Università di Cassino

Esercitazioni di Statistica 1 del 19 Febbraio 2010

Dott. Mirko Bevilacqua

DATASET STUDENTI

N	SESSO	ALTEZZA (cm)	PESO (kg)	CORSO LAUREA	NUMERO SCARPA	COLORE OCCHI	COLORE CAPELLI
1	M	179	65	INFORMAICA	43	scuri	Neri
2	M	180	62	INFORMATICA	43	scuri	Neri
3	F	165	50	INFORMATICA	39	scuri	Castani
4	F	160	49	INFORMATICA	37	verdi	Castani
5	F	160	47	MATEMATICA	37	azzurri	Biondi
6	F	160	48	MATEMATICA	36	scuri	Biondi
7	F	164	56	MATEMATICA	38	verdi	Castani
8	F	170	59	MATEMATICA	38	scuri	Castani
9	M	180	73	MATEMATICA	43	verdi	Castani
10	M	186	86	MATEMATICA	45	azzurri	Neri
11	F	170	66	MATEMATICA	42	scuri	Castani
12	M	180	68	INFORMATICA	41	scuri	Neri
13	M	180	85	INFORMATICA	43	scuri	Castani
14	F	176	56	INFORMAATICA	37	scuri	Biondi
15	M	170	72	INFORMATICA	42	verdi	Castani
16	M	180	65	INFORMATICA	42	scuri	Neri
17	M	170	72	INFORMATICA	41	scuri	Neri
18	F	172	70	INFORMATICA	40	scuri	Castani
19	M	178	80	INFORMATICA	45	scuri	Castani
20	F	162	49	BIOLOGIA	37	scuri	Neri

ESERCIZIO N° 1

Confontare, facendo uso del boxplot, la distribuzione delle donne e degli uomini per il carattere Peso.

Nelle esercitazioni precedenti abbiamo visto che le medie e gli indici di variabilità sono molto utili per descrivere sinteticamente alcune caratteristiche della distribuzione dei dati. Ora introduciamo un metodo di rappresentazione grafica, detto BOX PLOT, che si avvale di tali misure e che risulta estremamente maneggevole nella comparazione di due o più collettivi.

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.

Peso_donne	n_i	f_i	F_i	$(Peso\_donne)^2$
47	1	0,1	0,1	2209
48	1	0,1	0,2	2304
49	2	0,2	0,4	2401
50	1	0,1	0,5	2500
56	2	0,2	0,7	3136
59	1	0,1	0,8	3481
66	1	0,1	0,9	4356
70	1	0,1	1	4900
	$\sum_{i=1}^{8} n_i = 10$	$\sum_{i=1}^{8} f_i = 1$		$\sum_{i=1}^{8} X_i^2 \cdot n_i = 30824$

$\mu_{Peso\_donne} = 55$

$\mu_{(Peso\_donne)^2} = \frac{30824}{10} = 3082,4$

$\sigma^2_{Peso\_donne} = \mu_{Peso\_donne^2} - \left(\mu_{Peso\_donne}\right)^2 = 3082,4 - (55)^2 = 57,4$

$Me_{Peso\_donne} = \frac{50+56}{2} = 53$

$Q_{1(Peso\_donne)} = 49$

$Q_{3(\text{Peso\_donne})} = 59$

$a = Q_1 - 1,5(Q_3 - Q_1) = 49 - 1,5 \cdot (59 - 49) = 34$

da cui

$\alpha = 47$

(α è il minimo dei valori maggiori di a)

