Metodi Matematici e Statistici
Canale M-Z A.A 2024-2025
Appunti Semiseri

Disclaimer

Questo documento contiene gli appunti del corso di Metodi Matematici e Statistici del Corso di Laurea triennale in Informatica, canale M-Z, A.A. 2024-2025. Questi appunti devono considerarsi "semiseri", informali, per cui per ogni approfondimento o trattazione formale si rimanda ai libri di testo che trovate nella sezione "Testi di riferimento" del Syllabus.

Questo documento ovviamente sarà pieno di errori, per cui non esitate a mandarmi la lista di tutte le cose inesatte che troverete sparse in giro.

Questo documento è ancora "in fieri", per cui verrà periodicamente aggiornato in base al tempo che avrò a disposizione.

Avere a disposizione questo documento non vi autorizza a non venire a lezione, anche perché sarà ovviamente meno esaustivo di tutto il blabla che potrò raccontarvi in classe.

1 Eventi e Probabilità

In questa sezione introdurremo il concetto di *evento* e di *probabilità* .

1.1 Eventi

Un *evento* è una proposizione ben definita che può essere vera o falsa. Esempio:

- E = "Lancio una moneta ed esce testa"
- E = "Lancio un dado ed esce un numero pari"
- E = "Supero l'esame"

Possiamo identificare dei tipi specifici di eventi

- Evento **certo**: si indica con Ω, la proposizione può essere solo vera;
- Evento **impossibile**: si indica con $\emptyset$, la proposizione può essere solo falsa;
- Due eventi A e B si dicono **incompatibili** se uno è necessariamente vero quando l'altro è falso
- Due eventi A e B si dicono **uguali** se quando A è vero anche B è necessariamente vero e viceversa. L'uguaglianza tra eventi si indica con $A = B$.

Possiamo anche introdurre alcune operazioni tra eventi. Assumiamo che A e B siano due eventi, allora

- A **implica** B ($A \subseteq B$): A vero implica B vero. Se $A \subseteq B$ e $B \subseteq A$ allora $A = B$;
- A^c è l'evento **contrario** di A se è falso quando A è vero e viceversa;
- **Unione** o **somma logica** di A e B ($A \vee B$, $A \cup B$) è vero se almeno uno tra A e B è vero;
- **Intersezione** o **prodotto logico** di A e B ($A \wedge B$, $A \cap B$) è vero se entambi gli eventi A e B sono veri. Se due eventi A e B sono incompatibili avremo che $A \cap B = \emptyset$.

Esercizio 1

Un congegno idraulico è costituito da tre valvole come in figura 1. Il congegno funziona se almeno uno dei due rami del circuito funzioni. Il ramo superiore funziona se le valvole A e B funzionano contemporaneamente, il ramo inferiore funziona se la valvola C funziona. Descrivere l'evento E = "Il congegno funziona".

Introduciamo i tre eventi

1. F_A = "La valvola A funziona"
2. F_B = "La valvola B funziona"
3. F_C = "La valvola C funziona"

Introduciamo anche gli eventi

1. R_{sup} = "Il ramo superiore funziona" = $F_A \cap F_B$
2. R_{inf} = "Il ramo inferiore funziona" = F_C

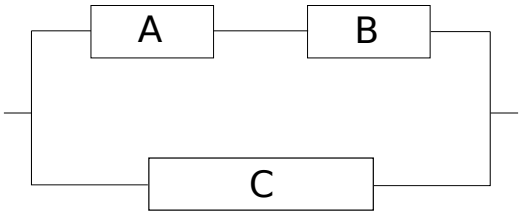

Figura 1: Esercizio 1

A questo punto l'evento E = "Il congegno funziona" si scriverà $E = R_{sup} \cup R_{inf} = (F_A \cap F_B) \cup F_C$.

1.2 Probabilità

In questa sottosezione proveremo a introdurre il concetto di probabilità. Daremo quattro definizione di probabilità.

1. Definizione Classica

Immaginiamo di avere n eventi elementari equiprobabili, conto quante volte h si verifica l'evento E. Allora la probabilità di E sarà

$$P(E) = \frac{h}{n} = \frac{\#\text{casi favorevoli}}{\#\text{casi totali}}. \tag{1}$$

2. Definizion Frequentista

Ho una sequenza di n istanze di un qualche fenomeno ripetibile. Conto quante volte k si verifica l'evento E che mi interessa. La probabilità dell'evento E sarà

$$P(E) = \frac{k}{n}. \tag{2}$$

OBS: più è grande n più è accurata come stima! Ma di questo parleremo poi.

3. Definizione soggettiva (B. De Finetti)

È una misura di fiducia del verificarsi di un evento. Prendiamo come esempio l'evento E = "passo l'esame". Non è possibile dare attribuire una probabilità né di tipo classica né frequentista.

Secondo questa definizione dirò che la probabilità che passi l'esame è del 70% se sono disposta a pagare 70 euro per vincerne 100 qualora dovesse passare l'esame.

Si può formalizzare questa cosa introducendo una quota $p \geq 0$ e una somma $S > 0$. Si dirà che si effettua una scommessa di quota p su un evento E se versando pS si riceve un importo S solo se E si verifica (scommessa vinta) e niente in caso contrario.

Introduciamo anche il guadagno G definito come segue

$$\begin{cases} G_1 = S - pS = S(1-p) & \text{se si verifica E} \\ G_2 = -Sp & \text{se non si verifica E.} \end{cases} \tag{3}$$

La scommessa si dice coerente se $G_1 G_2 \leq 0$, ovvero i due guadagni sono di segno discorde. E questo accade se $0 \leq p \leq 1$. In questo caso p sarà la probabilità di E.

OBS: Quando giocate la schedina alla SNAI fate esattamente questo. La partita del Catania è quotata $q > 1$. Voi andate a giocare una quota $\tilde{S}$, per cui se vincete ricevete $G_1 = q\tilde{S} - \tilde{S} = \tilde{S}(q-1)$. Se perdete, ovviamente, perderete $G_2 = -\tilde{S}$. Se confrontate quanto scritto prima si ha che $p = 1/q$ e $\tilde{S} = Sp$.

4. Definizione Assiomatica (Kolmogorov)

Questa è una definizione più formale di probabilità. Immaginiamo un qualche evento aleatorio e introduciamo tre elementi

- Ω l'insieme dei possibili esiti e $\mathcal{P}(\Omega)$ l'insieme delle sue parti;
- $\mathcal{A}$ la famiglia di sottoinsiemi di Ω che costituiscono gli eventi, $\mathcal{A} \subseteq \mathcal{P}(\Omega)$;
- P una misura della fiduciosità dell'evento della famiglia $\mathcal{A}$.

La terna $(\Omega, \mathcal{A}, P)$ si dice *spazio di probabilità*. Ci sono alcune proprietà degli eventi che è importante elencare:

1. $\emptyset, \Omega \in \mathcal{A}$, ovvero l'evento certo e l'evento impossibile sono eventi;
2. Se $A \in \mathcal{A}$ allora $A^c \in \mathcal{A}$, ovvero se esiste un evento esiste anche l'evento contrario;
3. Se ho tanti eventi $A_1, A_2, \ldots, A_n$, allora $\bigcup_{i=1}^n A_i \in \mathcal{A}$, ovvero l'unione di tanti eventi è ancora un evento.

Queste proprietà fanno sì che l'insieme $\mathcal{A}$ sia una σ-algebra.

A questo punto la probabilità sarà una funzione che va dalla famiglia degli eventi nell'intervallo $[0, 1]$. Ovvero

$$P : \mathcal{A} \to [0, 1].$$

Questa funzione, per essere coerente, deve godere (assiomaticamente) delle seguenti due proprietà:

1. $P(\Omega) = 1$;
2. $P(\bigcup_{i=1}^n A_i) = \sum_{i=1}^n P(A_i)$ se $A_i \cap A_j = 0$.

Da questo si possono dimostrare le seguente proprietà

- $P(A^c) = 1 - P(A)$;
- $P(A) \leq P(B) \forall A, B \in \mathcal{A} : A \subseteq B$;
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(\bigcup_{i=1}^n) = \sum_{i=1}^n P(A_i) - \sum_{(i,j)\in I_2} P(A_i \cap A_j) + \sum_{(i,j,k)\in I_3} P(A_i \cap A_j \cap A_k) + \cdots + (-1)^{n-1} P(A_1 \cap A_2 \cap \cdots \cap A_n)$.

1.3 Probabilità condizionale

Prima di parlare di probabilità condizionale dobbiamo definire un evento condizionato. Un evento condizionato $A|B$ (A noto B o A dato B), con $B \neq \emptyset$ è un ente a tre valori

1. **Vero** se essendo vero B lo è anche A;
2. **Falso** se essendo vero B A è falso;

3. **Indeterminato** se B è falso.

Ad esempio: A = "mangiare la pizza", B = "cenare fuori", $A|H$ = "mangiare la pizza sapendo di andare a cena fuori"

La probabilità condizionale di A noto H è

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

Due eventi A e B si dicono **indipendenti** se $P(A \cap B) = P(A)P(B)$ e, di conseguenza, $P(A|B) = P(A)$. OBS: quella sopra è condizione necessaria e sufficiente, ovvero se so che i due eventi sono indipendenti, allora la probabilità dell'intersezione sarà il prodotto delle probabilità. Se invece so che la probabilità dell'intersezione sarà il prodotto delle probabilità, allora posso dedurre che gli eventi sono indipendenti.

Siamo pronti per formulare il

Teorema delle probabilità totali

Siano $A_1, A_2, \ldots, A_n \in \mathcal{A} : \bigcup_i A_i = \Omega, A_i \cap A_j = \emptyset \ \forall i \neq j, P(A_i) > 0 \ \forall i$, allora

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i).$$

Dimostrazione

$$P(B) = P(B \cap \Omega) = P(B \cap \bigcup_{i=1}^{n} A_i) = P(\bigcup_{i=1}^{n}(B \cap A_i)) = \sum_{i=1}^{n} P((B \cap A_i)) = \sum_{i=1}^{n} P(B|A_i)P(A_i).$$

Esercizio 2

Una stazione radio riceve segnale da due sorgenti A e B. La probabilità che il segnale dalla stazione A sia distorto è 0.1. La probabilita che il segnale dalla stazione B sia distorto è 0.2. Qual è la probaiblità di ricevere un segnale distorto?

Introduciamo i due eventi S_A = "Segnale dalla stazione A" e S_B = "Segnale dalla stazione B". Immaginiamo che il segnale sia equidistribuito tra A e B, per cui $P(S_A)P(S_B) = 0.5 > 0$. Inoltre $S_A \cup S_B = \Omega$. Inoltre $S_A \cap S_B = \emptyset$.

Possiamo utilizzare il teorema delle probabilità totali per valutare la probabilità dell'evento S = "Segnale distorto". Dal testo sappiamo che $P(S|S_A) = 0.1$ e $P(S|S_B) = 0.2$. Per cui avremo

$$P(S) = P(S|S_A)P(S_A) + P(S|S_B)P(S_B) = 0.1 \cdot 0.5 + 0.2 \cdot 0.5 = 0.15.$$

Teorema di Bayes

Dati due eventi $A, B \in \mathcal{A}$ tali che $P(A), P(B) > 0$, si ha che

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

Dimostrazione

Dato che $P(A \cap B) = P(B \cap A)$, sappiamo che $P(B|A) = \frac{P(A \cap B)}{P(A)}$, da cui $P(A \cap B) = P(A|B)P(B)$. Possiamo sostituire questa espressione nella formula per la probabilità condizionale e trovare il nostro risultato.

1.4 Richiami di calcolo combinatorio

Richiamiamo rapidamente alcuni concetti di calcolo combinatorio. Dobbiamo per prima cosa introdurre due concetti fondamentali: disposizioni e combinazioni. Entrambi sono modi di prendere un tot di oggetti da un insieme di elementi, la differenza è che le disposizoni considerano gli elementi ordinati, mentre le combinazioni gli elementi senza ordinamento. Per capirci meglio: le disposizioni sono i possibili modi in cui si possono inserire i sei numeri del codice di sblocco del telefono (per cui il codice 123456 sarà diverso da 654321); le combinazioni invece sono i possibili modi in cui possono uscire i 15 numeri della cartella per fare tombola (non è importante se il 53 esce prima de 46, l'importante è che se nella mia cartella ho sia il 53 che il 46 questi escano entrambi, possibilmente prima che qualcuno faccia tombola). Più nel dettaglio avremo:

- r-disposizioni: r elementi ordinati scelti da un insieme di n elementi. Ne esistono n^r.
- r-disposizioni semplici: r elementi ordinati scelti da un insieme di n elementi senza ripetizioni. Ne esistono $D_{n,r} = \frac{n!}{(n-r)!}$. se $r = n$ abbiamo le permutazioni che sono $D_{n,n} = n!$.
- r-combinazioni: r elementi non ordinati scelti da un insieme di n elementi. Ne esistono $\binom{n+r-1}{r} = \frac{(n+r-1)!}{r!(n-1)!}$.
- r-combinazioni semplici: r elementi non ordinati scelti da un insieme di n elementi senza ripetizioni. Ne esistono $C_{n,r} = \binom{n}{r} = \frac{(n)!}{r!(n-r)!}$.

È importante ricordare che dato un insieme di ne elementi e dati degli interi $n_1, n_2, \ldots, n_k$ tali che $\sum_{i=1}^{k} n_i = n$, il numero di k partizioni con $n_1, n_2, \ldots n_k$ elementi è

$$C^n_{n_1,n_2,\ldots,n_k} = \frac{n!}{n_1!n_2!\ldots n_k!}$$

1.5 Esercizi

Esercizio 1

Un'urna contiene 5 palline bianche, 6 nere, 4 rosse. Se ne estraggono due. Calcolare che siano dello stesso colore nel caso in cui a) l'estrazione sia istantanea; b) l'estrazione sia con reinserimento.

Svolgimento

Iniziamo contando il numero di palline, quindi poniamo $n_{\text{palline}} = 15$.

Per calcolare questa probabilità dobbiamo semplicemente fare casi favorevoli su casi totali. Quello che è diverso è il modo di contare i casi. Nel primo caso dobbiamo utilizzare le combinazioni semplici, infatti ci interessa in quanti modi posso estrarre due palline (diverse, quindi senza ripetizione delle palline) da un insieme più grande. Nel secondo caso, invece, ci interessa tenere traccia del fatto che abbiamo un ordine nell'estrazione per cui userò le distribuzioni. Dato che ho reinserimento posso tranquillamente ripescare la stessa pallina, per cui uso le distribuzioni.

a)

Evento favorevole = "sono entrambe rosse" $\bigcup$"sono entrambe bianche"$\bigcup$ "sono entrambe nere"
Evento totale = "estraggo due palline di 15"

Il numero di possibili estrazioni è

$$\mathrm{n}_{\mathrm{totali}} = \binom{15}{2}.$$

Usando le regole del calcolo combinatorio abbiamo che il numero di combinazioni favorevoli affinché entrambe le palline siano rosse è $\mathrm{n}_{\mathrm{rosse}} = \binom{4}{2}$. Ragionando analogamente abbiamo che il numero di combinazioni restituenti due palline bianche è $\mathrm{n}_{\mathrm{bianche}} = \binom{5}{2}$, mentre quelle con due palline nere è $\mathrm{n}_{\mathrm{rosse}} = \binom{6}{2}$.

Per cui avremo che il numero di casi favorevoli sarà

$$\mathrm{n}_{\mathrm{favorevoli}} = \binom{5}{2} + \binom{6}{2} + \binom{4}{2}$$

Per cui la probaiblità sarà

$$p = \frac{\mathrm{n}_{\mathrm{favorevoli}}}{\mathrm{n}_{\mathrm{totali}}} = \frac{\binom{5}{2} + \binom{6}{2} + \binom{4}{2}}{\binom{15}{2}} = 0.2952$$

b)

Possiamo ripetere tutto per il caso con reinserimento. Quello che cambierà saranno le combinazioni danti vita agli eventi favorevoli Usando le regole del calcolo combinatorio abbiamo che il numero di combinazioni favorevoli affinché entrambe le palline siano rosse è $\mathrm{n}_{\mathrm{rosse}} = 4^2$. Ragionando analogamente abbiamo che il numero di combinazioni restituenti due palline bianche è $\mathrm{n}_{\mathrm{bianche}} = 5^2$, mentre quelle con due palline nere è 6^2, per cui avremo

$$\mathrm{n}_{\mathrm{favorevoli}} = 5^2 + 4^2 + 6^2.$$

Mentre il numero di casi totali sarà

$$\mathrm{n}_{\mathrm{totali}} = 15^2.$$

Pertanto la probabilità sarà

$$p = \frac{5^2 + 4^2 + 6^2}{15^2} = 0.3422$$

Potete risolvere questo stesso esercizio ragionando: la probabilità di pescare una pallina rossa alla prima e alla seconda estrazione sarà $p_R = 4/15$, per cui la probabilità di estrarne due a due estrazioni consecutive sarà $p_{RR} = (4/15)^2$ e quindi $p = p_{RR} + p_{BB} + p_{NN}$.

