CORSO DI STATISTICA (parte 1) - ESERCITAZIONE 5

Dott.ssa Antonella Costanzo

a.costanzo@unicas.it

Indici di forma, distribuzioni doppie di frequenza e studio del legame tra variabili

Il seguente data set riporta la rilevazione di alcuni caratteri su un collettivo di 20 studenti

Studente	Sesso	Età	Red	Istituto di provenienza	Voto al diploma	Statura (cm)	Colore occhi	Voto esame	Giud. sul corso
1	M	22	0,7	ITC	96	173	Nero	26	Pessimo
2	F	20	0,2	Liceo Classico	92	168	Marrone	26	Ottimo
3	F	30	1,6	Liceo Classico	90	165	Marrone	30	Buono
4	M	22	2,5	Liceo Scient	85	180	Nero	25	Buono
5	F	26	3,2	ITI	100	163	Azzurro	30	Pessimo
6	F	20	0,5	ITC	74	160	Nero	24	Pessimo
7	M	26	4,2	Liceo Scient	60	177	Marrone	20	Suff
8	M	30	1,3	ITC	76	164	Verde	18	Ottimo
9	F	27	1,2	Liceo Scient	100	158	Azzurro	29	Ottimo
10	F	25	1,7	ITI	95	170	Nero	25	Pessimo
11	F	25	1,9	ITI	85	167	Nero	25	Buono
12	M	22	0,7	ITC	97	159	Marrone	27	Buono
13	F	21	0,4	Liceo Classico	65	174	Azzurro	21	Ottimo
14	F	24	1,8	Liceo Scient	70	164	Verde	30	Suff
15	M	20	1,9	Liceo Scient	80	177	Nero	28	Suff
16	F	21	3,2	Liceo Classico	93	172	Nero	27	Pessimo
17	F	27	2,1	ITC	100	166	Marrone	26	Suff
18	F	22	0,1	ITI	84	160	Marrone	24	Buono
19	M	23	1,6	Liceo Scient	92	170	Azzurro	27	Ottimo
20	F	23	2,2	Liceo Scient	73	184	Verde	23	Buono

Esercizio 1. Il boxplot e gli indici di forma

1. Rappresentare graficamente, attraverso il boxplot, la distribuzione della variabile X=Voto all'esame condizionata al Sesso degli studenti. Commentare i risultati ottenuti.
2. Relativamente alla variabile Età organizzata in classi calcolare:
   - l'indice di asimmetria di Fisher
   - l'indice di Yule-Bowley
   - l'indice di Hotelling-Solomon

Soluzione Q.1

Distribuzione del carattere Voto all'esame rispetto al Sesso

Y=Voto esame|X=F	n_i^F	f_i^F	N_i^F	F_i^F
21	1	0.078	1	0.078
23	1	0.078	2	0.156
24	2	0.153	4	0.309
25	2	0.153	6	0.462
26	2	0.153	8	0.615
27	1	0.078	9	0.693
29	1	0.078	10	0.771
30	3	0.230	13	1
Totale (n)	13	1		
Y=Voto esame|X=M	n_i^M	f_i^M	N_i^M	F_i^M
18	1	0.143	1	0.143
20	1	0.143	2	0.286
25	1	0.143	3	0.429
26	1	0.143	4	0.572
27	2	0.286	6	0.858
28	1	0.143	7	1
Totale (n)	7	1		

Dati per costruire i boxplot (sintesi a cinque):

