CORSO DI STATISTICA (parte 2) - ESERCITAZIONE 1

Dott.ssa Antonella Costanzo

a.costanzo@unicas.it

A.Studio dell'interdipendenza tra variabili: riepilogo

Concetto relativo allo studio delle relazioni tra due variabili statistiche che rivestono lo stesso ruolo ai fini dell'analisi. A seconda della tipologia di caratteri oggetto di studio:

- Connessione o indipendenza assoluta se si tratta di due mutabili (Indice χ^2di Pearson, Indice φ^2, Indice V di Cramer)
- Indipendenza in media se si tratta di variabili miste (Indice η^2di Pearson)
- Associazione e correlazione se si tratta di due caratteri quantitativi (covarianza, coefficiente di correlazione di Pearson ρ_{xy})

Esercizio 1 A. La connessione tra mutabili

La tabella seguente riporta la distribuzione di turisti arrivati nel 2012 in una località turistica italiana secondo la nazionalità e il tipo di struttura ricettiva prescelta:

X=Nazionalità |Y=Struttura ricettiva prescelta	**Alberghiera**	**Extra-Alberghiera**	**Totale**
Austria	3560	1325	4885
Germania	6589	3420	10009
Regno Unito	2345	1200	3545
Francia	4267	2350	6617
Danimarca	2000	1984	3984
Totale	18761	10279	**29040**

Misurare, se esiste il grado di connessione tra i due caratteri considerati.

Sol.

I caratteri X=Nazionalità e Y=Tipo di struttura alberghiera scelta sono entrambi qualitativi. Tra essi esiste indipendenza (assoluta) se le modalità assunte da X non modificano la distribuzione di Y e viceversa. In altre parole la distribuzione condizionata in frequenze relative di $Y|X = x_i$ non cambia per ogni i=1,2,…, h ed è uguale alla distribuzione marginale di Y. Similmente la distribuzione condizionata $X|Y = y_j$ non cambia per ogni j=1,2,…, k ed è uguale alle distribuzione marginale di X. In particolare:

- $\frac{n_{ij}}{n_{i.}} = \frac{n_{.j}}{n}$ j=1,…,k (colonne)
- $\frac{n_{ij}}{n_{.j}} = \frac{n_{i.}}{n}$ i=1,…, h (righe)

Siccome, in caso di indipendenza assoluta tra X e Y deve valere la seguente:

$$\frac{n_{ij}}{n_{i.}} = \frac{n_{.j}}{n}$$

da un punto di vista operativo, i due caratteri X e Y si dicono indipendenti (in distribuzione) se le frequenze osservate sono uguali alle cosiddette frequenze teoriche per ogni cella (i, j) della distribuzione doppia.

Frequenze teoriche (sotto ipotesi di indipendenza):

$$\hat{n}_{ij} = \frac{n_{i.}n_{.j}}{n}$$

Tabella teorica (sotto l' ipotesi di indipendenza)

X=Nazionalità |Y=Struttura ricettiva prescelta	**Alberghiera**	**Extra-Alberghiera**	**Totale**
Austria	3155.905	1729.095	4885
Germania	6466.214	3542.786	10009
Regno Unito	2290.212	1254.788	3545
Francia	4274.846	2342.154	6617
Danimarca	2573.823	1410.177	3984
Totale	18761	10279	**29040**

A questo punto confronto la tabella delle frequenze teoriche con quella delle frequenze osservate. Le frequenze teoriche sono diverse da quelle osservate, quindi concludo che i caratteri Nazionalità e Tipo di struttura alberghiera scelta non sono indipendenti. Un indice che misura il loro grado di connessione è l'indice χ^2 di Pearson:

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

L'indice assume valore 0 in caso di indipendenza mentre tende a crescere al crescere del grado di connessione tra i caratteri. Dai calcoli risulta:

$$\chi^2 = 517.94$$

Esiste pertanto connessione tra i caratteri. Una misura relativa dell'indipendenza in distribuzione è data dall'indice di contingenza in media quadratica che si ottiene dividendo il χ^2 per la numerosità delle osservazioni n.

$$\varphi^2 = \frac{\chi^2}{n} = \frac{517.94}{29040} = 0.0178$$

Tale valore va confrontato con l'intervallo [0, 1] in quanto $0 \leq \varphi^2 \leq \min(h-1, k-1)$ dove h=numero di righe e k=numero di colonne. Possiamo dire che esiste un basso grado di connessione.

