Esplorazione dei dati

Introduzione

- L'analisi esplorativa dei dati evidenzia, tramite grafici ed indicatori sintetici, le caratteristiche di ciascun attributo presente in un dataset.
- Il processo di esplorazione consiste di tre fasi:
  - **Analisi univariata**: analisi degli attributi singoli,
  - **Analisi bivariata**: analisi delle coppie di attributi
  - **Analisi multivariata**: legami tra un sottoinsieme di attributi

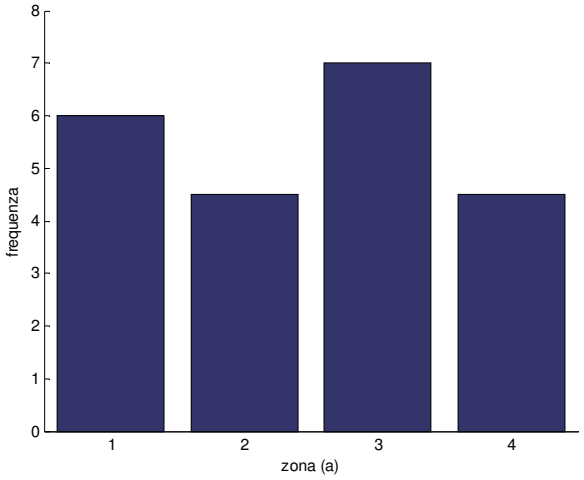
Analisi univariata

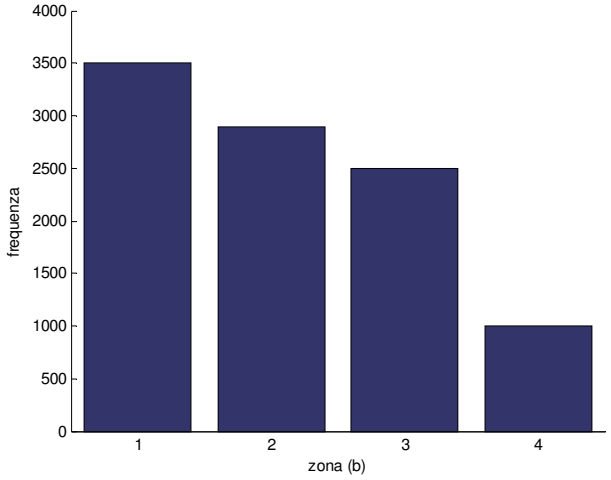
- Essa viene condotta valutando la tendenza dei valori dei singoli attributi a:
  - collocarsi in prossimità di un valore centrale (*posizionamento*),
  - assumere un certo range di valori (*dispersione*),
  - distribuirsi in maniera intelligibile.
- Obiettivi:
  - Verificare ipotesi statistiche
    - Un attributo che assume lo stesso valore nel 95% delle osservazioni può non fornire informazioni utili
  - Evidenziare anomalie e valori fuori scala.

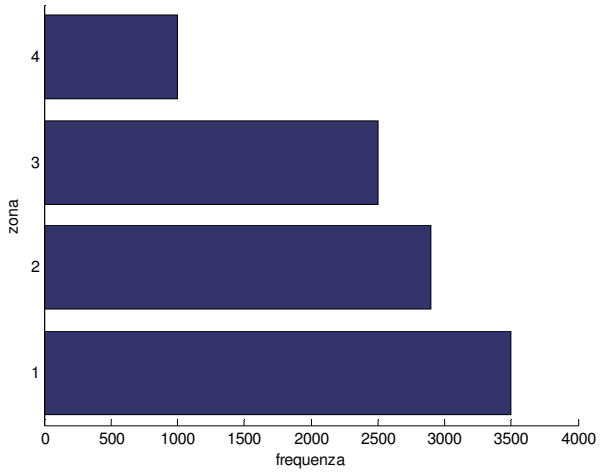
Attributi categorici

- Rappresentazione della ***frequenza empirica*** con cui i diversi valori $V=\{v_1, v_2, \cdots, v_H\}$ vengono assunti:

$$e_h=card\{i \in M : x_i=v_h\}, \quad h \in H$$







Attributi categorici

- Rappresentazione della ***frequenza empirica relativa***

$$f_h = \frac{e_h}{m} = \frac{card\{i \in M : x_i = v_h\}}{m}, \quad h \in H$$

- Per un campione sufficientemente numeroso:

$$f_h \approx p_h = Pr\{x = v_h\}, \quad h \in H$$

Attributi numerici

Istogrammi di densità empirica

- Si determina il numero R di classi, dipendente dal numero delle osservazioni m e dall'uniformità dei dati.
- Si definisce il range totale e l'ampiezza l_r di ciascuna classe.
  - Si può dividere il range per il numero di classi
- Si conta il numero di osservazioni in ciascun intervallo e si assegna a ciascun rettangolo altezza pari alla densità empirica:

