

Corso di Statistica Facoltà di Economia *francesco mola*

a.a. 2010-2011

Lezione n° 6 **Curata da Claudio Conversano**

Sommario

- Boxplot (*grafico a scatola*)
- Steam and Leaf (*grafico ramo-foglia*)
- Esempio riepilogativo sulla forma di una distribuzione

lez6_2010/2011

2

IL BOXPLOT

Permette di studiare graficamente la forma di una distribuzione

Costruzione del boxplot:

1. Si determinano i quantili: $x_{(1)}$, Q_1 , $Q_2 = Me$, Q_3 , $x_{(n)}$
2. Si disegna una scatola di estremi Q_1 e Q_3 tagliata sulla mediana



lez6_2010/2011

3

3. Si calcolano i valori:

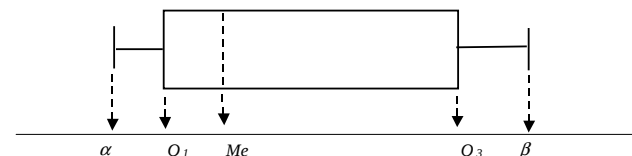
$$a = Q_1 - 1,5(Q_3 - Q_1); \quad b = Q_3 + 1,5(Q_3 - Q_1)$$

4. Si calcolano i valori:

α = minimo dei valori maggiore di a

β = massimo dei valori minori di b

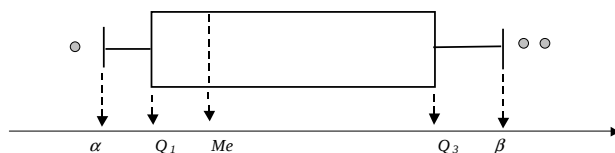
5. Si disegnano i baffi sui valori α e β



lez6_2010/2011

4

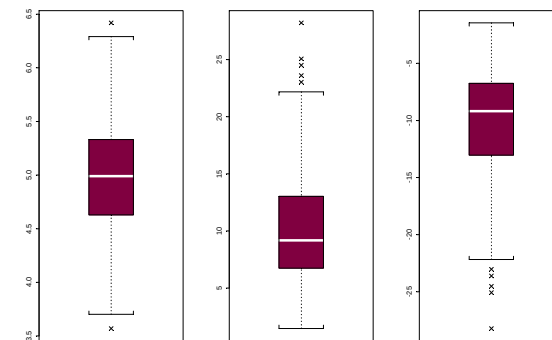
6. I valori oltre α e β sono disegnati con dei puntini (valori anomali)



Caratteristiche del boxplot:

- Permette di visualizzare le informazioni sulla forma e la variabilità di una distribuzione
- Consente di confrontare graficamente la forma di più distribuzioni

Esempi



Distribuzione
simmetrica

Distribuzione
asimmetrica
positiva

Distribuzione
simmetrica
negativa

STEAM AND LEAF PLOT

E' una rappresentazione grafica di dati che si ottiene disponendo intorno ad un ramo (*stem*) le cifre significative di una distribuzione di frequenza e diramando dal ramo alle foglie (*leaf*) le cifre meno significative.

Per rappresentare la distribuzione si costruisce una tabella a due colonne. Nella prima colonna, per ogni valore del cambiamento nei livelli principali della variabile osservata (ad es., le decine), si riportano le cifre finali corrispondenti, mentre nella seconda colonna le relative unità, queste ultime sono indicate tante volte quanti sono i cambiamenti assumenti tali valori,

I valori inferiori alla decina sono indicati antepoendo ad essi la cifra 0.

