- Probabilità e Statistica -

Federico Brutti

March 24, 2025

La generosità è una variabile aleatoria che si paga col denaro. - Franco Z.

Contents

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.

[1] Indicato 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 $(x_1, x_2, ..., x_n)$ 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.

  $$\textit{Dato l'insieme numerico } (1, 2, 3)$$
  $$\textit{La media è: } \frac{1+2+3}{3} = 2.$$

- **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.

  $$\textit{Dato l'insieme numerico } (1, 2, 3)$$
  $$\textit{La mediana è: } 2.$$

  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.

  $$\textit{Dato l'insieme numerico } (1, 2, 3, 4)$$
  $$\textit{La mediana è: } \frac{2+3}{2} = 2,5.$$

- **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 nosrto studio sono:

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

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

  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*

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

  Applichiamo ora la formula per un valore:

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

  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.

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

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 popolose 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:
  $np_1 = 25 \times 0,1 = 2,5$*,* $np_2 = 25 \times 0,8 = 20$

  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 $\overline{x}, \overline{y}$ 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:

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

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.

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 esperito 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 na 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 $A, B \subseteqq S$ 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 $P(A \cup B) = P(A) + P(B)$.*
4. *Siano A, B due eventi tali che $A \subseteqq B$, allora:*

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

5. *Se $A_1, A_2, ..., A_k$ sono eventi a due a due incompatibili, quindi disgiunti, allora:*

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

6. *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:

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

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:*

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

Puta caso, devi lavorare con una quantità di dati abnorme; utilizzare la tabella precedente per analizzare è impensabile, hai una vaga idea di quanto grande verrebbe? Viene ad aiutarci il seguente ragionamento, scrivibile tramite **coefficienti binomiali**:

Esempio 9. ***Calcolo combinatorio con coefficiente binomiale***
Diciamo di avere una gara a cui partecipano 10 atleti. In quanti modi possiamo assegnare i vari posti del podio? Avremo:

- *10 modi per il primo posto.*
- *9 modi per il secondo, in quanto il primo è già stato assegnato.*
- *8 modi per il terzo, per la medesima ragione.*

Supponiamo di voler uccidere tutti quelli che perdono e che quindi considereremo solo i tre posti del podio. Allora quali sarebbero i possibili esiti?

$$ABC,\ ACB,\ CAB,\ CBA,\ BAC,\ BCA,\ \text{quindi}\ 3 \times 2 \times 1 = 6\ \text{esiti.}$$

Questo calcolo si può esprimere più facilmente mediante l'utilizzo dei fattoriali, Gli esiti totali possibili saranno quindi:

$$\frac{(10 \times 9 \times 8)}{(3 \times 2 \times 1)} = \frac{10!}{7! \times 3!}$$

Perché è sbucato fuori un $7!$ *dal nulla? Ebbene, quelli sono tutti i numeri che non ci interessano, in quanto vogliamo solamente i posti del podio.*

Da questo esempio traiamo dunque una formula generale, in inglese chiamata **n choose k**, la quale ha due varianti dipendentemente se ci interessa (caso 1) o meno (caso 2) l'ordine dei dati:

$$\begin{cases} \dfrac{n!}{(n-k)!k!} \\ \dfrac{n!}{(n-k)!} \end{cases} \implies \binom{n}{k}$$

3.3 Probabilità condizionata

Finora abbiamo utilizzato l'approccio frequentista per il calcolo delle probabilità di un evento, considerandole come a loro stanti. È tuttavia possibile che la probabilità di un evento A possa essere influenzata da un altro B. Chiamiamo questo concetto **probabilità condizionata** e prende la formula matematica:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Personalmente leggo la formula come "probabilità di A sotto B". Quest'ultimo evento può quindi influenzare il primo positivamente, aumentandone la probabilità, oppure negativamente, diminuendola. Un'altra particolarità riguarda lo spazio campionario; essendo che stiamo valutando un'istanza dove B avviene sicuramente, sarà proprio questo lo spazio. Possiamo infatti spaccare le istanze:

- Succedono A e B; quindi la probabilità condizionata.
- Succede A, ma non B; quindi la probabilità senza influenze.

