Esplorazione dei dati

Lucidi e dataset tratti da
Turini - *Analisi dei Dati,* Dip. Inf. Unipi

Analisi mono e bivariata

- Si utilizzano indicatori sintetici che individuano, con un singolo valore, proprieta` statistiche di un campione della popolazione rispetto ad una sua variabile/attributo.
- Abbiamo fin qui visto:
  - indicatori di centralità: media aritmetica, moda, mediana;
  - Indicatori di dispersione: range, deviazione media, MAD;
  - indicatori di variabilita`: varianza, deviazione standard.

- Oggi vedremo:
  - indicatori di posizionamento: quartili, z-indice.;
  - indicatori di di asimmetria e curtosi;
  - diagrammi scatter, covarianza e correlazione.

Identificazione degli oulier

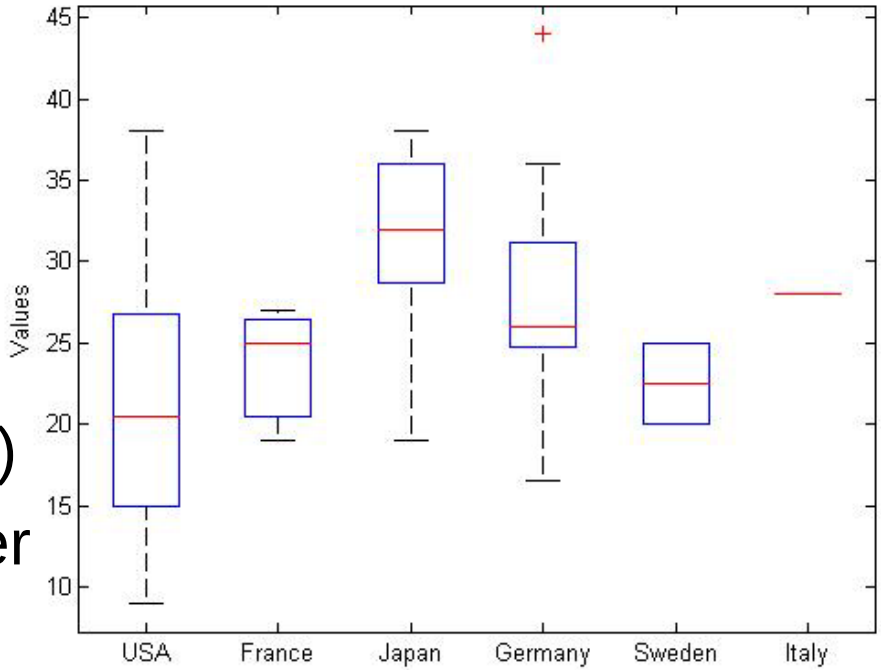
- Il teorema di Tchebysheff ha una semplice ed immediata applicazione per identificare gli outlier.
- Definiamo lo *z-indice* di una generica osservazione x_i:

$$z_i^{ind} = \frac{x_i - \bar{\mu}}{\bar{\sigma}}$$

- Possiamo ritenere x_i sospetto outlier $\Leftrightarrow |z_i^{ind}| > 3$

Diagrammi box&whiskers

- Il box si estende tra i due quartili inferiore e superiore.
- La linea centrale è la mediana
- I baffi sono il minimo e massimo valore in $(q_L - 1.5D_q, q_U + 1.5D_q)$
  - $D_q = q_U - q_L$ (dispersione dei dati)
- I valori esterni ai baffi sono outlier
  - Disponibile nello statistics toolbox del Matlab