Esercizio 2

Dati i due eventi A = “Lo studente ha studiato bene” e B = “Lo studente ha passato l’esame” tradurre in simboli i seguenti eventi composti:

1. La probabilità che uno studente abbia studiato bene e passi l'esame è 0.4
2. La probabilità che uno studente che ha studiato bene passi l'esame è 0.8
3. La probabilità che uno studente che non ha studiato bene non passi l'esame è 0.9
4. La probabilità che uno studente non ha studiato bene e passi l'esame è 0.05
5. La probabilità che uno studente che non ha passato l'esame non avesse studiato è di 0.82

Noti $P(\mathrm{A}) = 0.5$ e $P(\mathrm{B}) = 0.45$ calcolare a) $P(\overline{\mathrm{B}}|\mathrm{A})$ e b) $P(\overline{\mathrm{A}} \cap \overline{\mathrm{B}})$

Svolgimento

1. $P(\mathrm{A} \cap \mathrm{B}) = 0.4$
2. $P(\mathrm{B}|\mathrm{A}) = 0.8$
3. $P(\overline{\mathrm{B}}|\overline{\mathrm{A}}) = 0.9$
4. $P(\mathrm{B} \cap \overline{\mathrm{A}}) = 0.05$
5. $P(\overline{\mathrm{A}}|\overline{\mathrm{B}}) = 0.82$

Calcoliamo adesso le probabilità richieste al punto

a) $P(\overline{\mathrm{B}}|\mathrm{A}) = 1 - P(\mathrm{B}|\mathrm{A}) = 0.2$

b) $P(\overline{\mathrm{A}} \cap \overline{\mathrm{B}}) = P(\overline{\mathrm{A}}|\overline{\mathrm{B}})P(\overline{\mathrm{B}}) = P(\overline{\mathrm{B}}|\overline{\mathrm{A}})P(\overline{\mathrm{A}}) = 0.45$

Esercizio 3

Una "roulette" semplificata consiste in 12 numeri classigicati rossi o neri in base al seguente schema

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

Siano dati i seguenti eventi:

A = "esce un numero pari" B = "esce un numero rosso" C = "esce un numero ≤ 6 D = "esce un numero ≤ 8

Stabilire se

1. A, B, C sono a due a due indipendenti
2. A, B, C è una famiglia di eventi indipendenti
3. A, B, D è una famiglia di eventi indipendenti

Svolgimento

Ricordiamo che se degli eventi sono indipendenti se il prodotto dell'intersezione è scrivibile come prodotto delle probabilità. Le probabilità le calcoliamo valutando dalla tabella qual è il rapporto casi favorevoli su casi totali. I casi totali sono 12. Ad esempio, i casi favorevoli per l'evento $\mathrm{B} \cap \mathrm{A}$, ovvero "esce un numero rosso pari" sarà 3/12, in quanto gli unici numeri rossi pari sono 2, 8, 12.

Per cui

1. $P(\mathrm{A} \cap \mathrm{B}) = \frac{1}{4}$, $P(\mathrm{A})P(\mathrm{B}) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} = P(\mathrm{A} \cap \mathrm{B})$. Quindi A e B sono indipendenti!

   Ripetiamo il ragionamento per gli altri eventi e troviamo che

   $P(\mathrm{A} \cap \mathrm{C}) = P(\mathrm{A})P(\mathrm{C}) = \frac{1}{4}$ e $P(\mathrm{B} \cap \mathrm{C}) = P(\mathrm{B})P(\mathrm{C}) = \frac{1}{3}$

   Deduciamo quindi che gli eventi sono a due a due indipendenti.

2. Sfruttando il risultato precedente possiamo affermare che A, B, C sono eventi a due a due indipendenti. Per cui per verificare se sono una famiglia di eventi indipendenti è necessario calcolare se $P(\mathrm{A} \cap \mathrm{B} \cap \mathrm{C}) = P(\mathrm{A})P(\mathrm{B})P(\mathrm{C})$. Sempre ussando la tabella troviamo che $P(\mathrm{A} \cap \mathrm{B} \cap \mathrm{C}) = \frac{1}{12}$ e $P(\mathrm{A})P(\mathrm{B})P(\mathrm{C}) = \frac{1}{8}$, per cui gli eventi non costituiscono una famiglia di eventi indipendenti!

3. Provate voi a verificare se A, B, D costituiscono una famiglia di eventi indipendenti!

Esercizio 4

Tutte le borse che si imbarcano su un aereo vengono passate al metal detector allo scopo di individuare eventuali ordigni. È noto che:

- La probabilità che una borsa con un ordigno faccia scattare l'alalrme è 0.99;
- La probabiità che una borsa senza ordigno faccia scattare l'allarme è 0.05;
- Una borsa ogni 5000 contiene un ordigno.

Calcolare:

1. Con che probabilità scatterà l'allarme;
2. Qual è la probabilità che una borsa che ha fatto scattare l'allarme contenga un ordigno;
3. Di quanto aumenta la probabilità che una borsa contenga l'ordigno sapendo che ha fatto scattare l'allarme?

Svolgimento

Per risolvere il problema dobbiamo prima identificare gli eventi che ci interessano. Chiamiamo

A = "La borsa contiene l'ordigno"

B = "La borsa ha fatto scattare l'allarme"

Per ipotesi noi conosciamo che:

- $P(\mathrm{A}) = 1/5000$
- $P(\mathrm{B}|\mathrm{A}) = 0.99$
- $P(\mathrm{B}|\overline{\mathrm{A}}) = 0.05$

Siamo pronti per rispondere.

1. Per rispondere a questa domanda utilizziamo il teorema delle probabilità totali. Per cui

$$P(\mathrm{B}) = P(\mathrm{B}|\mathrm{A})P(\mathrm{A}) + P(\mathrm{B}|\overline{\mathrm{A}})P(\overline{\mathrm{A}})$$

   Ricordando che $P(\overline{\mathrm{A}}) = 1 - P(\mathrm{A})$ e sostituendo i numeri troviamo $P(\mathrm{B}) = 0.0502$.

2. Per rispondere a questa domanda dobbiamo calcolare la probabilità condizionata $P(\mathrm{A}|\mathrm{B})$. Per farlo utilizziamo il teorema di Bayes e troviamo che

$$P(\mathrm{A}|\mathrm{B}) = \frac{P(\mathrm{B}|\mathrm{A})P(\mathrm{A})}{P(\mathrm{B})} = 0.0039$$

3. Noi conosciamo la probaiblità che una borsa contenga l'ordigno, ovvero $P(\mathrm{A})$. Conosciamo anche la probabilità che una borsa che ha fatto scattare l'allarme contenga l'ordigno, ovvero $P(\mathrm{A}|\mathrm{B})$. Ovviamente mi aspetto che la seconda sia più grande della prima. Mi calcolo il fattore di proporzionalità tra le due quantità

$$\gamma = \frac{P(\mathrm{A}|\mathrm{B})}{P(\mathrm{A})} = 19.5.$$

Questo significa che l'informazione che l'allarme è scattato ha fatto aumentare di quasi venti volte la probabilità che la bomba sia scattata!

2 Variabili aleatorie

Il concetto di variabile aleatoria generalizza quello di evento.

Matematicamente, dato uno spazio di probabilità $(\Omega, \mathcal{A}, \mathcal{P})$, definiamo **variabile aleatoria** un'applicazione $X : \Omega \to \mathbb{R}$ t.c. $\forall t \in \mathbb{R}$ l'insieme $\{\omega : X(\omega) \leq t\} \in \mathcal{A}$, ovvero è un evento possibile.

Da ora in poi useremo una notazione abbreviata per cui $\{X \leq t\} \equiv \{\omega : X(\omega) \leq t\}$.

In parole povere, una variabile aleatoria è una funzione che prende un evento, ovvero una proposizione, e vi associa un numero reale.

Esempio. Se considero il lancio di un dado ho sei possibili eventi a cui posso associare un numero reale:

$$\text{A} = \text{“Esce il numero 1”} \to 1$$

$$\text{B} = \text{“Esce il numero 2”} \to 2$$

$$\text{C} = \text{“Esce il numero 3”} \to 3$$

$$\text{D} = \text{“Esce il numero 4”} \to 4$$

$$\text{E} = \text{“Esce il numero 5”} \to 5$$

$$\text{F} = \text{“Esce il numero 6”} \to 6$$

L'utilità delle variabili aleatorie è legata al fatto che con i numeri reali sappiamo operare, per cui è più facile ottenere determinati risultati.

In realtà usiamo il concetto di variabile aleatoria ogni giorno senza rendercene conto. Se vi chiedessi di calcolare il valore medio ottenuto dal lancio di un dado voi mi rispondereste 3.5, ottenuto banalmente usando la formula per la media $(1+2+3+4+5+6)/6$. Nel darmi questa risposta voi avete inconsapevolmente associato all'evento “esce il numero i” il numero i.

Chiamiamo $A \subset \mathbb{R}$ un generico sottoinsieme dei numeri reali. Definiamo quini **Legge o distribuzione di X**

$$P(X \in A) \quad \forall A \subset \mathbb{R}.$$

In sostanza, la legge di una variabile aleatoria (da ora in poi v.a.) mi dirà la probabilità che ogni evento si verifichi.

Ci sono due tipi di v.a.

- **Discreta**: L'insieme $X(\Omega)$ è finito o numerabile (ovvero il numero di esiti è discreto o numerabile) – Esempio: lancio di un dado;
- **Continua**: L'insieme $X(\Omega)$ è infinito (ovvero il numero di esiti è infinito) – Esempio: tempo decadimento sostanza radioattiva;

2.1 Variabili aleatorie discrete

Concentriamoci adesso sulle variabili aleatorie discrete. Sia X una v.a. discreta. Consideriamo $A \in \mathbb{R}$, sarà a sua volta un insieme discreto che si può scrivere come unione dei suoi elementi

a, ovvero $A = \bigcup_{a \in A} \{a\}$. Per cui avrò che

$$P(X \in A) = P(X \in \bigcup_{a \in A} \{a\}) = \sum_{a \in A} P(X = a).$$

Definisco **densità discreta** di X la funzione

$$p(x) = P(X = x).$$

La densità discreta gode di due proprietà:

1. $p(x) \geq 0 \quad \forall x \in \mathbb{R}$;
2. $\sum_{x \in \mathbb{R}} p(x) = \sum_{i=1}^{\infty} p(x_i) = 1$.

Nell'ultima proprietà ho sfruttato il fatto che l'insieme è discreto per cui è equivalente sommare sui valori in $\mathbb{R}$ o sui numeri naturali. Da ora in poi le considererò interscambiabili.

Definisco **funzione di ripartizione** (da ora in poi f.r.) $\forall t \in \mathbb{R}$

$$F_X(t) = P(X \leq t) = \sum_{x \leq t} p(x).$$

Osservo che $F_X : \mathbb{R} \to [0, 1]$, ovvero associa una probabilità a ogni istanza della nostra variabile aleatoria.

Definisco **speranza matematica** la quantità

$$E[X] = \sum_{i=1}^{\infty} x_i p(x_i).$$

La speranza matematica gode di alcune proprietà:

1. $E[cX] = cE[X] \ \forall c \in \mathbb{R}$;
2. $E[X + Y] = E[X] + E[Y]$;
3. se $p(x \leq y) = 1 \Rightarrow E[X] \leq E[Y]$;
4. $|E[X]| \leq E[|X|]$;
5. Se X e Y sono v.a. indipendenti (le definiremo dopo) $E[YX] = E[X]E[Y]$

In realtà la speranza matematica si può definire se e solo se *esiste finita*, ovvero se e solo se vale la proprietà che $E[X] = \sum_{i=1}^{\infty} |x_i| p(x_i) < \infty$, dove $|\cdot|$ indica il valore assoluto.

Se considero una generica funzione $f : \mathbb{R} \to \mathbb{R}$ t.c. $f(X)$ è una v.a. a sua volta, se esiste finita (non ripeterò l'espressione, avete capito il concetto), posso calcolarne la speranza matematica e ottenere $E[f(X)] = \sum_{i=1}^{\infty} f(x_i) p(x_i)$.

Utilizzo quanto sopra per definire il **momento di ordine k** come il valore di aspettazione di X^k, ovvero

$$E[X^k] = \sum_{i=1}^{\infty} x_i^k p(x_i).$$

Possiamo anche definire il **momento centrato di ordine k** come

$$E[(X - E[X])^k] = \sum_{i=1}^{\infty}(x_i - E[X])^k p(x_i).$$

Il più importante tra i momenti centrati è il momento di ordine 2 noto anche come **varianza**, $\sigma_X^2 = Var(X) = E[(X - E[X])^2]$. La varianza misura quanto in media i dati sono lontani dalla speranza matematica, ovvero ci dà un'indicazsione di quanto siano "sparpagliati". Si indica con il simbolo σ^2 e gode delle seguenti proprietà

1. $\text{Var}(X) = E[X^2] - E[X]^2$;
2. $\text{Var}(aX) = a^2\ \text{Var}(X)$;
3. $\text{Var}(a + X) = \text{Var}(X)$.

Definiamo **deviazione standardi** la radice quadrata della varianza $\sigma_X = \sqrt{\sigma_X^2}$.

Date due v.a. X, Y definiamo **covarianza tra X, Y**

$$\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y].$$

La covarianza tiene conto di come cambia X relativamente a Y. In particolare se $\text{Cov}(X, Y) = 0$ le due v.a. sono indipendenti.

Nota la covarianza, possiamo scrivere

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$$

.

Esempio: X v.a. che segue una distribuzione discreta uniforme, ovvero assume uno degli n elementi in $A = \{1, 2, \ldots, n\}$ con ugual probabilità $p(x) = \frac{1}{n}$.

La funzione di ripartizione sarà una funzione discontinua (a scalini) definita così

$$F(t) = \begin{cases} 0 & \text{se } t < 1 \\ \frac{1}{n} & \text{se } 1 \leq t < 2 \\ \frac{2}{n} & \text{se } 2 \leq t < 3 \\ \ldots & \\ 1 & \text{se} t \geq n \end{cases}$$

La speranza matemtica sarà

$$E[X] = \sum_{i \in A} i \cdot \frac{1}{n} = \frac{1}{n}\sum_{i=1}^{n} i = \frac{n(n+1)}{2n} = \frac{n+1}{2}$$

Analogamente possiamo calcolare la varianza

$$\sigma_X^2 = \frac{1}{n}\sum_{i=1}^{n} i^2 - E[X]^2 = \frac{n(n+1)(2n+1)}{6n} - \left(\frac{n+1}{2}\right)^2 = \frac{n^2 - 1}{12}$$

Nelle prossime sezioni vedremo più in dettaglio alcune distribuzioni note che ci torneranno comode per alcune applicazioni.

2.1.1 Distribuzione binomiale

La descrizione binomiale descrive un fenomeno che avviene secondo uno schema successo/insuccesso in cui si considerano n prove **indipendenti** con due soli esiti.

- n è il numero di prove
- k è il numero di successi ($k \leq n$)
- p è la probabilità di successo
- $q = 1 - p$ è la probabilità di insuccesso

X è la v.a. che conta il numero di successi ottenuti. Sia $p(k) = P(X = k)$ la probabilità di ottenere k successi. Proviamo a calcolare questa probabilità a partire dalla definizione classica.

Per farlo dobbiamo prima introdurre gli n eventi

A$-i$ = "successo all'i-esima prova".

Introduciamo pure l'evento

A = "k successi nelle prime k prove"

Calcoliamo la $P(A) \equiv P(\mathrm{A}_1 \cap \mathrm{A}_2 \cap \cdots \cap \mathrm{A}_k \cap \overline{\mathrm{A}}_{k+1} \cap \cdots \cap \overline{\mathrm{A}}_n)$. Dato che gli eventi sono indipendenti avremo

$$P(\mathrm{A}_1 \cap \mathrm{A}_2 \cap \cdots \cap \mathrm{A}_k \cap \overline{\mathrm{A}}_{k+1} \cap \cdots \cap \overline{\mathrm{A}}_n) = P(\mathrm{A}_1)P(\mathrm{A}_2)P(\mathrm{A}_k)P(\overline{\mathrm{A}}_{k+1}) \dots P(\overline{\mathrm{A}}_n) = p^k(1-p)^{n-k}.$$

Notiamo che, data l'indipendenza degli eventi, la probabilità di ottenere k successi nelle prime k prove sarà uguale alla probabilità di ottenere k successi in un qualsiasi ordine. Per cui, per trovare la probabilità di ottenere esattamente k successi in n prove, ci basterà contare in quanti modi è possibile avere k successi in n prove e moltiplicarlo per la probabilità che abbiamo appena calcolato. In questo modo arriveremo a

$$p(k) = \binom{n}{k} p^k(1-p)^{n-k}.$$

Diremo che una variabile aleatoria X segue una **legge binomiale** di parametri n e p se ha densità discreta

$$p(x) = \begin{cases} \binom{n}{x} p^x(1-p)^{n-x} & \text{se } \; x = 0, 1, \dots, n \\ 0 & \text{altrimenti} \end{cases}$$

In questo caso si scriverà $X \sim B(n, p)$.

Per calcolare la media conviene prima considerare un caso particolare, ovvero quello della legge di **Bernoulli**. La legge di Bernoulli non è altro che una binomiale con $n = 1$, ovvero descrive la probabilità di successo in un'unica prova (testa o croce?).

Chiamiamo $X_i \sim B(1, p)$ t.c

$$X_i = \begin{cases} 1 & \text{se successo all'iesima prova} \\ 0 & \text{altrimenti} \end{cases}$$

Da definizione si ha che

$$E[X_i] = \sum_{i=1}^{n} x_i p(x_i) = \binom{1}{1} 1 \cdot p + \binom{1}{0} 0 \cdot (1-p) = p$$

Sfruttando le proprietà della media possiamo quindi calcolare

$$E[X] = E[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} p = np$$

Possiamo operare analogamente per ricavare la varianza. Iniziamo valutando $\mathrm{Var}(X_i)$. Per farlo ci serve

$$E[X_i^2] = \sum_{i=1}^{n} x_i^2 p(x_i) = \binom{1}{1} 1^2 \cdot p + \binom{1}{0} 0^2 \cdot (1-p) = p$$

Da questo possiamo semplicemente calcolare

$$\mathrm{Var}(X_i) = E[X_i^2] - E[X_i]^2 = p - p^2 = p(1-p)$$

Sfruttando l'ipotesi che le prove debbano essere indipendenti, possiamo calcolare

$$\mathrm{Var}(X) = \mathrm{Var}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \mathrm{Var}(X_i) = \sum_{i=1}^{n} p(1-p) = np(1-p)$$

Prima di proseguire facciamo un paio di esempi per prendere familiarità.

Esercizio 1

Nella trascrizione di tre file si sono verificati 5 errori distribuiti randomicamente tra i tre file. Qual è la probabilità che in un singolo file a) ci sia un solo errore b) ci siano più di tre errori?