È necessario conoscere i valori di entrambe le istanze per il calcolo della probabilità effettiva, infatti compone la seguente:

Definizione 3. ***Formula delle probabilità totali***
Formula utilizzata per il calcolo delle probabilità di un evento il cui esperire è condizionato da un altro. Siano A, B due eventi generici.

$$P(A) = (A|B)P(B) + P(A|B^c)P(B^c)$$

Dove il primo addendo rappresenta la probabilità condizionata ed il secondo quella di A a sé stante.

Esempio 10. ***Calcolo di probabilità condizionata***
Siano due urne tali che:

- *A contiene 2 palline rosse e 4 verdi.*
- *B contiene 3 palline rosse e 2 verdi.*

Si lancia ora un dado; se esce 6 si estrae da A, altrimenti da B. Calcolare la probabilità di estrarre una pallina verde.
Introduciamo i seguenti due eventi in base al risultato del dado:

1. *E, Il dado mostra 6.*
2. *F, La pallina estratta è verde.*

Potremmo elencare ogni singola permutazione dati i pochi casi, ma useremo la formula della probabilità totale per pulizia. Attualmente deteniamo i seguenti dati:

- *$P(F|E) = \frac{4}{6}$, date le 6 palline di A, di cui 4 verdi.*
- *$P(F|E^c) = \frac{2}{5}$, date le 5 palline di B, di cui 2 verdi.*

- $P(E) = \frac{1}{6}$, *probabilità del dado di far uscire* 6.
- $P(E^c) = \frac{5}{6}$, *ogni altro numero del dado.*

Ciò che abbiamo è sufficiente per utilizzare la formula della probabilità totale. Risulterà infatti:

$$P(F) = P(F|E)P(E) + P(F|E^c)P(E^c) = \frac{4}{6} \times \frac{1}{6} + \frac{2}{5} \times \frac{5}{6} = \frac{4}{9}$$

Se pensi che sia possibile ottenere algebricamente le altre probabilità della formula, hai avuto un'ottima idea. Infatti per le probabilità condizionate abbiamo un nome apposito, che è:

Definizione 4. ***Formula e teorema di Bayes***
Necessaria per il calcolo della singola probabilità condizionata di un evento.

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

Sia ora $B_1, B_2, ..., B_n$ una partizione dello spazio campionario. Ne segue il teorema:

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{n} P(A|B_j)P(B_j)}$$

Esempio 11. ***Calcolo di probabilità con formula di Bayes***
Abbiamo un esame a 4 risposte multiple. Gli studenti iscritti si dividono in:

- *Preparato, corrispondente all'80% del totale. Risponde correttamente al 90%.*
- *Impreparato, il 20% rimanente, che risponde a caso. Quindi hanno un 25% di azzeccare la risposta.*

Qual è la probabilità di prendere l'esame di uno studente preparato fra tutti?

Definiamo gli eventi come A, ovvero che lo studente sia preparato, e B, quella di azzeccare una risposta, che è necessariamente condizionata dal primo evento. Elenchiamo i dati che abbiamo fin da subito:

- $P(A) = 0,8$
- $P(A^c) = 0,2$
- $P(B|A) = 0,9$

- $P(B|A^c) = 0,25$

Dobbiamo trovare $P(A|B)$, ma procediamo per passi. Innanzitutto ci serve $P(B)$, ovvero la probabilità di azzeccare la risposta in generale. Usiamo la formula delle probabilità totali:

$$P(B) = P(B|A)P(A) + P(B|A^c)P(A^c) = 0,9 \times 0,8 + 0,25 \times 0,2 = 0,72 + 0,05 = 0,77$$

Ora possiamo muoverci facendo delle asserzioni algebriche. Considera che $P(A|B)P(B) = P(A \cap B) = P(B \cap a) = P(B|A)P(A)$, da cui otteniamo la formula di Bayes:

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

La quale, sostituendo le variabili con i loro rispettivi valori, ci darà il risultato:

$$P(A|B) = \frac{0,72}{0,77} = 0.935$$

E se invece avessimo due eventi completamente **indipendenti**? Questi si dicono tali se, dati per esempio A, B, vale la relazione:

$$P(A \cap B) = P(A)P(B)$$

Per esempio, se lancio due dadi, la probabilità che escano i valori 6 e 5 separatamente è $\frac{1}{36}$, perché mi va bene una sola combinazione. Infatti:

$$P(A \cap B) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$$

Abbiamo appurato che i due dadi non si influenzano fra di loro. Se ci fossero invece tre eventi avremmo le seguenti relazioni, dati A, B, C:

- $P(A \cap B \cap C) = P(A)P(B)P(C)$
- $P(A \cap B) = P(A)P(B)$
- $P(A \cap C) = P(A)P(C)$
- $P(B \cap C) = P(B)P(C)$

Ovviamente, per casi richiedenti più di tre eventi, si dovranno verificare le istanze per tutti i successivi.

