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

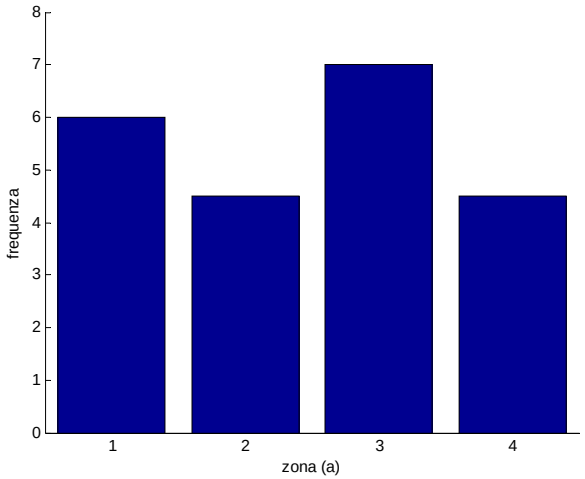
Analisi univariata

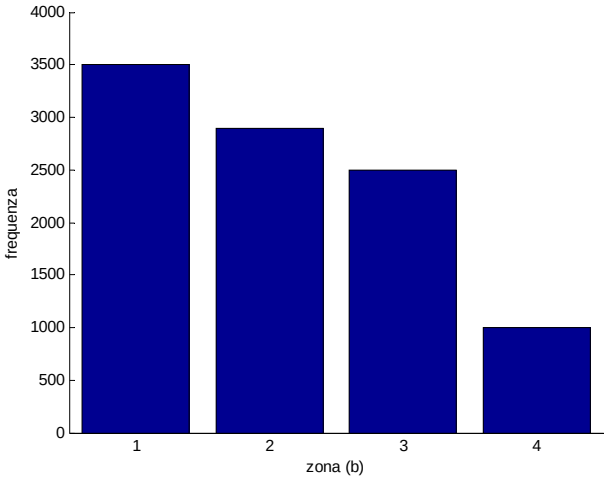
- Essa viene condotta valutando la tendenza dei valori dei singoli attributi a:
  - collocarsi in prossimità di un valore centrale (*posizionamento*),
  - assumere un certo range di valori (*dispersione*),
  - distribuirsi in maniera intelligibile.
- Obiettivi:
  - Verificare ipotesi statistiche
   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

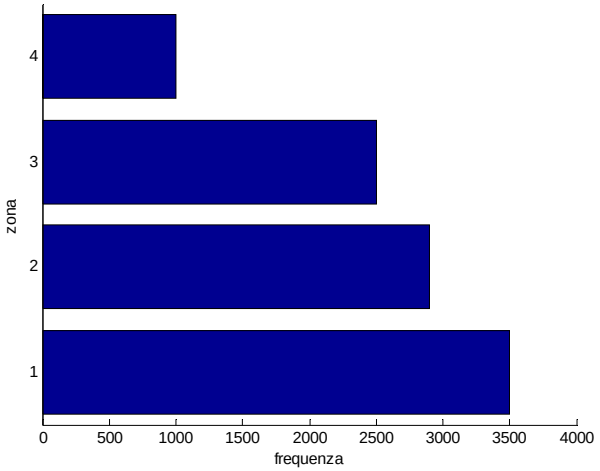
Attributi categorici

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

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







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:

