Metodi Quantitativi per Economia, Finanza e Management

Lezione n°4

Analisi Bivariata I° Parte

Statistica descrittiva bivariata

Indaga la relazione tra due variabili misurate. Si distingue rispetto alla tipologia delle variabili indagate:

- var. qualitative/quantitative discrete: tavole di contingenza (o a doppia entrata)
- var. quantitative: analisi di correlazione lineare
- una var. qualitativa e una quantitativa: confronto tra le medie

Tavole di contingenza

Sono tabelle a doppia entrata; i valori riportati all'interno della tabella sono le frequenze congiunte assolute, e la loro somma è pari al totale dei casi osservati.

Dalla tabella si possono ricavare inoltre le distribuzioni marginali, sommando per riga e per colonna le frequenze congiunte; le frequenze relative congiunte, pari al rapporto tra le frequenze assolute congiunte e il totale dei casi osservati.

Sesso * Età Crosstabulation

			Età				
			18-25	26-35	36-50	Over 50	Total
Sesso	M	Count	25	22	22	17	86
		% within Sesso	29.1%	25.6%	25.6%	19.8%	100.0%
		% within Età	32.1%	40.0%	53.7%	36.2%	38.9%
		% of Total	11.3%	10.0%	10.0%	7.7%	38.9%
	F	Count	53	33	19	30	135
		% within Sesso	39.3%	24.4%	14.1%	22.2%	100.0%
		% within Età	67.9%	60.0%	46.3%	63.8%	61.1%
		% of Total	24.0%	14.9%	8.6%	13.6%	61.1%
Total		Count	78	55	41	47	221
		% within Sesso	35.3%	24.9%	18.6%	21.3%	100.0%
		% within Età	100.0%	100.0%	100.0%	100.0%	100.0%
		% of Total	35.3%	24.9%	18.6%	21.3%	100.0%

Tavole di contingenza

Dalle tabelle di contingenza si possono ricavare ulteriori distribuzioni unidimensionali :

- *Frequenze subordinate* ovvero la frequenza di osservare il carattere x dato il carattere y e viceversa. Formalmente:

$$P\ y|x\ (x_i,y_j) = P\ (x_i,y_j) / P\ x(x_i)$$

$$P\ x|y\ (x_i,y_j) = P\ (x_i,y_j) / P\ y(y_j)$$

Indipendenza statistica se al variare di X le distribuzioni subordinate $(Y|X)= x_i$ sono tutte uguali tra loro,si può concludere che la distribuzione del carattere Y non dipende da X. Nel caso di indipendenza statistica, la frequenza relativa congiunta è pari al prodotto delle marginali corrispondenti

$$P(x_i,y_j)=P_x\ (x_i)P_y(y_j)$$

L'indipendenza stat. è un concetto simmetrico: se vale per X, vale anche per Y. Se si verifica, vuol dire che l'analisi bivariata di X (Y) non dà informazioni aggiuntive rispetto all'analisi univariata.

Tavole di contingenza

- *Perfetta dipendenza unilaterale* ad ogni valore di X corrisponde un solo valore di Y, ma non è detto che si verifichi il contrario. In generale, quando il numero di colonne (valori assunti dalla Y) è inferiore al numero di righe (valori assunti dalla X) non è mai possibile che X dipenda perfettamente da Y.

- *Perfetta dipendenza bilaterale* ad ogni valore di X corrisponde un solo valore di Y e viceversa; la perfetta dipendenza bilaterale si può avere allora solo per matrici quadrate.

Indici di connessione

Nella realtà è difficile che si verifichi la condizione di indipendenza statistica. Pertanto è utile disporre di indici che misurino il grado di connessione tra le variabili.

- χ^2 (chi-quadrato) assume valore nullo se i fenomeni X e Y sono indipendenti. Risente del numero delle osservazioni effettuate quindi al crescere di N, l'indice tende a crescere.

$$\chi^2 = N \Sigma \Sigma [P(x_i,y_j) - P_x(x_i)\ _y(y_j)]^{2} / P_x(x_i)\ P_y(y_j)$$

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	5.471[a]	3	.140
Likelihood Ratio	5.402	3	.145
N of Valid Cases	221		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 15.95.

Indici di connessione

– Un indice più efficace (perchè relativo, e dunque non risente del numero di osservazioni) è l'indice di Cramer V, basato sul χ^2. assume valori compresi tra 0 e 1: 0 nel caso di indipendenza statistica, 1 nel caso di perfetta dipendenza almeno unilaterale e tende a crescere all'aumentare del grado di dipendenza delle variabili considerate.

