Caso studio 10

Si consideri la seguente distribuzione degli occupati in Italia secondo il numero di ore settimanali effettivamente lavorate e il settore di attività (cfr. Italia in cifre, Anno 2008, pag. 17):

	Ore lavorate		
Settore di attività	Fino a 10	10-30	30-45
Agricoltura	23	158	688
Industria	63	616	5721
Servizi	413	3364	10184

Si può concludere che le ore lavorate siano connesse al settore di attività? Come si può sintetizzare il grado di dipendenza tra i due caratteri in un unico indice da poter utilizzare nei confronti (ad esempio in confronti temporali o con altri paesi)?

Dipendenza in media

- La *misurazione della connessione* tra due caratteri si basa solo sulle frequenze congiunte senza tenere conto delle modalità dei due caratteri.
- Quando uno dei due caratteri (tipicamente *Y*) è *quantitativo*, è possibile confrontare le distribuzioni condizionate di *Y* tramite le medie condizionate.
- Per l'*i*-esima modalità di *X* (x_i), la media condizionata di *Y* è data da

$$\bar{y}_{X=x_i} = \frac{1}{n_{io}} \sum_{j=1}^{K} y_j n_{ij}$$

- Nel caso in cui il carattere *Y* è in classi, occorre utilizzare i valori centrali delle classi al posto delle modalità.

- Il carattere *Y* si dice ***indipendente in media*** da *X* quando *X* non influenza la media di *Y*. In termini matematici:

$$\bar{y}_{X=x_1} = \bar{y}_{X=x_2} = \cdots = \bar{y}_{X=x_H} = \bar{y}$$

dove la *media marginale* di *Y* è data da

$$\bar{y} = \frac{1}{n} \sum_{j=1}^{K} y_j n_{oj}$$

- Altrimenti, *Y* è ***dipendente in media*** da *X*.

Esempio

	Reddito annuo (x 1.000€)		
Titolo di studio	0-10	10-30	30-100
Lic. Media	88	142	120
Diploma	9	39	38
Laurea	3	19	42

- La media della distribuzione condizionata del *reddito* quando il *titolo di studio* è la *licenza media* viene calcolata come

Licenza media

Reddito (y_j)	Frequenze (n_{1j})	Valori centrali (y_j)	$y_j n_{1j}$
0-10	88	5	440
10-30	142	20	2840
30-100	120	65	7800
Totale	350	---	11080

$$\bar{y}_{X=x_1} = \frac{1}{n_{1o}} \sum_{j=1}^{K} y_j n_{1j} = \frac{1}{350} 11080 = 31{,}66$$

Reddito annuo (x 1.000€) – val.centrali

Titolo di studio	5	20	65	Totale
Lic. Media	88	142	120	350
Diploma	9	39	38	86
Laurea	3	19	42	64
Totale	100	200	200	500

- Le varie medie condizionate si possono calcolare come:

$$\bar{y}_{X=x_1} = \frac{1}{n_{1o}} \sum_{j=1}^{K} y_j n_{1j} = \frac{5 \cdot 88 + 20 \cdot 142 + 65 \cdot 120}{350} = 31{,}66$$

$$\bar{y}_{X=x_2} = \frac{1}{n_{2o}} \sum_{j=1}^{K} y_j n_{2j} = \frac{5 \cdot 9 + 20 \cdot 39 + 65 \cdot 38}{86} = 38{,}31$$

$$\bar{y}_{X=x_3} = \frac{1}{n_{3o}} \sum_{j=1}^{K} y_j n_{3j} = \frac{5 \cdot 3 + 20 \cdot 19 + 65 \cdot 42}{64} = 48{,}83$$

e quindi il reddito dipende in media dal titolo di studio.

- La media della distribuzione marginale del *reddito* è

$$\bar{y} = \frac{1}{n} \sum_{j=1}^{K} y_j n_{oj} = \frac{5 \cdot 100 + 20 \cdot 200 + 65 \cdot 200}{500} = 35$$

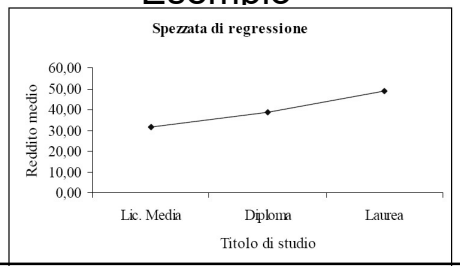
Spezzata di regressione

- La *spezzata di regressione* è un grafico che consiste nel rappresentare, e congiungere con dei segmenti, i punti di coordinate

$$\left(x_i, \bar{y}_{X=x_i}\right)$$

- La spezzata di regressione permette di intuire come varia la media di *Y* al variare della modalità di *X* .

