Concordanza tra due graduatorie

Dipendenza in media

Covarianza e correlazione

1

Misure di concordanza tra due graduatorie

- Si definiscono graduatorie per caratteri <u>ordinabili</u> (variabili e mutabili ordinabili).
- Due graduatorie sono concordanti se basse (alte) modalità di un carattere sono associate a basse (alte) modalità dell’altro carattere.
- In caso contrario sono discordanti.

2

Si considerano 4 studenti e i voti agli esami di Matematica (X) e Statistica (Y)

Unità	*X*	*Y*	*Rango X*	*Rango Y*
A	*24*	24	4	3
B	*26*	22	3	4
C	*28*	26	2	2
D	30	28	1	1

3

Indice di cograduazione di Spearman

- Definita d_i la differenza tra i ranghi per la *i*-esima unità, l'indice di Spearman è dato da

$$\rho_S = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

- $-1 \leq \rho_S \leq 1$

4

Interpretazione dell'indice di Spearman

- $\rho_S = -1$ se i ranghi sono in perfetta discordanza (l'unità in prima posizione in una graduatoria è ultima nell'altra graduatoria, e così via)
- $-1 < \rho_S < 0$ se vi è discordanza
- $\rho_S = 0$ se le due graduatorie non mostrano associazione
- $0 < \rho_S < 1$ se vi è concordanza
- $\rho_S = 1$ se i ranghi sono in perfetta concordanza (le unità hanno la stessa posizione nelle due graduatorie)

5

Unità	*X*	*Y*	*Rango X*	*Rango Y*	d_i	d_i^2
A	*24*	24	4	3	1	1
B	*26*	22	3	4	-1	1
C	*28*	26	2	2	0	0
D	*30*	28	1	1	0	0

$$\rho_S = 1 - \frac{6\boxed{\sum_{i=1}^{n} d_i^2}}{n(n^2-1)} = 1 - \frac{6 \cdot 2}{4(16-1)} = 1 - \frac{12}{60} = 0.80$$

Elevata concordanza tra le due graduatorie

6

Dipendenza in media di un carattere quantitativo da un carattere qualitativo

- Si considerano *n* unità, un carattere qualitativo (X) e un carattere quantitativo (Y).
- Si vuole verificare se la media di Y sia dipendente da X (se la media di Y cambia al cambiare di X).
- Si analizzano le distribuzioni di Y condizionate a X e si calcolano la medie condizionate, $\bar{y}_{X=x_i}$

7

- Un carattere quantitativo Y è indipendente in media da X se tutte le medie condizionate sono uguali tra loro e alla media marginale,

$$\bar{y}_{X=x_i} = \ldots = \bar{y}_{X=x_H} = \bar{y}$$

- Un carattere quantitativo Y è dipendente in media da X se per almeno una modalità

$$\bar{y}_{X=x_i} \neq \bar{y}$$

i = 1,2,…,H.

8

«Condizione occupazionale» e «Anni di laurea»
1378 giovani

		Anni di laurea (Y)			
Cond. Occup. (X)		*4*	*5*	*6*	*TOT*
	Occupati	512	216	102	830
	In cerca di lavoro	121	213	214	548
	TOT	633	429	316	1378

9

Distribuzione del carattere «Anni di laurea» condizionata a $X = x_1$ = Occupati

Anni (Y)	n_{1j}
4	512
5	216
6	102
TOT	830

$$\bar{y}_{X=x_1} = \frac{4 \cdot 512 + 5 \cdot 216 + 6 \cdot 102}{830} = 4{,}51$$

10

Distribuzione del carattere «Anni di laurea» condizionata a X = x_2 = In cerca di lavoro

Anni (Y)	n_{2j}
4	121
5	213
6	214
TOT	548

$$\bar{y}_{X=x_2} = \frac{4 \cdot 121 + 5 \cdot 213 + 6 \cdot 214}{548} = 5{,}17$$

11

Distribuzione marginale del carattere «Anni di laurea»

Anni (Y)	$n_{\cdot j}$
4	633
5	429
6	316
TOT	1378

$$\bar{y} = \frac{4 \cdot 633 + 5 \cdot 429 + 6 \cdot 316}{1378} = 4{,}77$$

12

- Poiché

$$\bar{y}_{X=x_i} \neq \bar{y}_{X=x_2} \neq \bar{y}$$

esiste dipendenza in media del carattere Y da X.

