- Probabilità e Statistica -

Federico Brutti

May 30, 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. $P_{X,Y}(x,y) = p_X(x)p_Y(y)$.

Siano ora due variabili aleatorie X, Y, definiamo il valore di **covarianza** come:

$$Cov(X,Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y)$$

Abbiamo inoltre le seguenti relazioni:

- Uguaglianza vera: $Cov(X,X) = Var(X)$.
- Se le due variabili aleatorie sono indipendenti e quindi **scorrelate**[2]: $Cov(X,Y) = 0$

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

[2] Notare che se due variabili sono scorrelate, non necessariamente ne implica anche l'indipendenza.

Inoltre, se il valore di covarianza è positivo, allora le variabili saranno positivamente correlate, mentre se è negativo varrà il contrario.

Esempio 13. ***Quando una variabile è scorrelata o indipendente?*** *Prendiamo un'urna con due palline numerate con* $1, 2$ *e definiamo due variabili aleatorie* X_1, X_2 *che avranno il loro valore pescato.*

- *Se le estrazioni avvengono con reimmissione allora* X_1, X_2 *saranno indipendenti e quindi scorrelate.*
- *Se le estrazioni avvengono senza reimmissione, le variabili saranno scorrelate, ma non indipendenti, perché la prima estrazione ha influenzato la seconda. Infatti per calcolare la covarianza useremo la funzione di massa congiunta.*

Questo concetto introduce quello di **coefficiente di correlazione**, un valore $Corr(X, Y) \in [-1, 1]$ che si calcola nel seguente modo:

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)} \times \sqrt{Var(Y)}}$$

Inoltre, avremo i due seguenti casi specifici:

- $Corr(X, Y) = 1 \iff \exists(a > 0 \wedge b) \in \mathbb{R} | Y = aX + b$
- $Corr(X, Y) = -1 \iff \exists(a < 0 \wedge b) \in \mathbb{R} | Y = aX + b$

Esiste anche la **funzione di ripartizione**, utile per calcolare la probabilità che una variabile aleatoria sia minore di un dato evento e per capire le modalità di crescita delle altre variabili aleatorie. Questo valore non è mai decrescente ed è definito (per una variabile aleatoria X) con:

$$F_X(x) = P(X \leq x)$$

Generalmente abbiamo le due scritture per i due tipi di variabile aleatoria:

- Per le discrete:

  $F_X(x) = \sum_{t \leq x} p_X(t) \implies p_X(x) = \Delta F_X(x) = F_X(x) - lim_{t \to x^+} F_X(t).$

- Per le continue:

  $$F_X(x) = \int_{-\infty}^{x} f_X(t)dt \implies F'_X(x) = f_X(x).$$

Esempio 14. *Abbiamo due variabili aleatorie X, Y definite come funzione lineare di X, quindi $aX + b$. Si calcoli la densità di Y avendo a disposizione quella di X. Eseguiamo i seguenti passaggi:*

1. $F_Y(y) = P(Y \leq y) = P(aX + b \leq y) = P(X \leq \frac{y-b}{a}) = F_X(\frac{y-b}{a})$.
2. *Deriviamo, ottenendo:* $F'_Y(y) = f_Y(y) = (\frac{y-b}{a}) \times \frac{1}{a} = \frac{1}{a} f \times (\frac{y-b}{a})$

In tal merito, abbiamo altre due proprietà della varianza:

- $Var(aX + b) = a^2 Var(X)$.
- $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$.
- Se le variabili aleatorie sono scorrelate: $Var(X + Y) = Var(X) + Var(Y)$.

3.6 Classi notevoli di variabili aleatorie

Dentro ai due tipi di variabili aleatorie ne possiamo riconoscere alcune con delle dinamiche notevoli e tocca impararle perché le chiede all'esame. Iniziamo:

- **Variabili aleatorie di Bernoulli**

  Una variabile aleatoria di questo tipo di parametro $p \in [0, 1]$ se assume solamente i valori $0, 1$. Inoltre:

  - Probabilità: $P(X = 1) = p = 1 - P(X = 0) \implies p_X(1) = p \wedge p_X(0) = 1 - p$.
  - Valore atteso: $E(X) = 0 * (1-p) + 1p = p \implies E(X^2) = E(X) = p$.
  - Varianza: $Var(X) = E(X^2) - E^2(X) = p - p^2 = p(1 - p)$.

- **Variabili aleatorie binomiali**

  Consideriamo un numero di prove ripetute indipendenti con probabilità di successo p e sia X la variabile aleatoria del numero di successi.

  Allora questa potrà assumere i valori da 0 a n. Definiamo la variabile X_i come:

$$X_i = \begin{cases} 1 \\ 0 \end{cases}$$

Dove risulta 1 se la prova ha esito positivo, altrimenti è uguale a 0. È fondamentalmente una sommatoria di variabili di Bernoulli. Inoltre:

- Funzione di massa: $p_X(k) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$
- Valore atteso: $E(X) = \sum_{i=1}^{n} E(X_i) = np$
- Varianza: $Var(X) = \sum_{i=1}^{n} Var(X_i) = np(1-p)$

- **Variabili aleatorie di Poisson**

  Una variabile aleatoria X di parametro $\lambda > 0$ è di questo tipo se può assumere i valori $(0, 1, ..., n, ...)$. Si utilizza per calcolare la probabilità di eventi rari in un totale di esperimenti ed è illimitata.

  Facciamo esempio che si effettui un numero spropositato di esperimenti aleatori, sui quali la probabilità che avvenga un evento X è infima, quasi impossibile. In tal caso, X sarà considerabile come una **variabile aleatoria di Poisson** di parametro np.