Esempio



Esercizio

Si considerino i dati del Caso Studio 10.

Il numero medio di ore lavorate dipende dal settore di occupazione? In che modo? Per rispondere si calcolino le medie del carattere «Ore lavorate» condizionatamente al settore di occupazione. Si rappresentino graficamente i risultati utilizzando la spezzata di regressione.

Misura della dipendenza in media

- Per verificare se c'è o meno *dipendenza in media* di *Y* da *X* si può utilizzare la ***varianza spiegata*** (o ***varianza delle medie condizionate***) che è definita come

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

- La varianza spiegata è sempre non negativa; in particolare
  - $\sigma^2_{Media(Y|X)} = 0 \quad \Rightarrow$ *indipendenza in media* di *Y* da *X*.
  - $\sigma^2_{Media(Y|X)} > 0 \quad \Rightarrow$ *dipendenza in media* di *Y* da *X*.
- Il massimo della varianza spiegata è dato dalla ***varianza totale*** di *Y*,

$$\sigma^2_Y = \frac{1}{n} \sum_{j=1}^{K} \left(y_j - \bar{y}\right)^2 n_{oj}$$

- La differenza tra la varianza totale e quella spiegata è chiamata varianza residua ed è sempre non negativa

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

- E' quindi possibile definire un *indice relativo* per misurare la dipendenza in media, chiamato ***rapporto di correlazione***, come

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

- Per l'interpretazione del valore assunto da $\eta^2_{Y|X}$ si consideri che:
  - $\eta^2_{Y|X} = 0 \Rightarrow$ *indipendenza in media* di *Y* da *X*.
  - $\eta^2_{Y|X} > 0 \Rightarrow$ *dipendenza in media* di *Y* da *X*.
  - $\eta^2_{Y|X} = 1 \Rightarrow$ *massima dipendenza in media* di *Y* da *X*.

- La *dipendenza in media* ***implica*** la *dipendenza statistica* e quindi se

$$\eta^2_{Y|X} > 0 \Rightarrow \chi^2 > 0$$

- L'indipendenza in media non implica l'indipendenza statistica e quindi può accadere che

$$\eta^2_{Y|X} = 0 \quad e \quad \chi^2 > 0$$

Esempio

Titolo di studio (x_i)	Medie condizionate ($\bar{y}_{X=x_i}$)	$n_{i\circ}$	$(\bar{y}_{X=x_i} - \bar{y})^2 n_{io}$
Licenza media	31,66	350	3911,14
Diploma	38,31	86	944,48
Laurea	48,83	64	12237,89
Totale	---	500	17093,51

$$\sigma^2_{Media(Y|X)} = \frac{1}{n}\sum_{i=1}^{H}(\bar{y}_{X=x_i} - \bar{y})^2 n_{io} = 17093{,}51/500 = 34{,}19$$

Reddito	Frequenze (n_{ij})	Valori centrali (y_j)	$(y_j - \bar{y})^2 n_{oj}$
0-10	100	5	90000
10-30	200	20	45000
30-100	200	65	180000
Totale	500	---	315000

$$\sigma^2_Y = \frac{1}{n}\sum_{j=1}^{K}(y_j - \bar{y})^2 n_{oj} = \frac{315000}{500} = 630 \qquad \eta^2_{Y|X} = \frac{34{,}19}{630} = 0{,}054$$

Esercizio

La tabella sottostante riporta la distribuzione per sesso e per voto al diploma degli studenti di un certo istituto superiore.

a) Determinare se il voto al diploma sia in media indipendente dal sesso
b) Misurare il grado di dipendenza in media attraverso un apposito indice.
c) Verificare la scomposizione della varianza totale in varianza spiegata e varianza residua

	Voto				
Sesso	60-69	70-79	80-89	90-100	Totale
Femmine	20	29	35	66	150
Maschi	46	40	36	78	200
Totale	66	69	71	144	350

a) Determinare se il voto al diploma sia in media indipendente dal sesso

	Voto				
Sesso	60-69	70-79	80-89	90-100	Totale
Femmine	20	29	35	66	150
Maschi	46	40	36	78	200
Totale	66	69	71	144	350