Svolgimento

Questo fenomeno rientra all'interno dello schema successo/insuccesso. Dato uno specifico file, il mio successo sarà trovare un errore. Dato che gli errori sono distribuiti randomicamente tra i tre file la mia probabilità di successo sarà $p = 1/3$. Il numero di prove, ovviamente, sarà il numero di errori, per cui $n = 5$.

La variabile aleatoria che descrive il numero di errori in un singolo file seguirà dunque una distribuzione binomiale di parametri n e p, ovvero $X \sim B(n, p)$. Per cui:

a) La probabilità che ci sia un solo errore è $P(X = 1)$.

$$P(X = 1) = \binom{5}{1} \left(\frac{1}{3}\right) \left(\frac{2}{3}\right)^4 = 0.3292$$

b) La probabilità che ci siano più di tre errori si riduce alla probabilità che ci siano o quattro o cinque errori, ovvero $P(X > 3) = P(X = 4) + P(X = 5)$. Sempre usando l'espressione della densità binomiale troviamo

$$P(X > 3) = \binom{5}{4} \left(\frac{1}{3}\right)^4 \left(\frac{2}{3}\right) + \binom{5}{5} \left(\frac{1}{3}\right)^5 = 0.0453$$

Esercizio 2

Un segnale è costituito da 5 bits. La probabilità di ricevere un singolo bit distorto è $p = 0.1$. Qual è la probabilità di ricevere un segnale distorto?

Svolgimento

Ecco un altro problema che si inquadra nello schema successo/insuccesso. In questo caso il successo è osservare un bit distorto, la probabilità di successo è $p = 0.1$ mentre il numero di prove è il numero di bit, ovvero $n = 5$.

La variabile aleatoria che conta il numero di bit distorti seguirà pertanto una distribuzione di Bernoulli $X \sim B(n, p)$.

La probabilità che cerchiamo è che il segnale sia distorto, ovvero che almeno un bit sia distorto. Introduciamo l'evento S="segnale distorto" e avremo

$$P(S) = 1 - P(X = 0) = 1 - \binom{5}{0} 0.1^0 \ 0.9^5$$

2.1.2 Distribuzione multinomiale

È una generalizzazione della distribuzione binomiale al caso in uci ho n prove ripetute e indipentneti e m possibili esiti (esempio: il lancio del dado). Definiamo q_i la probabilità di ottenere l'i-esimo risultato. ovviamente $\sum_{i=1}^{m} q_i = 1$.

Introduciamo la variabile aleatoria Y_i che conta quante volte si è verificato l'iesimo risultato. Possiamo quindi definire $Y = (Y_1, Y_2, \ldots, Y_m)$, che sarà una v.a. *multivariata* che conta quante volte si è verificato ciascun esito. L v.a. Y è essenzialmente un vettore di m v.a.! La v.a. seguirà una **legge multinomiale** e si indicherà così $Y \sim B(n, q_1, q_2, \ldots, q_m)$.

Dato un generico vettore $\overline{\omega} = (\omega_1, \omega_2, \ldots, \omega_m)$ siamo interessati in $P(Y = \overline{\omega}) = P(Y_1 = \omega_1, Y_2 = \omega_2, \ldots, Y_m = \omega_m)$.

Ragionando in maniera analoga al caso della binomiale (salteremo la dimostrazione), si può vedere che

$$P(Y = \overline{\omega}) = \frac{n!}{\omega_1!\omega_2! \ldots \omega_m!} q_1^{\omega_1} q_2^{\omega_2} \ldots q_m^{\omega_m}$$

.

Ovviamente, se $m = 2$ ritroviamo la distribuzione binomiale.

Esempio

Calcolare la probabilità che lanciando un dado non truccato quattro volte esca tre volte 6 e una volta 2.

Svolgimento

La variabile che descrive questo evento segue una multinomiale in cui ogni risultato ha la stessa probabilità di accadere $p = 1/6$. La probabilità che cerchiamo, quindi sarà

$$P(Y = \{0, 1, 0, 0, 0, 3\}) = \frac{4!}{0!1!3!0!0!0!3!} \frac{1}{6} \frac{1}{6^3} = \frac{4}{6^4}$$

OBS: dato che il dado ha tutti i risultati equiprobabili avremmo potuto ricavare lo stesso risultato utilizzando la definizione classica di probabilità!

2.1.3 Distribuzione di Poisson

Questa distribuzione si usa quando in uno schema successo/insuccesso si ha un alto numero di prove e una bassa possibilità di successo. Essa dipende da un unico parametro λ. La distribuzione di Poisson approssima la distribuzione binomiale per $n \gg 1$ e $p \ll 1$ e in questo caso $\lambda = np$.

Una v.a. X segue una legge di Poisson di parametro λ se ha distribuzione

$$p(x) = \begin{cases} e^{-\lambda}\frac{\lambda^x}{x!} & \text{se} \quad x = 0, 1, 2, \dots, n \\ 0 & \text{altrimenti} \end{cases}$$

In questo caso si scriverà $X \sim \text{Pois}(\lambda)$.

Dimostriamo intanto che è effettivamente una densità

$$\sum_{x\in\mathbb{R}} p(x) = e^{-\lambda}\sum_{x=0}^{\infty}\frac{\lambda^x}{x!} = e^{-\lambda}e^{\lambda}.$$

Nell'ultimo passaggio abbiamo risommato lo sviluppo in serie di McLaurin dell'esponenziale (in caso, cercate tra gli appunti di analisi 1!).

Adesso invece dimostriamo che questa distribuzione è esattamente la stessa che otterremmo da una binomiale nel limite $n \to \infty$ e $p \to 0$. Se $X \sim B(k, n)$. Poniamo $\lambda = np$, allora

$$\begin{aligned} P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} &= \frac{n!}{k!(n-k)!}\left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \frac{n!}{n^k(n-k)!}\frac{\lambda^k}{k!}\left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k}. \end{aligned}$$

Osserviamo che

- $\frac{n!}{n^k(n-k)!} = \frac{n(n-1)(n-2)\dots(n-k)}{n^k} \xrightarrow[n\to\infty]{} 1$
- $\left(1 - \frac{\lambda}{n}\right)^n \xrightarrow[n\to\infty]{} e^{-\lambda}$ (anche qui, appunti di analisi 1!)
- $\left(1 - \frac{\lambda}{n}\right)^k \xrightarrow[n\to\infty]{} 1$

Per cui avremo che

$$P(X = k) \xrightarrow[n\to\infty, p\to 0]{} \frac{\lambda^k}{k!}e^{-\lambda}$$

N.B. Anche se non abbiamo utilizzato esplicitamente, il limite $p \to 0$ è molto importante! Infatti questo limite permette di avere $\lambda \sim 1$, altrimenti divergerebbe al divergere di n!

Calcoliamo adesso la speranza matematica e la varianza

$$E[X] = e^{-\lambda}\sum_{x=0}^{\infty} x\frac{\lambda^x}{x!} = e^{-\lambda}\lambda\sum_{x=0}^{\infty}\frac{\lambda^{(x-1)}}{(x-1)!} = e^{-\lambda}\lambda e^{\lambda} = \lambda$$

Analogamente possiamo calcolare

$$\begin{aligned} E[X^2] =& e^{-\lambda}\sum_{x=0}^{\infty} x^2\frac{\lambda^x}{x!} = e^{-\lambda}\sum_{x=0}^{\infty} x(x-1+1)\frac{\lambda^x}{x!} \\ =& e^{-\lambda}\sum_{x=0}^{\infty} x\frac{\lambda^x}{x!} + e^{-\lambda}\sum_{x=0}^{\infty} x(x-1)\frac{\lambda^x}{x!} = \lambda + e^{-\lambda}\lambda^2\sum_{x=0}^{\infty}\frac{\lambda^{(x-2)}}{(x-2)!} \\ =& \lambda + e^{-\lambda}\lambda^2 e^{\lambda} = \lambda + \lambda^2 \end{aligned}$$

Di conseguenza possiamo calcolare la varianza usando

$$\mathrm{Var}(X) = E[X^2] - E[X]^2 = \lambda + \lambda^2 - \lambda^2 = \lambda.$$

Per concludere enunciamo una proprietà importante delle v.a. di Poisson. Se $X \sim \mathrm{Pois}(\lambda)$ e $Y \sim \mathrm{Pois}(\mu)$ sono variabili indipendenti, allora $X + Y \sim \mathrm{Pois}(\lambda + \mu)$.

2.1.4 Distribuzione ipergeometrica

Questa distribuzione si usa per descrivere il caso di estrazioni senza rimpiazzo, ovvero il caso in cui le singole estrazioni non sono indipendenti ma sono influenzate dalle precedenti.

Immaginiamo di avere m oggetti di cui b hanno una data proprietà e r no. Voglio estrarre $n \leq b + r$ di questi oggetti e voglio calcolare la probabilità che abbiano la proprietà b. Sia X la v.a. che descrive questo processo, voglio calcolare $P(X = x)\ \forall x = 0, 1, 2, \ldots, n$.

Sostanzialmente dobbiamo contare quanti sono i casi favorevoli e quanti quelli possibili. Usando la combinatoria stimo:

1. In quanti modi posso scegliere x oggetti tra $b \to \binom{b}{x}$
2. In quanti modi posso scegliere $n - x$ oggetti tra $r \to \binom{r}{n-x}$
3. In quanti modi posso scegliere n oggetti tra $b + r \to \binom{b+r}{n}$

Raccogliendo tutto avremo che

$$p(x) = \begin{cases} \dfrac{\binom{b}{x}\binom{r}{n-x}}{\binom{b+r}{n}} & \text{if} \quad x = 0, 1, \ldots, n \\ 0 & \text{altrimenti} \end{cases}$$

Senza entrare troppo nel dettaglio dei conti si può dimostrare che

$$E[X] = \frac{nb}{b+r} \quad \mathrm{Var}(X) = \frac{nbr}{(b+r)^2}\frac{b+r-n}{b+r-1}.$$

2.1.5 Distribuzione geometrica

Una variabile aleatoria X segue una legge geometrica di parametro p, con $0 \leq p \leq 1$ se ha densità

$$p(x) = \begin{cases} p(1-p)^{x-1} & \text{Se } \ x = 1, 2, \ldots, n \\ 0 & \text{altrimenti} \end{cases}$$

In questo caso si scriverà $X \sim \text{Geom}(p)$.

Notiamo subito che, a differenza delle precedenti distribuzioni, questa densità non è definita per $x = 0$. In realtà è possibile definire una densità geometrica, in maniera del tutto equivalente, che contenga anche lo 0, ma per evitare confusione ci limiteremo a considerare questa.

Per prima cosa dimostriamo che questa è realmente una densità

$$\sum_{x=1}^{\infty} p(x) = \sum_{x=1}^{\infty} p(1-p)^{x-1} = p \sum_{j=0}^{\infty} (1-p)^j = \frac{p}{1-(1-p)} = 1.$$

Questa distribuzione è intimamente legata alla distribuzione binomiale. Introduciamo T, v.a. che valuta il **tempo di primo successo** , ovvero il numero di tentativi necessario per osservare il primo successo in un processo binomiale. Vogliamo calcolare la probabilità $P(T = n)$. Per prima cosa osserviamo che $P(T > n-1) = P(T > n) + P(T = n)$. Pertanto, $P(T = n) = P(T > n-1) - P(T > n)$. Per valutare queste due probabilità, introduco $X^{(n)} \sim B(n, p)$.

Allora

$$P(T > n) = P(X^{(n)} = 0) = \binom{n}{0} (1-p)^n$$

Analogamente avremo

$$P(T > n-1) = (1-p)^{n-1}$$

Per cui, raccogliendo tutto assieme ritroviamo

$$P(T = n) = (1-p)^{n-1} - (1-p)^n = (1-p)^{n-1}(1-(1-p)) = p(1-p)^{n-1}.$$

La proprietà sicuramente più importante di cui gode la legge geometrica è la **proprietà di mancanza di memoria**. Ovvero, l'occorrere del primo successo è indipendente dalla storia precedente (esistono i numeri ricorrenti al lotto?).

In formule, questa proprietà si esprime così:

$$P(T = m + n \mid T > n) = P(T = m) \ \ \forall n, m > 0 \in \mathbb{N}.$$

Per dimostrarla basta utilizzare alcune proprietà delle probabilità condizionali che abbiamo già incontrato in precedenza, ovvero:

$$P(T = n+m \mid T > n) = \frac{P(T = n + m \ , T > n)}{P(T > n)} = \frac{P(T = m + n)}{P(T > n)} = \frac{p(1-p)^{m+n-1}}{(1-p)^n} = p(1-p)^{m-1}.$$

Al volo, segnaliamo che per questa distribuzione si ha

$$E[X] = \frac{1}{p}, \quad \text{Var}(X) = \frac{1-p}{p^2}.$$

2.2 Variabili aleatorie continue

Nel caso in cui $X(\Omega)$ sia continuo la v.a. si dice **continua**. Nel caso di v.a. continue, la funzione di ripartizione sarà una funzione

$$F_X : \mathbb{R} \to [0,1], \text{ t.c. } F_X(t) = P(X \leq t).$$

La funzione di ripartizione gode delle seguenti quattro proprietà

1. $0 \leq F_X(t) \leq 1 \ \forall t \in \mathbb{R}$;
2. $F_X(t_1) \leq F_X(t_2) \ \forall t_1, t_2 \in \mathbb{R}, \text{ t.c. } t_1 \leq t_2$ (monotonia);
3. $\lim_{t\to-\infty} F_X(t) = 0, \ \lim_{t\to\infty} F_X(t) = 1$;
4. $\lim_{\varepsilon\to0^+} F_X(t+\varepsilon) = F_X(t)$ (continuità a destra).

È importante notare che in generale la funzione di ripartizione può non essere continua. Sia t un punto di discontinuità, data l'ipotesi di monotonicità, sappiamo che esistono finiti i seguenti limiti

$$\lim_{\varepsilon\to0^-} F_X(t+\varepsilon) = F_X(t^-) \quad \lim_{\varepsilon\to0^+} F_X(t+\varepsilon) = F(t^+).$$

In questo caso avremo che

$$P(X = t) = F_X(t^+) - F_X(t^-).$$

OBS: se la funzione di ripartizione è continua, la probabilità che X assuma un singolo valore t è zero. Questo non deve sorprenderci, perché lavorando nel continuo quello che è importante non è il singolo valore quanto gli intervalli. In generala, se X è una v.a. continua, varrà la seguente catena di uguaglianze:

$$P(a \leq X \leq b) = P(a \leq X < b) = P(a < X \leq b) = P(a < X < b) = F_X(b) - F_X(a).$$

Una v.a. si dice **assolutamente continua** se ammette **densità** $f(x)$ tale che

$$F_X(t) = \int_{-\infty}^{t} f(x)dx,$$

con

$$f(x) \geq 0 \ \forall x \in \mathbb{R}, \text{ t.c. } \int_{-\infty}^{\infty} f(x)dx = 1.$$

OBS (puramente matematica): la densità non è unica, dato che è definita a meno di un sottoinsieme di misura nulla!

Se esiste finita, ovvero se $\int_{-\infty}^{\infty} |x| f(x)dx < \infty$, posso definire la speranza matematica di una v.a. continua come

$$E[X] = \int_{-\infty}^{\infty} x f(x)dx.$$

Analogamente, se esiste finita posso definire la varianza di una v.a. continua

$$\mathrm{Var}[X] = \int_{-\infty}^{\infty} (x - E[X])^2 f(x)dx < \infty.$$

Se esistono finiti, posso definire tutti i momenti centrati e non di ogni ordine analogamente a quanto fatto per le variabili discrete semplicemente sostituentdo $\sum_{x\in\mathbb{R}} \mapsto \int_{-\infty}^{\infty} dx$ e $p(x) \mapsto f(x)$.

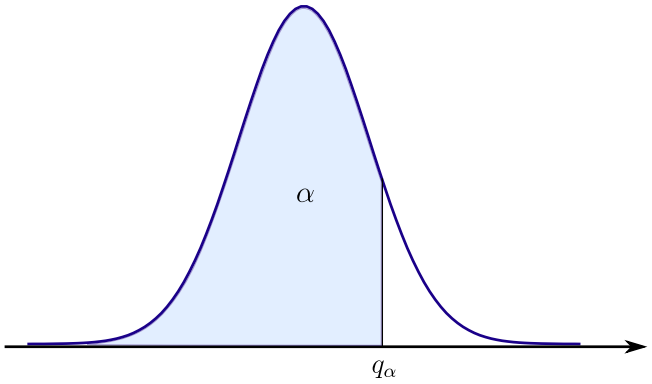

Figura 2: Definizione grafica di quantile di ordine α

Definisco **quantile di ordine** α la quantità

$$q_\alpha = \sup\{r \in \mathbb{R} : F_X(r) = \alpha\}.$$

In parole semplici, il quantile di ordine α è quel valore possibile della variabile aleatoria che divide l'area sottesa alla distribuzione in due parti di ampiezza α (a sinistra) e $1-\alpha$ (a destra). Va da sé che $F_X(q_\alpha) = \alpha$.

2.2.1 Distribuzione uniforme

Una v.a. segue una legge uniforme in $[a, b] \in \mathbb{R}$ se ha densità

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{se} \quad x \in [a, b] \\ 0 & \text{altrimenti.} \end{cases}$$

In questo caso si scriverà $X \sim U([a, b])$. Notiamo che questa distribuzione descrive una variabile aleatoria che può assumere tutti i valori nell'intervallo $[a, b]$ con uguale probabilità.

Verifichiamo che sia effettivamente una densità:

$$\int_{-\infty}^{\infty} dx f(x) = \frac{1}{b-a} \int_a^b dx = 1.$$

Possiamo calcolare, usando la definizione, la funzione di ripartizione $F_X(t) = \int_{-\infty}^{t} f(x)dx = \int_a^t f(x)dt$ e trovare

$$F_X(t) = \begin{cases} 0 & \text{se } t < a \\ \frac{t-a}{b-a} & \text{se } t \in [a, b] \\ 1 & \text{se } t > b \end{cases}$$

OBS: è una funzione continua in $\mathbb{R}$!