  Queste variabili aleatorie hanno anche la caratteristica di poter essere approssimate nel tipo binomiale Y visto prima di parametri n e $\frac{\lambda}{n}$, con $n >> 1$. Si può vedere con un valore approssimato del valore atteso e della varianza, come segue.

  - Funzione di massa: $p_X(n) = P(X = n) = e^{-\lambda} \frac{\lambda^n}{n!}$
  - Valore atteso: $E(X) \approx E(Y) = n \times \frac{\lambda}{n} = \lambda$
  - Varianza: $Var(X) \approx Var(Y) = n \times \frac{\lambda}{n}(1 - \frac{\lambda}{n}) \approx \lambda$

  È inoltre dimostrabile che $E(X) = \lambda = Var(X)$, la quale può dimostrarsi una relazione utile da ricordare.

- **Variabili aleatorie uniformi**

  Siano $\alpha < \beta$ due numeri reali. X è una variabile aleatoria uniforme in (α, β) se vale:

$$f_X(x) = \begin{cases} \alpha \leq x \leq \beta \implies \frac{1}{\beta - \alpha} \\ 0 \end{cases}$$

  - Valore atteso: $\int_{-\infty}^{+\infty} x f(x) dx = \int_{\alpha}^{\beta} \frac{x}{\beta-\alpha} dx = \frac{\alpha+\beta}{2}$
  - Valore atteso^2: $E(X^2) = \int_{-\infty}^{+\infty} x^2 f_X(x) dx = \int_{\alpha}^{\beta} \frac{x^2}{\beta-\alpha} dx = \frac{\alpha^2+\alpha\beta+\beta^2}{3}$
  - Varianza: $Var(X) = E(X^2) - E^2(X) = \frac{(\beta-\alpha)^2}{12}$

Una particolarità è come il valore atteso risulta essere il punto medio dell'intervallo considerato.

- **Variabili aleatorie normali**

  Fissiamo $\mu \in \mathbb{R}$ e $\sigma^2 > 0$. Diciamo che X è una tale variabile aleatoria di parametri $\mu e \sigma^2$ e scriviamo:

$$X \sim N(\mu, \sigma^2) \text{ se } f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

  Questa è la densità che caratterizza la curva gaussiana, di media μ e varianza σ^2; infatti, la funzione $f(x) = e^{-x^2}$ forma una curva gaussiana con un dato centro, spostabile modificando il valore μ.

INSERIRE IMMAGINE

  Notare inoltre che se una variabile aleatoria $Z \sim N(0, 1)$, diremo che la prima è una **normale standard** e la sua funzione risulta essere: $f_Z(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$; inoltre:

  - Valore atteso: $E(Z) = 0$
  - Varianza: $Var(Z) = 1$

  Ritornando invece al caso generale, ovvero $\mu \in \mathbb{R} e \sigma^2 > 0$ abbiamo che:

  - Variabile aleatoria: $X = \sigma Z + \mu \sim N(\mu, \sigma^2)$
  - Valore atteso: $E(X) = \sigma E(Z) + \mu = \mu$
  - Varianza: $Var(X) = \sigma^2 Var(Z) = \sigma^2$

  Ogni variabile aleatoria di questo tipo è facilmente trasformabile in una normale standard mediante la **standardizzazione**:

$$\frac{X - \mu}{\sigma} \sim N(0, 1)$$

  Sapendo questo, dichiariamo le seguenti due variabili aleatorie normali indipendenti:

$$X \sim N(\mu_1, \sigma_1^2), Y \sim N(\mu_2, \sigma_2^2) \implies X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

Il calcolo della probabilità di queste variabili si effettua con la standardizzazione. Calcoliamo $P(a \leq X \leq b)$:

$$\begin{aligned} P(a \leq X \leq b) &= P(\frac{a-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}) \\ &= P(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}) \\ &= \phi(\frac{b-\mu}{\sigma}) - \phi(\frac{a-\mu}{\sigma}) \end{aligned} \tag{3.1}$$

Dove $\phi(x) = F_Z(x)$ è la funzione di ripartizione $Z \sim N(0,1)$. Un'ultima cosa utile da ricordare è che, siccome stiamo esaminando una curva gaussiana, per la valutazione dell'integrale è possibile fare $1 - \int$ oppure, siccome è sicuramente una funzione pari, si può specchiare il punto richiesto e calcolare in sua funzione... oppure puoi scegliere il metodo sano ed utilizzare la **tabella di probabilità**.

Proposizione 1. ***Istruzioni per l'utilizzo della tabella***

Vogliamo calcolare la probabilità di $Z = 1.52$. Nota che sulla colonna all'estrema sinistra è indicato un intervallo di valori; prendi quello che più si avvicina a quello richiesto. Bisogna poi trovare un numero nella riga in cima alla tabella sicché la loro somma dia il valore di Z. La casella che interseca questa riga e colonna sarà il numero che ci serve.

3.7 Statistiche campionarie

Siano $X_1, ..., X_n$ variabili aleatorie indipendenti ed identicamente distribuite. Queste sono un modello per un dataset o un campione n di dati. Ogni combinazione di tali variabili si dice **statistica campionaria**. In particolare, abbiamo le due seguenti:

- **Media campionaria**:

$$\overline{X_n} = \frac{1}{m}(X_1 + ... + X_m)$$

- **Varianza campionaria**:

$$S_m^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \overline{X_n})^2$$

Tenderemo a considerare solo la media campionaria, in quanto la statistica più semplice con la quale lavorare. Più nello specifico, ricaviamo i suoi valori di:

- **Valore atteso**:

$$E(\overline{X}_n = \frac{1}{m}(E(X_1, ..., E(X_m))) = \mu$$

- **Varianza**:

$$Var(\overline{X}_n) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Inoltre, per avere ulteriori informazioni sulla distribuzione di $\overline{X_n}$, consideriamo il **teorema del limite centrale**:

Teorema 1. ***Teorema del limite centrale***
Qualunque sia la distribuzione delle X_i, la distribuxione di Z_n per n grande è ben approssimata da una $N(0,1)$.

$$Z_n = (\overline{X_n} - \mu)\frac{\sqrt{m}}{\sigma} \sim N(0,1) \equiv \overline{X_n} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Più precisamente, $\forall z \in \mathbb{R} \wedge n \to +\infty$:

$$F_{Z_n}(z) = P(Z_n \leq z) \to \phi(z)$$

Prestiamo particolare attenzione all'applicazione del teorema al caso X_i di Bernoulli di parametro p, ovvero

$$X_1 + ... + X_n = X$$

con X binomiale di parametri n, p. Il teorema del limite centrale afferma che:

$$X \sim N(np, np(1-p))$$

Questa approssimazione si dice **normale per una binomiale** ed è considerata buona se rispetta la seguente restrizione:

$$np \geq 5, n(1-p) \geq 5$$

Usando questa stima, essendo che si sta calcolando un valore di probabilità approssimando una distribuzione discreta con una continua, si deve applicare la **correzione di continuità**; una modifica dell'intervallo di integrazione ai suoi estremi di $\frac{1}{2}$.

Per esempio, data una variabile aleatoria X con distribuzione binomiale di parametri n, p, per n sufficientemente grande, si può assumere che:

$$P(X \leq x) \approx P\left(Y \leq x + \frac{1}{2}\right)$$

Dove Y è una variabile aleatoria che segue una distribuzione normale con parametri $\mu = np$ e $\sigma^2 = np(1-p)$. Ne risulta che la correzione dell'approssimazione è molto maggiore, rispetto a non utilizzarla.

Chapter 4

Statistica Inferenziale

Lo scopo delle statistica inferenziale è, data una popolazione, di capire la sua vera distribuzione a partire dai dati che si possono osservare e studiare in un campione da essa estratto.

Sarà quindi necessario formulare un **modello statistico** di cui è nota la distribuzione ed i dati osservati, ma con i valori dei parametri incogniti. Ciò ridurrà il problema intero a questi due passaggi:

1. Scegliere un valore plausibile per i valori dei parametri reali, quindi fare delle inferenze.

2. Testare ipotesi sui valori per verificarne l'attendibilità.

Le uniche inferenze statistiche di nostro interesse saranno quelle sulle **popolazioni normali** e di **Bernoulli**.

4.1 Stima dei parametri

Supponiamo che i dati siano realizzazioni di una variabile aleatoria X con una densità discreta o continua $f(x, \theta)$, dove θ è un parametro incognito. Dobbiamo stimarlo coi dati ottenuti dalle osservazioni su X; per farlo si usano due tipi di stima:

- **Stima puntuale**; Dove si ottiene un singolo valore come stima per il valore di θ.

- **Stima intervallare**; Si ottiene un intervallo di valori possibili per θ, associando ad ogni intervallo un livello di fiducia che θ vi appartenga.

Partiamo dal primo tipo e dalle sue basi; per prima cosa sarà richiesto introdurre qualche concetto fondamentale. Anzitutto, chiamiamo **campione** di

ampiezza n una collezione di variabili aleatorie X_n indipendenti e tutte con la stessa distribuzione $f(x;\theta)$, con θ parametro incognito.

Una **Statistica** T è invece una variabile aleatoria ottenuta come funzione del campione, ovvero $T = T(X_1, ..., X_n)$, ed uno **stimatore** è qualunque T indipendente da θ, è usata per stimarlo. Tendenzialmente, il ruolo è coperto dalla media campionaria per la varianza campionaria.

Infine, chiamiamo **stima** il valore numerico assunto dallo stimatore sui dati osservati $x_1, ..., x_n$, ovvero, $\theta = T(x_1, ..., x_n)$.

Sia ora $X_1, ..., X_n$ un campione casuale preso da una popolazione di densità $f(x, \theta)$, che dipende dal parametro θ. Se interpretiamo

$$f(x_1, ..., x_n; \theta)$$

come la verosimiglianza che si realizzi la n-upla $x_1, ..., x_n$ di dati quando θ è il vero valore del parametro, possiamo prendere come sua stima il *valore che rende massima la funzione*, chiamato **stimatore di massima verosimiglianza**. Consigliato inoltre vederlo come logaritmo, in quanto ha uno stesso massimo e facilita i calcoli.

Esempio 15. *INSERISCI ESEMPIO PER STIMATORE MAX VEROSIMIGLIANZA PER BERNOULLI E POPOLAZIONE NORMALE.*

Ma in che modo è possibile scegliere uno stimatore $T = T(X_1, ..., X_n)$? O meglio, come ne si valuta la **bontà**?

Bisogna cercare di minimizzare la deviazione dal valore reale del parametro attraverso valore atteso e varianza. Non potendo tuttavia essere onniscenti, è inevitabile incappare in errori e per questo si introduce il concetto di **distorsione** o bias.

Definizione 7. ***Bias***
Sia $T = T(X_1, ..., X_n)$ uno stimatore di θ. Allora $b(T) = E(T) - \theta$ è detto bias di T come stimatore di θ. Se è nullo, T è detto stimatore corretto di θ.

Uno stimatore buono e utile controlla sia varianza che bias in modo contenuto, con lo scopo di fornire un risultato quanto più vicino alla realtà possibile senza essere troppo permissivo.

Esempio 16. *INSERISCI ESEMPIO PER STIMATORE BUONO*

Inoltre, sia lo stesso stimatore $T = T(X_1, ..., X_n)$ del parametro θ. Chiamiamo **errore quadratico medio** il valore atteso del quadrato della differenza fra lo stimatore ed il θ. Si indica con:

$$MSE(T) = E[(T - \theta^2)^2] = Var(T) + b(T)^2$$

Se T è corretto, allora questo errore quadratico medio sarà uguale alla varianza dello stimatore.