3.4 Variabili aleatorie

Le variabili aleatorie sono quantità numeriche il cui valore dipende dall'esito di un esperimento aleatorio; si indicano con X, Y, Z. La loro primaria utilità sta nel consentirci di poter considerare un risultato specifico al posto di ogni singolo evento.

Per esempio, diciamo di lanciare due dadi e voler sapere la somma dei valori dei numeri che escono. Allora la variabile X potrà assumere un valore dell'evento x tale che:

$$[(x \in S) \wedge (2 \leq x \leq 12)], \text{ dove } S \text{ è lo spazio campionario.}$$

Possiamo inoltre ragionarci con la probabilità. Diciamo di voler effettuare un numero n di lanci e che ci interessi vedere le chances che esca un dato numero. In questo caso X sarà il valore totale dei successi ottenuti e la probabilità sarà data da $P(X = x)$.

Di variabili aleatorie ne esistono due tipi:

- **Discrete**; Se i valori che può assumere sono finiti o al più numerabili; quindi in un insieme $S = \{x_1, x_2, ..., x_n, ...\}$. Da questo spazio sappiamo che il calcolo della sua probabilità si effettua con la **funzione di massa**:

  $$p(x) = P(X = x), \text{ dove grazie all'algebra vale } p(x_1), p(x_2), ..., p(x_n)$$

  Siccome stiamo considerando valori reali, è possibile che X non possa assumere ogni numero. In questo caso il valore della probabilità dell'esito non possibile sarà, ovviamente, zero. Agli antipodi sta la probabilità massima, data dalla somma di tutte le chances di ogni esito. Sarà uguale ad uno, quindi $\sum_{x \in R} p(x) = 1$.

- **Continue**; Se esiste una funzione $f(x)$ detta **densità** della variabile aleatoria tale che per ogni insieme $A \subseteq R$ si ha:

  $$P(X \in A) = \int_A f(x)dx$$

  La ragione per cui è richiesto un integrale è che questo tipo di variabile aleatoria può assumere un range di valori reali. Essendo loro pressoché infiniti, rendendo anche la probabilità di esperirne uno solo infima, è necessario considerarli come una collettività. Altri casi sono:

$$\int_a^b f(x)dx = P(a \leqq X \leqq b), \int_0^{+\infty} f(x)dx = P(X \geqq 0), \int_{-\infty}^{+\infty} f(x)dx = 1$$

Ora possiamo introdurre un nuovo concetto importante:

Definizione 5. ***Valore atteso***
Si tratta della media pesata dei possibili valori che X può assumere e si scrive:

- *Per variabili discrete:*

$$E(X) = \sum_{x \in R} xp(x) = x_1p(x_1) + x_2p(x_2) + ... + x_np(x_n)$$

- *Per variabili continue:*

$$E(X) = \int_{-\infty}^{+\infty} xf(x)dx$$

Per eventuali diverse forme di X basta sostituire la forma alle x_i piccole. Per esempio, sostituiremo X^2 come x^2 dove stanno tutte le x nella formula.

Esempio 12. ***Calcolo del valore atteso della variabile aleatoria discreta*** X^2

$$E(X^2) = \sum_{x \in R} x^2p(x) = x_1^2p(x_1) + x_2^2p(x_2) + ... + x_n^2p(x_n)$$

Questa formula viene con delle proprietà, le quali sono differenti per i due tipi di variabili in gioco:

- $E(aX + b) = aE(X) + b$, dove $a, b \in \mathbb{R}$.
- Con X, Y variabili aleatorie dipendenti da uno stesso esperimento: $E(X + Y) = E(X) + E(Y)$.
- Valore atteso di una funzione di variabile aleatoria: $E(g(X)) = \sum_{x \in R} g(x)p(x)$.

