I Percentili, i Quartili,
il grafico Box-plot

Base concettuale:

Percentili: sono quei valori della distribuzione in grado di dividere la distribuzione in 100 parti uguali.

La mediana divide la distribuzione in due parti uguali => ottenendo ognuna il 50% della unità. Infatti, la mediana nella divisione percentile si posiziona (come già accennato) al 50-esimo percentile.

I percentili più frequenti sono il 25-esimo percentile e il 75-esimo percentile (chiamati anche primo quartile e terzo quartile) i quali insieme alla mediana dividono la distribuzione 4 parti uguali.

Il primo (Q1) e il terzo (Q3) quartile individuano un intervallo centrale che contiene circa il 50 % delle unità statistiche e che può essere considerato come misura di depressione dei valori più frequenti della popolazione/collettivo osservato.

Quartili: i quartili invece dividono la distribuzione i 4 parti uguali:

1° Quartile (**Q1**): 25% più piccolo, 75% più grande.

2° Quartile (**Q2**): 50% più piccolo, 50% più grande. Il secondo quartile rappresenta anche il punto centrale di una distribuzione = **mediana.**

3° Quartile (**Q3**): 75% più piccolo, 25% più grande.

Il calcolo dei quartili si effettua (come nella mediana) ordinando la distribuzione, si individua la posizione e si osserva la modalità presentata

Esempio (quartili)

Supponiamo di osservare la seguente distribuzione delle famiglie per numero di figli:

(nr figli) xi	**fi**	**Fi**
0	0,15	0,15
1	0,25	0,40
2	0,30	0,70
3	0,20	0,90
4	0,06	0,96
5	0,03	0,99
6	0,01	1
Totale	**1**	

Guardando la distribuzione di frequenze relative cumulate riusciamo ad affermare che:

- La mediana corrisponde al valore 2
- Il primo quartile corrisponde al valore 1
- Il terzo quartile corrisponde al valore 3
  Infatti, nel primo caso il 40% delle famiglie non ha più di un figlio (solo il 15 % delle famiglie non ha figli)
  Nel secondo caso il 90% delle famiglie non ha più di 3 figli
  Nel terzo caso il 70% non ha più di due figli

Attenzione

Nel caso in cui la distribuzione di frequenze suddivise in classi o intervalli non è possibile trovare l'esatto valore del quartile, ma, come per la mediana, possiamo avvalerci di una sua approssimazione. Consideriamo per esempio il primo quartile (il 25-esimo percentile) che è il valore a sinistra nel quale cade il 25% delle unità. in questo caso la formula è data da:

$$Q_1 \approx l_{Q_1} + \left(\frac{0,25 - F_{Q_1 - 1}}{F_{Q_1} - F_{Q_1 - 1}} \right) \Delta_{Q_1}$$

Dove:

l_Q1 => è l'estremo inferiore della classe dell'intervallo

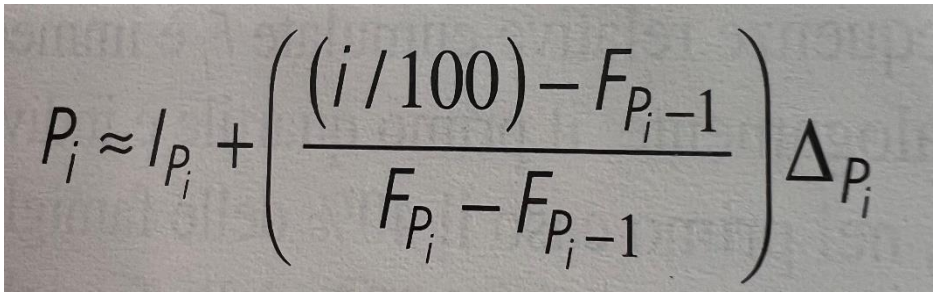
F_Q1-1 => è la frequenza relativa cumulata fino alla classe precedente a quella in cui cade il primo quartile

F_Q1 => è la frequenza relativa cumulata fino alla classe che contiene il primo quartile

Δ**_Q1** => è l'ampiezza della classe che contiene il primo quartile.

Per calcolare il secondo e il terzo quartile, basta sostituire i valori nella formula: **Q1=>Q2=>Q3.**

La formula per trovare invece l'approssimazione generico percentile è dato da:



Dove:

l_Pi => è l'estremo inferiore della classe dove cade l'i-esimo percentile.

F_Pi => è la frequenza relativa cumulata fino alla classe percentile a quella in cui cade l'i_esimo percentile.

F_Pi => è la frequenza relativa cumulata fino alla classe che contiene l'i_esimo percentile.

Δ**_Pi** => è l'ampiezza della classe che contiene l'i-esimo percentile.

Esempio

Calcolo quartile per un carattere suddiviso in classi?

