Calcolo delle Probabilità e Statistica

Luca De Paulis

15 ottobre 2020

INDICE

1 STATISTICA DESCRITTIVA

1.1 CONCETTI BASE

La statistica descrittiva è la branca della statistica che descrive fenomeni statistici senza sfruttare nozioni di probabilità. I concetti fondamentali della statistica descrittiva sono il concetto di *popolazione* e di *campione*: la popolazione è l'insieme delle entità e dei dati che vogliamo studiare, mentre il campione è un piccolo sottoinsieme della popolazione che verrà analizzato per fini statistici.

Altri concetti base sono il concetto di *frequenza assoluta* e *relativa*: si dice frequenza assoluta di un evento A il numero di volte che l'evento accade, senza considerare il numero di eventi (anche di tipo diverso) che accadono; invece si dice frequenza assoluta di un evento A il numero di volte che l'evento accade diviso il numero di eventi totali.

1.2 ANALISI NUMERICA DEI DATI

Supponiamo di avere un vettore $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ che rappresenta i nostri dati. Possiamo definire alcune operazioni fondamentali su questi dati.

Definizione 1.2.1 **Media (empirica.)** Dato $\boldsymbol{x}$ vettore di dati, si dice *media (empirica)* il valore

$$\bar{\boldsymbol{x}} := \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + \cdots + x_n}{n}. \tag{1}$$

Per descrivere quanto i dati contenuti in $\boldsymbol{x}$ si discostano dalla media $\bar{\boldsymbol{x}}$ si usa il concetto di varianza:

Definizione 1.2.2 **Varianza.** Dato $\boldsymbol{x}$ vettore di dati, si dice *varianza campionaria* il valore

$$\mathrm{var}(\boldsymbol{x}) := \sum_{i=1}^{n} \frac{(x_i - \bar{\boldsymbol{x}})^2}{n-1}; \tag{2}$$

si dice invece *varianza empirica* il valore

$$\mathrm{var}(\boldsymbol{x}) := \sum_{i=1}^{n} \frac{(x_i - \bar{\boldsymbol{x}})^2}{n}. \tag{3}$$

La varianza campionaria verrà usata quando i dati si riferiscono ad un campione, mentre la varianza empirica sarà più utile per trattare dati riferiti alle popolazioni.

In alcuni casi è utile conoscere la radice quadrata della varianza, quindi definiamo lo *scarto quadratico medio* (o *deviazione standard*) nel seguente modo:

$$\sigma(\boldsymbol{x}) := \sqrt{\mathrm{var}(\boldsymbol{x})}, \quad \sigma(\boldsymbol{x}) := \sqrt{\mathrm{var}(\boldsymbol{x})}. \tag{4}$$

Proposizione 1.2.3 *Vale la seguente uguaglianza:*

$$\sum_{i=1}^{n} (x_i - \bar{\boldsymbol{x}})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{\boldsymbol{x}}^2. \tag{5}$$

Dimostrazione.

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\bar{x})^2 &= \sum_{i=1}^{n}(x_i^2-2x_i\bar{x}+\bar{x}^2)\\ &= \sum_{i=1}^{n}x_i^2-2\bar{x}\sum_{i=1}^{n}x_i+\sum_{i=1}^{n}\bar{x}^2\\ &= \sum_{i=1}^{n}x_i^2-2\bar{x}\sum_{i=1}^{n}x_i+n\bar{x}^2\end{aligned}$$

Ricordando che $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$, ovvero $\sum_{i=1}^{n}x_i=2\bar{x}$:

$$\begin{aligned}&= \sum_{i=1}^{n}x_i^2-2n\bar{x}^2+n\bar{x}^2\\ &= \sum_{i=1}^{n}x_i^2-n\bar{x}^2.\end{aligned}$$

□

Usando la (5) otteniamo che

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

Proposizione 1.2.4 *Dato x vettore dei dati, la varianza (campionaria o empirica) di x è 0 se e solo se*

$$x_1=\cdots=x_n=\bar{x}.$$

La dimostrazione è ovvia: essendo la varianza definita come la somma di termini non negativi, essa è uguale a 0 se e soltanto se ogni termine è uguale a 0, ovvero se e solo se $x_i=\bar{x}$ per ogni $i=1,\ldots,n$. La varianza quindi rappresenta la "dispersione" dei dati: più è alta, più i dati sono diversi tra loro; più è bassa e più sono vicini (nel caso limite in cui sono tutti uguali la varianza è 0).

Possiamo generalizzare questa idea: fissata una soglia $d\in\mathbb{R}$ consideriamo il numero di elementi x_i la cui distanza dalla media $\bar{x}$ è maggiore o uguale a d:

$$\#\{\,x_1\;:\;|x_i-\bar{x}|\geqslant d\,\}.$$

È possibile dimostrare che vale la seguente disuguaglianza:

$$\#\{\,x_1\;:\;|x_i-\bar{x}|\geqslant d\,\}\leqslant\frac{1}{d}\sum_{i=1}^{n}(x_i-\bar{x})^2,$$

da cui, dividendo entrambi i membri per n, segue un caso particolare della cosiddetta *disuguaglianza di Chebyshev*:

$$\frac{\#\{\,x_1\;:\;|x_i-\bar{x}|\geqslant d\,\}}{n}\leqslant\frac{\mathrm{var}(x)}{d}. \tag{6}$$

Il membro sinistro rappresenta la *percentuale* dei dati che si discostano dalla media per un valore superiore alla soglia d.

Un altro metodo utile per ripartire i dati è utilizzare la cosiddetta *funzione di ripartizione empirica*.

Definizione 1.2.5 Dato $x\in\mathbb{R}^n$ vettore dei dati, la funzione di ripartizione empirica è una funzione $F_e:\mathbb{R}\to[0,1]$ tale che

$$F_e(t)=\frac{\#\{\,x_i\;:\;x_i\leqslant t\,\}}{n}.$$

Dunque per ogni soglia t la funzione di ripartizione empirica restituisce la percentuale dei dati che sono minori o uguali a t.

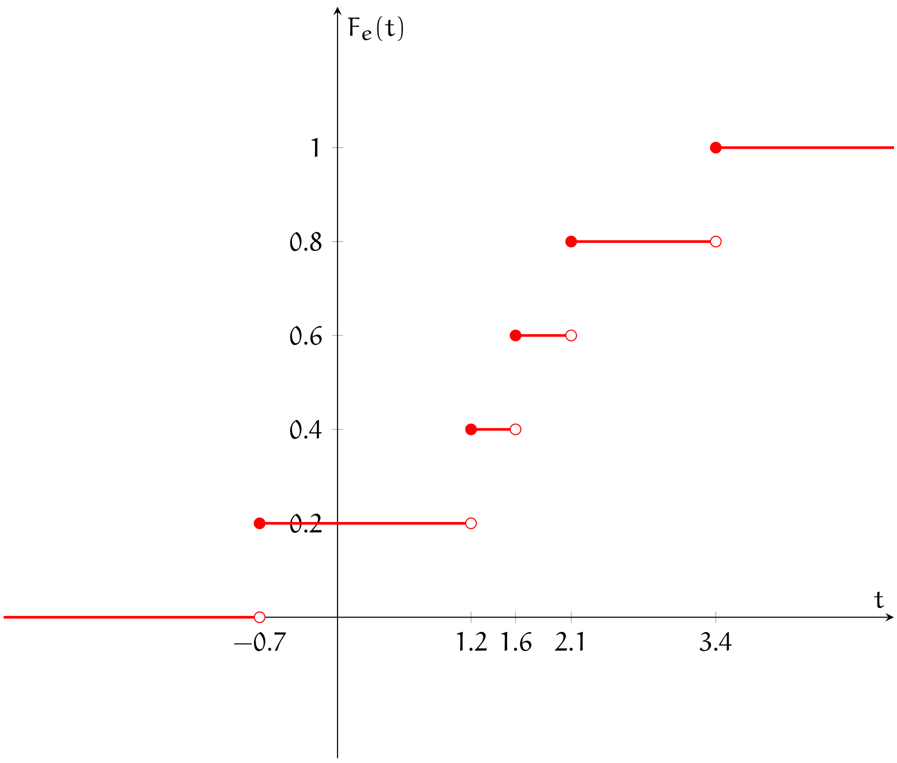
Esempio 1.2.6. Se il vettore dei dati è $\boldsymbol{x} = (1.2, -0.7, 3.4, 1.6, 2.1)$ per trovare la funzione di ripartizione empirica F_e mi conviene innanzitutto ordinarli, ottenendo

$$\boldsymbol{x} = (-0.7, 1.2, 1.6, 2.1, 3.4).$$

A questo punto posso descrivere molto semplicemente la funzione di ripartizione empirica:

- se $t < -0.7$ allora $F_e(t) = 0$: tutti i dati sono maggiori della soglia t;
- se $t \in [-0.7, 1.2)$ allora $F_e(t) = 1/5$: un solo dato è sicuramente minore o uguale a t (ovvero $x_1 = -0.7$), da cui dividendo per $n = 5$ si ottiene $1/5$;
- se $t \in [1.2, 1.6)$ allora $F_e(t) = 2/5$: due dati sono minori o uguali a t (ovvero -0.7 e 1.2), da cui dividendo per $n = 5$ si ottiene $2/5$;
- se $t \in [1.6, 2.1)$ allora $F_e(t) = 3/5$;
- se $t \in [2.1, 3.4)$ allora $F_e(t) = 4/5$;
- se $t \geqslant 3.4$ allora $F_e(t) = 1$ (tutti i dati sono minori o uguali a t, dunque la percentuale è 1);

Il grafico di questa funzione è quindi:



Percentili e quantili

Definizione 1.2.7 **Percentile.** Sia $k \in \mathbb{R}$ con $0 \leqslant k \leqslant 100$. Allora, dato un vettore dei dati $\boldsymbol{x}$ il k-esimo percentile è un qualsiasi numero $t \in \mathbb{R}$ tale che

- almeno $k/100$ dei dati sono minori o uguali di t,
- almeno $1-k/100$ dei dati sono maggiori o uguali a t.