Per le variabili continue:

- $$E(X) = \int_{-\infty}^{+\infty} xf(x)dx$$

- $$E(g(X)) = \int_{-\infty}^{+\infty} g(x)f(x)dx$$

In tal merito, sia una variabile aleatoria X con il rispettivo valore atteso μ, quindi $E(X) = \mu$. Da qui possiamo ottenere il concetto di **varianza**, ovvero il valore atteso degli scarti.

$$Var(X) = E[(X - \mu)^2] = E(X^2) - \mu^2 \text{ (grazie all'algebra)}$$

Per calcolare questo valore abbiamo bisogno di due elementi: il valore atteso $E(X)$ e al quadrato $E(X^2)$. I calcoli sono diversi in base al tipo di variabile preso in esame:

1. Valore atteso grado 1:

$$\mu = E(X) = \begin{cases} \sum_{x \in R} xp(x) \\ \int_{-\infty}^{+\infty} xf(x)dx \end{cases}$$

2. Valore atteso al quadrato:

$$E(X^2) = \begin{cases} \sum_{x^2 \in R} x^2p(x) \\ \int_{-\infty}^{+\infty} x^2f(x)dx \end{cases}$$

Notare infine come ultima cosa che la varianza è un valore compreso fra 0 e 1.

3.5 Distribuzioni congiunte

Può capitare di lavorare con variabili dipendenti da uno stesso esperimento. In tal caso avremo una funzione di massa **congiunta** scritta come:

$$p_{X,Y}(x, y) = P(X = x, Y = y), \text{ con } X, Y \text{ variabili aleatorie discrete.}$$

Per esempio, lanciamo un dado e definiamo le due variabili aleatorie discrete:

- X = punteggio più basso.
- Y = punteggio più alto.

Ne deriva necessariamente che, volendo sapere la probabilità che esca un certo numero:

- $P_{X,Y}(1,1) = P(X = 1, Y = 1) = \frac{1}{36}$, perché in ambo i lanci è uscito 1.
- $P_{X,Y}(1,2) = P(X = 1, Y = 2) = \frac{1}{18}$, perché stai sommando le probabilità che escano i numeri nell'ordine $(1,2), (2,1)$. Ciò risulta ovviamente più probabile piuttosto che due numeri singoli.

Un'altra cosa importante è che ci è possibile ottenere la funzione di massa di una singola variabile a partire da quella congiunta[1], infatti:

- $P(X = x) = p(x) = \sum_{y \in \mathbb{R}} p_{X,Y}(x, y)$
- $P(Y = y) = p(y) = \sum_{x \in \mathbb{R}} p_{X,Y}(x, y)$

In questo contesto, le funzioni di massa p_X, p_Y sono dette **marginali**. Come ben penserai, è possibile calcolare il valore atteso di una funzione di due variabili aleatorie nel seguente modo:

$$E[g(X,Y)] = \sum_{x,y \in R} g(x,y) p_{X,Y}(x,y)$$

Definizione 6. ***Variabili aleatorie indipendenti***
Due variabili aleatorie si dicono tali se valgono le seguenti relazioni:

1. $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$.
2. $P(X \in A | Y \in B) = P(X \in A)$, *equivalente alla prima.*
3. $P(X \in B | Y \in A) = P(X \in B)$, *equivalente alla prima.*
4. $P_{X,Y}(x,y) = p_X(x)p_Y(y)$.

3.6 Classi notevoli di variabili aleatorie

3.7 Statistiche campionarie

[1] Non è possibile il contrario, tuttavia. Soffri.

Chapter 4

Statistica Inferenziale

Pallw

Chapter 5

Regressione

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.

Basato su un codice open-source, la sua sintassi è estremamente flessibile e a tratti simile a C. È inoltre un linguaggio 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, utilizzare 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")
```

Prima di iniziare, tengo a ricordare che un codice scritto bene è nella maggior parte del tempo autoesplicativo e l'utilizzo dei commenti è richiesto solamente in casi dove sarebbe impossibile arrivare al senso da soli. Detto ciò:

```
# Questo è un commento
```

Ho utilizzato due scritture diverse. Nel corso della sezione sarà considerata scrittura al terminale qualunque cosa vada dopo il carattere "¿", altrimenti è uno script su file .R. Ora che abbiamo tutto pronto possiamo iniziare a lavorare col linguaggio.

6.2 Variabili, operatori e costrutti

Partiamo da dei semplici outputs. R ci consente di stampare a video direttamente stringhe, numeri ed operazioni senza che esse vengano contenute in una variabile:

