

- Probabilità e Statistica -

Federico Brutti

March 11, 2025

Inserire citazione inerente alla materia

Contents

1	Introduzione	5
2	Statistica Descrittiva	7
2.1	Organizzazione e descrizione dei dati	7
2.2	Grandezze per la sintesi dei dati	8
2.3	Campioni normali e correlazione	11
2.4	Riepilogo grafici e tabelle	12
2.5	Appunti	12
3	Probabilità	13
3.1	Elementi di probabilità	13
3.2	Calcolo della probabilità	14
3.3	Probabilità condizionata	16
3.4	Variabili aleatorie	16
3.5	Distribuzioni congiunte	16
3.6	Classi notevoli di variabili aleatorie	16
3.7	Statistiche campionarie	16
4	Statistica Inferenziale	17
5	Regressione	19
6	Il Linguaggio R	21
6.1	Introduzione	21
6.2	Funzioni di base	22
6.2.1	Variabili e funzioni di uso generale	22
6.2.2	Vettori e matrici	22

Chapter 1

Introduzione

La statistica si occupa della raccolta, descrizione ed analisi dei dati e ci aiuta a trarre delle conclusioni in base a quanto ottenuto.

Anzitutto, allo statista è richiesta l'ideazione dell'algoritmo ideale di valutazione per la raccolta dei dati, dopodiché, dato un sottoinsieme della **popolazione**¹, si effettuano delle **inferenze**, le quali saranno poi **descritte** mediante appositi grafici e tabelle.

Queste ultime due parole in neretto non sono evidenziate a caso, infatti distinguono le due parti della statistica, nostro oggetto di studio:

- **Statistica descrittiva**; Si occupa dell'illustrazione e sintetizzazione dei dati.
- **Statistica inferenziale**; Si occupa della ricerca e l'ottenimento dei dati.

Ci concentreremo poi sullo calcolo della **probabilità**, concetto strettamente legato alla statistica, in quanto ci consente di fare assunzioni sul risultato di un dato evento, come il lancio di un dado. Definiamo l'insieme di tali ipotesi come **modello probabilistico** e risulta utile per definire non solo le aspettative, ma anche per capire quali siano i risultati probabili dell'evento.

L'esame sarà di tipo informatizzato e comprenderà una parte di teoria come una parte di lavoro con il linguaggio di programmazione R.

¹Indicato con M , si tratta dell'insieme più grande che contiene ogni elemento. Presenta inoltre le caratteristiche reali, oggetto di studio ultimo degli statisti.

Chapter 2

Statistica Descrittiva

2.1 Organizzazione e descrizione dei dati

Repetita iuvant, la statistica descrittiva si occupa dei metodi di esposizione e sintesi dei dati. Si presuppone che questi siano rappresentati chiaramente ed esistono metodi standard come i seguenti:

Questi grafici svolgono la medesima funzione e sta al singolo capire quale sia il più adatto per mostrare le tendenze di un dato fenomeno. Osservando le immagini possiamo concludere che esistono due tipi di variabili: **numeriche**, che mostrano un dato in forma di numero e **categoriche**, le quali rappresentano una caratteristica. A partire da ciò, possiamo introdurre i concetti di:

- **Frequenza assoluta**; Occorrenze di un valore.
- **Frequenza relativa**; Rapporto fra la frequenza assoluta ed il numero di osservazioni effettuate.
- **Frequenza percentuale**; La frequenza relativa moltiplicata per 100.

In particolare, se si nota una certa pattern sulle variabili categoriche di un dato campione, è possibile utilizzarle per effettuare studi di correlazione, mentre per quanto riguarda le numeriche abbiamo una struttura più complessa. Queste infatti possono assumere due forme in un dato campione o intervallo:

- **Forma discreta**; Se assumono un singolo valore finito, come il numero degli studenti in una classe.
- **Forma continua**; Se possono assumere qualsiasi valore possibile, come altezza, età o temperatura.

Creando i grafici in base alle variabili ottenute, è possibile dare delle interpretazioni, come le **simmetriche**, **modali** o **bimodali**. Fondamentalmente si parla solo del modo in cui i dati sono mostrati. Seguono esempi:

Ora abbiamo tutti gli strumenti di base per effettuare calcoli statistici e mostrarli di conseguenza.

2.2 Grandezze per la sintesi dei dati

Piuttosto che buttarci a capofitto nella scrittura dei dati, è necessario capire in che modo essi devono essere presentati; infatti anche la statistica richiede una scrittura matematica formale. A partire da un dato campione di dati (x_1, x_2, \dots, x_n) abbiamo:

- **Media campionaria;** La semplice media aritmetica dei valori.