Usando la definizione possiamo calcolare anche la media

$$E[X] = \int_{-\infty}^{\infty} dx \; x f(x) = \int_a^b dx \; \frac{x}{b-a} = \frac{b^2 - a^2}{2(b-a)} = \frac{b+a}{2}$$

e la varianza

$$Var(X) = E[X^2] - E[x]^2 = \int_a^b dx\ \frac{x^2}{b-a} - \frac{(a+b)^2}{4} = \frac{b^3-a^3}{3(b-a)} - \frac{(a+b)^2}{4} = \frac{(b-a)^2}{12}$$

2.2.2 Legge normale standard

Eccoci alla distribuzione più importante di tutto il corso: la distribuzione normale standard. Una v.a. X segue una legge normale standard o gaussiana se ha densità

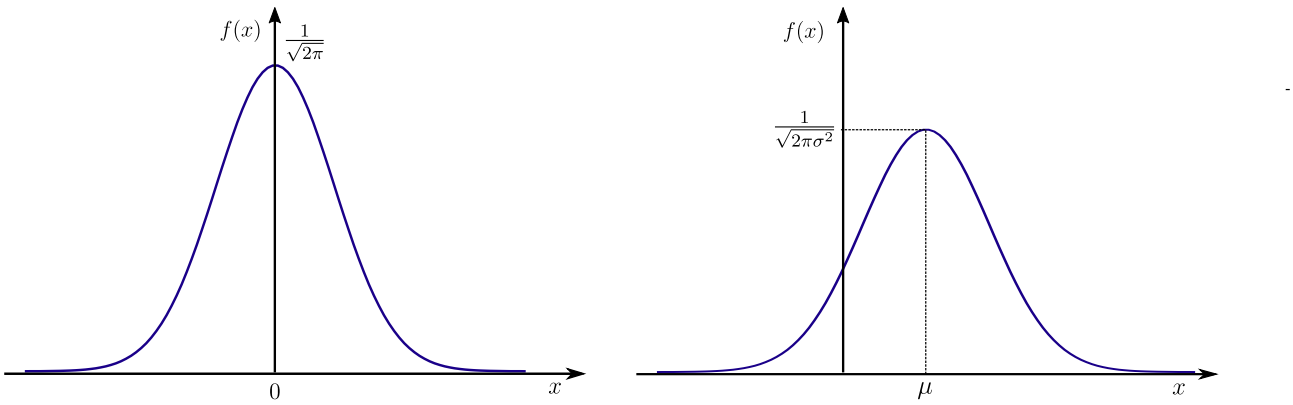

Figura 3: A sinistra, normale standard. A destra, normale di media μ e varianza σ^2

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

In questo caso si scriverà $X \sim N(0,1)$.

La distribuzione normale standard (pannello a sinistra in Fig. 3) è una distribuzione simmetrica rispetto allo 0. Per questo motivo si è ha che $E[X] = 0$. Proviamo a dimostrarlo

$$\begin{aligned} E[X] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dx\ x e^{-\frac{x^2}{2}} = -\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dx\ \frac{d}{dx} e^{-\frac{x^2}{2}} = -\frac{1}{\sqrt{2\pi}} \left[e^{-\frac{x^2}{2}} \right]_{-\infty}^{\infty} \\ &= -\frac{1}{\sqrt{2\pi}} \left[\lim_{x\to\infty} \left(e^{-\frac{x^2}{2}} \right) - \lim_{x\to-\infty} \left(e^{-\frac{x^2}{2}} \right) \right] = 0. \end{aligned}$$

Analogamente si può dimostrare che $\mathrm{Var}(X) = 1$.

Dalla simmetria della distribuzione avremo che i quantili della normale standard,che da ora in poi indicheremo con ϕ_α sono simmetrici rispetto allo zero, ovvero $\phi_\alpha = -\phi_{1-\alpha}$.

Data $X \sim N(0,1)$ possiamo costruire una v.a. Y che segue una normale di parametri reali μ e $\sigma^2 > 0$ attraverso la relazione

$$Y = \sigma X + \mu$$

. In questo caso si scriverà $Y \sim N(\mu, \sigma^2)$.

La densità di tale variabile aleatoria sarà

$$f_Y(X) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Notiamo che, notele proprietà di media è varianza, è molto semplice calcolare media

$$E[Y] = E[\sigma X + \mu] = \sigma E[X] + \mu = \mu$$

e varianza

$$\mathrm{Var}(Y) = \mathrm{Var}(\sigma X + \mu) = \sigma^2 \mathrm{Var}(X) = \sigma^2$$

Prima di concludere elenchiamo alcune proprietà interessanti (e utili) della distribuzione normale.

- In ottima approssimazione $f_Y(x) \approx 0$ quando $|x - \mu| > 3\sigma$. Questa si chiama **legge dei tre sigma**.
- Se $X \sim N(\mu_1, \sigma_1^2)$ e $Y \sim N(\mu_2, \sigma_2^2)$ sono indipendenti, allora $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$
- Se q_α è il quantile di ordine α di $Y \sim N(\mu_1, \sigma^2)$ e ϕ_α quello di $X \sim N(0, 1)$, allora $q_\alpha = \sigma\phi_\alpha + \mu$.

2.2.3 Legge esponenziale

Questa legge descrive la "durata di vita" di un fenomeno che "non invecchia", ossia di un fenomeno privo di memoria. Ad esempio, il tempo di decadimento di una particella radioattiva.

La densità di tale distribuzione è

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{se } x > 0 \\ 0 & \text{se } x \leq 0. \end{cases}$$

Notiamo per prima cosa che è una densità:

$$\int_{\mathbb{R}} f(x)dx = \int_0^\infty \lambda e^{-\lambda x} dx = -\int_0^\infty \frac{d}{dx} e^{-\lambda x} = -\left[e^{-\lambda x}\right]_0^\infty = 1.$$

Integrando per parti si può trovare

$$E[X] = \frac{1}{\lambda}, \qquad \mathrm{Var}(X) = \frac{1}{\lambda^2}.$$

È interessante calcolare la funzione di ripartizione

$$F_X(t) = \begin{cases} 0 & \text{se } \ t \leq 0 \\ 1 - e^{-\lambda t} & \text{se } \ t > 0. \end{cases}$$

È utile introdurre la **funzione di sopravvivenza** $S(t) = 1 - F_X(t) = e^{-\lambda t}$. Notiamo che essa non è altro che la probaiblità di osservare il fenomeno oltre un tempo t, infatti $S(t) = 1 - F_X(t) = 1 - P(X \leq t) = P(X > t)$.

Un'altra osservazione che va fatta è che la distribuzione esponenziale può essere messa in relazione con la distribuzione geometrica. In un certo senso è la "versione continua" della distribuzione geometrica. È possibile ritrovare dall'una l'altra sostituendo $p \leftrightarrow \lambda$ e $n \leftrightarrow x$.

Proprio come la distribuzione geometrica, la distribuzione esponenziale gode della **proprietà di mancanza di memoria**, ovvero $P(X \geq x + t | X \geq x) = P(X \geq t) = S(t)$.

$$P(X \geq x + t | X \geq x) = \frac{P(X \geq x + t, X \geq x)}{P(X \geq x)} = \frac{P(X \geq x + t)}{P(X \geq x)} = \frac{e^{-\lambda(x+t)}}{e^{-\lambda x}} = e^{-\lambda t}.$$

2.2.4 Legge del χ^2

Una v.a. X segue una legge del χ^2 a n gradi di libertà, $X \sim \chi^2(n)$, se ha densità

$$f(x) = \begin{cases} \frac{x^{\frac{n}{2}-1}e^{-\frac{x}{2}}}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} & \text{se } \ x > 0 \\ 0 & \text{altrimenti,} \end{cases}$$

dove $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx \ \forall \alpha > 0$ è la funzione Gamma di Eulero. Questa distribuzione è importante per applicazioni nei test d'ipotesi che vedremo in seguito. Pertanto elenchiamo alcune proprietà

- Se $X \sim N(0,1) \Rightarrow X^2 \sim \chi^2(1)$
- Se $X_1, X_2, \ldots, X_n \sim N(0,1) \Rightarrow \sum_{i=1}^n X_i^2 \sim \chi^2(n)$
- Se $X \sim \chi^2(n)$ e $Y \sim \chi^2(m)$, allora $X + Y \sim \chi^2(n+m)$.

2.2.5 Legge t di Student

Una v.a. X segue una legge ti di Student a n gradi di libertà, $X \sim T(n)$ se ha densità

$$f(x) = \frac{A_0}{\left(1 + \frac{x^2}{n}\right)^{\frac{n+1}{2}}}, \quad A_0 = \frac{1}{\sqrt{n\pi}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}.$$

Anche la legge t di Student ha un test d'ipotesi dedicato. Pertanto ricordiamo un paio di proprietà che ci torneranno comode:

- I quantili della t-Student si indicano con la lettera t e sono simmetrici, pertanto $t_\alpha = -t_{1-\alpha}$.
- Se $n \to \infty$, allora $T(n) \sim N(0,1)$. Nella pratica, basta $n > 30$.

2.3 Cenni di v.a. multivariate

Tutti i discorsi presentati fin'ora si possono estendere al caso di più v.a. (ad esempio: la distribuzione multinomiale è una distribuzione multivariata!). Immaginiamo di avere due v.a. X e Y. Per comodità le immaginiamo discrete ma tutto quanto diremo si può estendere al caso continuo sostituendo alle sommatorie gli integrali e alle densità discrete quelle continue.

Definiamo:

densità marginali

$$p_X(x) = P(X = x), \quad p_Y(y) = P(Y = y);$$

densità conginuntà

$$p(x,y) = P(X = x, Y = y);$$

densità condizionale

$$p_{X|Y}(x|y) = \frac{p(x,y)}{p_Y(y)} \quad \text{se } p_Y(y) > 0$$

Dalla conoscenza della congiunta possiamo ricavarci tutto ciò che ci serve, infatti:

$$p_X(x) = \sum_{y \in \mathbb{R}} p(x,y) \quad \text{e } p_Y(y) = \sum_{x \in \mathbb{R}} p(x,y).$$

Iteressante notare che se X e Y sono indipendenti, allora $p(x,y) = p_X(x)p_Y(y)$ e, di conseguenza, $p_{X|Y}(x|y) = p_X(x)$.

Infine introduciamo il **coefficiente di correlazione lineare**

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}, \quad \text{t.c. } -1 \leq \rho \leq 1.$$

In base al valore di questo coefficiente possiamo dedurre se c'è correlazione (NON CAUSALITÀ!) tra le due variabili aleatorie. Se il coefficiente è positivo la variabili saranno correlate, se negativo saranno anticorrelate. In particolare, si avrà anche che

- Se $|\rho| < 0.3$ la correlazione (o anticorrelazione) è debole;
- Se $0.3 < |\rho| < 0.7$ la correlazione (o anticorrelazione) è moderata;
- Se $0.7 < |\rho|$ la correlazione (o anticorrelazione) è forte;

2.4 Esercizi

Esercizio 1

In un libro di 500 pagine sono distribuiti a caso 300 errori di stampa. Qual è la probabilità che una data pagina contenga almeno due errori.

Svolgimento

Questo problema si inquadra nello schema successo-insuccesso, dove il successo è quello di trovare un errore in una data pagina. Gli errori sono 300, il numero di tentativi che ho è $n = 300$. Essendo gli errori distribiti a caso tra le cinquecento pagine, la probabilità di trovare un errore in una data pagina sarà $p = 1/500$. Pertanto, la v.a. che descrive questo problema sarà $X \sim B(n,p)$ e la probabilità q di avere almeno due errori in una pagina sarà

$$q = P(X \geq 2) = 1 - P(X = 0) - P(X = 1) = 0.121769.$$

Avremmo potuto risolvere lo stesso esercizio in maniera del tutto analoga notando che $n = 300 \gg 1$ e $p = 1/500 \ll 1$, per cui lo stesso problema può essere descritto da una v.a. $Y \sim \mathrm{Pois}(\lambda)$, con $\lambda = np$. Usando questo approccio avremmo trovato che

$$q = P(Y \geq 2) = 1 - P(Y = 0) - P(Y = 1) = 0.121901.$$

Osserviamo che i due numeri sono praticamente identici. Non sono esattamente uguali perché l'uguaglianza delle due distribuzioni vale strettamente solo nel limite $p \to 0$ e $n \to \infty$.

Esercizio 2

Lancio un dado tre volte. a) Qual è la probabilità di ottenere 6 almeno una volta? b) Qual è la probabilità che ciò avvenga al quinto tentativo? c) Quante volte devo lanciare il dado affinché

abbia una probabilità del 90% di avere 6? d) Questo risultato sarebbe stato diverso se avessimo considerato un altro numero?

Svolgimento

a) Schema successo-insuccesso in cui il successo è rappresentato dall'ottenere 6. Ho $n = 3$ lanci a disposizione e una probabilità di successo $p = 1/6$. Pertanto posso descrivere questo problema usando $X \sim B(n, p)$ e la probabilità richiesta sarà

$$q = P(X > 0) = 1 - P(X = 0) = 0.4219$$

b) Il tempo di primo successo T è una v.a. che segue una legge geometrica, ovvero $T \sim \text{Geom}(p)$. Pertanto la probabilità richiesta è

$$q = P(T = 5) = p(1 - p)^4 = 0.0803$$

c) Utilizzo sempre il fatto che il tempo di primo successo sia una v.a. che segue una distribuzione geometrica. La richiesta fatta dal problema, in altre parole, equivale a trovare un n tale per cui $P(T \leq n) = 0.9$. Pertanto

$$P(T \leq n) = 1 - P(T > n) = 1 - (1 - p)^n > 0.9 \Rightarrow (1 - p)^n < 0.1.$$

Per risolvere questa disequazione passiamo ai logaritmi

$$n \log(1 - p) > \log(0.1) \Rightarrow n > \frac{\log(0.1)}{\log(1 - p)} = 12.63.$$

Pertanto, avremo bisogno di lanciare il dado almeno 13 volte per avere una probabilità del 90% di ottenere 6.

d) Fintanto che il dado è equilibrato, il risultato di sopra non sarebbe cambiato se avessimo risolto il problema per un altro numero.

Esercizio 3

Su un tavolo ci sno due monete. Quando vengono lanciate, una moneta dà testa con $p = 0.5$, l'altra con $p = 0.6$. Una moneta scelta a caso viene lanciata. a) Qual è la probabilità che esca testa? b) Qual è la probabilità che, sapendo che è uscita testa, la moneta lanciata fosse quella equilibrata?

Svolgimento

a) Per risolvere questo esercizio utilizziamo il teorema delle probabilità totali.

Definiamo gli eventi

T = "Esce testa" A = "Lancio moneta equilibrata" B = "Lancio moneta truccata".

Dato che la scelta è casuale, la probabilità di lanciare la moneta equilibrata è identica a quella di lanciare la moneta truccata e quindi $P(A) = P(B) = 0.5$. Dato che l'intero spazio dell'eventi è costituito dal lancio di una o dell'altra moneta avremo che,

$$P(T) = P(T|A)P(A) + P(T|B)P(B) = 0.5 \cdot 0.5 + 0.5 \cdot 0.6 = 0.55$$

b) Usiamo il teorema di Bayes

$$P(A|T) = \frac{P(T|A)P(A)}{P(T)} = \frac{0.5 \cdot 0.5}{0.55} = 0.45454545.$$

Esercizio 4

Il numero di anni di funzionamento di una radio segue una exp(1/8). a) Qual è la probabilità che funzioni più di 10 anni? Qual è la probabilità che funzioni più di 10 anni se ne ha già funzionati 4?

Svolgimento

Questo problema è descritto da $X \sim \exp(1/8)$. Allora

a) $P(X > 10) = 1 - P(X \leq 10) = 1 - (1 - e^{-\frac{10}{8}}) = e^{-\frac{10}{8}} = 0.2865.$

b) Per la proprietà di assenza di memoria della distribuzione esponenziale, avremo che

$$P(X > 10|X > 4) = P(X > 6) = e^{-\frac{6}{8}} = 0.4724$$

Esercizio 5 L'altezza degli italiani segue una normale di media $\mu = 175cm$ e varianza $\sigma^2 = 8cm^2$. a) Qual è la percentuale di italiani con altezza superiore a $190cm$? b) Qual è la percentuale di italiani con altezza inferiore a $153cm$?

Svolgimento

Data $X \sim N(175, 8)$, avremo

a) $P(X > 190) = 1 - P(X \leq 190) = 1 - F_X(190) = 0.047.$

b) $P(X < 153) = F_X(153) = 0.0073.$

OBSNel calcolare queste probabilità, non è importante dove metto l'uguaglianza, tanto, essendo la normale una distribuzione continua, la $P(X = a) = 0$, per cui $P(X \geq 190) = P(X > 190)$.

Esercizio 6 Ad un esame la distribuzione dei voti segue una normale di medi $\mu = 24$ e deviazione standard è $\sigma = 4$. Calcolare a) la probabilità che il voto sia maggiore di 27; b) la probabilità che il voto sia minore di 22; c) la probabilità che il voto sia tra 23 e25; d) Il voto minimo riportato dal 70% degli studenti; e) Il voto massimo non superato dal 90% degli studenti.

Svolgimento

Per fare questo esercizio utilizziamo funzione di ripartizione di una v.a. $X \sim N(\mu, \sigma^2)$ e i quantili q_α.

a) $P(X > 27) = 1 - P(X \leq 27) = 1 - F(27) = 0.22$

b) $P(X < 22) = F(22) = 0.303$

c) $P(23 < X < 25) = P(X < 25) - P(X < 23) = F(25) - F(23) = 0.197$

d) $x_{\min} = q_{1-70\%}$

e) $x_{\max} = q_{90\%}$

3 Legge dei grandi numeri e Teorema del limite centrale

Ci occupiamo adesso di una legge e un teorema tra i più importanti e più utili della statistica: la legge dei grandi numeri e il teorema del limite centrale. Per fare questo ci servirà dimostrare due disuguaglianze, la disuguaglianza di Markov e la disuguaglianza di Čebišëv.