Calcoliamo i valori centrali delle classi:

$$\frac{(60+69)}{2} = 64{,}5 \;;\; \frac{(70+79)}{2} = 74{,}5 \;;\; \frac{(80+89)}{2} = 84{,}5 \;;\; \frac{(90+100)}{2} = 95$$

Calcoliamo il voto medio condizionatamente al sesso:

$$\bar{y}_{X=F} = \frac{64{,}5 \cdot 20 + 74{,}5 \cdot 29 + 84{,}5 \cdot 35 + 95 \cdot 66}{150} = \frac{12678}{150} = 84{,}52$$

$$\bar{y}_{X=M} = \frac{64{,}5 \cdot 46 + 74{,}5 \cdot 40 + 84{,}5 \cdot 36 + 95 \cdot 78}{200} = \frac{16399}{200} = 81{,}99$$

Calcoliamo il voto medio generale:

$$\bar{y} = \frac{64{,}5 \cdot 66 + 74{,}5 \cdot 69 + 8\,,5 \cdot 71 + 95 \cdot 144}{35} = \frac{29077}{350} = 83{,}08 \text{ o } \bar{y} = \frac{84{,}52 \cdot 150 + 81{,}99 \cdot 200}{350} = \frac{29077}{350} = 83{,}08$$

Essendo le medie condizionate diverse tra loro e diverse dalla media generale, possiamo concludere che il voto dipende in media dal sesso.

b) Misurare il grado di dipendenza in media attraverso un apposito indice.

	Voto				
Sesso	64,5	74,5	84,5	95	Totale
Femmine	20	29	35	66	150
Maschi	46	40	36	78	200
Totale	66	69	71	144	350

Calcoliamo la varianza delle medie condizionate:

$\bar{y}_{X=F} = 84{,}52$; $\bar{y}_{X=M} = 81{,}99$; $\bar{y} = 83{,}08$

$$\sigma^2_{Media(Y|X)} = \frac{(84{,}52 - 83{,}08)^2 \cdot 150 + (81{,}99 - 83{,}08)^2 \cdot 200}{350} = \frac{546{,}48}{350} = 1{,}56$$

Calcoliamo la varianza totale:

$$\sigma^2_Y = \frac{(64{,}5 - 83{,}08)^2 \cdot 66 + (74{,}5 - 83{,}08)^2 \cdot 69 + (84{,}5 - 83{,}08)^2 \cdot 71 + (95 - 83{,}08)^2 \cdot 144}{350}$$

$$= \frac{48467{,}4}{350} = 138{,}48$$

Calcoliamo il rapporto di correlazione: $\eta^2 = \frac{1{,}56}{138{,}48} = 0{,}01$

Il grado di dipendenza in media del voto d'esame dal sesso è bassissimo.

13

c) Verificare la scomposizione della varianza totale in varianza spiegata e varianza residua

	Voto				
Sesso	64,5	74,5	84,5	95	Totale
Femmine	20	29	35	66	150
Maschi	46	40	36	78	200
Totale	66	69	71	144	350

$\bar{y}_{X=F} = 84{,}52$; $\bar{y}_{X=M} = 81{,}99$; $\bar{y} = 83{,}08$

$\sigma^2_{Media(Y|X)} = \frac{(84{,}52-83{,}08)^2 \cdot 150 + (81{,}99-83{,}08)^2 \cdot 200}{350} = \frac{546{,}48}{350} = 1{,}56;$

$\sigma^2_Y = \frac{(64{,}5-83{,}08)^2 \cdot 66 + (74{,}5-83{,}08)^2 \cdot 69 + (84{,}5-83{,}08)^2 \cdot 71 \quad (95-83{,}08)^2 \cdot 144}{350} = 138{,}48;$

Calcoliamo la varianza residua come media delle varianze condizionate:

$\sigma^2_{Y|X=F} = \frac{(64{,}5-84{,}52)^2 \cdot 20 + (74{,}5-84{,}52)^2 \cdot 29 \quad (84{,}5-84{,}52)^2 \cdot 35 \quad (95-84{,}52)^2 \cdot 66}{150} = 121{,}2;$

$\sigma^2_{Y|X=M} = \frac{(64{,}5-81{,}99)^2 \cdot 46 + (74{,}5-81{,}99)^2 \cdot 40 + (84{,}5-81{,}99)^2 \cdot 36 + (95-81{,}99)^2 \cdot 78}{200} = 148{,}7;$

$Media(\sigma^2_{Y|X=x}) = \frac{121{,}2 \cdot 150 + 148{,}7 \cdot 200}{350} = 136{,}92;$

Decomposizione della varianza totale

$\sigma^2_Y = \sigma^2_{Media(Y|X)} + Media(\sigma^2_{Y|X=x}) \quad \rightarrow \quad 138{,}48 = 1{,}56 + 136{,}92$

14

Esercizio

Si considerino i dati del Caso Studio 10.