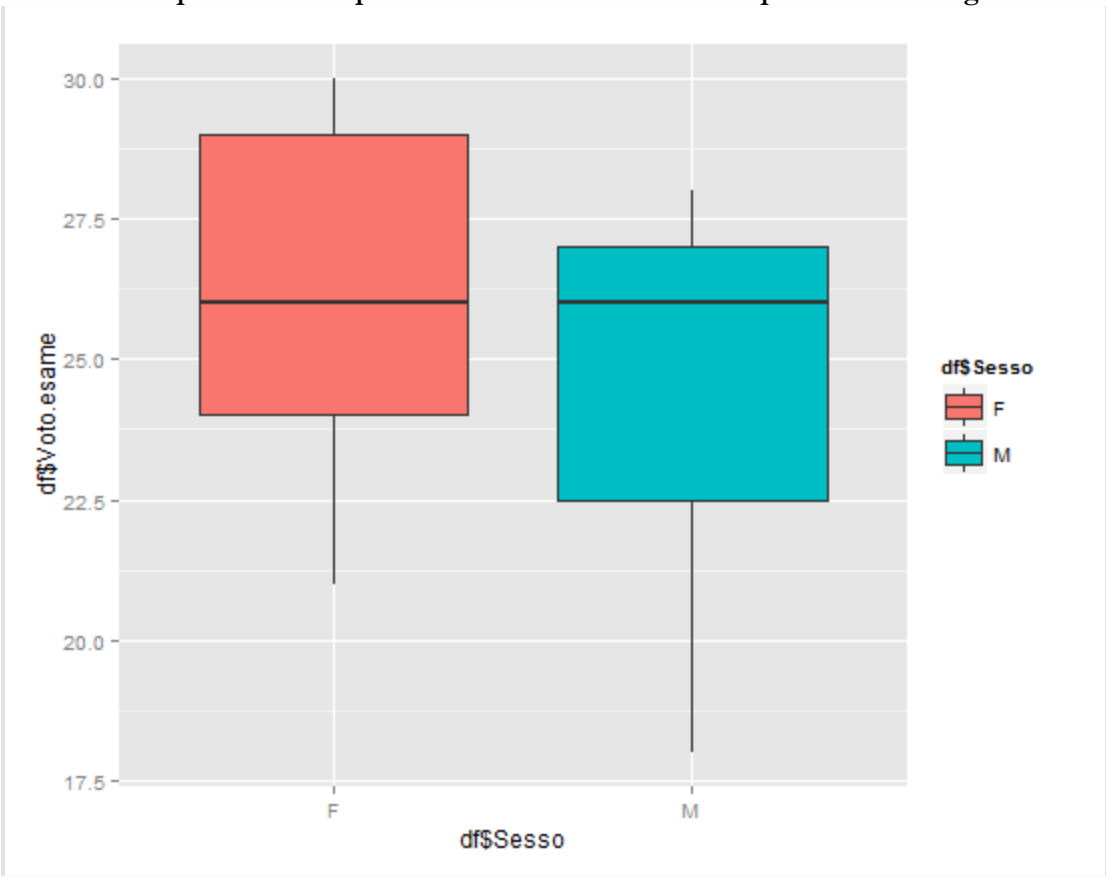
statistiche di sintesi Voto Esame condizionato a Sesso=F

- Min=21
- Q1=24
- Me=26
- Q3=29
- Max=30

statistiche di sintesi Voto Esame condizionato a Sesso=M

- Min=18
- Q1=20
- Me=26
- Q3=27
- Max=28

Grafico 1. Boxplot affiancati per il carattere Voto all'esame rispetto al sesso degli studenti



Commenti:

La mediana per i due gruppi è la stessa. La distribuzione dei voti all'esame per i maschi appare più asimmetrica negativa per cui c'è maggiore tendenza tra i maschi a prendere voti più bassi rispetto alle femmine. Il voto massimo per le femmine è 30 mentre per i maschi è 28. C'è però una maggiore variabilità nella distribuzione dei voti dei maschi rispetto alle femmine. La distribuzione dei voti all'esame per le femmine è più simmetrica rispetto a quella dei maschi. Il range dei voti all'esame (ovvero la differenza tra Q3 e Q1) è comunque simile tra maschi e femmine.

Soluzione Q.2

Partendo dalla distribuzione in classi del carattere Età calcoliamo l'indice di Fisher:

Y=Età (classi)	c_i	n_i	$c_i n_i$	$c_i^2 n_i$	$c_i - \mu$	$(\frac{c_i - \mu}{\sigma})^3 n_i$
(19, 22]	20.5	9	184.5	3782.25	-2.775	-7.416
(22, 24]	23	3	69	1587	-0.275	-0.0024
(24, 26]	25	4	100	2500	1.725	0.7917
(26,28]	27	2	54	1458	3.725	3.9860
(28, 30]	29	2	58	1682	5.725	14.470
		20	465.5	11009.25		11.829

$$\mu = \frac{465.5}{20} 23.275$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} c_i^2 * n_i - \mu^2 = \frac{1}{20}(11009.25) - (23.275)^2 = 8.736$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{8.736} = 2.96$$

Indice di Fisher:

$$\gamma = \frac{1}{N}\sum_{i=1}^{n} (\frac{c_i - \mu}{\sigma})^3 n_i = \frac{1}{20}(11.829) = 0.59$$