Disuguaglianza di Markov

Sia X una v.a. a valori non negativi (ovvero X può essere o zero o numeri positivi). Allora

$$P(X \geq \alpha) \leq \frac{E[X]}{\alpha} \quad \forall \alpha.$$

Dimostrazione

Per dimostrare questa disuguaglianza dobbiamo distinguere due casi:

- $E[X] \to +\infty$: in questo caso la disuguaglianza è sempre verificata, dato che la probabilità per definizione è minore o uguale di 1.
- $E[X] < +\infty$: ci serve introdurre una variabile aleatoria ausiliaria che chiamiamo I

  $$I = \begin{cases} 1 & \text{se } X \geq \alpha \\ 0 & \text{se } X < \alpha \end{cases}$$

  Notiamo che se $\frac{X}{\alpha} \geq 1$, allora $I = 1 > 0$. Invece, se $\frac{X}{\alpha} < 1$, allora $I = 0 \leq \frac{X}{\alpha}$ (vi ricordo che $X \geq 0$ per ipotesi!). Di conseguenza abbiamo $0 \leq I \leq \frac{X}{\alpha}$.

  Notiamo che

  $$E[I] = 0 \cdot P(I = 0) + 1 \cdot P(I = 1) = 1 \cdot P(X \geq \alpha) = P(X \geq \alpha).$$

  Ma, considerato che $0 \leq I \leq \frac{X}{\alpha}$, per le proprietà della media avremo

  $$E[I] \leq E\left[\frac{X}{\alpha}\right] = \frac{E[X]}{\alpha}$$

  Mettendo tutto assieme abbiamo dimostrato che

  $$P(X \geq \alpha) \leq \frac{E[X]}{\alpha}.$$

Disuguaglianza di Čebišëv

Sia X una v.a. che ammette media e varianza finiti ($E[X], \mathrm{Var}(X) < \infty$), allora

$$P(|X - E[X]| \geq \eta) \leq \frac{\mathrm{Var}(X)}{\eta^2}.$$

Dimostrazione

Dimostriamo questa disuguaglianza utilizzando una variabile aleatoria ausiliaria $Y = (X - E[X])^2$. Per definizione, $E[Y] = \mathrm{Var}(X)$. Notiamo che $Y \geq 0$ sempre (è un quadrato!), per cui vale certamente la disuguaglianza di Markov. Applichiamola usando $\alpha \equiv \eta^2$!

$$P(Y \geq \eta^2) \leq \frac{E[Y]}{\eta^2} = \frac{\mathrm{Var}(X)}{\eta^2}.$$

Ma noi abbiamo che

$$P(Y \geq \eta^2) = P(|X - E[X]|^2 \geq \eta^2) = P(|X - E[X]| \geq \eta^2).$$

Mettendo tutto assieme troviamo

$$P(|X - E[X]| \geq \eta) \leq \frac{\mathrm{Var}(X)}{\eta^2}.$$

Interludio

Pirma di andare avanti facciamo dei conticini che ci torneranno utili. i

Per prima cosa introduciamo il concetto di successione di v.a. $(X_n)_{n\ in\mathbb{N}}$, ovvero una successione di v.a. indipendenti definite tutte nello stesso spazio di probabilità.

Diremo che una successione $(X_n)_{n\ in\mathbb{N}}$ **converge in probabilità** a X se

$$\lim_{n\to\infty} P(|X_n - X| > \eta) = 0, \quad \forall \eta > 0.$$

In questo caso si scriverà

$$X_n \xrightarrow{\mathcal{P}} X.$$

Aggiungiamo qualcosa in più, ovvero chiamiamo $F_n(t)$ la funzione di ripartizione di X_n e $F(t)$ quella di X. Allora diremo che una successione $(X_n)_{n\ in\mathbb{N}}$ **converge in legge** a X se

$$\lim_{n\to\infty} F_n(t) = F(t).$$

In questo caso si scriverà

$$X_n \xrightarrow{\mathcal{L}} X.$$

Assumiamo ora che le X_n abbiano media μ e varianza σ^2 e consideriamo una nuova successione di v.a. costruita usando $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Per il futuro $\overline{X}_n$ è detta **media campionaria**.

Notiamo che

- $E[\overline{X}_n] = E\left[\frac{1}{n}\sum_{i=1}^n X_i\right] = \frac{1}{n}\sum_{i=1}^n E[X_i] = \frac{1}{n}\sum_{i=1}^n \mu = \frac{n\mu}{n} = \mu.$
- $\mathrm{Var}(\overline{X}_n) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(X_i) = \frac{1}{n^2}\sum_{i=1}^n \sigma^2 = \frac{n\sigma^2}{n} = \frac{\sigma^2}{n}.$

Legge dei grandi numeri in forma debole

Data una successione di v.a. $(X_n)_{n\ in\mathbb{N}}$ tale che le X_n abbiano media μ e varianza σ^2, allora:

$$\lim_{n\to\infty} P(|\overline{X}_n - \mu| > \eta) = 0, \quad \forall \eta > 0,$$

ovvero $\overline{X}_n$ converge in probabilità a μ

$$\overline{X}_n \xrightarrow{\mathcal{P}} \mu.$$

Dimostrazione Per la disguguaglianza di Čebišëv avremo

$$P(|\overline{X}_n - E[\overline{X}_n]| \geq \eta) \leq \frac{\mathrm{Var}(X_n)}{\eta^2}.$$

Per quanto dimostrato prima, però, avremo:

$$P(|\overline{X}_n - \mu| \geq \eta) \leq \frac{\sigma^2}{\eta^2 n} \xrightarrow[n\to\infty]{} 0.$$

Cosa significa in pratica questa legge? Che se io ho una variabile aleatoria di distribuzione ignota di cui voglio conoscere la media, allora prendo un numero molto grande di istanze e calcolo la v.a. $\overline{X}_n$: tante più istanze considero, tanto più questa v.a. tenderà alla media della distribuzione!

Esempio: voglio conoscere la media dei voti degli studenti UniCT ma non conosco la distribuzione. Un modo per stimarla è quella di prendere i voti di tanti studenti e calcoliamo la media campionaria: tanti più studenti considero, tanto più quella v.a. approssimerà la media dei voti degli studenti UniCT.

Teorema del limite centrale

Questo teorema verrà solo enunciato e non dimostrato. Sia data una successione di v.a. X_n di media μ e varianza σ^2, allora la v.a. **normale standardizzata**

$$S_n = \frac{\overline{X}_n - \mu}{\sigma}\sqrt{n}$$

convergerà in legge a una v.a. $S \sim N(0,1)$, ovvero

$$S_n \xrightarrow{\mathcal{L}} S.$$

Ma in pratica cosa significa $n \to \infty$? Per le applicazioni che interessano noi, sia la legge dei grandi numeri che il teorema del limite centrale sono giustificate se $n \geq 30$. Nel caso di distribuzioni binomiali, le due assunzioni sono giustificate se $np \geq 5$ e $n(1-p) \geq 5$.

4 Statistica

Dare la definizione di statistica non è immediato. La statistica è uno strumento del metodo scientifico che si propone di riassumere le caratteristiche di un fenomeno collettivo attraverso l'analisi di singole manifestazioni.

In termini più pratici viene utilizzata per riassumere e analizzare dei dati specifici non organizzati, in modo da caratterizzare l'insieme più grande da cui i dati sono stati estratti. Ci occuperemo di due branche della statistica: la statistica descrittiva e quella inferenziale.

In generale, avremo sempre una **popolazione**, che rappresenta l'insieme che vogliamo caratterizzare, e un suo sottoinsieme detto **campione**da cui estraiamo i dati a cui abbiamo accesso.

Tornando all'esempio della sezione precedente: voglio trovare la media delle medie degli studenti UniCT. La mia popolazione sarà l'insieme di tutti gli studenti UniCT. Se calcolassi la media delle medie su tutti gli studenti avrei la media esatta. Altrimenti potrei prelevare un sottoinsieme di tutti gli studenti e calcolare la media del sottoinsieme per avere un'indicazione della media di tutta la popolazione. In parole povere, la statistica ci fornisce indicazioni su come prelevare il campione e quale quantità guardare per poter estrarre informazioni che possano descrivere in maniera quanto più accurata l'intera popolazione.

4.1 Statistica descrittiva

Come suggerisce il nome, la statistica descrittiva si occupa di descrivere i dati raccolti che, quando non sono ancora organizzati, si definiscono **dati grezzi**.

Dato che un'analisi statistica si effettua su una o più caratteristiche della popolazione. Si distinguono

- Caratteri **qualitativi**: sono caratteri non numerici (es: colore delle macchine, indirizzo di studio...)
- Carattere **quantitativi**: grandezze numeriche (es: numero frequentanti corso, altezza delle persone..)

In quanto segue ci concentreremo principalmente sui caratteri quantitativi. Quest'ultimi si possono distinguere ancora in caratteri **discreti** (sono in numero finito o numerabile) e **continui** (sono un continuo).

Possiamo raggruppare i dati per **singoli valori** o per **classi di valori** .

x_i	f_i	p_i	F_i
x_1	f_1	p_1	F_1
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
x_n	f_n	p_n	F_n

4.1.1 Raggruppamento per singoli valori

Immaginiamo di avere N dati. Per prima cosa ordiniamoli in ordine crescente. Di questi N soltanto n sono distinti. Ad esempio, nell'insieme di $N = 10$ elementi $\{1, 1, 1, 1, 2, 7, 9, 9, 9, 10\}$, abbiamo solo $n = 5$ elementi distinti che sono $1, 2, 7, 9, 10$. Possiamo dare le seguenti definizioni.

- **Variabilità** o **Range**: Differenza tra valore maggiore e valore minore.
- **Frequenza assoluta**, f_i: molteplicità con cui il valore i si è presentato. Ovviamente $\sum_{i=1}^{n} f_i = N$.
- **Frequenze relative** (o probabilità empiriche), p_i: frequenze assolute normalizzate al numero di elementi $p_i = \frac{f_i}{N}$. Ovviamente $\sum_{i=1}^{n} p_i = 1$.
- **Frequenze relative cumulate**, F_i: rappresentano una approssimazione costante a tratti della funzione di ripartizione, $F_i = \sum_{k=1}^{i} p_k$.

I dati vengono in genere rappresentati in tabelle come quella rappresentata sopra.

Ci sono alcune quantità che possiamo valutare per caratterizzare i dati:

- **Media**: $\bar{x} = \frac{1}{N} \sum_{i=1}^{n} x_i f_i = \sum_{i=1}^{n} p_i x_i$
- **Moda**: Il valore più frequente. Se la moda è unica la distribuzione si dice *unimodale*, altrimenti si dice *multimodale*.
- **Mediana**: Il valore che divide l'insieme di dati in due gruppi di uguale numerosità. Esempio: la mediana di $\{13, 25, 17, 18, 22, 25, 26, 28, 30\}$ è $m_e = 22$. Nel caso in cui N sia pari la mediana è data dalla media aritmetica dei due numeri centrali. Esempio: la mediana di $\{13, 25, 17, 18, 22, 23, 25, 26, 28, 30\}$ è $m_e = (22 + 23)/2 = 22.5$. Inoltre, la mediana è quel valore che minimizza la funzione $S(a) = \sum_{i=1}^{N} |x_i - a|$, ovvero $S(m_e) \leq S(a)\ \forall a \in \mathbb{R}$.
- **Quantili di ordine** k: sono i $k - 1$ valori che dividono i dati in k gruppi di uguale numerosità
- **Varianza Empirica**: $\bar{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{n} (x_i - \bar{x})^2 f_i = \sum_{i=1}^{n} (x_i - \bar{x})^2 p_i$.
- **Deviazione standard empirica**: $\bar{\sigma} = \sqrt{\bar{\sigma}^2}$.
- **Momento empirico centrato di ordine** r: $\mu_r = \frac{1}{N} \sum_{i=1}^{n} (x_i - \bar{x})^r f_i = \sum_{i=1}^{n} (x_i - \bar{x})^r p_i$. Ovviamente, la varianza empirica è un momento empirico centrato di ordine $r = 2$. Tra i momenti centrati ce ne sono un paio che sono molto importanti

  1. **Skewness** $\gamma_1 = \frac{(\mu_3)^2}{(\mu_2)^3}$: quantifica l'asimmetria dei dati. La distribuzione di riferimento è la distribuzione normale che ha $\gamma_1 = 0$. Se una distribuzione ha $\gamma_1 < 0$ allora sarà asimmetrica verso sinistra, mentre se $\gamma_1 > 0$ allora sarà asimmetrica verso destra. Talvolta si utilizza il coefficiente $\beta_1 = \frac{(\mu_3)^2}{(\mu_2)^3}$, che però perde informazioni sul segno. **Importante**: l'annullarsi del coefficiente di simmetria è condizione necessaria per la simmetria (ovvero se la distribuzione è simmetrica l'indice di asimmetria sarà 0) ma non sufficiente (ovvero esistono delle distribuzioni asimmetriche con indice di asimmetria nullo).
  2. **Curtòsi** $\beta_2 = \frac{\mu_4}{\mu_2^2}$: misura quanto è "spanciata" la distribuzione. Nel caso della gaussiana si ha $\beta_2 = 3$ (regola dei tre sigma alert!), per cui di solito si preferisce introdurre il **coefficiente di curtòsi** $\gamma = \beta_2 - 3$. Se una distribuzione ha $\gamma = 0$

allora è spanciata quanto una normale (normocurtica), se $\gamma < 1$ la curva sarà più piatta di una normale (platicurtica), se $\gamma > 1$ la curva sarà più appuntita di una normale (leptocurtica).

4.1.2 Raggruppamento per classi di valori

Possiamo ripetere quanto detto prima raggruppando i dati in sottointervalli. Ciascun insieme dellla suddivisione costituisce una **classe**.

Tutte le definizioni precedenti si possono estendere, soltanto che invece di contare quanti determinati valori si ripetono si conterà quanti elementei cadono all'interno di una determinata classe.

L'esempio in tabella è quello dei dati relativi al peso di 100 studenti.

Peso (kg)	f_i	p_i	F_i
60-62	5	0.05	0.05
63-65	18	0.18	0.23
66-68	42	0.42	0.65
69-71	27	0.27	0.92
72-74	8	0.08	1.0

Possiamo dare le seguenti definizioni:

- **Estremi**: I limiti inferiore e superiore di una classe. Esempio: nella prima classe il limite inferiore è 60, quello superiore 62.
- **Valore centrale**: Il punto medio di ciascun sottointervallo. Esempio: nella prima classe il valore centrale è 61.
- Confini: Se le classi non sono contigue si definisce confine tra due classi il valore medio tra l'estremo superiore di una classe e quello inferiore della classe successiva. Esempio: Il confine tra la prima e la seconda classe è 62.5.

All'interno di ogni classe di valori possiamo approssimare la funzione di ripartizione linearmente: nella prima classe la fdr sarà una retta che interpola tra 0 e F_1, nella seconda tra F_1 e F_2, e via dicendo fino all'ultima classe in cui interpolerà tra F_{n-1} e $F_n = 1$. Con questa definizione di funzione di ripartizione, la definizione di quantile data prima coincide con quella teoricper cui $F(q_\alpha) = \alpha$.

4.1.3 Rappresentazione dei dati

In costruzione

4.2 Statistica inferenziale

Occupiamoci ora della statistica inferenziale, l'insieme di metodologie atte a trarre conclusioni su una popolazione statistica partendo dall'analisi dei campioni.

Per prima cosa definiamo **campione** di rango n un insieme di n variabili aleatorie X_i che provengono dalla stessa distribuzione. Io conosco che tipo di distribuzione ma non conosco i parametri. Ad esempio, io so che la distribuzione delle medie dei voti degli studenti UniCT segue una Gaussiana ma non conosco né la media né la varianza di questa Gaussiana. Posso dedurle prendendo un capione e analizzando quello?

4.2.1 Stimatori

Per fare quanto illustrato sopra si utilizzano delle opportune funzioni delle variabili aleatorie del campione. Una funzione di variabili aleatorie è detta **statistica**. Immaginiamo di voler stimare un parametro θ, allora utilizzerò una statistica a valori nell'immagine di θ. L'ultima affermazione si può banalmente leggere come segue: non posso prendere una combinazione di v.a. che non assume mai i valori assunti dal parametro θ. Ad esempio, ho un set di lanci di una moneta e voglio stimare se la moneta è truccata, ovvero la probabilità di successo p. Se scelgo una statistica che ha valori in $[2, +\infty[$ non riuscirò mai a stimare p che ha valori solo in $[0, 1]$.

Definiamo **stimatore puntuale** di θ una statistica $\hat{\theta}$ a valori nell'immagine di θ. Ovviamente posso anche stimare funzioni di θ.

$\hat{\theta}$ si dirà **stimatore non distorto** di θ se $E[\hat{\theta}] = \theta$.

Facciamo degli esempi per chiarire il concetto.

Stimatore non distorto della media

Ho un campione di n v.a. provenienti tutte dalla stessa distribuzione con media μ che non conosco e voglio stimare dal mio campione. La media campionaria

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

è uno stimatore non distorto della media.

In realtà questa cosa l'abbiamo già dimostrata nell'interludio precedente la legge dei grandi numeri, ma rivediamola anche qui:

Dimostrazione:

$$E[\overline{X}_n] = E\Big[\frac{1}{n}\sum_{i=1}^{n} X_i\Big] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{n\mu}{n} = \mu.$$

Stimatore non distorto della varianza in caso di media nota

Ho un campione di n v.a. provenienti tutte dalla stessa distribuzione con media μ nota e varianza σ^2 che non conosco e voglio stimare dal mio campione. La varianza campionaria

$$\bar{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

è uno stimatore non distorto della varianza nel caso in cui la media sia nota.