Esempio

Variabile X:

83,4 84,3 85,3 86,7 86,7 87,4 87,5 87,6 87,7 87,8 87,9 88,2 88,3 88,3 88,3 88,5 88,5 88,6
88,6 88,7 88,9 89,0 89,2 89,3 89,3 89,6 89,7 89,8 89,8 89,8 89,9 89,9 90,0 90,1 90,1 90,3
90,3 90,4 90,4 90,4 90,5 90,6 90,7 90,8 90,9 91,0 91,0 91,0 91,1 91,1 91,1 91,2 91,2 91,5
91,6 91,6 91,8 91,8 92,2 92,2 92,2 92,3 92,6 92,7 92,7 92,7 93,0 93,2 93,3 93,3 93,4 93,7
94,2 94,2 94,4 94,7 95,6 96,1 100,3

```
083 | 4
084 | 3
085 | 3
086 | 7
087 | 456789
088 | 2333556679
089 | 02336788899
090 | 0113344456789
091 | 0001112256688
092 | 22236777
093 | 023347
094 | 2247
095 | 6
096 | 1
097 |
098 |
099 |
100 | 3
```

ESEMPIO RIEPILOGATIVO

X: Carattere quantitativo discreto

0,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,2,2,2,3,3,3,3,3,3,4,4,4,4,4,4,5,5,6,6,7,8,8, 11

Analisi della forma mediante istogramma (5 classi equampie)

Fissiamo $k=5$ per la suddivisione in classi della variabile X .

L'ampiezza delle classi sarà data da: $b = \frac{11-0}{5} = 2,20 \approx 2,00$

Ogni classe dovrebbe avere ampiezza 2,00

Trattandosi di un carattere quantitativo discreto è possibile approssimare per difetto 2,20 e quindi considerare classi di ampiezza 2.

Le classi saranno:

$0 \leq x_i \leq 2$, Prima Classe
 $2 < x_i \leq 4$, Seconda Classe
 $4 < x_i \leq 6$, Terza Classe
 $6 < x_i \leq 8$, Quarta Classe
 $8 < x_i \leq 11$, Quinta Classe

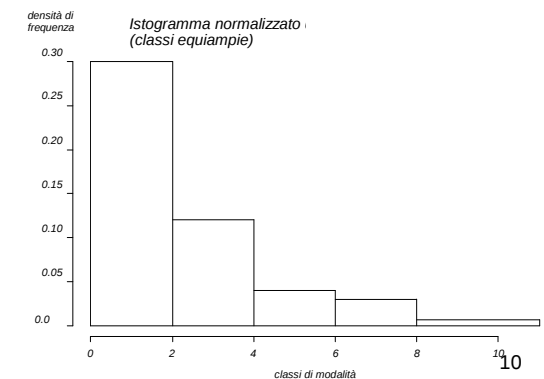
lez6_2010/2011

9

La distribuzione in classi equampie di X risulta:

Classi $x_i \leq x_l \leq x_{i+1}$	Frequenza assoluta n_i	Frequenza relativa $f_i = \frac{n_i}{n}$	Densità di frequenza $d_i = \frac{f_i}{b}$
0 -- 2	30	0,60	0,30
2 -- 4	12	0,24	0,12
4 -- 6	4	0,08	0,040
6 -- 8	3	0,06	0,030
8 -- 11	1	0,02	0,0067
Totale	50	1,00	

Distribuzione asimmetrica positiva



lez6_2010/2011

Steam and Leaf Plot

Variabile X:

0,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,2,2,2,3,3,3,3,3,3,4,4,4,4,4,4,5,5,6,6,7,8,8, 11

```

00 | XX
01 | XXXXXXXXXXXXXXXXXXXX
02 | XXXXXXXXXXXXXXXX
03 | XXXXXXX
04 | XXXXX
05 | XX
06 | XX
07 | X
08 | XX
09 |
10 |
11 | X
    
```

lez6_2010/2011

11

Disegnare il boxplot considerando:

- tutte le 50 osservazioni ($n=50$)
- escludendo l'ultima osservazione ($n=49$)

0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 5, 5, 6, 6, 7, 8, 8, 11

Caso a): $n=50$

$$Me = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} = \frac{x_{(25)} + x_{(25+1)}}{2} = \frac{2 + 2}{2} = 2$$

Per il calcolo del primo quartile possiamo considerare la mediana delle prime 25 osservazioni.