13

Il rapporto di correlazione η^2

- Per quantificare la dipendenza, introduciamo la varianza delle medie condizionate,

$$\sigma^2_{Media(Y|X)} = \frac{\sum_{i=1}^{H}\left(\bar{y}_{X=x_i} - \bar{y}\right)^2 n_{i.}}{n}$$

e il rapporto di correlazione

$$\eta^2_{Y|X} = \frac{\sigma^2_{Media(Y|X)}}{\sigma^2_Y}$$

- $0 \leq \eta^2_{Y|X} \leq 1$

14

Interpretazione del rapporto di correlazione

- $\eta^2_{Y|X} = 0$ indipendenza in media (medie condizionate sono uguali tra loro e alla media marginale)
- $0 < \eta^2_{Y|X} < 1$ dipendenza in media
- $\eta^2_{Y|X} = 1$ dipendenza in media perfetta (ad ogni modalità di X è associato un solo valore di Y)

15

«Condizione occupazionale» e «Anni di laurea» 1378 giovani

		Anni di laurea (Y)			
Cond. Occup. (X)		*4*	*5*	*6*	*TOT*
	Occupati	512	216	102	830
	In cerca di lavoro	121	213	214	548
	TOT	633	429	316	1378

16

Calcolo di η^2

$\bar{y}_{X=x_1} = 4{,}51$ $\bar{y} = 4{,}77$

$\bar{y}_{X=x_2} = 5{,}17$

$$\sigma^2_{Media(Y|X)} = \frac{(4{,}51 - 4{,}77)^2 830 + (5{,}17 - 4{,}77)^2 548}{1378} = 0{,}104$$

$$\sigma^2_Y = ?$$

17

Distribuzione marginale del carattere «Anni di laurea»

Anni (Y)	*$n_{\cdot j}$*
4	633
5	429
6	316
TOT	1378

$\bar{y} = 4{,}77$

$$\sigma^2_Y = \frac{(4 - 4{,}77)^2 633 + (5 - 4{,}77)^2 429 + (6 - 4{,}77)^2 316}{1378} = 0{,}636$$

18

- Il rapporto di correlazione è

$$\eta^2_{Y|X} = \frac{\sigma^2_{Media(Y|X)}}{\sigma^2_Y} = \frac{0{,}104}{0{,}636} = 0{,}163$$

Debole dipendenza in media

19

Misura dell'associazione tra due caratteri quantitativi

- Si considerano *n* unità e due caratteri quantitativi (X, Y).
- Rappresentazione grafica: grafico (o diagramma) di dispersione.

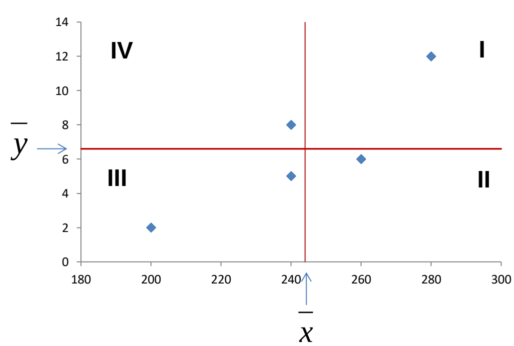
20

Tabella 12 – «Numero di parafarmacie» (X) e «Spesa per articoli sanitari (milioni euro)» (Y)

X	*Y*
200	2
240	5
260	6
240	8
280	12

21

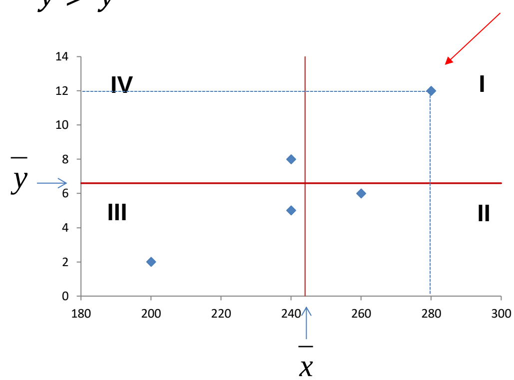
Grafico di dispersione



22

- I quadrante: valori di X e Y maggiori della media

  $x > \bar{x} \qquad y > \bar{y}$



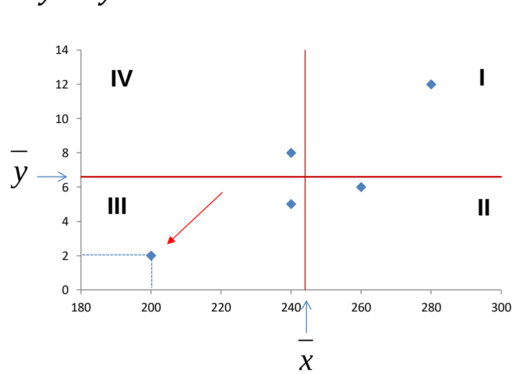
- Quindi gli scostamenti dalla media hanno lo stesso segno (positivo): scostamenti <u>concordi</u>