Data la distribuzione di frequenze relative e cumulate delle viti di acciaio prodotte da una industria meccanica secondo alcune classi di diametro espresse in millimetri:

(diametro mm) xi	fi	Fi
0-3	0,10	0,10
3-5	0,18	0,28
5-8	0,25	0,53
8-10	0,12	0,65
10-20	0,20	0,85
20-35	0,15	1
Totale	**1**	

Box-plot

Abbiamo visto che le medie e gli indici di variabilità per descrivere ed analizzare alcune caratteristiche sulla distribuzione dei dati. Adesso andiamo a vedere una rappresentazione grafica che si avvale di tali misure e che risulta ad essere estremamente maneggevole nella comparazione di due o tre collettivi.

[Il diagramma a ramo-e-foglia e l'istogramma danno una visione generale (qualitativa) di un insieme di dati. Singoli valori numerici come la media, la varianza o i quartili forniscono informazioni puntuali (quantitative) su uno specifico aspetto del campione.]

Il box plot di una distribuzione è un grafico caratterizzato da tre elementi principali:

- Una linea o un punto che indicano la posizione della media 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.

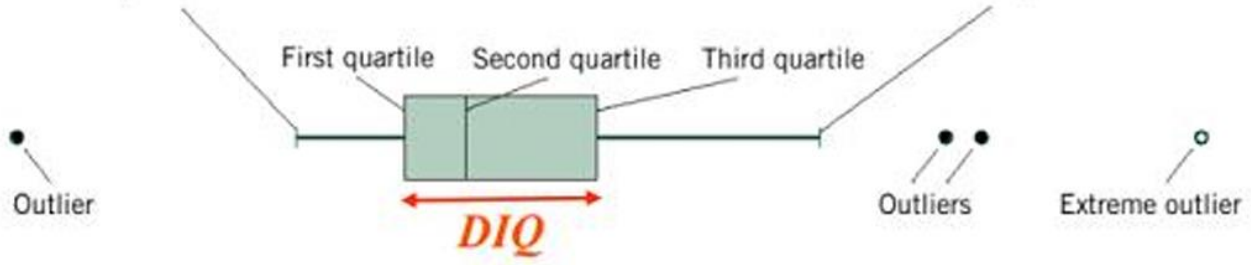
Un indicatore grafico (visione d'insieme) che descrive anche e contemporaneamente diverse importanti caratteristiche quantitative del campione è il box plot: **è una scatola**

delimitata dai quartili, e "tagliata" dalla mediana, che riporta anche due baffi estesi fino a "circa" 1.5 volte il range interquartile (DIQ=Q3-Q1).

Eventuali punti esterni ai baffi vengono riportati singolarmente (outliers). Un punto più lontano di 3 range interquartili, dal quartile corrispondente, è detto outlier estremo.

Estensione di 1.5 range dal "baffo" del primo quartile

Estensione di 1.5 range dal "baffo" del terzo quartile



Inter Quartile Range IQR = DIQ : delimita la "scatola" (box) che contiene il 50%, centrale, dei dati

DIQ= Q3-Q1 (Range interquartile)

WL,lim = Q1-1.5 DIQ limite inferiore per il baffo basso

WH,lim = Q3+1.5 DIQ limite superiore per il baffo alto

La Mediana : riguarda la linea al centro della scatola che rappresenta il punto centrale della distribuzione. Il valore centrale della distribuzione dei dati è utile in presenza di molti outliers in quanto sintetizza meglio il fenomeno rispetto ad una media.

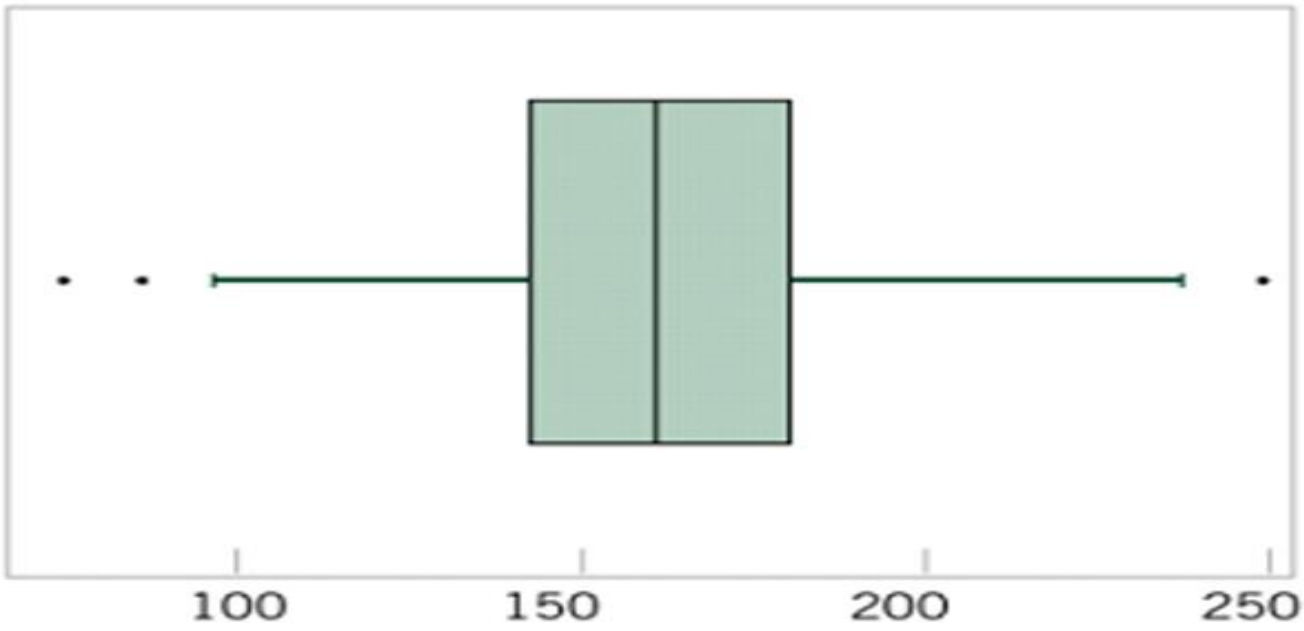
La Variabilità: osserva l'altezza della scatola e la lunghezza dei "baffi". Dimensioni maggiori corrispondono a una maggiore dispersione della variabile rispetto al valore mediano.

