

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
$$\bar{y}_{X=x_i} = \frac{1}{n_{i\cdot}} \sum_{j=1}^K y_j n_{ij}$$
- Nel caso in cui il carattere Y è in classi, occorre utilizzare i valori centrali delle classi al posto delle modalità.

Esempio

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

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

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

- Si ha anche

$$\bar{y}_{X=x_2} = 38,31 \quad (\text{titolo di studio} = \textit{diploma})$$

$$\bar{y}_{X=x_3} = 48,83 \quad (\text{titolo di studio} = \textit{laurea})$$

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

- La media della distribuzione marginale del *reddito* è

$$\bar{y} = 35$$

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

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

dove la *media marginale* di Y è data da

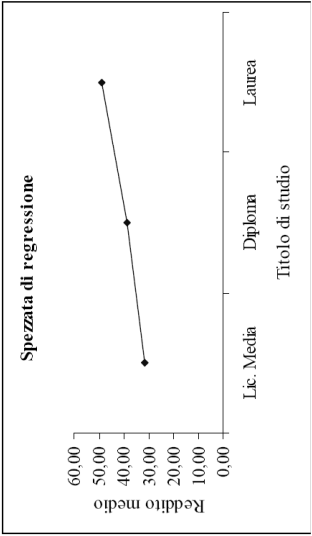
$$\bar{y} = \frac{1}{n} \sum_{j=1}^K y_j n_{\cdot j}$$

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

Spezzata di regressione

- La *spezzata di regressione* è un grafico che consiste nel rappresentare, e congiungere con dei segmenti, i punti di coordinate $(x_i, \bar{y}_{X=x_i})$
- La spezzata di regressione permette di intuire come varia la media di Y al variare della modalità di X .

Esempio



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
$$\sigma^2_{Media(Y|X)} = \frac{1}{n} \sum_{i=1}^H (\bar{y}_{X=x_i} - \bar{y})^2 n_{i_0}$$
- La varianza spiegata è sempre non negativa; in particolare
 - $\sigma^2_{Media(Y|X)} = 0 \Rightarrow$ *indipendenza in media* di Y da X.
 - $\sigma^2_{Media(Y|X)} > 0 \Rightarrow$ *dipendenza in media* di Y da X.
- Il massimo della varianza spiegata è dato dalla **varianza totale** di Y,

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

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

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

Esempio

Titolo di studio (x_i)	Medie condizionate ($\bar{y}_{X=x_i}$)	n_{i^*}	$(\bar{y}_{X=x_i} - \bar{y})^2 n_{i_0}$
Licenza media	31,66	350	3911,14
Diploma	38,31	86	944,48
Laurea	48,83	64	12237,89
Totale	---	500	17093,51

$$\sigma^2_{Media(Y|X)} = \frac{1}{n} \sum_{i=1}^H (\bar{y}_{X=x_i} - \bar{y})^2 n_{i_0} = 17093,51 / 500 = 34,19$$

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

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

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

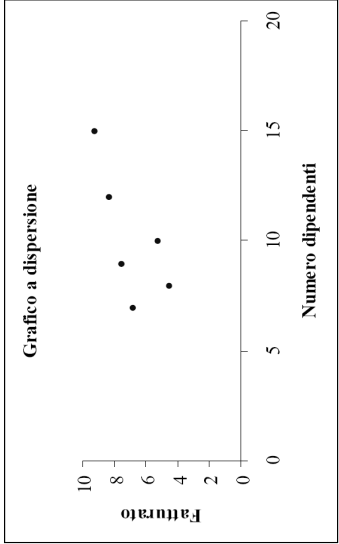
Correlazione

- 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_j) 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



Proprietà del coefficiente di correlazione

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

- 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

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

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

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

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

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

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

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

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

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

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