Passiamo ora alla **stima intervallare**. Sia un campione estratto da una popolazione. Ci si aspetta che la stima ottenuta valutando lo stimatore sui dati osservati non sia l'effettivo valore di θ, quindi è preferibile produrre un intervallo per il quale abbiamo una certa fiducia che il parametro vi appartenga. In tal merito, diamo le seguenti definizioni:

Definizione 8. ***Stimatore intervallare***
Sia $X_1, ..., X_n$ un campione casuale di una popolazione dove ci interessa stimare un parametro θ. Siano poi $L_1 = L_1(X_1, ..., X_n)$, $L_2 = L_2(X_1, ..., X_n)$ due statistiche non dipendenti da θ, tali che:

$$P(L_1 < \theta < L_2) = 1 - \alpha, \text{ con } \alpha \in (0, 1)$$

L'intervallo (L_1, L_2) si dice ***stimatore intervallare*** *del parametro θ e per costruirlo è necessario conoscere la distribuzione delle sue statistiche L_1, L_2.*

Definizione 9. ***Intervallo di confidenza***
Siano adesso $\hat{l}_1 = L_1(x_1, ..., x_n)$ e $\hat{l}_2 = L_2(x_1, ..., x_n)$ i valori assunti dalle statistiche L_1, L_2 sui dati osservati $x_1, ..., x_n$.

*Diremo quindi che $(\hat{l}_1, \hat{l}_2)$ è l'****intervallo di confidenza*** *di livello $1 - \alpha$ per il parametro θ.*

Notare che (L_1, L_2) è un intervallo aleatorio che contiene il valore di θ, mentre $(\hat{l}_1, \hat{l}_2)$ è una realizzazione del primo; data la sua natura non si presta ad alcuna valutazione probabilistica. Quindi, in sintesi:

1. Otteniamo un *campione casuale* $X_1, ..., X_n$ da una popolazione, il quale ci può far ottenere lo *stimatore intervallare* (L_1, L_2), composto da due variabili aleatorie, fra le quali è probabile accada il parametro θ di nostro interesse, quindi $P(L_1 < \theta < L_2) = 1 - \alpha$, che risulta dare il *coefficiente di fiducia*.

2. Dal campione casuale effettuiamo delle inferenze, ottenendo il *campione osservato* $x_1, ..., x_n$ che può avere solo valori numerici. Da questo possiamo ottenere l'*intervallo di confidenza* $(\hat{l}_1, \hat{l}_2)$.

Ultima cosa prima di passare ai vari casi di studio; per costruire lo stimatore intervallare è necessario conoscere la **distribuzione** delle statistiche L_1, L_2; quindi vediamo in che modo possono essere distribuite.

Sia un campione estratto da una popolazione normale $X_1, ..., X_n$ e diciamo che ha una media $\mu \in \mathbb{R}$ e varianza $\sigma^2 > 0$. Siamo interessati a studiarne la

distribuzione delle statistiche campionarie, ovvero la **media campionaria** $\overline{X}$ e la **varianza campionaria** S^2, per ottenere rispettivamente μ ed σ^2. Abbiamo che:

- **Densità χ^2 a $n-1$ gradi di libertà**

  Le variabili aleatorie $\overline{X}$, S^2 sono indipendenti e per il teorema di limite centrale abbiamo che la media campionaria si distribuisce con una normale di media μ incognita e varianza $\frac{\sigma^2}{n}$, quindi: $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$. Inoltre:

  $$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

  La cui densità è asimmetrica e non nulla solo sui numeri reali positivi.

- **Densità t di student a $n-1$ gradi di libertà**

  Si tratta di una densità con la forma a campana simmetrica rispetto ad $x = 0$ e si stima con:

  $$\frac{\overline{X} - \mu}{\sqrt{\frac{S^2}{n}}} \sim t_{n-1}$$

4.2 Intervalli di confidenza

Passiamo adesso ai vari casi di studio che possiamo trovare nello svolgimento degli esercizi. Premetto che si somigliano tutti abbastanza e sarà utile capire come agire per poter capire anche la sezione seguente e prendere una decisione ponderata riguardo alle richieste.

- **Intervalli di confidenza per la media di una popolazione normale, varianza nota**

  Sia il campione $X_1, ..., X_n$, la media incognita $\mu \in \mathbb{R}$ e la varianza nota $\sigma^2 > 0$. Con $\alpha \in (0, 1)$ dobbiamo ricavare intervalli di confidenza ad un livello $1 - \alpha$ per la media μ.

  Diciamo che $\frac{\overline{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} = Z \sim N(0, 1)$ e indichiamo con z_α il valore per cui $P(Z < z_\alpha) = \alpha$. Usando un pò di algebra noterai facilmente che:

  $$P(Z < z_\alpha) = \alpha \implies 1 - P(Z > z_\alpha) = \alpha \implies 1 - \alpha = P(Z < z_\alpha)$$

  E che quindi ci servirà capire quell'area della funzione dove $Z < z_\alpha$. È probabile che venga richiesto l'intervallo totale (bilaterale) oppure

una sola parte (unilaterale), quindi bisogna prendere la metà richiesta o tutto l'intervallo. Più in particolare, le due metà si ottengono con:

$$1-\alpha = P(-z_{\alpha/2} < Z < z_{\alpha/2}) \implies 1-\alpha = \left(\overline{X} - z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}\right)$$

Dove il primo elemento della disequazione è L_1, il secondo è il parametro della media da stimare e l'ultimo è L_2. Sapendo ora di aver osservato dei dati $x_1, ..., x_n$ tali che $\overline{X}(x_1, ..., x_n) = \overline{x}$, abbiamo che ad un livello di confidenza $1 - \alpha$, per la media μ, gli intervalli ottenibili sono:

