

Note del corso di Statistica per l'Ingegneria Fisica

Alessandro Toigo

21 aprile 2016

Indice

1	Calcolo delle Probabilità	5
1.1	Definizione e proprietà elementari della probabilità	5
1.2	Probabilità condizionata	11
1.3	Indipendenza	15
1.4	Prove di Bernoulli	18
1.5	Variabili aleatorie	20
1.6	La funzione di ripartizione di una variabile aleatoria	26
1.7	Funzioni di una variabile aleatoria e standardizzazione	29
1.8	Vettori aleatori	31
1.9	Media e varianza di una variabile aleatoria	38
1.10	Disuguaglianza di Chebyshev e legge dei grandi numeri	46
1.11	Teorema del Limite Centrale	50
1.12	Approssimazione normale e poissoniana della binomiale	51
2	Statistica	53
2.1	Intervalli di confidenza	53
2.2	Test d'ipotesi	66
2.3	Test non parametrici	79
3	Regressione lineare	83
3.1	Il metodo dei minimi quadrati	83
3.2	Il modello statistico	87

Capitolo 1

Calcolo delle Probabilità

1.1 Definizione e proprietà elementari della probabilità

In un *esperimento aleatorio*, un *evento* è una qualunque *proposizione riguardante il risultato dell'esperimento stesso*. Per chiarire le idee, consideriamo come esempio l'esperimento aleatorio consistente in tre lanci consecutivi di una stessa moneta. Allora tutte le proposizioni seguenti sono esempi di eventi:

$$\begin{aligned}T_i &= \text{"esce testa all'i-esimo lancio"} \quad (\text{dove } i = 1, 2, 3) \\E &= \text{"nei primi due lanci esce la stessa faccia"} \\F &= \text{"negli ultimi due lanci esce la stessa faccia"} \\G &= \text{"il risultato del primo e del terzo lancio sono diversi"}.\end{aligned}$$

Tramite le *operazioni logiche* di \wedge ("and"), \vee ("or") e $\bar{}$ ("not"), gli eventi possono essere combinati tra loro in modo da formare nuovi eventi oppure ottenere equazioni logiche. In questo modo, è facile vedere che l'insieme degli eventi acquista una struttura di *algebra booleana*. Senza entrare nei dettagli della definizione assiomatica precisa di un'algebra booleana, limitiamoci a osservare che, nell'esempio precedente dei tre lanci di una moneta, a partire dagli eventi T_1, T_2, T_3, E, F, G possiamo scrivere

$$\begin{aligned}\text{"esce croce al primo lancio"} &= \overline{T_1} \\ \text{"esce sempre testa"} &= T_1 \wedge T_2 \wedge T_3 \\ \text{"esce la stessa faccia in tutti i lanci"} &= E \wedge F\end{aligned}\tag{1.1}$$

e ancora

$$\begin{aligned}E &= (T_1 \wedge T_2) \vee (\overline{T_1} \wedge \overline{T_2}) \\ E \wedge F &= E \wedge \overline{G} = (T_1 \wedge T_2 \wedge T_3) \vee (\overline{T_1} \wedge \overline{T_2} \wedge \overline{T_3}) \\ T_1 \wedge G &= T_1 \wedge \overline{T_3} \\ (E \vee F) \wedge T_2 &= (T_1 \wedge T_2) \vee (T_2 \wedge T_3).\end{aligned}\tag{1.2}$$

Notiamo in particolare l'importanza dell'uso corretto delle parentesi quando sono coinvolte nella stessa espressione entrambe le operazioni logiche \wedge e \vee . Disporre le parentesi nel giusto ordine

è essenziale tra l'altro per enunciare nel modo corretto la *proprietà distributiva dell'“or” rispetto all'“and”*

$$A \vee (C \wedge D) = (A \vee C) \wedge (A \vee D)$$

e l'analoga *proprietà distributiva dell'“and” rispetto all'“or”*

$$A \wedge (C \vee D) = (A \wedge C) \vee (A \wedge D).$$

Anche l'operazione di “not” assume un diverso significato a seconda della sua posizione. Per convincersene, basta osservare che nel lancio delle monete i due eventi $\overline{T_1} \wedge \overline{T_2}$ e $\overline{T_1 \wedge T_2}$ sono completamente diversi. In generale, valgono infatti le *leggi di De Morgan*

$$\overline{A \wedge B} = \overline{A} \vee \overline{B} \quad \text{e} \quad \overline{A \vee B} = \overline{A} \wedge \overline{B}.$$

Per indicare l'*implicazione logica* tra due eventi si usa il simbolo \leq , cioè

$$A \leq B \quad \text{significa che l'evento } A \text{ implica l'evento } B.$$

Per esempio, nell'esperimento dei tre lanci di una moneta

$$\begin{aligned} T_1 \wedge T_2 &\leq T_1 \\ T_1 \wedge \overline{T_3} &\leq G. \end{aligned} \tag{1.3}$$

Infine, un ruolo particolare è giocato dall'*evento certo* (indicato con 1) e dall'*evento impossibile* (che denoteremo 0). Per chiarire il significato di questi due eventi, osserviamo che per esempio

$$1 = T_1 \vee \overline{T_1} = G \vee (T_1 \wedge T_3) \vee (\overline{T_1} \wedge \overline{T_3})$$

e

$$0 = \overline{1} = \overline{T_1} \wedge T_1 = E \wedge F \wedge G.$$

L'*algebra booleana* che si ottiene dotando gli eventi delle operazioni logiche \wedge , \vee e $\overline{}$ ricorda (anche visivamente!) le *operazioni di intersezione, unione e complementazione di insiemi*. Questa è in effetti ben più di una semplice somiglianza intuitiva. Infatti, *ciò che si fa in probabilità è proprio rappresentare l'algebra degli eventi in un'opportuna algebra di insiemi*. Più precisamente:

- (a) assegnato un esperimento aleatorio, si fissa un opportuno insieme Ω , detto *spazio campionario* (o spazio ambiente) di quel particolare esperimento;
- (b) *gli eventi dell'esperimento vengono rappresentati in sottoinsiemi di Ω* , cioè in insiemi $E, F, G \dots$ appartenenti all'insieme delle parti $\mathcal{P}(\Omega)$ di Ω ;
- (c) in questa rappresentazione, le operazioni logiche \wedge , \vee e $\overline{}$ vengono fatte corrispondere all'intersezione \cap , unione \cup e complementazione c di insiemi; inoltre, *l'implicazione logica \leq tra due eventi corrisponde al contenimento \subseteq di uno nell'altro*.

Per chiarire nuovamente le idee, torniamo ancora al nostro esempio dei tre lanci di una moneta. Una possibile scelta dello **spazio campionario** per tale esperimento è il **prodotto cartesiano**

$$\Omega = \{0, 1\}^3 = \{(\omega_1, \omega_2, \omega_3) \mid \omega_i \in \{0, 1\}\},$$

mentre gli eventi T_1, T_2, T_3, E, F, G si possono rappresentare nei sottoinsiemi

$$T_1 = \{1\} \times \{0, 1\}^2 = \{(\omega_1, \omega_2, \omega_3) \in \Omega \mid \omega_1 = 1\} = \{(1, 0, 0), (1, 1, 0), (1, 0, 1), (1, 1, 1)\}$$

$$T_2 = \{0, 1\} \times \{1\} \times \{0, 1\} = \{(\omega_1, \omega_2, \omega_3) \in \Omega \mid \omega_2 = 1\}$$

$$T_3 = \{0, 1\}^2 \times \{1\} = \{(\omega_1, \omega_2, \omega_3) \in \Omega \mid \omega_3 = 1\}$$

$$E = \{(\omega_1, \omega_2, \omega_3) \in \Omega \mid \omega_1 = \omega_2\}$$

$$F = \{(\omega_1, \omega_2, \omega_3) \in \Omega \mid \omega_2 = \omega_3\}$$

$$G = \{(\omega_1, \omega_2, \omega_3) \in \Omega \mid \omega_1 \neq \omega_3\}$$

(col lieve abuso di notazione di usare d'ora in poi lo stesso simbolo per l'evento e l'insieme che lo rappresenta!). Con questa scelta, si può facilmente verificare che valgono le seguenti relazioni analoghe delle (1.2), (1.3)

$$E = (T_1 \cap T_2) \cup (T_1^c \cap T_2^c)$$

$$E \cap F = E \cap G^c = (T_1 \cap T_2 \cap T_3) \cup (T_1^c \cap T_2^c \cap T_3^c)$$

$$T_1 \cap G = T_1 \cap T_3^c$$

$$(E \cup F) \cap T_2 = (T_1 \cap T_2) \cup (T_2 \cap T_3)$$

$$T_1 \cap T_2 \subseteq T_1$$

$$T_1 \cap T_3^c \subseteq G,$$

mentre le (1.1) danno le rappresentazioni in insiemi

$$\text{“esce croce al primo lancio”} = T_1^c = \{(\omega_1, \omega_2, \omega_3) \in \Omega \mid \omega_1 = 0\}$$

$$\text{“esce sempre testa”} = T_1 \cap T_2 \cap T_3 = \{(1, 1, 1)\}$$

$$\text{“esce la stessa faccia in tutti i lanci”} = E \cap F = \{(1, 1, 1), (0, 0, 0)\}.$$

Da notare che l'elemento $(1, 1, 1)$ di Ω *non rappresenta* un evento, mentre al contrario l'insieme $\{(1, 1, 1)\}$, che è un elemento di $\mathcal{P}(\Omega)$, *rappresenta* un evento. L'evento certo e l'evento impossibile sono rappresentati rispettivamente dall'insieme Ω e dall'insieme vuoto \emptyset .

A questo punto possiamo finalmente introdurre la nozione di probabilità.

Definizione 1. Sia Ω un insieme e sia $\mathcal{P}(\Omega)$ il suo insieme delle parti. Una **probabilità su Ω** è una funzione $\mathbb{P} : \mathcal{P}(\Omega) \rightarrow \mathbb{R}$ con le seguenti proprietà:

- (1) $\mathbb{P}(E) \geq 0$ per ogni evento $E \in \mathcal{P}(\Omega)$;
- (2) $\mathbb{P}(\Omega) = 1$;
- (3) per ogni famiglia $\{E_i\}_{i \in I}$ *finita o numerabile* di eventi $E_i \in \mathcal{P}(\Omega)$ tali che $E_i \cap E_j = \emptyset$ se $i \neq j$, si ha l'uguaglianza $\mathbb{P}\left(\bigcup_{i \in I} E_i\right) = \sum_{i \in I} \mathbb{P}(E_i)$.

Se gli eventi E_1, E_2, \dots soddisfano la condizione $E_i \cap E_j = \emptyset$ per ogni $i \neq j$, si dice che E_1, E_2, \dots sono mutuamente *incompatibili*. L'assioma (3) richiede pertanto che la probabilità dell'unione di eventi incompatibili sia la somma delle probabilità dei singoli eventi.

Di seguito sono riassunte le principali proprietà della probabilità che si possono direttamente ricavare dalla definizione.

Proposizione 1. *Siano $E, F \in \mathcal{P}(\Omega)$ due eventi.*

$$(i) \quad \mathbb{P}(\emptyset) = 0.$$

(ii) Se $E \subseteq F$, allora $\mathbb{P}(F \setminus E) = \mathbb{P}(F) - \mathbb{P}(E)$ (dove l'insieme $F \setminus E := F \cap E^c$ è la differenza di F meno E).

$$(iii) \quad \text{Se } E \subseteq F, \text{ allora } \mathbb{P}(E) \leq \mathbb{P}(F).$$

$$(iv) \quad \mathbb{P}(E^c) = 1 - \mathbb{P}(E).$$

$$(v) \quad \mathbb{P}(E) \leq 1.$$

$$(vi) \quad \mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(E \cap F).$$

Dimostrazione. (i) Si ha $\emptyset \cup \emptyset = \emptyset$ e $\emptyset \cap \emptyset = \emptyset$, dunque per l'assioma (3) della probabilità

$$\mathbb{P}(\emptyset) = \mathbb{P}(\emptyset \cup \emptyset) = \mathbb{P}(\emptyset) + \mathbb{P}(\emptyset) = 2\mathbb{P}(\emptyset) \quad \Rightarrow \quad \mathbb{P}(\emptyset) = 0.$$

(ii) Se $E \subseteq F$, allora $F = (F \setminus E) \cup E$ e $(F \setminus E) \cap E = \emptyset$, dunque, ancora per l'assioma (3),

$$\mathbb{P}(F) = \mathbb{P}(F \setminus E) + \mathbb{P}(E) \quad \Rightarrow \quad \mathbb{P}(F \setminus E) = \mathbb{P}(F) - \mathbb{P}(E).$$

(iii) Se $E \subseteq F$, per il punto precedente e per l'assioma (1)

$$\mathbb{P}(F) - \mathbb{P}(E) = \mathbb{P}(F \setminus E) \geq 0 \quad \Rightarrow \quad \mathbb{P}(F) \geq \mathbb{P}(E).$$

(iv) Si ha $E^c = \Omega \setminus E$ e $\mathbb{P}(\Omega) = 1$ per l'assioma (2), dunque

$$\mathbb{P}(E^c) = \mathbb{P}(\Omega \setminus E) = \mathbb{P}(\Omega) - \mathbb{P}(E) = 1 - \mathbb{P}(E)$$

come conseguenza del punto (ii) (notare che l'ipotesi $E \subseteq \Omega$ è chiaramente soddisfatta).

(v) $E \subseteq \Omega$ e $\mathbb{P}(\Omega) = 1$, dunque $\mathbb{P}(E) \leq 1$ segue dal punto (iii).

(vi) Abbiamo

$$E \cup F = [E \setminus (E \cap F)] \cup (E \cap F) \cup [F \setminus (E \cap F)]$$

e inoltre

$$\begin{aligned} [E \setminus (E \cap F)] \cap (E \cap F) &= \emptyset & (E \cap F) \cap [F \setminus (E \cap F)] &= \emptyset \\ [E \setminus (E \cap F)] \cap [F \setminus (E \cap F)] &= \emptyset. \end{aligned}$$

Ricaviamo pertanto

$$\mathbb{P}(E \cup F) = \mathbb{P}(E \setminus (E \cap F)) + \mathbb{P}(E \cap F) + \mathbb{P}(F \setminus (E \cap F))$$

per l'assioma (3). Applicando il punto (ii) agli insiemi $E \cap F \subseteq E$ e $E \cap F \subseteq F$, abbiamo

$$\begin{aligned}\mathbb{P}(E \setminus (E \cap F)) &= \mathbb{P}(E) - \mathbb{P}(E \cap F) \\ \mathbb{P}(F \setminus (E \cap F)) &= \mathbb{P}(F) - \mathbb{P}(E \cap F)\end{aligned}$$

e quindi, riprendendo l'equazione precedente,

$$\begin{aligned}\mathbb{P}(E \cup F) &= \mathbb{P}(E) - \mathbb{P}(E \cap F) + \mathbb{P}(E \cap F) + \mathbb{P}(F) - \mathbb{P}(E \cap F) \\ &= \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(E \cap F).\end{aligned}$$

■

Il punto (vi) della proposizione precedente si estende facilmente al caso di tre o più eventi. Infatti, iterandolo due volte,

$$\begin{aligned}\mathbb{P}(E \cup F \cup G) &= \mathbb{P}(E \cup (F \cup G)) \\ &= \mathbb{P}(E) + \mathbb{P}(F \cup G) - \mathbb{P}(E \cap (F \cup G)) \\ &= \mathbb{P}(E) + \mathbb{P}(F) + \mathbb{P}(G) - \mathbb{P}(F \cap G) - \mathbb{P}((E \cap F) \cup (E \cap G)) \\ &= \mathbb{P}(E) + \mathbb{P}(F) + \mathbb{P}(G) - \mathbb{P}(F \cap G) - [\mathbb{P}(E \cap F) + \mathbb{P}(E \cap G) - \mathbb{P}((E \cap F) \cap (E \cap G))] \\ &= \mathbb{P}(E) + \mathbb{P}(F) + \mathbb{P}(G) - \mathbb{P}(F \cap G) - \mathbb{P}(E \cap F) - \mathbb{P}(E \cap G) + \mathbb{P}(E \cap F \cap G),\end{aligned}$$

dove inoltre abbiamo usato la proprietà distributiva $E \cap (F \cup G) = (E \cap F) \cup (E \cap G)$. Il caso di $n \geq 4$ eventi è simile.

Osservazione 1. Per motivi di carattere tecnico che diventeranno più chiari nel modulo di Analisi, quando lo spazio campionario Ω ha **cardinalità non numerabile** (p.es., quando $\Omega = \mathbb{R}$, oppure $\Omega = \{0, 1\}^{\mathbb{N}}$) spesso non è necessario definire una probabilità *su tutto* l'insieme delle parti $\mathcal{P}(\Omega)$, ma è invece molto più **conveniente definirla solo su un particolare sottoinsieme $\mathcal{F} \subseteq \mathcal{P}(\Omega)$** . Per poter considerare solo \mathcal{F} (e non tutto $\mathcal{P}(\Omega)$) come la totalità degli eventi del nostro esperimento aleatorio, **il sottoinsieme \mathcal{F} deve naturalmente essere chiuso rispetto alle operazioni insiemistiche \cap , \cup e \cdot^c , corrispondenti delle operazioni logiche \wedge , \vee e \neg** . In altre parole, deve valere che

- (1) $\Omega \in \mathcal{F}$ e $\emptyset \in \mathcal{F}$;
- (2) se $E \in \mathcal{F}$, allora anche $E^c \in \mathcal{F}$;
- (3) per ogni famiglia $\{E_i\}_{i \in I}$ *finita o numerabile* di eventi $E_i \in \mathcal{F}$ (non necessariamente disgiunti) si ha $\bigcup_{i \in I} E_i \in \mathcal{F}$ e anche $\bigcap_{i \in I} E_i \in \mathcal{F}$.