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	.157	.140
	Cramer's V	.157	.140
N of Valid Cases		221	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Indici di connessione

Nella realtà è difficile che si verifichi la condizione di indipendenza statistica. Pertanto è utile disporre di indici che misurino il grado di connessione tra le variabili.

- χ^2 (chi-quadrato) assume valore nullo se i fenomeni X e Y sono indipendenti. Risente del numero delle osservazioni effettuate quindi al crescere di N, l'indice tende a crescere.

$$\chi^2 = N \Sigma \Sigma [P(x_i,y_j) - P_x(x_i)\ _y(y_j)]^{2} / P_x(x_i)\ P_y(y_j)$$

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	5.471[a]	3	.140
Likelihood Ratio	5.402	3	.145
N of Valid Cases	221		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 15.95.

Indici di connessione

– Un indice più efficace (perchè relativo, e dunque non risente del numero di osservazioni) è l'indice di Cramer V, basato sul χ^2. assume valori compresi tra 0 e 1: 0 nel caso di indipendenza statistica, 1 nel caso di perfetta dipendenza almeno unilaterale e tende a crescere all'aumentare del grado di dipendenza delle variabili considerate.

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	.157	.140
	Cramer's V	.157	.140
N of Valid Cases		221	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Statistica descrittiva bivariata

Indaga la relazione tra due variabili misurate. Si distingue rispetto alla tipologia delle variabili indagate:

- var. qualitative/quantitative discrete: tavole di contingenza (o a doppia entrata)
- var. quantitative: analisi di correlazione lineare
- una var. qualitativa e una quantitativa: confronto tra le medie

Correlazione lineare

Le misure di connessione possono essere applicate a variabili qualitative. Se si vuole misurare il grado di *concordanza* tra due variabili quantitative occorre utilizzare altri indici:

- *Covarianza* ***Cov(X,Y)*** è un indice che assume valori positivi se vi è concordanza tra X e Y (a modalità elevate dell'una, corrispondono modalità elevate dell'altra); assume valori negativi nel caso di discordanza (a modalità elevate dell'una non corrispondono modalità elevate dell'altra). Nel caso di indipendenza statistica, la covarianza assumerà valore nullo. È un indice assoluto, ovvero segnala la presenza e la direzione di un legame tra due variabili, ma nulla si può dire sul grado del loro legame.

$$Cov(X,Y)= \Sigma \Sigma (x_i-\mu_x) (y_j- \mu_y) p(x_i,y_j)$$

Correlazione lineare

- Covarianza tra due variabili:

Cov(x,y) > 0 → x e y tendono a muoversi nella stessa direzione

Cov(x,y) < 0 → x e y tendono a muoversi in direzioni opposte

Cov(x,y) = 0 → x e y no relazione lineare

  - Riguarda solo la forza della relazione, ma non implica un effetto causale

Correlazione lineare

– *Coefficiente di correlazione lineare* ***ρ(X,Y)*** è un indice relativo che ovvia al problema del precedente indice. Assume valori compresi tra -1 e 1. In particolare vale 1 se e solo se Y è funzione lineare di X (e viceversa) e in questo caso i punti corrispondenti alle osservazioni sono disposti su una retta con inclinazione positiva. Analogamente l'indice assume valore -1 nel caso in cui i punti siano disposti su una retta con inclinazione negativa. Assume valore nullo se tra le variabili non è presente alcun tipo di relazione lineare (indipendenti in correlazione).

Correlazione lineare

- *Coefficiente di correlazione lineare* ***ρ(X,Y)*** :

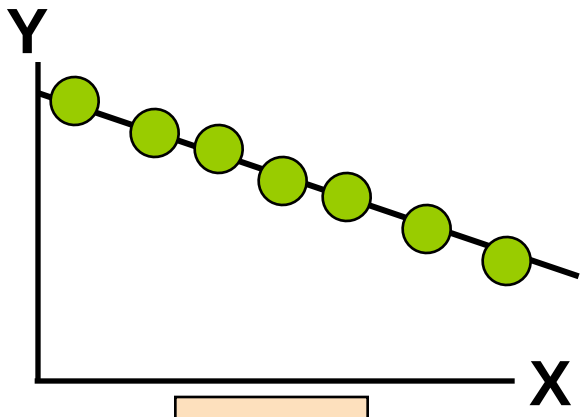
$$\rho = Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

- $\rho = 0$ → non c'è relazione lineare tra X e Y
- $\rho > 0$ → relazione lineare positiva tra X e Y
  - » quando X assume valori alti (bassi) allora anche Y probabilmente assume valori alti (bassi)
  - » $\rho = +1$ => dipendenza lineare perfetta positiva
- $\rho < 0$ → relazione lineare negativa tra X e Y
  - » quando X assume valori alti (bassi) allora Y probabilmente assume valori bassi (alti)
  - » $\rho = -1$ => dipendenza lineare perfetta negativa