Intuitivamente un numero reale t è il k-esimo percentile del nostro vettore di dati $\boldsymbol{x}$ se t è il più piccolo numero che è maggiore o uguale al k percento dei dati. Dato che preferiamo trattare numeri compresi tra 0 e 1 invece che tra 0 e 100 introduciamo il concetto di β-quantile: se t è un k-esimo percentile, allora t è un β-quantile per $\beta = k/100$.

Detto più direttamente, un numero t è un β-quantile se

- almeno β dei dati sono minori o uguali a t,
- almeno $1 - \beta$ dei dati sono maggiori o uguali a t.

Esempio 1.2.8. Dato il vettore $\boldsymbol{x} = (10, 20, 40, 60, 100)$, il dato $x_4 = 60$ corrisponde all'80-esimo percentile, o equivalentemente allo 0.80-quantile.

Alcuni quantili particolari hanno dei nomi specifici:

- lo 0.25-quantile è anche chiamato *primo quartile*,
- lo 0.50-quantile è anche chiamato *mediana*,
- lo 0.75-quantile è anche chiamato *terzo quartile*.

Dati multipli

In alcuni casi è necessario fare indagini statistiche su dati multipli: rappresentiamo i nostri dati come un vettore di coppie (o triple, o n-uple) di dati:

$$(\boldsymbol{x}, \boldsymbol{y}) = ((x_1, y_1), \dots, (x_n, y_n)).$$

Per studiare la *correlazione* tra i dati delle x e i dati delle y abbiamo bisogno di alcuni strumenti:

Definizione 1.2.9 **Covarianza.** Dato un vettore di coppie di dati $(\boldsymbol{x}, \boldsymbol{y})$ si dice *covarianza campionaria* il numero

$$\operatorname{cov}(\boldsymbol{x}, \boldsymbol{y}) := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{\boldsymbol{x}})(y_i - \bar{\boldsymbol{y}}); \tag{7}$$

si dice invece *convarianza empirica* il numero

$$\operatorname{cov}(\boldsymbol{x}, \boldsymbol{y}) := \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{\boldsymbol{x}})(y_i - \bar{\boldsymbol{y}}). \tag{8}$$

Definizione 1.2.10 **Coefficiente di correlazione.** Dato un vettore di coppie di dati $(\boldsymbol{x}, \boldsymbol{y})$, se $\sigma(\boldsymbol{x}), \sigma(\boldsymbol{y}) \neq 0$, si dice *coefficiente di correlazione* il numero

$$r(\boldsymbol{x}, \boldsymbol{y}) := \frac{\operatorname{cov}(\boldsymbol{x}, \boldsymbol{y})}{\sigma(\boldsymbol{x})\sigma(\boldsymbol{y})} = \frac{\sum_{i=1}^{n} (x_i - \bar{\boldsymbol{x}})(y_i - \bar{\boldsymbol{y}})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{\boldsymbol{x}})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{\boldsymbol{y}})^2}}.$$

Proposizione 1.2.11 *Dato un vettore di coppie di dati $(\boldsymbol{x}, \boldsymbol{y})$ con $\sigma(\boldsymbol{x}), \sigma(\boldsymbol{y}) \neq 0$, vale che*

$$0 \leqslant \left| r(\boldsymbol{x}, \boldsymbol{y}) \right| \leqslant 1.$$

Dimostrazione. Viene dalla disuguaglianza di Cauchy-Schwartz:

$$\sum_{i=1}^{n} \left| (x_i - \bar{\boldsymbol{x}})(y_i - \bar{\boldsymbol{y}}) \right| \leqslant \sqrt{\sum_{i=1}^{n} (x_i - \bar{\boldsymbol{x}})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{\boldsymbol{y}})^2},$$

da cui segue che

$$\left|r(\boldsymbol{x},\boldsymbol{y})\right| = \frac{\sum_{i=1}^{n}\left|(x_i-\bar{\boldsymbol{x}})(y_i-\bar{\boldsymbol{y}})\right|}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{\boldsymbol{x}})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{\boldsymbol{y}})^2}} \leqslant 1. \qquad \square$$

Intuitivamente il coefficiente di correlazione misura quanto è semplice approssimare la relazione tra le x e le y con una funzione lineare affine, ovvero con una retta: per vedere ciò cerchiamo di capire quale retta approssima meglio i nostri dati.

Per approssimare linearmente $(\boldsymbol{x},\boldsymbol{y})$ dobbiamo fare in modo che la nostra retta $a+bx$ sia il più vicino possibile ai punti (x_i,y_i) che formano i dati: vogliamo quindi che per ogni punto x_i la distanza tra y_i e $a+bx_i$ sia la minima possibile. Possiamo ottenere quello che vogliamo calcolando il seguente valore:

$$\min_{a,b\in\mathbb{R}^2}\sum_{i=1}^{n}(y_i-a-bx_i)^2. \tag{9}$$

(Eleviamo le distanze al quadrato in modo da renderle tutte positive e le sommiamo insieme poiché vogliamo che la distanza *complessiva* della retta sia minima.)

Teorema 1.2.12 *Il valore minimo della quantità in (9) si ottiene scegliendo*

$$b^* = \frac{\operatorname{cov}(\boldsymbol{x},\boldsymbol{y})}{\operatorname{var}(\boldsymbol{x})}, \quad a^* = -b\bar{\boldsymbol{x}}+\bar{\boldsymbol{y}}.$$

Inoltre vale che

$$\min_{a,b\in\mathbb{R}^2}\sum_{i=1}^{n}(y_i-a-bx_i)^2 = \sum_{i=1}^{n}(y_i-\bar{\boldsymbol{y}})^2\left(1-r(\boldsymbol{x},\boldsymbol{y})^2\right).$$

"Dimostrazione". Sia $Q:\mathbb{R}^2\to\mathbb{R}$ la funzione definita da

$$Q(a,b) \coloneqq \sum_{i=1}^{n}(y_i-a-bx_i)^2.$$

Per dimostrare che questa funzione ha minimo calcoliamo i limiti all'infinito: siccome vale che

$$\lim_{|a|,|b|\to+\infty}\sum_{i=1}^{n}(y_i-a-bx_i)^2 = +\infty$$

per il teorema di Weierstrass generalizzato questa funzone ha minimo. Per calcolarlo, imponiamo che le derivate parziali $\frac{\partial Q}{\partial a}$ e $\frac{\partial Q}{\partial b}$ siano uguali a 0, da cui ricaviamo le espressioni per a^* e b^*. Sostituendole in Q otteniamo l'espressione per il minimo di Q, che è la seconda parte della tesi. $\square$

La retta a^*+b^*x viene detta *retta di regressione* ed è la funzione lineare affine che meglio approssima i dati che abbiamo a nostra disposizione. Dato che la minima distanza tra la retta e il vettore dei dati (nel senso dato dalla formulazione in (9)) è proporzionale a $\left(1-r(\boldsymbol{x},\boldsymbol{y})^2\right)$, avremo che:

- più $r(\boldsymbol{x},\boldsymbol{y})^2$ si avvicina a 1, più la distanza minima si avvicina a 0 e quindi i dati sono correlati linearmente;

- più $r(\boldsymbol{x}, \boldsymbol{y})^2$ si avvicina a 0, più la distanza minima cresce e quindi i dati sono dispersi e non seguono una correlazione lineare.