Un sottoinsieme $\mathcal{F} \subseteq \mathcal{P}(\Omega)$ con tali proprietà si chiama **σ -algebra di sottoinsiemi di Ω** . Se inoltre **$\mathbb{P} : \mathcal{F} \rightarrow \mathbb{R}$ è una probabilità su \mathcal{F}** , cioè verifica i tre assiomi della probabilità, ma solo sugli insiemi

di \mathcal{F} anziché su tutto $\mathcal{P}(\Omega)$, la tripla $(\Omega, \mathcal{F}, \mathbb{P})$ si dice *spazio di probabilità*. Il requisito che il punto (3) valga anche quando gli eventi $\{E_i\}_{i \in I}$ sono un'infinita numerabile non è una mera complicazione matematica. Infatti, capita sovente di considerare intersezioni o unioni infinite numerabili di eventi. Per esempio, nell'esperimento aleatorio consistente in infiniti lanci ripetuti di una moneta, l'evento

$$\text{"non esce mai testa"} = \bigcap_{i \in I} T_i^c$$

è di tale tipo.

Esempio 1 (Spazi di probabilità uniforme). Se Ω è un insieme finito, con cardinalità $|\Omega| = N$, una probabilità su Ω è per esempio la seguente funzione $\mathbb{P} : \mathcal{P}(\Omega) \rightarrow \mathbb{R}$

$$\mathbb{P}(E) = \frac{|E|}{N} \quad \text{dove } |E| = \text{cardinalità di } E.$$

Infatti, è banale verificare che \mathbb{P} soddisfa gli assiomi (1) e (2), mentre (3) discende immediatamente dal fatto che *la cardinalità di un'unione di insiemi disgiunti è la somma delle cardinalità dei singoli insiemi*. Tale probabilità si chiama *probabilità uniforme* su Ω . Per esempio, nell'esperimento dei tre lanci consecutivi di una moneta, con $\Omega = \{0, 1\}^3$, abbiamo $|\Omega| = 2^3 = 8$

$$\begin{aligned} \mathbb{P}(T_1) &= \frac{|\{(1, 0, 0), (1, 1, 0), (1, 0, 1), (1, 1, 1)\}|}{8} = \frac{4}{8} = \frac{1}{2} \\ \mathbb{P}(E) &= \frac{|\{(0, 0, 0), (0, 0, 1), (1, 1, 0), (1, 1, 1)\}|}{8} = \frac{4}{8} = \frac{1}{2} \\ \mathbb{P}(G) &= \frac{|\{(1, 0, 0), (1, 1, 0), (0, 0, 1), (0, 1, 1)\}|}{8} = \frac{4}{8} = \frac{1}{2} \end{aligned}$$

e così via.

1.2 Probabilità condizionata

Definizione 2. Sia $(\Omega, \mathcal{F}, \mathbb{P})$ uno spazio di probabilità e siano $E, F \in \mathcal{F}$ due eventi. Supponiamo $\mathbb{P}(F) > 0$. La probabilità di E condizionata a F (o probabilità di E sapendo F) è il numero reale

$$\mathbb{P}(E | F) := \frac{\mathbb{P}(E \cap F)}{\mathbb{P}(F)}.$$

Dal fatto che $E \cap F \subseteq F$ segue $\mathbb{P}(E \cap F) \leq \mathbb{P}(F)$, e dunque $\mathbb{P}(E | F) \in [0, 1]$. Inoltre, è facile verificare che la funzione $\mathbb{P}_F : \mathcal{F} \rightarrow \mathbb{R}$ data da

$$\mathbb{P}_F(E) = \mathbb{P}(E | F) \quad \text{per ogni } E \in \mathcal{F}$$

è a sua volta una probabilità su Ω . Infatti

1. $\mathbb{P}_F(E) \geq 0$ per ogni evento $E \in \mathcal{F}$, in quanto sia $\mathbb{P}(E \cap F) \geq 0$ sia $\mathbb{P}(F) > 0$;
2. $\mathbb{P}_F(\Omega) = \mathbb{P}(\Omega \cap F) / \mathbb{P}(F) = \mathbb{P}(F) / \mathbb{P}(F) = 1$;
3. per ogni famiglia $\{E_i\}_{i \in I}$ finita o numerabile di eventi disgiunti $E_i \in \mathcal{F}$, si ha

$$\begin{aligned} \mathbb{P}_F\left(\bigcup_{i \in I} E_i\right) &= \frac{\mathbb{P}\left(\left(\bigcup_{i \in I} E_i\right) \cap F\right)}{\mathbb{P}(F)} && \text{definizione} \\ &= \frac{\mathbb{P}\left(\bigcup_{i \in I} (E_i \cap F)\right)}{\mathbb{P}(F)} && \text{proprietà distributiva di } \cap \text{ rispetto a } \cup \\ &= \frac{\sum_{i \in I} \mathbb{P}(E_i \cap F)}{\mathbb{P}(F)} && \text{assioma (3)} \\ &= \sum_{i \in I} \mathbb{P}_F(E_i) && \text{definizione.} \end{aligned}$$

Nella terza uguaglianza, abbiamo potuto usare l'assioma (3) perché $(E_i \cap F) \cap (E_j \cap F) = (E_i \cap E_j) \cap F = \emptyset$ se $i \neq j$ per ipotesi.

Esempio 2. Nell'esperimento aleatorio dei tre lanci consecutivi di una moneta, la probabilità che escano tre teste sapendo che nei primi due lanci è uscita la stessa faccia è

$$\begin{aligned} \mathbb{P}(T_1 \cap T_2 \cap T_3 | E) &= \frac{\mathbb{P}(T_1 \cap T_2 \cap T_3 \cap E)}{\mathbb{P}(E)} = \frac{\mathbb{P}(T_1 \cap T_2 \cap T_3)}{\mathbb{P}(E)} && \text{perché } T_1 \cap T_2 \cap T_3 \subseteq E \\ &= \frac{\mathbb{P}(\{(1, 1, 1)\})}{\mathbb{P}(\{(0, 0, 0), (0, 0, 1), (1, 1, 0), (1, 1, 1)\})} = \frac{1/8}{4/8} = \frac{1}{4}. \end{aligned}$$

Capita spesso che la probabilità che un certo evento E si realizzi sia nota solo sotto opportune condizioni a priori, cioè che, anziché conoscere $\mathbb{P}(E)$, siano note solo le probabilità di E condizionate ad una serie di altri eventi F_1, F_2, \dots, F_n . In questo caso, il teorema seguente risulta molto utile per calcolare la probabilità $\mathbb{P}(E)$.

Teorema 1 (Formula delle probabilità totali). Sia $(\Omega, \mathcal{F}, \mathbb{P})$ uno spazio di probabilità, e siano $F_1, F_2, \dots, F_n \in \mathcal{F}$ eventi che formano una **partizione** di Ω , cioè tali che

(a) $F_i \cap F_j = \emptyset$ se $i \neq j$;

(b) $\bigcup_{i=1}^n F_i = \Omega$.

Supponiamo inoltre che $\mathbb{P}(F_i) > 0$ per ogni $i = 1, 2, \dots, n$. Allora per ogni evento $E \in \mathcal{F}$ si ha

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

Dimostrazione. Si ha

$$E = E \cap \Omega = E \cap \bigcup_{i=1}^n F_i = \bigcup_{i=1}^n E \cap F_i$$

e

$$(E \cap F_i) \cap (E \cap F_j) = E \cap (F_i \cap F_j) = \emptyset \quad \text{se } i \neq j.$$

Per il terzo assioma della probabilità

$$\mathbb{P}(E) = \sum_{i=1}^n \mathbb{P}(E \cap F_i) = \sum_{i=1}^n \mathbb{P}(E \mid F_i) \mathbb{P}(F_i)$$

dove si è inoltre usata la definizione $\mathbb{P}(E \mid F_i) = \mathbb{P}(E \cap F_i) / \mathbb{P}(F_i)$. ■

Esempio 3. *Problema:* Supponiamo di avere un mazzo di 40 carte: 20 di queste sono rosse su un lato e nere sull'altro, mentre le altre 20 sono rosse su entrambi i lati. Pesca una carta a caso dal mazzo e la poso sul tavolo. Qual'è la probabilità che esibisca il colore rosso?

Soluzione: Introduciamo gli eventi

B = “la carta pescata è bicolore”

R = “la carta pescata esibisce il colore rosso”.

Sappiamo che

$$\mathbb{P}(B) = \mathbb{P}(B^c) = \frac{20}{40} = \frac{1}{2} \quad \mathbb{P}(R \mid B) = \frac{1}{2} \quad \mathbb{P}(R \mid B^c) = 1. \quad (1.4)$$

Allora i due eventi B, B^c formano una partizione di Ω , dunque si può applicare il teorema precedente e ottenere

$$\mathbb{P}(R) = \mathbb{P}(R \mid B) \mathbb{P}(B) + \mathbb{P}(R \mid B^c) \mathbb{P}(B^c) = \frac{1}{2} \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = \frac{3}{4}.$$

Abbiamo visto che la formula delle probabilità totali permette di calcolare la probabilità di un evento E una volta note le probabilità che E si realizzi sotto opportune condizioni a priori. La *formula di Bayes* data di seguito è utile invece nella situazione opposta, cioè quando abbiamo l'informazione a posteriori che l'evento E si è realizzato, e ci chiediamo con quale probabilità è avvenuto uno degli eventi che condizionavano E .

Teorema 2 (Formula di Bayes). Supponiamo che $F_1, F_2, \dots, F_n \in \mathcal{F}$ sia una partizione di Ω , e che $\mathbb{P}(F_i) > 0$ per ogni $i = 1, 2, \dots, n$. Allora per ogni $E \in \mathcal{F}$ con $\mathbb{P}(E) > 0$ si ha

$$\mathbb{P}(F_k | E) = \frac{\mathbb{P}(E | F_k) \mathbb{P}(F_k)}{\sum_{i=1}^n \mathbb{P}(E | F_i) \mathbb{P}(F_i)} \quad \text{per ogni } k \in \{1, 2, \dots, n\}.$$

Dimostrazione. Si ha

$$\mathbb{P}(F_k | E) = \frac{\mathbb{P}(F_k \cap E)}{\mathbb{P}(E)} = \frac{\mathbb{P}(F_k \cap E)}{\mathbb{P}(F_k)} \frac{\mathbb{P}(F_k)}{\mathbb{P}(E)} = \mathbb{P}(E | F_k) \frac{\mathbb{P}(F_k)}{\mathbb{P}(E)},$$

e $\mathbb{P}(E) = \sum_{i=1}^n \mathbb{P}(E | F_i) \mathbb{P}(F_i)$ per la formula delle probabilità totali. ■

Sottolineiamo che in alcuni testi per formula di Bayes si intende invece la più **semplice relazione**

$$\mathbb{P}(F | E) = \mathbb{P}(E | F) \frac{\mathbb{P}(F)}{\mathbb{P}(E)},$$

che è stata provata nel corso della dimostrazione precedente.

Esempio 4. Problema: Nell'Esempio 3 del mazzo di carte colorate, supponiamo ora che la carta pescata a caso dal mazzo e posata sul tavolo esibisca il colore rosso. Qual è la probabilità che l'altro suo lato (quello nascosto) sia nero?

Soluzione: Intuitivamente, verrebbe da rispondere che la probabilità è pari a $1/2$. Invece, applicando la formula di Bayes, si trova che la risposta corretta è

$$\mathbb{P}(B | R) = \frac{\mathbb{P}(R | B) \mathbb{P}(B)}{\mathbb{P}(R | B) \mathbb{P}(B) + \mathbb{P}(R | B^c) \mathbb{P}(B^c)} = \frac{1/2 \cdot 1/2}{1/2 \cdot 1/2 + 1 \cdot 1/2} = \frac{1}{3}.$$

Osservazione: Nella soluzione di questo problema, come anche in quella dell'Esempio 3, non si è mai dovuto ricorrere a una rappresentazione esplicita in uno spazio di probabilità degli eventi coinvolti, ma si è semplicemente utilizzata la conoscenza delle probabilità (1.4). Se lo si vuole, gli eventi R, R^c, B, B^c possono essere p.es. rappresentati nell'insieme $\Omega = \{0, 1\}^2$ con probabilità \mathbb{P} uniforme, ponendo

$$B = \{(0, 0), (0, 1)\} \quad R = \{(1, 0), (0, 1), (1, 1)\}$$

(verificare per esercizio che in questo modo si ottengono le probabilità (1.4)!). Tuttavia questo fatto non dà nessuna ulteriore informazione o semplificazione del problema, ma anzi lo complica inutilmente.

Un'altra situazione in cui la probabilità condizionata si rivela molto utile è quando si vuole conoscere la probabilità dell'intersezione di serie di eventi E_1, E_2, \dots, E_n che si verificano in successione uno dopo l'altro, in modo che ciascun evento E_i viene influenzato solo dagli eventi $E_{i-1}, E_{i-2}, \dots, E_1$ che sono avvenuti prima di esso. In tal caso, infatti, si può utilizzare il prossimo teorema.

Teorema 3 (Formula del prodotto). *Supponiamo che $E_1, E_2, \dots, E_n \in \mathcal{F}$ siano eventi qualsiasi. Allora*

$$\mathbb{P}(E_1 \cap E_2 \cap \dots \cap E_n) = \mathbb{P}(E_n \mid E_{n-1} \cap E_{n-2} \cap \dots \cap E_1) \mathbb{P}(E_{n-1} \mid E_{n-2} \cap E_{n-3} \cap \dots \cap E_1) \dots \cdot \mathbb{P}(E_1).$$

Dimostrazione. Per la definizione di probabilità condizionata

$$\mathbb{P}(E_1 \cap E_2 \cap \dots \cap E_n) = \mathbb{P}(E_n \mid E_{n-1} \cap E_{n-2} \cap \dots \cap E_1) \mathbb{P}(E_{n-1} \cap E_{n-2} \cap \dots \cap E_1)$$

e la formula del teorema ne segue per induzione. ■

Esempio 5. *Problema:* Un'urna contiene 10 palline, di cui 2 sono bianche e 8 sono nere. Estraiamo una dopo l'altra 3 palline, senza rimettere nell'urna nessuna delle palline estratte (*estrazione senza reimmissione*). Qual è la probabilità che tutte le palline estratte siano nere?

Soluzione: Indichiamo con N_i l'evento

$$N_i = \text{"l}'i\text{-esima pallina estratta è nera"} \quad (i = 1, 2, 3).$$

Vogliamo calcolare la probabilità dell'intersezione $N_1 \cap N_2 \cap N_3$. Per il teorema precedente

$$\mathbb{P}(N_1 \cap N_2 \cap N_3) = \mathbb{P}(N_3 \mid N_2 \cap N_1) \mathbb{P}(N_2 \mid N_1) \mathbb{P}(N_1).$$

D'altra parte,

$$\begin{aligned} \mathbb{P}(N_1) &= \frac{8}{8+2} && \text{perché all'inizio nell'urna ci sono 8 palline nere e 2 bianche} \\ \mathbb{P}(N_2 \mid N_1) &= \frac{7}{7+2} && \text{perché, se la prima pallina estratta è nera, allora per la seconda estrazione nell'urna ne restano 7 nere e 2 bianche} \\ \mathbb{P}(N_3 \mid N_2 \cap N_1) &= \frac{6}{6+2} && \text{perché, se le prime due palline estratte sono nere, allora per la terza estrazione nell'urna ne restano 6 nere e 2 bianche.} \end{aligned}$$

Pertanto la probabilità cercata è

$$\mathbb{P}(N_1 \cap N_2 \cap N_3) = \frac{8}{10} \cdot \frac{7}{9} \cdot \frac{6}{8} = \frac{7}{15}.$$

1.3 Indipendenza

La definizione di indipendenza per *due* eventi è molto semplice.

Definizione 3. Sia $(\Omega, \mathcal{F}, \mathbb{P})$ uno spazio di probabilità. Due eventi $E, F \in \mathcal{F}$ si dicono *indipendenti* se $\mathbb{P}(E \cap F) = \mathbb{P}(E) \mathbb{P}(F)$.