$\gamma = 0$ asimmetria nulla

$\gamma > 0$ asimmetria positiva → nel nostro caso, asimmetria positiva per il carattere Età

$\gamma < 0$ asimmetria negativa

Indice di Yule-Bowley

$$A_{YB} = \frac{(Q_3 - Me) - (Me - Q_1)}{(Q_3 - Me) + (Me - Q_1)} = \frac{Q_3 - 2Me + Q_1}{Q_3 - Q_1}$$

Questo indice si basa sul confronto tra i quartili e si concentra sugli sbilanciamenti che si verificano tra le modalità comprese nel 50% centrale della distribuzione

$A_{YB} = 0$ simmetria

$A_{YB} < 0$ asimmetria negativa, quindi dominano valori medio alti

$A_{YB} > 0$ asimmetria positiva, quindi dominano valori medio bassi

Nel nostro caso per la variabile Età:

Y=Età (classi)	c_i	n_i	f_i	F_i
(19, 22]	20.5	9	0.45	0.45
(22, 24]	23	3	0.15	0.60
(24, 26]	25	4	0.20	0.80
(26,28]	27	2	0.1	0.90
(28, 30]	29	2	0.1	1
		20	1	

Q1=20.33

Mediana=22.66

Q3=25.5

$A_{YB} = \frac{Q_3 - 2Me + Q_1}{Q_3 - Q_1} = \frac{25.5 - 2*22.66 + 20.33}{25.5 - 20.33} = \frac{0.51}{5.17} = 0.098$ asimmetria positiva

Nota: L'indice di YB è relativo ed è anche standardizzato ossia $-1 \leq A_{YB} \leq 1$

Il massimo negativo (valore pari a -1) è ottenuto per le distribuzione asimmetriche negative mentre il massimo positivo (valore pari a 1) è raggiunto da distribuzioni asimmetriche positive

Indice di Hotelling-Solomon

$$HS = \frac{\mu - Me}{\sigma}$$

Questo indice si basa sul concetto che, data una distribuzione unimodale,

1. se la distribuzione è simmetrica: Media=Mediana

2. Se la distribuzione è asimmetrica positiva: Media>Mediana

3. Se la distribuzione è asimmetrica negativa: Media<Mediana

Nel nostro caso, avremo quindi:

$HS = \frac{23.275-22.66}{2.96} = 0.207$ asimmetria positiva

Indipendenza e interdipendenza tra caratteri

Concetto relativo allo studio delle relazioni tra due variabili statistiche. Distinguiamo tre concetti di indipendenza a seconda della tipologia di caratteri oggetto di studio:

- Indipendenza assoluta (Indice χ^2di Pearson, Indice φ^2, Indice T di Tchuprov)
- Indipendenza in media (Indice η^2di Pearson)
- Associazione/correlazione lineare (covarianza, coefficiente di correlazione ρ_{xy})

Esercizio 2. La misura dell' indipendenza assoluta: il Chi-quadro

In un collettivo di 470 studenti sono state rilevate le seguenti informazioni relative al titolo di studio e all'attività preferita durante il tempo libero. Le variabili misurate sono quindi X=Titolo di Studio e Y=Tempo Libero.

Tabella 1. Distribuzione congiunta di X e Y (frequenze assolute)

Y=Tempo libero / X=Titolo di studio	Cinema	Teatro	Musica	Sport	**Totale (ni.)**
Lic.media	25	12	18	45	**100**
Diploma	76	58	49	67	**250**
Laurea	39	35	35	11	**120**
Totale (n.j)	**140**	**105**	**102**	**123**	**470**

Verificare se tra i due caratteri esiste indipendenza in distribuzione.

Soluzione

I caratteri Titolo di Studio e Tempo libero sono entrambi qualitativi. Tra essi esiste indipendenza assoluta se le modalità assunte da X non modificano la distribuzione di Y. In altre parole la distribuzione condizionata di $Y|X = x_i$ non cambia per ogni i=1,2,…, h. Similmente la distribuzione condizionata $X|Y = y_j$ non cambia per ogni j=1,2,…, k. In particolare:

- $\frac{n_{ij}}{n_{i.}} = \frac{n_{.j}}{n}$ j=1,…,k (colonne)
- $\frac{n_{ij}}{n_{.j}} = \frac{n_{i.}}{n}$ i=1,…, h (righe)

Quindi, dati due caratteri qualitativi Y e X, organizzati in una tabella doppia di frequenze dove i valori di X sono disposti per riga e i valori di Y in colonna, si può affermare che Y è indipendente in distribuzione da X se i profili colonna (distribuzioni condizionate di Y/X in frequenze relative) sono uguali tra loro e uguali al profilo medio di Y. Analogamente ciò è vero anche per i profili riga.

