PES - Probabilità e Statistica per l'informatica

Elia Ronchetti

Marzo 2022

Indice

Capitolo 1

Introduzione

Il corso di probabilità e statistica per l'informatica è diviso in 2 parti

1. Stastica Descrittiva - Descrivere e riassumere i dati
    (a) Probabilità - Descrivere matematicamente i fenomeni casuali
2. Statistica inferenziale - Trarre conclusioni dai dati

Capitolo 2

Analisi Descrittiva

2.1 Descrivere i dati

Per descrivere una raccolta dati in maniera chiara e immediata è utile utilizzare una **tabella delle frequenze** all'interno della quale sono contenuti:

- Valori
- Frequenze Assolute - Numero di volte in cui compare "i" nell'insieme di dati
- Frequenze Relative - Frazione di volte in cui compare i nell'insieme di dati
- Percentuali - (Frequenza relativa x 100)

Il dato che compare con frequenza più alta è detto **moda**.
I dati possono essere

- Qualitativi
- Quantitativi

Noi useremo i dati **quantitativi**

2.1.1 Rappresentazione dei dati

Per rappresentare le frequenze (assolute o relative) risulta efficace e immediato l'utilizzo di un grafico a barre detto istogramma, esso rappresenta in graficamente la tabella, chiaramente da esso è possibile risalire alla tabella stessa. Capita di avere degli insiemi di dati che assumono un valore elevato

di valori distini, per questo conviene suddividerli in classi e determinare la frequenza di ciascuna classe. In questo modo c'è una perdita d'informazioni (sui valori specifici), ma così facendo possiamo calcolare le frequenze delle classi e avere un'idea migliore della distribuzione dei dati.

2.1.2 Dati Bivariati

Quando per ciascun individuo vengono misurate due variabili ci troviamo un insieme di N dati a coppie detti **dati bivariati**. Anche in questo caso è possibile calcolare le frequenze, in questo caso detto **frequenze congiunte**.

è possibile, inoltre, misurare la correlazione tra le due variabili attraverso per esempio un diagramma di dispersione (detto anche scatterplot).

Correlazione non significa causalità! Non è detto che l'aumento di una variabile causi la diminuzione dell'altra o viceversa, potrebbe esserci una causa comune.

2.2 Riassumere i dati

Dopo aver rappresentato i dati vogliamo ora riassumerli mediante quantità numeriche, dette **Statistiche Campionarie**, al fine di sintetizzare le proprietà salienti dei dati.

2.2.1 Indici di posizione

Per definire il centro dell'insieme dei dati definiamo la

Media Campionaria $\dfrac{x_1+x_2+...+x_n}{N}$

Per misurare il valore in posizione centrale (considerando l'insieme di dati ordinato), utilizziamo la

Mediana

- Se N dispari - $X_{\frac{N+1}{2}}$
- Se N pari - $m = \frac{x_{\frac{N}{2}}+x_{(\frac{N}{2}+1)}}{2}$

La mediana è insesibile alle code

2.3 Coefficiente di correlazione lineare

Posso misurare il grado di correlazione tra una coppia di dati attraverso il coefficiente di correlazione lineare.

$$r = \frac{\sum_{k=1}^{N}(x_i - x)(y_i - y)}{(N-1)S_x S_y} \tag{2.1}$$

Si può mostrare che:

$$-1 <= r <= 1 \tag{2.2}$$

In generale $r > 0$ indica una correlazione positiva
$r < 0$ indica una correlazione negativa

2.3.1 Correlazioni significative

$|r| > 0.7$ Correlazione significativa
$|r| < 0.3$ Correlazione debole

2.4 Percentili e quantili

Per analizzare la distribuzione dei dati è utile fissare un numero k che rappresenta la posizione all'interno dato all'interno dell'insieme questo valore percentuale è detto **k-esimo Percentile Campionario**, valore t per cui

- almeno il k% dei dati è $<= t$
- almeno il $(100 - k)\%$ dei dati è $<= t$

I casi più importanti sono per k = 25, 50, 75
Risulta pratico scrivere $k = 100p$ dove $p = \frac{k}{100} \in [0, 1]$, dove i casi importanti sono per:

- $p = \frac{1}{4} : k = 100p$ = 25-esimo percentile = primo quartile q_1
- $p = \frac{1}{2} : k = 100p$ = 50-esimo percentile = secondo quartile q_2 = mediana m
- $p = \frac{3}{4} : k = 100p$ = 75-esimo percentile = terzo quartile q_3

Per calcolare il k-esimo percentile t è necessario:

1. Ordinare l'insieme di dati $x_1 <= x_2 <= \ldots <= x_n$

2. Se N_p non è intera $t = x_i$ è il dato la cui posizione i è l'intero successivo a N_p

3. Se N_p è intera $t = \frac{x_(Np)x_(Np+1)}{2}$ è la media aritmetica fra il dato in posizione N e il successivo

Nota per R Esistono diverse definizioni di quantile, R èer esempio ne utilizza una diversa di default

É possibile utilizzare i **Boxplot** per la rappresentazione dei quantili

Capitolo 3

Probabilità

3.1 Introduzione

Il calcolo delle probabilità è la teoria matematica che permette di descrivere e studiare esperimenti aleatori

Esperimento aleatorio $\rightarrow$ Fenomeno il cui esito non è prevedibile a priori

Capitolo 4

Esame

L'esame sarà strutturato nella seguente maniera

Parte 1 - Teoria 8 Domande a risposta multipla - Punteggio 10/30

Parte 2 - Pratica 4 Esercizi a risposta aperta - Punteggio 20/30

Progetto (facoltativo) Progetto R, da consegnare prima dell'esame, può fornire un massimo di 2/30