Tavole di contingenza e V di Cramér

Misurare l'associazione: il V di Cramér

Il **chi-quadro** misura se due caratteri sono associati, ma il suo valore dipende dalla dimensione della tavola e dalla numerosità campionaria, quindi non è direttamente interpretabile. Per ovviare a questo problema si usa il **V di Cramér**, che normalizza il chi-quadro su una scala da 0 a 1:

$$V = \sqrt{\frac{\chi^2}{n \times \min(r-1,\ c-1)}}$$

dove r è il numero di righe, c il numero di colonne e n la numerosità campionaria.

Interpretazione:

V	Intensità dell'associazione
0,00 – 0,10	Trascurabile
0,10 – 0,30	Debole
0,30 – 0,50	Moderata
0,50 – 1,00	Forte

Vantaggi rispetto al chi-quadro grezzo:

- Va sempre tra 0 e 1, indipendentemente dalla dimensione della tavola
- Non richiede tavole critiche
- Misura *quanto* sono associati due caratteri, non solo *se* lo sono

Esercizio 1

Dati

In un'indagine su 120 studenti universitari si rilevano due caratteri:

- **X** = Facoltà (Economia, Ingegneria)
- **Y** = Preferenza di lavoro (Pubblico, Privato, Startup)

	Pubblico	**Privato**	**Startup**	**Totale**
Economia	30	20	10	60
Ingegneria	10	35	15	60
Totale	40	55	25	120

Tabella 1: Frequenze osservate – Esercizio 1

Punto (a) – Frequenze attese

$$E_{ij} = \frac{\text{totale riga } i \times \text{totale colonna } j}{n}$$

	Pubblico	**Privato**	**Startup**	**Totale**
Economia	$\frac{60\times40}{120} = 20$	$\frac{60\times55}{120} = 27{,}5$	$\frac{60\times25}{120} = 12{,}5$	60
Ingegneria	$\frac{60\times40}{120} = 20$	$\frac{60\times55}{120} = 27{,}5$	$\frac{60\times25}{120} = 12{,}5$	60
Totale	40	55	25	120

Tabella 2: Frequenze attese – Esercizio 1

	Pubblico	**Privato**	**Startup**
Economia	$30/60 = 0{,}500$	$20/60 = 0{,}333$	$10/60 = 0{,}167$
Ingegneria	$10/60 = 0{,}167$	$35/60 = 0{,}583$	$15/60 = 0{,}250$

Tabella 3: Distribuzioni condizionate di Y dato X – Esercizio 1

Punto (b) – Distribuzioni condizionate di Y dato X

Le righe sono **diverse**: gli studenti di Economia preferiscono il lavoro pubblico, quelli di Ingegneria il privato. Questo suggerisce che X e Y non sono indipendenti.

Punto (c) – Chi-quadro e V di Cramér

Cella	**O**	**T**	**O - T**	$(O-T)^2$	$\frac{(O-T)^2}{T}$
Economia – Pubblico	30	20,00	+10,00	100,00	5,000
Economia – Privato	20	27,50	−7,50	56,25	2,045
Economia – Startup	10	12,50	−2,50	6,25	0,500
Ingegneria – Pubblico	10	20,00	−10,00	100,00	5,000
Ingegneria – Privato	35	27,50	+7,50	56,25	2,045
Ingegneria – Startup	15	12,50	+2,50	6,25	0,500
				Totale χ^2	**15,09**

Tabella 4: Schema di calcolo del χ^2 – Esercizio 1

Si calcola ora il V di Cramér con $n = 120$, $r = 2$, $c = 3$, quindi $\min(r-1, c-1) = \min(1, 2) = 1$:

$$V = \sqrt{\frac{15{,}09}{120 \times 1}} = \sqrt{0{,}1258} \approx \mathbf{0{,}35}$$

Interpretazione: $V \approx 0{,}35$ indica un'associazione **moderata** tra facoltà e preferenza lavorativa. Gli studenti dei due corsi tendono a preferenze lavorative diverse, ma con un'intensità non estrema.