Si misuri il grado di dipendenza in media del numero di ore lavorate dal settore di occupazione. Si verifichi inoltre se c'è stata una variazione nel grado di dipendenza in media con riferimento ai dati relativi al 2014, riportati sotto:

	Ore lavorate		
Settore di attività	Fino a 10	10-30	30-45
Agricoltura	69	182	649
Industria	748	810	5251
Servizi	2039	4593	10888

Si aggiunga la nuova spezzata di regressione al grafico disegnato precedentemente.

15

Caso studio 11

Un investitore ha nel suo portafoglio titoli il titolo A. Vuole scegliere tra il titolo B e il titolo C, un titolo da aggiungere al suo portafoglio. Per poter effettuare la scelta, osserva il logaritmo dei rendimenti registrati per i tre titoli negli ultimi 10 giorni. I risultati sono riportati in tabella:

Data	**Titolo A**	**Titolo B**	**Titolo C**
15/02/16	2,4	3,2	-0,7
16/02/16	1,1	2,3	-0,2
17/02/16	-0,3	0,4	1,1
18/02/16	-0,5	-0,2	2,5
19/02/16	-0,8	-0,3	3
22/02/16	0,3	0,6	1,2
23/02/16	2	1,7	0,8

Quale titolo conviene scegliere?

16

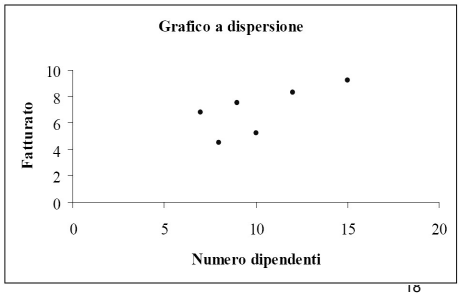
Correlazione

- Quando i caratteri *X* e *Y* studiati sono quantitativi, se ne può dare una rappresentazione grafica attraverso un ***grafico a dispersione*** (o ***nuvola di punti***).
- Un grafico a dispersione si ottiene riportando sul piano cartesiano i punti di coordinate (x_i, y_i) che rappresentano le modalità dei due caratteri presentate dall'unità *i*-esima.
- Si dice che *X* e *Y* sono ***correlati positivamente*** quando al crescere di *X* anche *Y* tende a crescere. Si dice che *X* e *Y* sono ***correlati negativamente*** quando al crescere di *X*, *Y* tende a decrescere.

Esempio

- Per un collettivo di $n = 6$ imprese sono state rilevate le modalità dei caratteri *numero di dipendenti* (*X*) e *fatturato* (*Y*).

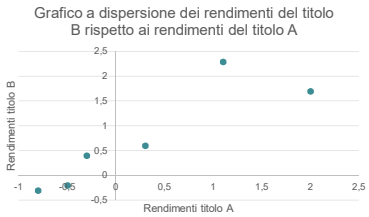
Dipendenti	Fatturato
15	9,2
10	5,2
8	4,5
7	6,8
12	8,3
9	7,5

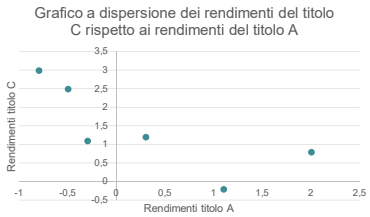


Esercizio

Si considerino i dati del Caso Studio 11.

Si disegni il grafico a dispersione del logaritmo dei rendimenti del titolo B rispetto al titolo A e del titolo C rispetto al titolo A.





