Metodi Quantitativi per Economia, Finanza e Management

Lezione n°5

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				Total
			18-25	26-35	36-50	Over 50	
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.

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

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.