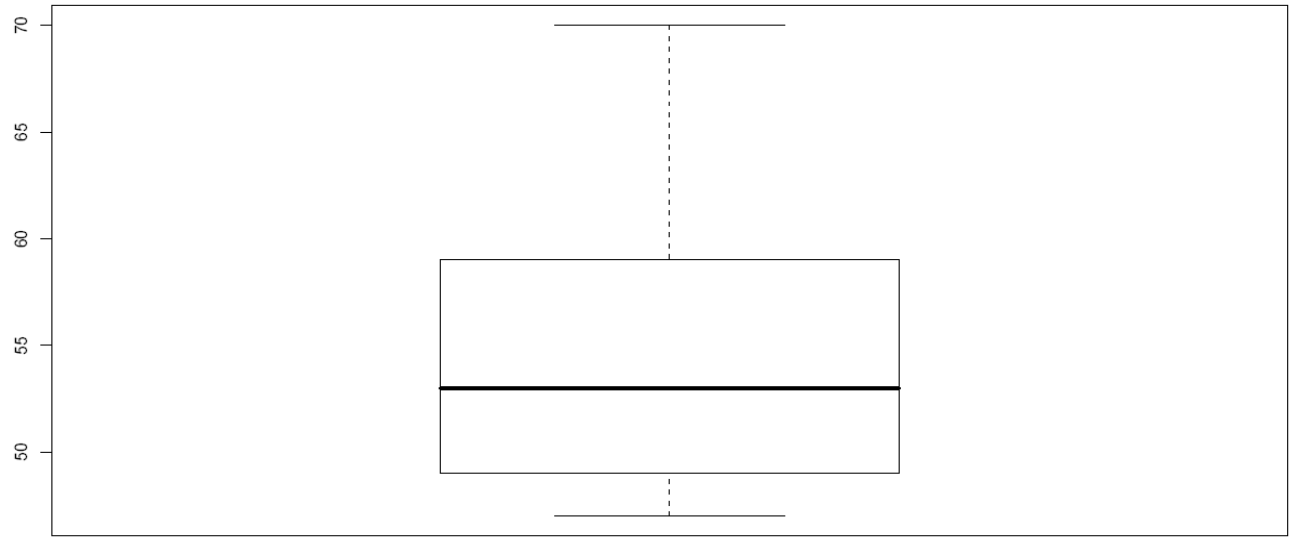
$b = Q_3 + 1,5(Q_3 - Q_1) = 59 + 1,5 \cdot (59 - 49) = 74$

da cui

$\beta = 70$

(β è il massimo dei valori minori di b)

Figura 1. Box Plot (Peso_donne)



Peso_uomini	n_i	f_i	F_i	$(Peso\_uomini)^2$
62	1	0,1	0,1	3844
65	2	0,2	0,3	4225
68	1	0,1	0,4	4624
72	2	0,2	0,6	5184
73	1	0,1	0,7	5329
80	1	0,1	0,8	6400
85	1	0,1	0,9	7225
86	1	0,1	1	7396
	$\sum_{i=1}^{8} n_i = 10$	$\sum_{i=1}^{8} f_i = 1$		$\sum_{i=1}^{8} X_i^2 \cdot n_i = 53636$

$\mu_{Peso\_uomini} = 72,8$

$\mu_{(Peso\_uomini)^2} = \frac{53636}{10} = 5363,6$

$\sigma^2_{Peso\_uomini} = \mu_{Peso\_uomini^2} - \left(\mu_{Peso\_uomini}\right)^2 = 5363,6 - (72,8)^2 = 63,76$

