Statistica descrittiva

Caso di 1 variabile: i dati si presentano in una tabella:

Nome soggetto	Dato
Alabama	11.6
……..	……

Per riassumere i dati si costruisce una distribuzione delle frequenze.

- Si determina il massimo M e il minimo m dei dati, quindi tutti i dati sono contenuti nell'intervallo [m,M] (range dei valori).
- Poi si suddivide [m,M] in intervalli (classi) disgiunti e si contano quanti dati cadono in ciascun intervallo, (frequenza della classe).
- Si costruisce la tabella:

Classe	Frequenza
0.0-2.9	5

Distribuzione di frequenze

È una lista di intervalli di possibili valori di una variabile insieme con una tabulazione del numero di osservazioni per ciascun intervallo (frequenza assoluta).

Le classi sono di solito scelte di uguale ampiezza e ogni dato deve cadere in una ed una sola classe.

frequenza relativa= frequenza assoluta diviso la dimensione del campione.

Frequenza relativa di una classe

è la percentuale di dati che stanno in quella classe.

La classe 0.0-2.9 contiene 5 dati su 50: quindi la frequenza relativa è 5/50=0.10=10%.

La somma delle frequenze relative = 1.

La somma delle percentuali = 100%.

Attenzione agli arrotondamenti!

Istogramma= grafico della distribuzione delle frequenze

Istogramma

Per costruire l'istogramma delle frequenze sono necessari i passi seguenti:

- Divisione del range di variazione dei dati in sottointervalli (classi) di uguale ampiezza
- Etichettatura dell'asse orizzontale con i sottointervalli selezionati;
- Determinazione delle frequenze relativa di ogni classe;
- Etichettatura dell'asse verticale con i valori di frequenza individuati;
- Disegno delle frequenze nel piano definito.

La scelta dell'ampiezza dei sottointervalli è cruciale infatti:

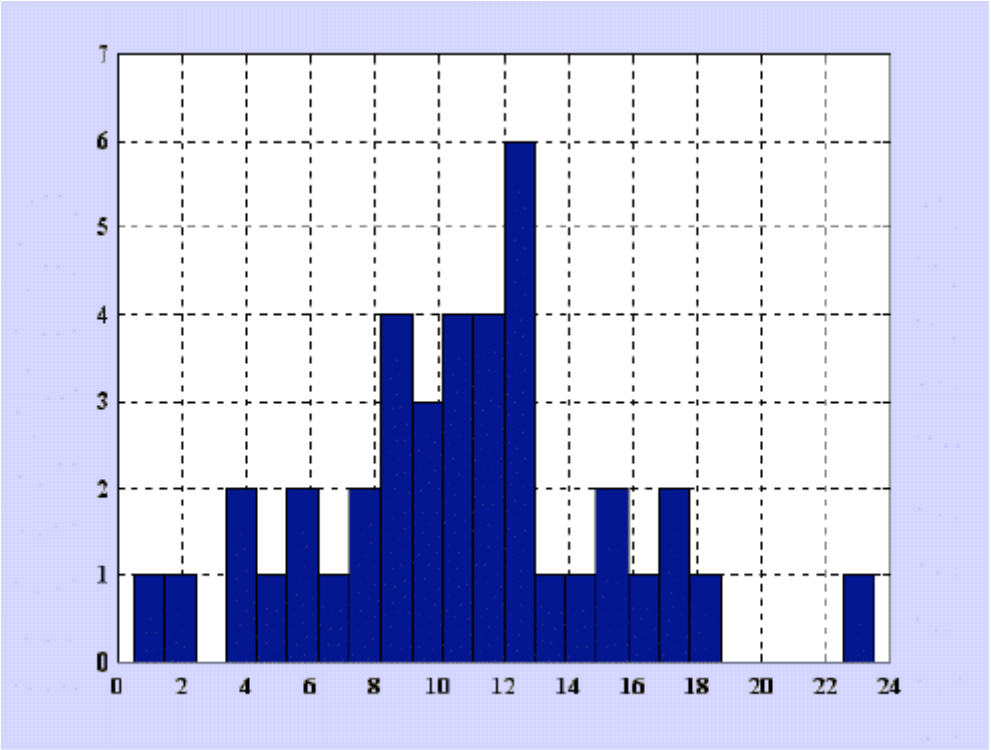
se l'intervallo è troppo ampio, l'istogramma risulta troppo aggregato e non consente di individuare una funzione di densità di probabilità;

se l'intervallo è troppo piccolo, l'istogramma evidenzia troppo eventuali picchi negativi e positivi risultando troppo "brusco".

Esempio: si vuole disegnare l'istogramma delle frequenze dei dati relativi alla durata dei viaggi aerei di un insieme di passeggeri.