Esempio 6. Nell'esempio dei tre lanci di una stessa moneta, è facile verificare che ciascuna coppia di eventi

$$\begin{array}{ccccccc} T_i, T_j & \text{con } i \neq j & T_i, E & T_i, F & T_i, G \\ & & E, F & E, G & F, G \end{array}$$

sono indipendenti. Infatti, p.es.

$$\begin{aligned} \mathbb{P}(T_1 \cap T_2) &= \mathbb{P}(\{(1, 1, 0), (1, 1, 1)\}) = \frac{2}{8} \equiv \frac{1}{2} \cdot \frac{1}{2} = \mathbb{P}(T_1) \mathbb{P}(T_2) \\ \mathbb{P}(T_1 \cap G) &= \mathbb{P}(\{(1, 0, 0), (1, 1, 0)\}) = \frac{2}{8} \equiv \frac{1}{2} \cdot \frac{1}{2} = \mathbb{P}(T_1) \mathbb{P}(G) \\ \mathbb{P}(E \cap G) &= \mathbb{P}(\{(1, 1, 0), (0, 0, 1)\}) = \frac{2}{8} \equiv \frac{1}{2} \cdot \frac{1}{2} = \mathbb{P}(E) \mathbb{P}(G) \end{aligned}$$

e così via per tutte le altre coppie. Un esempio di due eventi *non* indipendenti è la coppia $T_1 \cap T_2$ e E . Infatti

$$\mathbb{P}((T_1 \cap T_2) \cap E) = \mathbb{P}(T_1 \cap T_2 \cap E) = \mathbb{P}(\{(1, 1, 0), (1, 1, 1)\}) = \frac{2}{8} \neq \frac{1}{4} \cdot \frac{1}{2} = \mathbb{P}(T_1 \cap T_2) \mathbb{P}(E)$$

Notiamo che se E e F sono eventi indipendenti, allora sono indipendenti anche gli eventi in ciascuna coppia

$$E, F^c \quad E^c, F \quad E^c, F^c.$$

Infatti

$$\begin{aligned} \mathbb{P}(E \cap F^c) &= \mathbb{P}(E \setminus (E \cap F)) = \mathbb{P}(E) - \mathbb{P}(E \cap F) = \mathbb{P}(E) - \mathbb{P}(E) \mathbb{P}(F) = \mathbb{P}(E) (1 - \mathbb{P}(F)) \\ &\equiv \mathbb{P}(E) \mathbb{P}(F^c) \\ \mathbb{P}(E^c \cap F) &= \mathbb{P}(F \cap E^c) \equiv \mathbb{P}(F) \mathbb{P}(E^c) \quad \text{per il caso precedente} \end{aligned}$$

e, applicando il secondo caso agli eventi E e F^c che sono indipendenti per il primo,

$$\mathbb{P}(F^c \cap E^c) \equiv \mathbb{P}(F^c) \mathbb{P}(E^c).$$

Osservazione 2. Dire che due eventi E e F sono *indipendenti* è una cosa completamente diversa dall'affermare che sono *disgiunti* (o anche detti *incompatibili*), cioè che $E \cap F = \emptyset$. Infatti,

- se E e F sono indipendenti, allora $\mathbb{P}(E \cap F) = \mathbb{P}(E) \mathbb{P}(F)$;
- se E e F sono disgiunti, allora $\mathbb{P}(E \cap F) = \mathbb{P}(\emptyset) = 0$.

Le due cose possono avvenire contemporaneamente se e solo se $\mathbb{P}(E) = 0$ o $\mathbb{P}(F) = 0$ (in particolare, se $E = \emptyset$ o $F = \emptyset$), mentre in tutti gli altri casi indipendenza e incompatibilità sono due nozioni totalmente distinte.

La definizione di indipendenza per tre o più eventi è un po' più complessa, e richiede di considerare *tutte* le intersezioni possibili degli eventi.

Definizione 4. Sia $(\Omega, \mathcal{F}, \mathbb{P})$ uno spazio di probabilità. Una collezione di n eventi $E_1, E_2, \dots, E_n \in \mathcal{F}$ si dicono *indipendenti* se

$$\mathbb{P}(E_{i_1} \cap E_{i_2} \cap \dots \cap E_{i_k}) = \mathbb{P}(E_{i_1}) \mathbb{P}(E_{i_2}) \dots \mathbb{P}(E_{i_k}) \quad (1.5)$$

per ogni $k \leq n$ e per ogni sottoinsieme di indici $\{i_1, i_2, \dots, i_k\} \subseteq \{1, 2, \dots, n\}$.

Per chiarire la definizione precedente, applichiamola a tre eventi E, F, G . In tal caso, affinché E, F, G siano indipendenti *non basta* che valga $\mathbb{P}(E \cap F) = \mathbb{P}(E) \mathbb{P}(F)$, $\mathbb{P}(E \cap G) = \mathbb{P}(E) \mathbb{P}(G)$ e $\mathbb{P}(F \cap G) = \mathbb{P}(F) \mathbb{P}(G)$, ma deve anche essere $\mathbb{P}(E \cap F \cap G) = \mathbb{P}(E) \mathbb{P}(F) \mathbb{P}(G)$.

Esempio 7. Nell'esperimento aleatorio dei tre lanci di una moneta, i tre eventi T_1, E, F sono indipendenti. Infatti, abbiamo già visto che sono indipendenti a coppie. In più, abbiamo

$$\mathbb{P}(T_1 \cap E \cap F) = \mathbb{P}(\{(1, 1, 1)\}) = \frac{1}{8} \equiv \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \mathbb{P}(T_1) \mathbb{P}(E) \mathbb{P}(F).$$

Non sono invece indipendenti i tre eventi T_1, T_2, E . Infatti, benché siano indipendenti a coppie, si ha tuttavia

$$\mathbb{P}(T_1 \cap T_2 \cap E) = \mathbb{P}(\{(1, 1, 0), (1, 1, 1)\}) = \frac{2}{8} \neq \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \mathbb{P}(T_1) \mathbb{P}(T_2) \mathbb{P}(E).$$

Nel caso di n eventi, nella definizione di indipendenza le condizioni (1.5) sono in tutto $2^n - n - 1$, molte di più delle $\binom{n}{2}$ condizioni che coinvolgono solo le singole coppie di eventi.

Pur essendo complicata, la definizione di indipendenza per $n \geq 3$ eventi ha il pregio seguente: se E_1, E_2, \dots, E_n sono eventi indipendenti, allora, raggruppandoli in gruppi più piccoli e combinando gli eventi in ogni gruppo tramite le operazioni di \cap , \cup e c , le combinazioni provenienti da gruppi diversi continuano a essere fra loro indipendenti. Più formalmente, per ogni scelta di indici $\{i_1, i_2, \dots, i_k\} \subseteq \{1, 2, \dots, n\}$ con $1 = i_1 < i_2 < \dots < i_k = n$, se F_h è una combinazione degli eventi $E_{i_h}, E_{i_h+1}, E_{i_h+2}, \dots, E_{i_{h+1}}$, allora gli eventi F_1, F_2, \dots, F_{k-1} sono a loro volta indipendenti.

Per chiarire le idee, prendiamo il caso con $n = 3$, e supponiamo che E, F, G siano tre eventi indipendenti. Allora sono indipendenti anche le coppie

$$E \cap F, G \quad E \cup F, G \quad E^c, F \cup G^c \quad F, E \setminus G \quad \text{e così via.}$$

Infatti, prendiamo p.es. la prima coppia:

$$\mathbb{P}((E \cap F) \cap G) = \mathbb{P}(E \cap F \cap G) = \mathbb{P}(E) \mathbb{P}(F) \mathbb{P}(G) = \mathbb{P}(E \cap F) \mathbb{P}(G).$$

Un po' più complicato è dimostrare l'indipendenza della seconda:

$$\begin{aligned}
 \mathbb{P}((E \cup F) \cap G) &= \mathbb{P}((E \cap G) \cup (F \cap G)) && \text{(propr. distributiva di } \cap \text{ rispetto a } \cup) \\
 &= \mathbb{P}(E \cap G) + \mathbb{P}(F \cap G) - \mathbb{P}(E \cap F \cap G) && \text{(formula nota)} \\
 &= \mathbb{P}(E) \mathbb{P}(G) + \mathbb{P}(F) \mathbb{P}(G) - \mathbb{P}(E) \mathbb{P}(F) \mathbb{P}(G) && \text{(indipendenza)} \\
 &= (\mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(E) \mathbb{P}(F)) \mathbb{P}(G) \\
 &= (\mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(E \cap F)) \mathbb{P}(G) && \text{(indipendenza)} \\
 &\equiv \mathbb{P}(E \cup F) \mathbb{P}(G) && \text{(formula nota).}
 \end{aligned}$$

Notiamo infine che le definizioni di **indipendenza** e **speranza condizionata** sono in relazione fra loro nel modo che ci si aspetta: se E e F sono eventi tra loro indipendenti, allora la conoscenza a priori che si è realizzato F non cambia la probabilità che si realizzi E . In formule,

$$\mathbb{P}(E | F) = \mathbb{P}(E) \quad \text{se } E \text{ e } F \text{ sono indipendenti.}$$

Infatti, se E e F sono indipendenti, allora

$$\mathbb{P}(E | F) = \frac{\mathbb{P}(E \cap F)}{\mathbb{P}(F)} = \frac{\mathbb{P}(E) \mathbb{P}(F)}{\mathbb{P}(F)} = \mathbb{P}(E).$$

1.4 Prove di Bernoulli

In un esperimento aleatorio, supponiamo di **ripetere la medesima prova** (p.es., il lancio di una moneta, oppure la puntata sullo stesso numero della roulette...) per n volte, in modo che

- (a) **ciascuna prova non influenza le altre;**
- (b) ciascuna prova ha probabilità di **successo pari a $p \in [0, 1]$** (la stessa per tutte le prove).

Denotiamo con E_1, E_2, \dots, E_n gli eventi

$$E_i = \text{“la } i\text{-esima prova ha avuto successo”}.$$

Allora

- (a) gli eventi E_1, E_2, \dots, E_n sono indipendenti;
- (b) $\mathbb{P}(E_i) = p$ per ogni $i = 1, 2, \dots, n$.

Una sequenza di eventi con tali due proprietà si chiama **successione di prove di Bernoulli**.

Se $k \leq n$ denotiamo con B_k l'evento

$$B_k = \text{“si sono realizzati esattamente } k \text{ successi”}.$$

Si può scrivere B_k come l'unione

$$B_k = \bigcup_{\substack{I \subseteq \{1, 2, \dots, n\} \\ |I| = k}} \left[\left(\bigcap_{i \in I} E_i \right) \cap \left(\bigcap_{j \in I^c} E_j^c \right) \right].$$

In altre parole, **B_k è l'unione di tutti gli eventi in cui i successi si realizzano nelle prove $I = \{i_1, i_2, \dots, i_k\}$ e non si realizzano nelle rimanenti prove $I^c = \{j_1, j_2, \dots, j_{n-k}\}$** al variare dell'insieme di indici I in $\{1, 2, \dots, n\}$. Nell'insieme di n indici $\{1, 2, \dots, n\}$ sono possibili esattamente

$$\binom{n}{k} := \frac{n!}{k!(n-k)!}$$

scelte diverse del sottoinsieme I con k elementi. Inoltre, abbiamo

$$\left[\left(\bigcap_{i \in I} E_i \right) \cap \left(\bigcap_{j \in I^c} E_j^c \right) \right] \cap \left[\left(\bigcap_{i \in I'} E_i \right) \cap \left(\bigcap_{j \in I'^c} E_j^c \right) \right] = \emptyset \quad \text{se } I \neq I'.$$

Pertanto

$$\begin{aligned}
\mathbb{P}(B_k) &= \sum_{\substack{I \subseteq \{1,2,\dots,n\} \\ |I|=k}} \mathbb{P}\left(\left(\bigcap_{i \in I} E_i\right) \cap \left(\bigcap_{j \in I^c} E_j^c\right)\right) && \text{per l'assioma (3)} \\
&= \sum_{\substack{I \subseteq \{1,2,\dots,n\} \\ |I|=k}} \prod_{i \in I} \mathbb{P}(E_i) \prod_{j \in I^c} \mathbb{P}(E_j^c) && \text{per l'indipendenza} \\
&= \sum_{\substack{I \subseteq \{1,2,\dots,n\} \\ |I|=k}} p^k (1-p)^{n-k} && \text{perché } \mathbb{P}(E_i) = p, \mathbb{P}(E_i^c) = 1-p \\
&= \binom{n}{k} p^k (1-p)^{n-k}.
\end{aligned}$$

Osservazione 3. In tutta la discussione precedente, non abbiamo mai rappresentato esplicitamente gli eventi E_1, E_2, \dots, E_n in uno spazio di probabilità $(\Omega, \mathcal{F}, \mathbb{P})$, ma **abbiamo implicitamente assunto che esista sempre una tale rappresentazione**. In effetti, ciò è vero: per esempio, si può scegliere

$$\begin{aligned}
\Omega &= \{0, 1\}^n & \mathcal{F} &= \mathcal{P}(\Omega) \\
\mathbb{P}(\{(\omega_1, \omega_2, \dots, \omega_n)\}) &= p^{\sum_{k=1}^n \omega_k} (1-p)^{n-\sum_{k=1}^n \omega_k}
\end{aligned}$$

e rappresentare gli eventi E_i negli insiemi

$$E_i = \{(\omega_1, \omega_2, \dots, \omega_n) \in \Omega \mid \omega_i = 1\}.$$

Lasciamo come esercizio la verifica del fatto che con questa scelta gli eventi E_1, E_2, \dots, E_n sono indipendenti e $\mathbb{P}(E_i) = p$ per ogni $i = 1, 2, \dots, n$.

1.5 Variabili aleatorie

Rappresentare gli eventi di un esperimento aleatorio tramite sottoinsiemi di un opportuno spazio campionario Ω ha il seguente grosso vantaggio: su Ω si possono definire delle *funzioni reali* $X : \Omega \rightarrow \mathbb{R}$, ciascuna delle quali descrive il risultato di una *misura*.

Esempio 8. Per fissare le idee, consideriamo ancora una volta l'esperimento dei tre lanci di una moneta, con $\Omega = \{0, 1\}^3$ e l'usuale rappresentazione degli eventi. Se definiamo la funzione $X : \Omega \rightarrow \mathbb{R}$ data da

$$X(\omega_1, \omega_2, \omega_3) = \omega_1 + \omega_2 + \omega_3 \quad \text{per ogni } (\omega_1, \omega_2, \omega_3) \quad (1.6)$$

vediamo subito che X è la funzione che ‘conta’ o ‘misura’ il numero di volte in cui esce testa nei tre lanci.

Per il loro legame con il concetto di misura, le funzioni reali sullo spazio campionario Ω meritano un nome tutto loro.

Definizione 5. Supponendo per semplicità che $\mathcal{F} \equiv \mathcal{P}(\Omega)$, una qualunque funzione $X : \Omega \rightarrow \mathbb{R}$ si dice *variabile aleatoria*.

Notiamo che la controimmagine di un insieme $A \subseteq \mathbb{R}$ rispetto a una variabile aleatoria X è un *evento*:

$$X^{-1}(A) := \{\omega \in \Omega \mid X(\omega) \in A\} \in \mathcal{P}(\Omega) \quad \text{per ogni } A \subseteq \mathbb{R}.$$

Esempio 9. Sempre nell'esperimento dei tre lanci di una moneta e considerando la variabile aleatoria X definita nell'equazione (1.6), abbiamo le uguaglianze di eventi

$$\begin{aligned} X^{-1}(\{0\}) &= \{(0, 0, 0)\} = \text{“non è mai uscita testa”} \\ X^{-1}(\{1\}) &= \{(1, 0, 0), (0, 1, 0), (0, 0, 1)\} = \text{“è uscita testa una volta sola”} \\ X^{-1}(\{2\}) &= (T_1 \cap T_2 \cap T_3^c) \cup (T_1 \cap T_2^c \cap T_3) \cup (T_1^c \cap T_2 \cap T_3) \\ X^{-1}(\{3\}) &= T_1 \cap T_2 \cap T_3. \end{aligned}$$

e ancora

$$\begin{aligned} X^{-1}(\{2, 3\}) &= \{(1, 1, 0), (1, 0, 1), (0, 1, 1), (1, 1, 1)\} = \text{“è uscita testa almeno due volte”} \\ X^{-1}(\{0, 1, 2, 3\}) &= \Omega \\ X^{-1}(\{4\}) &= \emptyset \end{aligned}$$

e così via.

Per semplicità di scrittura, d'ora in poi useremo le notazioni

$$\begin{aligned} \{X \in A\} &= X^{-1}(A) \quad \text{per ogni sottoinsieme } A \subseteq \mathbb{R} \\ \{X = a\} &= X^{-1}(\{a\}) = \{\omega \in \Omega \mid X(\omega) = a\} \quad \text{per ogni numero } a \in \mathbb{R} \\ \{a < X \leq b\} &= X^{-1}((a, b]) \quad \text{per ogni coppia di numeri } a, b \in \mathbb{R} \end{aligned}$$

e simili. Sottolineiamo ancora una volta che **tutti gli insiemi precedenti sono *eventi*, cioè sottoinsiemi di Ω .** Quando tuttavia la σ -algebra \mathcal{F} non coincide con tutto $\mathcal{P}(\Omega)$, come si è supposto finora, **si aggiunge** nella definizione di variabile aleatoria il **requisito che $\{X \leq a\} \in \mathcal{F}$ per ogni $a \in \mathbb{R}$.** Tale requisito è infatti sufficiente a garantire che gli insiemi $\{X \in A\}$ stiano in \mathcal{F} per una vasta scelta di sottoinsiemi $A \subseteq \mathbb{R}$ (e non solo per insiemi della forma $A = (-\infty, a]$). A ogni modo, non entreremo nel dettaglio di questo fatto, anche perché d'ora in poi tutte le funzioni $X : \Omega \rightarrow \mathbb{R}$ con cui avremo a che fare soddisferanno automaticamente la condizione $\{X \leq a\} \in \mathcal{F}$ per ogni $a \in \mathbb{R}$.

