Elementi di Statistica

Definizioni

- Popolazione
  - insieme di elementi omogenei
- Campione
  - sottoinsieme di elementi della popolazione
- Parametro (o variabile)
  - caratteristica (generalmente – ma non necessariamente – numerica) della popolazione
- Campione di dati
  - insieme di valori di un dato parametro

Statistica

- Scopo della **Statistica** è lo studio delle caratteristiche di una data popolazione
- Spesso è troppo oneroso studiare l'intera popolazione (o impossibile se la popolazione è infinita)
- Si seleziona un campione rappresentativo della popolazione e lo si studia
  - il campione deve essere scelto in modo da essere *rappresentativo* dell'intera popolazione
  - si raccolgono i dati dal campione e li si analizza
  - si *inferisce* il comportamento della popolazione dall'analisi del campione

Variabili statistiche

- Qualitative
  - risultato di una categorizzazione o di un attributo della popolazione
    - » es: colore dei capelli, gruppo sanguigno, modello di vettura, città di residenza,…
- Quantitative
  - risultato di un conteggio o di una misura → sono sempre variabili numeriche
    - » es: pulsazioni cardiache, peso corporeo, abitanti di una città, prezzo di un oggetto,…
  - sono preferite perché è possibile usare strumenti matematici per analizzarle

Variabili statistiche quantitative

- Discrete
  - frutto di un conteggio
  - sono sempre numeri interi positivi (zero incluso)
    - » es: pulsazioni al minuto, studenti di un corso, messaggi ricevuti
- Continue
  - frutto di una misura (es. lunghezza, peso, tempo)
  - possono essere numeri interi, frazionari, decimali…
    - » es: peso corporeo, pressione arteriosa, durata di una chiamata telefonica

Rappresentazione dei dati

- Tabella di testo

Table 1: Basic stats

			Death	
	Total	Alive	Melanoma	Non-melanoma
Sex				
Male	126 (61 %)	91 (72 %)	28 (22 %)	7 (6 %)
Female	79 (39 %)	43 (54 %)	29 (37 %)	7 (9 %)
Age				
Mean (SD)	52 (± 17)	50 (± 16)	55 (± 18)	65 (± 11)
Ulceration				
Absent	115 (56 %)	92 (80 %)	16 (14 %)	7 (6 %)
Present	90 (44 %)	42 (47 %)	41 (46 %)	7 (8 %)
Thickness[a]				
Mean (SD)	2.9 (± 3.0)	2.2 (± 2.3)	4.3 (± 3.6)	3.7 (± 3.6)

