 = Perfetta dipendenza statistica

Nell'esempio analizzato:
$\sqrt{2,8/(2,8+50)} = 0,230$
Essendo diverso da 0 le variabili non sono statisticamente indipendenti
Essendo più vicino a 0 che a 1, la dipendenza è piuttosto debole.

PROVA TU

Dire il sesso dello studente e la scelta della facoltà università sono indipendenti fra loro.

	A	B	C	D	E
1		ECONOMIA	INGEGNERIA	LETTERE	
2	MASCHI	14	19	7	40
3	FEMMINE	21	9	15	45
4		35	28	22	85