Università di Pisa

Dipartimento di Informatica
Corso di Laurea Triennale in Informatica

Corso 2° anno - 6 CFU

Statistica

Professore:
Prof. Francesco Grotto

Autore:
Filippo Ghirardini

Anno Accademico 2023/2024

Contents

1 Statistica descrittiva . . . **3**
1.0.1 Campioni statistici . . . 3
1.0.2 Istogramma . . . 3
1.0.3 Indici statistici . . . 3
1.0.4 Quantili . . . 4
1.0.5 Dati multi-variati . . . 4

Statistica

Realizzato da: Filippo Ghirardini

A.A. 2023-2024

1 Statistica descrittiva

La statistica si occupa dello studio dei dati, ovvero della sua **raccolta**, **analisi** ed **interpretazione**. Le risposte dipendono dai dati e dalla conoscenza pregressa del problema, quindi da eventuali ipotesi ed assunzioni.

- Statistica **descrittiva**: quando i dati vengono analizzati senza fare assunzioni esterne per evidenziarne la struttura e rappresentarli in modo efficace
- **Inferenza statistica**: studia i dati utilizzando un modello probabilistico, ovvero supponendo che i dati siano valori assunti da *variabili aleatorie* con una certa *distribuzione di probabilità* dipendente da parametri non noti che devono essere stimati. Il modello potrà poi fare previsioni.

1.0.1 Campioni statistici

Definizione 1.0.1 (Popolazione)**.** *Insieme di oggetti o fenomeni che si vuole studiare su ognuno dei quali si può effettuare una stessa misura, ovvero un **carattere**. Può essere **ideale** o **reale**.*

Definizione 1.0.2 (Campione statistico)**.** *Un sottoinsieme della popolazione scelto per rappresentarla.*

Definizione 1.0.3 (Dati)**.** *Misure effettuate sul campione statistico.*

Definizione 1.0.4 (Frequenza)**.** *Può essere:*

- ***Assoluta**: il numero di volte in cui questo esito compare nei dati*
- ***Relativa**: frazione di volte in cui questo esito compare sul totale dei dati*

In generale dipendono dai dati e quindi non coincidono su tutta la popolazione.

Note 1.0.1. La scelta del campione in modo che sia rappresentativo è importante ma non verrà trattata.

1.0.2 Istogramma

Consiste in una serie di colonne ognuna delle quali ha per base un intervallo numerico e per area la frequenza relativa dei dati contenuti nell'intervallo.

Osservazione 1.0.1. La scelta delle ampiezze degli intervalli di base è cruciale. Un buon compromesso deve essere individuato sulla base della numerosità dei dati e sulla loro distribuzione.

Può avere varie forme:

- **Normale** se ha la forma di una *campana simmetrica*
- **Unimodale** se si concentra su una colonna più alta o **bimodale** se su due. Può essere asimmetrica a *destra* o a *sinistra* in base alla concentrazione dei dati in base al picco
- **Platicurtica** se i dati sono concentrati in un certo intervallo o **leptocurtica** se sono composti da un gruppo centrale e da molti *outliers*

1.0.3 Indici statistici

Dato un vettore $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ di dati numerici gli indici statistici sono quantità che riassumono alcune proprietà significative.

Definizione 1.0.5 (Media campionaria)**.** *La media aritmetica dei dati:*

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

Definizione 1.0.6 (Mediana)**.** *Il dato x_i tale che la metà degli altri valori è minore o uguale ad esso e l'altra metà maggiore o uguale.*

Osservazione 1.0.2. La **mediana** è utile nel caso di dati molto **asimmetrici** ed è robusta rispetto alle code delle distribuzione. Al contrario la **media campionaria** viene facilmente spostata da dati molto piccoli o grandi.

