Università di Cassino

Esercitazioni di Statistica 1 - *18 Febbraio 2010*

Dott. Mirko Bevilacqua

DATASET STUDENTI

	SEDE DISTACCATA	GENERE	VOTO DIPLOMA	TIPO SCUOLA	PROVINCIA SEDE SCUOLA	CITTA SCUOLA	PUNTEGGIO TEST
1	SEDE DI CASSINO	F	60	TC	FR	CASSINO	5
2	SEDE DI TERRACINA	M	60	LC	LT	FONDI	14,75
3	SEDE DI CASSINO	F	69	LS	FR	CECCANO	15
4	SEDE DI TERRACINA	M	100	GE	LT	FORMIA	15
5	SEDE DI CASSINO	M	73	TC	FR	CASSINO	17,25
6	SEDE DI CASSINO	M	70	LS	FR	VEROLI	17,5
7	SEDE DI CASSINO	F	72	MP	FR	CASSINO	8,75
8	SEDE DI CASSINO	M	64	LS	FR	PONTECORVO	17,25
9	SEDE DI TERRACINA	M	72	LS	FR	CECCANO	18,5
10	SEDE DI TERRACINA	M	70	TC	FR	CECCANO	9,25
11	SEDE DI CASSINO	M	63	LS	FR	SORA	15,75
12	SEDE DI CASSINO	F	76	MP	FR	CASSINO	7,25
13	SEDE DI CASSINO	F	80	TC	FR	CASSINO	8,5
14	SEDE DI TERRACINA	M	100	TC	LT	FORMIA	26,15
15	SEDE DI CASSINO	M	73	TC	LT	FORMIA	22,3
16	SEDE DI CASSINO	F	73	LS	LT	FORMIA	28,8

ESERCIZIO N° 1

Confrontare, facendo uso del boxplot, la distribuzione delle donne e degli uomini per il carattere “Punteggio test”.

BOX PLOT: metodo di rappresentazione grafica che, avvalendosi di alcuni indici di posizione e variabilità, permette di comparare due o più collettivi statistici.

Il box plot di una distribuzione è un grafico caratterizzato da tre elementi principali:

- Una linea o punto che indicano la posizione della mediana della distribuzione.
- Un rettangolo (BOX) la cui altezza indica la variabilità dei valori prossimi alla media.
- Due segmenti che partono dal rettangolo e i cui estremi sono determinati in base ai valori estremi della distribuzione.

Sesso	Punteggio Test **Y**	(Punteggio Test)^2 **Y^2**
F	5	25,00
F	7,25	52,56
F	8,5	72,25
F	8,75	76,56
F	15	225,00
F	28,8	829,44

$$\mu_{(PunteggioTest\_donne)} = 12,22 \quad ; \quad \mu_{(PunteggioTest\_donne^2)} = 213,47$$

$$\sigma^2_{(PunteggioTest\_donne)} = \mu_{(PunteggioTest\_donne^2)} - \left(\mu_{PunteggioTest\_donne}\right)^2 = 213,47 - (12,22)^2 = 64,22$$

$$Me_{(PunteggioTest\_donne)} = \frac{8,50 + 8,75}{2} = 8,63$$

$$Q_{1(PunteggioTest\_donne)} = 7,25 \quad ; \quad Q_{3(PunteggioTest\_donne)} = 15$$

$a = Q_1 - 1,5 \cdot (Q_3 - Q_1) = 7,25 - 1,5 \cdot (15 - 7,25) = -4,38$

da cui

$\alpha = 5$

(α è il minimo dei valori maggiori di a)