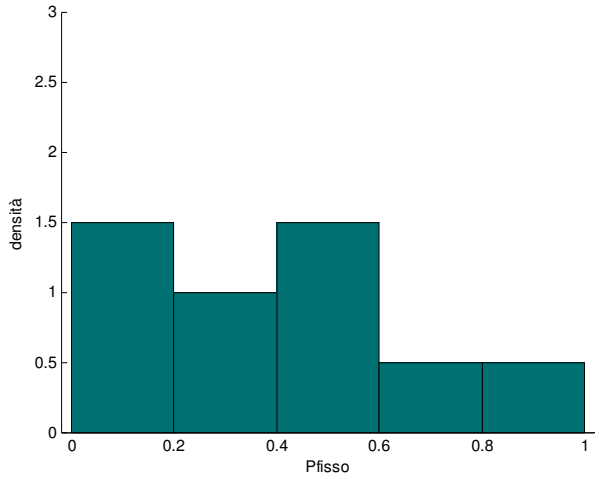
$$p_r = \frac{e_r}{m l_r}$$

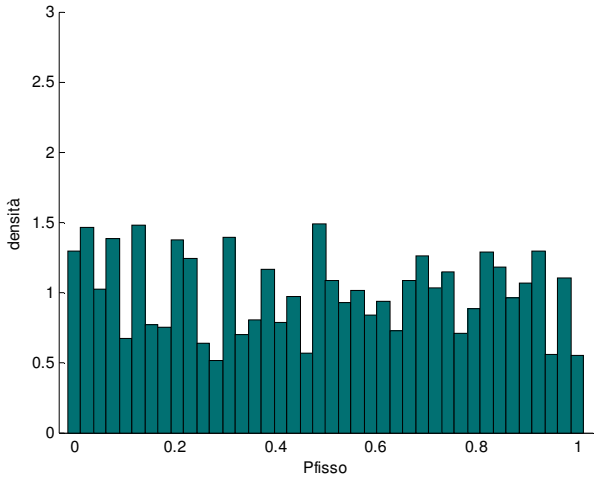
L'area totale dei rettangoli è 1:

$$\sum_{r=1}^{R} p_r l_r = \frac{1}{m} \sum_{r=1}^{R} e_r = 1$$

Attributi numerici

- La densità empirica può essere rappresentata con un diagramma simile a quello delle frequenze.





- Essa rappresenta la percentuale di campioni che si colloca in ciascuna classe e approssimano la probabilità che un nuovo campione cada nell'intervallo associato.

Indici di posizionamento

- *Media aritmetica campionaria*

$$\bar{\mu} = \frac{x_1 + x_2 + \cdots + x_m}{m} = \frac{1}{m} \sum_{i=1}^{m} x_i$$

- La somma algebrica degli *scarti* dalla media campionaria è nulla:

$$\sum_{i=1}^{m} (x_i - \bar{\mu}) = 0$$

- La media aritmetica rende minima la somma dei quadrati degli scarti da un valore di riferimento:

$$\sum_{i=1}^{m} (x_i - \bar{\mu})^2 = \min_c \sum_{i=1}^{m} (x_i - c)^2$$

- *Media campionaria pesata*:

$$\bar{\mu} = \frac{\omega_1 x_1 + \omega_2 x_2 + \cdots + \omega_m x_m}{m} = \frac{1}{m} \sum_{i=1}^{m} \omega_i x_i$$

Indici di posizionamento

- *Mediana*
  - m dispari: $x^{med}=x_{(m+1)/2}$
  - m pari: $x^{med} = \frac{(x_{m/2} + x_{(m/2+1)})}{2}$
- *Moda:* massimo della curva di densità empirica
- *Midrange*

$$x^{midr} = \frac{x^{max} + x^{min}}{2},\ x^{max} = \max_i x_i,\ x^{min} = \min_i x_i.$$