Definizione 1.0.7 (Varianza campionaria)**.** *Si usa per misurare la dispersione dei dati attorno alla media campionaria.*

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

È nulla se i dati sono tutti uguali. Possiamo mappare x diversamente:

- $x \mapsto x^2$ *misura la media dei punti della media campionaria*
- $x \mapsto x^3$ *misura la* ***sample skewness****, ovvero l'asimmetria della distribuzione*

$$b = \frac{1}{\sigma} \cdot \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3 \tag{3}$$

- $x \mapsto x^4$ *misura la piattezza della distribuzione dei dati, ovvero la* ***curtosi***

Definizione 1.0.8 (Scarto quadratico medio o deviazione standard)**.**

$$\sigma(x) = \sqrt{var(x)} \tag{4}$$

Proposizione 1.0.1. *Dato un campione di dati x ed un numero positivo d:*

$$\frac{\#\{x_i : |x_i - \bar{x}| > d\}}{n-1} \leq \frac{var(x)}{d^2} \tag{5}$$

Il termine a sinistra è la frazione di dati che differiscono dalla media campionaria più di d.

1.0.4 Quantili

Definizione 1.0.9 (Funzione di ripartizione empirica)**.** *Dato $x = (x_1, \dots, x_n) \in \mathbb{R}^n$:*

$$F_e(t) = \frac{\#\{i | x_i \leq t\}}{n} \tag{6}$$

Per ogni $t \in \mathbb{R}$ restituisce la frequenza relativa dei dati minori o uguali a t. È sempre ***non decrescente*** *e $F_e(-\infty) = 0$, $F(+\infty) = 1$.*

Definizione 1.0.10 (β-quantile)**.** *Il dato x_i tale che:*

- *almeno βn dati siano $\leq x_i$*
- *almeno $(1-\beta)n$ dati siano $\geq x_i$*

Inoltre:

- *Se βn non è intero vale $x_{(\lceil \beta n \rceil)}$*
- *Se βn è intero è la media aritmetica tra $x_{(\beta n)}$ e $x_{(\beta n+1)}$*

1.0.5 Dati multi-variati

Consideriamo coppie di dati **bivariati** del tipo

$$(x, y) = ((x_1, y_1), \dots, (x_n, y_n))$$

Definizione 1.0.11 (Covarianza campionaria)**.**

$$cov(x, y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1} \tag{7}$$

Definizione 1.0.12 (Coefficiente di correlazione)**.** *Dati $\sigma(x) \neq 0$ e $\sigma(y) \neq 0$:*

$$r(x,y) = \frac{cov(x,y)}{\sigma(x)\sigma(y)} \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{t})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{8}$$

Misura la presenza di una relazione lineare tra i dati x e y quantificata dalla ***retta di regressione****.*

Proposizione 1.0.2 (Disuguaglianza di Cauchy-Scwarz)**.**

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \leq \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{9}$$

e quindi

$$|r(x,y)| \leq 1 \tag{10}$$

La **retta di regressione** è un'approssimazione dei dati con y_i con una combinazione lineare affine a $a + bx_i$, ottenuta cercando il minimo della distanza dai dati da questa retta con i quadrati degli scarti. L'obiettivo è quindi di cercare i parametri a e b calcolando

$$\inf_{a,b\in\mathbb{R}} \sum_{i=1}^{n}(y_i - a - bx_i)^2 \tag{11}$$

Teorema 1.0.1 (Retta di regressione)**.** Se $\sigma(x) \neq 0$ e $\sigma(y) \neq 0$, esiste un unico minimo al variare di $a, b \in \mathbb{R}$ della quantità 11, dato da:

$$b^\star = \frac{(n-1)cov(x,y)}{n \cdot var(x)} \qquad a^\star = -b^\star\bar{x} + \bar{y} \tag{12}$$

e vale

$$\min_{a,b\in\mathbb{R}} \sum_{i=1}^{n}(y_i - a - bx_i)^2 = (1 - r(x,y)^2)\sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{13}$$

Quanto più $r(x,y)$ è vicino a 1, tanto più i valori tendono ad allinearsi con la retta. Se vale 1 vuol dire che i punti sono tutti sulla retta. Il segno di $r(x,y)$ corrisponde al segno del coefficiente angolare. Se è prossimo a zero allora non è una buona approssimazione.