Statistica e Analisi dei dati

Università degli studi di Milano - Informatica

Luca Favini , Matteo Zagheno

Ultima modifica: 01/06/2024 - Codice sorgente

Statistica e Analisi dei dati

Insegnamento del corso di laurea triennale in Informatica, Università degli studi di Milano. Tenuto dal Professore Dario Malchiodi, anno accademico 2023-2024.

La statistica si occupa di raccogliere, analizzare e trarre conclusioni su dati, attraverso vari strumenti:

- Statistica descrittiva: esposizione e **condensazione** dei dati, cercando di limitarne l'incertezza;
- Calcolo delle probabilità: creazione e analisi di modelli in situazioni di **incertezza**;
- Statistica inferenziale: **approssimazione** degli esiti mancanti, attraverso modelli probabilistici;
- *Appendice: Cheatsheet Python:* raccolta funzioni/classi Python utili ai fini dell'esame *(e non)*.

Indice

1. Statistica descrittiva

1.1. Introduzione

Popolazione insieme di elementi da analizzare, spesso troppo numerosa per essere analizzata tutta
Campione parte della popolazione estratta per essere analizzata, deve essere rappresentativo
Campione casuale (semplice) tutti i membri della popolazione hanno la stessa possibilità di essere selezionati

1.2. Classificazione dei dati: qualitativi e quantitativi

1.3. Frequenze

1.3.1. Frequenze assolute e relative

1.3.2. Frequenze cumulate

1.3.2.1. Funzione cumulativa empirica

1.3.3. Frequenze congiunte e marginali

1.3.4. Stratificazione

1.4. Grafici

1.5. Indici di centralità

Sono indici che danno un'idea approssimata dell'ordine di grandezza (quindi dove ricadono) dei valori esistenti.

1.5.1. Media campionaria

Viene indicata da $\overline{x}$, ed è la **media aritmetica** di tutte le osservazioni del campione.

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

La media opera linearmente, quindi può essere scalata ($\cdot\ a$) e/o traslata ($+b$):

$$\forall i\ y_i = ax_i + b \Rightarrow \overline{y} = a\overline{x} + b$$

Non è un stimatore robusto rispetto agli outlier. Può essere calcolata solo con dati quantitativi.

1.5.2. Mediana campionaria

È il valore a **metà** di un dataset ordinato in ordine crescente, ovvero un valore $\geq$ e $\leq$ di almeno la metà dei dati.

Dato un dataset di dimensione n la mediana è:
- l'elemento in posizione $\frac{n+1}{2}$ se n è dispari
- la media aritmetica tra gli elementi in posizione $\frac{n}{2}$ e $\frac{n}{2} + 1$ se n è pari

È robusta rispetto agli outlier ma può essere calcolata solo su *campioni ordinabili.*

1.5.3. Moda campionaria

È l'osservazione che compare con la maggior frequenza. Se più di un valore compare con la stessa frequenza allora tutti quei valori sono detti modali.

1.6. Indici di dispersione

Sono indici che misurano quanto i valori del campione si discostano da un valore centrale.

1.6.1. Scarto assoluto medio

Per ogni osservazione, lo scarto è la distanza dalla media: $x_i - \overline{x}$. La somma di tutti gli scarti farà sempre 0.

$$\sum_{i=1}^{n} x_i - \overline{x} \quad = \quad \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \overline{x} \quad = \quad n\overline{x} - n\overline{x} \quad = 0$$

1.6.2. Varianza campionaria

Misura di quanto i valori si discostano dalla media campionaria

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

Metodo alternativo per calcolare la varianza:

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

> **Nota**
>
> Verrebbe intuitivo applicare il *valore assoluto* ad ogni scarto medio, ma questo causa dei problemi. Per questo motivo la differenza viene elevata al *quadrato*, in modo da renderla sempre positiva.

La varianza *non* è un operatore lineare: la traslazione non ha effetto mentre la scalatura si comporta così:

$$s_y^2 = a^2 s_x^2$$

1.6.2.1. Varianza campionaria standard

È possibile applicare alla varianza campionaria la radice quadrata, ottenendo la varianza campionaria standard.

$$s = \sqrt{s^2}$$

> **Attenzione**
>
> Applicando la radice quadrata solo dopo l'elevamento a potenza, non abbiamo reintrodotto il problema dei valori negativi: $\sqrt{a^2} \quad \neq \quad \left(\sqrt{a}\right)^2 = a$

1.6.3. Coefficiente di variazione

Valore **adimensionale**, utile per confrontare misure di fenomeni con unità di misura differenti.