0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2

$$Q_1 = x_{\left(\frac{n+1}{2}\right)} = x_{\left(\frac{25+1}{2}\right)} = x_{(13)} = 1$$

lez6_2010/2011

12

Per il calcolo del terzo quartile possiamo considerare la mediana delle osservazioni che vanno dalla 26-esima alla 50-esima.

2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 5, 5, 6, 6, 7, 8, 8, 11

$$Q_3 = x_{\left(\frac{n+1}{2}\right)} = x_{\left(\frac{25+1}{2}\right)} = x_{(13)} = 4$$

Caso b): $n=49$

Eliminiamo l'ultima osservazione

0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 5, 5, 6, 6, 7, 8, 8, 11

$$Me = x_{\left(\frac{n+1}{2}\right)} = x_{\left(\frac{49+1}{2}\right)} = x_{(25)} = 2$$

Per il calcolo del primo quartile consideriamo le prime 24 osservazioni ($n=24$):

0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2

$$Q_1 = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} = \frac{x_{(12)} + x_{(12+1)}}{2} = \frac{1+1}{2} = 1$$

13

Per il calcolo del terzo quartile consideriamo le osservazioni che vanno dalla 26-esima alla 49-esima ($n=24$).

2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 5, 5, 6, 6, 7, 8, 8

$$Q_3 = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} = \frac{x_{(12)} + x_{(12+1)}}{2} = \frac{4+4}{2} = 4$$

Per disegnare i diagrammi a scatola i valori caratteristici da considerare sono i seguenti:

Caso a) $n=50$: $Q_1 = 1$, $Me = 2$, $Q_3 = 4$,

$$a = Q_1 - 1,5(Q_3 - Q_1) = 1 - 4,5 = -3,5 \Rightarrow \alpha = 0$$

$$b = Q_3 + 1,5(Q_3 - Q_1) = 4 + 4,5 = 8,5 \Rightarrow \beta = 8$$

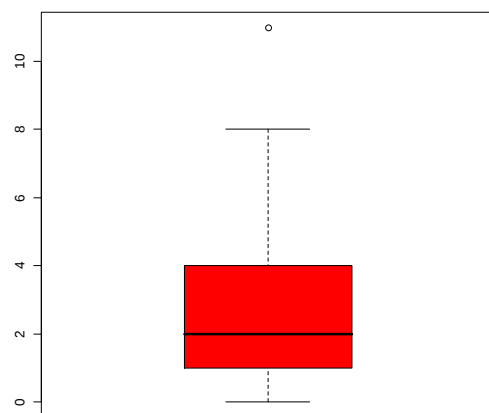
Caso b) $n=49$: $Q_1 = 1$, $Me = 2$, $Q_3 = 4$,

$$a = Q_1 - 1,5(Q_3 - Q_1) = 1 - 4,5 = -3,5 \Rightarrow \alpha = 0$$

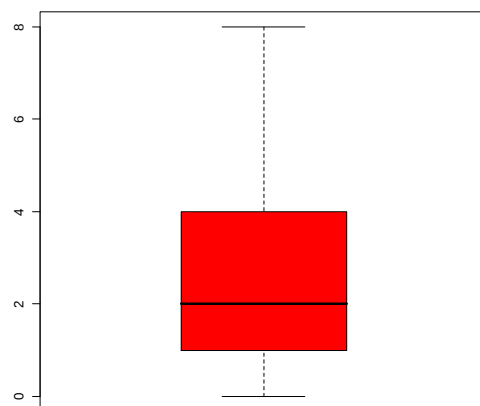
$$b = Q_3 + 1,5(Q_3 - Q_1) = 4 + 4,5 = 8,5 \Rightarrow \beta = 8$$

14

Boxplot di X ($n=50$)



Boxplot di X ($n=49$)



lez6_2010/2011

15

Con riferimento al carattere X procedere alla standardizzazione della distribuzione, calcolare media e varianza della variabile standardizzata e calcolare l'asimmetria e la curtosi.

