Statistica descrittiva

- La statistica descrittiva mette a disposizione il calcolo di indicatori sintetici che individuano, con un singolo valore, proprieta` statistiche di un campione/popolazione rispetto ad una sua variabile/attributo.
- In particolare:
  - indicatori di centralita`: media aritmetica, moda, mediana;
  - indicatori di variabilita`: varianza, deviazione standard;
  - misure di raggruppamento: quartili, percentili.

Misure di centralita`: media aritmetica

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

Media aritmetica in EXCEL

- **AVERAGE**(**number1**,number2,...)
  - Number1, number2, ... sono da 1 a 30 argomenti di cui calcolare la media aritmetica
  - Gli argomenti debbono essere o numeri o nomi, array, riferimenti a campi che contengono numeri.
  - se un argomento di tipo array o riferimento contiene testo, valori logici o celle vuote, questi valori sono ignorati; sono invece conteggiati gli zero presenti.

Media aritmetica: proprieta`

- Per effettuare la correzione di errori accidentali
  - permette di sostituire i valori di ogni elemento senza cambiare il totale

$$\frac{1}{n+k}\left(\sum_{i=1}^{n} x_i + k\bar{x}\right) = \bar{x}$$

Misure di centralita`: mediana

- Dato un campione di valori *ordinati*

  $\{X_1, X_1, \ldots, X_n\}$

- MEDIAN = $X_{((n+1)\ \%\ 2)}$
- la mediana e` quindi il valore centrale del campione.
- In EXCEL: **MEDIAN(number1,number2,...)**

Mediana: proprieta`

- Robusta: poco influenzata dalla presenza di dati anomali

$$7\ 12\ 18\ 23\ 34\ 54$$

$$\bar{x} = 21.3$$

$$M = 23$$

Misure di centralita`: moda

- la **moda** e` il valore con frequenza piu` alta nell'insieme delle osservazioni

- In EXCEL

MODE(number1,number2,...)

Varianza, deviazione standard

- misure di mutua variabilità tra i dati di una serie
- Devianza empirica

$$dev = \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2$$

- Varianza

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2$$

Misure di variabilita`: deviazione standard

$$\text{deviazione standard} = \sqrt{\text{varianza}}$$

In EXCEL:

- STDEV(range)
- VAR(range)

Quartili e percentili

- Dato un campione ordinato il percentile ennesimo e` il valore che separa n% dei dati dal resto
- p.e. esempio la mediana puo` essere interpretata come il 50-esimo percentile
- In Excel

  PERCENTILE(range, perc)

- Il Quartile separa un quarto dei dati dal resto.
- Si parla di primo (25%), secondo (50%), terzo (75%) quartile
- In Excel

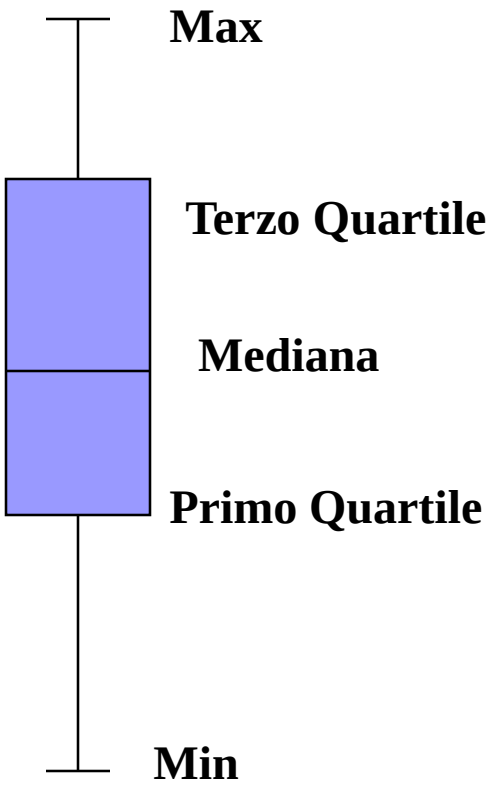
  QUARTILE(range,n) dove n puo` essere 1, 2 o 3

Altre misure descrittive

- MIN(range) calcola il valore minimo di un campione memorizzato in range
- MAX(range) calcola il valore massimo di un campione
- RANGE(range) calcola la differenza tra il valore minimo e massimo, ovvero il campo di variabilita` dei valori del campione.

Rappresentazioni 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
- non direttamente disponibile in Excel



Distribuzioni di frequenza

Intepretazione dei dati mediante la distribuzione dei valori degli attributi

- Nello studio di un campione di osservazioni in cui alcune variabili sono di tipo categorico o categorizzabile, puo` essere molto informativo vedere come le osservazioni si distribuiscono sulle categorie;

- Una *tabella di frequenze* riporta il numero di osservazioni che ricadono in ciascuna delle categorie stabilite;

- Un *istogramma* e` una tecnica di visualizzazione di una tabella di frequenza tramite un *diagramma a barre.*

Creazione di una *tabella di frequenze (1)*

1. Si crei una tabella Pivot che abbia l'attributo come dimensione e come fatto

1. Si usi come misura del fatto il conteggio ottenendo la tabella delle frequenze

1. Si usi il commando di generazione dell'istogramma per ottenere la visualizzazione della distribuzione di frequenza

Esercitazione

- La fabbrica di ascensori OTIS ha misurato il diametro dei cavi da ascensore prodotti (file OTIS1).

