Corso di Statistica - Esercitazione 3

Dott. Davide Buttarazzi

✉ *d.buttarazzi@unicas.it*

Esercizio 1

Si vuole studiare l'interdipendenza tra le ore di studio autonomo (carattere X: Poche, Abbastanza, Molte) ed il livello di attenzione in aula (carattere Y: Basso, Medio, Alto) degli studenti del corso di Economia Politica. I dati sono riportati nella seguente tabella, espressi in frequenze assolute congiunte.

		Y= Livello di attenzione			
		Basso	**Medio**	**Alto**	$\mathbf{n_{i.}}$
X= Ore di studio autonomo	**Poche**	8	12	60	80
	Abbastanza	4	54	12	70
	Molte	40	6	4	50
	$\mathbf{n_{.j}}$	52	72	76	$\mathbf{n_{..} = 200}$

1. Calcolare l'indice di connessione χ^2 e definire il suo campo di variazione
2. Calcolare l'indice di contingenza media Φ^2
3. Calcolare l'indice di connessione normalizzato V_{Cramer}

Soluzioni esercizio 1

1. Indicando con n_{ij} le frequenze assolute congiunte osservate, con $\hat{n}_{ij}$ le frequenze assolute congiunte teoriche, con r il numero di righe e con c il numero di colonne, l'indice χ^2 è così definito:

   $\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}-\hat{n}_{ij})^2}{\hat{n}_{ij}}$, con $0 \leq \chi^2 \leq max$

   Per poter calcolare tale indice, è possibile costruire passo passo una serie di tabelle con le quantità di interesse. Innanzitutto, occorre calcolare la tabella delle frequenze teoriche (quelle che si osserverebbero in caso di assenza di connessione tra i due caratteri).

   In particolare, $\hat{n}_{ij} = \frac{n_{i.} \times n_{.j}}{n_{..}}$, si avrà quindi:

Step 1: Tabella frequenze teoriche $\hat{n}_{ij}$

		Y= Livello di attenzione			
		Basso	**Medio**	**Alto**	$\mathbf{n_{i.}}$
X= Ore di studio autonomo	**Poche**	$\frac{80\times52}{200} = 20.8$	$\frac{80\times72}{200} = 28.8$	$\frac{80\times76}{200} = 30.4$	80
	Abbastanza	$\frac{70\times52}{200} = 18.2$	$\frac{70\times72}{200} = 25.2$	$\frac{70\times76}{200} = 26.6$	70
	Molte	$\frac{50\times52}{200} = 13$	$\frac{50\times72}{200} = 18$	$\frac{50\times76}{200} = 19$	50
	$\mathbf{n_{.j}}$	52	72	76	$\mathbf{n_{..} = 200}$

Occorre quindi calcolare la differenza tra le frequenze assolute congiunte osservate e le frequenze teoriche appena calcolare.

Le differenze semplici appena calcolare vanno ora trasformate in differenze quadratiche. Si avrà quindi:

Infine, occorre dividere le quantità appena calcolate per le frequenze assolute congiunte teoriche calcolate nello Step 1:

Step 2: Tabella differenze tra frequenze osservate e frequenze teoriche $(n_{ij} - \hat{n}_{ij})$

		Y= Livello di attenzione			
		Basso	**Medio**	**Alto**	**Tot.**
X= Ore di studio autonomo	**Poche**	8-20.8=-12.8	12-28.8=-16.8	60-30.4=29.6	0
	Abbastanza	4-18.2=-14.2	54-25.2=28.8	12-26.6=-14.6	0
	Molte	40-13=27	6-18=-12	4-19=-15	0
	Tot.	0	0	0	0

Step 3: Tabella dei quadrati delle differenze tra frequenze osservate e frequenze teoriche $(n_{ij} - \hat{n}_{ij})^2$

		Y= Livello di attenzione		
		Basso	**Medio**	**Alto**
X= Ore di studio autonomo	**Poche**	$-12.8^2 = 163.84$	$-16.8^2 = 282.24$	$29.6^2 = 876.16$
	Abbastanza	$-14.2^2 = 201.64$	$28.8^2 = 829.44$	$-14.6^2 = 213.16$
	Molte	$27^2 = 729$	$-12^2 = 144$	$-15^2 = 255$

Step 4: Tabella differenze tra frequenze osservate e frequenze teoriche $\frac{(n_{ij}-\hat{n}_{ij})^2}{\hat{n}_{ij}}$