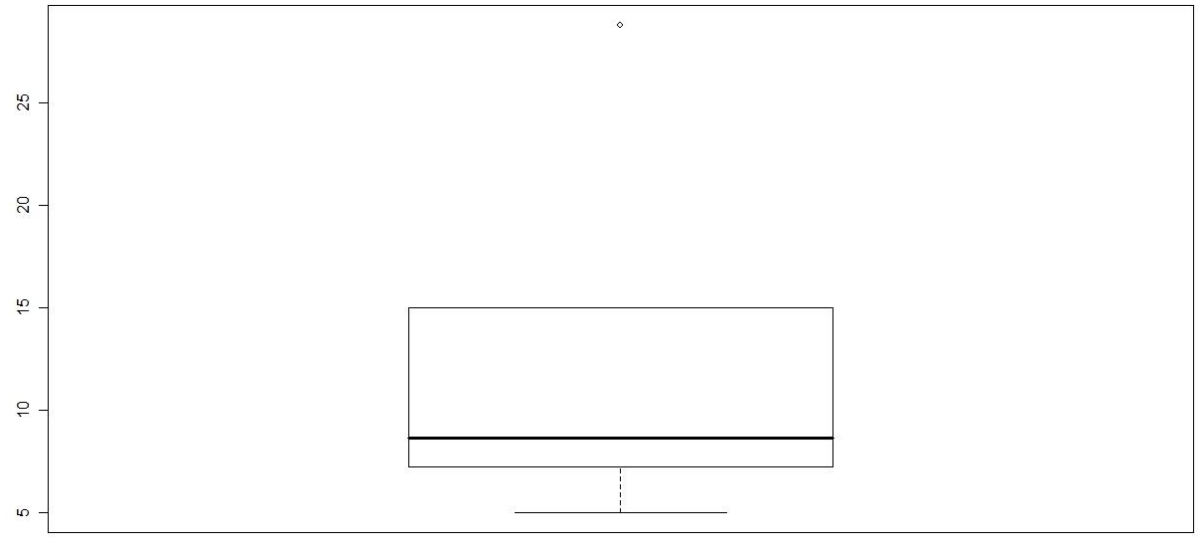
$b = Q_3 + 1,5(Q_3 - Q_1) = 15 + 1,5 \cdot (15 - 7,25) = 26,63$

da cui

$\beta = 15$

(β è il massimo dei valori minori di b)

Figura 1 - Box Plot (PunteggioTest_donne)



Sesso	Punteggio Test **X**	(Punteggio Test)^2 **x^2**
M	9,25	85,6
M	14,75	217,6
M	15	225,0
M	15,75	248,1
M	17,25	297,6
M	17,25	297,6
M	17,5	306,3
M	18,5	342,3
M	22,3	497,3
M	26,15	683,8

$\mu_{(PunteggioTest\_uomini)} = 17,37$; $\mu_{(PunteggioTest\_uomini^2)} = 320,09$

$\sigma^2_{(PunteggioTest\_uomini)} = \mu_{(PunteggioTest\_uomini^2)} - \left(\mu_{(PunteggioTest\_uomini)}\right)^2 = 320,09 - (17,37)^2 = 18,38$

$Me_{(PunteggioTest\_uomini)} = \frac{17,25 + 17,25}{2} = 17,25$

$Q_{1(PunteggioTest\_uomini)} = 15$; $Q_{3(PunteggioTest\_uomini)} = 18,50$

$a = Q_1 - 1,5 \cdot (Q_3 - Q_1) = 15 - 1,5 \cdot (18,5 - 15,0) = 9,75$

da cui

$\alpha = 14,75$

(α è il minimo dei valori maggiori di a)