Una variabile aleatoria X si dice **discreta** se esiste un sottoinsieme discreto $S \subset \mathbb{R}$ tale che $\mathbb{P}(\{X \in S\}) = 1$. In tal caso, posto

$$p_X(x) := \mathbb{P}(\{X = x\}) \quad \text{per ogni } x \in S,$$

la funzione $p_X : S \rightarrow [0, 1]$ si chiama **densità** (discreta) della variabile aleatoria X . Quando la densità p_X è nota, si possono calcolare tutte le probabilità di eventi del tipo $\{X \in A\}$ con $A \subseteq \mathbb{R}$. Infatti,

$$\begin{aligned} \mathbb{P}(\{X \in A\}) &= \mathbb{P}(\{X \in A \cap S\} \cup \{X \in A \cap S^c\}) \\ &= \mathbb{P}(\{X \in A \cap S\}) + \mathbb{P}(\{X \in A \cap S^c\}) \quad \text{per l'assioma (3)} \\ &= \mathbb{P}(\{X \in A \cap S\}) \quad \text{perché } \mathbb{P}(\{X \in A \cap S^c\}) \leq \mathbb{P}(\{X \in S^c\}) = 0 \\ &= \mathbb{P}\left(\bigcup_{x \in A \cap S} \{X = x\}\right) \quad \text{unione finita o numerabile di insiemi disgiunti} \\ &= \sum_{x \in A \cap S} \mathbb{P}(\{X = x\}) \quad \text{per l'assioma (3)} \\ &= \sum_{x \in A \cap S} p_X(x) \end{aligned}$$

e quindi si possono calcolare anche le probabilità

$$\mathbb{P}(\{X \leq a\}) = \sum_{\substack{x \in S \\ x \leq a}} p_X(x) \quad \mathbb{P}(\{a < X \leq b\}) = \sum_{\substack{x \in S \\ a < x \leq b}} p_X(x)$$

ecc.. **D'ora in poi, per evitare un'inutile proliferazione di parentesi che rendono illeggibili le formule, con lieve abuso di notazione scriveremo**

$$\mathbb{P}(X \in A) := \mathbb{P}(\{X \in A\}), \quad \mathbb{P}(X \leq a) := \mathbb{P}(\{X \leq a\}) \quad \dots$$

Si notino le due seguenti **proprietà fondamentali della densità p_X** :

- (a) p_X è una **funzione positiva**, cioè $p_X(x) \geq 0$ per ogni $x \in S$, in quanto $p_X(x) = \mathbb{P}(X = x)$ è il **valore preso da una probabilità**;
- (b) p_X è **normalizzata**, cioè $\sum_{x \in S} p_X(x) = 1$, perché $\sum_{x \in S} p_X(x) = \mathbb{P}(X \in S) = 1$.

Esempio 10. Consideriamo l'esperimento aleatorio consistente nel lancio di due dadi a sei facce equilibrati. Rappresentiamo gli eventi di questo esperimento nello spazio campionario

$$\Omega = \{1, 2, 3, 4, 5, 6\}^2 = \{(\omega_1, \omega_2) \mid \omega_i \in \{1, 2, 3, 4, 5, 6\}\},$$

associando agli eventi

$$\begin{aligned} E_i &= \text{"al primo lancio è uscito } i\text{"} \\ F_i &= \text{"al secondo lancio è uscito } i\text{"} \end{aligned} \quad (i = 1, 2, \dots, 6)$$

gli insiemi

$$\begin{aligned} E_i &= \{(\omega_1, \omega_2) \in \Omega \mid \omega_1 = i\} \\ F_i &= \{(\omega_1, \omega_2) \in \Omega \mid \omega_2 = i\}. \end{aligned}$$

Supponiamo che i due dadi siano equilibrati, e perciò consideriamo su Ω la probabilità \mathbb{P} uniforme. Siano X , Y e Z le variabili aleatorie

$$\begin{aligned} X &= \text{"risultato del primo lancio"} \\ Y &= \text{"risultato del secondo lancio"} \\ Z &= \text{"risultato più alto dei due lanci"}. \end{aligned}$$

Con la nostra rappresentazione, X , Y e Z sono le funzioni da Ω in \mathbb{R} date da

$$X(\omega_1, \omega_2) = \omega_1 \quad Y(\omega_1, \omega_2) = \omega_2 \quad Z(\omega_1, \omega_2) = \max\{\omega_1, \omega_2\}.$$

Le variabili aleatorie X , Y e Z sono discrete e prendono tutte e tre valori nell'insieme $S = \{1, 2, 3, 4, 5, 6\}$. Le loro densità sono date da

$$\begin{aligned} p_X(x) &= \mathbb{P}(X = x) = \mathbb{P}(X^{-1}(\{x\})) \\ &= \begin{cases} \mathbb{P}(\{(x, \omega_2) \mid \omega_2 \in \{1, 2, 3, 4, 5, 6\}\}) & \text{se } x \in \{1, 2, 3, 4, 5, 6\} \\ \mathbb{P}(\emptyset) & \text{se } x \notin \{1, 2, 3, 4, 5, 6\} \end{cases} \\ &= \begin{cases} \frac{|\{(x, \omega_2) \mid \omega_2 \in \{1, 2, 3, 4, 5, 6\}\}|}{|\Omega|} = \frac{6}{36} = \frac{1}{6} & \text{se } x \in \{1, 2, 3, 4, 5, 6\} \\ 0 & \text{se } x \notin \{1, 2, 3, 4, 5, 6\} \end{cases} \\ p_Y(y) &= \mathbb{P}(Y^{-1}(\{y\})) = \begin{cases} \mathbb{P}(\{(\omega_1, y) \mid \omega_1 \in \{1, 2, 3, 4, 5, 6\}\}) = \frac{1}{6} & \text{se } y \in \{1, 2, 3, 4, 5, 6\} \\ \mathbb{P}(\emptyset) = 0 & \text{se } y \notin \{1, 2, 3, 4, 5, 6\} \end{cases} \\ p_Z(z) &= \mathbb{P}(Z^{-1}(\{z\})) = \begin{cases} \mathbb{P}(\{(\omega_1, \omega_2) \mid \max\{\omega_1, \omega_2\} = z\}) = \frac{2z-1}{36} & \text{se } z \in \{1, 2, 3, 4, 5, 6\} \\ \mathbb{P}(\emptyset) = 0 & \text{se } z \notin \{1, 2, 3, 4, 5, 6\} \end{cases} \end{aligned}$$

(per calcolare la densità di Z abbiamo usato l'uguaglianza di eventi

$$\{(\omega_1, \omega_2) \mid \max\{\omega_1, \omega_2\} = z\} = \{(\omega_1, z) \mid \omega_1 < z\} \cup \{(z, \omega_2) \mid \omega_2 < z\} \cup \{(z, z)\}$$

in cui l'unione è disgiunta). Si osservi che, benché X e Y sono due variabili aleatorie diverse, esse tuttavia hanno la stessa densità.

Di seguito sono riportati alcuni esempi di **densità discrete** di uso molto frequente.

Esempio 11 (Densità **bernoulliana**). Sia E un evento con probabilità $\mathbb{P}(E) = p$, e sia $\mathbb{1}_E : \Omega \rightarrow \mathbb{R}$ la seguente **funzione indicatrice di E**

$$\mathbb{1}_E(\omega) = \begin{cases} 1 & \text{se } \omega \in E \\ 0 & \text{se } \omega \in E^c \end{cases}$$

Si vede subito che $\mathbb{1}_E$ è una variabile aleatoria **discreta che può prendere solo i due valori $S = \{0, 1\}$** . La sua densità è

$$\begin{aligned} p_{\mathbb{1}_E}(1) &= \mathbb{P}(\mathbb{1}_E = 1) = \mathbb{P}(E) = p \\ p_{\mathbb{1}_E}(0) &= \mathbb{P}(\mathbb{1}_E = 0) = \mathbb{P}(E^c) = 1 - \mathbb{P}(E) = 1 - p \end{aligned}$$

in quanto valgono le ovvie uguaglianze di eventi $E = \{\mathbb{1}_E = 1\}$ e $E^c = \{\mathbb{1}_E = 0\}$. La densità $p_{\mathbb{1}_E}$ così trovata si chiama **densità bernoulliana di parametro p** , e si denota con $B(1, p)$. Per indicare che la variabile aleatoria $\mathbb{1}_E$ ha tale densità, si scrive **$\mathbb{1}_E \sim B(1, p)$** .

Esempio 12 (Densità **uniforme discreta**). Sia $S = \{m, m+1, m+2, \dots, n-1, n\}$ un sottoinsieme dei numeri naturali. Una variabile aleatoria discreta X che prende valori in S e ha densità

$$p_X(x) = \begin{cases} \frac{1}{n-m+1} & \text{se } x \in S \\ 0 & \text{altrimenti} \end{cases}$$

si dice che ha **densità uniforme** sull'insieme S e si scrive **$X \sim \mathcal{U}(S)$** . Per esempio, le variabili aleatorie X e Y dell'Esempio 10 hanno entrambe densità uniforme sull'insieme $S = \{1, 2, 3, 4, 5, 6\}$.

Esempio 13 (Densità **binomiale**). Supponiamo che gli eventi E_1, E_2, \dots, E_n formino una successione di prove di Bernoulli con $\mathbb{P}(E_i) = p$, e indichiamo con X la variabile aleatoria

$$X = \mathbb{1}_{E_1} + \mathbb{1}_{E_2} + \dots + \mathbb{1}_{E_n},$$

intendendo con tale espressione che la **funzione X è la somma delle funzioni indicatrici $\mathbb{1}_{E_i}$** . In altre parole,

$$X(\omega) = \mathbb{1}_{E_1}(\omega) + \mathbb{1}_{E_2}(\omega) + \dots + \mathbb{1}_{E_n}(\omega) \quad \text{per ogni } \omega \in \Omega.$$

Allora X prende valori nell'insieme $S = \{0, 1, \dots, n\}$, ed è la funzione che **‘conta’ il numero di successi ottenuti nelle n prove ripetute**. Abbiamo pertanto l'uguaglianza di eventi

$$\{X = k\} = B_k = \text{“si sono realizzati esattamente } k \text{ successi”}$$

e la densità di X è

$$p_X(k) = \mathbb{P}(X = k) = \mathbb{P}(B_k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{per ogni } k = 1, 2, \dots, n$$

per quanto visto nella sezione 1.4. La densità precedente si chiama **densità binomiale di parametri n e p** , e si indica con **$B(n, p)$** . Si scrive inoltre **$X \sim B(n, p)$** .

Esempio 14 (Densità geometrica). Supponiamo di avere una successione infinita di prove di Bernoulli E_1, E_2, \dots , e indichiamo con T la variabile aleatoria

T = numero della prova in cui si ottiene il primo successo.

Per vedere che T è effettivamente una funzione $T : \Omega \rightarrow \mathbb{R}$, possiamo riscriverla esplicitamente come

$$T(\omega) = \inf\{n \geq 0 \mid \omega \in E_n\} \quad \text{per ogni } \omega \in \Omega.$$

La variabile aleatoria T prende valori nell'insieme $S = \mathbb{N}$, e si ha l'ovvia uguaglianza di eventi

$$\{T = k\} = E_1^c \cap E_2^c \cap E_3^c \cap \dots \cap E_{k-1}^c \cap E_k \quad \text{per ogni } k \in \mathbb{N}.$$

La sua densità è pertanto

$$\begin{aligned} p_T(k) &= \mathbb{P}(T = k) = \mathbb{P}(E_1^c \cap E_2^c \cap E_3^c \cap \dots \cap E_{k-1}^c \cap E_k) \\ &= \mathbb{P}(E_1^c) \mathbb{P}(E_2^c) \mathbb{P}(E_3^c) \dots \mathbb{P}(E_{k-1}^c) \mathbb{P}(E_k) \quad \text{per l'indipendenza degli } E_i \\ &= (1-p)^{k-1}p \quad \text{perché } \mathbb{P}(E_i) = p \text{ e } \mathbb{P}(E_i^c) = 1-p. \end{aligned}$$

La densità

$$p_T(k) = (1-p)^{k-1}p \quad \text{per ogni } k \in \mathbb{N}$$

si chiama *densità geometrica* di parametro p e si indica con $\mathcal{G}(p)$.

Fin qui abbiamo considerato solo variabili aleatorie discrete, che possono prendere un numero finito o numerabile di valori. Una variabile aleatoria X si dice invece *assolutamente continua* se esiste una funzione $f_X : \mathbb{R} \rightarrow \mathbb{R}$ tale che

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x) dx \quad \text{per ogni } a, b \in \mathbb{R} \cup \{+\infty, -\infty\} \text{ con } a < b.$$

La funzione f_X , detta anche in questo caso *densità di X* , soddisfa le due proprietà seguenti, analoghe delle corrispondenti proprietà delle densità discrete:

- (a) f_X è una *funzione positiva*, cioè $f_X(x) \geq 0$ per ogni $x \in \mathbb{R}$, in quanto $\int_a^b f_X(x) dx = \mathbb{P}(a \leq X \leq b)$ deve essere una quantità positiva per ogni $a < b$;
- (b) f_X è *normalizzata*, cioè $\int_{-\infty}^{+\infty} f_X(x) dx = 1$, perché $\int_{-\infty}^{+\infty} f_X(x) dx = \mathbb{P}(-\infty < X < +\infty) = 1$.

È da notare che la densità f_X non è necessariamente una funzione continua, e può anche assumere valori maggiori di 1. Inoltre, per una variabile aleatoria X assolutamente continua si ha sempre

$$\mathbb{P}(X = a) = \int_a^a f_X(x) dx = 0 \quad \text{per ogni } a \in \mathbb{R}$$

da cui seguono le uguaglianze

$$\mathbb{P}(a \leq X \leq b) = \mathbb{P}(a \leq X < b) = \mathbb{P}(a < X \leq b) = \mathbb{P}(a < X < b).$$

Esempio 15 (Densità uniforme continua). Siano $a, b \in \mathbb{R}$ con $a < b$. Una variabile aleatoria X assolutamente continua ha *densità uniforme* sull'intervallo $[a, b]$ se la sua densità è

$$f_X(x) = \frac{1}{b-a} \mathbb{1}_{[a,b]}(x) = \begin{cases} \frac{1}{b-a} & \text{se } x \in [a, b] \\ 0 & \text{se } x < a \text{ o } x > b \end{cases}$$

dove anche in questo caso abbiamo denotato con $\mathbb{1}_{[a,b]}$ la funzione indicatrice dell'intervallo $[a, b]$ (attenzione: questa volta $\mathbb{1}_{[a,b]}$ non è una variabile aleatoria, ma solo un'utile notazione per la densità!). Si scrive anche $X \sim \mathcal{U}([a, b])$.

Esempio 16. Sia $\lambda > 0$. Una variabile aleatoria T assolutamente continua ha *densità esponenziale* di parametro λ se

$$f_T(x) = \lambda e^{-\lambda x} \mathbb{1}_{[0,+\infty)}(x) = \begin{cases} \lambda e^{-\lambda x} & \text{se } x \geq 0 \\ 0 & \text{se } x < 0 \end{cases}$$

Si scrive in questo caso $T \sim \mathcal{E}(\lambda)$.

Esempio 17. Siano μ e σ due parametri reali, con $\sigma > 0$. Una variabile aleatoria Z assolutamente continua ha *densità normale* (o *gaussiana*) di parametri μ e σ se

$$f_Z(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{per ogni } x \in \mathbb{R}.$$

Dal momento che è impossibile trovare esplicitamente una primitiva della funzione f_Z , dimostrare che $\int_{-\infty}^{+\infty} f_Z(x) dx = 1$ è un problema non banale che lasceremo al corso di Analisi. La densità normale di parametri μ e σ si indica con $N(\mu, \sigma^2)$.

