Fonti e strumenti statistici per la comunicazione

Prof.ssa Isabella Mingo
A.A. 2016-2017

SAPIENZA
UNIVERSITÀ DI ROMA

A che punto siamo

- Lezione introduttiva : Introduzione all'analisi dei dati; (Mingo) (26 settembre)
- Parte I : I dati: introduzione alla terminologia; L'analisi monovariata (Bocci) (lezioni: 27 settembre – 24 ottobre)
- Parte II (4h) : Le fonti statistiche e l'analisi secondaria; Rappresentazioni grafiche (25 ottobre 15 novembre) (Mingo)
- Parte III : La costruzione e l'uso di variabili "complesse"; introduzione all'analisi bivariata per caratteri qualitativi (Mingo) (lezioni dal 16 novembre al 5 dicembre)
- Parte IV: L'analisi bivariata per caratteri quantitativi (Bocci) (lezioni dal 6 al 13 dicembre)

Esercizio

Prova di verifica : distribuzione dei voti

xi	ni	pi	Pi
5	1	.9	.9
8	1	.9	1.7
9	2	1.7	3.4
10	1	.9	4.3
11	1	.9	5.1
12	5	4.3	9.4
13	2	1.7	11.1
14	6	5.1	16.2
15	3	2.6	18.8
16	4	3.4	22.2
18	13	11.1	33.3
19	7	6.0	39.3
20	8	6.8	46.2
21	9	7.7	53.8
22	13	11.1	65.0
23	4	3.4	68.4
24	3	2.6	70.9
25	1	.9	71.8
26	4	3.4	75.2
27	6	5.1	80.3
28	4	3.4	83.8
29	10	8.5	92.3
30	3	2.6	94.9
31	6	5.1	100.0
Totale	117	100.0	

Calcolare:
Minimo, Massimo, Moda,
Mediana, Quartili

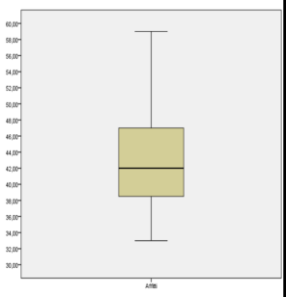
VOTO		
N		117
Minimo		5
Massimo		31
Moda		18 e 22
Quartili	1	18
	2	21
	3	26

Come si rappresenta una distribuzione tenendo conto delle caratteristiche espresse da questi valori ?

a.a 2016-2017

Rappresentare graficamente la distribuzione di un carattere: Il Box plot

È un grafico che consente di rappresentare la distribuzione di un carattere quantitativo (discreto o continuo) mettendone in evidenza la sua variabilità.



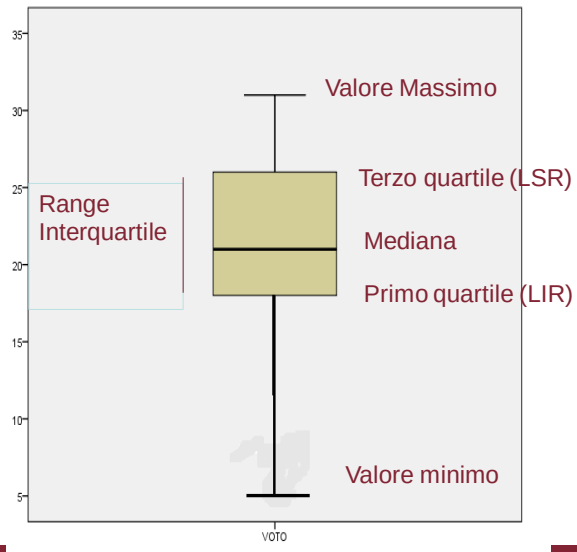
È caratterizzato da tre elementi:
1) una linea o un punto che individuano la posizione di un valore medio (la media aritmetica o la mediana) della distribuzione del carattere;
2) un rettangolo (box) la cui altezza rappresenta la variabilità (lo scarto quadratico medio oppure il range interquartile) dei valori prossimi alla media scelta
3) due segmenti che partono dai lati maggiori del rettangolo e giungono ad alcuni valori della distribuzione.