  $x - \bar{x} > 0 \qquad y - \bar{y} > 0$

23

- III quadrante: valori di X e Y minori della media

  $x < \bar{x} \qquad y < \bar{y}$



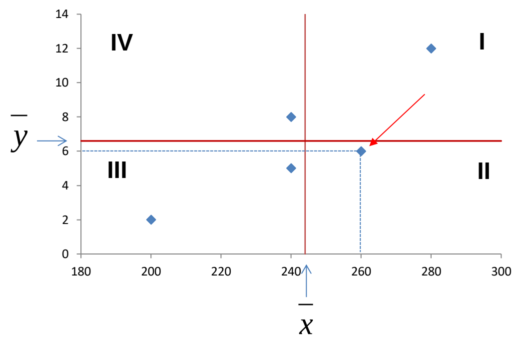
- Scostamenti <u>concordi</u>

  $x - \bar{x} < 0 \qquad y - \bar{y} < 0$

24

- II quadrante:

$x > \bar{x} \qquad y < \bar{y}$



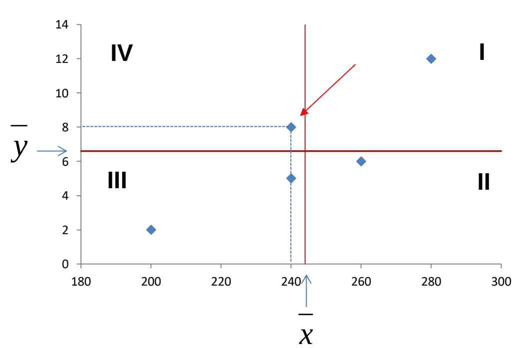
- Scostamenti <u>discordi</u>

$x - \bar{x} > 0 \qquad y - \bar{y} < 0$

25

- IV quadrante:

$x < \bar{x} \qquad y > \bar{y}$



- Scostamenti <u>discordi</u>

$x - \bar{x} < 0 \qquad y - \bar{y} > 0$

26

Concordanza

- Due caratteri quantitativi presentano concordanza se dominano gli scostamenti dalla media concordi.
- Concordanza (relazione diretta):

  all'aumentare di una variabile ci si aspetta un aumento dell'altra variabile, in pratica le due variabili tendono a muoversi nella stessa direzione

27

Discordanza

- Due caratteri quantitativi presentano discordanza se dominano gli scostamenti dalla media discordi.
- Discordanza (relazione inversa):

  all'aumentare di una variabile ci si aspetta un decremento dell'altra variabile, in pratica le due variabili tendono a muoversi in direzioni opposte

28

La covarianza

- È una misura del grado di concordanza/discordanza tra due caratteri quantitativi.

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

29

Interpretazione della covarianza

- $\sigma_{XY} < 0$ se esiste discordanza (relazione inversa, associazione negativa)
- $\sigma_{XY} = 0$ se le due variabili non sono associate
- $\sigma_{XY} > 0$ se esiste concordanza (relazione diretta, associazione positiva)

30

Tabella 12 – «Numero di parafarmacie» (X) e «Spesa per articoli sanitari (milioni di euro)» (Y)

X	*Y*
200	2
240	5
260	6
240	8
280	12

$\bar{x}$=244 $\bar{y}$=6,6

31

Calcolo di σ_{XY}

X	*Y*	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
200	2	-44	-4,6	202,4
240	5	-4	-1,6	6,4
260	6	16	-0,6	-9,6
240	8	-4	1,4	-5,6
280	12	36	5,4	194,4
TOT				388

32

$$\sigma_{XY} = \frac{388}{5} = 77{,}6$$

Esiste concordanza.

33

Formula alternativa di σ_{XY}

- La covarianza è anche data da

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

- σ_{XY} = media del prodotto – prodotto delle medie

34

Il coefficiente di correlazione lineare

- Il coefficiente di correlazione lineare di Bravais-Pearson è dato da

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

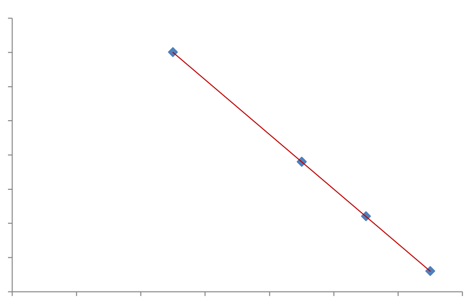
$$-1 \leq \rho_{XY} \leq 1$$

- Il coefficiente di correlazione lineare di Bravais-Pearson ha sempre lo stesso segno della covarianza.