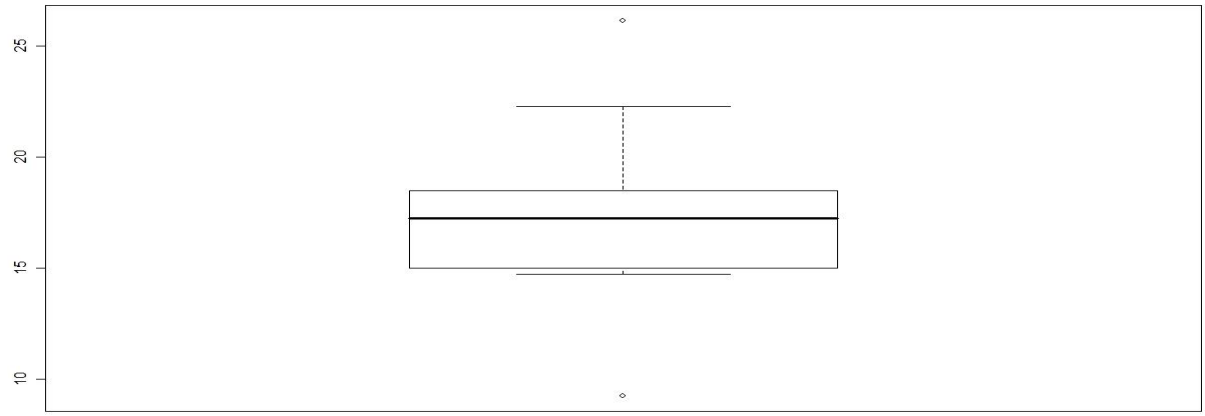
$b = Q_3 + 1,5(Q_3 - Q_1) = 15 + 1,5 \cdot (18,5 - 15) = 23,75$

da cui

$\beta = 22,3$

(β è il massimo dei valori minori di b)

Figura 2 - **Box Plot (PunteggioTest**_uomini**)**



ESERCIZIO N° 2

Calcolare gli indici di forma Yule-Bowley e Hotelling-Solomon per le due distribuzioni uomini e donne.

Una distribuzione si dice asimmetrica se non è possibile individuare un asse verticale che suddivida la distribuzione in due parti specularmente uguali. La nozione di asimmetria ha senso quindi solo se il carattere è almeno ordinabile.

Considerando un carattere quantitativo unimodale, si ha che

1. Se la distribuzione è simmetrica: *Media=Mediana=Moda*
2. Se la distribuzione è asimmetrica positiva: *Media>Mediana>Moda*
3. Se la distribuzione è asimmetrica negativa: *Media<Mediana<Moda*

- ***Indice di asimmetria Hotelling – Solomon:***

$$A_{HS} = \frac{\mu_x - Me}{\sigma_x}$$

$A_{HS}=0$ (simmetria) $-1<A_{HS}<0$ (asimmetria negativa) $0<A_{HS}<1$ (asimmetria positiva)

- ***Indice di asimmetria Yule – Bowley:***

$$A_{YB} = \frac{2 \cdot Me - Q_1 - Q_3}{Q_3 - Q_1}$$

$A_{YB}=0$ (simmetria) $A_{YB}<0$ (asimmetria positiva) $A_{YB}>0$ (asimmetria negativa)

$$A_{HS_{PunteggioTest\_donne}} = \frac{\mu_Y - Me}{\sigma_Y} = \frac{12,22 - 8,63}{8,01} = 0,45;$$

$$A_{HS_{peso\_uomini}} = \frac{\mu_x - Me}{\sigma_x} = \frac{17,37 - 17,25}{4,29} = 0,028.$$

$$A_{YB_{PunteggioTest\_donne}} = \frac{2 \cdot Me - Q_1 - Q_3}{Q_3 - Q_1} = \frac{2 \cdot 8,63 - 7,25 - 15}{15 - 7,25} = -0,645;$$

$$A_{YB_{PunteggioTest\_uomini}} = \frac{2 \cdot Me - Q_1 - Q_3}{Q_3 - Q_1} = \frac{2 \cdot 17,25 - 15 - 18,5}{18,5 - 15} = -0,286.$$

ESERCIZIO N° 3

Calcolare l'indice di eterogeneità di Gini per i caratteri "Tipo Scuola" e "Città Scuola".

TIPO SCUOLA	n_i	f_i	f_i^2
GE	1	0,06	0,004
LC	1	0,06	0,004
LS	6	0,38	0,141
MP	2	0,13	0,016
TC	6	0,38	0,141
TOTALE	16		**0,305**

Indice di eterogeneità di GINI:

$$G = 1 - \sum_{i=1}^{k} f_i^2 = 1 - 0,305 = 0,7$$

$$G_{MAX} = 1 - \frac{1}{K} = 1 - \frac{1}{5} = 0,8$$

Indice di eterogeneità di GINI normalizzato:

$$G^* = \frac{G}{G_{MAX}} = \frac{0,7}{0,8} = 0,87$$

G* è elevato, la distribuzione è molto eterogenea.