- **Bilaterale**: $(\overline{x} - z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}})$
- **Unilaterali**: $(\overline{x} - z_{\alpha} \times \frac{\sigma}{\sqrt{n}}, +\infty)$, $(-\infty, \overline{x} + z_{\alpha} \times \frac{\sigma}{\sqrt{n}})$

Per gli esercizi effettuare il seguente procedimento:

1. Calcola la media dei dati raccolti $\overline{x}$ e il livello di confidenza α.
2. Calcola il valore di z_α oppure $z_{\alpha/2}$.
3. Prendi il risultato di $1 - \alpha$ oppure $1 - \frac{\alpha}{2}$ e trova il corrispondente valore nella tavola degli $\phi(x)$
4. Hai tutto. Scrivi l'intervallo.

- **Intervalli di confidenza per la media di una popolazione normale, varianza incognita**

Sia il campione $X_1, ..., X_n$, la media $\mu \in \mathbb{R}$ e la varianza nota $\sigma^2 > 0$, ambo ignote. Con $\alpha \in (0, 1)$ dobbiamo ricavare intervalli di confidenza ad un livello $1 - \alpha$ per la media μ.

Teniamo anzitutto a mente due cose:

- $\frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$, con S deviazione standard campionaria.
- La densità t di student ha una forma a campana simmetrica rispetto a $x = 0$.

Se $X \sim t_n$, allora $t_{\alpha,n} \in \mathbb{R}$ è il valore per cui $P(X > t_{\alpha,n}) = \alpha$. Supponiamo nuovamente di avere dei dati $x_1, ..., x_n$ tali che $\overline{X}(x_1, ..., x_n) = \overline{x}$ ed $S(x_1, ..., x_n) = \hat{s}$.

Allora a livello di confidenza $1 - \alpha$ avremo i seguenti intervalli:

- **Bilaterale**: $(\overline{x} - t_{\alpha/2,n-1} \times \frac{\hat{s}}{\sqrt{n}}, \overline{x} + t_{\alpha/2,n-1} \times \frac{\hat{s}}{\sqrt{n}})$

- **Unilaterali**: $(\overline{x} - t_{\alpha,n-1} \times \frac{\hat{s}}{\sqrt{n}}, +\infty)$, $(-\infty, \overline{x} + t_{\alpha/2,n-1} \times \frac{\hat{s}}{\sqrt{n}})$

Per gli esercizi effettuare il seguente procedimento:

1. Calcola media $\overline{x}$ e deviazione standard $\hat{s}$.
2. Calcola α e $t_{\alpha,n-1}$, con n numero totale di elementi nel campione.
3. Ottieni il valore di $t_{\alpha,n-1}$ dalla tavola dei valori di t_n.
4. Scrivi l'intervallo.

- **Intervalli di confidenza per la varianza di una popolazione normale**

Siano $X_1, ..., X_n$, μ, $\sigma^2 > 0$, con media e varianza ignote. Possiamo creare degli intervalli di confidenza basandoci sul seguente fatto:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

Sia adesso $\alpha \in (0,1)$. Con $X \sim \chi^2_n$, allora $\chi^2_{\alpha,n} \in [0, +\infty)$ sarà il valore per cui $P(X > \chi^2_{\alpha,n}) = \alpha$.

Con una semplice sostituzione abbiamo che l'intervallo di confidenza $1 - \alpha$ è dato da:

$$P\left(\chi^2_{1-\alpha/2,n-1} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2,n-1}\right) = P\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right)$$

Dunque avendo i dati $x_1, ..., x_n$ tali che $S^2(x_1, ..., x_n) = \hat{s}^2$, abbiamo che ad un livello di confidenza $1 - \alpha$, per la varianza σ^2, gli intervalli ottenibili sono:

- **Bilaterale**: $\left(\frac{(n-1)\hat{s}^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)\hat{s}^2}{\chi^2_{1-\alpha/2,n-1}}\right)$
- **Unilaterali**: $\left(\frac{(n-1)\hat{s}^2}{\chi^2_{\alpha,n-1}}, +\infty\right)$, $\left(-\infty, \frac{(n-1)\hat{s}^2}{\chi^2_{1-\alpha,n-1}}\right)$

Per gli esercizi effetturare il seguente procedimento:

1. Calcolare $\hat{s}^2$ se non dato e poi α oppure $\frac{\alpha}{2}$.
2. Calcolare $n - 1$, poi ricercare il valore di $\chi^2_{\alpha,n-1}$ nella tabella.
3. Scrivere l'intervallo ottenuto.

- **Intervalli di confidenza per la media di una popolazione di Bernoulli**

  Qui abbiamo una popolazione di oggetti, ognuno dei quali, indipendentemente dagli altri, ha certi **requisiti** con una probabilità $q \in (0,1)$.

  Se testiamo n di questi oggetti, rilevando quanti di loro hanno tali requisiti, possiamo usare tale grandezza per ottenere un intervallo di confidenza per q.

  Consideriamo X come una variabile aleatoria binomiale di parametri n, q. Noi lavoreremo con casi semplici, altrimenti se $nq \geq 5$ e $nq(1-q) \geq 5$ si dice che il campione preso in considerazione è **numeroso**.