		Y= Livello di attenzione			
		Basso	**Medio**	**Alto**	**Tot.**
X= Ore di studio autonomo	**Poche**	$\frac{163.84}{20.8} = 7.87$	$\frac{282.24}{28.8} = 9.8$	$\frac{876.16}{30.4} = 28.82$	
	Abbastanza	$\frac{201.64}{18.2} = 11.08$	$\frac{829.44}{25.2} = 32.9$	$\frac{213.16}{26.6} = 8.01$	
	Molte	$\frac{729}{13} = 56.07$	$\frac{144}{18} = 8$	$\frac{255}{19} = 13.42$	
	Tot.				$\chi^2 = 175.97$

Sommando tutti gli elementi della tabella appena costruita si ottiene l'indice $\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij}-\hat{n}_{ij})^2}{\hat{n}_{ij}} = 175.97$,

$0 \leq \chi^2 \leq max$, con $max = n_{..}[min(r,c) - 1] = 200(3-1) = 400$

Tale indice dipende quindi sia da $n_{..}$ (numero totale di osservazioni) che da $[min(r,c)-1]$ (struttura della tabella di contingenza).

2. L'indice di contingenza media Φ^2 può essere utilizzato per rimuovere l'effetto di $n_{..}$

   Si avrà quindi $\Phi^2 = \frac{\chi^2}{n_{..}} = \frac{175.97}{200} = 0.88$

3. L'indice di connessione normalizzato V_{Cramer} invece, rimuove anche l'effetto della struttura.

   $V_{Cramer} = \sqrt{\frac{\Phi^2}{min(r,c)-1}} = \sqrt{0.44} = 0.66$, con $0 \leq V_{Cramer} \leq 1$.

   Si osserva pertanto una sostanziale interdipendenza (o connessione) tra i due caratteri X= Ore di studio autonomo e Y=Livello di attenzione in aula.

Esercizio 2

Si vuole studiare se l'apporto calorico (Y=Calorie, quantitativa) dei panini serviti in un noto ristorante dipenda dalla tipologia di carne (Burger o Chicken) da essi contenuta (X=Tipo, qualitativa).

		Y= Calorie				
		[400,500]	**(500-600]**	**(600-700]**	**(700-800]**	$\mathbf{n_{i.}}$
X= Tipo	**Burger**	1	3	3	0	**7**
	Chicken	7	2	1	1	**11**
	$\mathbf{n_{.j}}$	**8**	**5**	**4**	**1**	$\mathbf{n_{..} = 18}$

1. Calcolare l'indice $\eta^2_{Y|X}$

Soluzioni esercizio 2

1. Date le variabili X (qualitativa) ed Y (quantitativa), l'indice $\eta^2_{Y|X}$ può essere così definito:

   $\eta^2_{Y|X} = \frac{\sigma^2_{EST}}{\sigma^2} = \frac{DEV_{EST}}{DEV(Y)}$, con $0 \leq \eta^2_{Y|X} \leq 1$

   Occorre quindi calcolare le quantità DEV_{EST} e $DEV(Y)$ per la tabella di frequenze assolute congiunte data.

   Innanzitutto si noti che la variabile Y è organizzata in classi. Occorre quindi operare con i centri di tali classi, rispettivamente uguali a 450, 550, 650, 750.

   La devianza esterna $DEV_{EST} = \sum_{i=1}^{G} (\bar{y_i} - \mu_y)^2 \; n_i$

   Il primo passo è quindi quello di individuare le quantità $\bar{y_i}$ (medie di gruppo, ovvero medie condizionate) e μ_y (media marginale). Si ha quindi che:

   $\bar{y}|(X = Burger) = \sum_{j=1}^{c} \frac{y_j n_{1j}}{n_{1.}} = \frac{450\times1+550\times3+650\times3+750\times0}{7} = 578.6$

   $\bar{y}|(X = Chicken) = \sum_{j=1}^{c} \frac{y_j n_{2j}}{n_{2.}} = \frac{450\times7+550\times2+650\times1+750\times1}{11} = 513.6$

   $\bar{\mu_y} = \sum_{j=1}^{c} \frac{y_j n_{.j}}{n_{..}} = \frac{450\times8+550\times5+650\times4+750\times1}{18} = 538.9$

   È possibile ora calcolare le due componenti della devianza esterna:

   $DEV_{Y|X=Burger} = [(\bar{y}|X = Burger) - \mu_y]^2 \; n_{1.} = (578.6 - 538.9)^2 \times 7 = 11032.63$

   $DEV_{Y|X=Chicken} = [(\bar{y}|X = Chicken) - \mu_y]^2 \; n_{2.} = (513.6 - 538.9)^2 \times 11 = 7040.99$

   Sommando queste componenti si ottiene la varianza esterna:

   $DEV_{EST} = \sum_{i=1}^{G} (\bar{y_i} - \mu_y)^2 \; n_i = 11032.63 + 7040.99 = 18073.62$