Inoltre il coefficiente angolare della retta di regressione b^* può essere riscritto come

$$b^* = \frac{\operatorname{cov}(\boldsymbol{x}, \boldsymbol{y})}{\operatorname{var}(\boldsymbol{x})} = \frac{\operatorname{cov}(\boldsymbol{x}, \boldsymbol{y})}{\sigma(\boldsymbol{x})\sigma(\boldsymbol{y})} \frac{\sigma(\boldsymbol{x})\sigma(\boldsymbol{y})}{\operatorname{var}(\boldsymbol{x})} = r(\boldsymbol{x}, \boldsymbol{y}) \frac{\sigma(\boldsymbol{x})\sigma(\boldsymbol{y})}{\operatorname{var}(\boldsymbol{x})}.$$

Dato che $\frac{\sigma(\boldsymbol{x})\sigma(\boldsymbol{y})}{\operatorname{var}(\boldsymbol{x})} \geqslant 0$ il segno del coefficiente angolare dipende solamente dal coefficiente di correlazione:

- se $r(\boldsymbol{x}, \boldsymbol{y}) > 0$ la retta di regressione ha coefficiente angolare positivo, dunque è crescente e al crescere delle x tendenzialmente crescono anche le y;
- se $r(\boldsymbol{x}, \boldsymbol{y}) < 0$ la retta di regressione ha coefficiente angolare negativo, dunque è decrescente e al crescere delle x tendenzialmente le y decrescono.

Il coefficiente di correlazione dunque ci dice quanto sono correlate le due quantità che stiamo esaminando (più è vicino ad 1 e più sono correlate) e se al crescere della prima cresce anche la seconda (se è di segno positivo), oppure al crescere della prima la seconda diminuisce (se è di segno negativo).

2 PROBABILITÀ

2.1 SPAZIO DI PROBABILITÀ

Definizione 2.1.1 **Algebra delle parti.** Sia Ω un insieme. Una famiglia $\mathcal{F}$ di sottoinsiemi di Ω si dice *algebra delle parti su* Ω se valgono le seguenti proprietà:

(i) $\varnothing, \Omega \in \mathcal{F}$;

(ii) se $A \in \mathcal{F}$ allora $A^C \in \mathcal{F}$;

(iii) se $A, B \in \mathcal{F}$ allora $A \cup B \in \mathcal{F}$, $A \cap B \in \mathcal{F}$.

Un'algebra di parti su Ω modella bene l'insieme dei possibili eventi:

(i) l'insieme vuoto è un evento, ed in particolare corrisponde all'evento "non accade nessuno degli eventi nel nostro universo";

(ii) l'insieme universo è un evento, ed in particolare corrisponde all'evento "accade una cosa qualsiasi nel nostro universo";

(iii) se A è un evento, allora A^C corrisponde all'evento "non accade A";

(iv) se A, B sono eventi, allora $A \cup B$ corrisponde all'evento "accade A oppure accade B";

(v) se A, B sono eventi, allora $A \cap B$ corrisponde all'evento "accadono sia A che B".

Definizione 2.1.2 **Probabilità.** Sia Ω un insieme e $\mathcal{F}$ un'algebra delle parti su Ω. Si dice *probabilità* una funzione

$$\mathbb{P} : \mathcal{F} \to [0, 1]$$

tale che

1. $\mathbb{P}(\Omega) = 1$,
2. $\mathbb{P}$ è *finitamente additiva*: per ogni $A, B \in \mathcal{F}$ disgiunti (ovvero $A \cap B = \varnothing$) vale che
$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B).$$

Proposizione 2.1.3 **Proprietà della probabilità.** *Sia Ω un insieme e $\mathcal{F}$ un'algebra delle parti su Ω. Allora la funzione probabilità $\mathbb{P}$ soddisfa le seguenti proprietà:*

1. $\mathbb{P}(\varnothing) = 0$;
2. $\mathbb{P}(A^C) = 1 - \mathbb{P}(A)$;
3. *se* $B \subseteq A$ *allora* $\mathbb{P}(A \setminus B) = \mathbb{P}(A) - \mathbb{P}(B)$;
4. *per ogni* $A, B \in \mathcal{F}$ *vale il* principio di inclusione-esclusione*:*
$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

Dimostrazione. Dimostriamo le quattro affermazioni separatamente:

1. Siccome

- $\varnothing \cap \varnothing = \varnothing$,
- $\varnothing \cup \varnothing = \varnothing$,
- la funzione probabilità è finitamente additiva

segue che

$$\mathbb{P}(\varnothing) = \mathbb{P}(\varnothing \cup \varnothing) = \mathbb{P}(\varnothing) + \mathbb{P}(\varnothing),$$

da cui, sottraendo $\mathbb{P}(\varnothing)$ ad entrambi i membri, otteniamo $\mathbb{P}(\varnothing) = 0$.

2. Per definizione di complementare segue che $\Omega = A \cup A^C$. Inoltre un insieme e il suo complementare sono sempre disgiunti, dunque vale l'additività finita:

$$1 = \mathbb{P}(\Omega) = \mathbb{P}(A \cup A^C) = \mathbb{P}(A) + \mathbb{P}(A^C)$$

da cui segue che $\mathbb{P}(A^C) = 1 - \mathbb{P}(A)$.

3. Siccome $B \subseteq A$ possiamo scrivere $A = B \cup (A \setminus B)$. Inoltre B e $A \setminus B$ sono ovviamente disgiunti, dunque vale l'additività finita:

$$\mathbb{P}(A) = \mathbb{P}\big(B \cup (A \setminus B)\big) = \mathbb{P}(B) + \mathbb{P}(A \setminus B)$$

da cui segue che $\mathbb{P}(A \setminus B) = \mathbb{P}(A) - \mathbb{P}(B)$.

4. Separiamo gli insiemi A e B in tre sottoinsiemi particolari:
   - $C_1 = A \setminus (A \cap B)$, ovvero C_1 contiene solo gli elementi che sono in A e non in B,
   - $C_2 = B \setminus (A \cap B)$, ovvero C_2 contiene solo gli elementi che sono in B e non in A,
   - $C_3 = A \cap B$, ovvero C_3 contiene tutti e soli gli elementi che appartengono sia ad A che a B.

   Notiamo che
   - $A \cup B = C_1 \cup C_2 \cup C_3$,
   - $A = C_1 \cup C_3$, $B = C_2 \cup C_3$,
   - i tre insiemi C_1, C_2, C_3 sono tutti e tre disgiunti.

   Dunque per additività finita:

$$\begin{aligned} \mathbb{P}(A \cup B) &= \mathbb{P}(C_1 \cup C_2 \cup C_3) \\ &= \mathbb{P}(C_1) + \mathbb{P}(C_2 \cup C_3) \end{aligned}$$

   Aggiungiamo e sottraiamo $\mathbb{P}(C_3)$:

$$\begin{aligned} &= \mathbb{P}(C_1) + \mathbb{P}(C_2 \cup C_3) + \mathbb{P}(C_3) - \mathbb{P}(C_3) \\ &= \mathbb{P}(C_1 \cup C_3) + \mathbb{P}(C_2 \cup C_3) - \mathbb{P}(C_3) \end{aligned}$$

   Infine, siccome $C_1 \cup C_3 = A$, $C_2 \cup C_3 = B$, $C_3 = A \cap B$:

$$= \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B). \qquad \square$$

Questa formalizzazione dei concetti di evento e probabilità è però limitata: possiamo calcolare la probabilità di unioni e intersezioni di un numero finito di eventi, ma non di un numero infinito (anche se numerabile). Abbiamo quindi bisogno di un'estensione di questi concetti che ci permetta di "andare al limite".

Definizione 2.1.4 **Sigma-algebra.** Sia Ω un insieme. Una famiglia $\mathcal{F}$ di sottoinsiemi di Ω si dice *σ-algebra su* Ω se valgono le seguenti proprietà:

(i) $\varnothing, \Omega \in \mathcal{F}$;

(ii) se $A \in \mathcal{F}$ allora $A^C \in \mathcal{F}$;

(iii) se $(A_n)_{n \in \mathbb{N}}$ è un insieme numerabile di elementi di $\mathcal{F}$ (ovvero per ogni $n \in \mathbb{N}$ vale che $A_n \in \mathcal{F}$), allora

$$\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}, \qquad \bigcap_{n=1}^{\infty} A_n \in \mathcal{F}.$$

Definizione 2.1.5 **Probabilità.** Sia Ω un insieme e $\mathcal{F}$ una σ-algebra su Ω. Si dice *probabilità* una funzione

$$\mathbb{P} : \mathcal{F} \to [0, 1]$$

tale che

(i) $\mathbb{P}(\Omega) = 1$,

(ii) $\mathbb{P}$ è *numerabilmente additiva*: data una successione $(A_n)_{n \in \mathbb{N}}$ a valori in $\mathcal{F}$ (ovvero con $A_i \in \mathcal{F}$ per ogni $i \in \mathbb{N}$) a due a due disgiunti (ovvero $A_i \cap A_j = \varnothing$ per ogni $i \neq j$) vale che

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(A_i).$$

OSSERVAZIONE. Questa nuova definizione di probabilità è un'*estensione* della definizione iniziale: infatti dalla additività numerabile segue necessariamente l'additività finita.

Dimostrazione. Siano $A, B \in \mathcal{F}$ disgiunti. Allora possiamo costruire la successione

$$C_n \coloneqq \begin{cases} A, & \text{se } n = 1, \\ B, & \text{se } n = 2, \\ \varnothing, & \text{altrimenti.} \end{cases}$$

Gli insiemi che formano questa successione sono a due a due disgiunti, in quanto A e B sono disgiunti e l'intersezione tra un insieme qualunque e l'insieme vuoto è sempre vuota.