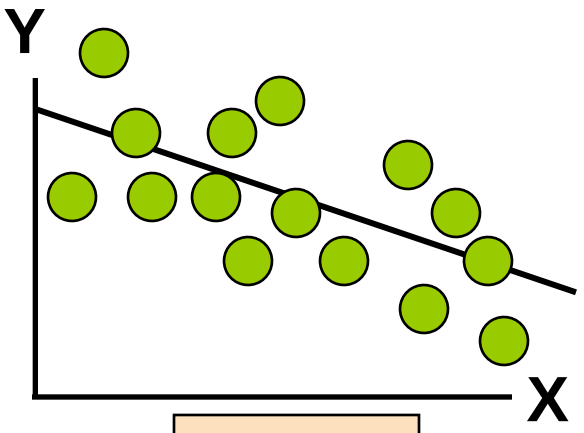
Correlazione lineare

- Senza unità di misura
- Campo di variazione fra –1 e 1
- Quanto più è vicino a –1, tanto più è forte la relazione lineare negativa
- Quanto più è vicino a 1, tanto più è forte la relazione lineare positiva
- Quanto più è vicino a 0, tanto più è debole la relazione lineare

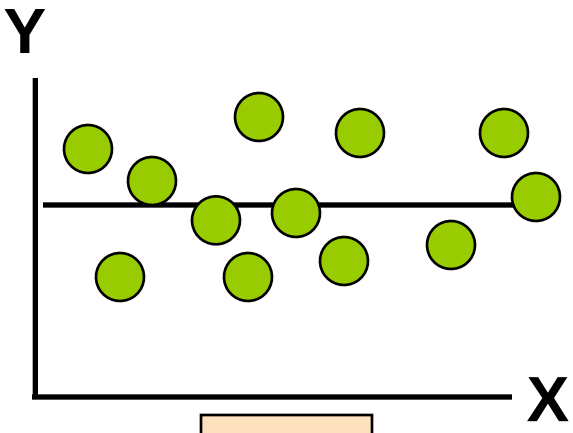
Correlazione lineare



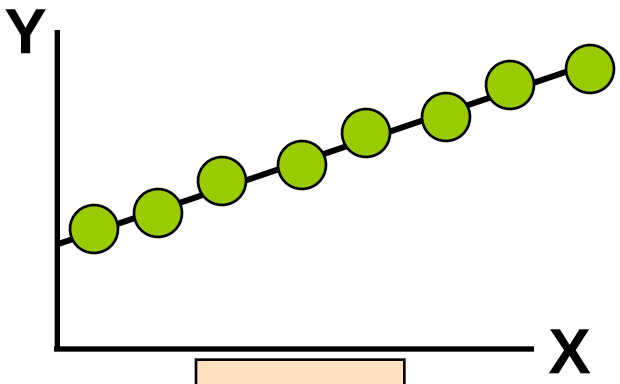
r = -1



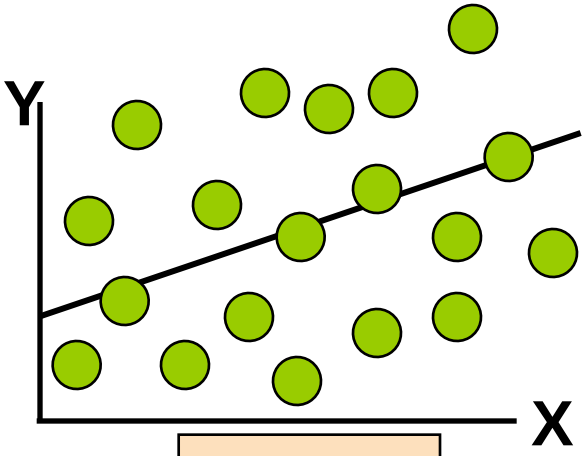
r = -0.6



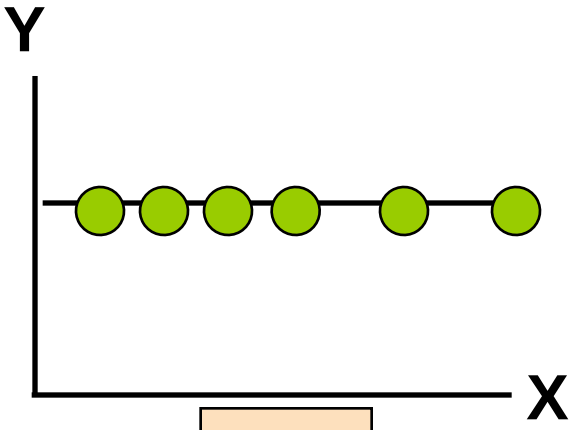
r = 0



r = +1



r = +0.3



r = 0

Correlazione lineare

Correlations

		Qualità degli ingredienti	Genuinità	Leggerezza	Sapore/gusto
Qualità degli ingredienti	Pearson Correlation	1	.629**	.299**	.232**
	Sig. (2-tailed)		.000	.000	.001
	N	220	220	218	220
Genuinità	Pearson Correlation	.629**	1	.468**	.090
	Sig. (2-tailed)	.000		.000	.181
	N	220	220	218	220
Leggerezza	Pearson Correlation	.299**	.468**	1	.030
	Sig. (2-tailed)	.000	.000		.657
	N	218	218	219	219
Sapore/gusto	Pearson Correlation	.232**	.090	.030	1
	Sig. (2-tailed)	.001	.181	.657	
	N	220	220	219	221

**. Correlation is significant at the 0.01 level (2-tailed).

Statistica descrittiva bivariata

Indaga la relazione tra due variabili misurate. Si distingue rispetto alla tipologia delle variabili indagate:

- var. qualitative/quantitative discrete: tavole di contingenza (o a doppia entrata)
- var. quantitative: analisi di correlazione lineare
- una var. qualitativa e una quantitativa: confronto tra le medie

Confronto tra le medie

Per misurazione della connessione tra una variabile quantitativa Y e una qualitativa X, è possibile confrontare le distribuzioni condizionate di Y tramite le medie condizionate.

Confronto tra le medie

Se si vuole incrociare una variabile quantitativa con una variabile qualitativa, la loro relazione può essere descritta confrontando le medie della variabile numerica all'interno delle categorie definite dalla variabile misurata a livello nominale/ordinale.

Rapidità *Tipo cliente*	*Media*	*N*
Persone fisiche	7.8403	357
Aziende	8.5132	76
Totale	7.9584	433