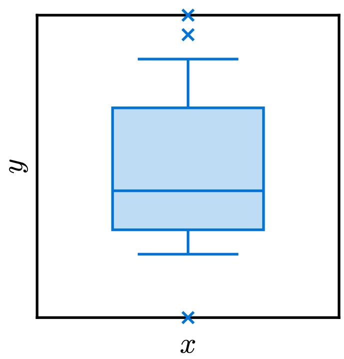
$$s^* = \frac{s}{|\overline{x}|}$$

> **Nota**
>
> Sia la varianza campionaria standard che la media campionaria sono dimensionali, ovverro hanno unità di misura. Dividendoli tra loro otteniamo un valore adimensionale.

1.6.4. Quantile

Il quantile di ordine α (con α un numero reale nell'intervallo $[0, 1]$) è un valore q_α che divide la popolazione in due parti, proporzionali in numero di elementi ad α e (1-α) e caratterizzate da valori rispettivamente minori e maggiori di q_α.

Percentile quantile descritto in percentuale
Decile popolazione divisa in 10 parti con ugual numero di elementi
Quartile popolazione divisa in 4 parti con ugual numero di elementi

> **Nota**
>
> È possibile visualizzare un campione attraverso un **box plot**, partendo dal basso composto da:
> - eventuali *outliers*, rappresentati con le x prima del baffo
> - il *baffo* "inferiore", che parte dal valore minimo e raggiunge il primo quartile
> - il *box* (scatola), che rappresenta le osservazioni comprese tra il primo e il terzo quartile
> - la linea che divide in due il box, che rappresenta la *mediana*
> - il *baffo* "superiore", che parte terzo quartile e raggiunge il massimo
> - eventuali *outliers* "superiori", rappresentati con le x dopo il baffo
>
> 
>

1.7. Indici di correlazione

Campione bivariato campione formato da coppie $\{(x_1, y_1), ..., (x_n, y_n)\}$.
Correlazione relazione tra due variabili tale che a ciascun valore della prima corrisponda un valore della seconda seguendo una certa regolarità.

1.7.1. Covarianza campionaria

È un valore numerico che fornisce una misura di quanto le due variabili varino assieme. Dato un campione bivariato definiamo la **covarianza campionaria** come:

$$\text{Cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

Metodo alternativo di calcolo:

$$\text{Cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i y_i - n\overline{x}\overline{y})$$

Informalmente

Intuitivamente c'è una **correlazione diretta** se al crescere di x cresce anche y o al descrescere di x decresce anche y, dato che il contributo del loro prodotto alla sommatoria sarà positivo. Quindi se x e y hanno segno concorde allora la correlazione sarà *diretta*, altrimenti *indiretta*.

- $\text{Cov}(x, y) > 0$ probabile correlazione diretta
- $\text{Cov}(x, y) \simeq 0$ correlazione improbabile
- $\text{Cov}(x, y) < 0$ probabile correlazione indiretta

Nota

Una relazione diretta/indiretta non è necessariamente *lineare*, può essere anche *logaritmica* o seguire altre forme.

1.7.2. Indice di correlazione di Pearson (indice di correlazione lineare)

Utilizziamo l'indice di correlazione di Pearson per avere un valore *adimensionale* che esprime una correlazione. Possiamo definirlo anche come una misura normalizzata della covarianza nell'intervallo $[-1, +1]$. ρ è **insensibile** alle trasformazioni lineari.

$$\rho(x, y) = \frac{1}{n-1} \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{s_x s_y} = \frac{s_{XY}}{s_X s_Y}$$

Dove s è la varianza campionaria standard.

- $\rho \sim +1$ probabile correlazione linearmente diretta
- $\rho \sim 0$ correlazione improbabile
- $\rho \sim -1$ probabile correlazione linearmente indiretta

Attenzione

L'indice di correlazione lineare (ρ) cattura **solo** relazioni dirette/indirette *lineari* ed è insensibile alle trasformazioni lineari.

Attenzione

La covarianza campionaria o l'indice di correlazione lineare $\simeq 0$ non implicano l'indipendenza del campione, ma è vero il contrario:

$$\text{Cov}(x, y) \simeq 0 \quad \not\Rightarrow \quad \text{Indipendenza}$$

$$\rho(x, y) \simeq 0 \quad \not\Rightarrow \quad \text{Indipendenza}$$

$$\text{Indipendenza} \quad \Rightarrow \quad \rho(x, y) \simeq \text{Cov}(x, y) \simeq 0$$

1.8. Indici di eterogeneità

Massima eterogeneità il campione è composto da tutti elementi diversi
Minima eterogeneità il campione non contiene due elementi uguali *(campione omogeneo)*

L'eterogeneità può essere calcolata anche su un insieme di dati qualitativi.