  Poniamo lo stimatore di massima verosimiglianza di q: $Q = \frac{X}{n}$ e, sapendo di avere dati tali per cui $X = x \wedge Q(x) = \hat{q}$, usando il teorema del limite centrale si può ottenere un intervallo di confidenza bilaterale approssimato di livello $1-\alpha$ per q come segue:

  $$\left(\hat{q} - z_{\alpha/2}\sqrt{\frac{\hat{q}(1-\hat{q})}{n}}, \hat{q} + z_{\alpha/2}\sqrt{\frac{\hat{q}(1-\hat{q})}{n}}\right)$$

  Questa scrittura è resa valida dal fatto che lo stimatore si approssima ad una distribuzione normale $N(q, \frac{q(1-q)}{n})$ e $z_{\alpha/2}$ è il quantile di ordine $\frac{\alpha}{2}$ di $Z \sim N(0,1)$.

  Per gli esercizi effettuare il seguente procedimento:

  1. Calcolare la percentuale di elementi coi requisiti adatti sul campione $\hat{q}$.
  2. Calcolare $z_{\alpha/2}$ ed n.
  3. Sostituisci i valori ottenuti alla formula dell'intervallo.

4.3 Verifica di ipotesi

In questa sezione useremo quanto appena visto per decidere se i risultati ottenuti sono sufficientemente affidabili da rappresentare verità. Noi faremo quindi delle **ipotesi statistiche**, ovvero delle affermazioni su un parametro θ da cui dipende un campione $X_1, ..., X_n$. Matematicamente si tratta di un'asserzione di tipo:

$$\theta = \theta_0,\ \theta \leq \theta_0,\ \theta \leq \theta_0$$

Con θ_0 un certo valore del parametro. E possibile fare due tipi di ipotesi:

- **Semplice**: Specifica un solo valore di θ.
- **Composta**: Specifica un insieme di valori di θ.

Il processo di verifica di un'ipotesi statistica è detto **test statistico**, dal quale avremo due alternative da cui scegliere, in base al nostro grado di confidenza; la **convinzione di partenza** H_0 e l'**affermazione contrapposta** H_1.

È chiaro che ha senso fare testing solamente se si nutre qualche dubbio sulla veridicità di H_0 e vi è la possibilità che venga contraddetta. Tuttavia il test statistico non è la risposta assoluta all'affidabilità, bensì consente solo di vedere se i dati ottenuti sono compatibili o meno con H_0; la decisione finale sta al singolo, ed è ponderata da un certo grado di tolleranza di errore.

Creiamo anzitutto uno stimatore di θ, indicato con $ST = ST(X_1, ..., X_n)$. Per avere un test quantitativo è necessario creare la **regione critica** C, una soglia da non varcare per confermare la veridicità della convinzione iniziale. Segue le due regole:

- Confermiamo H_0 se $st = ST(x_1, ..., x_n) \notin C$.
- Rifiutiamo H_0 se $st = ST(x_1, ..., x_n) \in C$.

Ma in che modo si costruisce la regione critica? Non è un valore fisso; bisogna prendere una decisione in base alla presenza di valori, troppo alti o troppo bassi, che fanno dubitare della veridicità di H_0. In tal merito, è possibile commettere due tipi di errore: rifiutare H_0 quando è vera si dice errore di **prima specie**, mentre accettarla quando è falsa si dice di **seconda specie**.

Per evitare di sbagliare si prendono in considerazione i due seguenti valori:

- **Livello di significatività** α: Un livello tale per cui la probabilità che H_0 sia nella regione critica è ad esso minore o uguale. In parole povere, dice la probabilità che si stia sbagliando a stimare.

$$P_{H_0}(ST \in C) \leq \alpha$$

- **Valore-p dei dati**: Si tratta di un valore che indica il livello di significatività **critico**, ovvero il numero estremo superiore o estremo inferiore che non deve essere superato per far sì che valga H_0.

$$ValP = sup\{\alpha : ST \notin C\} = inf\{\alpha : ST \in C\}$$

Fare attenzione ad usare queste restrizioni con cautela; è necessario prima osservare i dati del campione, per poi scegliere i relativi livelli di significatività.

4.4 Testing su una popolazione

- **Test per la media di una popolazione normale con varianza nota**

  Sia $X_1, ..., X_n$ campione estratto da una popolazione normale, con media $\mu \in \mathbb{R}$ ignota e varianza $\sigma^2 > 0$ nota. La dinamica decisionale è la seguente:

H_0	H_1	Statistica test ST	Rifiuto H_0 a livello $1-\alpha$ se
$\mu = \mu_0$	$\mu \neq \mu_0$	$\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$	$\lvert st\rvert > z_{\alpha/2}$
$\mu \leq \mu_0$	$\mu > \mu_0$		$st > z_\alpha$
$\mu \geq \mu_0$	$\mu < \mu_0$		$st < -z_\alpha$

- **Test per la media di una popolazione normale con varianza ignota**

  Sia $X_1, ..., X_n$ campione estratto da una popolazione normale, con media $\mu \in \mathbb{R}$ e varianza $\sigma^2 > 0$, ambo ignote. La dinamica decisionale è la seguente:

H_0	H_1	Statistica test ST	Rifiuto H_0 a livello $1-\alpha$ se
$\mu = \mu_0$	$\mu \neq \mu_0$	$\frac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$	$\lvert st\rvert > t_{\alpha/2,n-1}$
$\mu \leq \mu_0$	$\mu > \mu_0$		$st > t_{\alpha,n-1}$
$\mu \geq \mu_0$	$\mu < \mu_0$		$st < -t_{\alpha,n-1}$

- **Test per la varianza di una popolazione normale**

  Sia $X_1, ..., X_n$ campione estratto da una popolazione normale, con media $\mu \in \mathbb{R}$ e varianza $\sigma^2 > 0$, ambo ignote. La dinamica decisionale è la seguente:

H_0	H_1	Statistica test ST	Rifiuto H_0 a livello $1-\alpha$ se
$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$\frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}$	$st < \chi^2_{1-\alpha/2,n-1} \vee st > \chi^2_{\alpha/2,n-1}$
$\sigma^2 \leq \sigma_0^2$	$\sigma^2 > \sigma_0^2$		$st > \chi^2_{\alpha,n-1}$
$\sigma^2 \geq \sigma_0^2$	$\sigma^2 < \sigma_0^2$		$st < \chi^2_{1-\alpha,n-1}$

- **Test asintotici per la media di una popolazione di Bernoulli**

  Qui useremo il teorema del limite centrale per costruire dei test asintotici per il parametro $q \in (0,1)$ di una popolazione di Bernoulli.