L'indice V di Cramer (anche noto come indice T di Tchuprov) è la versione normalizzata dell'indice φ^2 ed è più semplice da interpretare, poiché varia tra 0 e 1.

$$Indice\ V = \sqrt{\frac{\chi^2}{\mathrm{n} * \min[(h-1), (k-1)]}} = \sqrt{\frac{517.94}{29040 * \min[(5-1), (2-1)]}} = 0.134$$

Possiamo concludere che il grado di connessione tra i due caratteri è pari al 13.4% del massimo valore che la connessione può raggiungere.

Esercizio 2 A. Indipendenza in media

La tabella seguente riporta la distribuzione di 700 studenti secondo il voto conseguito all'esame di statistica e secondo la frequenza delle lezioni al corso:

X=Frequenza Y=Voto all'esame	**[18, 20]**	**[21, 23]**	**[24, 25]**	**[26, 28]**	**[29, 30]**	**Totale**
Si	11	23	38	116	75	263
No	55	163	107	48	64	437
Totale	66	186	145	164	139	**700**

Quanta parte della variabilità totale dei voti è attribuibile alla dipendenza del voto medio dalla frequenza alle lezioni?

Sol.

Siamo interessati a verificare se la distribuzione dei voti riportati all'esame è indipendente in media dalla scelta degli studenti di frequentare o meno le lezioni. L'obiettivo è quindi un confronto fra le medie della distribuzione dei voti condizionata alla scelta in merito alla frequenza ai corsi e della distribuzione marginale del voto riportato all'esame. Esiste indipendenza in media tra i caratteri se, in particolare vale la seguente:

$$\mu(Y|X = x_1) = \mu(Y|X = x_2) \dots \mu(Y|X = x_k) = \mu(Y)$$

nel nostro caso, esiste indipendenza in media tra i caratteri se:

$$\mu(Y|X = SI) = \mu(Y|X = NO) = \mu(Y)$$

Nota: indipendenza in distribuzione → indipendenza in media (ma <u>non</u> vice-versa).

Riporto la tabella di partenza con l'indicazione per Y dei valori centrali delle classi:

X=Frequenza Y=voto all'esame	**19**	**22**	**24.5**	**27**	**29.5**	**Totale**
Si	11	23	38	116	75	263
No	55	163	107	48	64	437
Totale	66	186	145	164	139	**700**

Calcolo le medie condizionate di Y|X:

La media condizionata di Y dato che X=Si è pari a:

$$\mu_{y|X=Si} = \frac{1}{n_{SI}} \sum_{i=1}^{n_{SI}} c_i * n_{1.} = \frac{1}{263}(19 * 11 + 22 * 23 + 24.5 * 38 + 27 * 116 + 29.5 * 75) = 26.58$$

La media condizionata di Y dato che X=No è pari a:

$$\mu_{y|X=No} = \frac{1}{n_{NO}} \sum_{i=1}^{n_{NO}} c_i * n_{2.} = \frac{1}{437}(19 * 55 + 22 * 163 + 24.5 * 107 + 27 * 48 + 29.5 * 64) = 23.882$$

La media generale del voto all'esame (Y) è pari a:

$$\mu = \frac{1}{n} \sum_{j=1}^{n} c_i * n_{.j} = \frac{1}{700}(19 * 66 + 22 * 186 + 24.5 * 145 + 27 * 164 + 29.5 * 139) = 24.896$$

Le medie di Y condizionate alle modalità di X non sono costanti e sono diverse dalla media generale. Tra i due caratteri non esiste indipendenza in media.