Esercizio 2

Dati

In un'indagine su 200 persone si rilevano due caratteri:

- **X** = Titolo di studio (Licenza media, Diploma, Laurea)
- **Y** = Utilizzo quotidiano dei social media (Basso, Medio, Alto)

	Basso	**Medio**	**Alto**	**Totale**
Licenza media	5	20	35	60
Diploma	15	40	25	80
Laurea	30	20	10	60
Totale	50	80	70	200

Tabella 5: Frequenze osservate – Esercizio 2

Punto (a) – Frequenze attese

	Basso	**Medio**	**Alto**	**Totale**
Licenza media	$\frac{60\times50}{200} = 15$	$\frac{60\times80}{200} = 24$	$\frac{60\times70}{200} = 21$	60
Diploma	$\frac{80\times50}{200} = 20$	$\frac{80\times80}{200} = 32$	$\frac{80\times70}{200} = 28$	80
Laurea	$\frac{60\times50}{200} = 15$	$\frac{60\times80}{200} = 24$	$\frac{60\times70}{200} = 21$	60
Totale	50	80	70	200

Tabella 6: Frequenze attese – Esercizio 2

Punto (b) – Distribuzioni condizionate di Y dato X

	Basso	**Medio**	**Alto**
Licenza media	$5/60 = 0{,}083$	$20/60 = 0{,}333$	$35/60 = 0{,}583$
Diploma	$15/80 = 0{,}188$	$40/80 = 0{,}500$	$25/80 = 0{,}313$
Laurea	$30/60 = 0{,}500$	$20/60 = 0{,}333$	$10/60 = 0{,}167$

Tabella 7: Distribuzioni condizionate di Y dato X – Esercizio 2

Si osserva una chiara tendenza inversa: chi ha la licenza media usa i social prevalentemente in modo alto; chi ha la laurea prevalentemente in modo basso.

Punto (c) – Chi-quadro e V di Cramér

Si calcola ora il V di Cramér con $n = 200$, $r = 3$, $c = 3$, quindi $\min(r-1, c-1) = \min(2,2) = 2$:

$$V = \sqrt{\frac{41{,}67}{200 \times 2}} = \sqrt{\frac{41{,}67}{400}} = \sqrt{0{,}1042} \approx \mathbf{0{,}32}$$

Interpretazione: $V \approx 0{,}32$ indica un'associazione **moderata** tra titolo di studio e utilizzo dei social media. La tendenza è chiara e sistematica, ma non così intensa da essere definita forte.

Esercizio 3

Dati

In uno studio del 2002 è stata analizzata la relazione tra l'attitudine verso il pensionamento anticipato e la soddisfazione sull'ambiente di lavoro (misurata su scala da 0 a 10, dove 0 = per nulla soddisfatto e 10 = pienamente soddisfatto).

Cella	O	T	O - T	$(O-T)^2$	$\frac{(O-T)^2}{T}$
Lic. media – Basso	5	15,00	−10,00	100,00	6,667
Lic. media – Medio	20	24,00	−4,00	16,00	0,667
Lic. media – Alto	35	21,00	+14,00	196,00	9,333
Diploma – Basso	15	20,00	−5,00	25,00	1,250
Diploma – Medio	40	32,00	+8,00	64,00	2,000
Diploma – Alto	25	28,00	−3,00	9,00	0,321
Laurea – Basso	30	15,00	+15,00	225,00	15,000
Laurea – Medio	20	24,00	−4,00	16,00	0,667
Laurea – Alto	10	21,00	−11,00	121,00	5,762
				Totale χ^2	**41,67**

Tabella 8: Schema di calcolo del χ^2 – Esercizio 2

Attitudine pens. anticipato	0–3	3–7	7–10	Totale
A favore	267	152	52	471
Neutro	68	30	11	109
A sfavore	26	24	10	60
Totale	361	206	73	640

Tabella 9: Tavola di contingenza completa – Esercizio 3

Verificare se esiste una relazione (associazione) tra attitudine al pensionamento anticipato e soddisfazione sull'ambiente di lavoro.