Inoltre l'unione di tutti questi insiemi è uguale all'unione di A e B, da cui segue che

$$\begin{aligned} \mathbb{P}(A \cup B) &= \mathbb{P}\left(\bigcup_{n=1}^{\infty} C_n\right) && \text{(per addit. num.)} \\ &= \sum_{i=1}^{\infty} \mathbb{P}(C_n) \\ &= \mathbb{P}(C_1) + \mathbb{P}(C_2) + \sum_{i=3}^{\infty} \mathbb{P}(C_n) \end{aligned}$$

Siccome $C_n = \varnothing$ per ogni $n \geqslant 3$ e la probabilità dell'insieme vuoto è 0 la sommatoria vale 0 e dunque segue che

$$\begin{aligned} &= \mathbb{P}(C_1) + \mathbb{P}(C_2) \\ &= \mathbb{P}(A) + \mathbb{P}(B) \end{aligned}$$

cioè la funzione probabilità è finitamente additiva. □

OSSERVAZIONE. Dato che la probabilità estesa è finitamente additiva, continua a valere la Proposizione 2.1.3.

Definizione 2.1.6 **Spazio di probabilità.** Sia Ω un insieme, $\mathcal{F}$ una σ-algebra su Ω e $\mathbb{P}$ una funzione probabilità definita su $\mathcal{F}$: la tripla $(\Omega, \mathcal{F}, \mathbb{P})$ si dice *spazio di probabilità*.

Definizione 2.1.7 **Eventi trascurabili e quasi certi.** Sia $(\Omega, \mathcal{F}, \mathbb{P})$ uno spazio di probabilità. Un evento $A \in \mathcal{F}$ si dice

- *trascurabile* se $\mathbb{P}(A) = 0$;
- *quasi certo* se $\mathbb{P}(A) = 1$.

La definizione "estesa" di probabilità ci consente di "passare al limite", ovvero di calcolare in modo semplice la probabilità di un evento definito come unione o intersezione numerabile di eventi.

Proposizione 2.1.8 *Sia* $(\Omega, \mathcal{F}, \mathbb{P})$ *uno spazio di probabilità e sia*

$$A_1 \subseteq A_2 \subseteq A_3 \subseteq \dots$$

una catena di insiemi, con $A_i \in \mathcal{F}$ *per ogni* $i \in \mathbb{N}$. *Sia inoltre*

$$A \coloneqq \bigcup_{i=1}^{\infty} A_i.$$

Allora vale che

$$\mathbb{P}(A) = \lim_{n \to +\infty} \mathbb{P}(A_i). \tag{10}$$

Dimostrazione. Definiamo $B_1 = A_1$, $B_n = A_n \setminus A_{n-1}$. L'unione di queste due successioni di insiemi è la stessa:

$$\bigcup_{i=1}^{\infty} B_i = \bigcup_{i=1}^{\infty} A_i.$$

Dalla definizione di B_n e dalla Proposizione 2.1.3 (in particolare dal terzo punto, siccome $A_{n-1} \subseteq A_n$) segue che

$$\mathbb{P}(B_n) = \mathbb{P}(A_n \setminus A_{n-1}) = \mathbb{P}(A_n) - \mathbb{P}(A_{n-1}).$$

Inoltre i B_n sono a due a due disgiunti, dunque vale l'additività numerabile:

$$\begin{aligned}
\mathbb{P}(A) &= \mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) \\
&= \mathbb{P}\left(\bigcup_{i=1}^{\infty} B_i\right) \\
&= \sum_{i=1}^{\infty} \mathbb{P}(B_n) \\
&= \lim_{n \to +\infty} \sum_{i=1}^{n} \mathbb{P}(B_i) \\
&= \lim_{n \to +\infty} \mathbb{P}(B_1) + \mathbb{P}(B_2) + \dots \\
&\qquad + \mathbb{P}(B_{n-1}) + \mathbb{P}(B_n) \\
&= \lim_{n \to +\infty} \mathbb{P}(A_1) + (\mathbb{P}(A_2) - \mathbb{P}(A_1)) + \dots \\
&\qquad + (\mathbb{P}(A_{n-1}) - \mathbb{P}(A_{n-2})) \\
&\qquad + (\mathbb{P}(A_n) - \mathbb{P}(A_{n-1})) \\
&= \lim_{n \to +\infty} \mathbb{P}(A_n).
\end{aligned}$$

□

INSIEME FONDAMENTALE FINITO Nel caso in cui Ω sia un insieme finito possiamo scriverlo come

$$\Omega := \{ w_1, \dots, w_n \}$$

dove n è la cardinalità di Ω. Come σ-algebra su Ω possiamo sempre prendere l'insieme delle parti $\mathcal{P}(\Omega)$, ovvero l'insieme di tutti i sottoinsiemi di Ω: in questo modo tutti i sottoinsiemi di Ω sono eventi.

Chiamiamo p_i la probabilità che accada l'evento $\{ w_i \} \in \mathcal{P}(\Omega)$, ovvero

$$p_i := \mathbb{P}\big(\{ w_i \}\big).$$

Sicuramente $p_i \geqslant 0$ (in quanto la funzione probabilità restituisce numeri reali tra 0 e 1); inoltre vale che

$$\sum_{i=1}^{n} p_i = p_1 + p_2 + \cdots + p_n = 1.$$

Infatti:

$$\begin{aligned} p_1 + p_2 + \cdots + p_n &= \mathbb{P}\big(\{ w_1 \}\big) + \mathbb{P}\big(\{ w_2 \}\big) + \cdots + \mathbb{P}\big(\{ w_n \}\big) \\ &= \mathbb{P}\big(\{ w_1 \} \cup \{ w_2 \} \cup \ldots \cup \{ w_n \}\big) \\ &= \mathbb{P}\big(\{ w_1, w_2, \dots, w_n \}\big) \\ &= \mathbb{P}(\Omega) \\ &= 1. \end{aligned}$$

Se Ω è finito e tutti gli eventi sono circa equiprobabili si può prendere

$$p_i = \frac{1}{n}.$$

In questo caso si parla di *distribuzione uniforme di probabilità*. Dato un evento $A \in \mathcal{P}(\Omega)$, nel caso di distribuzione uniforme, si ha che

$$\mathbb{P}(A) = \frac{\#A}{\#\Omega} = \frac{\#\text{casi favorevoli}}{\#\text{casi totali}}.$$

2.2 PROBABILITÀ CONDIZIONATA

ESEMPIO 2.2.1. Immaginiamo di star giocando alla roulotte e sappiamo che è appena uscito un numero pari. Qual è la probabilità che questo numero sia 4?

Sia A l'evento "è uscito 4" e sia B l'evento "è uscito un numero pari". Siccome sappiamo che è uscito un numero pari, e i numeri pari alla roulotte sono 18, la probabilità che sia uscito proprio 4 (sapendo che il numero uscito è pari) equivale a $1/18$.

Possiamo tuttavia pensarla in questo modo: $1/18$ è la probabilità che gli eventi A e B accadano contemporaneamente (ovvero è la probabilità di $A \cap B$) diviso la probabilità che sia avvenuto l'evento B:

$$\frac{1}{18} = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{1/37}{18/37}.$$

Definizione 2.2.2 **Probabilità condizionata.** Sia $(\Omega, \mathcal{F}, \mathbb{P})$ uno spazio di probabilità e sia $B \in \mathcal{F}$ non trascurabile. Si dice *probabilità di* A *condizionata a* B la quantità

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

OSSERVAZIONE. La funzione $\mathbb{P}' : \mathcal{F} \to [0, 1]$ data da

$$\mathbb{P}'(A) := \mathbb{P}(A \mid B)$$

è una probabilità.

Proposizione 2.2.3 **Formula di Bayes.** *Siano A, B due eventi non trascurabili. Allora vale che*

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

Dimostrazione. Per definizione di probabilità condizionata

$$\mathbb{P}(A \mid B)\mathbb{P}(B) = \mathbb{P}(A \cap B) = \mathbb{P}(B \mid A)\mathbb{P}(A),$$

da cui la tesi. □

Proposizione 2.2.4 **Formula del condizionamento ripetuto.** *Siano $A_1, \dots, A_n$ eventi non trascurabili. Allora*

$$\mathbb{P}(A_1 \cap \ldots \cap A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2 \mid A_1) \cdots \mathbb{P}(A_n \mid A_{n-1} \cap \ldots \cap A_1).$$

Dimostrazione. Per definizione di probabilità condizionata

$$\begin{aligned}
&\mathbb{P}(A_1)\mathbb{P}(A_2 \mid A_1) \cdots \mathbb{P}(A_n \mid A_{n-1} \cap \ldots \cap A_1) \\
&= \mathbb{P}(A_1)\frac{\mathbb{P}(A_1 \cap A_2)}{\mathbb{P}(A_1)}\frac{\mathbb{P}(A_3 \cap A_2 \cap A_1)}{\mathbb{P}(A_2 \cap A_1)} \cdots \frac{\mathbb{P}(A_n \cap \ldots \cap A_1)}{\mathbb{P}(A_{n-1} \cap \ldots \cap A_1)} \\
&= \mathbb{P}(A_1 \cap \ldots \cap A_n).
\end{aligned}$$

□

Definizione 2.2.5 **Sistema di alternative.** Sia $(\Omega, \mathcal{F}, \mathbb{P}$ uno spazio di probabilità e siano $B_1, \dots, B_n \in \mathcal{F}$. Allora se $B_1, \dots, B_n$ formano una partizione di Ω (ovvero la loro unione è Ω e la loro intersezione a due a due è vuota) si dice che $B_1, \dots, B_n$ è un *sistema di alternative*.