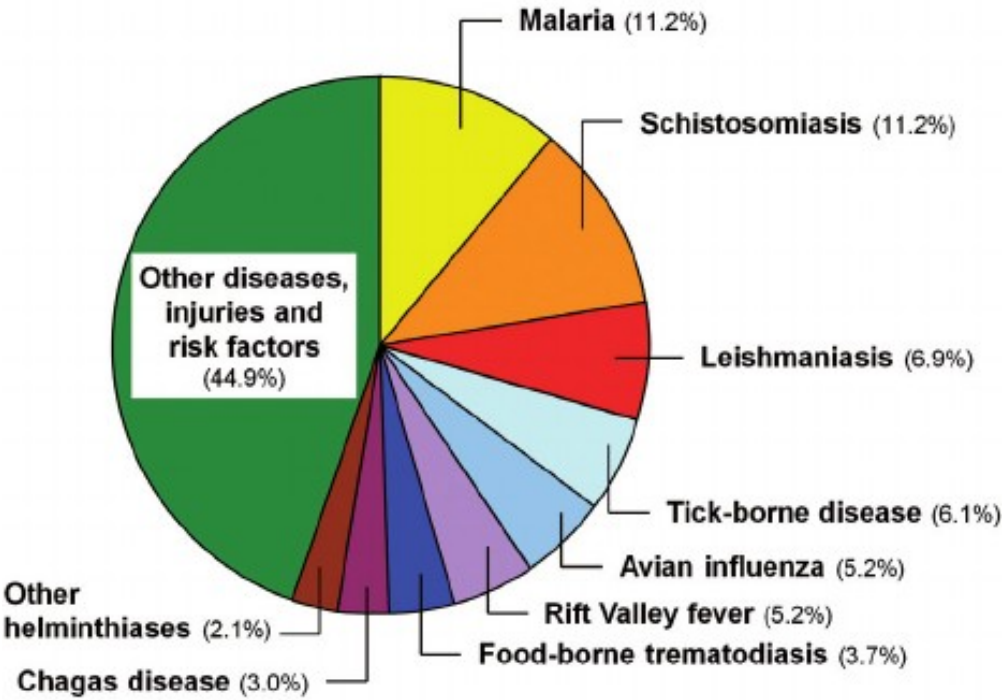
[a] Also known as Breslow thickness

☺ (solitamente) più ricca di dati

☹ non di immediata lettura

Rappresentazione dei dati

- Diagramma a torta (pie chart)

