Università di Cassino
Corso di Statistica 1
Esercitazione del 14/11/2006
Dott. Alfonso Piscitelli

Esercizio 1

Il seguente *data set* riporta la rilevazione di alcuni caratteri su un collettivo di 20 soggetti.

Soggetto	Sesso	Età	Reddito (Migliaia di €)	Titolo di studio	Nucleo familiare	Statura (cm)	Colore degli occhi
1	M	22	0,7	Diploma	3	173	NERO
2	F	18	0,2	Lic. Media	4	168	MARRONE
3	F	34	1,6	Diploma	2	165	MARRONE
4	M	42	2,5	Laurea	5	180	NERO
5	F	50	3,2	Diploma	3	163	AZZURRO
6	F	12	0,1	Lic. Elementare	4	160	NERO
7	M	46	3,8	Lic. Media	4	177	MARRONE
8	M	72	1,3	Nessun Titolo	2	164	VERDE
9	F	27	1,2	Laurea	3	158	AZZURRO
10	F	48	1,7	Lic. Media	5	170	NERO
11	F	35	1,9	Laurea	1	167	NERO
12	M	84	0,8	Nessun Titolo	1	159	MARRONE
13	F	21	0,4	Diploma	5	174	AZZURRO
14	F	44	1,8	Diploma	4	164	VERDE
15	M	56	1,9	Lic. Media	2	177	NERO
16	F	58	3,2	Lic. Media	3	172	NERO
17	F	37	2,1	Diploma	1	166	MARRONE
18	F	16	0,1	Lic. Media	4	160	MARRONE
19	M	73	1,6	Lic. Elementare	2	170	AZZURRO
20	M	64	2,2	Lic. Elementare	3	184	VERDE

a) Calcolare il Box-Plot per il carattere **Statura**.
b) Calcolare l'indice di Curtosi di Pearson per il carattere **Statura**.

Soluzioni

a) Il box-plot è una particolare rappresentazione di una distribuzione, dove è possibile evidenziare: la simmetria della distribuzione; la variabilità e la presenza di eventuali valori anomali.

Soggetto	Statura
1	158
2	159
3	160
4	160
5	163
6	164
7	164
8	165
9	166
10	167
11	168
12	170
13	170
14	172
15	173
16	174
17	177
18	177
19	180
20	184

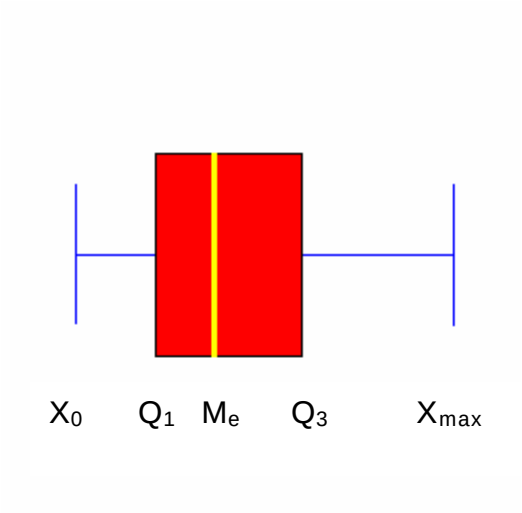
A partire dalla distribuzione ordinata del carattere **Statura**, si calcolano i valori della Mediana, del primo e del terzo Quartile. M_e=167,5; Q_1=163,5 e Q_3=173,5.
La Mediana, il primo e il terzo Quartine ci consentono di disegnare la scatola, mentre per disegnare i baffi dobbiamo prima calcolarci i Limiti Superiori ed Inferiori del box plot.

$$L_I = Q_1 - 1{,}5 * (Q_3 - Q_1)$$

Il baffo di sinistra (o di sotto) sarà il valore più grande tra L_I e X_{min}. Nel nostro caso L_I =148,5 mentre X_{min}=158. Il più grande è X_{min}, quindi il primo baffo coinciderà con il valore di X_{min}.

$$L_S = Q_3 + 1{,}5 * (Q_3 - Q_1)$$

Il baffo di destra (o quello di sopra) sarà il valore più piccolo tra L_S e X_{max}. Nel nostro caso L_S =188,5 mentre X_{max}=184. Il più piccolo è X_{max}, quindi il secondo baffo coinciderà con il valore di X_{max}.



b) La Curtosi riguarda un maggiore o minore appiattimento della forma della distribuzione. L'indice di Pearson, è un indice di forma basato sui momenti quarti standardizzati.