CITTÀ SCUOLA	n_i	f_i	f_i^2
CASSINO	5	0,313	0,098
CECCANO	3	0,188	0,035
FONDI	1	0,063	0,004
FORMIA	4	0,250	0,063
PONTECORVO	1	0,063	0,004
SORA	1	0,063	0,004
VEROLI	1	0,063	0,004
	16		0,211

Indice di eterogeneità di GINI:

$$G = 1 - \sum_{i=1}^{k} f_i^2 = 1 - 0,211 = 0,79$$

$$G_{MAX} = 1 - \frac{1}{K} = 1 - \frac{1}{7} = 0,86$$

Indice di eterogeneità di GINI normalizzato:

$$G^* = \frac{G}{G_{MAX}} = \frac{0,79}{0,86} = 0,92$$

G* è elevato, la distribuzione è molto eterogenea.

ESERCIZIO N° 4

Si discuta se esiste connessione tra i caratteri Tipo Scuola e Sesso.

Per misurare l'associazione tra due caratteri qualitativi sconnessi si utilizza l'indice *<<chi-quadrato>>.* Quest'indice si basa sulla differenza tra le frequenze osservate n_{ij} e le frequenze teoriche $\hat{n}_{ij}$ (queste differenza vengono anche dette *contingenze*) che corrispondono alle frequenze che avremmo dovuto avere se, date le distribuzioni semplici, i due caratteri fossero stati indipendenti.
Indice Chi-quadrato di Pearson:

$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}-\hat{n}_{ij})^2}{\hat{n}_{ij}}$ dove r e c rappresentano, rispettivamente, il numero di righe e di colonne della tabella di frequenza a doppia entrata.

Se i due caratteri sono perfettamente indipendenti tutte le contingenze devono essere nulle e dunque l'indice Chi-quadrato assumerà valore pari a zero. Se al contrario i due caratteri sono associati, l'indice sarà positivo, assumendo valori tanto più grandi quanto più le frequenze osservate si differenziano da quelle teoriche.
Le frequenze osservate sono:

Sesso / Tipo Scuola	M	F	Totale
GE	1	0	1
LC	1	0	1
LS	4	2	6
MP	0	2	2
TC	4	2	6
Totale	10	6	16

Le frequenze teoriche $\hat{n}_{ij}$ $\hat{n}_{ij} = \frac{n_{i.} \times n_{.j}}{N}$ sono raccolte nella seguente tabella:

Sesso / Tipo Scuola	M	F	Totale
GE	0,63	0,37	1
LC	0,63	0,37	1
LS	3,75	2,25	6
MP	1,25	0,75	2
TC	3,75	2,25	6
Totale	10	6	16

$$\chi^2 = \frac{(1-0,63)^2}{0,63} + \frac{(0-0,37)^2}{0,37} + \frac{(1-0,63)^2}{0,63} + \frac{(0-0,37)^2}{0,37} + \frac{(4-3,75)^2}{3,75} +$$
$$+ \frac{(2-2,25)^2}{2,25} + \frac{(0-1,25)^2}{1,25} + \frac{(2-0,75)^2}{0,75} + \frac{(4-3,75)^2}{3,75} + \frac{(2-2,25)^2}{2,25} = 4,62$$

L'indice chi-quadrato dipende dalla numerosità del collettivo cosicché, a parità di associazione, il suo valore aumenta all'aumentare di *n*. In genere si preferisce utilizzare degli indici che diano misure non dipendenti dalle frequenze delle distribuzioni marginali o dal totale della tabella, come l'indice di Fisher.

Indice di Fischer:

$$\phi^2 = \frac{\chi^2}{n} = \frac{4,62}{16} = 0,29$$

Tale valore va confrontato con l'intervallo [0, 1], in quanto

$$0 \le \varphi^2 \le \min(r-1;\ c-1) \qquad 0 \le \varphi^2 \le 1$$