1.6 La funzione di ripartizione di una variabile aleatoria

Sia X una variabile aleatoria qualsiasi. La **funzione di ripartizione** di X è la funzione

$$F_X : \mathbb{R} \rightarrow [0, 1] \quad F_X(x) = \mathbb{P}(X \leq x).$$

Notiamo che F_X ha le seguenti proprietà:

(a) è una funzione **monotona non decrescente**, in quanto se $x < y$

$$F_X(x) = \mathbb{P}(X \leq x) \leq \mathbb{P}(X \leq y) = F_X(y)$$

perché $\{X \leq x\} \subseteq \{X \leq y\}$;

(b) **$F_X(x) \rightarrow 0$ per $x \rightarrow -\infty$ e $F_X(x) \rightarrow 1$ per $x \rightarrow +\infty$** , poiché

$$\begin{aligned} \lim_{x \rightarrow +\infty} F_X(x) &= \mathbb{P}(X < \infty) = 1 \\ \lim_{x \rightarrow -\infty} F_X(x) &= \lim_{x \rightarrow -\infty} (1 - \mathbb{P}(X > x)) = 1 - \mathbb{P}(X > -\infty) = 0; \end{aligned}$$

(c) per ogni $a < b$, la funzione F_X può essere **usata per calcolare le probabilità del tipo $\mathbb{P}(a < X \leq b)$** , in quanto

$$\mathbb{P}(a < X \leq b) = \mathbb{P}(\{X \leq b\} \setminus \{X \leq a\}) = \mathbb{P}(X \leq b) - \mathbb{P}(X \leq a) = F_X(b) - F_X(a).$$

Se X è discreta a valori nell'insieme $S = \{x_1, x_2, \dots\}$ e con densità p_X , esplicitamente

$$F_X(x) = \sum_{x_i \leq x} p_X(x_i) \quad \text{per ogni } x \in \mathbb{R}.$$

Enumerando gli elementi di S in ordine crescente, con $x_i < x_{i+1}$ per ogni i , la funzione di ripartizione F_X è una funzione a salti che è costante su ogni intervallo del tipo $[x_i, x_{i+1})$. La sua densità può essere ricavata da

$$p_X(x_i) = F(x_i) - F(x_{i-1}).$$

Esempio 18 (**Funzione di ripartizione della densità bernoulliana**). È il caso più semplice: se $X \sim B(1, p)$, allora

$$F_X(x) = \begin{cases} 0 & \text{se } x < 0 \\ 1 - p & \text{se } 0 \leq x < 1 \\ 1 & \text{se } x \geq 1 \end{cases}$$

Esempio 19 (**Funzione di ripartizione della densità geometrica. Assenza di memoria**). Sia T la variabile aleatoria

T = numero della prova in cui si ottiene il primo successo

in una successione infinita di prove di Bernoulli E_1, E_2, \dots , ciascuna con probabilità $p = \mathbb{P}(E_i)$ di successo. Abbiamo visto nell'Esempio 14 che $T \sim \mathcal{G}(p)$. Per calcolare la funzione di ripartizione di T osserviamo che, per ogni $k \in \{1, 2, 3, \dots\}$,

$$F_T(k) = 1 - \mathbb{P}(T > k) = 1 - \mathbb{P}(E_1^c \cap E_2^c \cap \dots \cap E_k^c) = 1 - \mathbb{P}(E_1^c) \mathbb{P}(E_2^c) \dots \mathbb{P}(E_k^c) = 1 - (1 - p)^k,$$

mentre chiaramente $F_T(x) = 0$ per $x < 1$. Pertanto,

$$F_T(x) = \begin{cases} 1 - (1 - p)^{\lfloor x \rfloor} & \text{se } x \geq 1 \\ 0 & \text{se } x < 1 \end{cases}$$

dove abbiamo indicato con $\lfloor x \rfloor$ la parte intera di x . La variabile aleatoria T gode della proprietà di *assenza di memoria*, cioè

$$\mathbb{P}(T > n + \Delta n \mid T > n) = \mathbb{P}(T > \Delta n) \quad \text{per tutti gli interi } n, \Delta n \geq 0. \quad (1.7)$$

In altre parole, la probabilità che il primo successo si verifichi dopo $n + \Delta n$ prove sapendo che fino alla prova n -esima non si è ancora verificato coincide con la probabilità che il primo successo avvenga dopo Δn prove senza nessuna informazione a priori. Infatti, supponendo $n, \Delta n$ interi,

$$\begin{aligned} & \mathbb{P}(T > n + \Delta n \mid T > n) \\ &= \frac{\mathbb{P}(\{T > n + \Delta n\} \cap \{T > n\})}{\mathbb{P}(T > n)} \quad \text{per la definizione di probabilità condizionata} \\ &= \frac{\mathbb{P}(T > n + \Delta n)}{\mathbb{P}(T > n)} \quad \text{perché } \{T > n + \Delta n\} \subseteq \{T > n\} \\ &= \frac{1 - F_T(n + \Delta n)}{1 - F_T(n)} = \frac{(1 - p)^{n + \Delta n}}{(1 - p)^n} = (1 - p)^{\Delta n} = 1 - F_T(\Delta n) \\ &= \mathbb{P}(T > \Delta n). \end{aligned}$$

La proprietà di assenza di memoria della densità geometrica spiega il motivo per cui, per esempio, al lotto non ha senso puntare sui numeri ritardatari. Infatti, se $i \in \{1, 2, \dots, 90\}$ è uno qualunque dei numeri del lotto, la variabile aleatoria

$T =$ settimana in cui il numero i viene estratto per la prima volta

ha densità $T \sim \mathcal{G}(1/90)$. Per l'equazione (1.7), *condizionatamente all'evento*

$$\{T > n\} = \text{"all'}n\text{'-esima estrazione il numero } i \text{ non è ancora uscito"},$$

la variabile aleatoria $T - n$ ha la stessa funzione di ripartizione di T , e pertanto $T - n \sim \mathcal{G}(1/90)$ condizionatamente all'evento $\{T > n\}$. In altre parole, la variabile aleatoria T senza nessuna informazione a priori e la variabile aleatoria $T - n$ con l'informazione a priori che $\{T > n\}$ hanno la stessa densità.

Se invece X è una variabile aleatoria assolutamente continua con densità f_X , si ha

$$F_X(x) = \int_{-\infty}^x f_X(t) dt \quad \text{per ogni } x \in \mathbb{R}.$$

Osserviamo che in questo caso F_X è una funzione continua. Inoltre, per il teorema fondamentale del calcolo integrale la densità di X si può calcolare derivando la funzione di ripartizione F_X

$$f_X(x) = \frac{dF_X(x)}{dx} \quad \text{per ogni } x \in \mathbb{R}.$$

Esempio 20 (Funzione di ripartizione della densità esponenziale. Assenza di memoria). Se T è una variabile aleatoria assolutamente continua con densità $\mathcal{E}(\lambda)$, la sua funzione di ripartizione è $F_T(x) = 0$ per ogni $x < 0$, mentre per $x > 0$ abbiamo

$$F_T(x) = \int_0^x \lambda e^{-\lambda t} dt = [-e^{-\lambda t}]_{t=0}^{t=x} = 1 - e^{-\lambda x}.$$

Da questa espressione ricaviamo che T gode della proprietà di assenza di memoria esattamente come nel caso della densità geometrica. Infatti, se $x, \Delta x \geq 0$,

$$\begin{aligned} \mathbb{P}(T > x + \Delta x \mid T > x) &= \frac{\mathbb{P}(\{T > x + \Delta x\} \cap \{T > x\})}{\mathbb{P}(T > x)} = \frac{\mathbb{P}(T > x + \Delta x)}{\mathbb{P}(T > x)} \\ &= \frac{1 - F_T(x + \Delta x)}{1 - F_T(x)} = \frac{e^{-\lambda(x + \Delta x)}}{e^{-\lambda x}} = e^{-\lambda \Delta x} = 1 - F_T(\Delta x) \\ &= \mathbb{P}(T > \Delta x). \end{aligned}$$

La densità esponenziale viene in genere associata al tempo di guasto di una macchina non soggetta a usura. Ciò è giustificato proprio dall'assenza di memoria dell'esponenziale: infatti, la relazione $\mathbb{P}(T > t + \Delta t \mid T > t) = \mathbb{P}(T > \Delta t)$ significa che sapere a priori che la macchina al tempo t non si è ancora rotta non cambia la probabilità che essa duri ancora per un altro intervallo Δt . In altre parole, l'età della macchina non influenza la sua durata successiva, che è esattamente la caratteristica di assenza di usura.

1.7 Funzioni di una variabile aleatoria e standardizzazione

La conoscenza di F_X permette di calcolare molto facilmente la densità di funzioni arbitrarie della variabile aleatoria X . Più precisamente, se $g : \mathbb{R} \rightarrow \mathbb{R}$ è una qualunque funzione, si indica con $g(X)$ la variabile aleatoria ottenuta componendo g con X :

$$g(X) := g \circ X \quad \text{cioè} \quad g(X)(\omega) := g(X(\omega)) \quad \text{per ogni } \omega \in \Omega.$$

In tal senso, la variabile aleatoria $g(X)$ è una *funzione* di X .

Esempio 21. Scegliamo come $g : \mathbb{R} \rightarrow \mathbb{R}$ la funzione $g(x) = x^2$. Allora $g(X)$ è la variabile aleatoria $Y = X^2$. Se X è assolutamente continua con densità f_X , abbiamo

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(X^2 \leq y) = \begin{cases} \mathbb{P}(-\sqrt{y} \leq X \leq \sqrt{y}) & \text{se } y \geq 0 \\ \mathbb{P}(\emptyset) & \text{se } y < 0 \end{cases}$$

Nel primo caso

$$F_Y(y) = \mathbb{P}(-\sqrt{y} \leq X \leq \sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y})$$

e dunque derivando e ricordando che $F'_X(x) = dF_X(x)/dx = f_X(x)$ troviamo

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{1}{2\sqrt{y}}F'_X(\sqrt{y}) + \frac{1}{2\sqrt{y}}F'_X(-\sqrt{y}) = \frac{1}{2\sqrt{y}}f_X(\sqrt{y}) + \frac{1}{2\sqrt{y}}f_X(-\sqrt{y}).$$

Nel secondo caso invece

$$F_Y(y) = 0 \quad \Rightarrow \quad f_Y(y) = \frac{dF_Y(y)}{dy} = 0.$$

Mettendo insieme i due casi si trova che la densità di Y è

$$f_Y(y) = \begin{cases} \frac{1}{2\sqrt{y}}f_X(\sqrt{y}) + \frac{1}{2\sqrt{y}}f_X(-\sqrt{y}) & \text{se } y \geq 0 \\ 0 & \text{se } y < 0 \end{cases}$$

Esempio 22 (Trasformata affine di una variabile aleatoria). Fissati due numeri reali a, b con $a \neq 0$, scegliamo la funzione $g(x) = ax + b$. Se X è una qualunque variabile aleatoria, $g(X)$ è pertanto la *trasformata affine* $Y = aX + b$ di X . Se X è assolutamente continua, usiamo ancora la funzione di ripartizione per determinare la densità di Y :

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(aX + b \leq y) = \begin{cases} \mathbb{P}(X \leq \frac{y-b}{a}) = F_X\left(\frac{y-b}{a}\right) & \text{se } a > 0 \\ \mathbb{P}(X \geq \frac{y-b}{a}) = 1 - F_X\left(\frac{y-b}{a}\right) & \text{se } a < 0 \end{cases}$$

Derivando come nell'esempio precedente otteniamo

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \begin{cases} \frac{1}{a}F'_X\left(\frac{y-b}{a}\right) = \frac{1}{a}f_X\left(\frac{y-b}{a}\right) & \text{se } a > 0 \\ -\frac{1}{a}F'_X\left(\frac{y-b}{a}\right) = -\frac{1}{a}f_X\left(\frac{y-b}{a}\right) & \text{se } a < 0 \end{cases}$$

In altre parole,

$$f_Y(y) = \frac{1}{|a|}f_X\left(\frac{y-b}{a}\right).$$

Applichiamo l'esempio precedente al calcolo della trasformata affine $Y = aZ + b$ di una variabile aleatoria normale $Z \sim N(\mu, \sigma^2)$. Abbiamo

$$f_Y(y) = \frac{1}{|a|} f_Z\left(\frac{y-b}{a}\right) = \frac{1}{\sqrt{2\pi}\sigma|a|} \exp\left[-\frac{\left(\frac{y-b}{a} - \mu\right)^2}{2\sigma^2}\right] = \frac{1}{\sqrt{2\pi}\sigma|a|} e^{-\frac{(y-a\mu-b)^2}{2(\sigma a)^2}}$$

da cui si deduce che $Y \sim N(a\mu + b, (|a|\sigma)^2)$. In particolare, la trasformata affine di una variabile aleatoria normale è ancora normale. Notiamo che scegliendo $a = 1/\sigma$ e $b = -\mu/\sigma$ troviamo che la variabile aleatoria $Y = \frac{Z-\mu}{\sigma}$ ha densità normale $N(0, 1)$. Tale densità si chiama *normale standard*, e la sua funzione di ripartizione $\Phi := F_Y$ è tabulata in qualunque libro di probabilità e statistica. La conoscenza di Φ permette di trovare tutte le probabilità del tipo $\mathbb{P}(Z \leq z)$ ecc. mediante la seguente *standardizzazione* di Z

$$\mathbb{P}(Z \leq z) = \mathbb{P}(Z - \mu \leq z - \mu) = \mathbb{P}\left(\underbrace{\frac{Z - \mu}{\sigma}}_{\sim N(0,1)} \leq \frac{z - \mu}{\sigma}\right) = \Phi\left(\frac{z - \mu}{\sigma}\right) \quad \text{per ogni } z \in \mathbb{R}.$$

1.8 Vettori aleatori

Un *vettore aleatorio* a n componenti è una qualunque funzione $\vec{X} : \Omega \rightarrow \mathbb{R}^n$. Esplicitamente,

$$\vec{X}(\omega) = (X_1(\omega), X_2(\omega), \dots, X_n(\omega)) \quad \text{per ogni } \omega \in \Omega.$$

Ciascuna componente X_i è una funzione da Ω in \mathbb{R} , e pertanto un vettore aleatorio può essere visto come una n -upla di variabili aleatorie (X_1, X_2, \dots, X_n) .

Quando la σ -algebra degli eventi non coincide con tutto l'insieme delle parti $\mathcal{P}(\Omega)$, si richiede in più che $\{X_i \leq a\} \in \mathcal{F}$ per ogni $a \in \mathbb{R}$ e $i = 1, 2, \dots, n$. Tuttavia, come al solito non approfondiremo questo dettaglio, e supporremo che tutti i vettori aleatori con cui avremo a che fare soddisfino tale requisito.

Come nel caso scalare, anche qui useremo le notazioni

$$\begin{aligned} \{\vec{X} \in A\} &= \vec{X}^{-1}(A) && \text{per ogni dominio } A \subseteq \mathbb{R}^n \\ \{\vec{X} = \vec{x}\} &= \vec{X}^{-1}(\{\vec{x}\}) && \text{per ogni } \vec{x} \in \mathbb{R}^n \end{aligned}$$

e inoltre

$$\begin{aligned} \{X_1 \in A_1, X_2 \in A_2, \dots, X_n \in A_n\} &= \{\vec{X} \in A_1 \times A_2 \times \dots \times A_n\} \\ &= \{X_1 \in A_1\} \cap \{X_2 \in A_2\} \cap \dots \cap \{X_n \in A_n\} \end{aligned}$$

e simili.

Un vettore aleatorio \vec{X} si dice *discreto* quando tutte le sue componenti sono variabili aleatorie discrete. In tal caso, se S_1, S_2, \dots, S_n sono i sottoinsiemi discreti di \mathbb{R} in cui prendono valore ciascuna delle componenti X_1, X_2, \dots, X_n , rispettivamente, la *densità* di \vec{X} è la funzione

$$p_{\vec{X}} : S_1 \times S_2 \times \dots \times S_n \rightarrow [0, 1]$$

data da

$$p_{\vec{X}}(x_1, x_2, \dots, x_n) = \mathbb{P}(\vec{X} = (x_1, x_2, \dots, x_n)) = \mathbb{P}(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n)$$

(come al solito, per semplicità di scrittura nella formula precedente abbiamo rimosso le parentesi graffe dagli eventi). La funzione $p_{\vec{X}}$ si chiama anche *densità congiunta* delle variabili aleatorie X_1, X_2, \dots, X_n . È immediato verificare che la densità congiunta $p_{\vec{X}}$ ha proprietà analoghe alla densità di una singola variabile aleatoria. In particolare, per ogni dominio $A \subseteq \mathbb{R}^n$ si ha

$$\mathbb{P}(\vec{X} \in A) = \sum_{\substack{x_1 \in S_1, x_2 \in S_2, \dots, x_n \in S_n \\ (x_1, x_2, \dots, x_n) \in A}} p_{\vec{X}}(x_1, x_2, \dots, x_n)$$

e quindi la densità $p_{\vec{X}}$

(a) è positiva: $p_{\vec{X}}(x_1, x_2, \dots, x_n) \geq 0$ per ogni (x_1, x_2, \dots, x_n) ;

(b) è normalizzata: $\sum_{x_1 \in S_1} \sum_{x_2 \in S_2} \dots \sum_{x_n \in S_n} p_{\vec{X}}(x_1, x_2, \dots, x_n) = 1$.

La densità di ciascuna componente X_i si può ricavare dalla densità congiunta semplicemente sommando quest'ultima rispetto a tutte le variabili tranne la i -esima. Per esempio, per calcolare la densità di X_1 abbiamo

$$\begin{aligned} p_{X_1}(x) &= \mathbb{P}(X_1 = x) = \mathbb{P}(X_1 = x, X_2 \in S_2, X_3 \in S_3, \dots, X_n \in S_n) \\ &= \sum_{x_2 \in S_2} \sum_{x_3 \in S_3} \dots \sum_{x_n \in S_n} p_{\vec{X}}(x, x_2, \dots, x_n). \end{aligned}$$

Per questo motivo, la densità p_{X_i} si chiama anche *densità marginale* di $p_{\vec{X}}$ rispetto all' i -esima componente.

Osservazione 4. Quando si ha a che fare con un vettore aleatorio a $n = 2$ componenti $\vec{X} = (X, Y)$, è molto utile riassumere la sua densità congiunta nella tabella

$X \setminus Y$	y_1	y_2	\dots
x_1	$p_{(X,Y)}(x_1, y_1)$	$p_{(X,Y)}(x_1, y_2)$	\dots
x_2	$p_{(X,Y)}(x_2, y_1)$	$p_{(X,Y)}(x_2, y_2)$	\dots
\dots	\dots	\dots	\dots

dove $S_1 = \{x_1, x_2, \dots\}$ e $S_2 = \{y_1, y_2, \dots\}$ sono i valori possibili di X e di Y , rispettivamente. Poiché la densità congiunta è positiva e normalizzata, tutte le caselle interne devono essere ≥ 0 e la loro somma deve fare 1.