```
> "Hello World!" # Stampa la stringa Hello World!
> 27 # Stampa il numero 27
> 5 + 5 # Stampa il numero 10
```

La dichiarazione di variabili è molto semplice; bisogna solamente tenere a mente tre restrizioni del linguaggio:

- R è case-sensitive. Quindi "Palle" $\neq$ "palle".
- Le variabili dichiarate non possono iniziare con un numero.
- Le variabili dichiarate non possono essere dichiarate con un nome uguale ad una parola chiave.

Chiarito ciò, è possibile dichiararle come segue:

```
> A <- 1+1  # Dichiaro A col valore 2
> A    # Stampo la variabile A a video
> A <- NULL  # Rendo la variabile vuota
> remove(A)  # Rimuove la variabile dall'ambiente
```

R supporta i seguenti tipi di dato, assegnati automaticamente all'inizializzazione della variabile:

- Numerico, come $[2, 10, 3.75]$. Considera quindi anche i razionali.
- Intero, come $[10L, 32L, 99L]$. La lettera L è necessaria per dichiarare il numero come intero.
- Complesso, come $[2i, 4i, 28i]$. La i indica l'unità immaginaria.

- Carattere, come [”a”, ”$Pallw$”, ”30”, ”$TRUE$”]. Notare che qualunque cosa all'interno di apici verrà visto come stringa.
- Booleano, come $[TRUE, FALSE]$. Autoesplicativo, dai.

L'allocazione della memoria è gestita dal linguaggio stesso e la variabile viene salvata nell'**ambiente** di lavoro. Questa rimane utilizzabile fin quando non la si rimuove.

È possibile controllare il tipo di una variabile con la funzione ”class(nomeVariabile)”. Vediamo ora un esempio di stampa di stringhe insieme a variabili utilizzeremo la funzione di concatenazione:

```
> c1 <- c2 <- c3 <- "A" # Associa il carattere "A" a multiple variabili.
> class(c1)
> cat("Questo gioco è un:", c1, c2, c3)
```

Passiamo alla matematica del linguaggio. Come già menzionato è possibile effettuare le quattro operazioni elementari, esponente, modulo e divisione intera con rispettivamente $[+, -, *, /, \%\%, \%/\%]$, ma esistono anche funzioni diverse per facilitarci la vita:

```
...
abs(-4.2)   # Torna il valore assoluta del dato inserito
sqrt(9)    # Torna la radice quadrata del dato inserito
min(10, 5, 15)  # Torna il valore minimo dei dati inseriti
max(10, 5, 15)  # Torna il massimo dei dati inseriti
ceiling(1.4)  # Torna il numero approssimato per eccesso (2)
floor(1.4)   # Torna il numero approssimato per difetto (1)
...
```

Abbiamo poi i comparatori la cui sintassi è uguale spiccicata a C:

```
x == y  # x uguale a y
x != y # x diverso da y
x < y # x minore di y
x <= y # x minore o uguale a y
x > y # x maggiore di y
x >= y # x maggiore o uguale a y
```

Come anche gli operatori logici:

```
&& # AND logico, torna vero se lo sono ambo gli elementi
|| # OR logico, torna vero se lo è almeno un elemento
! # NOT logico, inverte il valore dell'elemento
```

Passiamo a qualcosa di più serio, ma ugualmente semplice; R mantiene i costrutti condizionali e i cicli, quindi abbiamo le costruzioni "if else", "while" e "for".

