Analisi dei dati

Un goliardico riassunto

Ollari Dmitri

9 maggio 2023

Indice

Elenco delle figure

Capitolo 1

Statistica descrittiva

1.1 Grandezze che sintetizano i dati

1.1.1 Media

Avendo un'insieme di dati $x_1, x_2, \ldots, x_n$, la **media campionaria** è definita da:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{1.1}$$

1.1.2 Mediana

Avendo un'insieme di dati **ordinati(crescente)** con ampiezza n, la **mediana** è il valore che **occupa la posizione**:

- $\frac{n+1}{2}$ se n è **dispari**
- Media tra il valore $\frac{n}{2}$ e $\frac{n}{2} + 1$ se n è **pari**

La mediana campionaria (o mediana campionaria ordinata) è un indice di tendenza centrale utilizzato per descrivere il valore centrale di un insieme di dati numerici.

1.1.3 Moda

Unico valore che ha frequenza massima.

1.1.4 Varianza

La varianza è una misura della dispersione o della variabilità dei dati all'interno di un insieme di dati numerici. In altre parole, la varianza indica quanto i dati si distribuiscono intorno alla media.

Dato un'insieme di dati $x_1, x_2, \ldots, x_n$, la **varianza campionaria (s^2)** è:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{1.2}$$

1.1.5 Deviazione standard

La deviazione standard campionaria (o deviazione standard campionaria corretta) è una misura della dispersione dei dati all'interno di un campione di dati numerici. Come la varianza, la deviazione standard è un'importante misura di dispersione dei dati, ma a differenza della varianza, ha la stessa unità di misura degli elementi dell'insieme di dati e quindi è più facilmente interpretabile.

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{1.3}$$

1.1.6 Percentili campionari

I percentili campionari sono una misura statistica utilizzata per suddividere un insieme di dati in 100 parti uguali. In altre parole, i percentili campionari suddividono l'insieme di dati in modo tale da poter analizzare la distribuzione dei dati in modo più dettagliato.

Per calcolare i percentili campionari, si segue il seguente procedimento:

1. Si ordina l'insieme di dati in ordine crescente.
2. Si moltiplica il numero totale di elementi nell'insieme di dati per la percentuale desiderata (ad esempio, se si vuole trovare il 25° percentile, si moltiplica il numero totale di elementi per 0,25).
3. Si arrotonda il risultato ottenuto al punto 2 all'intero successivo, se il risultato non è già un intero.
4. Si individua l'elemento nell'insieme di dati che si trova all'indice calcolato al punto 3. Questo valore corrisponde al percentile desiderato.

1.1.7 Box plotting

Il diagramma a scatola si presenta come un rettangolo orientato verticalmente, che ha come estremi inferiori e superiori il primo quartile (Q1) e il terzo quartile (Q3) della distribuzione. All'interno del rettangolo si trova una linea orizzontale che rappresenta la mediana (o secondo quartile, Q2), mentre le linee che si estendono dalla parte superiore e inferiore del rettangolo (i cosiddetti baffi) rappresentano la variazione massima e minima dei dati.

1.2 Disuguaglianza di Chebyschev

La Disuguaglianza di Chebyshev è un importante teorema della teoria della probabilità che fornisce una stima superiore sulla percentuale di dati che si discostano dalla media, per ogni distribuzione di probabilità.

La disuguaglianza afferma che, per qualsiasi distribuzione di probabilità con media μ e varianza finita σ^2, la percentuale di dati che si discostano dalla media per più di k volte la deviazione standard (cioè la distanza dalla media in termini di deviazioni standard) è inferiore o uguale a $1/k^2$, ovvero:

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

dove X è una variabile aleatoria con media μ e varianza σ^2, e k è un qualsiasi valore positivo.

Questa disuguaglianza implica che, per qualsiasi distribuzione di probabilità, almeno il 75% dei dati si trova entro due deviazioni standard dalla media, e almeno il 89% dei dati si trova entro tre deviazioni standard dalla media. In altre parole, la disuguaglianza di Chebyshev fornisce un limite superiore alla dispersione dei dati, indipendentemente dalla loro distribuzione.

La disuguaglianza di Chebyshev è utile per stimare la probabilità di osservare un valore estremo in un insieme di dati, senza conoscere la loro distribuzione. Ad esempio, se si vuole stimare la percentuale di studenti di una scuola che hanno un QI inferiore a 130, sapendo solo la media e la deviazione standard del QI degli studenti, la disuguaglianza di Chebyshev può essere utilizzata per fornire una stima superiore sulla percentuale di studenti che hanno un QI inferiore a 130.

1.3 Insieme di dati bivariati

In statistica, un insieme di dati bivariati è un insieme di dati che contiene informazioni su due variabili per ogni osservazione nel campione. In altre parole, ogni osservazione nel campione è rappresentata da una coppia ordinata di numeri, uno per ogni variabile. Ad esempio, un insieme di dati bivariati potrebbe contenere informazioni sul reddito e sull'età di un gruppo di persone.

Un insieme di dati bivariati può essere rappresentato graficamente utilizzando un grafico di dispersione, dove i valori delle due variabili sono rappresentati sui due assi cartesiani. Ciò consente di visualizzare la relazione tra le due variabili e identificare eventuali pattern o tendenze.

1.4 Coefficiente di correlazione campionaria

Il coefficiente di correlazione campionaria è una misura di quanto due variabili aleatorie siano linearmente correlate tra loro in un campione di dati. Il coefficiente di correlazione può assumere valori compresi tra -1 e 1, dove un valore di 1 indica una correlazione positiva perfetta, un valore di -1 indica una correlazione negativa perfetta, e un valore di 0 indica assenza di correlazione.

Il coefficiente di correlazione campionaria si calcola come il rapporto tra la covarianza campionaria delle due variabili e il prodotto delle deviazioni standard campionarie delle due variabili. In formule:

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

dove n è la dimensione del campione, x_i e y_i sono i valori osservati delle due variabili nel campione, $\bar{x}$ e $\bar{y}$ sono le medie campionarie delle due variabili, e $\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ e $\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}$ sono le deviazioni standard campionarie delle due variabili.

Un valore di r vicino a 1 indica una forte correlazione positiva tra le due variabili, mentre un valore di r vicino a -1 indica una forte correlazione negativa tra le due variabili. Un valore di r vicino a 0 indica che le due variabili non sono correlate linearmente.