Statistica

Definizione

Studia le popolazioni di individui, ovvero gruppi che condividono una caratteristica specifica
Si occupa di analizzare ed interpretare i dati

Campione

Nell'impossibilità di osservare l'intera popolazione spesso si osserva un campione, ovvero un sottoinsieme ottenuto tramite un processo di selezione
Deve essere rappresentativo, quindi ogni elemento deve avere la stessa probabilità di essere considerato nel campione

Parametro

Misura numerica calcolata sull'intera popolazione, solitamente fissa ma sconosciuta

Stimatore

O statistica, è una misura calcolata su un campione nota ma variabile da campione a campione

---

Statistica descrittiva

Definizione

Si concentra sull'analisi dei dati osservati con l'obiettivo di sintetizzare le informazioni contenute nei dati di un campione
Si possono rappresentare queste informazioni tramite grafici e tabelle

Variabili

Definizione

Si distinguono tra:

- Numeriche (quantitative):
  - Discrete
  - Continue
- Categoriche (qualitative)

Indici di posizione e dispersione

Definizione

Media aritmetica

Anche chiamata media campionaria

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

Per i dati raggruppati in classi si moltiplica il valore centrale per la frequenza assoluta e si divide per il numero totale di osservazioni

Mediana

Valore che divide il campione (dati ordinati) in due parti

- dispari: valore centrale
- pari: media aritmetica dei due valori centrali

Quartile e percentile

Valori che dividono il campione in quarti e percentuali
Si applica la stessa regola della mediana

Range

Distanza tra il minimo ed il massimo
$\text{Range} = x_{\max} - x_{\min}$

IQR

Range interquartile
$\text{IQR} = Q_3 - Q_1$

Deviazione Standard

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Tabelle

Definizione

Di frequenza

Si suddividono i dati in classi e si calcolano la frequenza relativa e cumulativa

Grafici

Definizione

Istogramma

Ogni classe è rappresentata da un rettangolo la cui area è proporzionale alla frequenza

Box and whiskers plot

Fornisce una rappresentazione visiva della distribuzione dei dati, permette di confrontare la distribuzione dei dati tra diversi gruppi o categorie e di identificare simmetrie o asimmetrie

- Whiskers o baffi: linee verticali che si estendono dal box fino al massimo e minimo escludendo gli outliers
- Box o scatola: rappresenta il 50% centrale dei dati compreso tra Q_1 e Q_3, una linea all'interno rappresenta la mediana
- Outliers: punti al di fuori dei baffi che rappresentano i valori che distano più di $1.5 \cdot \text{IQR}$ dalla scatola

Esempio ›

Tabella di frequenza, istogramma e box plot

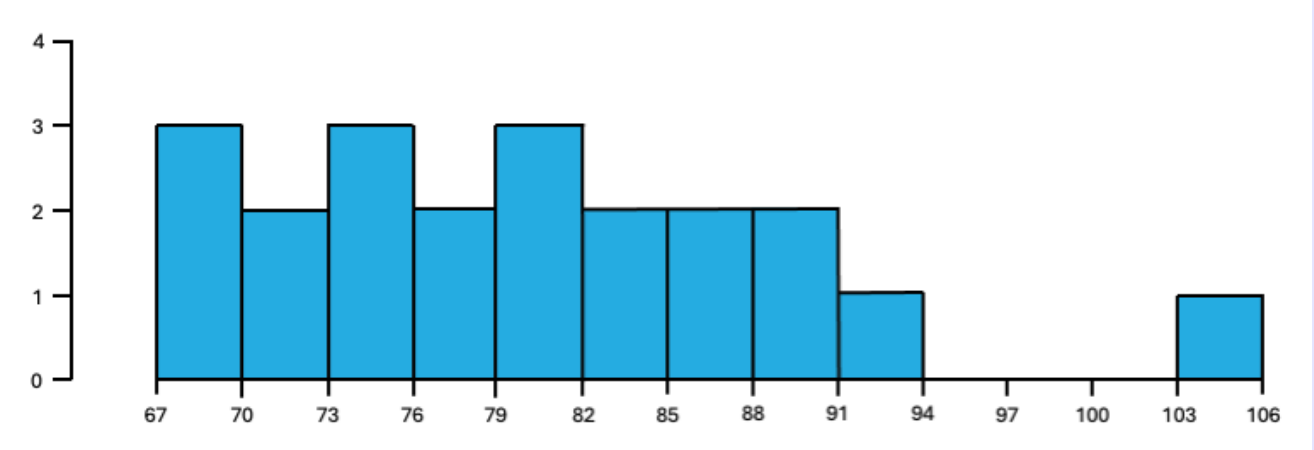
Dati ordinati:
$67.8, 68.2, 69.3, 70.5, 72.1, 73.4, 75.5, 75.0, 76.4, 77.9, 79.0,$
$80.6, 81.5, 82.3, 83.7, 85.6, 86.4, 88.7, 90.1, 91.2, 105.8$