```
a <- 200
b <- 33

# Funzionamento del costrutto if else
if (b > a) {
 print("b is greater than a")
} else if (a == b) {
 print("a and b are equal")
} else {
 print("a is greater than b")
}

c <- 40

# Funzionamento del costrutto while
while(b < c) {
 # Esegui il codice
 b <- b+1
}

dice <- c(1, 2, 3, 4, 5, 6)

# Funzionamento del costrutto for
for (x in dice) {
 print(x)
}
```

Per la strutturazione corretta di uno script più complesso sarà necessario dividere il tutto in funzioni:

```
my_function <- function(x) {
 return (5 * x)
}

print(my_function(3))
print(my_function(5))
print(my_function(9))
```

6.3 Strutture dati di R

R consente la scrittura di semplici programmi con i costrutti appena visti, ma detiene anche sei strutture dati di tipo più astratto:

- Vettori; Una lista di oggetti dello stesso tipo.
- Liste; Collezioni di dati ordinate e modificabili.
- Matrici; Dataset a due dimensioni composto da righe e colonne.
- Array; Collezione di elementi dello stesso tipo che può avere più di due dimensioni.
- Dataframes; Collezione di dati di vario tipo salvata in un format matriciale.
- Fattori; Utilizzati per la categorizzazione dei dati.

Andiamo con ordine; i **vettori** sono tali e quali a quelli visti nel linguaggio C; ciò significa che rispetteranno le stesse dinamiche, ma in R detengono anche modalità di accesso e modifica più ampie. Anzitutto, abbiamo più modi per dichiarare un vettore:

```
# Vettore di elementi 5, 10, 15, 20, 25
> v1 <- c(5, 10, 15, 20, 25)
# Vettore di elementi in sequenza da 1 a 5
> v2 <- 1:5
# Vettore di elementi da 5 a 25 a passi di 5
> v3 <- seq(from = 5, to = 25, by = 5)
# Vettore degli elementi di v3 ripetuti tre volte
> v4 <- rep(v3, times = 3)
# Accesso al valore del vettore v1 in posizione 1
> v1[1]
```

Come puoi vedere, è possibile utilizzare il vettore come una semplice variabile senza che il compilatore si lagni. Ciò significa che è possibile dare in pasto alle funzioni la variabile e se la gestirà da sola. In tal merito, abbiamo alcune funzioni utili per lavorare coi vettori:

- length(): Calcola la lunghezza del vettore.
- max(), min(): Calcolano valore massimo e minimo del vettore.
- sum(): Calcola la somma di tutti gli elementi del vettore.

- cumsum(): Calcola la somma cumulativa di ogni elemento.
- sort(): Ordina il vettore.

Le **liste** hanno un comportamento uguale ai vettori ed è possibile aggiungerne elementi come anche concatenare più liste:

```
# Dichiarazione delle liste l1, l2
l1 <- list("Dio", "Gesù", "Maria")
l2 <- list("Pietro", "Paolo", "Giovanni")

# Accesso all'elemento in posizione 3 di l1
l1[3]

# Aggiungo l'elemento "Cane" alla lista l1
append(l1, "Cane")

# Aggiungo l'elemento "Mascio" in posizione 2 alla lista
append(l1, "Mascio", after = 2)

# Concatenazione di l1 e l2 in l3
l3 <- c(l1, l2)
```

Passiamo ora alle **matrici**, le quali, come i vettori, detengono le funzioni e accessi presenti in C e molto altro. Abbiamo anche qui vari modi per dichiarare e accedere:

```
# Dichiarazione di una matrice, lega i due vettori come colonne.
mat1 <- cbind(c(1, 2, 3), c(1, 2, 3))

# Dichiarazione di una matrice, lega i due vettori come righe.
mat2 <- rbind(c(1, 2, 3), c(1, 2, 3))

# Accesso a elementi in colonne da 1 a 2 e righe da 2 a 3.
mat1[2:3,1:2]

# Rimuove la prima riga, nessun comando per le colonne.
mat1[-1,]

# Riempie una matrice di i righe e j colonne del valore n
mat3 <- matrix(n, i, j)

# Crea una matrice identità di dimensione k
matIdt <- diag(k)
```

Le operazioni utilizzabili dalle matrici seguono quelle viste in algebra lineare, vale a dire somma, sottrazione fra vettori o matrici e divisione, moltiplicazione per scalari, vettori o matrici. Il risultato deve essere salvato in una variabile diversa e l'operazione si scrive semplicemente come fossero due numeri. Seguono altre funzioni utili:

- det(): calcola il determinante di una matrice.
- qr(): Calcola la decomposizione QR.
- solve(): Fa l'inversa di una matrice.
- nrow(), ncol(): Ritornano il numero di righe o colonne di una matrice.
- rowSums(), colSums(): Fanno la somma dei valori di ogni riga o colonna.
- rowMeans(), colMeans(): Calcola la media dei valori di ogni riga o colonna.
- dim(): Calcola la dimensione di una matrice.

Gli **array** hanno la possibilità di andare da una a più dimensioni, senza un vero limite. Capirai ovviamente che per una e due dimensioni è preferibile usare le strutture apposite viste prima, quindi mostrerò solo come lavorare in dimensioni superiori.