Box Plot con mediana

- ✓ Rappresentazione grafica della distribuzione di un carattere quantitativo che mette in evidenza la sua variabilità
- ✓ Elementi caratteristici
  - ✓ un punto che individua la posizione della mediana della distribuzione
  - ✓ un rettangolo (box) la cui altezza rappresenta il range interquartile
    - ✓ il limite inferiore del rettangolo (LIR) corrisponde al primo quartile,
    - ✓ il limite superiore del rettangolo (LSR) corrisponde al terzo quartile
  - ✓ 2 segmenti che partono dai lati maggiori del rettangolo e i cui estremi sono rappresentati dai valori minimo e massimo della distribuzione

BoxPlot costruito attorno alla mediana: esercizio

VOTO		
N		117
Minimo		5
Massimo		31
Moda		18 e 22
Quartili	1	18
	2	21
	3	26



Range interquartile

$$W = Q_3 - Q_1$$

- ✓ Quantifica l'estensione del 50% della distribuzione del carattere che si trova attorno alla mediana: il 50% delle unita statistiche che presentano una modalità prossima a quella centrale.
- ✓ Più ampio è il range interquartile, maggiore è la dispersione delle unità statistiche attorno alla mediana.
- ✓ E' espresso nella stessa unità di misura del carattere
- ✓ Non è influenzato dall'eventuale presenza di valori estremi o anomali assunti dal carattere nel collettivo in esame.

Esempio: Range interquartile

- Riprendendo l'esempio precedente riguardante i voti ottenuti nella prova di verifica

$$Me = 21$$

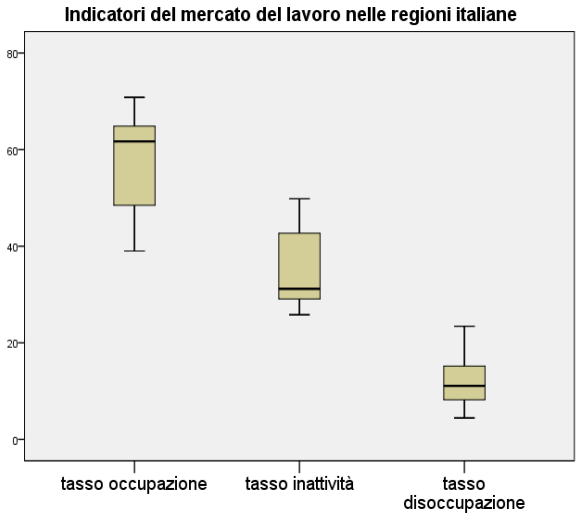
$$Q_1 = 18 \qquad Q_3 = 26$$

$$W = 26 - 18 = 8$$

Cosa vuol dire?

Vuol dire che il 50% degli studenti che hanno preso un voto che si attesta intorno al valore mediano (21) si differenziano per al massimo 8 punti.

Confronto tra tre distribuzioni: box plot



Il grafico fornisce informazioni sulle diverse distribuzioni dei tassi di occupazione, disoccupazione e inattività delle regioni italiane.
Visualizza per ciascuno di essi le mediane, i valori minimo e massimo e la dispersione attorno al valore mediano.
Consente di rispondere ai seguenti quesiti:
-quale tasso assume valori più bassi?
-quale i valori più alti?
- Per ciascun tasso possiamo stabilire il valore minimo e massimo assunto da almeno la metà delle regioni italiane?
-rispetto a quale tasso la situazione delle regioni italiane è più eterogenea? E più omogenea?

Box Plot con valori anomali

- maggiori del Valore Soglia Superiore (VSS)
- VSS= LSR + λ (LSR - LIR)
- oppure minori del Valore Soglia Inferiore (VSI)

  VSI= LIR - λ (LSR - LIR)

λ è una costante positiva, in genere posta uguale a 1,5

Se vengono individuati eventuali valori anomali allora gli estremi superiore e inferiore dei segmenti del Box Plot diventano i due valori della distribuzione più vicini ai valori anomali individuati.

Esercizio : box plot con valori anomali (1/2)

Prova di verifica : distribuzione dei voti

xi	ni	pi	Pi
5	1	.9	.9
8	1	.9	1.7
9	2	1.7	3.4
10	1	.9	4.3
11	1	.9	5.1
12	5	4.3	9.4
13	2	1.7	11.1
14	6	5.1	16.2
15	3	2.6	18.8
16	4	3.4	22.2
18	13	11.1	33.3
19	7	6.0	39.3
20	8	6.8	46.2
21	9	7.7	53.8
22	13	11.1	65.0
23	4	3.4	68.4
24	3	2.6	70.9
25	1	.9	71.8
26	4	3.4	75.2
27	6	5.1	80.3
28	4	3.4	83.8
29	10	8.5	92.3
30	3	2.6	94.9
31	6	5.1	100.0
Totale	117	100.0	

VOTO		
N		117
Minimo		5
Massimo		31
Moda		18 e 22
Quartili	1	18
	2	21
	3	26

A partire dalle caratteristiche della distribuzione della variabile Voto, costruire un box plot con la mediana, individuando eventuali valori anomali.