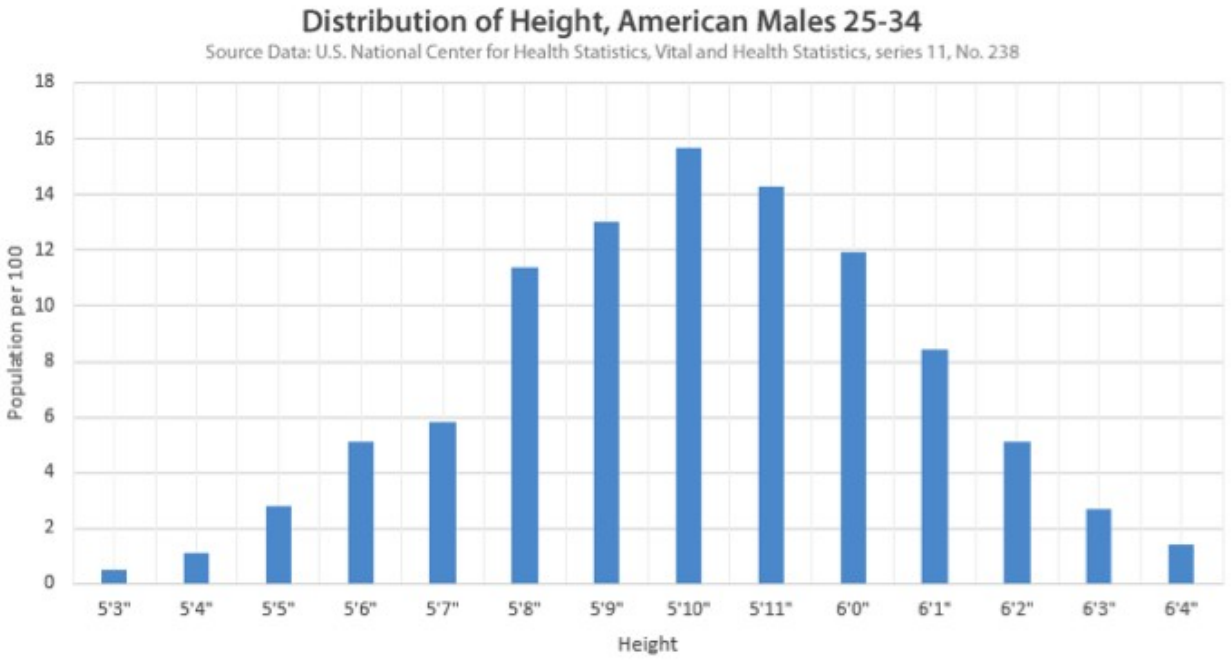


☺ immediata lettura delle proporzioni

☹ non chiari gli andamenti

Rappresentazione dei dati

- Istogramma

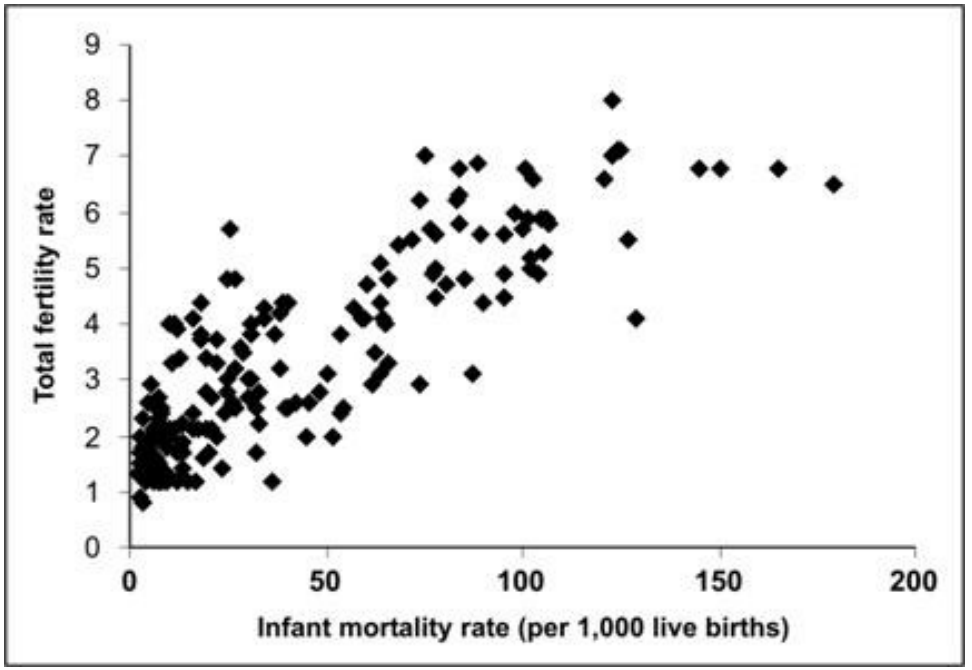


☺ immediata lettura delle proporzioni

☹ non sempre chiari gli andamenti

Rappresentazione dei dati

- Grafico a dispersione (scatter plot)