Verifichiamo che è effettivametne non distorto

Dimostrazione:

$$E(\bar{\sigma}_n^2) = E\Big(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2\Big) = \frac{1}{n^2}\sum_{i=1}^{n} E(X_i^2) + \mu^2 - 2\mu E[X_i] = \frac{1}{n^2}\sum_{i=1}^{n} E(X_i^2) - \mu^2.$$

A questo sostituiamo l'identità $\mu^2 = \frac{1}{n}\sum_{i=1}^{n} E[X_i]^2$ e troviamo

$$E(\bar{\sigma}_n^2) = \frac{1}{n}\sum_{i=1}^{n} E(X_i^2) - \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\sum_{i=1}^{n}\big(E(X_i^2) - E[X_i]^2\big) = \frac{1}{n}\sum_{i=1}^{n}\sigma^2 = \sigma^2.$$

(nell'ultimo passaggio abbiamo usato la definizione di varianza, ricordandoci che ogni X_i ha varianza σ^2.)

Stimatore non distorto della varianza in caso di media non nota

Ho un campione di n v.a. provenienti tutte dalla stessa distribuzione con media μ e varianza σ^2 entrambe ignote da stimare dal mio campione. In questo caso come stimatore della media si usa la media campionaria, mentre la varianza si stima con

$$\bar{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

Notare: al denominatore abbiamo $n-1$ perché altrimenti lo stimatore sarebbe distorto. Dimostriamolo!

Dimostrazione:

Prima di dimostrare che lo stimatore non è distorto ci conviene riscrivere questo stiimatore in un modo più comodo

$$\bar{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}X_i^2 - \frac{2\overline{X}_n}{n-1}\sum_i X_i \cdot \frac{n}{n} + \frac{1}{n-1}\sum_{i=1}^{n}\overline{X}_n^2$$

Da cui possiamo riscrivere

$$\bar{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}X_i^2 - \frac{2n\overline{X}_n}{n-1}\cdot\frac{1}{n}\sum_i X_i + \frac{n}{n-1}\overline{X}_n^2.$$

Raccogliendo il tutto abbiamo

$$\bar{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}X_i^2 - \frac{2n\overline{X}_n^2}{n-1} + \frac{n}{n-1}\overline{X}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i)^2 - \frac{n\overline{X}_n^2}{n-1}.$$

Adesso osserviamo anche che, dalla definizione di varianza, abbiamo che:

$$E[X_i^2] = \mathrm{Var}(X_i) + \mu^2 = \sigma^2 + \mu^2, \quad E[X_n^2] = \mathrm{Var}(X_n) + \mu^2 = \frac{\sigma^2}{n} + \mu^2.$$

Notate che $\mathrm{Var}(X_n) = \frac{\sigma}{n}$ l'abbiamo dimostato nell'interludio che precede la legge dei grandi numeri.

A questo punto avremo

$$E[\bar{S}_n^2] = \frac{1}{n-1}\sum_i E[X_i^2] - \frac{n}{n-1}E[X_n^2] = \frac{1}{n-1}\sum_i(\sigma^2+\mu) - \frac{n}{n-1}\Big(\mu + \frac{\sigma}{n}\Big).$$

E raccogliendo

$$E[\bar{S}_n^2] = \frac{n\sigma^2 + n\mu}{n-1}\sum_i(\sigma^2+\mu) - \frac{n\mu+\sigma^2}{n-1} = \sigma^2.$$

Ma come li usiamo questi concetti? Banalmente prendiamo i dati del nostro campione e li usiamo per calcolare gli stimatori per trarre informazioni sulla nostra popolazione. Più è grande il campione tanto più accurato sarà il nostro stimatore!

4.2.2 Intervalli di confidenza

Possiamo anche non dare una stima puntuale dei parametri della distribuzione ma ammettere che la nostra stima sia corretta con un errore $(1-\alpha)\%$. In questo caso si parlerà di **intervalli di confidenza con livello di confidenza** α . $1-\alpha$ verrà detto **livello di significatività**.

In formule, questo significa che, dato un campion $X_1, X_2, \ldots, X_n$ e un $\alpha \in [0,1]$, definirò $I_X = [I_1, I_2]$ intervallo di confidenza di livello $1-\alpha$ del parametro θ se

$$P(\theta \in [I_1, I_2]) = P(I_1 \leq \theta \leq I_2) = 1-\alpha.$$

In quello che segue determineremo gli intervalli di confidenza per media e varianza. Per farlo utilizzeremo delle v.a. ausiliarie dalla distribuzione nota.

Intervallo di confidenza per la media nel caso di varianza σ^2 nota

La variabile ausiliaria di cui ci serviamo è la variabile normale standardizzata

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma}\sqrt{n}.$$

Se il campione è distribuito Gaussianamente questa variabile seguirà una normale standard, ovvero $Z_n \sim N(0,1)$. In alternativa, basta avere $n > 30$ per poter sfruttare il teorema del limite centrale e assumere comunque che $Z_n \sim N(0,1)$.

Dalla teoria sappiamo che, per la definizione di quantile

$$P(\phi_{\frac{\alpha}{2}} \leq Z_n \leq \phi_{1-\frac{\alpha}{2}}) = 1-\alpha,$$

dove ϕ_α indica il quantile di ordine α della normale standard. Sappiamo anche che i quantili della normale standard sono simmetrici, per cui $\phi_\alpha = -\phi_{1-\alpha}$. Sostituendo troviamo che

$$P\left(-\phi_{1-\frac{\alpha}{2}} \leq \frac{\overline{X}_n - \mu}{\sigma}\sqrt{n} \leq \phi_{1-\frac{\alpha}{2}}\right) = 1-\alpha.$$

Da questo possiamo isolare la μ

$$P\left(\overline{X}_n - \frac{\sigma}{\sqrt{n}}\phi_{1-\frac{\alpha}{2}} \leq \mu \leq \overline{X}_n + \frac{\sigma}{\sqrt{n}}\phi_{1-\frac{\alpha}{2}}\right) = 1-\alpha.$$

A questo punto è facile identificare $I_1 = \overline{X}_n - \frac{\sigma}{\sqrt{n}}\phi_{1-\frac{\alpha}{2}}$ e $I_2 = \overline{X}_n + \frac{\sigma}{\sqrt{n}}\phi_{1-\frac{\alpha}{2}}$, pertanto il nostro intervallo di confidenza sarà

$$I_\mu = \left[\overline{X}_n - \frac{\sigma}{\sqrt{n}}\phi_{1-\frac{\alpha}{2}}, \overline{X}_n + \frac{\sigma}{\sqrt{n}}\phi_{1-\frac{\alpha}{2}}\right].$$

Osservazione. Quest'intervallo è simmetrico rispetto alla media campionaria! Chiamando $\varepsilon_\mu = \frac{\sigma}{\sqrt{n}}\phi_{1-\frac{\alpha}{2}}$ possiamo riscriverlo

$$I_\mu = [\overline{X}_n - \varepsilon_\mu, \overline{X}_n + \varepsilon_\mu].$$

Notiamo inoltre che la larghezza dell'intervallo dipende proporzionalmente dalla varianza (ovvero maggiore è la variabilità dei dati maggiore sarà la mia incertezza sul parametro) e in maniera

inversamente proporzionale dall'ampiezza del campion (ovvero più grande è il campione migliore è la mia stima).

Intervallo di confidenza per la media nel caso di varianza non nota

In questo caso la varianza del problema non è nota, per cui dobbiamo stimarla usando lo stimatore introdotto poco fa. Dato che anche la media non è nota è necessario utilizzare, come stimatore, $\bar{S}_n^2$.

In tal caso la nosta variabile d'ausilio sarà

$$T_n = \frac{\overline{X}_n - \mu}{\bar{S}_n}\sqrt{n}.$$

Si può fare vedere che, se il campione è gaussiano o è abbastanza numeroso, allora $T_n \sim t(n-1)$, ovvero segue una t di Student a $n-1$ gradi di libertà.

In tal caso avremo

$$P(-t_{1-\frac{\alpha}{2}} \leq T_n \leq t_{1-\frac{\alpha}{2}}) = 1-\alpha,$$

dove t_α indica il quantile di ordine α della t di Student. Sappiamo anche che i quantili della t di Student sono simmetrici, per cui $t_\alpha = -t_{1-\alpha}$.

Si può ripetere quindi passo passo il discorso fatto prima per arrivare a identificare l'intervallo di confidenza per la media

$$I_\mu^{\bar{S}_n} = \left[\overline{X}_n - \frac{\bar{S}_n}{\sqrt{n}}t_{1-\frac{\alpha}{2}}, \overline{X}_n + \frac{\bar{S}_n}{\sqrt{n}}t_{1-\frac{\alpha}{2}}\right].$$

Osservazione. Anche quest'intervallo è simmetrico rispetto alla media campionaria! Chiamando $\varepsilon_\mu^{\bar{S}_n} = \frac{\bar{S}_n}{\sqrt{n}}t_{1-\frac{\alpha}{2}}$ possiamo riscriverlo

$$I_\mu^{\bar{S}_n} = [\overline{X}_n - \varepsilon_\mu^{\overline{S}_n}, \overline{X}_n + \varepsilon_\mu^{\bar{S}_n}].$$

Intervallo di confidenza per la varianza σ^2.

Dato che la varianza non è nota, la stimiamo con $\bar{S}_n^2$.

Per introdurre la variabile ausiliaria dobbiamo richiedere che il campione sia strettamente gaussiano. Infatti, sotto tale ipotesi si può dimostrare che

$$W_n = \frac{\bar{S}_n^2}{\sigma^2}(n-1) \sim \chi^2(n-1),$$

ovvero la variabile W_n segue una χ^2 a $n-1$ gradi di libertà. Chiamiamo $\chi^2_\alpha(n-1)$ il relativo quantile. Dato che la distribuzione χ^2 non è simmetrica avremo $\chi^2_\alpha(n-1) \neq -\chi^2_{1-\alpha}(n-1)$.

Pertanto avremo

$$P(\chi^2_{\frac{\alpha}{2}} \leq W_n \leq \chi^2_{1-\frac{\alpha}{2}}) = 1-\alpha$$

Di conseguenza, procedendo come prima, avremo

$$I_{\sigma^2} = \left[\frac{\bar{S}_n^2(n-1)}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}, \frac{\bar{S}_n^2(n-1)}{\chi^2_{\frac{\alpha}{2}}(n-1)}\right].$$

Intervallo di confidenza sulla proporzione

Immaginiamo di avere un campione di ampiezza n con $X_i \sim B(1,p)$ e di voler stimare p. Sappiamo dalla teoria che $\mu = p$ e $\sigma^2 = p(1-p)$. Se n è abbastanza grande (ovvero tale che $np > 5$ e $n(1-p) > 5$, allora il campione sarà approssimabile a un campione Gaussiano e la variabile

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma}\sqrt{n} = \frac{\overline{X}_n - p}{\sqrt{p(1-p)}}\sqrt{n} \sim N(0,1).$$

Invertendo questa relazione qui arriviamo a trovare

$$I_p = \left[\overline{X}_n - \frac{\sqrt{\overline{X}_n(1-\overline{X}_n)}}{\sqrt{n}}\phi_{1-\frac{\alpha}{2}}, \overline{X}_n + \frac{\sqrt{\overline{X}_n(1-\overline{X}_n)}}{\sqrt{n}}\phi_{1-\frac{\alpha}{2}}\right].$$

4.3 Test d'ipotesi

Immginiamo di voler testare un'ipotesi statistica sul parametroi θ che vogliamo stimare. Proviamo a cercare una procedura per respingere o avallare l'ipotesi fatta. Imamginiamo di avere due insiemi di possibili valori di Θ_0 e Θ_1. Per prima cosa definiamo:

- **Ipotesi nulla**: $H_0 : \theta \in \Theta_0$
- **Ipotesi alternativa**: $H_1 : \theta \in \Theta_1$.

Aggiungiamo inoltre che, se $\Theta_0 = \{\theta_0\}$ è un insieme con un solo elemento, l'ipotesi nulla diventa $H_0 : \theta = \theta_0$ e si parla di ipotesi nulla **semplice**. In quanto segue ci focalizzeremo solo su questo caso.

L'idea è quella di cercare un criterio per rigettare l'ipotesi nulla a favore di quella alternativa.

ATTENZIONE: H_0 e H_1 non hanno un ruolo simmetrico. Infatti, se trovo argomenti necessari per rigettare H_0, allora potrò affermare che H_1 è vera. Se invece non trovo argomenti per rigettare H_0, il test non è concludente e NON posso assolutamente affermare né che H_1 sia vera né che sia falsa.

ESEMPIO: per testare se una moneta è truccata faccio un'ipotesi statistica sulla probabilità p che esca testa e utilizzo come insiemi $\Theta_0 = \{\frac{1}{2}\}$ e $\Theta_1 = [0,1]/\{\frac{1}{2}\}$.

In pratica quello che si fa è la seguente cosa:

1. Si preleva un campione di dati
2. Si calcola uno stimatore $\hat{\theta}$ della quantità su cui si fa l'ipotesi
3. Si definisce una zona di rigetto C_R
4. Se $\theta \in C_R$ si rifiuta l'ipotesi nulla e si accetta quella alternativa. Altrimenti il test è inconcludente.

Il modo in cui viene determinata la zona di rigetto dipende da test a test, come vedremo tra poco.

Ovviamente ogni test d'ipotesi può essere soggetto a errori. Identifichiamo due errori differenti:

1. **Errore di prima specie**α: è la probabilità di rigettare l'ipotesi nulla a torto, ovvero

$$\alpha = P(\hat{\theta} \in C_R | \theta = \theta_0).$$

   Questo errore è anche detto **livello di significatività** del test.

2. **Errore di seconda specie**β: è la probabilità di non rigettare H_0 quando è falsa, ovvero

$$\beta = P(\hat{\theta} \notin C_R | \theta \in \Theta_1).$$

Si definisce **potenza del test** la quantità

$$\pi = \begin{cases} \alpha & \text{se} \quad \theta = \theta_0, \\ \beta & \text{se} \quad \theta \in \Theta_1. \end{cases}$$

In generale ci piacerebbe che il test avesse potenza massima (ovvero 1), quando l'ipotesi alternativa è vera. Per fare ciò dobbiamo richiedere che $\alpha, \beta = 0$. In quanto segue ci dimenticheremo dell'errore di seconda specie assumendo $\beta = 0$ e ci concentreremo soltanto su quello di prima specie. Non possiamo mai richiedere che quello di prima specie sia nullo, ma possiamo richiedere che sia molto piccolo. In generale si assume $\alpha = 0.05$ o $\alpha = 0.01$.

Si definisce π**-value** il più piccolo valore di α per cui si rigetta l'ipotesi nulla.

Passiamo alla determinazione delle zone di rigetto. Immaginiamo di voler stimare il parametro θ e sia $\hat{\theta}$ un suo stimatore. Scegliamo come ipotesi nulla

$$H_0 : \theta = \theta_0.$$

Possiamo identificare tre tipi di test:

1. **Test bilatero**: $H_1 : \theta \neq \theta_0$. Fissato α, rigettiamo l'ipotesi nulla se $\hat{\theta} > \theta_0 + \delta$ o $\hat{\theta} < \theta_0 - \delta$, ovvero se $|\hat{\theta} - \theta_0| > \delta$. Scelgo δ in modo tale che

$$P(|\hat{\theta} - \theta_0| > \delta) = \alpha.$$

2. **Test unilatero a sinistra**: $H_1 : \theta < \theta_0$ Fissato α, rigettiamo l'ipotesi nulla se $\hat{\theta} < \theta_0 - \delta$. Scelgo δ in modo tale che

$$P(\hat{\theta} < \theta_0 - \delta) = \alpha.$$

3. **Test unilatero a destra**: $H_1 : \theta > \theta_0$ Fissato α, rigettiamo l'ipotesi nulla se $\hat{\theta} > \theta_0 + \delta$. Scelgo δ in modo tale che

$$P(\hat{\theta} > \theta_0 + \delta) = \alpha.$$

Per identificare nel dettaglio δ e quindi le zone di rigetto di ciascun test dobbiamo di nuovo, come fatto per gli intervalli di confidenza, servirci di una variabile ausialiaria di cui conosciamo la distribuzione.

Nel seguito vedremo tre possibili test:

1. Z-test: basati su una variabile ausiliaria $Z_n \sim N(1, 0)$;
2. T-test: basati su una variabile ausiliaria $T_n \sim t(n-1)$;

3. test del χ^2: basati su una variabile ausiliaria $W_n \sim \chi^2(n-1)$.

Illustriamo il caso del calolo della zona di rigetto nel caso di uno Z test. Per farlo utilizziamo un caso specifico, ovvero:

Test d'ipotesi sulla media con varianza nota

In questo caso il parametro da stimare è la media della distribuzione μ. Come stimatore, ovviamente utilizziamo la media campionaria $\overline{X}_n$. Come fatto in precedenza sappiamo che la variabile $Z_n = \frac{\overline{X}_n - \mu}{\sigma}\sqrt{n} \sim N(0,1)$. Utilizziamola per calcolare le zone di rigetto con livello di significatività α. L'ipotesi nulla è $H_0 : \mu = \mu_0$.

