CORSO DI STATISTICA (parte 2) - ESERCITAZIONE 4

Dott.ssa Antonella Costanzo

a.costanzo@unicas.it

Esercizio 1. Valore atteso e varianze condizionate, indipendenza tra variabili casuali

E' stata svolta una indagine per studiare la relazione fra il peso dei neonati e l'abitudine al fumo in gravidanza da parte della madre. Per il peso dei neonati sono state individuate quattro classi: la prima ha valore centrale 2Kg, la seconda ha valore centrale 2.75Kg, la terza ha valore centrale 3.25Kg ed, infine, la quarta ha valore centrale 4Kg. Sia X una variabile casuale che assume valore 1 se la madre ha fumato durante la gravidanza e 0 in caso contrario e sia Y una variabile casuale che assume valori 2, 2.75, 3.25 e 4 a seconda della classe di appartenenza del peso del neonato. E' stata stimata la funzione di probabilità congiunta. I valori di tale funzione sono riportati nella seguente tabella:

X|Y	2	2.75	3.25	4
0	0.12	0.15	0.15	0.19
1	0.16	0.09	0.08	0.06

a) Calcolare le funzioni di probabilità marginali di X e Y .
b) Calcolare i valori attesi condizionati della Y , per X = 0 e X = 1, e confrontarli con il valore atteso marginale, commentando i risultati, calcolare le varianze condizionate della Y.
c) Valutare se le variabili casuali X e Y sono indipendenti.

Sol.

a) le distribuzioni marginali si possono ottenere direttamente semplicemente sommando sulle righe (distribuzione marginale della X) e sommando lungo le colonne (distribuzione marginale della Y) ottenendo cosi:

X|Y	2	2.75	3.25	4	Totale
0	0.12	0.15	0.15	0.19	**0.61**
1	0.16	0.09	0.08	0.06	**0.39**
Totale	**0.28**	**0.24**	**0.23**	**0.25**	1

b) Il valore atteso marginale della Y è dato da:

$$E(Y) = \sum_{j=1}^{k=4} y_j P(Y = y_j) = \sum_{j=1}^{k=4} y_j p_j = 0.28(2) + 0.24(2.75) + 0.23(3{:}25) + 0.25(4) = 2.9675$$

Per calcolare i valori attesi condizionati occorre innanzitutto determinare le due distribuzioni condizionate. Tali distribuzioni sono date da:

$$P(Y = y|X = 0) = 0.12/0.61 = 0.197\ se\ y = 2$$

$$0.15/0.61 = 0.246\ se\ y = 2.75$$

$$0.15/0.61 = 0.246\ se\ y = 3.25$$

$$0.19/0.61 = 0.311\ se\ y = 4$$

$$P(Y = y|X = 1) = 0.16/0.39 = 0.410\ se\ y = 2$$

$$0.09/0.39 = 0.231\ se\ y = 2.75$$

$$0.08/0.39 = 0.205\ se\ y = 3.25$$

$$0.06/0.39 = 0.154\ se\ y = 4$$

Il valori attesi condizionati della Y sono dati da:

$$E(Y|X = 0) = \sum_{j=1}^{k=4} y_j P\big(Y = y_j \big| X = 0\big) = 0.197 * 2 + 0.246 * 2.75 + 0.246 * 3.25 + 0.311 * 4 = 3.114$$

$$E(Y|X = 1) = \sum_{j=1}^{k=4} y_j P\big(Y = y_j \big| X = 1\big) = 0.410 * 2 + 0.231 * 2.75 + 0.205 * 3.25 + 0.154 * 4 = 2.7375$$

Confronto il valore atteso delle distribuzioni condizionate di Y ad X con il valore atteso marginale E(Y):

$E\big(E(Y|X)\big) = P(X = 0)E(Y|X = 0) + P(X = 1)E(Y|X = 1) = 0.61 * 3.114 + 0.39 * 2.7375 = 2.9675 = E(Y)$ *(proprietà delle medie reiterate)*