Indici di dispersione

- *Range*

$$x^{range} = x^{max} - x^{min},\ x^{max} = \max_i x_i,\ x^{min} = \min_i x_i.$$

- *Deviazione media*

$$s_i = x_i - \bar{\mu}, \quad i \in M$$

  - vale la relazione:

$$\sum_{i=1}^{m} s_i = 0$$

- *Deviazione media assoluta*

$$MAD = \frac{1}{m} \sum_{i=1}^{m} |s_i| = \sum_{i=1}^{m} |x_i - \bar{\mu}|$$

Varianza campionaria

- *Varianza campionaria*

$$\bar{\sigma}^2 = \frac{1}{m-1}\sum_{i=1}^{m} s_i^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{\mu})^2$$

- Una varianza campionaria inferiore comporta una minore dispersione dei valori attorno alla media campionaria.
- Dilata gli errori più grandi. Per riportare la misura di dispersione alla scala originale delle osservazioni, si ricorre alla *deviazione standard campionaria*:

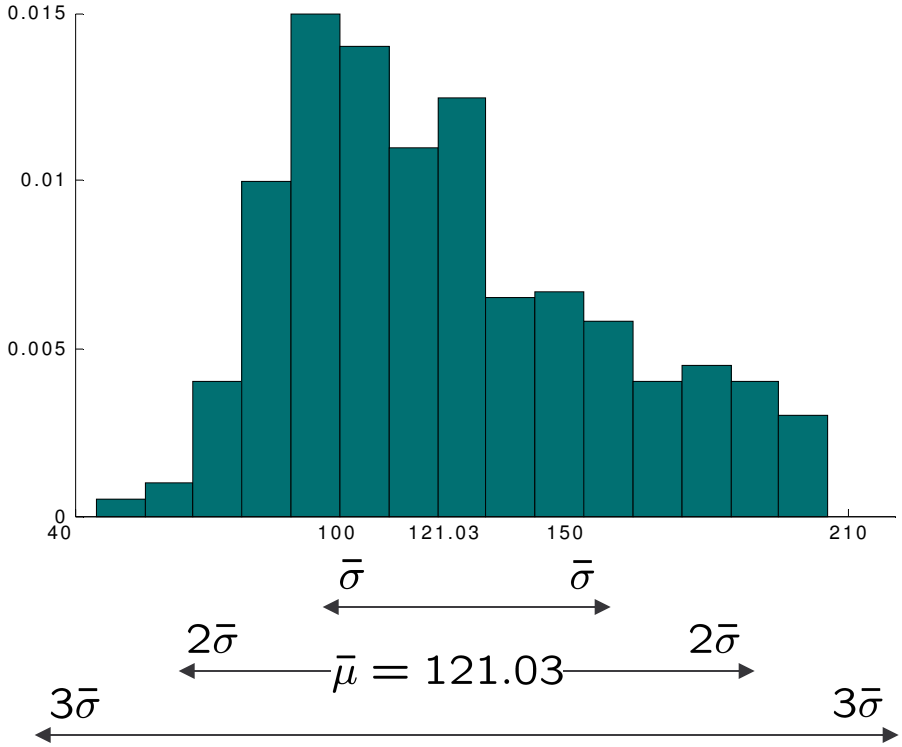
$$\bar{\sigma} = \sqrt{\bar{\sigma}^2}$$

Varianza campionaria

- La varianza può essere impiegata per delimitare l'intervallo intorno alla media campionaria in cui è ragionevole attendersi che cadano i valori del campione.
- **Distribuzione normale**
  - L'intervallo $(\mu \pm \sigma)$ contiene circa il 68% dei valori osservati
  - L'intervallo $(\mu \pm 2\sigma)$ contiene circa il 95% dei valori osservati
  - L'intervallo $(\mu \pm 3\sigma)$ contiene circa il 100% dei valori osservati
- **Distribuzione arbitraria**
  - Anche se la distribuzione è significativamente diversa dalla normale è ancora possibile ricavare intervalli entro cui ci si può attendere che cadano i valori del campione.

Teorema di Tchebysheff

Dato un numero $\gamma \geq 1$ e un insieme di m valori $\boldsymbol{a}=(x_1,x_2,\cdots,x_m)$, una percentuale pari ad almeno $(1\text{-}1/\gamma^2)$ dei valori si colloca all'interno dell'intervallo $(\mu \pm \gamma\sigma)$, ossia a non più di γ deviazioni standard dalla media campionaria.



x_i	$x_i - \bar{\mu}$	$\lvert x_i - \bar{\mu} \rvert$	$(x_i - \bar{\mu})^2$
45.0	-4.3	4.3	18.5
20.0	-29.3	29.3	858.5
69.0	19.7	19.7	388.1
66.0	16.7	16.7	278.9
11.0	-38.3	38.3	1466.9
42.0	-7.3	7.3	53.3
126.0	76.7	76.7	5882.9
47.0	-2.3	2.3	5.3
43.0	-6.3	6.3	39.7
24.0	-25.3	25.3	640.1
Σ = 493.0	Σ = 0.0	Σ = 226.2	Σ = 9632.1
$\bar{\mu}$ = 49.3			$\bar{\sigma}^2$ = 1070.23