Per valutare quanta parte della variabilità totale dei voti è attribuibile alla dipendenza del voto medio dalla frequenza alle lezioni, consideriamo la proprietà di scomposizione della devianza:

$$Dev_{TOT}(Y) = Dev_{INT}(Y) + Dev_{EST}(Y)$$

dove:

$Dev_{EST}(Y) = \sum_i (\mu_{Y|X=x_i} - \mu)^2 * n_{i.}$ variabilità delle medie di gruppo rispetto alla media generale

$Dev_{INT}(Y) = \sum_i (Y_i - \mu_{Y|X=x_i})^2 * n_{i.}$ variabilità dei valori di Y rispetto alle medie condizionate di Y|X

$Dev_{TOT}(Y) = \sum_i \sum_j (Y_i - \mu)^2 * n_{.j}$ variabilità totale di Y

e calcoliamo il rapporto di correlazione tra Y e X noto come indice η^2 di Pearson:

$$\eta^2_{Y|X} = \frac{Dev_{EST}(Y)}{Dev_{TOT}(Y)} = \frac{\sum_{i=1}^{h} (\mu_{Y|X=x_i} - \mu)^2 \, n_{i.}}{\sum_i \sum_j (Y_i - \mu)^2 \, n_{.j}}$$

E' un indice normalizzato che varia tra 0 (massima indipendenza in media) a 1 (massima dipendenza in media)

L'indice descrive quanta parte della devianza totale è spiegata dalla variabilità delle medie parziali rispetto alla media generale. Occorre dunque calcolare la devianza esterna ai gruppi:

$$Dev_{EST}(Y) = \sum_j (\mu_{Y|X=x_i} - \mu)^2 * n_{i.}$$

$$= (26.58 - 24.896)^2 * 263 + (23.882 - 24.896)^2 * 437 = 1194.883$$

La devianza totale del carattere voto calcolata sulla distribuzione marginale è pari a:

$$Dev_{TOT}(Y) = \sum_i \sum_j (c_i - \mu)^2 * n_{.j} =$$

$$= (19 - 24.896)^2 * 66 + (22 - 24.896)^2 * 186 + (24.5 - 24.896)^2 * 145 +$$

$$+ (27 - 24.896)^2 * 14 + (29.5 - 24.896)^2 * 139 = 7549.387$$

Quindi:

$$\eta^2_{Y|X} = \frac{1194.883}{7549.387} = 0.158$$

Il 15.8% della variabilità del carattere voto all'esame è spiegato dal suo dipendere in media dalla frequenza alle lezioni.

Nota: possiamo ottenere per differenza la devianza interna ai gruppi:

$$Dev_{INT}(Y) = Dev_{TOT}(Y) - Dev_{EST}(Y) = 7549.387 - 1194.883 = 6354.504$$

B. Richiami di calcolo delle probabilità e teorema di Bayes

Probabilità condizionata. Dati due eventi A e B $\in \Omega$, la probabilità che si verifichi l'evento B dato che si è verificato A è data da:

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

Da cui: $P(A \cap B) = P(B|A) * P(A)$ probabilità intersezione eventi dipendenti e compatibili

Teorema delle probabilità totali. Dato uno spazio campione Ω composto dagli eventi $E_1, E_2 \dots E_i$ allora per ogni evento A $\in \Omega$ si ha che:

$$P(A) = \sum_{i=1}^{\infty} P(A|E_i)P(E_i)$$

La probabilità di A è data dalla media ponderata delle probabilità condizionate con pesi dati da $P(E_i)$.

Partizione di Ω. Dati due eventi A e B essi definiscono una partizione dello spazio campionario se insieme sono incompatibili $A \cap B = \emptyset$ e necessari $A \cup B = \Omega$ e sono inoltre equiprobabili P(A)=P(B)

Esercizio 1 B. Applicazione del Teorema di Bayes

Il tasso di diffusione territoriale di una patologia cronica è pari all'1%. Il test di laboratorio impiegato per la rilevazione precoce dà esito positivo nel 99% dei soggetti affetti e nel 3% dei soggetti sani. Il test di un individuo preso a caso dai residenti nel territorio è positivo. Qual è la probabilità che l'individuo sia affetto da quella patologia?

Sol.

Indichiamo con E l'evento "l'individuo è affetto dalla patologia cronica" e con F l'evento " il test da esito positivo", per cui:

$P(E) = 0.01$

$P(F|E) = 0.99$

$P(F|\bar{E}) = 0.03$

La probabilità richiesta è $P(E|F)$ per cui si ricorre al teorema di Bayes:

$$P(E|F) = \frac{P(F|E)P(E)}{P(F|E)P(E) + P(F|\bar{E})P(\bar{E})} = \frac{0.99 \times 0.01}{0.99 \times 0.01 + 0.03 \times (1 - 0.01)} = 0.25$$