Esercizio standard di Biostatistica

Siano

$$\begin{aligned} \mathcal{C}_N(X) &= \{x_1, .., x_N\} \\ \mathcal{C}'_{N'}(X) &= \{x'_1, .., x'_N\} \ , \end{aligned}$$

due campioni distinti di misure della quantità X e

$$\mathcal{C}_N(Y) = \{y_1, .., y_N\}$$

un campione di misure della quantità Y, tale che $\mathcal{C}_N(X)$ e $\mathcal{C}_N(Y)$ siano i campioni marginali associati al campione bivariato

$$\mathcal{C}_N(X, Y) = \{(x_1, y_1), .., (x_N, y_N)\}.$$

- Tracciare il diagramma ramo-foglia del campione $\mathcal{C}_N(X)$.

  Risp. vedi appunti del corso.

- Calcolare la media campionaria $\bar{x}$ dei dati di $\mathcal{C}_N(X)$.

  $$Risp.\ \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- Calcolare mediana campionaria $\hat{x}$ di $\mathcal{C}_N(X)$.

  Risp. una volta riordinati i dati del campione dal più piccolo al più grande si ottiene una collezione $\{\hat{x}_1, .., \hat{x}_N\}$ di misure di X dove, $\forall i = 1, .., N,\ \hat{x}_i \leq \hat{x}_{i+1}$. Allora, la mediana campionaria $\hat{x}$ risulta

  $$\hat{x} := \begin{cases} \frac{\left(\hat{x}_{\frac{N}{2}} + \hat{x}_{\frac{N}{2}+1}\right)}{2} & se\, N\, è\, pari \\ \hat{x}_{\frac{N+1}{2}} & se\, N\, è\, dispari \end{cases} .$$

- Calcolare i valori modali campionari del campione $\mathcal{C}_N(X)$.

  Risp. trovare i valori dei dati di $\mathcal{C}_N(X)$ che hanno frequenza massima, cioè che ripetono più volte come elementi del campione.

- Calcolare la varianza campionaria s_X^2 e la deviazione standard capionaria s_X dei dati di $\mathcal{C}_N(X)$.

  $$Risp.\ s_X^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2\ ,\ s_X = \sqrt{s_X^2}.$$

- Trovare il più piccolo intorno U_α della media campionaria tale che la percentuale dei dati di $\mathcal{C}_N(X)$ che vi ricade sia maggiore del $p\%$.

  Risp. dalla disuguaglianza di Chebichev si ha

  $$\frac{|\mathcal{S}_N^\alpha(X)|}{N} > 1 - \frac{1}{\alpha^2} > p\%.$$

  Invertendo la precedente disuguaglianza si ha $U_\alpha = [\bar{x} - \alpha s_X, \bar{x} + \alpha s_X]$ dove

  $$\alpha > \sqrt{\frac{1}{1 - p\%}}.$$

- Utilizzare i valori dei campioni $\mathcal{C}_N(X)$ e $\mathcal{C}'_{N'}(X)$ per dare una stima $\breve{x}$ del valore della quantità X.

  Risp. calcolare la media campionaria

  $$\bar{x}' = \frac{1}{N'} \sum_{i=1}^{N'} x'_i$$

  e la varianza campionaria

  $$\left(s'_X\right)^2 = \frac{1}{N' - 1} \sum_{i=1}^{N'} \left(x'_i - \bar{x}'\right)^2$$

  di $\mathcal{C}'_N(X)$ e applicare il metodo della media pesata

  $$\breve{x} = \frac{\frac{1}{\left(s'_X\right)^2}\bar{x}' + \frac{1}{s_X^2}\bar{x}}{\frac{1}{\left(s'_X\right)^2} + \frac{1}{s_X^2}}.$$

- Graficare il diagramma di dispersione dei dati di $\mathcal{C}_N(X, Y)$, calcolare il coefficiente di correlazione $r_{X,Y}$ e, compatibilmente col suo valore, calcolare la retta di regressione tra i dati di $\mathcal{C}_N(X)$ e $\mathcal{C}_N(Y)$.

  Risp. per quanto riguarda tracciare il diagramma di dispersione associato al campione $\mathcal{C}_N(X, Y)$ si consultino gli appunti sull'argomento.

  $$\begin{aligned}
  s_Y &= \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} \left(y_i - \bar{y}\right)^2}, \\
  s_{X,Y} &= \frac{1}{N-1} \sum_{i=1}^{N} \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right), \\
  r_{X,Y} &= \frac{s_{X,Y}}{s_X s_Y} = \frac{\sum_{i=1}^{N} \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{N} \left(x_i - \bar{x}\right)^2}\sqrt{\sum_{i=1}^{N} \left(y_i - \bar{y}\right)^2}}.
  \end{aligned}$$

Se il valore di $r_{X,Y}$ in modulo è abbastanza prossimo a 1 allora esistono A e B tali che $Y = AX + B$ (retta di regressione) i cui valori stimati dai dati risultano

$$\begin{cases} \check{A} = \frac{s_{X,Y}}{s_X^2} \\ \check{B} = \bar{y} - \frac{s_{X,Y}}{s_X^2} \bar{x} \end{cases},$$

dove

$$\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i.$$

- Testare l'ipotesi di gaussianità delle misure della quantità X.

  Risp. applicare il test del chi quadro ai dati del campione $\mathcal{C}_N(X)$ (cfr gli appunti). La partizione più piccola di $\mathbb{R}$ che si può scegliere è $A_{-2} = (-\infty, \bar{x} - s_X]$, $A_{-1} = (\bar{x} - s_X, \bar{x}]$, $A_1 = (\bar{x}, \bar{x} + s_X]$, $A_2 = (\bar{x} + s_X, +\infty)$ con $\mathbb{P}_{\bar{x},s_X}(A_{-1}) = \mathbb{P}_{\bar{x},s_X}(A_1) = 0,34$; $\mathbb{P}_{\bar{x},s_X}(A_{-2}) = \mathbb{P}_{\bar{x},s_X}(A_2) = 0,16$.