Le varianze condizionate risultano le seguenti:

$$Var(Y|X = 0) = \sum_{j=1}^{4} \Big(y_j - E(Y|X = 0)\Big)^2 P(Y = y_j|X = 0)$$

$$= 0.197(2 - 3.114)^2 + 0.246(2.75 - 3.114)^2 + 0.246(3.25 - 3.114)^2 +$$

$$+0.311(4 - 3.114)^2 = 0.53$$

$$Var(Y|X=1) = \sum_{j=1}^{4} \Big(y_j - E(Y|X=1)\Big)^2 P(Y=y_j|X=1)$$

$$= 0.410(2-2.7375)^2 + 0.231(2.75-2.7375)^2 + 0.205(3.25-2.7375)^2 +$$

$$+0.154(4-2.7375)^2 = 0.52$$

Nota: in analogia al caso non condizionato, la varianza condizionata può essere espressa come la seguente differenza:

$$Var(Y|Y=x) = E[Y^2|X=x] - [E(Y|X=x)]^2$$

Inoltre, come visto nella prima parte del corso, per la proprietà di scomposizione della devianza:

$$Var(Y) = Var_W + Var_B$$

dove:

$$Var(Y) = E(Y^2) - [E(Y)]^2 = 0.5583$$

$$Var_W = \sum_{i=1}^{2} Var(Y|X=x_i) * P(X=x) = 0.53 * 0.61 + 0.52 * 0.39 = 0.5261$$

$$Var_B = \sum_{i=1}^{2} [E(Y|X=x_i) - E(Y)]^2 P(X=x) = (3.114-2.9675)^2 * 0.61 + (2.7375-2.9675)^2 * 0.39$$

$$= 0.033$$

c) due variabili X e Y sono indipendenti se la loro distribuzione congiunta può essere espressa come prodotto delle distribuzioni marginali. Tale ipotesi nel caso proposto può essere espressa come:

$$P(Y=y, X=x) = P(X=x) * P(Y=y) \qquad x = 0,1 \quad y = 2, 2.75, 3.25, 4$$

La distribuzione teorica sotto l'ipotesi di indipendenza è pari a:

X | Y	2	2.75	3.25	4	totale
0	0.171	0.146	0.140	0.153	0.61
1	0.109	0.094	0.09	0.097	0.39
totale	0.28	0.24	0.23	0.25	1

si può, quindi, concludere che le variabili non sono indipendenti.

Inoltre in caso di indipendenza tra X e Y dovrebbe verificarsi che:

$P(Y=y_j|X=0)$ e $P(Y=y_j|X=1)$ sono uguali alla distribuzione marginale P(Y).

Nel nostro caso invece:

$P(Y = y_j|X = 0)$ e $P(Y = y_j|X = 1)$ non sono uguali a P(Y) perciò X e Y non sono indipendenti

Esercizio 2. Il problema del collezionista

Un ragazzo deve riempire un album di 10 figurine. Egli acquista le figurine in busta chiusa; ciascuna busta contiene una sola figurina, e si suppone che le figurine contenute nelle buste siano del tutto casuali ed indipendenti l'una dall'altra. Ovviamente la prima busta che acquista contiene una figurina che egli metterà sicuramente nell'album. Sia X il numero di buste che deve acquistare, dopo la prima, per trovare la prima figurina diversa da quella già inserita nell'album; sia poi Y il numero di buste che deve acquistare successivamente per trovare la prima figurina diversa dalle prime due già inserite nell'album.

a) Calcolare la funzione di probabilità della variabile X. E' una distribuzione nota?
b) Calcolare la densità della variabile Y. Quanto vale P(X+Y=3)?
c) Calcolare il numero totale di buste che il ragazzo deve acquistare per riempire l'album.