Punto (a) – Frequenze attese

$$T_{ij} = \frac{n_{i\cdot} \times n_{\cdot j}}{N}$$

Ad esempio: $T_{11} = \dfrac{471 \times 361}{640} = 265{,}67$

Attitudine pens. anticipato	0–3	3–7	7–10
A favore	265,67	151,60	53,72
Neutro	61,48	35,08	12,43
A sfavore	33,84	19,31	6,84

Tabella 10: Frequenze attese – Esercizio 3

Punto (b) – Distribuzioni condizionate di Y dato X

Le righe sono simili tra loro, il che lascia già intuire una debole associazione tra le due variabili.

	0–3	**3–7**	**7–10**
A favore	$267/471 = 0{,}567$	$152/471 = 0{,}323$	$52/471 = 0{,}110$
Neutro	$68/109 = 0{,}624$	$30/109 = 0{,}275$	$11/109 = 0{,}101$
A sfavore	$26/60 = 0{,}433$	$24/60 = 0{,}400$	$10/60 = 0{,}167$

Tabella 11: Distribuzioni condizionate di Y dato X – Esercizio 3

Cella	**O**	**T**	**O - T**	$(O-T)^2$	$\frac{(O-T)^2}{T}$
A favore – 0–3	267	265,67	+1,33	1,77	0,007
A favore – 3–7	152	151,60	+0,40	0,16	0,001
A favore – 7–10	52	53,72	−1,72	2,96	0,055
Neutro – 0–3	68	61,48	+6,52	42,51	0,691
Neutro – 3–7	30	35,08	−5,08	25,81	0,735
Neutro – 7–10	11	12,43	−1,43	2,04	0,164
A sfavore – 0–3	26	33,84	−7,84	61,47	1,816
A sfavore – 3–7	24	19,31	+4,69	22,00	1,140
A sfavore – 7–10	10	6,84	+3,16	9,99	1,461
				Totale χ^2	**6,07**

Tabella 12: Schema di calcolo del χ^2 – Esercizio 3

Punto (c) – Chi-quadro e V di Cramér

Si calcola ora il V di Cramér con $n = 640$, $r = 3$, $c = 3$, quindi $\min(r-1, c-1) = \min(2,2) = 2$:

$$V = \sqrt{\frac{6{,}07}{640 \times 2}} = \sqrt{\frac{6{,}07}{1280}} \approx \sqrt{0{,}00474} \approx \mathbf{0{,}069}$$

Interpretazione: $V \approx 0{,}07$ è molto vicino a 0, ben al di sotto della soglia di 0,10. **Non possiamo concludere che vi sia associazione** tra attitudine verso il pensionamento anticipato e soddisfazione sull'ambiente di lavoro.

Testo

In un'indagine su 10 studenti universitari sono state rilevate le ore settimanali di studio (X) e il voto ottenuto all'esame (Y):

Studente	1	2	3	4	5	6	7	8	9	10
Ore di studio X	4	7	5	12	9	3	15	10	6	11
Voto Y	18	22	19	29	25	17	30	27	20	28

a. Costruire lo scatterplot, riportando sull'asse delle ascisse le ore di studio e sull'asse delle ordinate il voto. Cosa suggerisce visivamente il grafico?

b. Calcolare la media e la deviazione standard di X e di Y.

c. Calcolare la covarianza tra X e Y:

$$\sigma_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

d. Calcolare il coefficiente di correlazione di Pearson:

$$r = \frac{\sigma_{XY}}{\sigma_X \, \sigma_Y}$$

e. Interpretare il valore di r ottenuto. La relazione è forte? È positiva o negativa? Ha senso dal punto di vista sostanziale?

Nota: tenere presente che un valore $|r| < 1$ non implica necessariamente l'assenza di un legame perfetto tra le variabili, ma l'assenza di un legame *lineare* perfetto. Analogamente, $r = 0$ non implica assenza di relazione tra le variabili, ma assenza di relazione lineare (più in generale, monotona).