Nota: si può dimostrare che dati due caratteri qualitativi (X e Y) organizzati in una tabella doppia di frequenze dove i valori di X sono disposti per riga e i valori di Y in colonna, il profilo medio di Y può essere ottenuto come media ponderata delle distribuzioni condizionate di Y rispetto ai valori di X utilizzando come pesi i marginali di riga (profilo medio riga di X). Allo stesso modo, si può dimostrare che il profilo medio di X può essere ottenuto come media ponderata delle distribuzioni condizionate di X rispetto ai valori di Y utilizzando come pesi i marginali di colonna (profilo medio colonna di Y). Nel nostro caso :

Tabella 2. Distribuzione condizionata di Y=Tempo libero rispetto alle modalità di X=Titolo di Studio

Y=Tempo libero / X=Titolo di studio	Cinema	Teatro	Musica	Sport	**Totale**	**Profilo medio fi.** *(distribuzione marginale di riga)*
Lic.media	0.25	0.12	0.18	0.45	**1**	**0.213**
Diploma	0.304	0.232	0.196	0.268	**1**	**0.532**
Laurea	0.325	0.292	0.292	0.091	**1**	**0.255**
Profilo medio f.j *(distribuzione marginale di colonna)*	**0.298**	**0.223**	**0.217**	**0.262**		**1**

$f_{.1}$ =(0.25)*0.213+(0.304)*0.532+(0.325)*0.255=0.298

E cosi via per le altre frequenze del profilo medio:

- $f_{.2}$ = (0.12)* 0.213+(0.232)*0.532+(0.292)*0.255=0.223 etc.

Siccome, in caso di indipendenza assoluta tra X e Y deve valere la seguente:

$$\frac{n_{ij}}{n_{i.}} = \frac{n_{.j}}{n}$$

da un punto di vista operativo, i due caratteri X e Y si dicono indipendenti (in distribuzione) se le frequenze osservate sono uguali alle cosiddette frequenze teoriche per ogni cella (i, j) della distribuzione doppia.

Frequenze teoriche (sotto ipotesi di indipendenza): $\hat{n}_{ij} = \frac{n_{i.}n_{.j}}{n}$

Tabella teorica (ipotesi di indipendenza)

Tempo libero / Titolo di studio	Cinema	Teatro	Musica	Sport	**Totale (ni.)**
Lic.media	29.79	22.34	21.70	26.17	**100**
Diploma	74.47	55.85	54.26	65.43	**250**
Laurea	35.74	26.81	26.04	31.40	**120**
Totale (n.j)	**140**	**105**	**102**	**123**	**470**

Es. calcoli

$$n_{11} = \frac{100 * 140}{470} = 29.79$$

$$n_{21} = \frac{250 * 140}{470} = 74.47$$

$$n_{31} = \frac{120 * 140}{470} = 35.74$$

..e cosi via fino a riempire tutte le altre celle.

A questo punto confronto la tabella delle frequenze teoriche con quella delle frequenze osservate. Le frequenze teoriche sono diverse da quelle osservate, quindi concludo che i caratteri Titolo di Studio e Tempo Libero non sono indipendenti. Un indice che misura il grado di connessione tra due caratteri qualitativi è l'indice χ^2 di Pearson:

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

L'indice assume valore 0 in caso di indipendenza mentre tende a crescere al crescere del grado di connessione tra i caratteri.