Caso bilatero

L'ipotesi alternativa è $H_1 : \mu \neq \mu_0$. La zona di rigetto che vogliamo calcolare è definita da

$$P(|\overline{X}_n - \mu| > \delta) = \alpha.$$

Questa probabilità è identica alla seguente probabilità

$$P\left(\left|\frac{\overline{X}_n - \mu}{\sigma}\sqrt{n}\right| > \frac{\delta\sqrt{n}}{\sigma}\right) = \alpha.$$

O, in altri termini

$$P\left(|Z_n| > \frac{\delta\sqrt{n}}{\sigma}\right) = \alpha.$$

Notiamo, tuttavia, che, dato che $Z_n \sim N(0,1)$, per definizione di quantile, non può che essere $\phi_{1-\frac{\alpha}{2}} = \frac{\delta\sqrt{n}}{\sigma}$. Pertanto la condizione all'interno della probabilità si può riscrivere per ottenere

$$P\left(|Z_n| > \phi_{1-\frac{\alpha}{2}}\right) = \alpha.$$

Questo porta alle due zone di rigetto:

$$Z_n > \phi_{1-\frac{\alpha}{2}} \Rightarrow \overline{X}_n > \mu + \frac{\sigma}{\sqrt{n}}\phi_{1-\frac{\alpha}{2}}$$

e

$$Z_n < -\phi_{1-\frac{\alpha}{2}} \Rightarrow \overline{X}_n < \mu - \frac{\sigma}{\sqrt{n}}\phi_{1-\frac{\alpha}{2}}.$$

Caso unilatero a sinistra

L'ipotesi alternativa è $H_1 : \mu < \mu_0$. Ragionando in maniera analoga a quanto fatto prima avremo che la zona di rigetto è

$$Z_n < \phi_\alpha \Rightarrow \overline{X}_n < \mu + \frac{\sigma}{\sqrt{n}}\phi_\alpha.$$

Caso unilatero a destra

L'ipotesi alternativa è $H_1 : \mu > \mu_0$. La zona di rigetto sarà

$$Z_n < \phi_\alpha \Rightarrow \overline{X}_n < \mu + \frac{\sigma}{\sqrt{n}}\phi_\alpha.$$

Per calcolare il p-value basta osservare che il minimo alpha necessario per rigettare l'ipotesi nulla si può trovare, per definizione, calcolando la funizone di ripartizione sulla variabile. In tal modo otteniamo:

1. bilatero: p-value $= 2(1 - F_Z(|Z_n|))$
2. unilatero a sx: p-value $= F_Z(Z_n)$
3. unilatero a dx: p-value $= 1 - F_Z(Z_n)$.

La teoria dei test d'ipotesi si può riassumere in quanto segue.

Z-test- Z_n

Zona rigetto:

$$C_R : \begin{cases} |Z_n| > \phi_{1-\frac{\alpha}{2}} \text{ bilatero} \\ Z_n < \phi_\alpha \text{ unilatero a sx} \\ Z_n > \phi_{1-\alpha} \text{ unilatero a dx} \end{cases}$$

p-value:

$$\text{p-value} = \begin{cases} 2(1 - \Phi(|Z_n|) \text{ bilatero} \\ \Phi(Z_n) \text{ unilatero a sx} \\ 1 - \Phi(Z_n) \text{ unilatero a dx} \end{cases}$$

- **Media con varianza nota**: vogliamo testare se la media di una distribuzione di varianza nota σ^2 è uguale a una media nota μ_0. L'ipotesi nulla è $H_0 : \mu = \mu_0$. La variabile ausiliaria è:
$$Z_n = \frac{\overline{X}_n - \mu_0}{\sigma}\sqrt{n}.$$
- **Proporzione**: vogliamo stimare la probabilità p di una distribuzione di Bernulli è verificare che sia diversa da una data p_0. L'ipotesi nulla è $H_0 : p = p_0$. La variabile ausiliaria è:
$$Z_n = \frac{\overline{X}_n - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n}.$$
- **Media di coppie di poplazioni**: abbiamo due popolazioni X, Y indipendenti che possiamo assumere gaussiani con varianza σ_X^2 e σ_Y^2 nota e vogliamo verificare se le loro medie sono μ_X e μ_Y sono uguali. l'ipotesi nulla è $H_0 : \mu_X = \mu_Y$. La variabile ausiliaria è:
$$Z_n = \frac{\overline{X}_n - \mu_X - (\overline{Y}_m - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}.$$

---

t-test- T_n

Zona rigetto:

$$C_R : \begin{cases} |T_n| > t_{1-\frac{\alpha}{2}} \text{ bilatero} \\ T_n < t_\alpha \text{ unilatero a sx} \\ T_n > t_{1-\alpha} \text{ unilatero a dx} \end{cases}$$

p-value:

$$\text{p-value} = \begin{cases} 2(1 - \Phi(|T_n|) \text{ bilatero} \\ \Phi(T_n) \text{ unilatero a sx} \\ 1 - \Phi(T_n) \text{ unilatero a dx} \end{cases}$$

- **Media con varianza non nota**: vogliamo testare se la media di una distribuzione di varianza non nota è uguale a una media nota μ_0. L'ipotesi nulla è $H_0 : \mu = \mu_0$. La variabile ausiliaria è:
$$T_n = \frac{\overline{X}_n - \mu_0}{\bar{S}_n}\sqrt{n}.$$

- **Media di popolazioni accoppiate**: abbiamo due popolazioni accoppiate (sono della forma (X_i, Y_i) – es: velocità e tempo) X, Y di media μ_X e μ_Y e varianze σ_X^2 e σ_Y^2. Definite le differenze $D_i = X_i - Y_i$, assumiamo che $D_i \sim N(\mu_D, \sigma_D^2)$, con $\mu_D = \mu_X - \mu_Y$ e σ_D^2 non nota. Stimiamo σ_D^2 con la varianza campionaria $\bar{S}_n$. L'ipotesi nulla è: $H_0 : \mu_D = 0$. La variabile ausialiaria è:

$$T_n = \frac{\overline{D}_n}{\bar{S}_n}\sqrt{n}.$$

---

test del χ^2- W_n

Zona rigetto:

$$C_R : \begin{cases} W_n > \chi^2_{1-\frac{\alpha}{2}}(n-1) \text{ o } W_n < \chi^2_{\frac{\alpha}{2}}(n-1) \text{ bilatero} \\ W_n < \chi^2_{\alpha}(n-1) \text{ unilatero a sx} \\ W_n > \chi^2_{1-\alpha}(n-1) \text{ unilatero a dx} \end{cases}$$

- **Varianza**: vogliamo testare se la varianza di una distribuzione σ^2 è uguale a una media nota σ_0^2. L'ipotesi nulla è $H_0 : \sigma = \sigma_0$. La variabile ausiliaria è:

$$W_n = \frac{\bar{S}_n(n-1)}{\sigma_0^2}.$$

- **Multinomiale m-dimensionale (Pearson)**: ho un campione di legge multinomiale $X = (X_1, X_2, \ldots, X_n)$ che può assumere i valori $\{x_1, x_2, \ldots, x_m\}$ con probabilità $p = \{p_1, p_2, \ldots, p_m\}$. Vogliamo testare se il vettore di probabilità è uguale a un vettore noto $p_0 = \{p_1^0, p_2^0, \ldots, p_m^0\}$. L'ipotesi nulla è $H_0 = p = p_0$. La variabile ausiliaria è:

$$W_n = n\sum_{k=1}^{m} \frac{(\bar{p}_k - p_k^0)^2}{p_k^0} \sim \chi^2(m-1).$$

4.4 Regressione Lineare

Immaginiamo adesso di voler determinare se c'è una relazione tra due o più variabili. Spesso alcune grandezze camminano in coppia. Per cui sarà abbastanza ragionevole aspettarsi una relazione tra il peso di una persona e il numero di calorie che assume quotidianamente. Probabilmente sarà meno ragionevole assumere una relazione tra il numero di uova mangiate nella vita e il voto preso a Metodi. In generale le grandezze con rapporti di causalità avranno anche una relazione di qualche tipo. Vi ricordo ancora una volta che **correlazione non implica causalità**. Infatti, esistono quelle che si chiamano **correlazioni spurie**, ovvero degli andamenti correlati tra variabili che non hanno nessun legame di causa-effetto. Per alcuni esempi guardate qui: https://www.tylervigen.com/spurious-correlations

In generale, possiamo quantificare *a priori* la presenza di correlazioni calcolando il **coefficiente di correlazione lineare** definito come

$$\rho = \frac{\sum_{i=1}^n x_i y_i - \bar{x}\bar{y}}{\sqrt{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i^2}} = \frac{\bar{\sigma}_{xy}}{\bar{\sigma}_x \bar{\sigma}_y}.$$

Questo coefficiente è sempre $-1 \leq \rho \leq 1$. Tanto più il coefficiente di correlazione lineare è vicino a zero, tanto meno le variabili sono correlate. Se $\rho < 0$ allora le variabili saranno correlate antilinearmente, se $\rho > 0$ saranno correlate linearmente. Detto questo identifichiamo tre gradi di correlazione:

- Correlazione **debole** se $0 < |\rho| \leq 0.3$;
- Correlazione **moderata** se $0.3 < |\rho| \leq 0.7$;
- Correlazione **forte** se $0.7 < |\rho| \leq 1$.

Detto questo torniamo al problema del come trovare la relazione tra due variabili. Di questo si occupa **l'analisi di regressione**. In quanto segue ci occuperemo solo di regressione lineare tra due sole variabili e trascureremo il caso multidimensionale.

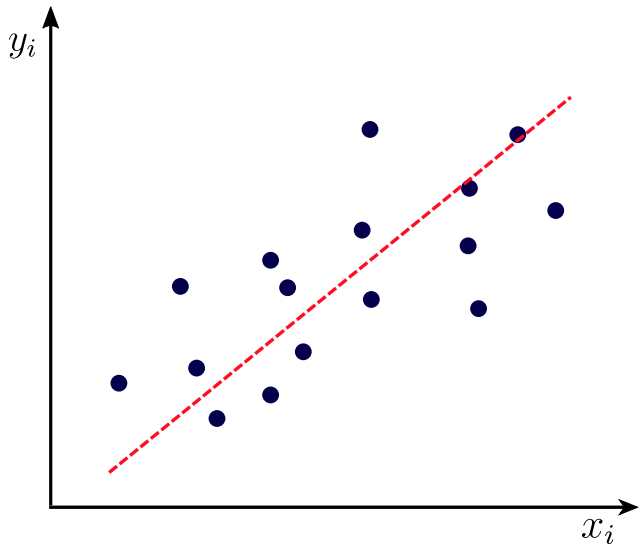

Figura 4

Immaginiamo di avere dei n dati come quelli in Fig. 4, ovvero per ogni valore x_i dell'asse delle ascisse abbiamo un valore y_i dell'asse delle ordinate associato. Guardando l'andamento dei punti si intuisce che c'è una relazione di qualche tipo, dato che all'aumentare di x_i aumenta y_i. Per cui, si può bene di sovrapporre una retta immaginaria che descriva bene i dati, tipo la retta trattegiata in rosso. Ma come scegliere la retta giusta tra le infinite rette? In formule: stiamo cercando un modo sensato di determinare β_0 e β_1, tali che la retta

$$y = \beta_0 + \beta_1 x$$

sia quella che meglio si adatti ai miei dati (x_i, y_i). Questo problema è anche noto come problema di **best fit**.

Questa retta non passerà per tutte le coppie di punti, ma passerà per la media dei miei dati. In altre parole, descrive l'andamento in media dei dati.

Il metodo utilizzato per determinare i coefficienti è quello **metodo dei minimi quadrati**. Concettualmente si vuole minimizzare la somma delle distanza da ogni punto da quello predetto dal modello. Per ogni x_i, il mio modello lineare prevede un $\hat{y}_i = \beta_0 + \beta_1 x_i$. Cosideriamo la differenza tra il punto predetto e quello reale, ovvero $d_i = y_i - \hat{y}_i$. Costruiamo la funzione

$$S^2 = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2,$$

che sostanzilmente quantifica la distanza tra i dati e quelli ottenuti dal mio modello.

S^2 è una funzione di β_0 e β_1, ovvero cambia valore cambiando i parametri del modello. In formule $S^2 \equiv S^2(\beta_1, \beta_2)$. Vogliamo trovare i parametri che rendono minima S^2, ovvero minimizzano la distanza del modello dai dati reali. Ricordando Analisi 2, sappiamo che per fare questo dobbiamo per prima cosa annullare le derivate parziali rispetto a β_0 e β_1. Annulliamo la prima, ovvero

$$\frac{\partial S^2}{\partial \beta_0} = 2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0.$$

Sviluppando i conti troviamo

$$\begin{aligned}
&\frac{\partial S^2}{\partial \beta_0} = 0 \iff \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0 \iff \\
&\sum_{i=1}^{n} y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i = 0 \iff \\
&n\bar{y} - n\beta_0 - n\beta_1 \bar{x} = 0,
\end{aligned}$$

dove abbiamo posto $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ e $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. Da cui segue

$$\beta_0 = \bar{y} - \beta_1 \bar{x}.$$

Facciamo qualcosa di analogo per l'altra variabile

$$\frac{\partial S^2}{\partial \beta_1} = 2\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) = 0.$$

Sviluppando i conti troviamo

$$\begin{aligned}
&\frac{\partial S^2}{\partial \beta_1} = 0 \iff \sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) = 0 \iff \\
&\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \beta_0 - \beta_1 \sum_{i=1}^{n} x_i^2 = 0 \iff \\
&\sum_{i=1}^{n} x_i y_i - n\bar{x}\beta_0 - \beta_1 \sum_{i=1}^{n} x_i^2 = 0 \iff \\
&\sum_{i=1}^{n} x_i y_i - n\bar{x}(\bar{y} - \beta_1 \bar{x}) - \beta_1 \sum_{i=1}^{n} x_i^2 = 0 \iff \\
&\sum_{i=1}^{n}(x_i y_i - \bar{x}\bar{y}) - \beta_1(\sum_{i=1}^{n} x_i^2 - \bar{x}^2) = 0 \iff \\
&n\bar{\sigma}_{xy} - n\beta_1 \bar{\sigma}_x^2 = 0.
\end{aligned}$$

dove abbiamo posto $\bar{\sigma}_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}$ e $\bar{\sigma}_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$. Da qui segue

$$\beta_1 = \frac{\bar{\sigma}_{xy}}{\bar{\sigma}_x^2}.$$

Possiamo sostituire nel risultato precedente e trovare:

$$\beta_0 = \bar{y} - \frac{\bar{\sigma}_{xy}}{\bar{\sigma}_x^2}\bar{x} = 0.$$

Come vi ricorderete, annullare le derivate ci garantisce solo di trovare un punto di estremo della nostra funzione. Per essere sicuri che sia un minimo dobbiamo prima valutare il segno della matrice Hessiana. Ometteremo questa dimostrazione, e ci fideremo ciecamente che i due punti trovati siano di minimo.

Ricapitolando, da questi conti abbiamo trovato che la retta che meglio approssima i nostri dati è

$$y = \left(\bar{y} - \frac{\bar{\sigma}_{xy}}{\bar{\sigma}_x^2}\bar{x}\right) + \frac{\bar{\sigma}_{xy}}{\bar{\sigma}_x^2}x$$

Per assicurarci che il nostro modello di regressione lineare è adeguato possiamo calcolare il **coefficiente di determinazione**

$$R = \frac{\bar{\sigma}_{xy}^2}{\bar{\sigma}_x^2\bar{\sigma}_y^2} = \rho^2.$$

Tanto più il coefficiente di determinazione è vicino a uno tanto migliore sarà il mio modello.

Osservazione: il punto $(\bar{x}, \bar{y})$ sta sulla retta di regressione!

Interpretazione: come si deve interpretare questo risultato? Io ho trovato un modello che descrive i dati: lo posso usare per fare predizioni! Ad esempio, una volta trovata la retta di regressione di un campione di dati (x_i, y_i) potrò predire la y associata a ogni x, anche se non presente nel mio campione (vedete esercizi).

Osservazione 2: anche se abbiamo visto soltanto analisi lineare, quest'analisi permette di fittare dipendenze anche più variegate. Ad esempio: se i miei dati sono correlati e mostrano un andamento parabolico (si adagiano in media su una parabola), allora un modello credibile per descrivere questi dati sarebbe $y = \beta_0 + \beta_1 x^2$, che non è lineare. Tuttavia, ponendo $z = x^2$ e il nostro modello si riduce a $y = \beta_0 + \beta_1 z$ che è esattamente il modello che abbiamo studiato fin'ora!

5 Generazione di numeri pseudorandom e Metodi Monte Carlo

In questa sezione ci occuperemo della generazione di algoritmi di generazione di numeri pseudorandom e faremo degli esempi di algoritmi Monte Carlo.

5.1 Generazione di numeri pseudorandom

Per prima cosa chiediamoci: perché parliamo di generazione di numeri pseudorandom?

L'unica possibilità per generare numeri random è quella di avere un generatore che evolve secondo un processo puramente stocastico. Ad esempio, un generatore di numeri random tra 0 e 1 potrebbe essere un omino che estrae da un contenitore infinite palline numerate da 0 a 9: ciascuna pallina sarà una delle infinite cifre decimanli del nostro numero random. Comodo, no?

In alternativa possiamo pensare di generare dei numeri che non siano puramente casuali, generati da un algoritmo deterministico che però, si comportino da numeri casuali per un tempo abbastanza lungo. Il tempo abbastanza lungo è detto **periodo** ed è essenzialmente il tempo che ci vuole per ritornare al primo elemento della sequenza generato (e quindi a ripetere di nuovo la sequenza per sempre). Nonostante questo inconveniente, gli algoritmi pseudocasuali sono preferibili sia per la loro "leggerezza" ma anche per la loro velocità.

Concentriamoci su numeri pseudorandom tra 0 e 1. Uno dei metodi più diffusi è il **metodo delle congruenze lineari** . L'idea è semplice: dato un numero X_0 puramente casuale, detto **seed**, (ad esempio generato dal clock di sistema) possiamo costruire una sequenza di numeri pseudorandom compresi tra 0 e 1 usando la sequente espressione ricorsiva:

$$X_{n+1} = (aX_n + b)\text{mod } m.$$

In questa formula chiameremo a **modulo**, b **incremento**, e m **modulo**.

Facciamo un esempio. Partiamo da $X_0 = 1$ e poniamo $a = 3$, $b = 1$, $m = 5$. Allora avremo:

$$\begin{aligned} X_1 &= (3X_0 + 1)\text{mod } 5 = 4 \\ X_2 &= (3X_1 + 1)\text{mod } 5 = 3 \\ X_3 &= (3X_2 + 1)\text{mod } 5 = 0 \\ X_4 &= (3X_3 + 1)\text{mod } 5 = 1 \\ X_5 &= (3X_4 + 1)\text{mod } 5 = 4, \end{aligned}$$

ovvero dopo 4 iterazioni siamo al punto di partenza. Questo generatore ha un periodo $T = 4$, quindi forse non un ottimo generatore.

Giusto per darvi un'idea dell'importanza dell'avere un periodo lungo, molti degli articoli scientifici scritti utilizzando i primissimi generatori di numeri random sono sotto esame perché potrebbero avere degli errori generati dalla non perfetta stocasticità!