$Me_{Peso\_uomini} = 72$

$Q_{1_{(Peso\_uomini)}} = 65$

$Q_{3_{(Peso\_uomini)}} = 80$

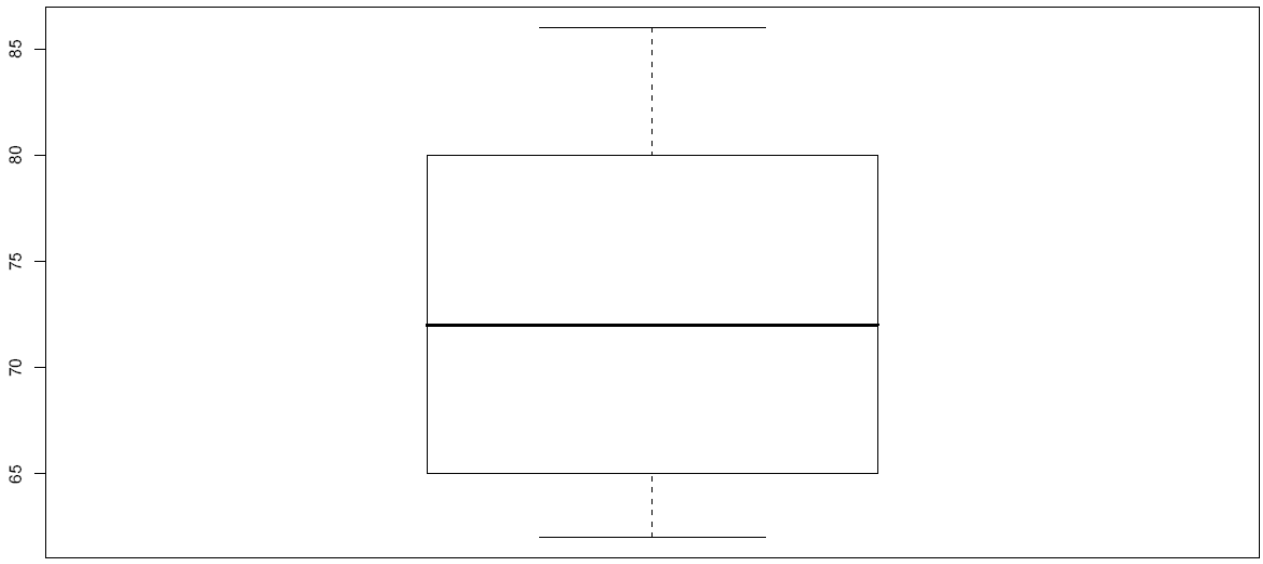
$a = Q_1 - 1,5(Q_3 - Q_1) = 65 - 1,5 \cdot (80 - 65) = 42,5$
da cui
$\alpha = 62$

(α è il minimo dei valori maggiori di a)

$b = Q_3 + 1,5(Q_3 - Q_1) = 80 + 1,5 \cdot (80 - 65) = 102,5$
da cui
$\beta = 86$

(β è il massimo dei valori minori di b)

Figura 2. Box Plot (Peso_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.

Una distribuzione di frequenza $n_1, n_2, \ldots\ldots, n_k$ è simmetrica se:

$n_1 = n_k$; $n_2 = n_{k-1}$; ..... ; $n_j = n_{k-j+1}$

Una distribuzione non simmetrica potrà mostrare asimmetria positive o negative a seconda che siano più frequenti nella distribuzione le modalità più piccole o più grandi .

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)

$$A_{HS_{peso\_donne}} = \frac{\mu_x - Me}{\sigma_x} = \frac{55 - 53}{7,57} = 0,26$$

$$A_{HS_{peso\_uomini}} = \frac{\mu_x - Me}{\sigma_x} = \frac{72,8 - 72}{7,98} = 0,1$$

$$A_{YB_{peso\_donne}} = \frac{2 \cdot Me - Q_1 - Q_3}{Q_3 - Q_1} = \frac{2 \cdot 53 - 49 - 59}{59 - 49} = -0,2$$

$$A_{YB_{peso\_uomini}} = \frac{2 \cdot Me - Q_1 - Q_3}{Q_3 - Q_1} = \frac{2 \cdot 72 - 65 - 80}{80 - 65} = -0,07$$

ESERCIZIO N° 3

Calcolare l'indice di eterogeneità di Gini per i caratteri "Colore Occhi" e "Colore Capelli".

Nell'esercitazione di venerdì scorso, abbiamo visto che gli indici di concentrazione misurano se una distribuzione di quantità si trova più vicino al caso di equidistribuzione oppure al caso di massima concentrazione. Un concetto analogo può essere applicato anche a una distribuzione di frequenza. In questo caso si dirà che si è in presenza di massima omogeneità quando tutte le unità del collettivo presentano stessa modalità, per esempio la j-esima, ottenendo quindi:

max omogeneità: $f1=f2=\ldots\ldots\ldots\ldots=f_{j-1}=\ldots\ldots\ldots\ldots f_k=0$ $\quad$ $e\ f_j=1$

Viceversa, si è in presenza di minima omogeneità o max eterogeneità, se tutte le modalità sono presenti con la stessa frequenza nel collettivo:

max eterogeneità: $f1=f2=\ldots\ldots\ldots\ldots=f_j=\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots f_k=1/k.$

Indice di eterogeneità di GINI:

Dato un carattere qualitativo X con k modalità:

$$G = 1 - \sum_{i=1}^{k} f_i^2$$

In caso di max eterogeneità $f_i = 1/K$ per ogni i.