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

$\beta > 3 \rightarrow$ *Leptocurtica*;
$\beta = 3 \rightarrow$ *Normocurtica*;
$\beta < 3 \rightarrow$ *Platicurtica*;

Partendo dalla distribuzione del carattere **Statura** si calcola la media e lo scarto quadratico medio:

$$\mu = 168{,}55 \qquad \sigma = 7{,}145$$

Soggetto	**Statura**	$(x_i - \mu)$	$Z_i = (x_i - \mu)/\sigma$	$Z_i = \left((x_i - \mu)/\sigma\right)^4$
1	158	-10,55	-1,47661	4,754019
2	159	-9,55	-1,33665	3,192012
3	160	-8,55	-1,19668	2,050763
4	160	-8,55	-1,19668	2,050763
5	163	-5,55	-0,77679	0,364102
6	164	-4,55	-0,63683	0,164474
7	164	-4,55	-0,63683	0,164474
8	165	-3,55	-0,49687	0,060949
9	166	-2,55	-0,35691	0,016226
10	167	-1,55	-0,21694	0,002215
11	168	-0,55	-0,07698	3,51E-05
12	170	1,45	0,202946	0,001696
13	170	1,45	0,202946	0,001696
14	172	3,45	0,482872	0,054366
15	173	4,45	0,622835	0,150484
16	174	5,45	0,762798	0,338561
17	177	8,45	1,182686	1,956491
18	177	8,45	1,182686	1,956491
19	180	11,45	1,602575	6,595884
20	184	15,45	2,162426	21,86578
			Tot	**45,74148**

$$\beta = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^4 = \frac{45{,}74148}{20} = 2{,}287$$

La distribuzione è Platicurtica.

Esercizio 2

Un'indagine del 2005, sulle iscrizioni universitarie di due città, rileva i seguenti dati:

	Facoltà		
Città	Umanistica	Scientifica	Totale
Milano	250	170	**420**
Napoli	180	220	**400**
Totale	**430**	**390**	**820**

Si stabilisca se la scelta del tipo di facoltà è indipendente dal luogo di residenza e si commenti il risultato.

L'indice più opportuno per lo studio della relazione tra due mutabili è il χ^2, in quanto si tratta di due variabili entrambe di natura qualitativa e quindi non possiamo studiare altro legame che la connessione.

$$\chi^2 = \sum_{i=1}^{R}\sum_{j=1}^{C}\frac{c_{ij}^2}{n_{ij}^{'}}$$

Dove R sono le righe e C le colonne della Tabella; n_{ij} la frequenza doppia; $n_{i.}$ la frequenza marginale di riga; $n_{.j}$ la frequenza marginale di colonna;

$$c_{ij} = n_{ij} - n_{ij}^{'} \qquad \text{e} \qquad n_{ij}^{'} = {(n_{i.}n_{.j})}/{N}$$

Calcolo delle frequenze teoriche $n_{ij}^{'}$:

	Facoltà		
Città	Umanistica	Umanistica	Totale
Milano	${(420*430)}/{820} = 220{,}24$	${(420*390)}/{820} = 199{,}76$	**420**
Napoli	${(400*430)}/{820} = 209{,}76$	${(400*390)}/{820} = 190{,}24$	**400**
Totale	**430**	**390**	**820**

Riportando le frequenze osservate e le frequenze teoriche in un'unica tabella si passa al calcolo del χ^2

	n_{ij}	n'_{ij}	$(n_{ij} - n'_{ij})^2 / n'_{ij}$
	250	220,24	4,02
	170	199,76	4,43
	180	209,76	4,22
	220	190,24	4,65
Totale (*N*)	820	820	

χ^2	17,32

L'indice χ^2 dipende dalla numerosità del collettivo, cosicché, a parità di associazione, il suo valore aumenta all'aumentare di *N*.

Generalmente si preferisce utilizzare degli indici "normalizzati" che diano misure non dipendenti dalla numerosità.

Perarson:

$$\Phi^2 = \frac{\chi^2}{N} \qquad = 0{,}021$$

In caso di indipendenza assume il suo valore minimo che è zero. Il valore massimo è pari a 1 solo quando il numero di righe o il numero di colonne è uguale a 2, altrimenti risulta maggiore di 1.

Cramer:

$$V = \frac{\Phi^2}{\min[(R-1),(C-1)]} \qquad = 0{,}021$$

Proprietà:

$0 \le V \le 1$

V=0 se vi è indipendenza assoluta o stocastica o in distribuzione tra i caratteri.
V=1 se vi è massima connessione tra i caratteri.