

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. List of Notes	1
2. Statistica descrittiva	2
2.1. Introduzione	2
2.2. Descrivere i dati	2
2.2.1. Dati a coppie (bivariati)	2
2.3. Riassumere i dati	2
2.3.1. Indici di posizione	2

1. List of Notes

<input type="checkbox"/> Correggere la formula (manca un pezzo)	4
<input type="checkbox"/> Aggiungere linearità della media	5
<input type="checkbox"/> Aggiungere formula	5
<input type="checkbox"/> Terminare formula	5

2. Statistica descrittiva

2.1. 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

2.2. Descrivere i dati

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

$$x_1, x_2, x_3, \dots, x_n$$

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**.

2.2.1. 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**:

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

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

2.3. Riassumere i dati

2.3.1. Indici di posizione

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

Media campionaria

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_N}{N}$$

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

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

Correggere la formula (manca un pezzo)

Aggiungere linearità della media

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.

Aggiungere formula

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

k -esimo percentile campionario

Fissando un numero $k \in [0, 100]$, il ...

Terminare formula

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

Scriviamo $k = 100p$, con $p = \frac{k}{100}$.

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*)