1.8.1. Indice di Gini (per l'eterogeneità)

$$I = 1 - \sum_{j=1}^{n} f_j^2$$

Dove f_j è la frequenza relativa di j ed n è il numero di elementi distinti. Quindi $\forall j, 0 \leq f_j \leq 1$. Prendiamo in considerazione i due estremi:

- eterogeneità *minima* (solo un valore con frequenza relativa 1):
$$I = 1 - 1 = 0$$
- eterogeneità *massima* (tutti i valori hanno la stessa frequenza relativa $\frac{1}{n}$ dove n è la dimensione del campione):

$$I = 1 - \sum_{j=1}^{n} \left(\frac{1}{n}\right)^2 = 1 - \frac{n}{n^2} = \frac{n-1}{n}$$

Generalizzando, I non raggiungerà mai 1:

$$0 \leq I \leq \frac{n-1}{n} < 1$$

Dal momento che l'indice di Gini tende a 1 senza mai arrivarci introduciamo l'**indice di Gini normalizzato**, in modo da arrivare a 1 nel caso di eterogeneità massima:

$$I' = \frac{n}{n-1} I$$

1.8.2. Entropia

$$H = \sum_{j=1}^{n} f_j \log\left(\frac{1}{f_j}\right) = \sum_{j=1}^{n} -f_j \log(f_j)$$

Dove f_j è la frequenza relativa e n è il numero di elementi distinti. L'entropia assume valori nel range $[0, \log(n)]$ quindi utilizziamo l'**entropia normalizzata** per confrontare due misurazioni con diverso numero di elementi distinti n.

$$H' = \frac{1}{\log(n)} H$$

Nota

In base alla base del logaritmo utilizzata, l'entropia avrà unità di misura differente:
- $\log_2$: bit
- $\log_e$: nat
- $\log_{10}$: hartley

Informalmente

Intuitivamente sia l'indice di Gini che l'entropia sono una *"media pesata"* tra la frequenza relativa di ogni elemento ed un peso: la *frequenza stessa* nel caso di Gini e il *logaritmo del reciproco* nell'entropia. La frequenza relativa è già nel range $[0, 1]$, quindi non c'è bisogno di dividere per il numero di elementi.

1.9. Indici di concentriazione

Un indice di concentrazione è un indice statistico che misura in che modo un *bene* è distribuito nella *popolazione*.

Distruzione del bene $a_1, a_2, ...a_n$ indica la quantità ordinata in modo **non decrescente**, del bene posseduta dall'individuo i

Media $\overline{a}$ indica la quantità media posseduta da un individuo

Totale $\text{TOT} = n\overline{a}$ indica il totale del bene posseduto

- Concentrazione *massima* (**sperequato**): un individuo possiete tutta la quantità $a_{1..n-1} = 0, \quad a_n = n\overline{a}$
- Concentrazione *minima* (**equo**): tutti gli individui possiedono la stessa quantità $a_{1..n} = \overline{a}$

1.9.1. Curva di Lorentz

Dati:
- $F_i = \frac{i}{n}$: posizione percentuale dell'osservazione i nell'insieme
- $Q_i = \frac{1}{\text{TOT}} \sum_{k=1}^{i} a_k$

La tupla (F_i, Q_i) indica che il $100 \cdot F_i\%$ degli individui detiene il $100 \cdot Q_i\%$ della quantità totale.

Inoltre: $\forall i 0 \leq Q_i \leq F_i \leq 1$.

1.9.2. Indice di Gini (per la concentrazione)

Dato che la curva di Lorenz non assume mai alcun valore nella parte di piano superiore alla retta che collega $(0, 0)$ a $(1, 1)$, allora introduciamo l'indice di Gini, che invece assume valori nel range $[0, 1]$.

$$G = \frac{\sum_{i=1}^{n-1} F_i - Q_i}{\sum_{i=1}^{n-1} F_i}$$

$$\sum_{i=1}^{n-1} F_i = \frac{1}{n} \sum_{i=1}^{n-1} i = \frac{1}{n} \frac{n(n-1)}{2} = \frac{n-1}{2}$$

$$G = \frac{2}{n-1} \sum_{i=1}^{n-1} F_i - Q_i$$

2. Calcolo delle probabilità

3. Statistica inferenziale

4. Cheatsheet Python

Varianza campionaria `Series.var()`

Devianza standard campionaria `Series.std()`

Indici di centralità e dispersione `Series.describe()`

Quantile `Series.quantile()`

Covarianza `Series.cov()`