$Q_0 = x_{\min} = 67.8$
$Q_1 = 72.1 \cdot 0.25 + 73.4 \cdot 0.75 = 73.075$
$Q_2 = 79.0$
$Q_3 = 85.6 \cdot 0.75 + 86.4 \cdot 0.25 = 85.8$
$Q_4 = x_{\max} = 105.8$
$\bar{x} = \frac{1}{21} \sum_{i=1}^{21} x_i = 80$
$\text{Range} = 38$
$\text{IQR} = 12.725$
Outliers: $x \notin [Q_1 - 1.5 \cdot \text{IQR}, Q_3 + 1.5 \cdot \text{IQR}] = [53.9875, 104.8875]$

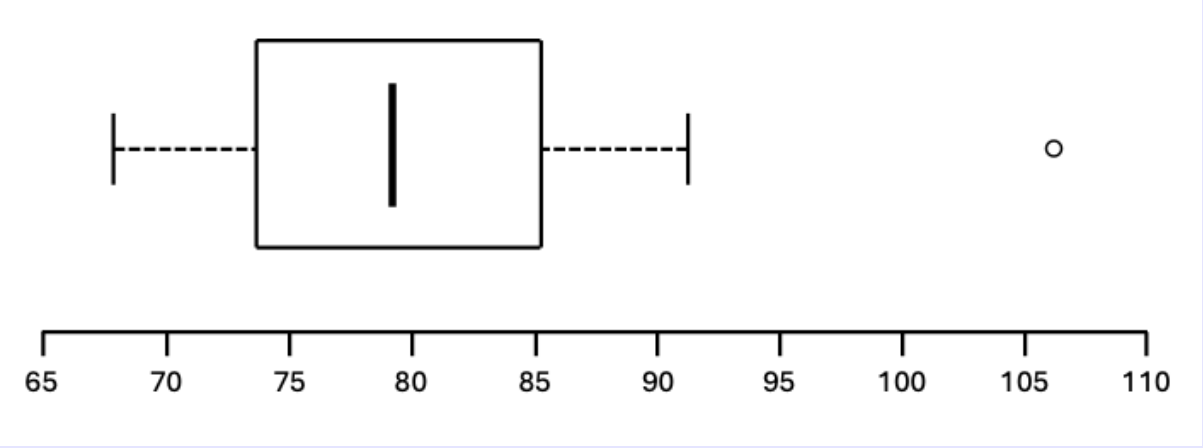
Suddivisione in classi di larghezza 3

Classe	Frequenza assoluta	Frequenza relativa	Frequenza cumulativa
$[67, 70)$	3	$\frac{3}{21}$	$\frac{3}{21}$
$[70, 73)$	2	$\frac{2}{21}$	$\frac{5}{21}$
$[73, 76)$	3	$\frac{3}{21}$	$\frac{8}{21}$
$[76, 79)$	2	$\frac{2}{21}$	$\frac{10}{21}$
$[79, 82)$	3	$\frac{3}{21}$	$\frac{13}{21}$
$[82, 85)$	2	$\frac{2}{21}$	$\frac{15}{21}$
$[85, 88)$	2	$\frac{2}{21}$	$\frac{17}{21}$
$[88, 91)$	2	$\frac{2}{21}$	$\frac{19}{21}$
$[91, 94)$	1	$\frac{1}{21}$	$\frac{20}{21}$
$[94, 97)$	0	0	$\frac{20}{21}$
$[97, 100)$	0	0	$\frac{20}{21}$
$[100, 103)$	0	0	$\frac{20}{21}$
$[103, 106)$	1	$\frac{1}{21}$	1

Istogramma:



Box plot:



Statistica inferenziale

Definizione

Cerca di fare affermazioni riguardanti l'intera popolazione basandosi sulle informazioni ricavate da un campione, espresse in termini di probabilità

Stimatore

Definizione

Si identifica il campione con una famiglia di variabili aleatorie $X_1, \ldots, X_n$ stocasticamente indipendenti con stessa distribuzione le cui CDF/PDF/PMF dipendono da un parametro θ spesso ignoto, l'obiettivo è determinarlo
Una regola che dice come calcolare la stima di θ o $\psi(\theta)$ dal campione $x_1, \ldots, x_n$ è detta stimatore

T funzione delle variabili aleatorie $X_1, \ldots, X_n$ è variabile aleatoria, si indica con $T = t(X_1, \ldots X_n)$ ed è detta statistica o stimatore di θ o $\psi(\theta)$ se osservato il campione e si utilizza $t(x_1, \ldots, x_n)$ al posto del parametro

Esempi ›

$T_1 : x_1$
$T_2 : x_1 + \ldots + x_n$
$T_3 : \frac{1}{n}(x_1, \ldots, x_n)$
$T_4 : \sqrt[n]{x_1 \cdot \ldots \cdot x_n}$