Esempio 1. *Calcolo della media aritmetica*

Somma ogni valore e dividi il risultato per il totale dei numeri nell'insieme.

Dato l'insieme numerico $(1, 2, 3)$

$$\text{La media è: } \frac{1 + 2 + 3}{3} = 2.$$

- **Mediana campionaria;** Il valore centrale, assumendo che i dati siano scritti in ordine crescente.

Esempio 2. *Calcolo della mediana se cardinalità dispari*

Ordina i tuoi valori in ordine crescente. In questo caso non è necessario svolgere calcoli, prendi direttamente il valore al centro.

Dato l'insieme numerico $(1, 2, 3)$

La mediana è: 2.

Esempio 3. *Calcolo della mediana se cardinalità pari*

Ordina i tuoi valori in ordine crescente e prendi i due centrali. Effettuando la media aritmetica fra di loro otterrai la tua mediana.

Dato l'insieme numerico $(1, 2, 3, 4)$

$$\text{La mediana è: } \frac{2 + 3}{2} = 2,5.$$

- **Moda campionaria;** Il valore che compare più frequentemente. Se più mode sono presenti, si dicono **valori modali**.

Esempio 4. Calcolo della moda

Dato l'insieme numerico (1, 2, 2, 3, 5, 7)

La moda è: 2.

Queste tre misure danno informazioni in merito al valore attorno al quale si posizionano i dati. Tuttavia è possibile che questi compaiano anche in modo sparso, ed è per questo che hanno introdotto gli **indici di dispersione**, i quali hanno lo scopo opposto, ovvero di mostrare quanto i dati si disperdano intorno ad un dato valore centrale. Quelli utili al nostro studio sono:

- **Varianza campionaria;** La media aritmetica del valore della distanza dei dati dalla media campionaria elevato al quadrato.

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

Dove gli elementi nella formula sono:

- n : Numero di elementi nell'insieme.
- x_i : Un elemento dell'insieme.

Esempio 5. Calcolo della varianza campionaria

Dato il campione (3, 4, 6, 7, 10), calcoliamo prima la media

$$\bar{x} = \frac{3 + 4 + 6 + 7 + 10}{5} = 6$$

Applichiamo ora la formula per un valore:

$$s_{x_1}^2 = \frac{1}{5-1} (3-6)^2 = \frac{(-3)^2}{4}$$

Applica lo stesso procedimento per tutti gli altri. La varianza campionaria è:

$$s^2 = \frac{[(-3)^2 + (-2)^2 + 0^2 + 1^2 + 4^2]}{4} = 7,5$$

- **Deviazione standard campionaria;** La radice della varianza campionaria. Mantiene l'unità di misura iniziale.

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

Quando si lavora coi grafici, risulta utile avere dei checkpoints per delimitare i dati in percentuali; la funzione è svolta dagli **indici di posizione relativi**. Ne esistono due tipi:

- **Percentili;** Diciamo tale un valore p ($0 \leq p \leq 100$), il quale è maggiore di una percentuale p dei dati e minore della restante percentuale $100 - p$. Se questo dato risulta unico (relativo), allora diremo che è il *percentile p -esimo* dell'insieme. Se invece non è unico (intero), allora sono esattamente due valori ed il percentile effettivo è dato dalla loro media aritmetica.

Esempio 6. Calcolo del p -esimo percentile

Dato l'insieme ordinato delle 25 città più popolate d'America, calcolare il 10° e l'80° percentile. Per calcolarli, abbiamo già a disposizione che $n = 25$, ovvero la numerosità (totale degli elementi) dell'insieme. I percentili sono invece rispettivamente $p_1 = 0,1$ e $p_2 = 0,8$. Abbiamo ora tutti i dati che ci servono.

Ricerchiamo la posizione da prendere per entrambi:

$$np_1 = 25 \times 0,1 = 2,5, \quad np_2 = 25 \times 0,8 = 20$$

Per p_1 il 10° percentile è il terzo dato più piccolo per arrotondamento per eccesso.

Per p_2 , siccome è un numero intero, l'80° percentile è la media degli elementi in posizioni 20, 21 a partire dai più piccoli.

- **Quartili;** Questi sono come dei percentili notevoli. Separano in quattro parti un campione numerico. Questi sono il **25°**, **50°**, corrispondente alla mediana campionaria, ed il **75°**; vengono chiamati rispettivamente primo, secondo e terzo quartile. Inoltre, la differenza fra il primo ed il terzo quartile viene detta **scarto interquartile**.