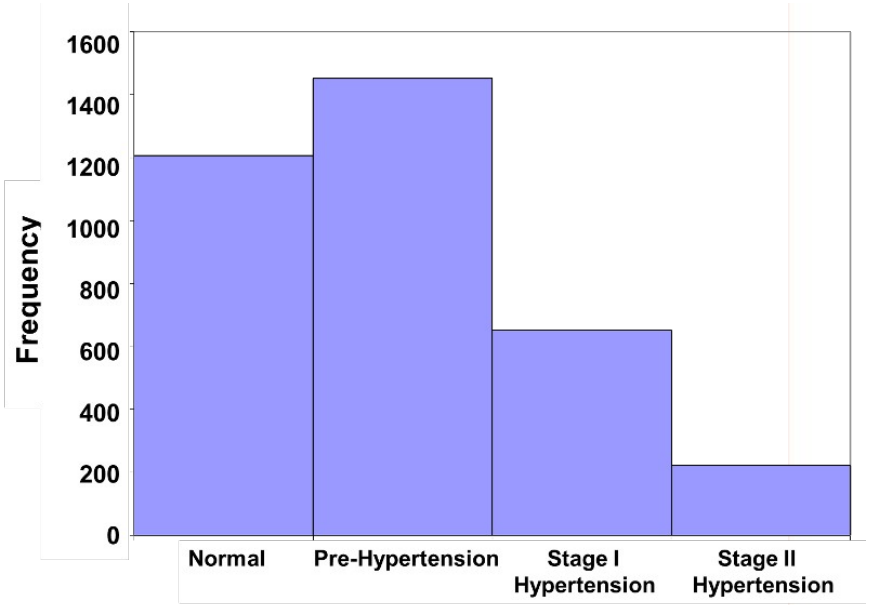


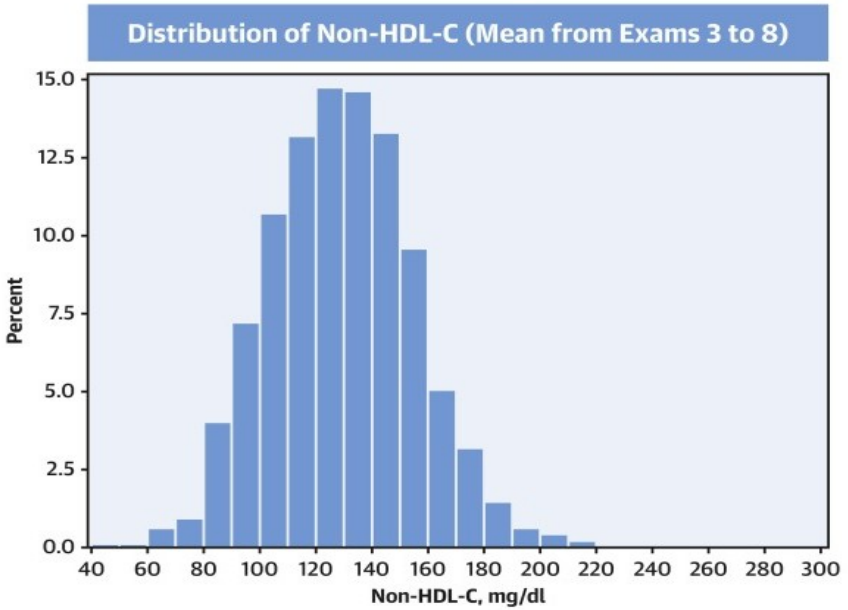
☺ immediata lettura degli andamenti

☹ poco adatto per rappresentare proporzioni

Rappresentazione dei dati

- Istogrammi: adatti sia a variabile qualitative che quantitative

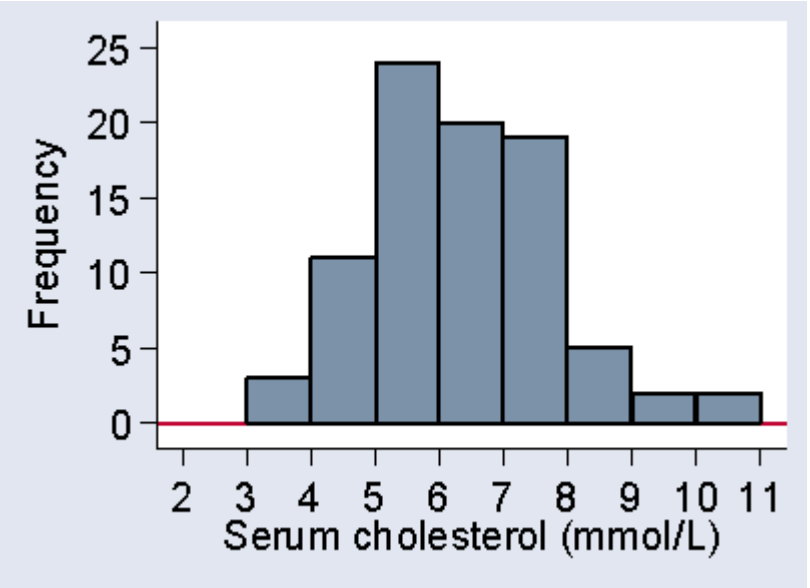


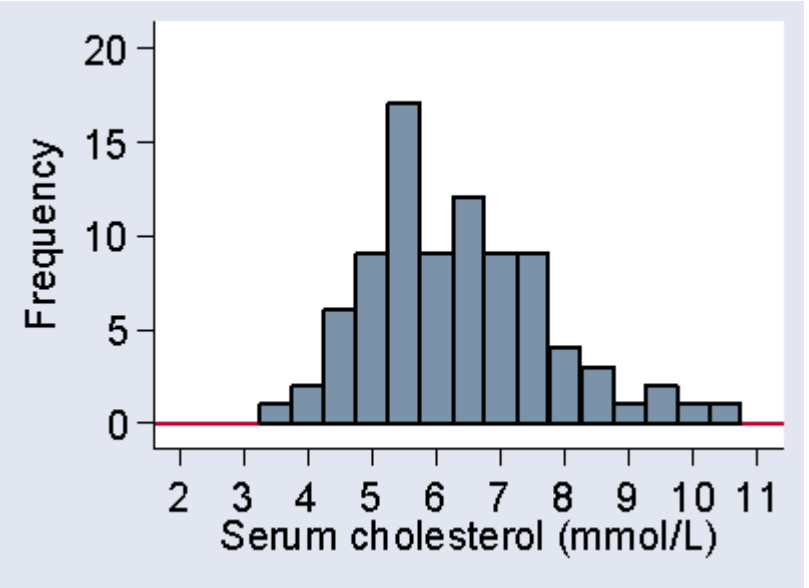


Pencina, K.M. et al. J Am Coll Cardiol. 2019;74(1):70-9.

Rappresentazione dei dati

- Suddivisione delle classi ← maggiori dettagli





Media, mediana, moda

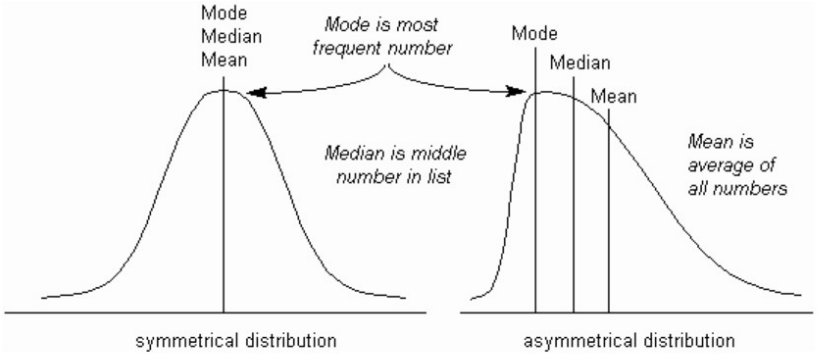
- Data una popolazione e una certa variabile x
- Media aritmetica

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  - dove la somma è estesa a *tutti* gli elementi della popolazione
- Mediana
  - valore x_m che divide in due parti uguali la distribuzione della variabile x
    - » metà dei dati è $< x_m$ metà è $> x_m$

Media, mediana, moda

- Moda
  - valore x_M per cui la distribuzione di x è massima, ovvero valore che si presenta con maggior frequenza
    - » distribuzione unimodale: x_M è unico (una sola moda)
    - » distribuzione multimodale: più di un x_M (ci sono più picchi)
      - distribuzione bimodale: due mode
- In generale media, mediana e moda **non** coincidono
  - a meno che la distribuzione sia simmetrica (e unimodale)



Media, mediana, moda

- Dato un campione di N dati $\{x_1, x_2, \ldots, x_N\}$ di una certa variabile x
- Media empirica

$$\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

- La media empirica $\bar{x}$ calcolata sul campione $\{x_i\}$ è uno stimatore della media μ della popolazione
  - al crescere del numero N di dati la media empirica tende alla media della popolazione

Media, mediana, moda

- Mediana
  - valore x_m che divide in due parti uguali il campione di dati
    - » metà dei dati è $< x_m$ metà è $> x_m$
    - » può essere $x_m \notin \{x_i\}$
- Moda
  - valore $x_M \in \{x_i\}$ per cui la distribuzione è massima, ovvero valore che si presenta con maggior frequenza
    - » anche in questo caso la distribuzione può essere unimodale o multimodale
- In generale media, mediana e moda **non** coincidono
  - a meno che la distribuzione sia simmetrica (e unimodale)