Il modo migliore per allungare il periodo di questo generatore è quello di scegliere opportunamente a, b, m.

Ad oggi il generatore di numeri casuali migliore che abbiamo è quello di Mersenne Twister che fornisce un periodo di $2^{19937} - 1$.

Importante: una volta trovato un generatore di numeri random tra 0 e 1 possiamo generare il mondo.

Per esempio possiamo generare numeri distribuiti secondo una distribuzione **binomiale**. La prescrizione è semplice. Per prima cosa generiamo delle variabili $X_i \sim B(1,p)$. Per generare ciascuna X_i, generiamo prima un numero random $\xi \in [0,1]$ e poi poniamo

$$X_i = \begin{cases} 1 & \text{se } \xi \in [0,p] \\ 0 & \text{se } \xi \in [p,1]. \end{cases}$$

Fatto questo avremo $X = \sum_{i=1}^{n} X_i \sim B(n,p)$.

Possiamo ragionare analogamente per generare una v.a. $Y \sim B(n, p_1, \ldots, p_m)$ distribuita secondo una **multinomiale**. Come prima, generiamo prima $Y_i \sim B(1, p_1, \ldots, p_n)$. Per farlo, dividiamo l'intervallo in m sottointervalli di ampiezza proporzionale alla probabilità di accadimento di quell'evento. Possiamo definire $F_k = \sum_{i=1}^{k} p_i$ e porre:

$$i = \begin{cases} 1 & \text{if } \xi \in [0, F_1] \\ 2 & \text{se } \xi \in [F_1, F_2] \\ \vdots & \\ m & \text{se } \xi \in [F_{m-1}, 1]. \end{cases}$$

Fatto questo avremo che la v.a.

$$Y_j = \begin{cases} 1 & \text{se } j = i \\ 0 & \text{se altrimenti,} \end{cases}$$

è tale che $Y_i \sim B(1, p_1, \ldots, p_m)$.

Generate n v.a. distribuite secondo una multinomiale avremo che $Y = \sum_{i=1}^{n} Y_i \sim B(n, p_1, \ldots, p_m)$.

Per generare ulteriori distribuzioni si possono utilizzare alcuni metodi che vedremo a breve.

5.1.1 Metodo di inversione della funzione di ripartizione

Questo metodo si basa sul seguente

Teorema: Se X è una v.a. di densità $f(x)$ con funzione di ripartizione $F_X(t)$ strettamente crescente, allora $F_X(t) \sim U([0,1])$. (dimostrabile ma ometteremo la dimostrazione).

Possiamo usare questo teorema per dimostrare (lo ometteremo) la seguente

Proprietà: Sia $X \sim U([0,1])$ e sia F una funzione invertibile. Allora la v.x. $Y = F^{-1}(X)$ ha funzione di ripartizione F.

Come si applica in pratica? Voglio generare dei numeri random distribuiti secondo $f(x)$ con f.d.r $F(t)$ (invertibile). Allora calcolo $F^{-1}(t)$ e la applico a $X \sim U([0,1])$ e ottengo Y distribuito secondo $f(x)$.

Esempio 1: generazione di numeri distribuiti uniformemente tra $[a,b]$, ovvero $Y \sim U([a,b])$. Tale distribuzione ha, in $[a,b]$, f.d.r $F_Y(t) = \frac{t-a}{b-a}$. Chiamiamo $x = \frac{t-a}{b-a}$. L'inversa della funzione di ripartizione è quella funzione che applicata a y mi restituisce t, ovvero $F^{-1}(x) = a + (b-a)x$.

Pertanto se genero $X \sim U([0,1])$, in accordo con la proprietà avrò che $Y = a + (b-a)X \sim U([a,b])$.

Esempio 2: generazione di numeri distribuiti secondo una distribuzione esponenziale.

Sappiamo che se $Y \sim \text{Exp}(\lambda)$, allora $F_Y(t) = 1 - e^{-\lambda t}$. Ragionando come troviamo che $F^{-1}(x) = -\frac{\log(1-x)}{\lambda}$.

Pertanto, genero $X \sim U([0,1])$ e allora avrò $Y = -\frac{\log(1-X)}{\lambda} = -\frac{\log(X)}{\lambda} \sim \text{Exp}(\lambda)$. Nell'ultimo passaggio ho usato il fatto che se $X \sim U([0,1])$ allora $1 - X \sim U([0,1])$.

Esempio 3: generazione di numeri distribuiti secondo una normale standard.

Questa è un'applicazione indiretta del metodo illustrato sopra, infatti non sappiamo scrivere la f.d.r. di una distribuzione normale. Tuttavia possiamo sfruttare alcune proprietà delle distribuzioni normali. Ad esempio sappiamo che se $X, Y \sim N(0,1)$ allora $Z^2 = X^2 + Y^2 \sim \chi^2(2)$. Fortunatamente noi sappiamo che $\chi^2(2) = \text{Exp}(\frac{1}{2})$. Li sappiamo generare! Siamo quindi in grado di generare $Z^2 = -2\log(\Lambda)$ con $\Lambda \sim U([0,1])$.

Una volta generato Z^2 posso generare X e Y osservando che, data la loro relazione, giacciono su un cerchio di raggio Z. Mi serve quindi generare un angolo θ distribuito random tra $[0, 2\pi]$ e per farlo genero $\Gamma \sim U([0,1])$ e porre

$$X = \sqrt{-2\log(\Lambda)}\cos(2\pi\Gamma) \quad \text{e} \quad Y = \sqrt{-2\log(\Lambda)}\sin(2\pi\Gamma).$$

Inoltre, dato $X \sim N(0,1)$ potrò generare

- $W = \mu + \sigma X \sim N(\mu, \sigma^2)$
- $W = X^2 \sim \chi^2(1)$

5.1.2 Metodo del rigetto

Immaginiamo di voler generare dei numeri secondo una distirbuzione $f(x) : [ab] \to \mathbb{R}$ continua. Si può dimostrare che questo obiettivo si può ragigungere implementando il seguente algoritmo:

(a) Genero una coppia (ξ, η) con $\xi \sim U([a,b])$ e $\eta \sim U([0,M])$ con $M = \max_{[a,b]} f(x)$

(b) Se $0 \leq \eta \leq f(\xi)$ accetto ξ e pongo $X = \xi$. Altrimenti ricomincio dal punto precedente.

Si può dimostrare che $P(X \leq t) = \int_{-\infty}^{t} f(x)dx$.

5.2 Metodi Monte Carlo

I metodi Monte Carlo sono dei metodi che utilizzano variabili aleatorie per risolvere problemi di matematica che si basano sulla legge dei grandi numeri e sul teorema del limite centrale. I metodi montecarlo si basano su quattro passaggi:

1. Definisco un dominio di input;
2. Genero numeri random nel dominio;
3. Eseguo un calcolo deterministico;
4. Aggrego i risultati.

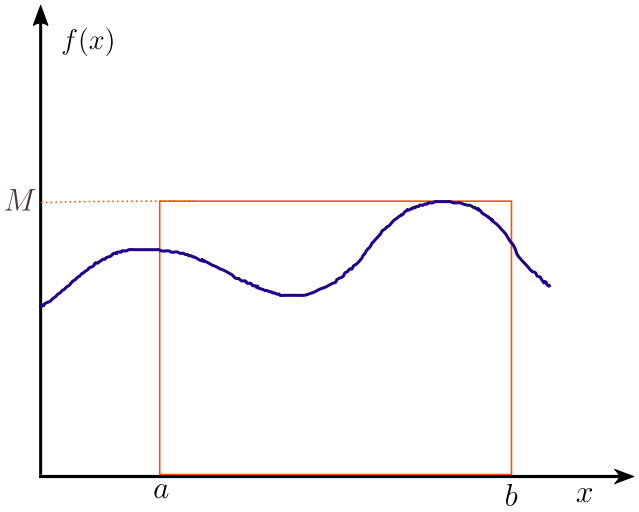

Figura 5

5.2.1 Metodo Hit or Miss

Il metodo Hit or Miss è un metodo Monte Carlo che serve per valutare l'integrale di funzioni continue positive. Immaginiamo di avere una qualsiasi funzione $f(x) : \mathbb{R} \to \mathbb{R}$ continua e tale che $f(x) \geq 0 \quad \forall x \in [a, b]$. Come esempio possiamo guardare quella in Fig. 5.Il nostro obiettivo è quello di calcolare un integrale definito del tipo $I = \int_a^b dx f(x)$.

Per fare questo genero per prima cosa un rettangolo (arancione in ifgura) $R = [a, b] \times [0, M]$, dove $M = \max_{[a,b]} f(x)$. Osservo poi che, pescati due numeri a caso nel rettangolo la probabilità che questi stiano sotto la curva della funzione è $p = \frac{I}{(b-a)M}$, ovvero il rapporto tra le aree. Un modo per stimare questa probabilità è quella di ripetere tante volte l'esperimento di estrarre una coppia di numeri nel rettangolo. Siano N le coppie estratte e N_s quelle "andate a segno". Allora la probaiblità sarà approssimabile con $p \sim \frac{N_s}{N}$.Tanto più è grande N, tanto più è accurata la stima.

Una volta che abbiamo una stima della p possiamo calcolarci il nostro integrale, infatit

$$\frac{N_s}{N} = \frac{I}{(b-a)M} \Rightarrow I = \frac{N_s}{N}(b-a)M.$$

6 Catene di Markov

Le catene di Markov sono un esempio di **processo stocastico**. Un procesos stocastico è una famiglia di v.a. che dipendono da un parametro t che vive in un insieme T. In generale un processo stocastico si indica con

$$\{X(t)\ t \in T\} \equiv (X_t)_{t \in T}.$$

Questo parametro può essere visto come un indice temporale e riflettere l'evoluzione del processo stocastico nel tempo (es: fluttuazioni dei titoli in finanza, passi fatti dall'ubriaco). Se T è un insieme continuo, allora il processo si dirà a **tempi continui**. Se T è un insieme numeravile si dirà a **tempi discreti**. In quanto segue ci concentreremo solo su procesis a tempi discrtei.

Sia X_t un processo stocastico a tempi discreti. I valori che X_t può assumere appartengono a un insieme E detto **insieme degli stati**. L'insieme E può essere infinito, ovvero $E = \mathbb{N}$, o finito, ovvero $E = \{1, 2, \ldots, n\}$. Per semplicità, in quanto segue ci concentreremo solo sul caso di catene di Markov finite.

Un processo stocastco è detto **catena di Markov** se gode della proprietà di assenza di memoria (o proprietà di Markov):

$$P(X_{t+1} = j | X_t = i, X_{t-1} = i_{t-1}, \ldots, X_1 = i_1) = P(X_{t+1} = j | X_t = i) = p_{ij}(t).$$

Data questa probabilità posso costruire la **matrice di transizione** $P(t)$ che sarà una matrice il cui elemento $i, j-$esimo contiene la probabilità di passare, all'istante di tempo t, dallo stato i allo stato j, ovvero $(P(t))_{ij} = p_{ij}(t)$. Gli elementi della matrice di transizione hanno le seguenti proprietà:

1. $p_{ij}(t) \geq 0 \quad \forall i, j \in E$;
2. $\sum_{j \in E} p_{ij}(t) = 1 \quad \forall i \in E$.

Una catena di Markov si dice **omogenea** se la sua matrice di transizione non dipende dall'istante di tempo t, ossia se ognuno degli elementi è tale che $p_{ij}(t) \equiv p_{ij}$. Cosa significa? Che la probabilità di saltare da uno stato all'altro sarà identica a ogni istante di tempo.

Da ora in poi assumiamo che le catene di Markov siano omogenee. Chiediamoci: posso conoscere la probabilità di saltare da i a j in m passi? Ovvero posso conoscere

$$p_{ij}^{(m)} = P(X_{t+m} = j | X_t = i)?$$

Senza entrare nei dettagli, questa risposta ce la dà il **Teorema di Chapman - Kolmogorov** che ci dice che la matrice di transizione a m passi $P^{(m)}$ si può ricavare dalla matrice di transizione semplicemente facendo

$$P^{(m)} = P^m,$$

dove il prodotto sul lato destro dell'uguaglianza è da intendersi come prodotto riga per colonna.

Diamo un'altra definizion. Ovvero P si dirà **regolare** se $\exists\ m$ t.c. $p_{ij}^{(m)} > 0 \quad \forall\ i, j \in E$. Esiste un criterio di regolarità che afferma che se E è finito e la catena è irriducibile condizione sufficiente affinché P sia regolare è che $\exists\ h \in E$ t.c. $p_{hh} > 0$.

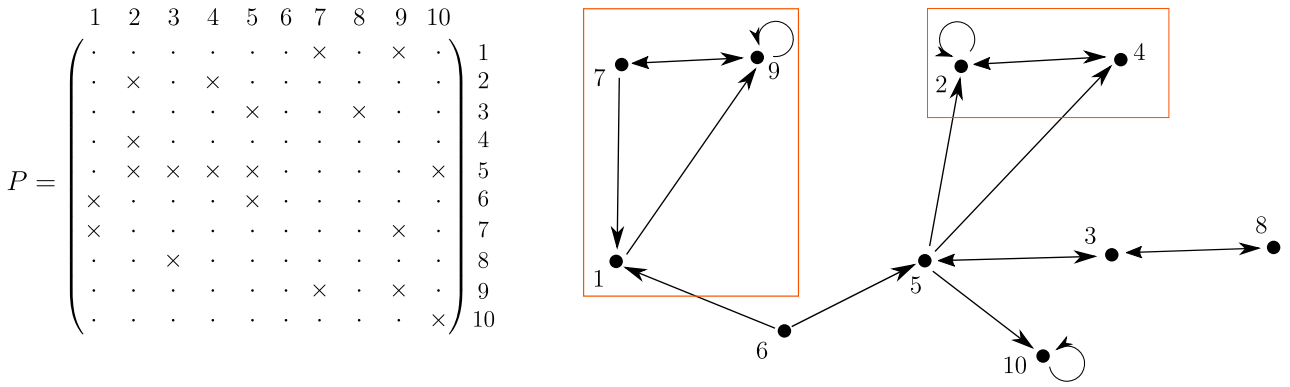

Figura 6: La matrice di transizione P è semplificata così: puntino non ho collegamento tra gli stati, $\times$ ho collegamento tra gli stati. Al momento non ci interessa qual è il peso dell'arco, ci interessa sapere solo se comunicano o no. In questo esempio abbiamo che: a) $\{1, 7, 9\}$ è una classe irriducibile; b) $\{2, 4\}$ è una classe irriducibile; c) $\{3, 5, 6, 8\}$ sono transitori; d) $\{10\}$ è assorbente.

6.1 Classificazione degli stati

È importante dare una classificazione degli stati. In quanto segue daremo un po' di definizioni.

- Si dice che i **comunica** con j se $\exists\ m > 0$ t.c. $p_{ij}^{(m)} > 0$.. In tal caso si indica con $i \longrightarrow j$. Se questo non accade i e j non comunicano e si indica con $i \not\longrightarrow j$.

  Ci sono due osservazioni:

  - Non c'è simmetria: se $i \longrightarrow j$ non significa che $j \longrightarrow i$.
  - Vale la proprietà transitiva: se $i \longrightarrow h$ e $h \longrightarrow j$, allora $i \longrightarrow j$.

- Un insieme $C \subset E$ si dice **classe chiusa** se gli stati di C non comunicano con quelli di E/C.
- Una classe chiusa è **irriducibile** se tutti i suoi stati comunicano
- Una catena di Markov si dice irriducibile se tutti i suoi stati comunicano
- Se una classe chiusa è costituita da uno stato solo, lo stato si dirà **assorbente**
- Uno stato si dice **ricorrente** se ho la certezza di ritornare, ad un certo tempo t, nello stato stesso
- Uno stato si dice **transitorio** se non è ricorrente.

  Se una catena è finita, condizione necessaria e sufficiente affinché uno stato i sia transitorio è che $i \longrightarrow j$ ma $j \not\longrightarrow i$.

  Se una catena di Markov è irriducibile gli stati sono o tutti transitori o tutti ricorrenti. Se la catena di Markov è finita sono tutti ricorrenti.

Per maggiori chiarimenti si può guardare l'esempio in Fig. 6.

6.2 Distribuzioni stazionarie

Ad ogni passo t della nostra catena definiamo il vettore $w = (w_1, w_2, \ldots, w_n)$ legge di X_t, se $w_k = P(X_t = k) \ \forall \ k \in E$. In altre parole, la legge di X è il vettore che contiene la probabilità che la mia v.a. abbia valore k.

Immagiano che all'istante iniziale $t = 0$ la mia catena abbia distribuzione $v = (v_1, v_2, \ldots, v_n)$ con $v_k = P(X_0 = k) \ \forall \ k \in E$. Si può dimostrare allora che $w = vP^t$.

Una distribuzione v si dice **stazionaria** se è tale che

$$v = vP,$$

ossia se rimane invariata quando si applica la matrice di transizione. Se siamo attenti noteremo che una distribuzione stazionaria è un autovettore sinistro con autovalore 1 della matrice di transizione. Parte del nostro interesse è quello di riuscire a trovare le distribuzioni stazionarie. La prima domanda che ci facciamo è: siamo sicuri che esista? Il **Teorema di Markov-Kakutani** ci viene incontro dicendoci che: se E è finito, allora esiste almeno una distribuzione stazionaria.

Per trovarla ovviamente possiamo sempre risolvere il problema agli autovalori e calcolare v come autovettore sinistro di P. Ma non sempre questa è cosa agevole, pertanto si pososno sfruttare alcuni trucchi.

Intanto definiamo π distribuzione **reversibile** se $\pi_i p_{ij} = \pi_j p_{ji} \ \ \forall \ i, j \in E$. Vale quindi la seguente

Proprietà: se una distribuzione è reversibile, allora è stazionaria.

E infine questo ci permette di sfruttare il

Teorema di Markov: se P è regolare e E è finito allora esiste una sola distribuzione stazionaria

$$\pi_j = \lim_{t \to \infty} p_{ij}(t) \ \ \forall \ j \in E.$$