Esercitazione 4 del corso di Statistica (parte 1)

Dott.ssa Paola Costantini

5 Febbraio 2009

La seguente tabella riporta le informazioni relative a 25 laureati nell'anno 2005 in Economia, ad un anno dal conseguimento del titolo.

Genere	Eta	Voto*	Durata	Tipo Contratto	Utilizzo della Laurea	Efficacia della Laurea
Uomini	47,8	83	4,4	Stabile	In misura elevata	Efficace
Uomini	26,6	113	7,4	Stabile	In misura ridotta	Poco efficace
Uomini	31,5	91	12,4	Atipico	In misura ridotta	Abb. efficace
Uomini	23,6	102	4,4	Inserimento/Form	In misura elevata	Efficace
Uomini	25,9	94	6,4	Stabile	Per niente	Per nulla efficace
Donne	23,6	108	4,7	Inserimento/Form	In misura ridotta	Abb. efficace
Donne	28,6	108	5,7	Atipico	In misura ridotta	Abb. efficace
Uomini	42,1	100	3	Stabile	Per niente	Efficace
Donne	24,3	113	3,4	Atipico	Per niente	Per nulla efficace
Donne	26,3	113	3,4	Atipico	Per niente	Per nulla efficace
Uomini	24,9	106	3,7	Inserimento/Form	In misura elevata	Molto efficace
Donne	24	95	4,5	Stabile	In misura elevata	Efficace
Uomini	34,7	92	4,4	Stabile	In misura elevata	Efficace
Donne	24,7	106	5,14	Atipico	In misura ridotta	Abb. efficace
Uomini	25,9	100	5,94	Atipico	In misura ridotta	Abb. efficace
Uomini	25,4	92	6,14	Atipico	In misura elevata	Molto efficace
Donne	27,6	113	4,14	Senza contratto	Per niente	Poco efficace
Donne	23,4	113	4,4	Inserimento/Form	In misura elevata	Molto efficace
Uomini	31,3	105	3,4	Stabile	In misura elevata	Molto efficace
Uomini	29,7	110	4,14	Atipico	In misura ridotta	Abb. efficace
Uomini	27	93	7,3	Stabile	In misura elevata	Efficace
Uomini	35,6	97	15,4	Atipico	Per niente	Poco efficace
Uomini	23,1	101	3,7	Senza contratto	Per niente	Poco efficace
Uomini	25,3	91	6,5	Atipico	Per niente	Per nulla efficace
Uomini	32,6	92	13,5	Stabile	In misura ridotta	Abb. efficace

* Il 113 indica il 110 con Lode.

Esercizio 1 A partire dalla distribuzione di frequenza del carattere Durata del corso di studi, determinare la differenza interquartile e rappresentare il boxplot.

Dati ordinati	n_i	fi	Ni	Fi
3	1	0,077	1	0,077
3,4	2	0,154	3	0,231
3,7	1	0,077	4	0,308
4,4	3	0,23	7	0,538
4,5	1	0,077	8	0,615
4,7	1	0,077	9	0,692
5,7	1	0,077	10	0,769
6,4	1	0,077	11	0,846
7,4	1	0,077	12	0,923
12,4	1	0,077	13	1
totale	13	1		

Soluzione

Me = 4,4 Q_1=3,7 Q_2=5,7

Si definisce differenza interquartile la differenza tra il terzo e il primo quartile.