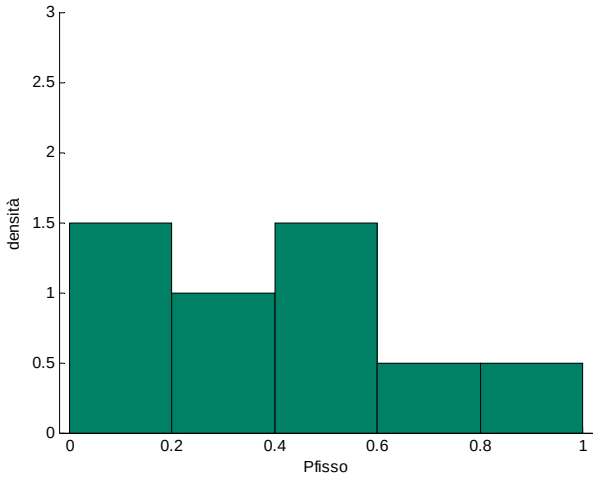
$$p_r = \frac{e_r}{m l_r}$$

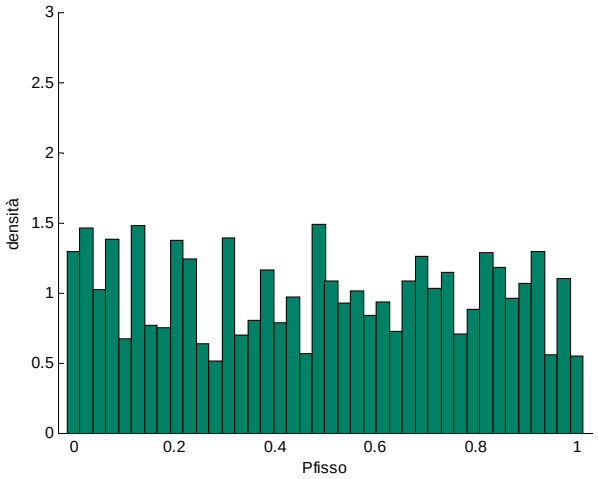
L'area totale dei rettangoli è 1:

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

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:

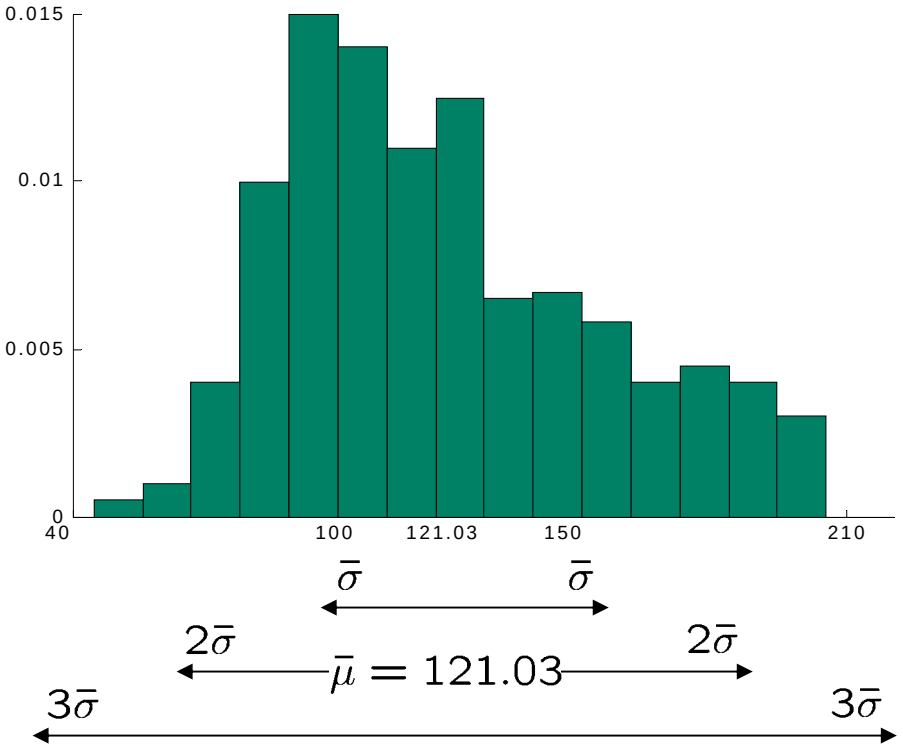
$$\bar{\sigma} = \sqrt{\bar{\sigma}^2}$$

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.