  Sia $X_1, ..., X_n$ un campione casuale estratto dalla suddetta popolazione con parametro incognito q. La dinamica decisionale è la seguente:

H_0	H_1	Statistica test ST	Rifiuto H_0 a livello $1-\alpha$ se
$q = q_0$	$q \neq q_0$	$\frac{Q - q_0}{\sqrt{\frac{q_0(1-q_0)}{n}}} \sim N(0,1)$	$\lvert st \rvert > z_{\alpha/2}$
$q \leq q_0$	$q > q_0$		$st > z_\alpha$
$q \geq q_0$	$q < q_0$		$st < -z_\alpha$

4.5 Testing su due popolazioni

Chapter 5

Regressione

5.1 Regressione lineare semplice

5.2 Stima dei coefficienti di regressione

5.3 Inferenza statistica sul coefficiente angolare

5.4 Coefficiente di determinazione e analisi dei residui

Chapter 6

Il Linguaggio R

R è linguaggio di programmazione case-sensitive. 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. I files con i quali lavoreremo hanno l'estensione ".R" e 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**.

Prima di iniziare è bene installare ogni pacchetto necessario per lo studio di probabilità e statistica; il comando da terminale è:

```
# Questo è un commento
> install.package("nomePacchetto")
```

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.

6.1 Componenti base del linguaggio

6.1.1 Variabili, operatori e strutture dati

Il linguaggio R consente di stampare a video stringhe, numeri ed operazioni sia che essi siano contenuti in una variabile dichiarata o meno:

```
A <- "Hello World!"
A # Stampa il contenuto della variabile A
B <- 27
B # Stessa roba, ma per B
C <- 5 + 5
C # Avrai capito ora, no?
A <- NULL # Rendo A vuota
remove(A) # Rimuovo A dall'ambiente
```

Bisogna solo tenere a mente che non è possibile dichiarare variabili il cui nome inizia con un numero o hanno lo stesso nome di una parola chiave. 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. Rimane utilizzabile fin quando non la si rimuove. Una cosa di cui tener conto è che i caratteri si possono concatenare per creare delle stringhe, in questo modo:

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

Per quanto riguarda gli **operatori**, abbiamo la possibilità di effettuare le operazioni elementari, potenze, 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
```

Non fermiamoci qui, il vero potenziale di R sta anche nelle sue sei strutture dati innate, le quali sono:

- **Vettori**; Una lista di oggetti dello stesso tipo. Presentano le stesse dinamiche viste in C con migliori modalità di accesso.

  ```
  # 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]
  ```

  È evidente che è possibile utilizzare il vettore come una semplice variabile, rendendo possibile darlo in pasto alle funzioni senza problemi di memoria. Abbiamo inoltre alcune funzioni innate utili per lavorare con questa struttura di dati, dove "v" indica il vettore:

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

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

- **Liste**; Collezioni di dati ordinate e modificabili. Hanno un comportamento uguale ai vettori ed è possibile aggiungervi 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)
```

- **Matrici**; Dataset a due dimensioni composto da righe e colonne.

```
# 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, con "A" matrice:

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

- **Array**; Collezione di elementi dello stesso tipo che può avere più di due dimensioni. Usarli solo se è necessario avere più di due dimensioni.

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

# Stampo l'intero array
a2

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

- **Dataframes**; Collezione di dati di vario tipo salvata in un formato matriciale. Ne consegue che siano composti da 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.

- **Fattori**; Utilizzati per lavorare con dati categorizzati. Questa struttura può contenere solo un insieme di valori fissati detti **livelli**, i quali vengono 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
```

6.1.2 Costrutti condizionali, cicli e funzioni

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) {
 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**:

```
# Dichiarazione della funzione my_function
my_function <- function(x) {
 return (5 * x)
}

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

6.1.3 Grafici e salvataggio immagini

R ha varie funzioni di base per la rappresentazione dei dati su grafici; possiamo creare i seguenti tipi:

- **Grafici a barre**

```
ages = 20:29
students = c(2,1,5,3,4,2,0,2,1,0)

barplot(students,
 xlab = "Age of students",  # Nome asse x
 ylab = "Number of students", # Nome asse y
 names.arg = ages    # Dati posti sull'asse x
)
```

- **Grafici a linea**

```
x <- c(1, 2, 3, 4, 5)
y <- c(2, 4, 6, 8, 10)

# Creo un grafico di tipo lineplot, con dati x,y
plot(type = "l", x, y,
xlab = "X values",  # Nome dei valori sulle x
ylab = "Y values",  # Nome dei valori sulle y
main = "X and Y values",# Nome del grafico
col = "red"    # Colore dei dati
)
```

- **Istogrammi**; hist()
- **Boxplots**; boxplot()
- **Grafici di dispersione**; plot()
- **Mappe di calore**; heatmap()
- **Grafici di densità**; plot(density(...))

È possibile aggiungere più grafici in un singolo foglio, a discapito dello spazio. Per farlo è necessario usare il seguente comando:

```
x = 1:6
y = c(33, 11, 5, 9, 22, 30)

# Gestione del layout. Ho scritto una matrice 2x2.
par(mfrow = c(2,2))
```

```
# Le posizioni si ordinano da sole.
barplot(y, main = "Barplot", names.arg = x)
barplot(y, main = "Horizontal Barplot",
names.arg = x, horiz = TRUE)
plot(x, y, type = "l", main = "Line plot")
plot(x, y, main = "Scatter plot", col = "red")
```

Per quanto riguarda l'export, basta cliccare il pulsante corretto nell'interfaccia di R-Studio. Nessun bisogno di stressarsi con vari comandi. Inoltre per fare grafici migliori è preferibile usare il pacchetto **ggplot2**, del quale parleremo nelle prossime sezioni.

6.2 Gestione scripts e pacchetti aggiuntivi

6.2.1 Salvataggio, caricamento e fonti

La voglia di ridichiarare le variabili ogni volta che si vuole lavorare con R è pari a zero; infatti è possibile salvarle in due formati:

- **.RData**; Formato di file binario usato per salvare uno o più oggetti R in un singolo file.

