Appunti di Probabilità e Statistica per l'Informatica

Federico Zotti

2° Semestre, 2° A.A. 2024-25
03 Mar 2025

Università degli Studi di Milano - Bicocca
CdL Informatica

Prof. Francesco Caravenna & Federica Masiero

Indice

1. Statistica descrittiva

1.1. Introduzione

Statistica arte di «imparare dai dati»

Si divide in due parti:

1. La **statistica descrittiva** descrive e riassume i dati
2. La **statistica inferenziale** trae conclusioni dai dati

1.2. Descrivere i dati

Misuriamo una certa variabile (qualitativa o quantitativa) in un campione, ottenendo un insieme di dati:

$$x_1, x_2, x_3, ..., x_n$$

con n il numero dei dati.

Se i dati sono distinti si possono rappresentare in una tabella.

Frequenza assoluta f_i è il numero di volte in cui compare un valore nell'insieme.
Frequenza relativa $p_i = \frac{f_i}{N}$.

I dati possono essere **quantitativi** se sono categorie o nomi, oppure **quantitativi** se sono numeri.

Per rappresentare le frequenze si può utilizzare un **istogramma** (grafico a barre). Esso è una rappresentazione equivalente a una tabella.

Se i valori distinti dei dati sono in numero elevato, conviene suddividere i valori in intervalli detti **classi**.

1.2.1. Dati a coppie (bivariati)

Generalmente gli insiemi di dati si riferiscono a una singola variabile. Se si misurano due dati al posto di uno, ogni dato è una coppia di numeri. Questi vengono detti **dati bivariati**:

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), ..., (x_n, y_n)$$

In questo caso al posto di usare un istogramma è meglio utilizzare un **diagramma a dispersione**, rappresentando le coppie in un piano cartesiano.

1.3. Riassumere i dati

1.3.1. Indici di posizione

1.3.1.1. Media campionaria

Per descrivere il *centro* dell'insieme dei dati, definiamo la

Media campionaria

$$\overline{x} = \frac{x_1 + x_2 + ... + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Linearità della media

Se si applica una trasformazione lineare ai dati

$$(y_i = ax_i + b)_{i=1,\dots,N} \qquad a, b \in \mathbb{R}$$

la media rimane lineare

$$\overline{y} == a\overline{x} + b$$

Con i valori z_i e le loro relative frequenze f_i la formula diventa

$$\overline{x} = \frac{z_1 f_1 + z_2 f_2 + \dots + z_M f_M}{M}$$

1.3.1.2. Mediana campionaria

Un'altra misura del *centro* dell'insieme dei dati alternativa alla media è la **mediana campionaria**.

Mediana campionaria

Avendo i dati in ordine crescente, la mediana è il valore in posizione centrale.

$$m = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$$

1.3.1.3. k-esimo percentile campionario

k-esimo percentile campionario

Fissando un numero $k \in [0, 100]$, il k-esimo percentile campionario è il valore t per il quale:

- almeno il $k\%$ dei dati è $\leq t$
- almeno il $(100 - k)\%$ dei dati è $\geq t$

Casi più importanti: $k = 25 \quad 50 \quad 75$

Scriviamo $k = 100p$, con $p = \frac{k}{100} \in [0, 1]$.

Dunque possiamo definire:

- $p = \frac{1}{4}$: $k = 25$-esimo percentile (*primo quartile*)
- $p = \frac{1}{2}$: $k = 50$-esimo percentile (*secondo quartile* o mediana)
- $p = \frac{3}{4}$: $k = 75$-esimo percentile (*terzo quartile*)

Per calcolare il k-esimo percentile t si ordina l'insieme dei dati

$$x_1 \leq x_2 \leq \dots \leq x_N$$

- Se Np non è intera, $t = x_i$ è il dato la cui posizione i è l'intero successivo a Np.
- Se Np è intera, $t = \frac{x_{Np} + x_{Np+1}}{2}$ è la media aritmetica del dato in posizione Np e il dato successivo.

Attenzione: esistono definizioni alternative di percentile nel caso in cui Np è intero (per esempio in R).