IQR= Q_3 - Q_1 = 5,7-3,7 =2

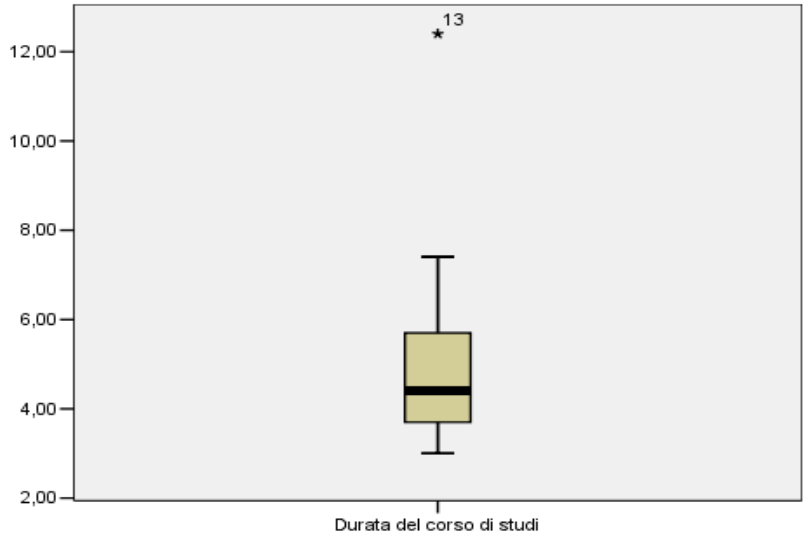
La Mediana, il primo e il terzo Quartine ci consentono di disegnare la scatola, mentre per disegnare i baffi dobbiamo prima calcolarci i Limiti Superiori ed Inferiori del box plot.

$L_I = Q_1 - 1{,}5*(Q_3 - Q_1) = 3{,}7 - 1{,}5*2 = 0{,}7$

Il baffo di sinistra (o inferiore) sarà il valore più grande tra L_I e X_{min}. Nel nostro caso L_I=0,7 mentre Xmin=3. Il più grande è X_{min}, quindi il primo baffo coinciderà con il valore di X_{min}.

$L_s = Q_3 + 1{,}5*(Q_3 - Q_1) = 5{,}7 + 1{,}5*2 = 8{,}7$

Il baffo di destra (o superiore) sarà il valore più piccolo tra L_S e X_{max}. Nel nostro caso L_S =8,7 mentre X_{max}=12,4. Il più piccolo è L_S, quindi il secondo baffo coinciderà con il valore di L_S.



Esercizio 2

Calcolare l'indice di Asimmetria di Fisher per il carattere Durata (prime 13 osservazioni) suddiviso in 3 classi equifrequenti.

x_i	n_i	f_i	N_i	F_i	x_i^c	a_i	di
C_1 = [3; 3,7]	4	0,31	4	0,31	3,35	0,7	0,44
C_2 =] 3,7; 4,5]	4	0,31	8	0,62	4,1	0,8	0,38
C_3 =] 4,5; 12,4]	5	0,38	13	1	8,45	7,9	0,048
Totali	**13**	**1,00**					

Soluzione

L'indice di Fisher, è un indice di forma basato sui momenti terzi standardizzati:

$$\gamma = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^3 n_i$$

$$\gamma > 0 \rightarrow AsimmetricaPositiva;$$
$$\gamma = 0 \rightarrow Simmetrica;$$
$$\gamma < 0 \rightarrow AsimmetricaNegativa;$$

Partendo dalla distribuzione in classi del carattere Durata,

x_i	n_i	x_i^c
C_1 = [3; 3,7]	4	3,35
C_2 =] 3,7; 4,5]	4	4,1
C_3 =] 4,5; 12,4]	5	8,45
Totali	**13**	

Calcoliamo dapprima la media aritmetica

$$\mu = \frac{1}{N}\sum_{i=1}^{C} x_i^c * n_i \quad = 5{,}54$$

Poi lo scarto quadratico medio

$$\sigma = \sqrt[2]{\frac{1}{N}\sum_{i=1}^{C}(x_i^c - \mu)^2 * n_i} = \sqrt[2]{\sigma^2} = \quad \sqrt{5{,}74} = 2{,}4$$

A questo punto abbiamo tutti gli elementi utili per calcolare l'indice di asimmetria di Fisher.