  Generare la distribuzione (scegliendo opportunamente le categorie) e l'istogramma relativo e discuterne la forma

- Il file BANK elenca l'intervallo di tempo che separa l'arrivo dei clienti in banca in un dato giorno. Se ne calcoli la distribuzione (scegliendo opportunamente le categorie) e l'istogramma relativo e se ne discuta la forma.

Distribuzioni di frequenza teoriche

- La rappresentazione con istogramma delle frequenze dei valori di una variabile numerica su un intervallo partizionato in categorie continue e` la rappresentazione di una *distribuzione di frequenza*

- Può essere interpretata come una approssimazione del grafico di una funzione che descrive come varia la frequenza di un valore rispetto al crescere (decrescere) del valore stesso

- Le distribuzioni di frequenza osservate spesso sono così regolari da poter essere viste come casi particolari di *distribuzioni di frequenze teoriche*, descritte da ben precise funzioni matematiche

Distribuzioni di frequenza teoriche

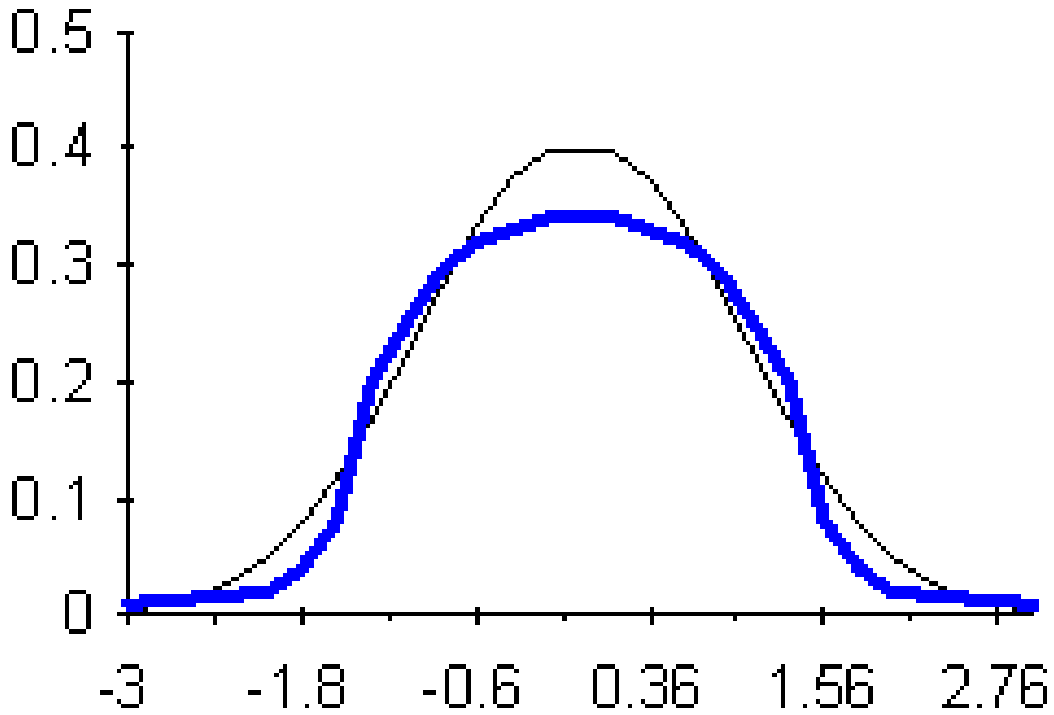
- Distribuzione uniforme: ogni evento (categoria) ha la stessa frequenza
- Distribuzione normale (o Gaussiana, o a campana)

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(y-\mu)^2/2\sigma^2}$$

$$\mu = \text{media}$$

$$\sigma = \text{deviazione standard}$$

Esempio di gaussiana



Simmetria

- Si ha simmetria quando media, moda e mediana coincidono
  - condizione necessaria, non sufficiente
  - Asimmetria sinistra: moda, mediana, media
  - Asimmetria destra: media, mediana, moda

Misure della forma della distribuzione: skewness

- Indice di asimmetria (skewness): applicato ai dati misura la mancanza di simmetria della distribuzione di frequenza:
  - Un valore positivo indica una distribuzione in cui i valori sono raggruppati nel range dei valori bassi con una lunga coda che si estende verso I valori maggiori
  - Un valore negativo indica la situazione opposta
- Esempio: calcolare il valore di asimmetria per la distribuzione dei tempi di interarrivo in banca (BANK.XML)

Misure della forma della distribuzione: curtosi

- L’indice di curtosi misura la “pesantezza” delle code di una distribuzione dei dati:
  - Una curtosi positiva indica che ci sono piu` valori agli estremi della distribuzione di quanto aspettato
  - Una curtosi negativa indica ci sono meno avlori agli estremi della distribuzione di quanto aspettato.

Curtosi (cont.)

- Coefficiente di curtosi
  - Una distribuzione leptocurtica ha K ~ 1/2
  - platicurtosi: k~0

$$K = \frac{\frac{1}{2}\left(x_{.75} - x_{.25}\right)}{\left(x_{.90} - x_{.10}\right)}$$

Esercitazione

- Usando il file ACTORS.XLS
  1. Calcolare la distribuzione su salary
  2. Costruire un istogramma diviso per categorie tenendo conto del sesso

- Usando il file HOMEDATA.XLS
  1. Calcolare la distribuzione dei prezzi delle case
  2. Costruire l'istogramma della distribuzionetenendo conto della posizione (NE_sector)
  3. Calcolare gli indici di asimmetria e di curtosi
  4. Calcolare i quartili