Confronto tra le medie

Un indice sintetico dell'intensità della relazione si basa sulla scomposizione della varianza per la variabile quantitativa Y, di cui viene studiata la dipendenza nei confronti della variabile categorica X. La variabilità totale di Y è

$$\mathbf{SQT_y = SQ_{tra} + SQ_{nei}}$$

dove

- **SQT_y** (somma dei quadrati tot) è la variabilità totale,
- **SQ_{tra}** variabilità tra i gruppi (somma dei quadr. tra i gruppi) esprime quanta variabilità di Y può essere legata al variare delle categorie di X,
- **SQ_{nei}** variabilità interna ai gruppi (somma dei quadr. nei gruppi) esprime la variabilità nell'andamento di Y indipendente da X.

Confronto tra le medie

E' quindi possibile definire un indice relativo per misurare la dipendenza in media, come

$$\eta^2 = SQ_{tra} / SQT_y = 1 - (SQ_{nei} / SQT_y)$$

Per l'interpretazione del valore assunto da η^2 si consideri che:

- $\eta^2 = 0 \Rightarrow$ indipendenza in media
- $\eta^2 > 0 \Rightarrow$ dipendenza in media
- $\eta^2 = 1 \Rightarrow$ massima dipendenza in media

η^2 è sempre compreso tra 0 e 1.

Confronto tra le medie

Report

Produzione artigianale

Età	Mean	N	Std. Deviation
18-25	5.01	78	2.224
26-35	5.53	55	2.609
36-50	6.00	41	2.098
Over 50	6.09	47	2.320
Total	5.55	221	2.352

Measures of Association

	Eta	Eta Squared
Produzione artigianale * Età	.191	.036

Modesta dipendenza in media della produzione artigianale dall'età

In caso di **indipendenza in media le medie dei diversi gruppi** (medie condizionate ai diversi livelli della variabile qualitativa) saranno tutte uguali tra loro e quindi la variabilità tra i gruppi sarà nulla. Viceversa qualora ad ogni livello della variabile qualitativa sia associato un unico valore della variabile quantitativa, si parlerà di massima dipendenza in media e si avrà variabilità interna ai gruppi nulla. **Per misurare l'intensità della dipendenza in media si può utilizzare l'indice Eta (radice quadrata di Eta Squared) considerando 0.2 come valore soglia oltre il quale si può asserire che esiste dipendenza in media tra le variabili. Aumentando il valore di Eta aumenta la dipendenza in media.**