Nella tabella precedente, la densità marginale di X si ricava sommando gli elementi in ciascuna riga. Ciò si fa di solito aggiungendo una colonna contenente i valori della marginale p_X , nel modo che segue:

$X \setminus Y$	y_1	y_2	\dots	p_X
x_1	$p_{(X,Y)}(x_1, y_1)$	$p_{(X,Y)}(x_1, y_2)$	\dots	$p_X(x_1) = p_{(X,Y)}(x_1, y_1) + p_{(X,Y)}(x_1, y_2) + \dots$
x_2	$p_{(X,Y)}(x_2, y_1)$	$p_{(X,Y)}(x_2, y_2)$	\dots	$p_X(x_2) = p_{(X,Y)}(x_2, y_1) + p_{(X,Y)}(x_2, y_2) + \dots$
\dots	\dots	\dots	\dots	\dots

Analogamente, la marginale p_Y si ricava sommando le colonne corrispondenti, e aggiungendo un'ulteriore riga in fondo alla tabella:

$X \setminus Y$	y_1	y_2	\dots	p_X
x_1	$p_{(X,Y)}(x_1, y_1)$	$p_{(X,Y)}(x_1, y_2)$	\dots	$p_X(x_1)$
x_2	$p_{(X,Y)}(x_2, y_1)$	$p_{(X,Y)}(x_2, y_2)$	\dots	$p_X(x_2)$
\dots	\dots	\dots	\dots	\dots
p_Y	$p_Y(y_1)$	$p_Y(y_2)$	\dots	

La normalizzazione di $p_{(X,Y)}$ si riflette allora nel fatto che sia la riga corrispondente a p_Y sia la colonna corrispondente a p_X sommano a 1.

Osservazione 5. È importante osservare che le densità marginali *non determinano mai univocamente la densità congiunta* $p_{\vec{X}}$. In altre parole, assegnate le densità $p_{X_1}, p_{X_2}, \dots, p_{X_n}$, esistono un'infinità di densità congiunte $p_{\vec{X}}$ diverse che danno le stesse marginali $p_{X_1}, p_{X_2}, \dots, p_{X_n}$. Per esempio, con $n = 2$, le due densità congiunte $p_{(X,Y)}$ e $q_{(X,Y)}$ sull'insieme $S = \{0, 1\}^2$ descritte rispettivamente dalle tabelle

$$p_{(X,Y)} = \begin{array}{c|cc} X \setminus Y & 0 & 1 \\ \hline 0 & 0 & 1/2 \\ \hline 1 & 1/2 & 0 \end{array} \quad q_{(X,Y)} = \begin{array}{c|cc} X \setminus Y & 0 & 1 \\ \hline 0 & 1/4 & 1/4 \\ \hline 1 & 1/4 & 1/4 \end{array}$$

danno entrambe come marginali la densità uniforme

$$p_X(i) \equiv q_X(i) = \frac{1}{2} \quad \text{per ogni } i \in \{0, 1\},$$

eppure $p_{(X,Y)}$ e $q_{(X,Y)}$ sono diverse. Ne concludiamo che la conoscenza delle sole marginali non permette di calcolare le probabilità del tipo $\mathbb{P}(\vec{X} \in A)$ quando $A \subseteq \mathbb{R}^n$ è un generico insieme.

Una successione di n variabili aleatorie X_1, X_2, \dots, X_n si dicono *indipendenti* se la loro densità congiunta è il prodotto delle marginali, cioè

$$p_{\vec{X}}(x_1, x_2, \dots, x_n) = p_{X_1}(x_1)p_{X_2}(x_2) \dots p_{X_n}(x_n) \quad \text{per ogni } (x_1, x_2, \dots, x_n). \quad (1.8)$$

Una successione di variabili aleatorie indipendenti costituisce un'importante eccezione all'Osservazione 5. Infatti, la fondamentale informazione aggiuntiva che le X_1, X_2, \dots, X_n sono indipendenti permette di determinare la loro densità congiunta a partire dalla conoscenza delle sole marginali attraverso la formula (1.8) precedente.

Osserviamo che le definizioni di indipendenza per eventi e per variabili aleatorie sono consistenti tra loro: infatti, se p.es. X e Y sono due variabili aleatorie indipendenti, allora tutte le coppie di eventi della forma $\{X \in A\}, \{Y \in B\}$ sono indipendenti, in quanto

$$\begin{aligned} \mathbb{P}(\{X \in A\} \cap \{Y \in B\}) &= \sum_{x \in A} \sum_{y \in B} p_{(X,Y)}(x, y) = \sum_{x \in A} \sum_{y \in B} p_X(x) p_Y(y) = \sum_{x \in A} p_X(x) \sum_{y \in B} p_Y(y) \\ &= \mathbb{P}(X \in A) \mathbb{P}(Y \in B). \end{aligned}$$

La generalizzazione al caso di $n \geq 3$ variabili aleatorie è simile (provare per esercizio!).

Esempio 23. Riprendiamo l'Esempio 10 del lancio di due dadi. I vettori $\vec{U} = (X, Y)$ e $\vec{W} = (X, Z)$ sono entrambi vettori aleatori discreti a valori nell'insieme $S = \{1, 2, 3, 4, 5, 6\}^2$. Esplicitamente,

$$\vec{U}(\omega_1, \omega_2) = (\omega_1, \omega_2) \quad \vec{W}(\omega_1, \omega_2) = (\omega_1, \max\{\omega_1, \omega_2\})$$

per ogni $(\omega_1, \omega_2) \in \Omega$. La densità congiunta di X e Y è

$$p_{\vec{U}}(x, y) = \mathbb{P}(\vec{U} = (x, y)) = \mathbb{P}(\{(x, y)\}) = \frac{1}{36} \quad \text{per ogni } (x, y) \in S.$$

Si vede che

$$p_{\vec{U}}(x, y) = \frac{1}{36} \equiv \frac{1}{6} \cdot \frac{1}{6} = p_X(x)p_Y(y) \quad \text{per ogni } (x, y) \in S$$

da cui segue che le variabili aleatorie X e Y sono indipendenti. Per quanto riguarda il vettore aleatorio \vec{V} , invece, per ogni $(x, z) \in S$ si ha

$$p_{\vec{V}}(x, z) = \mathbb{P}(\vec{V} = (x, z)) = \begin{cases} \mathbb{P}(\emptyset) = 0 & \text{se } x > z \\ \mathbb{P}(\{(z, \omega_2) \mid \omega_2 \leq z\}) = \frac{z}{36} & \text{se } x = z \\ \mathbb{P}(\{(x, z)\}) = \frac{1}{36} & \text{se } x < z \end{cases}$$

In questo caso

$$p_X(x)p_Z(z) = \frac{1}{6} \cdot \frac{2z-1}{36} \neq p_{\vec{V}}(x, z)$$

e quindi le variabili aleatorie X e Z *non sono* indipendenti.

Un vettore aleatorio $\vec{X} = (X_1, X_2, \dots, X_n)$ si dice *assolutamente continuo* se esiste una funzione $f_{\vec{X}} : \mathbb{R}^n \rightarrow \mathbb{R}$ tale che

$$\mathbb{P}(\vec{X} \in A) = \iint \dots \int_A f_{\vec{X}}(x_1, x_2, \dots, x_n) dx_1 dx_2 \dots dx_n \quad \text{per ogni dominio } A \subseteq \mathbb{R}^n.$$

La funzione $f_{\vec{X}}$ si chiama anche in questo caso *densità* del vettore aleatorio \vec{X} o *densità congiunta* delle variabili aleatorie X_1, X_2, \dots, X_n .

Valgono per $f_{\vec{X}}$ proprietà analoghe a quelle di una densità discreta, naturalmente a patto di scambiare le somme con i corrispondenti integrali:

- (a) positività: $f_{\vec{X}}(x_1, x_2, \dots, x_n) \geq 0$ per ogni $(x_1, x_2, \dots, x_n) \in \mathbb{R}^n$;
- (b) normalizzazione: $\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} f_{\vec{X}}(x_1, x_2, \dots, x_n) dx_1 dx_2 \dots dx_n = 1$.

In modo simile al caso discreto, la densità f_{X_i} di ciascuna componente X_i si ottiene come i -esima *marginale* della congiunta $f_{\vec{X}}$ secondo la formula

$$f_{X_i}(x) = \int_{-\infty}^{+\infty} dx_1 \dots \int_{-\infty}^{+\infty} dx_{i-1} \int_{-\infty}^{+\infty} dx_{i+1} \dots \int_{-\infty}^{+\infty} dx_n f_{\vec{X}}(x_1, \dots, x_{i-1}, x, x_{i+1}, \dots, x_n).$$

Infatti, vediamo per esempio per la prima componente X_1 ; per farlo, ricaviamo la funzione di ripartizione F_{X_1} dalla densità congiunta e poi deriviamola:

$$\begin{aligned} F_{X_1}(x) &= \mathbb{P}(X_1 \leq x) = \mathbb{P}(X_1 \leq x, X_2 \in (-\infty, +\infty), X_3 \in (-\infty, +\infty), \dots, X_n \in (-\infty, +\infty)) \\ &= \int_{-\infty}^x dx_1 \int_{-\infty}^{+\infty} dx_2 \int_{-\infty}^{+\infty} dx_3 \dots \int_{-\infty}^{+\infty} dx_n f_{\vec{X}}(x_1, x_2, x_3, \dots, x_n) \\ f_{X_1}(x) &= \frac{dF_{X_1}(x)}{dx} = \int_{-\infty}^{+\infty} dx_2 \int_{-\infty}^{+\infty} dx_3 \dots \int_{-\infty}^{+\infty} dx_n f_{\vec{X}}(x, x_2, x_3, \dots, x_n). \end{aligned}$$

Anche nel caso di un vettore aleatorio assolutamente continuo \vec{X} , le componenti X_1, X_2, \dots, X_n si dicono *indipendenti* se la loro densità congiunta si fattorizza nel prodotto delle marginali, cioè se

$$f_{\vec{X}}(x_1, x_2, \dots, x_n) = f_{X_1}(x_1)f_{X_2}(x_2) \dots f_{X_n}(x_n) \quad \text{per ogni } (x_1, x_2, \dots, x_n) \in \mathbb{R}^n.$$

Esempio 24 (Densità uniforme sul cerchio). *Problema:* Consideriamo un esperimento aleatorio in cui si sceglie a caso con probabilità uniforme un punto nel cerchio unitario $C = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq 1\}$. Indichiamo con (X, Y) il vettore aleatorio

$$(X, Y) = \text{coordinate del punto}$$

cioè X e Y sono le variabili aleatorie

$$X = \text{ascissa del punto} \qquad Y = \text{ordinata del punto.}$$

La densità del vettore aleatorio (X, Y) è pertanto la seguente *densità uniforme* sul cerchio C :

$$f_{(X,Y)}(x, y) = \begin{cases} \frac{1}{\pi} & \text{se } x^2 + y^2 \leq 1 \\ 0 & \text{altrimenti} \end{cases}$$

(verificare per esercizio che $f_{(X,Y)}$ è normalizzata). Per ogni $r > 0$, qual è la probabilità che il punto scelto a caso disti meno di r dall'origine? Le variabili aleatorie X e Y sono indipendenti?

Soluzione: Si ha l'uguaglianza di eventi

$$\text{"il punto dista meno di } r \text{ dall'origine"} = \{\sqrt{X^2 + Y^2} < r\} = \{X^2 + Y^2 < r^2\} = \{(X, Y) \in A\}$$

dove $A \subset \mathbb{R}^2$ è l'insieme

$$A = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq r^2\}.$$

La prima domanda richiede pertanto di calcolare la probabilità

$$\mathbb{P}((X, Y) \in A) = \iint_A f_{(X,Y)}(x, y) \, dx \, dy.$$

Se $r \geq 1$, l'integrale precedente vale chiaramente 1. Se invece $r < 1$, integrando per sezioni

$$\begin{aligned} \mathbb{P}((X, Y) \in A) &= \iint_A \frac{1}{\pi} \, dx \, dy = \int_{-r}^r dx \int_{-\sqrt{r^2-x^2}}^{\sqrt{r^2-x^2}} dy \frac{1}{\pi} = \frac{2}{\pi} \int_{-r}^r \sqrt{r^2-x^2} \, dx \\ &= \frac{2r^2}{\pi} \int_{-\pi/2}^{\pi/2} \cos^2 \theta \, d\theta \quad \text{con la sostituzione } x = r \sin \theta \\ &= r^2 \end{aligned}$$

(notare che l'integrale precedente si poteva calcolare ancora più facilmente passando in coordinate polari).

Per rispondere invece alla seconda domanda, dobbiamo calcolare le densità marginali

$$f_X(x) = \int_{-\infty}^{+\infty} f_{(X,Y)}(x,y) dy = \begin{cases} \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \frac{1}{\pi} dy = \frac{2}{\pi} \sqrt{1-x^2} & \text{se } -1 \leq x \leq 1 \\ 0 & \text{altrimenti} \end{cases}$$

$$f_Y(y) = \int_{-\infty}^{+\infty} f_{(X,Y)}(x,y) dx = \begin{cases} \int_{-\sqrt{1-y^2}}^{\sqrt{1-y^2}} \frac{1}{\pi} dx = \frac{2}{\pi} \sqrt{1-y^2} & \text{se } -1 \leq y \leq 1 \\ 0 & \text{altrimenti} \end{cases}$$

e confrontare il loro prodotto $f_X(x)f_Y(y)$ con la densità congiunta $f_{(X,Y)}(x,y)$. Poiché $\frac{2}{\pi} \sqrt{1-x^2} \cdot \frac{2}{\pi} \sqrt{1-y^2} \neq \frac{1}{\pi}$, se ne deduce che X e Y *non sono* indipendenti.

Se X e Y sono due variabili aleatorie, la loro somma $Z = X + Y$ è ancora una variabile aleatoria. In generale, però, la conoscenza delle sole densità di X e di Y non è sufficiente a determinare la densità di Z (per questo occorrerebbe infatti conoscere tutta la densità congiunta del vettore (X, Y)). Se tuttavia abbiamo in più l'informazione che X e Y sono indipendenti, allora la loro densità congiunta è determinata dalle marginali, e la densità di Z si può effettivamente calcolare a partire solo da queste. Vale infatti il risultato seguente.

Proposizione 2. *Siano X e Y due variabili aleatorie assolutamente continue e indipendenti, con densità f_X e f_Y , rispettivamente. Sia $Z = X + Y$ la loro somma. Allora Z è una variabile aleatoria assolutamente continua con densità*

$$f_Z(z) = \int_{-\infty}^{+\infty} f_X(z-y)f_Y(y) dy. \quad (1.9)$$

Dimostrazione. Calcoliamo la funzione di ripartizione di Z e poi deriviamola per ottenere f_Z . Si ha

$$\begin{aligned} F_Z(z) &= \mathbb{P}(Z \leq z) = \mathbb{P}(X + Y \leq z) = \mathbb{P}((X, Y) \in A) \quad \text{dove } A = \{(x, y) \in \mathbb{R}^2 \mid x + y \leq z\} \\ &= \iint_A f_{(X,Y)}(x, y) dx dy = \int_{-\infty}^{+\infty} dy \int_{-\infty}^{z-y} dx f_{(X,Y)}(x, y) \quad \text{integrando per sezioni} \\ &= \int_{-\infty}^{+\infty} dy \int_{-\infty}^{z-y} f_X(x)f_Y(y) dx \quad \text{per l'indipendenza di } X \text{ e } Y. \end{aligned}$$

Nell'integrale più interno, facciamo il cambio di variabili $x = x' - y$ come segue

$$\int_{-\infty}^{z-y} f_X(x)f_Y(y) dx = \int_{-\infty}^z f_X(x' - y)f_Y(y) dx'$$

e quindi

$$\begin{aligned} F_Z(z) &= \int_{-\infty}^{+\infty} dy \int_{-\infty}^z f_X(x' - y)f_Y(y) dx' \\ &= \int_{-\infty}^z \left(\int_{-\infty}^{+\infty} f_X(x' - y)f_Y(y) dy \right) dx' \quad \text{scambiando i due integrali.} \end{aligned}$$

Derivando quest'espressione rispetto a z otteniamo la (1.9). ■

L'integrale nella formula (1.9) si chiama *prodotto di convoluzione* delle densità f_X e f_Y . Quando f_X e f_Y sono entrambe gaussiane, dall'equazione (1.9) si ricava la seguente importante proprietà.