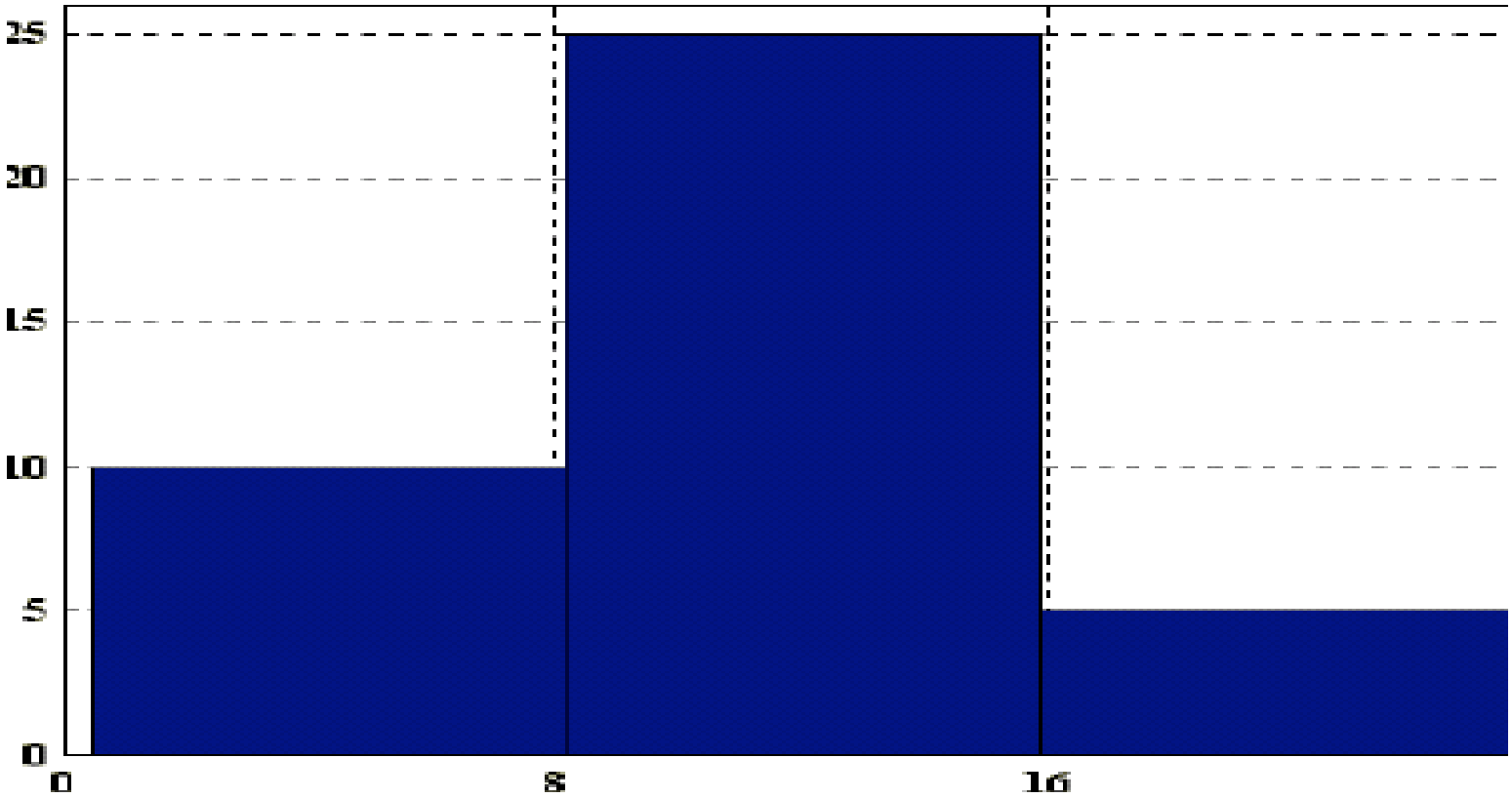
I campioni raccolti sono (unità di misura=ora):

0.5, 2.1, 3.4, 4.1, 4.6, 5.7, 6.2, 6.6, 7.8, 8.1,
8.3, 8.4, 8.6, 8.9, 9.2, 9.8, 10.0, 10.3, 10.5,
10.6,10.8, 11.2, 11.3, 11.6, 11.7, 12.1, 12.5,
12.6, 12.8, 12.9, 12.9, 13.2, 14.4, 15.0, 15.5,
16.3,17.0, 17.3, 18.5, 23.5.