```
# Dichiarazione di un array a più dimensioni
a2 <- array(c(1:12), dim = c(2, 3, 2))

# Stampo l'array in toto
a2

# Accedo alle posizioni i-1st dim, j-2nd dim, k-3rd dim.
a2[2, 3, 2]
```

I **dataframes** sono strutture bidimensionali che possono salvare dati in un formato matriciale. Si compongono quindi di righe e colonne, le quali possono essere viste come singoli vettori che contengono un tipo di dato; ciò comporta che la struttura può contenere diversi tipi allo stesso tempo:

```
# Dichiaro un dataframe
df1 <- data.frame(
 mount = c("Everest", "K2", "Fuji"),
 height = c(8848, 8611, 3776),
```

```
 todo = c(TRUE, TRUE, FALSE)
)

# Stampo il dataframe
df1

# Accedo alla singola colonna mount
df1$mount
```

Essendo inoltre strutturata come una matrice, sarà possibile utilizzare le sue stesse funzioni. Basta immaginarla come una matrice multitipo.

Utilizziamo poi i **fattori** per lavorare con dati categorizzati. Nel creare una struttura dati tale, potrà contenere solamente un insieme di valori fissati detti *livelli*, dati nella dichiarazione.

```
# Dichiaro il fattore f1 col set marital_status di relativi elementi
f1 <- marital_status <- factor(c("married", "single", "single",
"divorced", "married"))

# Stampo il fattore
f1

# Stampo i singoli livelli (ignora ripetizioni)
levels(f1)

# Provo ad inserire "hey" come livello, fallendo.
f1[1] <- "Hey" # In questo caso genererà NA
```

6.4 Funzioni di R per i grafici

R detiene varie funzioni per la rappresentazione dei dati su grafici; oltre a quelle built-in del linguaggio, ne abbiamo aggiunte altre col pacchetto ggplot che ti ho fatto installare ad inizio lettura. Possiamo creare grafici dei seguenti tipi:

- **Barplots**; barplot()
- **Lineplots**; plot(..., type = l)
- **Istogrammi**; hist()
- **Boxplots**; boxplot()

- **Scatterplots**; plot()
- **Heatmaps**; heatmap()
- **Density plots**; plot(density(...))

Tutte queste funzioni condividono le stesse particolarità, ovvero prendono tutte gli stessi valori e li elaborano allo stesso modo. L'unica differenza reale sta nella rappresentazione dei dati. Quindi una scrittura usata per plot() varrà anche per hist(), per esempio.

```
# Dichiaro i dati
x <- c(1, 2, 3, 4, 5)
y <- c(6, 7, 8, 9, 10)

# Creo il grafico, in questo caso uno scatterplot
plot(x,y)
```

Le caratteristiche del grafico sono modificabili a proprio piacimento; in particolare puoi modificare nomi degli assi, del grafico intero e i punti che rappresentano i singoli dati.

```
...
# Creo un grafico customizzato
```

6.5 Funzioni di R per la statistica

Ovviamente, se le variabili contengono valori è possibile effettuare operazioni aritmetiche. Inoltre, esistono due semplici funzioni per familiarizzare con la loro logica:

- Per il calcolo della media: mean(x, ...)
- Per il calcolo della mediana: median(x, na.rm = FALSE, ...)

Nella funzione della mediana notiamo la parola "na.rm"; si tratta di un valore logico con il quale decidere se contare o meno i valori "NA", ovvero **Not Available**.

Segue ora esempio di codice con quanto esposto finora:

```
# Assegno i valori 24, 25, 28 alle rispettive variabili A, B, C
> A <- 24
> B <- 25
> C <- 28
```

```
# Calcolo la media e la mediana e le salvo in due variabili
> mn <- mean(A, B, C)
> md <- median(c(24, 25, 28))

# Stampo a video i risultati
> cat("Media:", mn, "Mediana:", md)
```

La riga della mediana fa da apripista per il prossimo argomento.