Per agevolare i calcoli dell'indice di Pearson posso costruire la seguente tabella dove ogni cella contiene la differenza al quadrato tra frequenze osservate e teoriche diviso la corrispondente frequenza teorica:

Tempo libero / Titolo di studio	Cinema	Teatro	Musica	Sport
Lic.media	0.77	4.79	0.63	13.55
Diploma	0.03	0.08	0.51	0.04
Laurea	0.30	2.50	3.08	13.26

Si procede alla somma di tutti gli incroci relativi a tale tabella e si ottiene:

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}} = 39.54$$

Si conclude che tra i caratteri non esiste indipendenza in distribuzione. Una misura relativa dell'indipendenza in distribuzione è data dall'indice di contingenza in media quadratica che si ottiene dividendo il χ^2 per la numerosità delle osservazioni n.

$$\varphi^2 = \frac{\chi^2}{n} = \frac{39.54}{470} = 0.084$$

Tale valore va confrontato con l'intervallo [0, 2] in quanto $0 \leq \varphi^2 \leq \min(h-1, k-1)$ dove h=numero di righe e k=numero di colonne. Possiamo dire che esiste un basso grado di connessione.

L'indice T di Tchuprov è la versione normalizzata dell'indice φ^2 ed è più semplice da interpretare, poiché varia tra 0 e 1.
Se prossimo a 0 esiste indipendenza assoluta tra i caratteri. Se, d'altro canto, il suo valore è prossimo a 1 vuol dire che i caratteri sono perfettamente associati

$$Indice\ T = \frac{\varphi^2}{\min[(h-1), (k-1)]} = \frac{\chi^2}{\mathrm{n} * \min[(h-1), (k-1)]} = 0.168$$

Esercizio 3. Misura dell'indipendenza in media tra caratteri

In un collettivo di giovani si è osservato l'atteggiamento verso il fumo per classi di età ottenendo la seguente distribuzione di frequenze:

Y= Età (classi)	**Fuma**	**Non Fuma**
[16, 18]	7	16
(18, 22]	8	18
(22, 25]	21	9
(25, 30]	30	10

Quesiti:

1. Verificare se esiste indipendenza in media tra l'età e l'abitudine al fumo
2. Calcolare il rapporto di correlazione dell'età all'atteggiamento verso il fumo

Soluzione Q. 1

Dati due caratteri X qualitativo e Y quantitativo si dice che Y è indipendente in media da X se alla variare delle modalità della X le medie delle distribuzioni condizionate di Y rimangono costanti, ovvero:

$$\mu(Y|X = x_1) = \mu(Y|X = x_2) \dots \mu(Y|X = x_k) = \mu(Y)$$

Nota: indipendenza in distribuzione →indipendenza in media (ma non vice-versa)

Distribuzione congiunta della variabile Età rispetto all'abitudine al fumo

Y= Età (classi)	**c_i=valore centrale**	**Fuma**	**Non Fuma**	**Totale**
[16, 18]	17	7	16	**23**
(18, 22]	20	8	18	**26**
(22, 25]	23.5	21	9	**30**
(25, 30]	27.5	30	10	**40**
Totale		**66**	**53**	**119**

Indichiamo con $\mathrm{n_F}$ il totale dei soggetti fumatori e con $\mathrm{n_{NF}}$ il totale dei soggetti non fumatori. La media condizionata di Y dato che X=fuma è pari a:

$$\mu_{y|X=Fuma} = \frac{1}{n_F}\sum_{i=1}^{n_F} c_i * n_{i1} = \frac{1}{66}(17 * 7 + 20 * 8 + 23.5 * 21 + 27.5 * 30) = 24.205$$

La media condizionata di Y dato che X=non fuma è pari a:

$$\mu_{y|X=Non\ Fuma} = \frac{1}{n_{NF}} \sum_{i=1}^{n_{NF}} c_i * n_{i2} = \frac{1}{53}(17 * 16 + 20 * 18 + 23.5 * 9 + 27.5 * 10) = 21.103$$

La media generale è pari a:

$$\mu = \frac{1}{N} \sum_{i=1}^{n} c_i * n_{i.} = \frac{1}{119}(17 * 23 + 20 * 26 + 23.5 * 30 + 27.5 * 40) = 22.823$$

Le medie di Y condizionate alle modalità di X non sono costanti e sono diverse dalla media generale. Tra i due caratteri non esiste indipendenza in media.