x_i	n_i	x_i^c	$x_i^c - \bar{x}$	$Z_i = \frac{(x_i^c - \bar{x})}{\sigma}$	$Z_i = \left(\frac{(x_i^c - \bar{x})}{\sigma}\right)^3$	$(Z_i)^3 \cdot n_i$
C_1 = [3; 3,7]	4	3,35	-2,19	-0,91	-0,76	-3,04
C_2 =] 3,7; 4,5]	4	4,1	-1,44	-0,6	-0,216	-0,864
C_3 =] 4,5; 12,4]	5	8,45	2,91	1,21	1,78	8,9
Totali	**13**					**4,996**

$$\gamma = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma}\right)^3 \cdot n_i = \frac{4{,}996}{13} = 0{,}38$$

Possiamo concludere che la distribuzione è caratterizzata da un'asimmetria positiva (indice maggiore di zero).
Tale risultato è confermato dal confronto tra la mediana e la media aritmetica.

$$\bar{x} = 5{,}54 > Me = 4{,}4$$

Esercizio 3
Calcolare l'indice di Curtosi di Pearson per il carattere Durata (prime 13 osservazioni).

Soluzione
La Curtosi riguarda un maggiore o minore appiattimento della forma della distribuzione.
L'indice γ_c è un indice di forma basato sui momenti quarti standardizzati.

$$\gamma_C = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^4 - 3$$

$$\gamma_C > 0 \rightarrow Leptocurtica;$$
$$\gamma_C = 0 \rightarrow Normocurtica;$$
$$\gamma_C < 0 \rightarrow Platicurtica;$$

Calcoliamo media e scarto quadratico medio a partire dai dati grezzi (prime 13 osservazioni).

$\bar{x} = 5{,}2$
$\sigma = 2{,}4$

Dati ordinati	**n_i**	$x_i - \bar{x}$	$Z_i = \frac{(x_i - \bar{x})}{\sigma}$	$Z_i = \left(\frac{(x_i^c - \bar{x})}{\sigma}\right)^4$	$(Z_i)^4 \cdot n_i$
3	1	-2,2	-0,91	0,685	0,685
3,4	2	-1,8	-0,75	0,316	0,632
3,7	1	-1,5	-0,625	0,152	0,152
4,4	3	-0,8	-0,33	0,012	0,036
4,5	1	-0,7	-0,29	0,007	0,007
4,7	1	-0,5	-0,21	0,002	0,002
5,7	1	0,5	0,21	0,002	0,002
6,4	1	1,2	0,5	0,062	0,062
7,4	1	2,2	0,91	0,685	0,685
12,4	1	7,2	3	81	81
totale	13				83,263

$$\gamma_c = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma}\right)^4 - 3 = \left(\frac{83{,}263}{13}\right) - 3 = 3{,}4$$

La distribuzione è Leptocurtica.

Esercizio 4

Si calcoli l'indice di Eterogeneità di Gini per il carattere *Efficacia della Laurea*.

Nel caso di variabili qualitative la variabilità del carattere è espressa in termini di mutabilità, definita come l'attitudine di un carattere ad assumere differenti modalità qualitative. Quando tutte le unità statistiche assumono la stessa modalità, si ha una perfetta omogeneità. (minima eterogeneità) Quando le modalità del carattere hanno tutte

la stessa frequenza assoluta o relativa, si ha la massima disomogeneità. L'Eterogeneità misura la variabilità delle frequenze delle k modalità del carattere.

Soluzione

L'Indice di Eterogeneità (G) di Gini si basa sulle frequenze relative.

Efficacia della Laurea	ni	fi
Per nulla eff.	4	0,16
Poco eff.	4	0,16
Abbastanza eff.	7	0,28
Efficace	6	0,24
Molto efficace	4	0,16
Totale	25	1

$$G = 1 - \sum_{i=1}^{k} f_i^2 = 1 - \left(0{,}16^2 + 0{,}16^2 + 0{,}28^2 + 0{,}24^2 + 0{,}24^2\right) = 1\text{-}0{,}2128 = 0{,}7872$$

Se vogliamo normalizza re l'indice lo dividiamo per il suo massimo dato da
Gmax=1-1 / k=1 = 1 -0,2=0,8

G*=G / Gmax= 0,984

Conclusione

G* molto prossimo ad 1, la distribuzione è molto eterogenea.