OSSERVAZIONE. Un sistema di alternative non è necessariamente finito: data una famiglia numerabile $(B_i)_{i \in \mathbb{N}}$ di sottoinsiemi di Ω, essa rappresenta un sistema di alternative se è una partizione di Ω.

Proposizione 2.2.6 **Formula di fattorizzazione.** *Sia $B_1, \dots, B_n$ un sistema di alternative e $A \in \mathcal{F}$ un evento. Allora vale che*

$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A \mid B_i)\mathbb{P}(B_i).$$

Proposizione 2.2.7 **Formula delle probabilità delle cause.** *Sia $B_1, \dots, B_n$ un sistema di alternative e $A \in \mathcal{F}$ un evento. Allora per ogni $i \in \{1, \dots, n\}$ vale che*

$$\mathbb{P}(B_i \mid A) = \frac{\mathbb{P}(A \mid B_i)\mathbb{P}(B_i)}{\mathbb{P}(A)}.$$

2.3 INDIPENDENZA STOCASTICA

Vogliamo rendere l'idea che talvolta conoscere se un evento A si è verificato oppure no non influenza la probabilità che l'evento B si verifichi e viceversa.

Sfruttando le probabilità condizionate, possiamo esprimerlo con una di queste due formule:

$$(1): \ \mathbb{P}(B \mid A) = \mathbb{P}(A). \qquad (2): \ \mathbb{P}(A \mid B) = \mathbb{P}(A).$$

È facile dimostrare che queste due condizioni sono equivalenti; tuttavia nessuna delle due modella precisamente la nostra condizione, poiché non sono simmetriche e richiedono che A e B siano non trascurabili.

Definizione 2.3.1 **Indipendenza stocastica.** Siano $A, B \in \mathcal{F}$ due eventi. Allora A e B sono indipendenti stocasticamente se

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

Proposizione 2.3.2 **Conseguenze dell'indipendenza stocastica.** *Sia* $(\Omega, \mathcal{F}, \mathbb{P}$ *uno spazio di probabilità e siano* A, B *due eventi. Valgono le seguenti affermazioni.*

1. *Se* A, B *sono indipendenti, allora* A^C *e* B *sono indipendenti.*
2. *Se* A *è trascurabile oppure quasi certo, allora* A *e* B *sono indipendenti qualunque sia* B*.*
3. *Se* A *e* B *sono non trascurabili e incompatibili (ovvero* $A \cap B = \varnothing$*) allora* A *e* B *non sono indipendenti.*

Dimostrazione. Mostriamo le tre affermazioni separatamente.

1. Mostriamo che A^C e B sono indipendenti tramite la definizione:

$$\begin{aligned}
\mathbb{P}(A^C \cap B) &= \mathbb{P}(B \setminus (A \cap B)) && \text{(per la 2.1.3)} \\
&= \mathbb{P}(B) - \mathbb{P}(A \cap B) && \text{(A, B indip.)} \\
&= \mathbb{P}(B) - \mathbb{P}(A)\mathbb{P}(B) \\
&= \mathbb{P}(B)(1 - \mathbb{P}(A)) \\
&= \mathbb{P}(B)\mathbb{P}(A^C).
\end{aligned}$$

2. Sia B un evento qualunque: allora se A è quasi certo segue che

$$\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B)\mathbb{P}(B) = 1 \cdot \mathbb{P}(B) = \mathbb{P}(A)\mathbb{P}(B).$$

Dimostrazione analoga se A è trascurabile.

3. Se $A \cap B = \varnothing$ allora $\mathbb{P}(A \cap B) = \mathbb{P}(\varnothing) = 0$, dunque se fossero indipendenti avremmo che $\mathbb{P}(A)\mathbb{P}(B) = 0$, il che è assurdo poiché li abbiamo supposti entrambi non trascurabili. □

2.4 PROBABILITÀ SULLA RETTA REALE

Se il nostro insieme Ω è l'insieme $\mathbb{R}$ dei numeri reali, possiamo definire diversi tipi di probabilità.

Probabilità discreta

Il primo tipo di probabilità che possiamo definire su $\mathbb{R}$ è la *probabilità discreta*: dato un insieme finito o numerabile di punti $(x_i)_{i \in \mathbb{N}}$ poniamo

$$p(x_i) \coloneqq \mathbb{P}(\{x_i\}).$$

Siccome vogliamo definire la nostra probabilità in modo che questi sottoinsiemi siano gli unici non trascurabili, imponiamo inoltre che

$$\sum_{i \in \mathbb{N}} p(x_i) = 1.$$

La probabilità è quindi *concentrata* in un insieme numerabile di punti: dato $A \subseteq \mathbb{R}$ la probabilità dell'insieme A è

$$\mathbb{P}(A) = \sum_{x_i \in A} p(x_i).$$

Definizione 2.4.1 **Funzione di massa.** La funzione $p : \mathbb{R} \to \mathbb{R}$ tale che

$$p(x) := \begin{cases} \mathbb{P}(\{x_i\}) & \text{se } x = x_1 \text{ per qualche } x_i \\ 0 & \text{altrimenti} \end{cases}$$

è detta *funzione di massa* oppure *densità discreta*.

Osservazione. Una probabilità discreta può essere definita su ogni sottoinsieme di $\mathbb{R}$: come vedremo, questo non è il caso per altri tipi di probabilità.

Probabilità definita da una densità

Definizione 2.4.2 **Densità di probabilità.** Si dice *densità di probabilità* una funzione $f : \mathbb{R} \to [0, +\infty)$ integrabile e tale che

$$\int_{-\infty}^{+\infty} f(x)dx = 1.$$

Definizione 2.4.3 **Probabilità definita da una densità.** Sia $f : \mathbb{R} \to [0, +\infty)$ una densità, $A \subseteq \mathbb{R}$. La probabilità definita da f è tale che

$$\mathbb{P}(A) = \int_A f(x)dx.$$

Notiamo che questa funzione definisce davvero una probabilità:

- la probabilità di tutto $\mathbb{R}$ è

$$\mathbb{P}(\mathbb{R}) = \int_{\mathbb{R}} f(x)dx = \int_{-\infty}^{+\infty} f(x)dx = 1.$$

- è finitamente additiva: se $A, B \subseteq \mathbb{R}$ sono disgiunti, allora

$$\mathbb{P}(A \cup B) = \int_{A \cup B} f(x)dx = \int_A f(x)dx + \int_B f(x)dx = \mathbb{P}(A) + \mathbb{P}(B).$$

- è anche numerabilmente additiva (deriva dal Teorema di Beppo-Levi).

Tuttavia questa probabilità, al contrario della probabilità discreta, non è definibile su tutti i sottoinsiemi di $\mathbb{R}$.

Esempio 2.4.4. Supponiamo di volere una funzione che restituisce un numero reale a caso tra 0 e 1.

Lo spazio di probabilità più naturale per questa funzione è $\Omega = [0, 1]$, in quanto sicuramente il numero scelto sarà in questo intervallo. Per definire la probabilità di un sottoinsieme di Ω, iniziamo studiando il caso in cui il sottoinsieme sia un intervallo chiuso $[a, b]$.

La probabilità che un numero casuale sia in $[a, b]$ può essere pensata come la lunghezza dell'intervallo $[a, b]$ diviso la lunghezza dello spazio universo Ω: dunque

$$\mathbb{P}([a, b]) = \frac{b-a}{1} = b - a.$$

Questa probabilità può essere definita come *probabilità data da una densità*: la densità ad essa relativa è la funzione

$$f(x) := \begin{cases} 1 & \text{se } 0 \leqslant x \leqslant 1 \\ 0 & \text{altrimenti.} \end{cases}$$

Questo esempio innanzitutto mostra la differenza tra *eventi trascurabili* e *eventi impossibili*: preso un qualunque $x \in [0, 1]$ si ha che

$$\{x\} \subseteq \left[x - \frac{1}{n}, x + \frac{1}{n}\right],$$

qualunque sia $n \in \mathbb{N}$. Questo significa che la probabilità del singoletto $\{x\}$ deve essere minore o uguale della probabilità dell'insieme in cui è contenuto, ovvero

$$\mathbb{P}(\{x\}) \leqslant \mathbb{P}\left(\left[x - \frac{1}{n}, x + \frac{1}{n}\right]\right) = \frac{2}{n},$$

ma siccome ciò deve valere per n arbitrariamente grande segue che $\mathbb{P}(\{x\}) = 0$. L'evento definito da $\{x\}$ è quindi trascurabile, ma certamente non è impossibile in quanto x può essere il risultato dell'estrazione di un numero reale casuale.