Soluzione Q.2

Il rapporto di correlazione tra Y e X rappresenta l'indice η^2 di Pearson, nel nostro caso definito nel modo seguente:

$$\eta^2_{Y|X} = \frac{Dev_{between}}{Dev_{tot}} = \frac{\sum_{j=1}^{J}(\mu_{Y|X=x_j} - \mu)^2\, n_{.j}}{\sum_i \sum_j (c_i - \mu)^2\, n_{i.}}$$

E' un indice normalizzato che varia tra 0 (massima indipendenza in media) a 1 (massima dipendenza in media)

L'indice descrive quanta parte della devianza totale è spiegata dalla variabilità delle medie parziali rispetto alla media generale. In caso di massima dipendenza in media la devianza totale coincide con la devianza esterna per cui la variabilità del fenomeno è unicamente spiegata dalla variabilità delle medie condizionate rispetto alla media generale. Allo stesso modo, se X e Y sono perfettamente indipendenti in media, la devianza complessiva coincide con la varianza interna ai gruppi essendo la devianza esterna esattamente pari a zero (in caso di indipendenza, le medie condizionate saranno tutte costanti e la variabilità ad esse associata sarà quindi nulla).

Per comodità, calcoliamo la devianza esterna ai gruppi:

$$Dev_{between} = \sum_j (\mu_{Y|X=x_j} - \mu)^2 * n_{.j}$$

$$= (24.205 - 22.823)^2 * 66 + (21.103 - 22.823)^2 * 53 = 282.5$$

La devianza totale è pari a:

$$Dev_{tot} = \sum_i \sum_j (c_i - \mu)^2 * n_{i.} =$$

$$= (17 - 22.823)^2 * 23 + (20 - 22.823)^2 * 26 + (23.5 - 22.823)^2 * 30 + (27.5 - 22.823)^2 * 40$$
$$= 1875.87$$

Quindi:

$$\eta^2_{Y|X} = \frac{Dev_{between}}{Dev_{tot}} = \frac{282.5}{1875.87} = 0.1506$$

Nota: Ricordando che vale la seguente: $Dev_{tot} = Dev_{within} + Dev_{between}$ possiamo ottenere per differenza la devianza interna ai gruppi: $D_{within} = Dev_{tot} - Dev_{between} = 1875.87 - 282.5 = 1593.37$

Appendice

Il boxplot e la gestione dei valori anomali

Consideriamo la seguente tabella di frequenza relativa alla variabile X=punteggio all'esonero per un collettivo di 31 studenti

X=punteggio	n_i	f_i	N_i	F_i
23	3	0.097	3	0.097
24	6	0.194	9	0.291
25	8	0.258	17	0.549
26	5	0.161	22	0.71
27	5	0.161	27	0.871
28	3	0.097	30	0.968
88	1	0.032	31	1
Totale (n)	31	1		

Il grafico a scatola (box-plot) è una particolare rappresentazione di una distribuzione. E'ottenuto a partire da 5 numeri di sintesi: minimo, 1° quartile (Q1), mediana, 3° quartile (Q3), massimo.
Il box plot o diagramma a scatola e baffi si ottiene riportando su un asse verticale (oppure orizzontale) i 5 numeri di sintesi. La scatola del box plot ha come estremi inferiore e superiore rispettivamente Q1 e Q3. La differenza tra Q3 e Q1 costituisce il campo di variazione interquartile, indicato con CVI=Q3-Q1. La mediana divide la scatola in due parti. I baffi si ottengono congiungendo Q1 al minimo osservato e Q3 al massimo osservato nella distribuzione della variabile di interesse.

Dati (sintesi a 5):