  ```
  # Per salvare l'environment
  save(nomeOggetto)
  # Per caricarlo in un file diverso
  load(nomeOggetto)
  ```

- **.RDS**; Formato di file binario usato per salvare un singolo oggetto R in un file.

  ```
  # Per salvare l'oggetto in RDS
  saveRDS(nomeOggetto)
  # Per caricarlo in un file diverso
  readRDS(nomeOggetto)
  ```

È possibile dividere un progetto in più files, per poi chiamare lo script nel momento necessario tramite la funzione **source()**. Un grande insieme di files .R si dice **pacchetto**.

6.2.2 Dplyr

Il pacchetto **Dplyr** è creato per la manipolazione dei dati, consentendo di trasformare e riassumere tabelle di dati usando delle funzioni apposite. Nel nostro caso, lo utilizzeremo per l'organizzazione degli elementi nei dataframes. Le principali funzioni di dplyr sono:

- **Operatore pipe** %>%; Invia il risultato della funzione precedente alla successiva.
- **filter()**; Filtra i dati in base ad una condizione specificata.
- **select()**; Seleziona la colonna indicata in base al nome.
- **mutate()**; Modifica colonne esistenti o ne crea nuove.
- **arrange()**; Ordina le righe.
- **group_by()**; Raggruppa i dati per effettuare operazioni in ogni gruppo.
- **summarize()**; Fa un sommario delle statistiche date.

6.2.3 Statistica descrittiva e Ggplot2

Abbiamo discusso dello scopo della statistica descrittiva nella parte di teoria; con R possiamo usare le seguenti funzioni e studiare i dati in nostro possesso:

- **Media**: mean().
- **Mediana**: median().
- **Differenza fra valore minimo e massimo**: range().
- **Scarto interquartile**: IQR().
- **Varianza**: var().
- **Deviazione standard**: sd().
- **Skewness**: skewness().
- **Curtosi**: kurtosis().
- **Covarianza**: cov().
- **Coefficiente di correlazione**: cor().

Altre funzioni di R base:

- **summary()**: Fa un sommario di un dataset, comprende minimo, massimo, quartili, media e mediana di ogni variabile.
- **table()**: Crea una tavola di contingenza per le variabili categoriche, mostrando la frequenza di ogni combinazione di categoria.
- **prop.table()**: Crea una tavola di contingenza per le variabili categoriche, mostrando la proporzione di ogni combinazione di categoria.
- **quantile()**: Calcola il valore di un quantile specifico.

Per quanto riguarda le operazioni matematiche, bisogna prendere la funzione che svolge quel compito ed eventualmente stampare i risultati a video in forma leggibile utilizzando le verbs di dplyr come **summarize()**. Per la rappresentazione grafica seria, invece, useremo il pacchetto **ggplot2**, il quale consente, tramite una scrittura chiara e stratificata, di creare grafici altamente personalizzabili, di ogni tipo.

La struttura a strati necessita che le parti siano scritte nel seguente ordine:

1. **ggplot(data = ...)**: Inizializza il grafico.
2. **aes()**: Dà nome agli assi e setta i colori.
3. **geom_*()**: Determina il tipo di grafico.
4. **scale_color_manual(values = ...)**: Setta manualmente i colori dei dati nel grafico.
5. **labs()**: Dà i nomi alle varie parti del grafico.

Vediamo un esempio:

```
# Usa le librerie utili
library("palmerpenguins")
library("ggplot2")

# Dichiarazione della variabile per il grafico
mass_flipper <- ggplot(data = penguins) +
 aes(x = flipper_length_mm, y = body_mass_g) +
 geom_point(aes(color = species, shape = species), size = 3, alpha = 0.8) +
 scale_color_manual(values = c("darkorange", "purple", "cyan4")) +
 labs(title = "Flipper length and body mass",
 subtitle = "Colored by Adelie, Chinstrap and Gentoo Penguins",
```

```
  x = "Flipper length (mm)",
  y = "Body mass (g)",
  color = "Penguin species",
  shape = "Penguin species") +
  theme(legend.position = "bottom")

mass_flipper
```

Con questa formula è possibile costruire ogni tipo di grafico per ogni evenienza. Questo comprende anche tabelle di frequenza, boxplots, mappe di calore, istogrammi etc. Familiarizzarci è più che consigliato.

6.3 Elementi di probabilità

6.3.1 Costrutti per il calcolo della probabilità

6.3.2 Coefficiente binomiale

6.3.3 Formula di Bayes

6.4 Variabili aleatorie

6.4.1 Binomiali

6.4.2 Di Poisson

6.4.3 Uniformi

6.4.4 Normali

6.4.5 Esponenziali

6.5 Statistica inferenziale

e stima di parametri

6.5.1 Approssimazione delle distribuzioni binomiali e normali

6.5.2 Teorema del limite centrale

6.5.3 Stimatori di massima verosimiglianza

6.5.4 Intervalli di confidenza e testing delle ipotesi

6.6 Regressione lineare semplice