Il secondo punto, molto più difficile da mostrare, è che questa probabilità non è definita per ogni sottoinsieme di $\mathbb{R}$, ma solo per i sottoinsiemi *misurabili*: il *controesempio di Vitali* mostra che esistono sottoinsiemi di $[0, 1]$ per cui questa probabilità non può essere definita.

D'ora in avanti considereremo soltanto sottoinsiemi misurabili di $\mathbb{R}$.

DENSITÀ ESPONENZIALE Un tipo di densità molto utile è quella definita dalla seguente funzione:

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

Questa funzione è una probabilità in quanto è integrabile e

$$\begin{aligned} \int_{-\infty}^{+\infty} f(x)dx &= \int_{-\infty}^{0} 0dx + \int_{0}^{+\infty} e^{-x} \\ &= \lim_{M\to+\infty} \left[-e^{-x}\right]_0^M \\ &= \lim_{M\to+\infty} -e^{-x} + 1 \\ &=1. \end{aligned}$$

2.5 VARIABILI ALEATORIE

Per introdurre il concetto di variabili aleatorie e la loro utilità useremo il seguente esempio:

ESEMPIO 2.5.1. Supponiamo di giocare alla roulotte e di aver puntato 1£ sul numero 28 e 1£ sull'uscita di un numero pari. Possiamo ad esempio domandarci quanto sia la probabilità di vincere più di 10£, oppure quanto sia la probabilità di perdere soldi.

Lo spazio più naturale per questo problema è l'insieme dei possibili risultati della roulotte, ovvero $\Omega = \{0, 1, \ldots, 36\}$ munito della distribuzione uniforme di probabilità; tuttavia gli eventi che vogliamo considerare ("vincere più di 10£", "perdere soldi") non sono sottoinsiemi di Ω, bensì di $\mathbb{R}$.

Possiamo quindi definire una funzione $X : \Omega \to \mathbb{R}$ che indichi nel seguente modo la "vittoria netta":

$$X(\omega) = \begin{cases} 36, & \text{se } \omega = 28 \\ 0, & \text{se } \omega \neq 28, \omega \text{ è pari} \\ -1, & \text{se } \omega = 0 \\ -2, & \text{se } \omega \text{ è dispari.} \end{cases}$$

Usando la funzione X possiamo *trasportare* la probabilità da Ω a $\mathbb{R}$: la risposta alla prima domanda è dunque

$$\mathbb{P}(\{\omega \in \Omega : X(\omega) > 10\}) = \mathbb{P}\Big(X^{-1}([0, +\infty))\Big) = \frac{1}{37};$$

la risposta alla seconda domanda è invece

$$\mathbb{P}(\{\omega \in \Omega : X(\omega) < 0\}) = \mathbb{P}\Big(X^{-1}((-\infty, 0))\Big) = \frac{19}{37}.$$

Definizione 2.5.2 **Variabili aleatorie.** Sia $(\Omega, \mathcal{F}, \mathbb{P})$ uno spazio di probabilità. Una funzione $X : \Omega \to \mathbb{R}$ si dice *variabile aleatoria* e la funzione

$$\mathbb{P}_X(A) = \mathbb{P}(X^{-1}(A))$$

viene detta *legge di probabilità* associata alla variabile aleatoria X.

Dato $A \subseteq \mathbb{R}$, introduciamo la notazione $\{X \in A\}$ per indicare l'insieme di tutti i valori dello spazio fondamentale Ω la cui immagine cade in A, ovvero

$$\{X \in A\} \coloneqq \{\omega \in \Omega : X(\omega) \in A\} = X^{-1}(A).$$

Definizione 2.5.3 **Variabili equidistribuite.** Siano $X, Y : \Omega \to \mathbb{R}$. X e Y si dicono equidistribuite se hanno la stessa legge di probabilità.

Definizione 2.5.4 **Variabili discrete e continue.** Sia $X : \Omega \to \mathbb{R}$ una variabile aleatoria. Allora X si dice:

1. *discreta* se la sua distribuzione è discreta, ovvero se l'immagine di X è un sottoinsieme finito o numerabile di $\mathbb{R}$;
2. *con densità* se la sua distribuzione è con densità, ovvero se esiste una densità $f : \mathbb{R} \to [0, +\infty)$ tale che

$$\mathbb{P}_X(A) = \mathbb{P}(\{X \in A\}) = \int_A f(x)dx.$$

Questi due tipi non sono tutti i tipi possibili di variabili aleatorie, tuttavia ci occuperemo solo di questi due casi (anche perché sono i casi più rilevanti nelle applicazioni).

Definizione 2.5.5 **Funzione di ripartizione.** Data una variabile aleatoria $X : \Omega \to \mathbb{R}$ si dice *funzione di ripartizione* di X (o anche c.d.f., da *cumulative distribution function*) la funzione $F_X : \mathbb{R} \to [0, 1]$ tale che

$$F_X(x) = \mathbb{P}(\{X \leqslant x\}) = \mathbb{P}(\{X \in (-\infty, x]\}).$$

Notiamo che la funzione di ripartizione ha due comportamenti diversi a seconda del tipo di variabile aleatoria che stiamo considerando:

DISCRETA La funzione di ripartizione è data da

$$F_X(x) = \sum_{x_i \leqslant x} p(x_i).$$

CON DENSITÀ La funzione di ripartizione è data da

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

dove f è la densità associata alla variabile X. Notiamo che siccome f è integrabile, allora la corrispondente F_X è necessariamente continua.

Non vale invece il viceversa: è possibile trovare una funzione F_X continua la cui variabile aleatoria associata X non è con densità.

La funzione di ripartizione ha delle interessanti proprietà che racchiuderemo nella prossima proposizione.

Proposizione 2.5.6 **Proprietà della funzione di ripartizione.** *Sia $X : \Omega \to \mathbb{R}$ una variabile aleatoria e sia F_X la sua funzione di ripartizione. Valgono le seguenti affermazioni:*

- *F_X è crescente.*
- $\lim_{x \to -\infty} F_X(x) = 0, \lim_{x \to +\infty} F_X(x) = 1.$
- *F_X è sempre continua a destra; ovvero per ogni $x_0 \in \mathbb{R}$ vale che*

$$\lim_{x \to x_0^+} F_X(x) = F_X(x_0).$$

Dimostrazione. Dimostriamo ad esempio che $\lim_{x \to -\infty} F_X(x) = 0$.

Sia (x_n) una successione decrescente che diverge negativamente (ovvero $x_n \to -\infty$). Sia A_n l'insieme definito da

$$A_n := \{X \leqslant x_n\} = X^{-1}\big((-\infty, x_n]\big).$$

Siccome x_n è decrescente segue che $A_n \supseteq A_{n-1}$; dunque (sfruttando il fatto che $x_n \to -\infty$)

$$\bigcap_{n=1}^{\infty} X^{-1}\big((-\infty, x_n]\big) = X^{-1}\left(\bigcap_{n=1}^{\infty} A_n\right) = X^{-1}(\varnothing) = \varnothing.$$

Vale quindi che

$$\begin{aligned} \lim_{x \to -\infty} F_X(x) &= \lim_{n \to +\infty} F_X(x_n) = \lim_{n \to +\infty} \mathbb{P}(A_n) \\ &= \mathbb{P}\left(\lim_{n \to +\infty} A_n\right) = \mathbb{P}(\varnothing) = 0. \end{aligned}$$ □

Un fatto importante è che la proposizione precedente può essere in qualche modo "invertita": data una funzione $F : \mathbb{R} \to [0, 1]$ che rispetta le tre proprietà della Proposizione 2.5.6 esiste una e una sola probabilità $\mathbb{P}()$ sui sottoinsiemi di $\mathbb{R}$ tale che F sia la funzione di ripartizione di una variabile aleatoria con legge di probabilità $\mathbb{P}()$.

Da ciò segue che tutte le variabili aleatorie che hanno la stessa funzione di ripartizione hanno a maggior ragione la stessa legge di probabilità, ovvero sono *equidistribuite*. Nel caso generale è difficile ricavare la legge di probabilità dalla c.d.f., ma se ci limitiamo al caso delle variabili discrete e con densità è fattibile. Possiamo notrare infatti che in generale vale che

$$F_X(x_0) - \lim_{x \to x_0^-} F_X(x) = \mathbb{P}(\{X = x_0\}).$$

Nel caso discreto questo implica che $\mathbb{P}(\{X = x_i\}) = p(x_i)$: in ogni punto di salto l'*ampiezza del salto* è data proprio dal valore della funzione di massa nel punto.