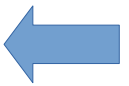
Media pesata

- Media pesata
  - se ogni dato x_i ha un peso diverso p_i

$$\overline{x} = \frac{\sum_{i=1}^{N} p_i x_i}{\sum_{i=1}^{N} p_i}$$

  - es: se ogni dato x_i ha un errore s_i diverso

$$\overline{x} = \frac{\sum_{i=1}^{N} \frac{x_i}{s_i}}{\sum_{i=1}^{N} \frac{1}{s_i}}$$



peso maggiore
ai dati con
errore minore

Scarto

- Per ogni x_i: scarto

$$\sigma_i = \mu - x_i$$

$$s_i = \bar{x} - x_i \qquad \text{(scarto empirico)}$$

- Per definizione di media

$$\sum_{i=1}^{n} \sigma_i = \sum_{i=1}^{n} (\mu - x_i) = n\mu - \sum_{i=1}^{n} x_i = 0$$

$$\sum_{i=1}^{N} s_i = \sum_{i=1}^{N} (\bar{x} - x_i) = N\,\bar{x} - \sum_{i=1}^{N} x_i = 0$$

  - quindi la media degli scarti è sempre nulla!

Varianza

- Varianza di una popolazione

$$\sigma^2 = \frac{\sum_{i=0}^{n} (\mu - x_i)^2}{n}$$

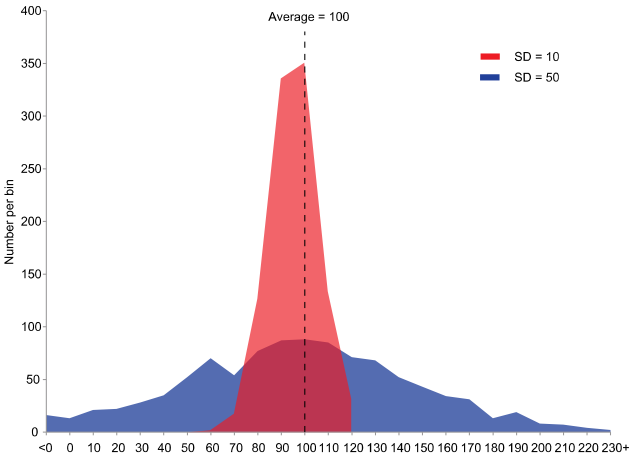
  - è la media del quadrato degli scarti
    - » la somma è estesa a *tutti* gli elementi della popolazione

Deviazione standard

- Deviazione standard (o scarto quadratico medio) di una popolazione

$$\sigma = \sqrt{\frac{\sum_{i=0}^{n} (\mu - x_i)^2}{n}}$$

  - è la radice quadrata della varianza σ^2
  - è un indicatore di quanto i valori della variabile *x* siano "dispersi" intorno al valor medio μ
    - » se è piccola vuol dire che i dati sono in gran parte raccolti vicino al valor medio; se è grande i dati sono molto dispersi



Varianza empirica

- Varianza (empirica) di un campione di dati

$$s^2 = \frac{\sum_{i=0}^{N} \left(\overline{x} - x_i\right)^2}{N-1}$$

  - è la media del quadrato degli scarti
  - è uno stimatore della varianza della popolazione
    - » al crescere del numero *N* di dati la varianza empirica tende alla varianza della popolazione

Deviazione standard empirica

- Deviazione standard empirica (o scarto quadratico medio empirico) di un campione di dati

$$s = \sqrt{\frac{\sum_{i=0}^{N} \left(\overline{x} - x_i\right)^2}{N-1}}$$

  - è la radice quadrata della varianza empirica s^2
  - è un indicatore di quanto i dati $\{x_i\}$ del campione siano "dispersi" intorno al valor medio empirico $\overline{x}$
  - è uno stimatore della varianza della popolazione
    - » al crescere del numero N di dati la deviazione standard empirica tende alla deviazione standard della popolazione

Deviazione standard della media

- Deviazione standard della media (o errore della media)
  - per una popolazione

$$\sigma_e = \frac{\sigma}{\sqrt{n}}$$

  - per un campione

$$s_e = \frac{s}{\sqrt{N}}$$