- Una misura della *concordanza* tra *X* e *Y* è data dalla ***covarianza***, un indice simmetrico calcolato come:

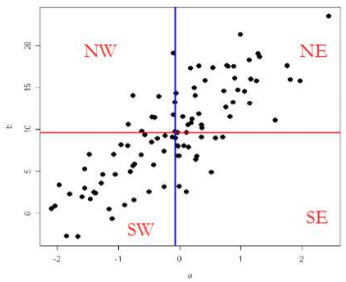
$$\sigma_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = Media(XY) - \bar{x}\bar{y}$$

- La covarianza assume valori nell'intervallo

$$-\sigma_X\sigma_Y \le \sigma_{XY} \le \sigma_X\sigma_Y$$

- Dividendo la covarianza per $\sigma_X\sigma_Y$ si ottiene il coefficiente di correlazione lineare di Bravais-Pearson:

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



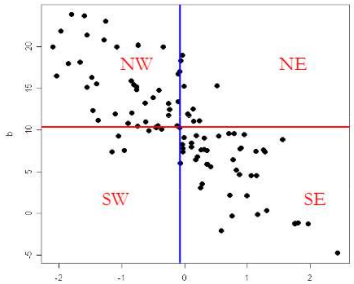
CONCORDANZA

NW $(x_i - \bar{x})(y_i - \bar{y}) < 0$ NE $(x_i - \bar{x})(y_i - \bar{y}) > 0$

SE $(x_i - \bar{x})(y_i - \bar{y}) < 0$ SW $(x_i - \bar{x})(y_i - \bar{y}) > 0$

I punti si trovano in maggioranza nei quadranti NE e SW

→ covarianza POSITIVA (associazione **concordante**)



DISCORDANZA

NW $(x_i - \bar{x})(y_i - \bar{y}) < 0$ NE $(x_i - \bar{x})(y_i - \bar{y}) > 0$

SE $(x_i - \bar{x})(y_i - \bar{y}) < 0$ SW $(x_i - \bar{x})(y_i - \bar{y}) > 0$

I punti si trovano in maggioranza nei quadranti NW e SE

→ covarianza NEGATIVA (associazione **discordante**)

Proprietà del coefficiente di correlazione

- $-1 \le \rho_{XY} \le 1$
- $\rho_{XY} = 1$ se *Y* = a+b*X* e i due caratteri sono concordi.
- $\rho_{XY} = -1$ se *Y* = a+b*X* e i due caratteri sono discordi.
- $\rho_{XY} = 0$ se *X* e *Y* sono indipendenti, oppure se la loro relazione non è lineare.

Esempio

Dipendenti (x_i)	**Fatturato** (y_i)	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$y_i x_i$
15	9,2	4,833	2,283	11,036	138
10	5,2	-0,167	-1,717	0,286	52
8	4,5	-2,167	-2,417	5,236	36
7	6,8	-3,167	-0,117	0,369	47,6
12	8,3	1,833	1,383	2,536	99,6
9	7,5	-1,167	0,583	-0,681	67,5
61	41,5	0	0	18,783	440,7

$\bar{x} = 61/6 = 10{,}17$ $\sigma_{XY} = 18{,}783/6 = 3{,}13$

$\bar{y} = 41{,}5/6 = 6{,}92$ $\sigma_{XY} = 440{,}7/6 - 10{,}17 \cdot 6{,}92 = 3{,}13$

Dipendenti (x_i)	**Fatturato** (y_i)	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
15	9,2	4,833	2,283	23,361	5,214
10	5,2	-0,167	-1,717	0,028	2,947
8	4,5	-2,167	-2,417	4,694	5,840
7	6,8	-3,167	-0,117	10,028	0,014
12	8,3	1,833	1,383	3,361	1,914
9	7,5	-1,167	0,583	1,361	0,340
61	41,5	0	0	42,833	16,268

$\sigma_X = \sqrt{42,833/6} = 2,672$

$\sigma_Y = \sqrt{16,268/6} = 1,647$

$\rho_{XY} = \frac{3,13}{2,672 \cdot 1,647} = 0,712$

25

Esercizio

Si considerino i dati del Caso Studio 11.

Si calcoli l'indice di correlazione di Bravais-Pearson tra i log-rendimenti del titolo A e quelli del titolo B, e poi tra quelli del titolo A e quelli del titolo C. In base al risultato si stabilisca quale titolo sia meglio aggiungere al portafoglio.

26

Dove e come studiare

- Libro di testo: S. Borra, A. Di Ciaccio (2014), Cap. 6 (escluso paragrafo 6.7)
- Svolgere 'Esercitazione 3', esercizi non precedentemente svolti.
- Svolgere gli esercizi nel file 'Esercizi di statistica bivariata.xls', fogli 2, 6, 11, 12, 13.

27