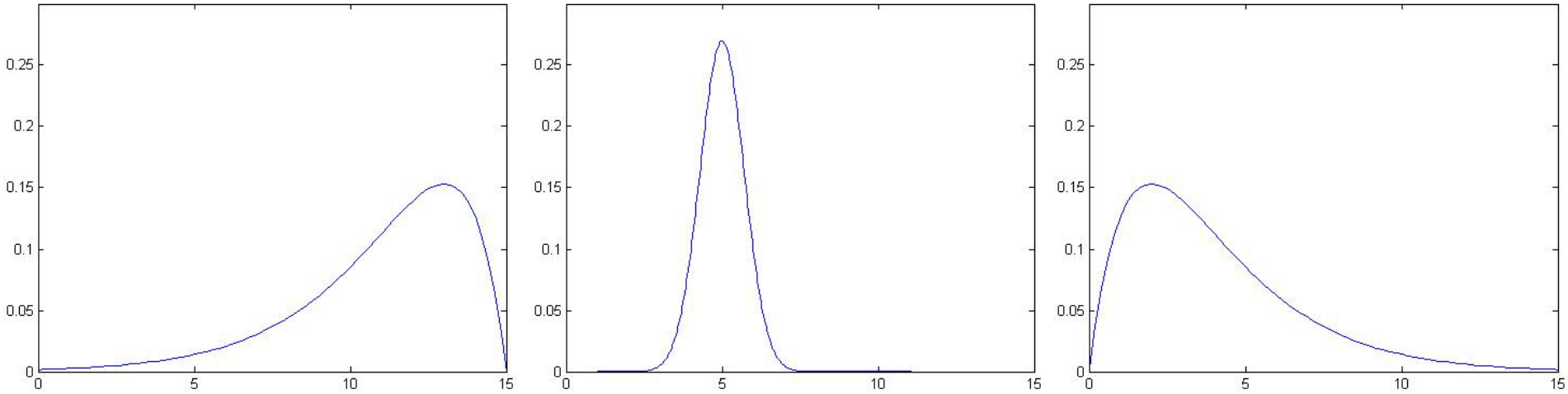



```
>> load carsmall
>> boxplot(MPG, Origin)
```

Asimmetria della curva di densità

- *Momento terzo campionario*: $\bar{\mu}_3 = \frac{1}{m} \sum_{i=1}^{m} (x_i - \bar{\mu})^3$

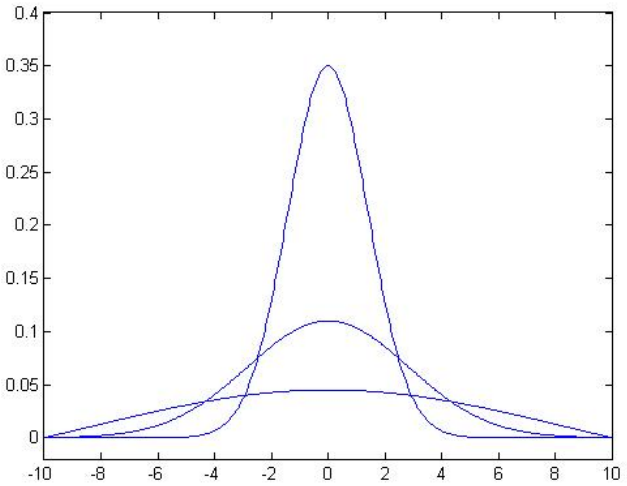
- *Indice di asimmetria* (*skewness*): $I_{as} = \frac{\bar{\mu}_3}{\bar{\sigma}^3}$
  - Misura la mancanza di simmetria



Curtosi della curva di densità

- *Momento quarto campionario*: $\bar{\mu}_4 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \bar{\mu})^4$
- *Indice di curtosi*: $I_{curt} = \frac{\bar{\mu}_4}{\bar{\sigma}^2} - 3$
  - In Matlab, kurtosis()

- L'indice di curtosi misura la "pesantezza" delle code :
  - I_{curt}>0 : meno valori agli estremi di quanto aspettato
    - Minore dispersione
  - I_{curt}<0 : più valori agli estremi di quanto aspettato.
    - Maggiore dispersione



Analisi bivariata

- L'*analisi di regressione* permette di esplorare le relazioni tra due insiemi di valori (es. i valori di due attributi di un campione) alla ricerca di associazioni.

- Per esempio possiamo usare l'analisi di regressione per determinare se:
  - le spese in pubblicità sono associate con le vendite
  - il fumo è associato con le malattie cardiache
  - la dieta mediterranea è associata con la durata della vita

Scatter plots

- Un primo approccio all'analisi di regressione è la creazione di uno scatter plot, che mostra su un piano XY un punto per ogni coppia di valori
  - In Matlab, scatter().