Proposizione 3. *Siano X e Y due variabili aleatorie assolutamente continue e indipendenti. Supponiamo che sia X sia Y abbiano densità normale. Allora la loro somma $Z = X + Y$ ha anch'essa densità normale.*

Dimostrazione. Supponiamo $X \sim N(\mu_X, \sigma_X^2)$ e $Y \sim N(\mu_Y, \sigma_Y^2)$. Possiamo riscrivere Z nella forma

$$Z = \sigma_X \left(\frac{X - \mu_X}{\sigma_X} + \frac{Y - \mu_Y}{\sigma_X} \right) + (\mu_X + \mu_Y) \equiv \sigma_X(X_0 + Y_0) + (\mu_X + \mu_Y)$$

in cui $X_0 := (X - \mu_X)/\sigma_X$ e $Y_0 := (Y - \mu_Y)/\sigma_X$ sono indipendenti e $X_0 \sim N(0, 1)$ e $Y_0 \sim N(0, (\sigma_Y/\sigma_X)^2)$ in quanto trasformazioni affini di variabili aleatorie normali. Sempre perché trasformazioni affini di normali sono ancora normali, è dunque sufficiente dimostrare che la somma $Z_0 = X_0 + Y_0$ è gaussiana quando $X_0 \sim N(0, 1)$ e $Y_0 \sim N(0, \sigma^2)$. In tal caso, applicando la formula (1.9) troviamo

$$f_{Z_0}(z) = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-y)^2}{2}} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{y^2}{2\sigma^2}} dy = C \int_{-\infty}^{+\infty} e^{-\frac{1}{2}(z^2 - 2zy + \frac{\sigma^2+1}{\sigma^2}y^2)} dy$$

dove $C = 1/(2\pi\sigma)$ è una costante. Completando il quadrato nell'esponentiale,

$$f_{Z_0}(z) = C \int_{-\infty}^{+\infty} e^{-\frac{z^2}{2(\sigma^2+1)} - \frac{1}{2}\left(\frac{\sigma}{\sqrt{\sigma^2+1}}z - \frac{\sqrt{\sigma^2+1}}{\sigma}y\right)^2} dy = C e^{-\frac{z^2}{2(\sigma^2+1)}} \int_{-\infty}^{+\infty} e^{-\frac{1}{2}\left(\frac{\sigma}{\sqrt{\sigma^2+1}}z - \frac{\sqrt{\sigma^2+1}}{\sigma}y\right)^2} dy.$$

Sostituendo infine nell'integrale precedente $y' = \frac{\sigma}{\sqrt{\sigma^2+1}}z - \frac{\sqrt{\sigma^2+1}}{\sigma}y$, abbiamo

$$f_{Z_0}(z) = \left(C \int_{-\infty}^{+\infty} e^{-\frac{y'^2}{2}} dy' \right) e^{-\frac{z^2}{2(\sigma^2+1)}} = C' e^{-\frac{z^2}{2(\sigma^2+1)}},$$

dove $C' = C \int_{-\infty}^{+\infty} e^{-\frac{y'^2}{2}} dy'$ è una costante numerica che non dipende da z . Riconosciamo pertanto in f_{Z_0} la forma di una densità gaussiana. ■

1.9 Media e varianza di una variabile aleatoria

Definizione 6. Sia X una variabile aleatoria. La *media* (o *speranza*) di X è il numero reale $\mathbb{E}[X]$ definito come segue:

- se X è discreta con densità $p_X : S \rightarrow [0, 1]$,

$$\mathbb{E}[X] = \sum_{x \in S} x p_X(x); \quad (1.10)$$

- se X è assolutamente continua con densità f_X ,

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} x f_X(x) dx. \quad (1.11)$$

Esempio 25 (Media della bernoulliana). Se $X \sim B(1, p)$, abbiamo

$$\mathbb{E}[X] = \sum_{x \in \{0,1\}} x p_X(x) = 0 \cdot (1-p) + 1 \cdot p = p.$$

Esempio 26 (Media dell'esponenziale). Se $T \sim \mathcal{E}(\lambda)$, risolvendo l'integrale per parti ricaviamo

$$\mathbb{E}[T] = \int_0^{+\infty} x \lambda e^{-\lambda x} dx = \left[-x e^{-\lambda x} \right]_{x=0}^{x=+\infty} + \int_0^{+\infty} e^{-\lambda x} dx = \left[-\frac{1}{\lambda} e^{-\lambda x} \right]_{x=0}^{x=+\infty} = \frac{1}{\lambda}.$$

Se $\vec{X} = (X_1, X_2, \dots, X_n)$ è un vettore aleatorio, ogni sua componente X_i è una variabile aleatoria, e dunque ne possiamo calcolare la media $\mathbb{E}[X_i]$. Più in generale, possiamo calcolare la media di una qualunque funzione scalare del vettore \vec{X} , e non solo di ciascuna sua componente. Ma cosa intendiamo precisamente per *funzione scalare* di \vec{X} ?

La nozione di funzione di un vettore aleatorio è l'estensione naturale del concetto di funzione di una singola variabile aleatoria: se $g : \mathbb{R}^n \rightarrow \mathbb{R}$ è una qualunque funzione, definiamo $g(X_1, X_2, \dots, X_n) := g \circ \vec{X}$. Esplicitamente, $g(X_1, X_2, \dots, X_n)$ è la funzione data da

$$[g(X_1, X_2, \dots, X_n)](\omega) = g(X_1(\omega), X_2(\omega), \dots, X_n(\omega)) \quad \text{per ogni } \omega \in \Omega.$$

Essendo definita su Ω e a valori in \mathbb{R} , la funzione $g(X_1, X_2, \dots, X_n)$ è *a tutti gli effetti una variabile aleatoria*.

Esempio 27. (i) Per la componente X_i abbiamo

$$X_i = g(X_1, X_2, \dots, X_n) \quad \text{dove} \quad g(x_1, x_2, \dots, x_n) = x_i.$$

(ii) La *norma* del vettore aleatorio \vec{X} è la variabile aleatoria

$$R = g(X_1, X_2, \dots, X_n) \quad \text{dove} \quad g(X_1, X_2, \dots, X_n) = \sqrt{x_1^2 + x_2^2 + \dots + x_n^2}.$$

Dal momento che $g(X_1, X_2, \dots, X_n)$ è una variabile aleatoria, possiamo chiederci quanto vale la sua media. Se dovessimo calcolare $\mathbb{E}[g(X_1, X_2, \dots, X_n)]$ usando direttamente la Definizione 6 della media, dovremmo prima ricavare la densità di $g(X_1, X_2, \dots, X_n)$ a partire dalla densità congiunta delle X_1, X_2, \dots, X_n , e poi applicare una delle due formule (1.10) o (1.11) per trovare $\mathbb{E}[g(X_1, X_2, \dots, X_n)]$. Tuttavia, la proposizione seguente ci dice che in realtà il calcolo è molto più semplice e trovare la densità di $g(X_1, X_2, \dots, X_n)$ non è necessario.

Proposizione 4. *Se $\vec{X} = (X_1, X_2, \dots, X_n)$ è un vettore aleatorio discreto con densità $p_{\vec{X}}$ [rispettivamente, assolutamente continuo con densità f_X] e $g : \mathbb{R}^n \rightarrow \mathbb{R}$ è una funzione qualunque, allora*

$$\mathbb{E}[g(X_1, X_2, \dots, X_n)] = \sum_{x_1} \sum_{x_2} \dots \sum_{x_n} g(x_1, x_2, \dots, x_n) p_{\vec{X}}(x_1, x_2, \dots, x_n)$$

[risp.,

$$\mathbb{E}[g(X_1, X_2, \dots, X_n)] = \int_{-\infty}^{+\infty} dx_1 \int_{-\infty}^{+\infty} dx_2 \dots \int_{-\infty}^{+\infty} dx_n g(x_1, x_2, \dots, x_n) f_{\vec{X}}(x_1, x_2, \dots, x_n)$$

]

Dimostrazione. Dimosteremo solo il caso discreto. Inoltre, per semplificare le notazioni supporremo di avere solo $n = 2$ componenti. Se X, Y sono due variabili aleatorie discrete a valori negli insiemi S_X e S_Y , rispettivamente, la variabile aleatoria $g(X, Y)$ prende valori nell'insieme discreto $S = \{g(x, y) \mid x \in S_X, y \in S_Y\}$. La Definizione 6 della media dà

$$\mathbb{E}[g(X, Y)] = \sum_{z \in S} z p_{g(X, Y)}(z) = \sum_{z \in S} z \mathbb{P}(g(X, Y) = z).$$

Inoltre,

$$\mathbb{P}(g(X, Y) = z) = \mathbb{P}((X, Y) \in g^{-1}(\{z\})) = \sum_{\substack{x, y \\ (x, y) \in g^{-1}(\{z\})}} p_{(X, Y)}(x, y) = \sum_{\substack{x, y \\ g(x, y) = z}} p_{(X, Y)}(x, y).$$

Inserendo quest'espressione nella precedente, otteniamo

$$\mathbb{E}[g(X, Y)] = \sum_{z \in S} z \sum_{\substack{x, y \\ g(x, y) = z}} p_{(X, Y)}(x, y) = \sum_{z \in S} \sum_{\substack{x, y \\ g(x, y) = z}} g(x, y) p_{(X, Y)}(x, y) = \sum_{x, y} g(x, y) p_{(X, Y)}(x, y).$$

■

La media gode inoltre delle seguenti proprietà fondamentali.

Proposizione 5 (Proprietà della media). *(i) Se $c \in \mathbb{R}$ e $X \equiv c$ è la variabile aleatoria identicamente uguale a c (cioè $X(\omega) = c$ per ogni $\omega \in \Omega$), allora $\mathbb{E}[X] = c$.*

(ii) Se X e Y sono due variabili aleatorie e a, b sono numeri reali, allora $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$. In particolare, $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$.

(iii) Se X e Y sono due variabili aleatorie indipendenti, allora $\mathbb{E}[XY] = \mathbb{E}[X] \mathbb{E}[Y]$.

Dimostrazione. (i) X è una variabile aleatoria discreta a valori nell'insieme $S = \{c\}$ e con densità $p_X(c) = 1$. Per la formula (1.10), $\mathbb{E}[X] = c \cdot p_X(c) = c$.

(ii) Per fissare le idee supponiamo che il vettore aleatorio (X, Y) sia discreto, e usiamo la Proposizione 4 con $g : \mathbb{R}^2 \rightarrow \mathbb{R}$ data da $g(x, y) = ax + by$. Abbiamo

$$\begin{aligned} \mathbb{E}[aX + bY] &= \mathbb{E}[g(X, Y)] = \sum_{x, y} (ax + by) p_{(X, Y)}(x, y) \\ &= a \sum_x x \sum_y p_{(X, Y)}(x, y) + b \sum_y y \sum_x p_{(X, Y)}(x, y) \\ &= a \sum_x x p_X(x) + b \sum_y y p_Y(y) \quad \text{perché } \sum_y p_{(X, Y)}(x, y) = p_X(x) \text{ e } \sum_x p_{(X, Y)}(x, y) = p_Y(y) \\ &= a\mathbb{E}[X] + b\mathbb{E}[Y]. \end{aligned}$$

Il caso assolutamente continuo è del tutto simile. Infine, per quanto appena dimostrato,

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b\mathbb{E}[1] = a\mathbb{E}[X] + b \quad \text{perché } \mathbb{E}[1] = 1.$$

(iii) Supponiamo questa volta che il vettore aleatorio (X, Y) sia assolutamente continuo. Allora, scegliendo $g : \mathbb{R}^2 \rightarrow \mathbb{R}$ data da $g(x, y) = xy$ e applicando di nuovo la Proposizione 4, otteniamo

$$\begin{aligned} \mathbb{E}[XY] &= \mathbb{E}[g(X, Y)] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xy f_{(X, Y)}(x, y) dx dy \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xy f_X(x) f_Y(y) dx dy \quad \text{per l'indipendenza} \\ &= \int_{-\infty}^{+\infty} x f_X(x) dx \int_{-\infty}^{+\infty} y f_Y(y) dy = \mathbb{E}[X] \mathbb{E}[Y]. \end{aligned}$$

La proprietà (ii) significa che la media \mathbb{E} è un *funzionale lineare* sullo spazio vettoriale delle variabili aleatorie. Osserviamo inoltre che, nella proprietà (iii), l'indipendenza di X e Y è solo un requisito *sufficiente* per aversi $\mathbb{E}[XY] = \mathbb{E}[X] \mathbb{E}[Y]$, ma *non è necessario*. In altre parole, esistono variabili aleatorie X e Y *non indipendenti* per cui vale comunque $\mathbb{E}[XY] = \mathbb{E}[X] \mathbb{E}[Y]$ (trovarne una coppia per esercizio!).

Esempio 28 (Media della normale). Supponiamo per cominciare che $Z_0 \sim N(0, 1)$. Osservando che la funzione $xe^{-\frac{x^2}{2}}$ è antisimmetrica rispetto all'asse delle y troviamo

$$\mathbb{E}[Z_0] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} xe^{-\frac{x^2}{2}} dx = 0.$$

Se invece $Z \sim N(\mu, \sigma^2)$, allora abbiamo l'identità

$$Z = \sigma \frac{Z - \mu}{\sigma} + \mu =: \sigma Z_0 + \mu$$

dove la variabile aleatoria $Z_0 := (Z - \mu)/\sigma$ è la standardizzazione di Z e ha pertanto densità $N(0, 1)$. Quindi, per la linearità della media e per quanto appena visto per la normale standard,

$$\mathbb{E}[Z] = \mathbb{E}[\sigma Z_0 + \mu] = \sigma \mathbb{E}[Z_0] + \mu = \sigma \cdot 0 + \mu = \mu.$$

Definizione 7. La *covarianza* di due variabili aleatorie X e Y è il numero reale $\text{Cov}(X, Y)$ definito come segue

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

Da notare che, al contrario della media, la covarianza ha come argomento *due* variabili aleatorie, e non una sola.

Dalla Proposizione 4 possiamo ricavare l'espressione esplicita di $\text{Cov}(X, Y)$ nei due casi in cui (X, Y) è un vettore aleatorio discreto

$$\text{Cov}(X, Y) = \sum_{x, y} (x - \mathbb{E}[X])(y - \mathbb{E}[Y])p_{(X, Y)}(x, y)$$

oppure assolutamente continuo

$$\text{Cov}(X, Y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x - \mathbb{E}[X])(y - \mathbb{E}[Y])f_{(X, Y)}(x, y) dx dy.$$

La proposizione seguente riassume le principali proprietà della covarianza.

Proposizione 6 (Proprietà della covarianza). (i) $\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$ (*formula alternativa della covarianza*).

(ii) $\text{Cov}(X, Y) = \text{Cov}(Y, X)$ per ogni coppia di variabili aleatorie X, Y (*simmetria*).

(iii) Se X è una variabile aleatoria costante, allora $\text{Cov}(X, Y) = 0$.

(iv) $\text{Cov}(aX + bY, Z) = a\text{Cov}(X, Z) + b\text{Cov}(Y, Z)$ per ogni tripla di variabili aleatorie X, Y, Z e coppia di numeri reali a, b ; la stessa proprietà vale anche per il secondo argomento (*bilinearità*).

(v) Se X e Y sono indipendenti, allora $\text{Cov}(X, Y) = 0$.

Dimostrazione. (i) Abbiamo

$$\begin{aligned} \text{Cov}(X, Y) &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= \mathbb{E}[XY - \mathbb{E}[Y]X - \mathbb{E}[X]Y + \mathbb{E}[X]\mathbb{E}[Y]] && \text{svolgendo il prodotto} \\ &= \mathbb{E}[XY] - \mathbb{E}[Y]\mathbb{E}[X] - \mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y]\mathbb{E}[1] && \text{per la linearità della media} \\ &= \mathbb{E}[XY] - \mathbb{E}[Y]\mathbb{E}[X] - \mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y] && \text{perché } \mathbb{E}[1] = 1 \\ &= \mathbb{E}[XY] - \mathbb{E}[Y]\mathbb{E}[X]. \end{aligned}$$

(ii) Per la definizione,

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

(iii) Se $c \in \mathbb{R}$ e $X \equiv c$, allora sappiamo che $\mathbb{E}[c] = c$, e pertanto

$$\text{Cov}(X, Y) = \mathbb{E}[(c - \mathbb{E}[c])(Y - \mathbb{E}[Y])] = \mathbb{E}[0 \cdot (Y - \mathbb{E}[Y])] = \mathbb{E}[0] = 0.$$

(iv) Usando la formula alternativa della covarianza, abbiamo

$$\begin{aligned} \text{Cov}(aX + bY, Z) &= \mathbb{E}[(aX + bY)Z] - \mathbb{E}[aX + bY]\mathbb{E}[Z] \\ &= \mathbb{E}[aXZ + bYZ] - \mathbb{E}[aX + bY]\mathbb{E}[Z] \\ &= a\mathbb{E}[XZ] + b\mathbb{E}[YZ] - (a\mathbb{E}[X] + b\mathbb{E}[Y])\mathbb{E}[Z] \quad \text{linearità di } \mathbb{E} \\ &= a(\mathbb{E}[XZ] - \mathbb{E}[X]\mathbb{E}[Z]) + b(\mathbb{E}[YZ] - \mathbb{E}[Y]\mathbb{E}[Z]) \\ &= a\text{Cov}(X, Z) + b\text{Cov}(Y, Z). \end{aligned}$$

La proprietà analoga per il secondo argomento segue dalla simmetria di Cov .

(v) Se X e Y sono indipendenti, abbiamo visto fra le proprietà della meda che $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$. Perciò, usando ancora la formula alternativa della covarianza, $\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0$. ■

Osservazione 6. Attenzione! La proprietà (v) della covarianza significa l'implicazione

$$X, Y \text{ sono indipendenti} \Rightarrow \text{Cov}(X, Y) = 0$$

ma l'implicazione inversa *non è vera!* Per esempio, se (X, Y) è un vettore aleatorio discreto a valori in $S = \{0, 1\} \times \{0, 1, 2\}$ e con densità congiunta data dalla seguente tabella

$X \setminus Y$	0	1	2
0	1/12	1/3	1/12
1	1/4	0	1/4

allora le variabili aleatorie X e Y *non* sono indipendenti, tuttavia $\text{Cov}(X, Y) = 0$ (verificarlo!).

Definizione 8. La *varianza* di una variabile aleatoria X è il numero reale $\text{Var}(X)$ dato da

$$\text{Var}(X) = \text{Cov}(X, X) = \mathbb{E}[(X - \mathbb{E}[X])^2].$$

L'espressione esplicita di $\text{Var}(X)$ nei due casi in cui X è discreta o assolutamente continua può ancora essere ricavata tramite la Proposizione 4, ottenendo

$$\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \begin{cases} \sum_x (x - \mathbb{E}[X])^2 p_X(x) & \text{se } X \text{ è discreta} \\ \int_{-\infty}^{+\infty} (x - \mathbb{E}[X])^2 f_X(x) dx & \text{se } X \text{ è assolutamente continua} \end{cases}$$

Notiamo in particolare che in entrambi i casi $\text{Var}(X)$ è *sempre un numero positivo*, in quanto somma o integrale di quantità positive. Possiamo prenderne pertanto la radice quadrata: la quantità $\sqrt{\text{Var}(X)}$ si chiama *deviazione standard* di X .