Per rappresentare al meglio i percentili si utilizza un grafico **boxplot**, il quale introduce anche il concetto di **outliers**, ovvero valori estremamente piccoli o grandi rispetto al resto dei dati. La media ne è particolarmente suscettibile, ed è per questo che si tende a preferire la mediana.

2.3 Campioni normali e correlazione

Il concetto di pattern-recognition aiuta molto nello studio dei dati. Sarà capitato infatti di osservare grafici, in particolare istogrammi, che presentano qualche somiglianza, oppure che i dati prendano la forma di una curva. Ciò non è casuale, infatti esiste addirittura un tipo di grafico che si presenta spesso, dalle seguenti caratteristiche:

- Presenta un solo massimo ed è in corrispondenza della mediana.
- Decresce da ambo i lati simmetricamente, creando una curva a campana.

Sotto queste restrizioni possiamo dichiarare un dato campione **normale**, il quale ha la tendenza ad avere media e mediana con valori simili. Ovviamente avere grafici perfettamente simmetrici risulta impossibile, quindi si tende ad approssimare, ma esistono anche altre forme come la **skewed form**, ovvero che presenta una curva più ripida da una parte, oppure la **bimodale**, che presenta due massimi, quindi una curva che ricorda le gobbe di un cammello.

Capiterà poi di dover lavorare con sequenze di coppie di numeri; in tal caso risulta utile l'utilizzo di uno **scatter plot**, o grafico di dispersione. Il pregio in primis di questa rappresentazione è la possibilità di vedere se esiste una correlazione fra i dati raccolti, e se è così, sarà possibile notare che i punti nel grafico prenderanno (circa) la forma di una retta. Il concetto di correlazione si può infatti ricondurre ad una funzione lineare.

Ma in che modo possiamo dire che i dati sono correlati? Ebbene, esiste un coefficiente apposito, dalla formula particolarmente dolorosa.

Definizione 1. Coefficiente di correlazione campionaria

Sia dato un campione bivariato (x_i, y_i) , dove $i \in \mathbf{N}$, con medie campionarie \bar{x}, \bar{y} e deviazioni standard campionarie s_x, s_y per i soli dati x, y rispettivamente. Allora si dice coefficiente di correlazione campionaria r la quantità:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

Il coefficiente può assumere solo la forma di $-1 \leq r \leq 1$. Più il valore è alto, più positivamente sono correlati i dati, altrimenti si dicono correlati negativamente.

2.4 Riepilogo grafici e tabelle

Line graph, grafico a barre, grafico a linee, box plot, scatter plot e tant'altro.

2.5 Appunti

Chapter 3

Probabilità

3.1 Elementi di probabilità

La probabilità è una branca della matematica che si occupa dello studio e descrizione degli **esperimenti aleatori**, ovvero delle inferenze il cui esito non è del tutto prevedibile. Esistono due metodi per l'espressione del concetto di probabilità:

- **Approccio frequentista**; Determinazione della probabilità mediante esperimenti ripetuti. Risulta quindi come il rapporto fra il totale in cui si è verificato un esito e il totale degli esperimenti.
- **Approccio soggettivista**; Dove la probabilità è vista come un livello di fiducia nel verificarsi di un dato esito. È una roba da filosofi, non fa per noi.

Abbiamo parlato di una totalità di esperimenti; questi vengono formalmente chiamati **eventi** E e detengono informazioni riguardo al loro esito. Ogni evento è un sottoinsieme dello **spazio campionario** S , che li comprende tutti.

Esempio 7. Spazio campionario

Un esempio di spazio campionario è dato dalla totalità dei valori delle facce di un dado, mentre gli eventi sono i singoli valori usciti da un esperimento.

$$S = \{1, 2, 3, 4, 5, 6\}, E = \{4\}$$

Alle operazioni logiche sulle affermazioni corrispondono quelle insiemistiche, le quali si mostrano mediante diagrammi di eulero venn. Siano $A, B \subseteq S$ due eventi:

- **Intersezione**
Quando "Avviene A e avviene B "
- **Unione**
Quando "Avviene A oppure B "
- **Sottrazione**
Quando "Avviene A , ma non B "
- **Complementare**
Quando "Non avviene A "

Inoltre, se gli insiemi A, B sono tali che la loro intersezione sia vuota, si dicono **incompatibili**.