35

Interpretazione di ρ_{XY}

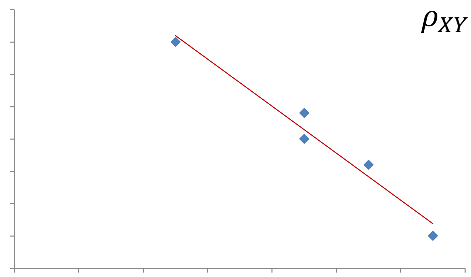
- $\rho_{XY} = -1$ se esiste perfetta discordanza (o relazione inversa). I punti del grafico di dispersione sono allineati su una retta con pendenza negativa



36

- $-1 < \rho_{XY} < 0$ se esiste discordanza (o relazione inversa). I punti del grafico di dispersione sono <u>intorno</u> ad una retta con pendenza negativa

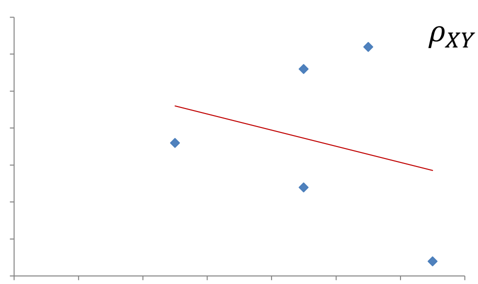
$\rho_{XY} = -0{,}98$



37

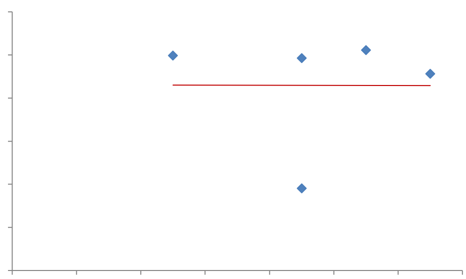
- $-1 < \rho_{XY} < 0$ se esiste discordanza (o relazione inversa). I punti del grafico di dispersione sono <u>intorno</u> ad una retta con pendenza negativa

$\rho_{XY} = -0{,}27$



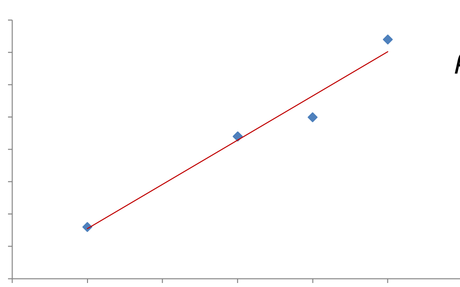
38

- $\rho_{XY} = 0$ se non esiste associazione. I punti del grafico di dispersione sono intorno ad una retta con pendenza nulla



39

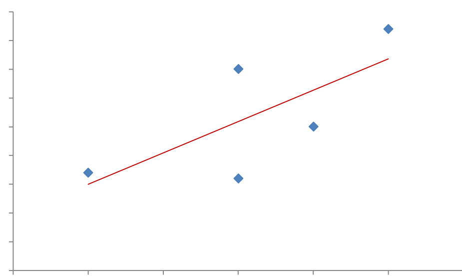
- $0 < \rho_{XY} < 1$ se esiste concordanza (o relazione diretta). I punti del grafico di dispersione sono intorno ad una retta con pendenza positiva



$\rho_{XY} = 0{,}98$

40

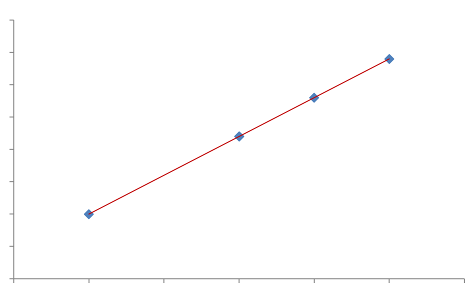
- $0 < \rho_{XY} < 1$ se esiste concordanza (o relazione diretta). I punti del grafico di dispersione sono intorno ad una retta con pendenza positiva



$\rho_{XY} = 0{,}71$

41

- $\rho_{XY} = 1$ se esiste perfetta concordanza (o relazione diretta). I punti del grafico di dispersione sono allineati su una retta con pendenza positiva



42

Calcolo di ρ_{XY}

X	Y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
200	2	-44	-4,6	1936	21,16
240	5	-4	-1,6	16	2,56
260	6	16	-0,6	256	0,36
240	8	-4	1,4	16	1,96
280	12	36	5,4	1296	29,16
TOT				3520	55,2

43

$$\sigma_X = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} = \sqrt{\frac{3520}{5}} = 26{,}53$$

$$\sigma_Y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}} = \sqrt{\frac{55{,}2}{5}} = 3{,}32$$

$$\rho_{XY} = \frac{77{,}6}{26{,}53 \cdot 3{,}32} = 0{,}88$$

Esiste una forte relazione diretta.

44

Ordine delle variabili

- L'ordine delle due variabili non è rilevante per la covarianza e la correlazione:

$$\sigma_{XY} = \sigma_{YX}$$

$$\rho_{XY} = \rho_{YX}$$

45