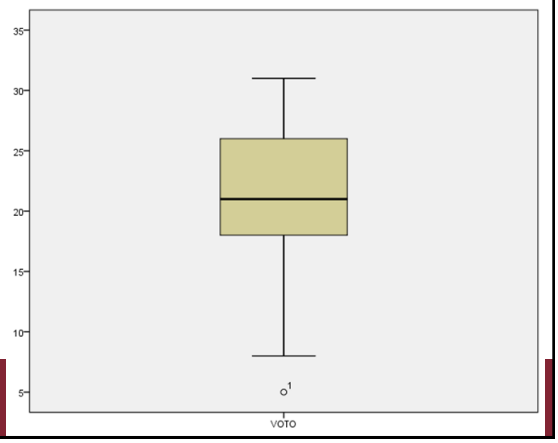
Esercizio : box plot con valori anomali (2/2)

Individuare i Valori soglia:

VSS= LSR + λ (LSR - LIR) = 26 + 1,5 (26-18) = 38

VSI= LIR - λ (LSR - LIR) = 18 - 1,5 (26-18)= 6

E' anomalo il solo valore 5

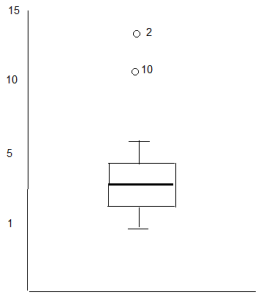


Esercizio

Sulla base della tabella seguente un cui viene riportata la distribuzione della variabile numero di cellulari per famiglia. Costruire un box plot considerando come valore medio di riferimento la mediana, controllando la presenza di valori anomali.

x_j	n_j	N_j	P_j
1	75	75	17,01
2	100	175	39,68
3	120	295	66,89
4	95	390	88,44
5	24	414	93,88
6	15	429	97,28
11	10	439	99,55
13	2	441	100,00
	441		

Me= 3 Q_1=2 Q_3=4
VSS= LSR + λ (LSR - LIR) = 4 + 1,5 (4 - 2) = 7
VSI= LIR - λ (LSR - LIR) = 2- 1,5 (4 - 2) = -1



FSSC

Box Plot con media aritmetica

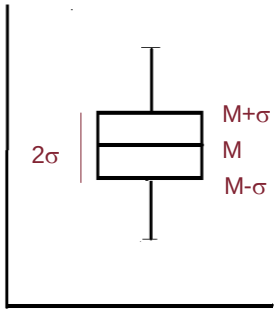
Il Box Plot può essere costruito considerando anche la media aritmetica come punto centrale del rettangolo.

In tal caso:

- ✓ Il punto o il segmento centrale è la media aritmetica
- ✓ L'estremo superiore (LSR)= $M+\sigma$
- ✓ L'estremo inferiore (LIR) =$M-\sigma$
- ✓ L'altezza box è pari a 2σ
- ✓ Gli estremi dei segmenti

  Superiore=$M+1{,}96\,\sigma$

  Inferiore=$M-1{,}96\,\sigma$



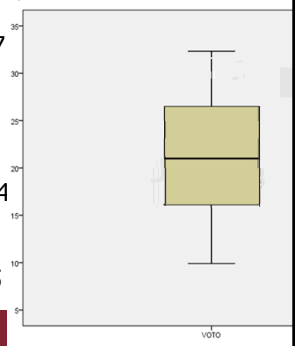
Box Plot con media aritmetica: esempio

Riprendendo l'esempio precedente riguardante la distribuzione dei voti degli studenti:
M=21,08 σ=5,98

Segmenti:
Estremo superiore= **M+1,96σ** =21,08+1,96*5,98=32,80
Estremo inferiore= **M-1,96σ** =21,08-1,96*5,98=9,36
Rettangolo:
LSR=Estremo superiore = **M+σ**=21,08+5,98= 27
LIR=Estremo inferiore= **M-σ** =21,08- 5,98= 15

Valori anomali:
Valori minori di:
VSI=M-σ-1,5*(2σ)=**M-4σ**=21,08-4*(5,98)=-2,84

Valori maggiori di :
VSS=M+σ+1,5(2σ)=**M+4σ**=21,08+4*(5,98)=45



Box Plot con mediana e media a confronto

Box Plot con mediana	**Box Plot con media aritmetica**
1) media = mediana 2) altezza box = differenza interquartile ($Q_3 - Q_1$) estremo sup. box = Q_3 estremo inf. box = Q_1 3) estremi dei segmenti superiore = valore max inferiore = valore min **Valori anomali:** Valori minori di VSI=Q_1-1,5(Q_3-Q_1) Valori maggiori di : VSS=Q_3+1,5(Q_3-Q_1)	1) media = media aritmetica (M) 2) altezza box = 2 σ estremo sup. box = $M + \sigma$ estremo inf. box = $M - \sigma$ 3) estremi dei segmenti superiore = $M + 1{,}96\sigma$ inferiore = $M - 1{,}96\sigma$ **Valori anomali:** Valori minori di VSI=M-σ-1,5*(2σ)=M-4σ Valori maggiori di : VSS=M+σ+1,5*(2σ)=M+4σ

FSC