Correttezza

Definizione

Uno stimatore si dice corretto o **non distorto** per θ o $\psi(\theta)$ se $\mathbb{E}[T] = \theta \ \ \forall \theta$

Esempi ›

$\mathbb{E}[T_1] = \theta$, corretto
$\mathbb{E}[T_2] = \mathbb{E}[X_1 + \ldots + X_n] = \sum_{i=1}^{n} X_i = n\theta$, non corretto
$\mathbb{E}[T_3] = \frac{1}{n} n\theta = \theta$, corretto
$\mathbb{E}[T_4] \leq \theta$, non corretto

Consistenza

Definizione

Uno stimatore si dice consistente se $\lim_{n\to+\infty} \mathrm{Var}[T] = 0$ ovvero più grande è il campione più si riduce l'incertezza

Esempi ›

$\lim_{n\to+\infty} \mathrm{Var}[T_1] = \mathrm{Var}[X_1]$, non consistente
$\lim_{n\to+\infty} \mathrm{Var}[T_3] = \lim_{n\to+\infty} \mathrm{Var}\left[\frac{1}{n}(X_1 + \ldots + X_n)\right] = \lim_{n\to+\infty} \frac{\sigma^2}{n} = 0$, consistente

Stimatori per la media

Formule

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

$$\frac{1}{\bar{x}} = \frac{n-1}{\sum_{i=1}^{n} x_i}$$

Stimatori per la varianza

Formule

Se μ è noto:

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

Dimostrazione >

$$\begin{aligned}\mathbb{E}[T] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2\right] = \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n}(X_i^2 - 2X_i\mu + \mu^2)\right] \\ &= \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n}X_i^2 - 2\mu\sum_{i=1}^{n}X_i + n\mu^2\right] = \frac{1}{n}\left[\sum_{i=1}^{n}\mathbb{E}[X_i^2] - 2\mu\sum_{i=1}^{n}\mathbb{E}[X_i] + n\mu^2\right] \\ &= \frac{1}{n}\left[\sum_{i=1}^{n}\mathbb{E}[X_i^2] - 2n\mu^2 + n\mu^2\right] = \frac{1}{n}\left[\sum_{i=1}^{n}(\mathbb{E}[X_i^2] - \mathbb{E}[X_i]^2)\right] = \frac{1}{n}n\sigma^2 = \sigma^2\end{aligned}$$

Se μ non è noto:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n}x_i^2 - n\bar{x}^2\right]$$

anche chiamata varianza campionaria

Intervallo di confidenza

Definizione

Un intervallo di confidenza o **Confidence Interval** di livello $\alpha \in (0,1)$ per un parametro θ o $\psi(\theta)$ è un intervallo I_X : $\mathbb{P}(\theta \in I_X) = \alpha$, dipende dalle osservazioni $X_1, \ldots, X_n$ quindi è un'intervallo aleatorio

Formule

Intervallo di confidenza per la media se σ è nota

$$\Phi_{\frac{\alpha+1}{2}} = F_U^{-1}\left(\frac{\alpha+1}{2}\right)$$

$$\mu \in \left(\bar{x} - \Phi_{\frac{\alpha+1}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + \Phi_{\frac{\alpha+1}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

Definizione

Intervallo di confidenza di una proporzione
$X_i \sim \mathrm{B}(p)$, $X_i \perp\!\!\!\perp X_j \;\; i \neq j \; (i, j = 1, \ldots, n)$
Se $n \cdot p > 5$ e $n \cdot p \cdot (1 - p) > 5$ si può utilizzare l'approssimazione gaussiana e stimare p con $\bar{x}$ considerando $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \sim \mathrm{N}\left(p, \frac{p \cdot (1-p)}{n}\right)$

Formule

Intervallo di confidenza di una proporzione

$$p \in \left(\bar{x} - \Phi_{\frac{\alpha+1}{2}} \cdot \sqrt{\frac{\bar{x} \cdot (1 - \bar{x})}{n}}, \bar{x} + \Phi_{\frac{\alpha+1}{2}} \cdot \sqrt{\frac{\bar{x} \cdot (1 - \bar{x})}{n}} \right)$$

Definizione

$X_i \sim \mathrm{N}(0, 1)$, $X_i \perp\!\!\!\perp X_j \;\; i \neq j$, $(i, j = 1, \ldots, n)$
$Y = \sum_{i=1}^{n} \sim \chi^2(n)$
$Z \sim \mathrm{N}(0, 1)$, $Z \perp\!\!\!\perp Y$

$$T = \frac{Z}{\sqrt{\frac{Y}{n}}}$$

è detta T di Student con n gradi di libertà

Formule

Stimatore per la media se σ non è nota

$$\mu \in \left(\bar{x} - t_{\frac{\alpha+1}{2}(n-1)} \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha+1}{2}(n-1)} \cdot \frac{s}{\sqrt{n}} \right)$$