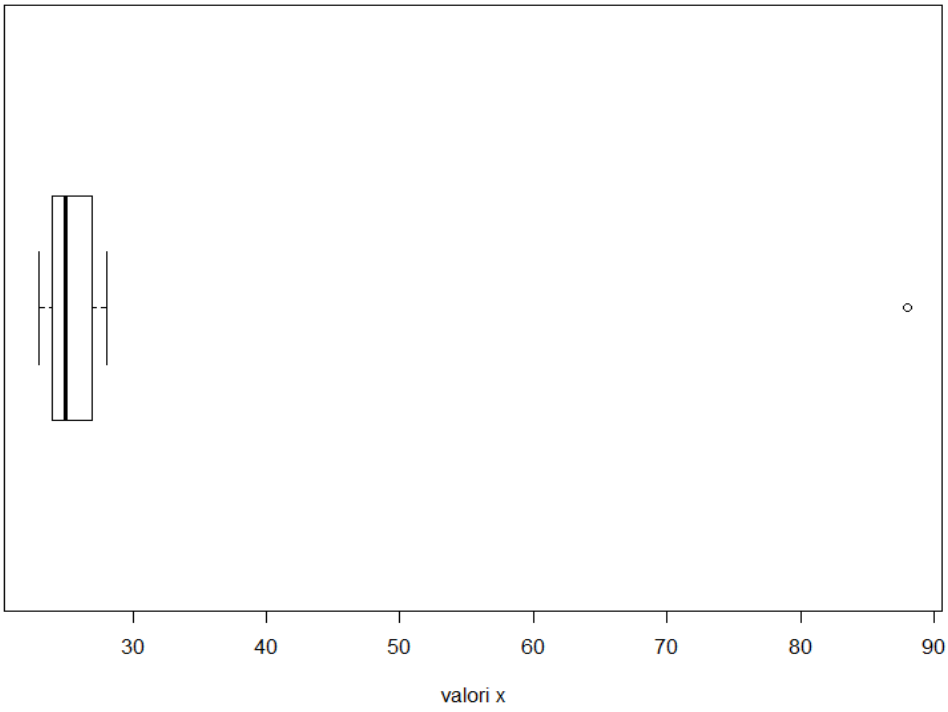
Min(x)	Q_1	Q_2	Q_3	Max(x)
23	24	25	27	88

$$CVI = Q_3 - Q_1 = 27 - 24 = 3$$

La distanza tra il terzo ed il primo quartile (CVI), è una misura della dispersione della distribuzione. Il 50% delle osservazioni si trovano comprese tra questi due valori. Se il campo di variazione interquartile è piccolo, tale metà delle osservazioni si trova fortemente concentrata intorno alla mediana; all'aumentare della distanza interquartilica aumenta la dispersione del 50% delle osservazioni centrali intorno alla mediana.
Le distanze tra ciascun quartile e la mediana forniscono informazioni relativamente alla forma della distribuzione. Se una distanza è diversa dall'altra allora la distribuzione è asimmetrica (vedi indice di Yule-Bowley che sfrutta proprio queste considerazioni).

Rappresentiamo mediante il boxplot variabile X oggetto di studio.

Grafico 1. Boxplot variabile X



La rappresentazione evidenzia la presenza di un valore *anomalo*. I valori anomali (distanti rispetto a tutti gli altri valori che caratterizzano la distribuzione) vengono determinati dal confronto con il campo di variazione interquartile. In particolare vengono considerate due soglie:

- il valore al di sotto del quale una modalità viene considerata outlier:
  $Q_1 - 1.5(Q_3 - Q_1) = 24 - 1.5(3) = 19.5$
- il valore al di sopra del quale una modalità viene considerata outlier:
  $Q_3 + 1.5(Q_3 - Q_1) = 27 + 1.5(3) = 31.5$

I valori al di fuori di queste soglie, costituiscono appunto un' "anomalia" rispetto alla maggior parte dei valori osservati e pertanto è necessario non solo identificarli ma anche analizzarne le caratteristiche e le eventuali cause che li hanno determinati. Essi infatti forniscono informazioni ulteriori sulla dispersione e sulla forma della distribuzione.

Nota operativa per non confondersi: nel boxplot, i baffi vengono tracciati congiungendo, rispettivamente, il minimo valore osservato (non anomalo) al primo quartile e il massimo valore osservato (non anomalo) al terzo quartile della distribuzione ordinata di X. Nel nostro caso, il massimo valore osservato a sinistra della soglia è 31.5 mentre il minimo osservato sempre rispetto alla soglia è 19.5. In sostanza, i baffi individuano gli intervalli in cui sono posizionati i valori rispettivamente minori di Q1 e maggiori di Q3; i punti estremi dei "baffi" evidenziano i limiti ovvero i valori di minimo e massimo propri della distribuzione. Il confronto di un valore particolarmente distante rispetto alle soglie consente di individuare i valori esterni a questi limiti. Questi ultimi costituiscono

infatti una "anomalia" rispetto alla maggior parte dei valori osservati e rispetto ai valori limite della distribuzione di X.

Nel nostro caso il valore $x_i = 88$ risulta al di sopra della soglia max quindi può essere considerato un outlier. A differenza del grafico 1 in cui si isola il particolare valore anomalo e la distribuzione di X appare fortemente asimmetrica, se non avessimo avuto questo valore nella distribuzione, il boxplot sarebbe stato il seguente:

Grafico 2. Boxplot della variabile X (senza outlier)