Nel caso con densità possiamo ricavare la legge di probabilità invertendo l'operazione di integrazione: in ogni punto in cui la densità f è continua vale che

$$f(x) = \frac{dF_X(x)}{dx}.$$

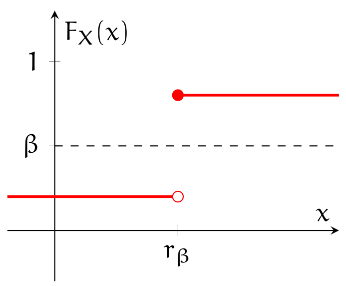
Vediamo ora la definizione di quantile:

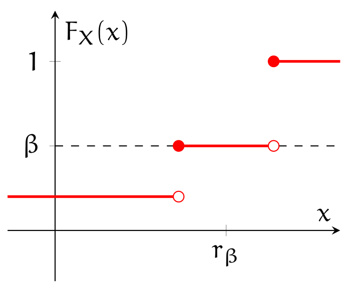
Definizione 2.5.7 **Quantile.** Sia X una variabile aleatoria con legge di probabilità $\mathbb{P}()$. Sia inoltre $\beta \in (0, 1)$. Si dice β-*quantile* un numero r_β tale che

$$\mathbb{P}(\{X \leqslant r_\beta\}) \geqslant \beta, \quad \mathbb{P}(\{X \geqslant r_\beta\}) \geqslant 1 - \beta.$$

Nel caso in cui la variabile sia discreta ci sono due possibilità:

1. La funzione non assume mai il valore β poiché β è compreso tra gli estremi di un punto di salto. In questo caso l'unico r_β possibile è il punto di salto, il cui valore è il minimo valore assunto dalla funzione superiore a β.
2. La funzione assume il valore β su tutto un intervallo. In questo caso per convenzione si sceglie r_β come il punto medio dell'intervallo, anche se tutti i punti dell'intervallo andrebbero ugualmente bene.





Se invece la variabile X ha densità, *solitamente* esiste uno e un solo β-quantile, ed è il valore $r_\beta \in \mathbb{R}$ tale che $F_X(r_\beta) = \beta$.

Esempio 2.5.8. Sia X una variabile aleatoria con densità esponenziale, ovvero

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

La sua funzione di ripartizione è

$$F_X(x) = \begin{cases} 0, & \text{se } x \leqslant 0 \\ \int_0^x e^{-t} dt = 1 - e^{-x}, & \text{se } x > 0. \end{cases}$$

Il β-quantile di X è quindi quel numero r_β tale che

$$F_X(r_\beta) = 1 - e^{-r_\beta} = \beta \iff r_\beta = -\log(1 - \beta).$$

Definiamo ora i tipi più comuni di variabili aleatorie.

2.5.1 Variabile binomiale

La situazione concreta da cui nasce questa variabile aleatoria è ad esempio la seguente: vogliamo ripetere n volte, in condizioni di indipendenza, un esperimento che può avere come risultato o il successo (con probabilità p_0), o l'insuccesso (con probabilità $1 - p_0$). La variabile X deve quindi contare il numero di successi in n tentativi.

È chiaro quindi che i possibili valori che la variabile X può assumere sono $0, 1, \dots, n$: negli n tentativi posso avere 0, 1, ..., oppure n successi.

Calcoliamo ora la funzione di massa p: dato un qualsiasi $k \in \{0, \dots, n\}$ vogliamo calcolare la probabilità che la variabile aleatoria valga esattamente k, ovvero che su n tentativi k di essi siano successi.

La funzione di massa è la seguente:

$$p(k) = \mathbb{P}(\{X = k\}) = \binom{n}{k} p_0^k (1 - p_0)^{n-k}.$$

Dimostrazione. Consideriamo una stringa di successi/insuccessi di lunghezza n. Esistono esattamente $\binom{n}{k}$ stringhe con k successi; inoltre la probabilità ad essa associata è $p_0^k(1-p_0)^{n-k}$, in quanto ogni successo ha probabilità p_0 e ogni insuccesso ha probabilità $1 - p_0$, da cui la tesi. □

Per esser certo che p sia effettivamente una funzione di massa verifico che $\displaystyle\sum_{k=0}^{n} p(k) = 1$.

Dimostrazione. Per definizione di p:

$$\sum_{k=0}^{n} p(k) = \sum_{k=0}^{n} \binom{n}{k} p_0^k (1-p_0)^{n-k} = \Big(p_0 + (1 - p_0)\Big)^n = 1,$$

dove il terzo passaggio è giustificato dal binomio di Newton. □

Se $n = 1$ la variabile X viene detta *di Bernoulli* di parametro p_0.

2.5.2 Variabile geometrica

Vogliamo ripetere, in condizioni di indipendenza, un esperimento che ha come risultati il successo (con probabilità p_0) oppure l'insuccesso (con probabilità $1 - p_0$) finché l'esperimento non ha successo.

Come nel caso della variabile binomiale, la variabile aleatoria X conta il numero di tentativi necessari: in questo caso però i valori possibili sono tutti i possibili valori naturali positivi insieme a $\{+\infty\}$, in quanto a priori il successo potrebbe non arrivare mai.

Calcoliamo ora la funzione di massa p: dato un qualsiasi $k \in \{1, 2, \dots, +\infty\}$ vogliamo calcolare la probabilità che la variabile aleatoria valga esattamente k, ovvero che i primi $k - 1$ tentativi siano insuccessi e il K-esimo sia un successo. Nel caso $k = +\infty$, questo significa che tutti i tentativi sono insuccessi.

La funzione di massa legata a questa variabile è la seguente:

$$p(k) = \mathbb{P}\big(\{X = k\}\big) = (1 - p_0)^{k-1} \cdot p_0.$$

Dimostrazione. Siccome la probabilità di ogni insuccesso è di $1 - p_0$ e abbiamo $k - 1$ insuccessi, la probabilità di ogni successo è p_0 e ne abbiamo uno solo, e gli eventi sono tutti indipendenti segue la tesi. □

Verifichiamo che $\sum p(k) = 1$:

$$\begin{aligned}\sum_{k=1}^{+\infty} p(k) &= \sum_{k=1}^{+\infty} \mathbb{P}(\{X = k\}) \\ &= \sum_{k=1}^{+\infty} (1-p_0)^{k-1} \cdot p \\ &= p \cdot \sum_{k=1}^{+\infty} (1-p_0)^{k-1}\end{aligned}$$

Per ricondurmi alla serie geometrica pongo $h := k - 1$, ottenendo

$$\begin{aligned}&= p \cdot \sum_{h=0}^{+\infty} (1-p_0)^{h} \\ &= p \cdot \frac{1}{1-(1-p)} \\ &= 1.\end{aligned}$$

La variabile geometrica ha un'altra interessante proprietà, detta *proprietà dell'assenza di memoria*.

Proposizione 2.5.9 **Assenza di memoria della variabile geometrica.** *Siano* $n, h \in \mathbb{N}$, $n, h > 0$ *e sia* X *una variabile aleatoria geometrica. Allora vale che*

$$\mathbb{P}\big(\{X = n+h\} \mid \{X > n\}\big) = \mathbb{P}\big(\{X = h\}\big).$$

Dimostrazione. Per definizione di probabilità condizionata

$$\mathbb{P}\big(\{X = n+h\} \mid \{X > n\}\big) = \frac{\mathbb{P}\big(\{X = n+h\}\big)}{\mathbb{P}\big(\{X > n\}\big)}. \tag{11}$$

Calcoliamo il denominatore di questa espressione:

$$\begin{aligned}\mathbb{P}\big(\{X > n\}\big) &= \sum_{h=n+1}^{\infty} p(h) \\ &= \sum_{h=n+1}^{\infty} (1-p_0)^{h-1} \cdot p_0\end{aligned}$$

Ponendo $k := h - n$, ovvero $h = k + n$:

$$\begin{aligned}&= \sum_{k=1}^{\infty} (1-p_0)^{k+n-1} \cdot p_0 \\ &= (1-p_0)^n \cdot \sum_{k=1}^{\infty} (1-p_0)^{k-1} \cdot p_0 \\ &= (1-p_0)^n,\end{aligned}$$

dove l'ultimo passaggio è giustificato dai calcoli fatti nel verificare la correttezza della funzione di massa della variabile binomiale.