- Per esempio se abbiamo un campione che riporta per ciascuna famiglia le entrate mensili, le spese per attività culturali, le spese per attività sportive ecc., possiamo creare uno scatter plot che usa le coppie entrate-spese culturali per indagare l'esistenza di una relazione

Esempio: entrate e spese familiari

- A partire dal file
  http://www.di.unipi.it/~turini/Analisi%20dei%20dati/dati/Expenses.xls
  trovare:
  - l'associazione tra entrate e spese per cultura
  - l'associazione tra entrate e spese per sport
  - l'associazione tra spese per sport e spese per cultura

Covarianza

- La *covarianza* quantifica la forza della relazione tra due insiemi di valori, ovvero misura quanto lineare è la dipendenza tra i due attributi;
- La covarianza è la media del prodotto delle deviazioni dei valori dalla media degli insiemi dei dati

$$v_{jk} = cov(a_j, a_k) = \frac{1}{m-2} \sum_{i=1}^{m} (x_{ij} - \bar{\mu}_j)(x_{ik} - \bar{\mu}_k)$$

  - In Matlab cov()
- un valore positivo indica una variazione di X e Y nella stessa direzione, un valore negativo l'opposto

Correlazione

- Un limite della covarianza è la sua dipendenza dall'unità di misura.
- Per esempio possiamo aumentare il fattore covarianza di 1000, semplicemente usando come unità di misura € in luogo di K€
  - Nel caso le unità siano appropriate
- La misura di *correlazione* risolve il problema producendo un risultato indipendente dalle unità di misura e compreso tra –1 e 1

$$r_{jk} = corr(a_j, a_k) = \frac{v_{jk}}{\bar{\sigma}_j \bar{\sigma}_k}$$

Correlazione

- Un valore della correlazione è vicino a –1 indica che i due insiemi di valori tendono a variare in senso opposto
- Un valore della correlazione vicino a +1 indica che i due insiemi di valori tendono a variare nello stesso senso
- Una indipendenza nelle variazioni dei due valori produce un indice di correlazione uguale a 0
  - L'indice di correlazione è rilevante solo per relazioni *lineari*
  - L'indice può risultare vicino a 0 anche se esiste una relazione non lineare tra i due insiemi di valori.

Serie temporali

- Una serie temporale è una tabella in cui una delle variabili è una variabile che assume valori su una scala temporale in modo regolare, ovvero a intervalli fissi.
- Una serie temporale può essere rappresentata con uno scatter plot con il tempo sull'asse orizzontale e la variabile di cui studiare l'andamento sull'asse verticale.
- Osservando la serie temporale è possibile rispondere a domande come:
  - i dati hanno un andamento regolare?
  - ci sono schemi ricorrenti (es. le vendite hanno un andamento stagionale?)
- Esempio: http://www.di.unipi.it/~turini/Analisi%20dei%20dati/dati/Toys.xls

Esercitazione (1)

- Esempio: calcolare il valore di asimmetria per la distribuzione dei tempi di interarrivo in banca (http://www.di.unipi.it/~turini/Analisi%20dei%20dati/dati/Bank.xls)
- Usando il file http://www.di.unipi.it/~turini/Analisi%20dei%20dati/dati/Actors.xls
  1. Calcolare la distribuzione su salary
  2. Costruire un istogramma diviso per categorie tenendo conto del sesso

- Usando il file http://www.di.unipi.it/~turini/Analisi%20dei%20dati/dati/HomeData.xls
  1. Calcolare la distribuzione dei prezzi delle case
  2. Costruire l'istogramma della distribuzione tenendo conto della posizione (NE_sector)
  3. Calcolare gli indici di asimmetria e di curtosi
  4. Calcolare i quartili

Esercitazione (2)

- Il responsabile del personale della Beta Tecnologies Inc. sta cercando di individuare la variabile che meglio spiega le variazioni di stipendio degli impiegati usando un campione che riporta i dati di 52 impiegati a tempo pieno.
- I dati sono nel file http://www.di.unipi.it/~turini/Analisi%20dei%20dati/dati/IMPIEGATI.xls.
- Si generino diagrammi XY per determinare quale delle seguenti variabili ha la relazione lineare *più forte* con lo stipendio annuale:
  - sesso
  - età
  - numero di anni di esperienza lavorativa prima dell'assunzione in azienda
  - numero di anni di impiego in azienda
  - numero di anni di educazione post-secondaria.

Esercitazione (3)

- Usando il file
  http://www.di.unipi.it/~turini/Analisi%20dei%20dati/dati/BEER.XLS

  che riporta i dati di produzione negli anni 1980-1991 di una anonima fabbrica di birra

  Studiare l'andamento temporale della produzione, sia a livello annuale che a livello mensile.