Sostituendo la precedente la condizione di massima eterogeneità, si ottiene il valore massimo che l'indice può assumere:

$$G_{MAX} = 1 - \frac{1}{K}$$

Indice di Gini normalizzato (ossia compreso tra 0 e 1):

$$G^* = \frac{G}{G_{MAX}} = G \cdot \frac{K}{K-1}$$

COLORE OCCHI	n_i	f_i	$(f_i)^2$
AZZURRI	2	0,1	0,01
SCURI	14	0,7	0,49
VERDI	4	0,1	0,04
	tot. =20		$\sum_{i=1}^{k} f_i^2 = 0,54$

$$G = 1 - \sum_{i=1}^{k} f_i^2 = 1 - 0,54 = 0,46$$

$$G_{MAX} = 1 - \frac{1}{K} = 1 - \frac{1}{3} = 0,66$$

$$G^* = \frac{G}{G_{MAX}} = \frac{0,46}{0,66} = 0,69$$

G* è abbastanza elevato. La distribuzione è abbastanza eterogenea: entrambe le modalità sono presenti, ma con frequenze non equilibrate tra loro.

COLORE CAPELLI	n_i	f_i	$(f_i)^2$
BIONDI	3	0,15	0,02
CASTANI	10	0,5	0,25
NERI	7	0,35	0,12
	tot. =20		$\sum_{i=1}^{k} f_i^2 = 0,39$

$$G = 1 - \sum_{i=1}^{k} f_i^2 = 1 - 0,39 = 0,61$$

$$G_{MAX} = 1 - \frac{1}{K} = 1 - \frac{1}{3} = 0,66$$

$$G^* = \frac{G}{G_{MAX}} = \frac{0,61}{0,66} = 0,9$$

G* molto prossimo ad 1. La distribuzione è molto eterogenea: tutte le modalità sono presenti e con frequenze molto simili tra loro.

ESERCIZIO N° 4

Considerando il DATASET STUDENTI, si discuta se esiste connessione tra i caratteri CORSO DI LAUREA e SESSO.

SESSO / CORSO DI LAUREA	MASCHIO	FEMMINA	Totale
BIOLOGIA	1	0	**1**
MATEMATICA	5	2	**7**
INFORMATICA	4	8	**12**
Totale	**10**	**10**	**20**

La misura dell'associazione tra due caratteri qualitativi sconnessi avviene analizzando la distribuzione congiunta delle frequenze dei due caratteri.

Per questo tipo di analisi, si ricorre, tipicamente, all'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:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(n_{ij} - \hat{n}_{ij}\right)^2}{\hat{n}_{ij}}$$

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.

Poiché "Corso di laurea" e "Sesso" sono due caratteri qualitativi sconnessi, il loro grado di connessione si misura attraverso l'indice χ^2.

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

SESSO / CORSO DI LAUREA	MASCHIO	FEMMINA	Totale
BIOLOGIA	0,5	0,5	**1**
MATEMATICA	3,5	3,5	**7**
INFORMATICA	6	6	**12**
Totale	**10**	**10**	**20**

$$\chi^2 = \frac{(1-0,5)^2}{0,5} + \frac{(0-0,5)^2}{0,5} + \frac{(5-3,5)^2}{3,5} + \frac{(2-3,5)^2}{3,5} + \frac{(4-6)^2}{6} + \frac{(8-6)^2}{6} = 3,62$$

L'indice chi-quadrato dipende dalla numerosità del collettivo cosicché, a parità di associazione, il suo valore aumenta all'aumentare di *n* (cfr. formula dell'indice).

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 e l'indice T-chuprov:

$$\phi^2 = \frac{\chi^2}{n}$$

$$T = \frac{\phi^2}{\min\{r-1; c-1\}} = \frac{\chi^2}{n \cdot \min\{r-1; c-1\}}$$

Indice di Fischer:

$$\phi^2 = \frac{\chi^2}{n} = \frac{3,62}{20} = 0,18$$

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

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

Indice di T-chuprov:

$$T = \frac{\phi^2}{\min\{r-1; c-1\}} = \frac{\chi^2}{n \cdot \min\{r-1; c-1\}} = \frac{3,62}{20 \cdot 1} = 0,18$$

poiché $0 < T < 1$,

possiamo affermare che vi è un basso grado di connessione.