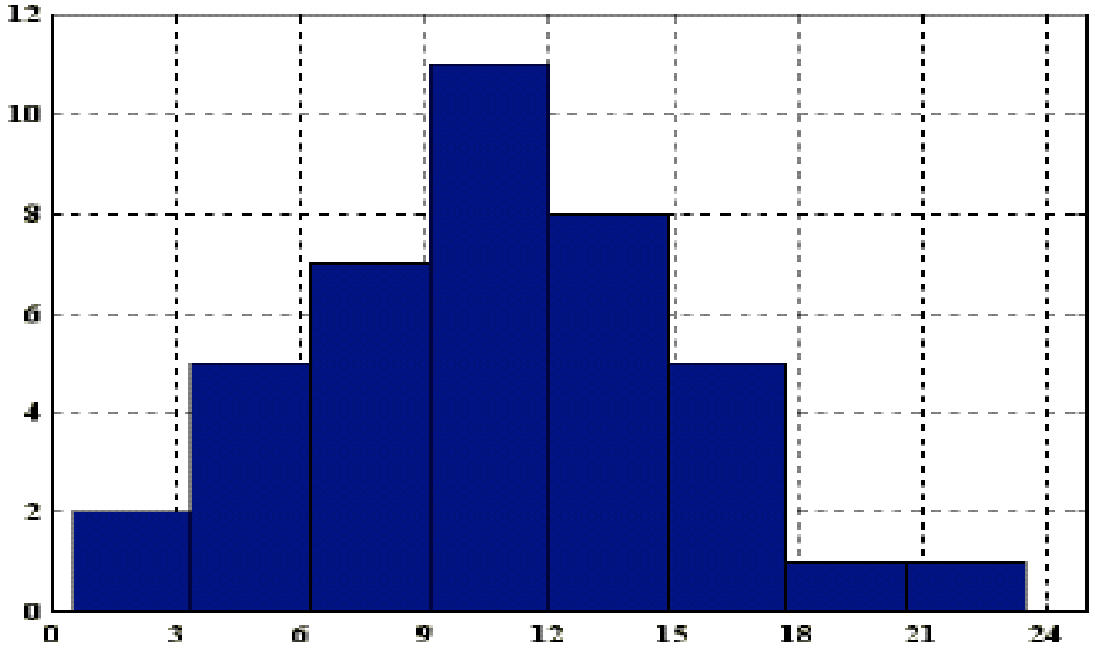


L’ampiezza dei sottointervalli è troppo piccola

Intervalli di 8 ore



Intervalli di 3 ore



La statistica descrittiva mette a disposizione il calcolo di indicatori sintetici che individuano, con un singolo valore, proprietà statistiche di un campione rispetto ad una sua variabile.

In particolare:

- indicatori di centralità: media aritmetica, moda, mediana;
- indicatori di variabilità: varianza, deviazione standard;
- misure di raggruppamento: quartili, percentili.

Mediana

E’ necessario che la lista dei dati sia ordinata in ordine crescente o decrescente:

la mediana è un numero *x* tale che almeno il 50% dei dati è minore di *x* e il 50% è maggiore di *x*.

E’ robusta: cioè, poco influenzata dalla presenza di dati anomali

7 12 18 23 34 54

Si può calcolare anche nel caso di dati ordinali.

Media

È adatta per variabili quantitative ed è influenzata da dati anomali, (outliers).

$$mean = \frac{\sum_{i=1}^{n} X_i}{n}$$

Esempio: Abbiamo i 6 dati

7 12 18 23 34 54

La loro media è….

Moda

La **moda** è il valore con frequenza più alta nell'insieme delle osservazioni.

E' adatta a tutti i tipi di dati.

Non è posta necessariamente vicino al centro della distribuzione e perciò non è propriamente una misura della tendenza centrale.

Nel caso di variabili che esprimono una attitudine

si presentano spesso distribuzioni bimodali.

Esempi

• In un negozio lavorano 7 impiegati, la media dei loro stipendi è $37900 e la mediana $11200. Cosa significa?

•

Media inf	38012
Media sup	65291
Laurea I	33191
Laurea II	7570
Dottorato	1311
Specializz azione	3110
Master	7599

Insieme ordinato di categorie.

Quale è la dimensione del campione?

Quale è la classe mediana?

Quartili e percentili

Il primo quartile è un numero tale che il 25% dei dati è inferiore e il 75% è superiore.

Il terzo quartile è un numero tale che il 75% dei dati è inferiore e il 25% è superiore.

I quartili insieme alla mediana dividono i dati in 4 parti ciascuna contenente un quarto dei dati.

La differenza fra il primo e il terzo quartile si dice distanza interquartilica.

Boxplot

- Rappresentano
  - il grado di dispersione o variabilità dei dati (w.r.t. mediana e/o media)
  - la simmetria
  - la presenza di valori anomali
- Le distanze tra i quartili definiscono la dispersione dei dati