Soluzione

b) Media e deviazione standard

$$\bar{x} = \frac{4+7+5+12+9+3+15+10+6+11}{10} = \frac{82}{10} = \mathbf{8{,}2}$$

$$\bar{y} = \frac{18+22+19+29+25+17+30+27+20+28}{10} = \frac{235}{10} = \mathbf{23{,}5}$$

Tavola degli scarti per il calcolo delle varianze:

i	x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
1	4	18	$-4{,}2$	$-5{,}5$	17,64	30,25
2	7	22	$-1{,}2$	$-1{,}5$	1,44	2,25
3	5	19	$-3{,}2$	$-4{,}5$	10,24	20,25
4	12	29	3,8	5,5	14,44	30,25
5	9	25	0,8	1,5	0,64	2,25
6	3	17	$-5{,}2$	$-6{,}5$	27,04	42,25
7	15	30	6,8	6,5	46,24	42,25
8	10	27	1,8	3,5	3,24	12,25
9	6	20	$-2{,}2$	$-3{,}5$	4,84	12,25
10	11	28	2,8	4,5	7,84	20,25
$\sum$					**133,60**	**214,50**

$$s_X^2 = \frac{133{,}60}{9} \approx 14{,}844 \quad \Rightarrow \quad s_X = \sqrt{14{,}844} \approx \mathbf{3{,}853}$$

$$s_Y^2 = \frac{214{,}50}{9} \approx 23{,}833 \quad \Rightarrow \quad s_Y = \sqrt{23{,}833} \approx \mathbf{4{,}882}$$

c) Covarianza

Aggiungiamo la colonna dei prodotti degli scarti:

i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
1	$-4{,}2$	$-5{,}5$	23,10
2	$-1{,}2$	$-1{,}5$	1,80
3	$-3{,}2$	$-4{,}5$	14,40
4	3,8	5,5	20,90
5	0,8	1,5	1,20
6	$-5{,}2$	$-6{,}5$	33,80
7	6,8	6,5	44,20
8	1,8	3,5	6,30
9	$-2{,}2$	$-3{,}5$	7,70
10	2,8	4,5	12,60
$\sum$			**166,00**

$$\sigma_{XY} = \frac{166{,}00}{9} \approx \mathbf{18{,}444}$$

d) Coefficiente di correlazione di Pearson

$$r = \frac{\sigma_{XY}}{\sigma_X \, \sigma_Y} = \frac{18{,}444}{3{,}853 \times 4{,}882} = \frac{18{,}444}{18{,}811} \approx \mathbf{0{,}980}$$

$r \approx 0{,}98$

e) Interpretazione

Il coefficiente di correlazione $r \approx 0{,}98$ indica una **relazione lineare positiva molto forte** tra le ore di studio e il voto all'esame: all'aumentare delle ore di studio corrisponde sistematicamente un aumento del voto.

Il valore è molto prossimo a 1 (correlazione lineare perfetta positiva), il che suggerisce che quasi tutta la variabilità dei voti è spiegata linearmente dalle ore di studio.

Attenzione alle limitazioni di r:

- Un valore $|r| < 1$ **non implica** l'assenza di un legame perfetto tra X e Y, ma soltanto l'assenza di un legame *lineare* perfetto. Potrebbe esistere una relazione deterministica di tipo non lineare (ad esempio quadratica o esponenziale) che r non è in grado di rilevare.
- Un valore $r = 0$ **non implica** l'assenza di qualsiasi relazione tra le variabili, ma soltanto l'assenza di una relazione *lineare* (più in generale, monotona). Le due variabili potrebbero essere legate da una relazione non monotona — ad esempio una parabola con vertice al centro del campo di variazione di X — che produce covarianza nulla pur in presenza di un legame funzionale perfetto.

In sintesi, r misura esclusivamente la **forza e la direzione della relazione lineare** tra due variabili. Lo scatterplot rimane quindi indispensabile per verificare visivamente la forma della relazione prima di affidarsi al solo valore di r.