   Resta da calcolare quindi la devianza di Y, ovvero:

   $DEV(Y) = \sum_{j=1}^{c} (y_j - \mu_y)^2 \times n_{j.} = [(450-538.9)^2 \times 8] + [(550-538.9)^2 \times 5] + [(650-538.9)^2 \times 4] + [(750-538.9)^2 \times 1] = 157777.8$

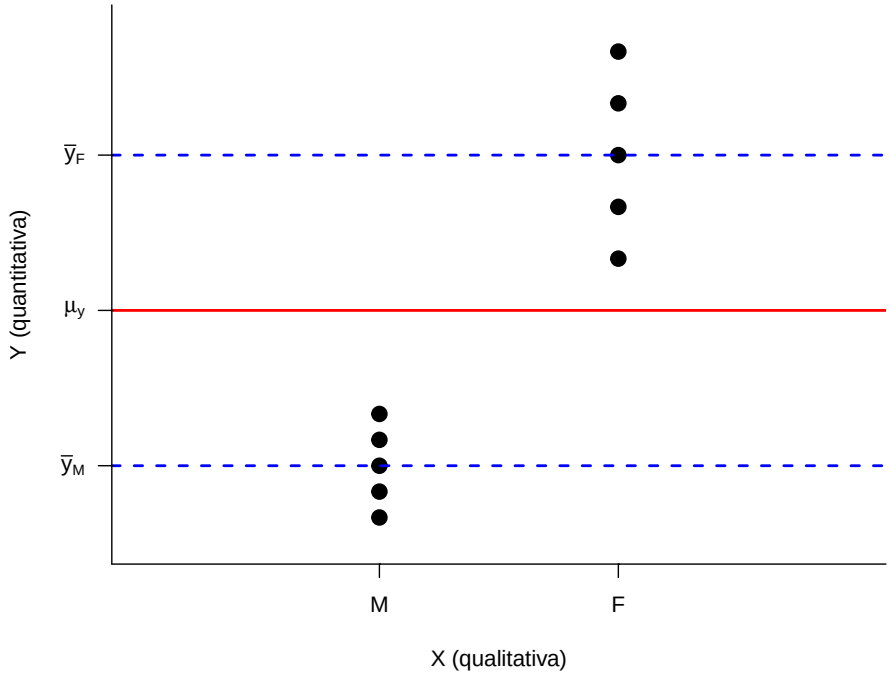
   Non resta quindi che calcolare l'indice

   $\eta^2_{Y|X} = \frac{DEV_{EST}}{DEV(Y)} = \frac{18073.62}{157777.8} = 0.11$, il quale segnala assenza di dipendenza in media della variabile Y=Calorie dalla variabile X=Tipo.

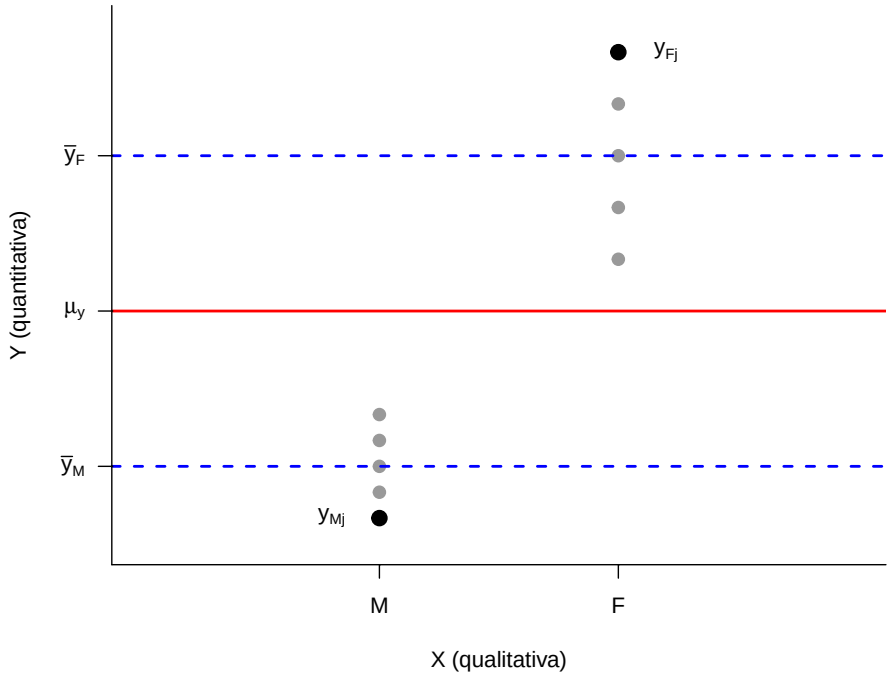
Nota per assenti o non frequentanti: durante l'esercitazione 3 sono stati discussi i concetti di variabilità totale, variabilità interna e variabilità esterna (ovvero la relazione $\sigma^2_{TOT} = \sigma^2_{INT} + \sigma^2_{EST}$). In appendice i grafici discussi a lezione.