Sol.

a)

La variabile X conta il numero di prove necessarie per ottenere un successo in un esperimento con probabilità p=9/10 di successo. Si tratta quindi di una v.c. geometrica di parametro 9/10. Si ha quindi:

$$P(X = k) = p(1-p)^{k-1} \qquad \text{k=1,2,…} \qquad \text{quindi:} \quad P(X = k) = (\tfrac{1}{10})^{k-1} * \tfrac{9}{10}$$

b)

In modo analogo si vede che la v.c. Y è indipendente da X e anch'essa con distribuzione geometrica di parametro 8/10 (infatti questa volta 8 figurine su 10 danno un risultato favorevole in quanti sono differenti dalle prime due).

La v.c. X+Y prende valori 2,3… e si ha per $n \geq 2$

$$P(X+Y=n) = \sum_{k=1}^{n-1} P(X=k)P(Y=n-k) = \sum_{k=1}^{n-1} \left[\frac{9}{10}\left(\frac{1}{10}\right)^{k-1}\right]\left[\frac{8}{10}\left(\frac{2}{10}\right)^{n-k-1}\right]$$

In particolare:

$$P(X+Y=3) = P(X=1)P(Y=2) + P(X=2)P(Y=1)$$

$$P(X+Y=3) = \frac{9}{10}\frac{1}{10}\frac{2}{10} + \frac{9}{10}\frac{1}{10}\frac{8}{10} = 0.09$$

c)

La v.c. X ha valore atteso $E(X) = \frac{1}{p} = \frac{10}{9}$, la v.c. Y ha valore atteso $E(Y) = \frac{1}{p} = \frac{10}{8}$ e così via. Indichiamo con T la variabile che conta il numero totale di buste acquistate: essa è pari alla costante 1 sommata a X, a Y ecc. e pertanto il suo valore atteso è:

$$E(T) = 1 + E(X) + E(Y) + \cdots$$

$$E(T) = 1 + \frac{10}{9} + \frac{10}{8} + \frac{10}{7} + \frac{10}{6} + \frac{10}{5} + \frac{10}{4} + \frac{10}{3} + \frac{10}{2} + \frac{10}{1} = 29.27$$

Esercizio 3. Approssimazione normale di una binomiale (TLC)

Una prova d'esame consiste di 10 quesiti a risposta multipla. Ogni quesito prevede 4 possibili risposte, una sola delle quali è corretta. Per superare l'esame è necessario rispondere correttamente ad 8 quesiti. Uno studente, analizzato il testo della prova, è assolutamente certo della risposta da dare a 5 dei 10 quesiti.

A) Lo studente decide di consegnare la prova d'esame tirando ad indovinare sui restanti 5 quesiti; che probabilità ha di superare l'esame?
B) Come cambia la risposta se i quesiti sono 100 e lo studente è convinto di poter rispondere correttamente a 50 di essi e se la soglia per superare l'esame è di 80 risposte corrette?

Sol.

A) Il numero delle risposte corrette nei 5 quesiti per i quali la risposta è data a caso si distribuisce secondo una legge binomiale di parametri n=5 e p=0.25 (essendo 4 le risposte, ciascuna ha la stessa probabilità di essere corretta e i risultati sono indipendenti). Condizionatamente al fatto che lo studente dia la risposta corretta ai 5 quesiti per i quali ritiene di conoscerla, egli supera la prova se risponde ad almeno altri tre, ossia se $X \geq 3$. La probabilità di superare l'esame è quindi:

$$X \sim Bin(n = 5, p = 0.25)$$

$$P(X \geq 3) = P(X = 3) + P(X = 4) + P(X = 5)$$

$$P(X \geq 3) = \binom{5}{3} 0.25^3 0.75^2 + \binom{5}{4} 0.25^4 0.75^1 + \binom{5}{5} 0.25^5 = 0.1035$$

B) Indichiamo con S la variabile "numero di risposte corrette sulle 50 domande per le quali lo studente è incerto". Essa è distribuita secondo una legge binomiale di parametri n=50 e p=0.25. La probabilità di superare l'esame è

$$P(S \geq 30) = \sum_{i=30}^{50} P(S = i)$$

Possiamo tuttavia adottare l'approssimazione normale e per il Teorema del Limite Centrale (TLC) ragionare assumendo S distribuito secondo una legge normale con media np=50*0.25=12.5 e varianza np(1-p)=50(0.25)(1-0.25)=9.375. Quindi:

$$P(S \geq 30) = 1 - P(S \leq 30) = 1 - \varphi\left(\frac{30 - 12.5}{\sqrt{9.375}}\right) = 1 - \varphi(5.7155) = 5.47 * 10^{-9} \approx 0$$