Sostituendo nella (11):

$$\begin{aligned}\frac{\mathbb{P}\big(\{X = n+h\}\big)}{\mathbb{P}\big(\{X > n\}\big)} &= \frac{(1+p_0)^{n+h-1} \cdot p_0}{(1-p_0)^n} \\ &= (1-p_0)^{h-1} \cdot p_0 \\ &= \mathbb{P}\big(\{X = h\}\big).\end{aligned}$$

$\square$

2.5.3 Variabile di Poisson

Si dice *variabile di Poisson di parametro* λ (con $\lambda \in \mathbb{R}$, $\lambda > 0$) la variabile aleatoria con dominio $\mathbb{N}$ definita dalla seguente funzione di massa:

$$p(h) = \mathbb{P}(\{X = h\}) \coloneqq e^{-\lambda}\frac{\lambda^h}{h!}.$$

Verifichiamo che $\sum p(h) = 1$:

$$\begin{aligned}\sum_{h=0}^{\infty} p(h) &= \sum_{h=0}^{\infty} e^{-\lambda}\frac{\lambda^h}{h!}\\ &= e^{-\lambda}\sum_{h=0}^{\infty}\frac{\lambda^h}{h!}\\ &= e^{-\lambda}e^{\lambda}\\ &= 1.\end{aligned}$$

2.5.4 Densità esponenziale

Si definisce *densità esponenziale di parametro* λ la funzione $f : \mathbb{R} \to [0, +\infty)$ tale che

$$f(x) \coloneqq \begin{cases}\lambda e^{-\lambda x}, & x \geqslant 0\\ 0, & x < 0.\end{cases}$$

Questa funzione è effettivamente una densità: infatti

$$\int_{-\infty}^{+\infty} f(x)dx = \int_{-\infty}^{0} 0dx + \int_{0}^{+\infty} \lambda e^{-\lambda x}dx$$

Sostituendo $t \coloneqq \lambda x$ (da cui $dt = \lambda dx$) l'integrale è equivalente a

$$\begin{aligned}&= \int_0^{+\infty} e^{-t}dt\\ &= 1.\end{aligned}$$

La funzione di ripartizione corrispondente a questa densità è data da

$$F_X(x) = \int_{-\infty}^{x} f(t)dt = \begin{cases}0, & x < 0\\ \int_0^x \lambda e^{-\lambda t}dt = 1 - e^{-\lambda x}, & x \geqslant 0.\end{cases}$$

La variabile esponenziale è *senza memoria*, esattamente come la variabile binomiale, ovvero per ogni $s, t > 0$ vale che

$$\mathbb{P}(\{X \leqslant s+t\} \mid \{X > s\}) = \mathbb{P}(\{X \leqslant t\}).$$

Dimostrazione. Per definizione di probabilità condizionata:

$$\begin{aligned}\mathbb{P}(\{X \geqslant s+t\} \mid \{X > s\}) &= \frac{\mathbb{P}(\{X \leqslant s+t\} \cap \{X > s\})}{\mathbb{P}(\{X > s\})}\\ &= \frac{\mathbb{P}(\{s < X \leqslant s+t\})}{\mathbb{P}(\{X > s\})}.\end{aligned}$$

Facciamo alcune osservazioni generali:

$$\begin{aligned}(1) : \mathbb{P}(\{X > s\}) &= 1 - \mathbb{P}(\{X \leqslant s\}) = 1 - F_X(s).\\ (2) : \mathbb{P}(\{s < X \leqslant s+t\}) &= \mathbb{P}(\{X \leqslant s+t\}) - \mathbb{P}(\{X \leqslant s\})\\ &= F_X(s+t) - F_X(s).\end{aligned}$$

Sostituendo nel nostro caso si ha

$$\begin{aligned}
\frac{\mathbb{P}\big(\{s < X \leqslant s+t\}\big)}{\mathbb{P}\big(\{X > s\}\big)} &= \frac{F_X(s+t) - F_X(s)}{1 - F_X(s)} \\
&= \frac{1 - e^{-\lambda(s+t)} - (1 - e^{-\lambda s})}{1 - (1 - e^{-\lambda s})} \\
&= \frac{-e^{-\lambda s} e^{-\lambda t} + e^{-\lambda s}}{e^{-\lambda s}} \\
&= \frac{e^{-\lambda s}\big(1 - e^{-\lambda t}\big)}{e^{-\lambda s}} \\
&= 1 - e^{-\lambda t} \\
&= \mathbb{P}\big(X \leqslant t\big).
\end{aligned}$$

□

A PREREQUISITI

A.1 SERIE

Definizione A.1.1 **Successioni delle somme parziali.** Sia $(a_n)_{n\in\mathbb{N}}$ una successione. Si dice *successione delle somme parziali* la successione $(s_n)_{n\in\mathbb{N}}$ tale che

$$s_n \coloneqq a_0 + a_1 + \cdots + a_n = \sum_{i=0}^{n} a_i.$$

Definizione A.1.2 **Serie.** Sia $(a_n)_{n\in\mathbb{N}}$ una successione e $(s_n)_{n\in\mathbb{N}}$ la sua successione delle somme parziali. Allora si dice serie il limite per $n \to +\infty$ delle somme parziali, e lo si indica con

$$\sum_{i=0}^{\infty} a_n \coloneqq \lim_{n\to+\infty} s_n.$$

Spesso si usa anche la notazione $\sum a_n$ quando l'indice di partenza è sottointeso oppure non è importante nel calcolo della serie.

Definizione A.1.3 **Comportamento di una serie.** Sia (a_n) una successione. Allora si dice che

(1) $\sum a_n$ converge se $\lim\limits_{n\to+\infty} s_n = l$ per qualche $l \in \mathbb{R}$;

(2) $\sum a_n$ diverge positivamente se $\lim\limits_{n\to+\infty} s_n = +\infty$;

(3) $\sum a_n$ diverge negativamente se $\lim\limits_{n\to+\infty} s_n = -\infty$;

(4) $\sum a_n$ è indeterminata se $\lim\limits_{n\to+\infty} s_n$ non esiste.

Proposizione A.1.4 **Condizione necessaria per la convergenza.** *Sia (a_n) una successione. Allora se $\sum a_n$ converge segue che*

$$\lim_{n\to+\infty} a_n = 0.$$

Dimostrazione. Per definizione della successione delle somme parziali

$$\begin{aligned} s_n &\coloneqq a_0 + \cdots + a_{n-1} + a_n \\ s_{n-1} &\coloneqq a_0 + \cdots + a_{n-1} \end{aligned}$$

dunque

$$a_n = s_n - s_{n-1}.$$

Supponiamo che $\sum a_n$ converga al valore reale l: il limite della successione (a_n) sarà quindi

$$\begin{aligned} \lim_{n\to+\infty} a_n &= \lim_{n\to+\infty} s_n - s_{n-1} \\ &= l - l \\ &= 0. \end{aligned}$$

□

Proposizione A.1.5 *Sia (a_n) una successione crescente e a termini non negativi (ovvero $a_n \geqslant 0$ per ogni $n \in \mathbb{N}$): allora la serie $\sum a_n$ è convergente oppure divergente positivamente.*

Dimostrazione. La successione delle somme parziali è debolmente crescente (ad ogni passo aggiungiamo un numero positivo o nullo), dunque non può divergere negativamente o essere indeterminata. □

Definizione A.1.6 **Serie assolutamente convergente.** Sia (a_n) una successione. Se $\sum |a_n|$ converge, allora la serie $\sum a_n$ si dice assolutamente convergente.

Osservazione. Siccome $|a_n|$ è una successione a termini positivi o nulli, la serie $\sum |a_n|$ (in virtù della Proposizione A.1.5) può soltanto convergere o divergere positivamente.

Proposizione A.1.7 **Proprietà delle serie assolutamente convergenti.** *Sia (a_n) una successione la cui serie converge assolutamente. Allora valgono le seguenti affermazioni:*

(i) la serie $\sum a_n$ converge;

(ii) se cambio l'ordine dei termini della successione, la serie converge allo stesso valore della serie relativa alla successione originale;

(iii) data una partizione di $\mathbb{N}$ della forma $A_1, A_2, \dots$ vale che

$$\sum_{n=1}^{\infty} a_n = \sum_{n=1}^{\infty} \left(\sum_{k \in A_n} a_k \right).$$

Serie geometrica

Dato $a \in \mathbb{R}$ tale che $|a| < 1$ si dice *serie geometrica* la serie

$$\sum_{n=0}^{\infty} a^n = 1 + a + a^2 + \dots$$

Proposizione A.1.8 *Sia $a \in \mathbb{R}$, $|a| < 1$. Allora*

$$\sum_{n=0}^{\infty} a^n = \frac{1}{1-a}. \tag{12}$$

Dimostrazione. Mostriamo innanzitutto per induzione su n che

$$s_n = \frac{a^{n+1} - 1}{a - 1}.$$

CASO BASE Se $n = 0$ allora $s_0 = a^0 = 1$.

PASSO INDUTTIVO Supponiamo che la formula valga per $n - 1$ e dimostriamo che vale per n.

$$\begin{aligned} s_n &= s_{n-1} + a^n \\ &= \frac{a^n - 1}{a - 1} + a^n \\ &= \frac{a^n - 1 + a^n(a-1)}{a - 1} \\ &= \frac{a^n - 1 + a^{n+1} - a^n}{a - 1} \\ &= \frac{a^{n+1} - 1}{a - 1}. \end{aligned}$$

Dunque la formula vale per ogni $n \in \mathbb{N}$: quando n tende a $+\infty$ allora avremo che

$$\sum a^n = \lim_{n \to +\infty} s_n = \lim_{n \to +\infty} \frac{a^{n+1} - 1}{a - 1} = \frac{-1}{a - 1} = \frac{1}{1 - a},$$

dove abbiamo usato il fatto che $a^n \to 0$ se $|a| < 1$. $\square$

Serie esponenziale

Vale la seguente formula per l'esponenziale: per ogni $x \in \mathbb{R}$

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}. \tag{13}$$