X_i	0	1	2	3	4	5	6	7	8	11
n_i	2	16	12	6	6	2	2	1	2	1

Da calcoli effettuati in precedenza risulta:

$$\mu = 2,76$$

$$\sigma = 2,25$$

Variabile standardizzata: $Z = \frac{X - \mu_x}{\sigma_x}$

I valori standardizzati $z_i = \frac{x_i - \mu}{\sigma}$ risultano:

X_i	0	1	2	3	4	5	6	7	8	11
Z_i	-1,23	-0,78	-0,34	0,11	0,55	1,00	1,44	1,88	2,33	3,66

Calcoliamo la media

Z_i	-1,23	-0,78	-0,34	0,11	0,55	1,00	1,44	1,88	2,33	3,66
n_i	2	16	12	6	6	2	2	1	2	1

lez6_2010/2011

16

$$\mu_z = \frac{1}{n} \sum_{i=1}^k z_i n_i = \frac{1}{50} [-1,23 \times 2 + (-0,78 \times 16) + (-0,34 \times 12) + 0,11 \times 6 + 0,55 \times 6 + 1 \times 2 + 1,44 \times 2 + 1,88 \times 1 + 2,33 \times 2 + 3,66 \times 1] = \frac{0}{50} = 0$$

Calcoliamo la varianza

$$\mu_z^2 = \frac{1}{n} \sum_{i=1}^k z_i^2 n_i = \frac{1}{50} [-1,23^2 \times 2 + (-0,78)^2 \times 16 + (-0,34)^2 \times 12 + 0,11^2 \times 6 + 0,55^2 \times 6 + 1^2 \times 2 + 1,44^2 \times 2 + 1,88^2 \times 1 + 2,33^2 \times 2 + 3,66^2 \times 1] = \frac{50}{50} = 1$$

$$\mu_z^2 = (0)^2 = 0$$

$$\sigma_z^2 = \mu_{z^2} - (\mu_z)^2 = 1 - 0 = 1$$

Calcoliamo l'asimmetria

z_i	n_i	z_i^3	$z_i^3 n_i$
-1,23	2	-1,846	-3,69
-0,78	16	-0,479	-7,66
-0,34	12	-0,039	-0,46
0,11	6	0,001	0,01
0,55	6	0,167	1,00
1,00	2	0,987	1,97
1,44	2	2,986	5,97
1,88	1	6,692	6,69
2,33	2	12,631	25,26
3,66	1	49,117	49,12
Totale	50		78,22

$$\gamma_1 = \frac{1}{n} \sum_{i=1}^k z_i^3 n_i = \frac{78,22}{50} = 1,56 \quad (\text{asimmetria positiva})$$

Calcoliamo la curtosi

z_i	n_i	z_i^4	$z_i^4 n_i$
-1,23	2	2,289	4,58
-0,78	16	0,370	5,92
-0,34	12	0,013	0,16
0,11	6	0,000	0,00
0,55	6	0,092	0,55
1,00	2	1,000	2,00
1,44	2	4,300	8,60
1,88	1	12,492	12,49
2,33	2	29,473	58,95
3,66	1	179,442	179,44
Totale	50		272,69

$$P_2 = \frac{1}{n} \sum_{i=1}^n z_i^4 n_i = \frac{272,69}{50} = 5,45$$

$$\gamma_2 = P_2 - 3 = 5,45 - 3 = 2,45$$

Distribuzione leptocurtica

Misura il rapporto relativo tra "corpo centrale" e "code" di una distribuzione di frequenza.

La distribuzione empirica è confrontata con una distribuzione teorica (normale standardizzata)

L'indice di curtosi va interpretato con cautela, soprattutto se le distribuzioni sono asimmetriche e/o plurimodali