Appendice: variabilità totale, variabilità interna e variabilità esterna

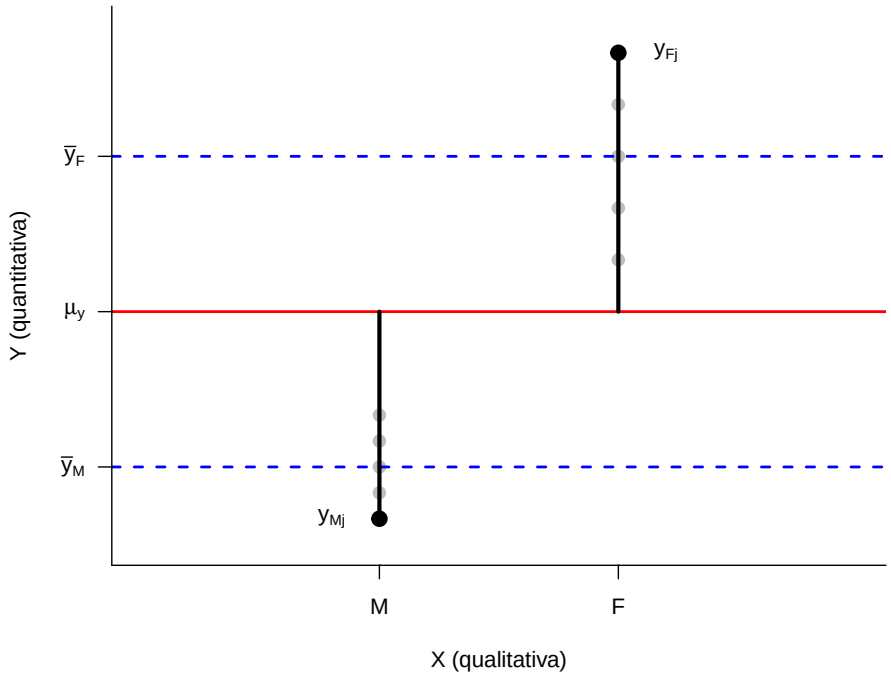
Si consideri la variabile Y quantitativa (Altezza)
rispetto alla variabile X qualitativa (Genere: M,F)



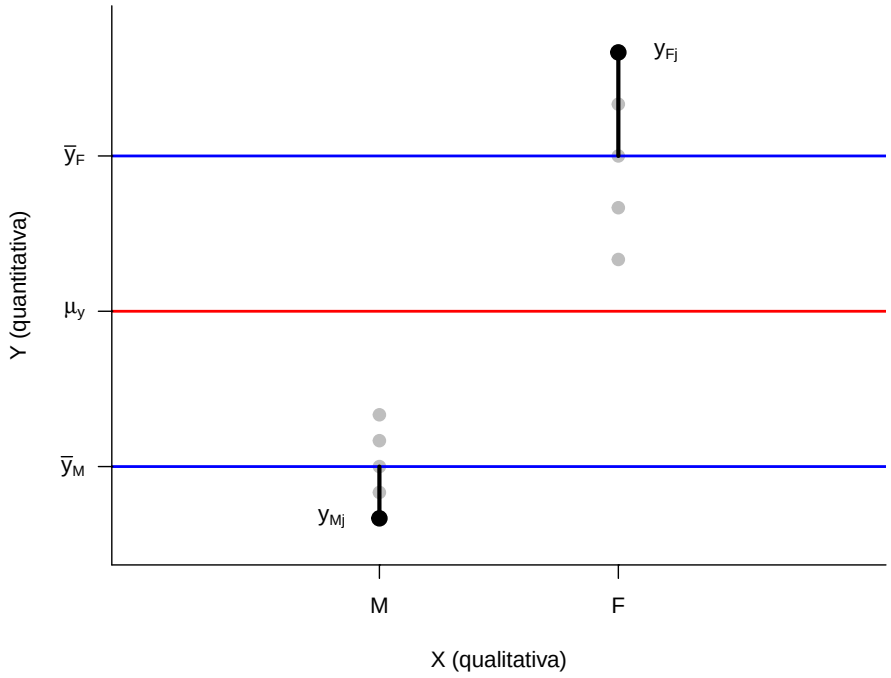
Si considerino le generiche osservazioni y_{Mj} ed y_{Fj}



Variabilità totale



Variabilità interna



Variabilità esterna