Asimmetria: si osserva nel caso in cui uno dei due "baffi" è più lungo rispetto all'altro, una tendenza dei dati a disperdersi verso valori più grandi o più piccoli rispetto a quello centrale. In particolare, se il baffo inferiore è più pronunciato si ha una asimmetria sinistra, ossia i valori più piccoli della variabile sono più dispersi; viceversa, se il baffo superiore è più lungo si avrà una asimmetria destra e quindi i valori più dispersi saranno quelli più alti.

Utilità dei box-plot?

- I box-plot sono molto utili per il confronto diretto (visivo) di dati provenienti da campioni diversi.
- Il box plot è un indicatore grafico che fornisce importanti informazioni quantitative su un insieme di dati: **Posizione e tendenza centrale**; **Variabilità e dispersione**; **Simmetria o asimmetria; Identificazione dei punti esterni**

Esempio esposizione.

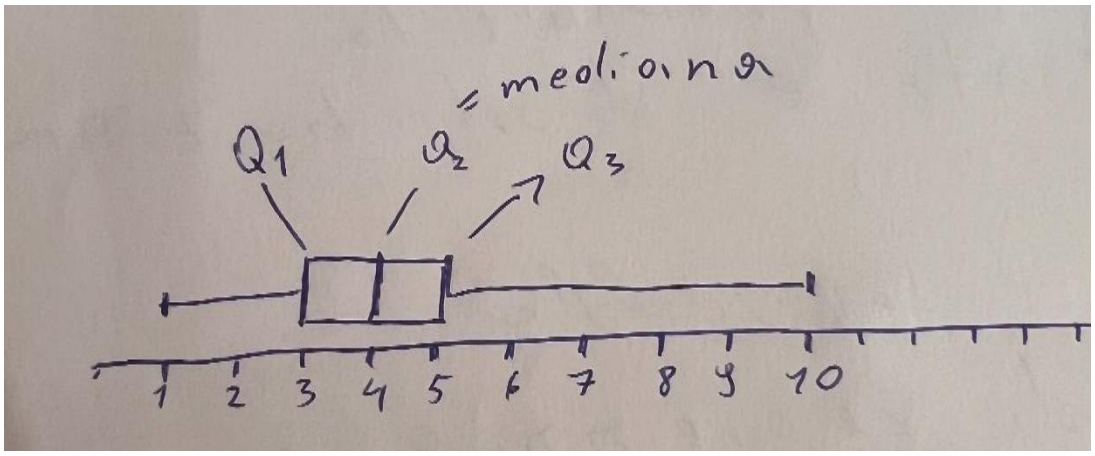


Il box plot è un indicatore grafico che fornisce importanti informazioni quantitative su un insieme di dati: – Posizione e tendenza centrale – Variabilità e dispersione – Simmetria o asimmetria – Identificazione dei punti esterni

Esempio: Box-plot

x_i	n_i	f_i	F_i
1	3	0,02	0,02
2	8	0,06	0,08
3	30	0,22	0,30
4	45	0,33	0,63
5	22	0,16	0,79
6	12	0,09	0,88
7	10	0,07	0,94
8	5	0,04	0,98
9	2	0,01	0,99
10	1	0,01	1
Totale	138		

Q1=3 Q2=4 Q3=5



Esempio: Box-plot in presenza di outliers

xi	ni	Fi
1	7	0,30
2	5	0,52
3	8	0,87
10	3	1
totale	23	

$Q_1 = 1$

$Q_2 = Me = 2$

$Q_3 = 3$

$DI = Q_3 - Q_1 = 2$

$Pt\ Inf. = Q_1 - 1.5 \cdot DI = 1 - 1.5 \cdot 2 = -2$ $\quad L_1 = max(x_{min}, Pt\ Inf.) = max(1, -2) = 1$

$Pt\ Sup. = Q_3 + 1.5 \cdot DI = 3 + 1.5 \cdot 2 = 6$ $\quad L_2 = min(x_{max}, Pt\ Sup.) = min(10, 6) = 6$