3.2 Calcolo della probabilità

Fortunatamente esiste una concezione standard sulle caratteristiche assunte dalla probabilità. Associamo infatti ad ogni evento E sullo spazio campionario S , un valore denotato con $P(E)$, detto **probabilità dell'evento E** . Il comportamento della funzione è dato dai seguenti **assiomi di Kolmogorov**:

Definizione 2. Assiomi di Kolmogorov

1. $P(A)$ è un valore compreso fra 0 e 1.
2. $P(S) = 1$.
3. Se A e B sono incompatibili, allora $P(A \cup B) = P(A) + P(B)$.

Questi detengono inoltre le seguenti proprietà:

- Siano A, B due eventi tali che $A \subseteq B$, allora:

$$\begin{cases} B = S \implies S/A = A^c \implies P(A^c) = 1 - P(A) \\ P(B/A) = P(B) - P(A) \end{cases}$$

- Se A_1, A_2, \dots, A_k sono eventi a due a due incompatibili, quindi disgiunti, allora:

$$P\left(\bigcup_{i=1}^k A_i\right) = \sum_{i=1}^k P(A_i)$$

- Siano A, B due eventi generici, allora:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Un primo caso di studio per la probabilità è il suo calcolo ad **esiti equiprobabili**; ciò significa che ogni evento ha la stessa chance di avvenire rispetto agli altri. L'esempio classico è il lancio di un dado; implicando che questo non sia truccato, ogni faccia ha $\frac{1}{6}$ di possibilità di uscire. Formalmente la definiamo con la seguente scrittura:

$$P(A) = \frac{|A|}{|S|}$$

Esempio 8. Quali sono le probabilità che lanciando due volte un dado esca il valore 7?

Innanzitutto dobbiamo chiederci quale sia lo spazio campionario e gli eventi. Sappiamo che è un dado, quindi avremo rispettivamente:

- $S = \{1, 2, 3, 4, 5, 6\}$
- $E_1 = \{1\}, \dots, E_6 = \{6\}$

Ora, potremmo fare bruteforcing facendoci del male, ma il trucco per questi esercizi (entro certi limiti) è disegnare una tabella dei risultati, prendere il totale di quante volte si presenta il valore richiesto e poi applicare la formula dell'approccio frequentista. In questo caso:

-	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Notiamo che il valore 7 compare 6 volte ed il totale degli esiti ottenibili è $6 \times 6 = 36$. Il risultato sarà dato quindi da:

$$\frac{6}{36} = \frac{1}{6}, \text{ soluzione dell'esercizio.}$$

3.3 Probabilità condizionata**3.4 Variabili aleatorie****3.5 Distribuzioni congiunte****3.6 Classi notevoli di variabili aleatorie****3.7 Statistiche campionarie**

Chapter 4

Statistica Inferenziale

Pallw

Chapter 5

Regression

Pallw

Chapter 6

Il Linguaggio R

6.1 Introduzione

R è linguaggio di programmazione. Viene utilizzato per l'analisi statistica dei dati e la loro visualizzazione, statistica descrittiva, machine learning, manipolazione dei dati, datamining e anche nella ricerca scientifica.

Si tratta di un codice open-source, la cui sintassi è estremamente flessibile e a tratti simile a C. R è poi diviso in pacchetti, ovvero insiemi di funzioni create da altri utenti. In particolare vedremo:

- **ggplot2**; usato per la creazione dei grafici.
- **dplyr**; usato per la manipolazione dei dati.

I files con dominio ".r" sono detti **R-Scripts**; infatti ciò che andremo a scrivere e far elaborare dal software sono fondamentalmente dei testbench per ritornare determinati risultati o comportamenti. Per lavorare su di essi è necessario l'uso del terminale, e dove è possibile utilizzare l'interfaccia di R innata, è consigliato, e così faremo nel corso, l'ambiente di sviluppo **R Studio**. Spiegare un IDE esula dallo scopo della dispensa, arrangiati e leggi le relative documentazioni.

Installato R Studio avremo davanti il terminale, l'ambiente dei dati e una terza interfaccia variabile per i grafici, manuale e tant'altro. Abbiamo nominato dei pacchetti con i quali andremo a lavorare; questi devono essere installati con il seguente comando:

```
> install.package("nomePacchetto")
```

Abbiamo tutto e adesso possiamo iniziare a lavorare col linguaggio.

6.2 Funzioni di base

6.2.1 Variabili e funzioni di uso generale

esempi di codice, variabili, operazioni, media (mean, trim)/moda/mediana, ?funz, n:k (sequenze), NA, na.rm,

6.2.2 Vettori e matrici