Nota: quest'approssimazione viene solitamente accettata quando np>5 e np(1-p)>5

Esercizio 3bis. Approssimazione normale di una binomiale

L'ufficio delle risorse umane di una grande azienda seleziona ogni anno 1000 CV di neolaureati per uno stage. Sapendo che nell'anno precedente sul totale di neolaureati il 32% risultano donne, qual è la probabilità che la proporzione di donne nel campione selezionato dall'azienda sia tra 280 e 300?

Sol.

Popolazione: tutti i neolaureati

Campione: n=1000 CV

Variabile di interesse: numero di donne che hanno presentato il CV all'azienda

Si tratta di distribuzione binomiale dove l'evento "successo"= essere donna, quindi il nostro obiettivo è calcolare la probabilità che il numero di successi sia compreso tra 280 e 300 considerando le 1000 prove consecutive e tenendo conto che la probabilità di successo è pari al 32%.

Grazie al teorema del limite centrale sappiamo che per un numero elevato di osservazioni, quindi dato un n molto grande, la distribuzione binomiale tende a una normale, cioè può essere considerata come una normale.

E(X) = np

Var(X) = np (1-p)

La nostra X, che abbiamo detto si distribuisce come una normale, ha i seguenti parametri:

X ~ N (np, np(1-p)) cioè

X ~ N (1000 * 0.32 ; 1000 * 0.32(1-0.32))

X ~ N (320; 217.6)

Calcoliamo la probabilità che X sia compreso tra 280 e 300.

Standardizzazione:

$$Z_1 = \frac{X_1 - \mu}{\sigma} = \frac{280 - 320}{14.75} = -2.71$$

$$Z_2 = \frac{X_2 - \mu}{\sigma} = \frac{300 - 320}{14.75} = -1.36$$

A questo punto quello che dobbiamo calcolare è P (- 2.71 < Z < - 1.36) ma sapendo che sulle tavole

troviamo solo valori positivi, questo equivale a calcolare P (1.36 < Z < 2.71)

P (Z < 2.71) = 0.9966

P (Z < 1.36) = 0,9131 quindi: P (1.36 < Z < 2.71) = 0.9966 – 0.9131 = 0.084

Esercizio 4. Media campionaria e sua distribuzione

La durata delle telefonate urbane segue una distribuzione normale di media μ = 10 minuti e scarto quadratico medio σ = 3 minuti. Selezionato un campione casuale di 20 telefonate, trovare la distribuzione della media campionaria e la probabilità che la durata media delle telefonate sia compresa fra 9.5 e 10.3 minuti.

Sol.

Siccome le osservazioni X_i seguono una legge normale, anche la media campionaria $\bar{X}$ seguirà la stessa legge di probabilità. La media campionaria avrà il valore atteso pari a μ, e la varianza pari a $\frac{\sigma^2}{n}$ (risultato del TLC).

Di conseguenza se X= durata delle telefonate urbane è tale che: X ~ N(10, 9), allora

$$\bar{\bar{X}} \sim N(10, 0.45)$$

Sia $Z = \frac{X_i - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$

Allora:

$$P(9.5 \leq \bar{X} \leq 10.3) = P\left(\sqrt{20}\frac{9.5 - 10}{3} \leq Z \leq \sqrt{20}\frac{10.3 - 10}{3}\right) =$$

$$= P(-0.74536 \leq Z \leq 0.44721) = P(Z \leq 0.44721) - P(Z \leq -0.74536) =$$

$$= P(Z \leq 0.44721) - [1 - P(Z \leq 0.74356)] = 0.67634 - [1 - 0.7704] = 0.4467$$