Esempio di media, scarti, MAD e varianza

Indici di posizionamento relativo

- Se abbiamo m valori $x_1 \leq x_2 \leq \cdots \leq x_m$, un *quantile di ordine* p è un valore q_p tale che pm osservazioni cadono alla sinistra di q_p e le rimanenti $(1\text{-}p)m$ alla sua destra.
  - Il quantile di ordine 0.5 coincide con la mediana
  - q_L quartile di ordine 0.25 (*inferiore*)
  - q_U quartile di ordine 0.75 (*superiore*)
- I due quartili e la mediana dividono le osservazioni in quattro porzioni di numerosità equivalente.

Analisi monovariata

- Si utilizzano indicatori sintetici che individuano, con un singolo valore, proprietà 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 variabilità: varianza, deviazione standard;

- Adesso vedremo:
  - indicatori di posizionamento: quartili, z-indice.
  - indicatori di eterogeneità: indice di entropia.
  - indicatori di di asimmetria e curtosi

Identificazione degli oulier

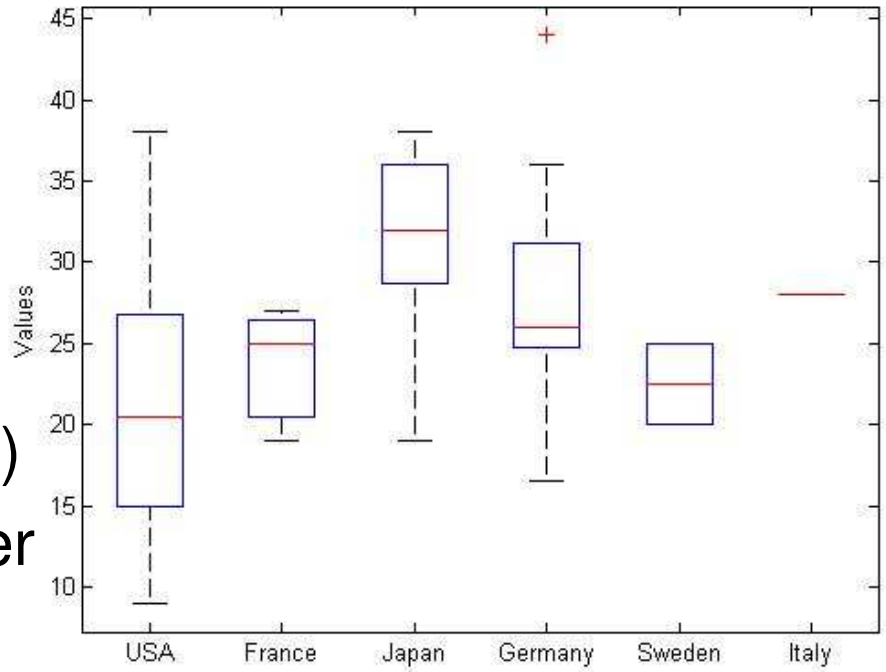
- Il teorema di Tchebysheff ha una semplice ed immediata applicazione per identificare gli outlier.
- Definiamo lo *z-indice* di una generica osservazione x_i:

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

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

Diagrammi box&whiskers

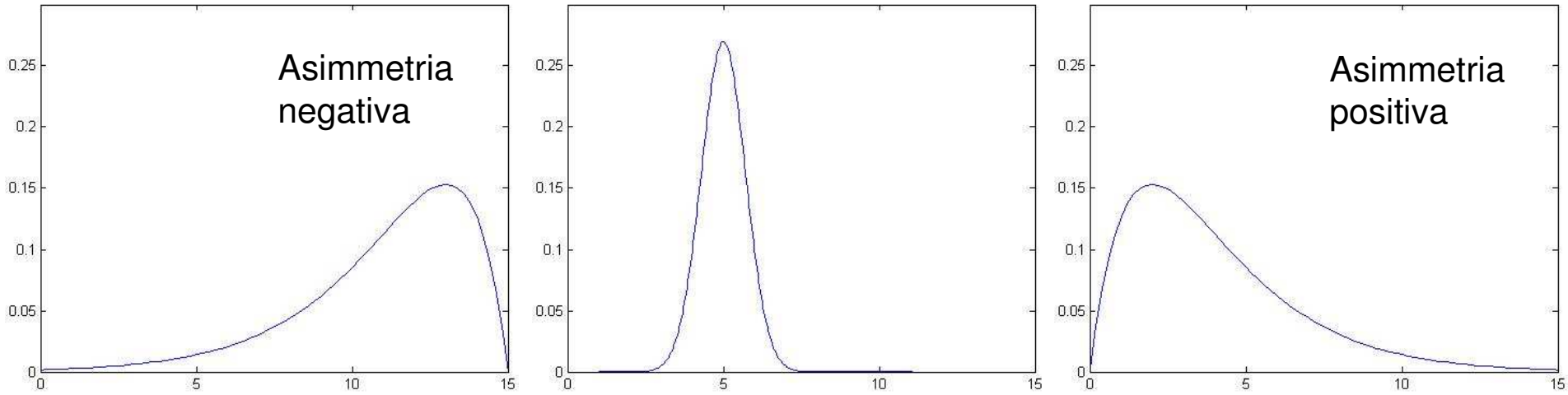
- Il box si estende tra i due quartili inferiore e superiore.
- La linea centrale è la mediana
- I baffi sono il minimo e massimo valore in $(q_L - 1.5D_q, q_U + 1.5D_q)$
  - $D_q = q_U - q_L$ (dispersione dei dati)
- I valori esterni ai baffi sono outlier
  - Disponibile nello statistics toolbox del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



Asimmetria della curva di densità

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

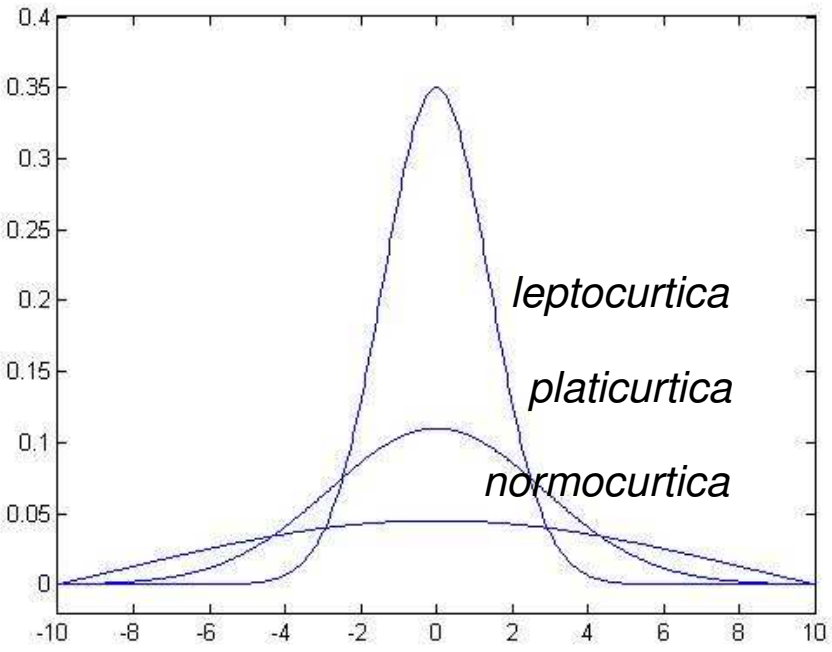
- *Indice di asimmetria* (*skewness*): $I_{as} = \frac{\bar{\mu}_3}{\bar{\sigma}^3}$
  - Misura la mancanza di simmetria



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 :

Icurt>0 : meno valori agli estremi di quanto aspettato

Minore dispersione

Icurt<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 (p.e. 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 (diagrammi a punti)

- 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

- 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 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 sono appropriate
- La misura di *correlazione* risolve il problema producendo un risultato indipendente dalle unita` 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 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
- Ma, attenzione: 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 (p.e. le vendite hanno un andamento stagionale?)
- Esempio: TOYS.xls

Esercitazione

- Esempio: calcolare il valore di asimmetria per la distribuzione dei tempi di interarrivo in banca (BANK.XML)
- 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

Esercitazione (1)

- 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 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 (2)

- Usando il file 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.

Sommario

- In questa lezione abbiamo visto:
  - **Analisi univariata**:
    - Indici di posizionamento
    - Indici di dispersione
    - Varianza campionaria
  - **Analisi bivariata**:
    - Covarianza
    - Correlazione
  - **Analisi multivariata**
    - Legami tra un sottoinsieme di attributi