La varianza gode delle seguenti proprietà fondamentali, che vanno confrontate con le analoghe proprietà della media.

Proposizione 7 (Proprietà della varianza). (i) $\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$ (formula alternativa della varianza).

(ii) Se X è una variabile aleatoria costante, allora $\text{Var}(X) = 0$.

(iii) Se a, b sono numeri reali, allora $\text{Var}(aX + b) = a^2 \text{Var}(X)$.

(iv) Se X_1, X_2, \dots, X_n sono n variabili aleatorie, la varianza della loro somma è

$$\text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var}(X_i) + 2 \sum_{\substack{i,j=1 \\ i < j}}^n \text{Cov}(X_i, X_j). \quad (1.12)$$

In particolare, se le variabili aleatorie X_1, X_2, \dots, X_n sono indipendenti

$$\text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var}(X_i). \quad (1.13)$$

Dimostrazione. (i) Segue dalla formula alternativa della covarianza $\text{Cov}(X, X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.

(ii) Immediata dall'analogia proprietà della covarianza.

(iii) Abbiamo

$$\begin{aligned} \text{Var}(aX + b) &= \text{Cov}(aX + b, aX + b) \\ &= a^2 \text{Cov}(X, X) + a \text{Cov}(b, X) + a \text{Cov}(X, b) + \text{Cov}(b, b) \quad \text{bilinearità di Cov} \\ &= a^2 \text{Cov}(X, X) \quad \text{perché Cov}(X, Y) = 0 \text{ se } Y \text{ è costante} \\ &= a^2 \text{Var}(X). \end{aligned}$$

(iv) Si ha

$$\begin{aligned} \text{Var}\left(\sum_{i=1}^n X_i\right) &= \text{Cov}\left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i\right) = \sum_{i,j=1}^n \text{Cov}(X_i, X_j) \quad \text{bilinearità di Cov} \\ &= \sum_{i=1}^n \text{Cov}(X_i, X_i) + \sum_{\substack{i,j=1 \\ i \neq j}}^n \text{Cov}(X_i, X_j) \\ &= \sum_{i=1}^n \text{Var}(X_i) + 2 \sum_{\substack{i,j=1 \\ i < j}}^n \text{Cov}(X_i, X_j) \quad \text{definizione di Var e simmetria di Cov} \end{aligned}$$

Se in più le X_1, X_2, \dots, X_n sono indipendenti, allora $\text{Cov}(X_i, X_j) = 0$ per ogni $i \neq j$, e la formula precedente si riscrive come in (1.13). ■

Nel caso di due sole variabili aleatorie X e Y , la formula (1.12) per la varianza della loro somma si riscrive

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

Esempio 29 (Varianza della bernoulliana). Se $X \sim B(1, p)$, allora X può prendere solo i due valori 0 oppure 1, e di conseguenza $X^2 = X$. Perciò, usando la formula alternativa della varianza,

$$\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathbb{E}[X] - \mathbb{E}[X]^2 = p - p^2.$$

Esempio 30 (Media e varianza della binomiale). Se $X \sim B(n, p)$, allora possiamo vedere X come la variabile aleatoria che conta il numero di successi in n prove di Bernoulli E_1, E_2, \dots, E_n con probabilità di successo $\mathbb{P}(E_i) = p$ per ciascuna prova. In questo modo, X è la somma

$$X = \mathbb{1}_{E_1} + \mathbb{1}_{E_2} + \dots + \mathbb{1}_{E_n},$$

dove le variabili aleatorie $\mathbb{1}_{E_1}, \mathbb{1}_{E_2}, \dots, \mathbb{1}_{E_n}$ sono tutte indipendenti e identicamente distribuite (i.i.d.), ciascuna con densità $\mathbb{1}_{E_i} \sim B(1, p)$. Per la linearità della media abbiamo pertanto

$$\begin{aligned} \mathbb{E}[X] &= \mathbb{E}[\mathbb{1}_{E_1} + \mathbb{1}_{E_2} + \dots + \mathbb{1}_{E_n}] = \mathbb{E}[\mathbb{1}_{E_1}] + \mathbb{E}[\mathbb{1}_{E_2}] + \dots + \mathbb{E}[\mathbb{1}_{E_n}] = p + p + \dots + p \\ &= np. \end{aligned}$$

Dalla formula (1.13) per la varianza della somma di variabili aleatorie *indipendenti* otteniamo invece

$$\begin{aligned} \text{Var}(X) &= \text{Var}(\mathbb{1}_{E_1}) + \text{Var}(\mathbb{1}_{E_2}) + \dots + \text{Var}(\mathbb{1}_{E_n}) = p(1-p) + p(1-p) + \dots + p(1-p) \\ &= np(1-p). \end{aligned}$$

Esempio 31 (Varianza dell'esponenziale). Se $T \sim \mathcal{E}(\lambda)$, integrando per parti due volte abbiamo

$$\begin{aligned} \mathbb{E}[T^2] &= \int_0^{+\infty} x^2 \lambda e^{-\lambda x} dx = [-x^2 e^{-\lambda x}]_{x=0}^{x=+\infty} + \int_0^{+\infty} 2x e^{-\lambda x} dx = \int_0^{+\infty} 2x e^{-\lambda x} dx \\ &= \left[-\frac{2x}{\lambda} e^{-\lambda x} \right]_{x=0}^{x=+\infty} + \int_0^{+\infty} \frac{2}{\lambda} e^{-\lambda x} dx = \int_0^{+\infty} \frac{2}{\lambda} e^{-\lambda x} dx \\ &= \left[-\frac{2}{\lambda^2} e^{-\lambda x} \right]_{x=0}^{x=+\infty} = \frac{2}{\lambda^2} \end{aligned}$$

e quindi

$$\text{Var}(T) = \mathbb{E}[T^2] - \mathbb{E}[T]^2 = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}.$$

Esempio 32 (Varianza della normale). Supponiamo innanzitutto $Z_0 \sim N(0, 1)$. Sappiamo allora che $\mathbb{E}[Z_0] = 0$, e quindi

$$\begin{aligned} \text{Var}(Z_0) &= \mathbb{E}[Z_0^2] - \mathbb{E}[Z_0]^2 = \mathbb{E}[Z_0^2] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} x^2 e^{-\frac{x^2}{2}} dx \\ &= \frac{1}{\sqrt{2\pi}} \left[-x e^{-\frac{x^2}{2}} \right]_{x=-\infty}^{x=+\infty} + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} dx \quad \text{integrando per parti} \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} dx = \int_{-\infty}^{+\infty} f_{Z_0}(x) dx \\ &= 1 \quad \text{perché } f_{Z_0} \text{ è normalizzata.} \end{aligned}$$

Se ora $Z \sim N(\mu, \sigma^2)$, procedendo come nell'Esempio 28 abbiamo

$$Z = \sigma Z_0 + \mu \quad \text{con } Z_0 := \frac{Z - \mu}{\sigma} \sim N(0, 1)$$

e quindi per la proprietà (iii) della varianza

$$\text{Var}(Z) = \text{Var}(\sigma Z_0 + \mu) = \sigma^2 \text{Var}(Z_0) = \sigma^2.$$

Per comodità del lettore, nella seguente tabella riassuntiva mettiamo a confronto le principali proprietà della media, della varianza e della covarianza.

	$\mathbb{E}[Y]$	$\text{Var}(Y)$	$\text{Cov}(Y, Z)$
$Y = c$	c	0	0
$Y = aX + b$	$a\mathbb{E}[X] + b$	$a^2\text{Var}(X)$	$a\text{Cov}(X, Z)$
$Y = X_1 + X_2$	$\mathbb{E}[X_1] + \mathbb{E}[X_2]$	$\text{Var}(X_1) + \text{Var}(X_2) + 2\text{Cov}(X_1, X_2)$	$\text{Cov}(X_1, Z) + \text{Cov}(X_2, Z)$
Y, Z indep.	$\mathbb{E}[Y]$	$\text{Var}(Y)$	0

1.10 Disuguaglianza di Chebyshev e legge dei grandi numeri

La varianza di una variabile aleatoria X è un numero reale che misura quanto la densità di X si disperde intorno al suo valor medio. Ciò è abbastanza chiaro dalla formula esplicita di $\text{Var}(X)$: per esempio, se X è discreta, allora

$$\text{Var}(X) = \sum_x (x - \mathbb{E}[X])^2 p_X(x)$$

è tanto più grande quanto più la densità p_X ‘pesa’ i punti x che si trovano lontano da $\mathbb{E}[X]$, cioè i punti per i quali la distanza $|x - \mathbb{E}[X]|$ al quadrato è grande. Ciò si vede particolarmente bene nell’esempio in cui X ha densità normale $N(\mu, \sigma^2)$; in tal caso, infatti, semplicemente disegnando il grafico della densità $f_X(x) = e^{-\frac{(x-\mu)^2}{2\sigma^2}}/(\sqrt{2\pi}\sigma)$, si osserva subito che f_X è tanto più allargata intorno alla media $\mathbb{E}[X] = \mu$ quanto più la varianza $\text{Var}(X) = \sigma^2$ è grande. Un esempio ancora più semplice e illuminante è quello di una variabile aleatoria X discreta che prende valori nell’insieme di due numeri reali $\{-a, a\}$ (dove $a > 0$) con densità $p_X(-a) = p_X(a) = 1/2$. In tal caso, infatti, $\mathbb{E}[X] = 0$ per ogni a , mentre la varianza $\text{Var}(X) = a^2$ cresce col quadrato della distanza delle due masse.

Un modo alternativo di quantificare il grado di dispersione di una variabile aleatoria X intorno al suo valor medio è il seguente: fissiamo un numero k a nostro piacimento, e calcoliamo la probabilità che X si discosti da $\mathbb{E}[X]$ per più di k volte la sua deviazione standard. Tanto più grande è tale probabilità, tanto maggiore sarà la dispersione di X intorno al suo valor medio. Per esempio, se fissiamo $k = 3$ la nostra misura di dispersione è la quantità $\mathbb{P}\left(|X - \mathbb{E}[X]| > 3\sqrt{\text{Var}(X)}\right)$, cioè la probabilità che X si discosti dalla sua media $\mathbb{E}[X]$ per più di 3 volte la deviazione standard. Tanto maggiore è tale probabilità, tanto più ci aspettiamo che X sia dispersa intorno a $\mathbb{E}[X]$.

Le due misure di dispersione precedenti sono strettamente legate tra di esse. Infatti, la *disuguaglianza di Chebyshev* che ora enunceremo stabilisce la loro relazione.

Proposizione 8 (Disuguaglianza di Chebyshev). *Sia X una variabile aleatoria qualsiasi. Allora, per ogni $k > 0$,*

$$\mathbb{P}\left(|X - \mathbb{E}[X]| > k\sqrt{\text{Var}(X)}\right) \leq \frac{1}{k^2}.$$

Dimostrazione. Supporremo per semplicità che X sia una variabile aleatoria discreta a valori nell’insieme S (il caso assolutamente continuo è del tutto analogo). Denotiamo in breve $\mu = \mathbb{E}[X]$

e $\sigma^2 = \text{Var}(X)$. Si ha

$$\begin{aligned}
 \sigma^2 &= \sum_{x \in S} (x - \mu)^2 p_X(x) = \sum_{x: |x - \mu| \leq k\sigma} (x - \mu)^2 p_X(x) + \sum_{x: |x - \mu| > k\sigma} (x - \mu)^2 p_X(x) \\
 &\geq \sum_{x: |x - \mu| > k\sigma} (x - \mu)^2 p_X(x) \quad \text{perché} \quad \sum_{x: |x - \mu| \leq k\sigma} (x - \mu)^2 p_X(x) \text{ è positiva} \\
 &\geq \sum_{x: |x - \mu| > k\sigma} (k\sigma)^2 p_X(x) \quad \text{perché se } |x - \mu| > k\sigma \text{ allora } (x - \mu)^2 > (k\sigma)^2 \\
 &= k^2 \sigma^2 \sum_{x: |x - \mu| > k\sigma} p_X(x) \\
 &= k^2 \sigma^2 \mathbb{P}(|X - \mu| > k\sigma)
 \end{aligned}$$

e l'enunciato segue immediatamente dividendo ambo i membri per $k^2 \sigma^2$. ■

Con $k = 3$, la disuguaglianza di Chebyshev ci dice per esempio che la probabilità che X disti dal suo valor medio per più di 3 volte la deviazione standard è minore o uguale a $1/9$, o, equivalentemente, che gli $8/9$ della densità di X sono concentrati entro una distanza di $3\sqrt{\text{Var}(X)}$ dalla media. Da notare che questo vale *qualunque* sia la densità di X , quindi anche nel caso in cui la densità è incognita e l'unico dato disponibile è la sua varianza.

Introduciamo ora un concetto che diventerà fondamentale nella parte di statistica.

Definizione 9. Un *campione aleatorio* di numerosità n è una successione di variabili aleatorie X_1, X_2, \dots, X_n tali che

- (a) le variabili aleatorie X_1, X_2, \dots, X_n sono indipendenti;
- (b) tutte le variabili aleatorie X_1, X_2, \dots, X_n hanno la stessa densità.

Un campione aleatorio è pertanto una successione di variabili aleatorie *indipendenti e identicamente distribuite* (i.i.d.). In particolare, per l'uguaglianza delle loro densità, tutte le X_i hanno la stessa media e la stessa varianza: $\mathbb{E}[X_i] = \mathbb{E}[X_j]$ e $\text{Var}(X_i) = \text{Var}(X_j)$ per ogni $i \neq j$, in quanto sia la media che la varianza dipendono solo dalla densità.

Per chiarire ancora meglio la definizione, supponiamo per fissare le idee che le X_i siano tutte assolutamente continue e ciascuna abbia densità f_{X_i} . Allora il punto (b) richiede che $f_{X_i} = f_{X_j} =: f$ per ogni $i \neq j$, dove f è la densità comune. Il punto (a) significa invece che la densità congiunta è

$$f_{(X_1, X_2, \dots, X_n)}(x_1, x_2, \dots, x_n) = f_{X_1}(x_1) f_{X_2}(x_2) \dots f_{X_n}(x_n) = f(x_1) f(x_2) \dots f(x_n).$$

Esempio 33. In ciascuno degli esperimenti aleatori seguenti, la successione di variabili aleatorie X_1, X_2, \dots, X_n è un esempio di campione aleatorio.

- (i) Prendiamo a caso un gruppo di n maschi adulti della stessa popolazione e indichiamo con X_i la variabile aleatoria

$$X_i = \text{altezza dell}'i\text{-esimo individuo.}$$

- (ii) Se abbiamo una scatola contenente n lampadine tutte della stessa marca e dello stesso modello, poniamo

$$X_i = \text{durata dell}'i\text{-esima lampadina prima di bruciarsi.}$$

- (iii) Lanciamo uno dado per n volte e definiamo

$$X_i = \text{risultato dell}'i\text{-esimo lancio.}$$

- (iv) In un imballaggio contenente n scatole da 10 DVD ciascuna, chiamiamo

$$X_i = \text{numero di DVD guasti nell}'i\text{-esima scatola.}$$

Nei primi due esempi, le X_i hanno densità assolutamente continua (tipicamente, gaussiana in (i) e esponenziale in (ii)), mentre negli ultimi due la loro densità è discreta (uniforme in (iii) e binomiale in (iv)).

La *media campionaria* del campione X_1, X_2, \dots, X_n è la variabile aleatoria

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

Notiamo che la media di \bar{X}_n è la stessa di una qualunque delle X_i . In altre parole, abbiamo

$$\mathbb{E}[\bar{X}_n] = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[X_i] = \frac{1}{n} \cdot n \mathbb{E}[X_1] = \mathbb{E}[X_1].$$

La varianza di \bar{X}_n è invece la varianza delle X_i *riscalata di un fattore* $1/n$. Infatti, ricordando che le X_i sono indipendenti, la formula (1.13) ci dà

$$\begin{aligned} \text{Var}(\bar{X}_n) &= \text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \left(\frac{1}{n}\right)^2 \text{Var}\left(\sum_{i=1}^n X_i\right) \quad \text{perché } \text{Var}(aX + b) = a^2 \text{Var}(X) \\ &= \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) \quad \text{per la formula (1.13)} \\ &= \frac{1}{n^2} \cdot n \text{Var}(X_1) = \frac{\text{Var}(X_1)}{n}. \end{aligned}$$

Pertanto, mentre la media di \bar{X}_n rimane la stessa delle X_i , la sua dispersione si riduce di un fattore $1/n$. In altre parole, la densità di probabilità della media campionaria si “stringe” intorno al valor medio delle X_i con una larghezza (= deviazione standard) che scala come $1/\sqrt{n}$. Ciò è alla base del seguente teorema.

Teorema 4 (Legge dei Grandi Numeri). *Supponiamo che X_1, X_2, \dots sia un campione aleatorio. Allora per ogni $\epsilon > 0$ si ha*

$$\lim_{n \rightarrow \infty} \mathbb{P}(|\bar{X}_n - \mathbb{E}[X_1]| > \epsilon) = 0.$$

Dimostrazione. Usiamo la disuguaglianza di Chebyshev per la variabile aleatoria \bar{X}_n , che abbiamo visto avere media $\mathbb{E}[\bar{X}_n] = \mathbb{E}[X_1]$ e varianza $\text{Var}(\bar{X}_n) = \text{Var}(X_1)/n$. Ponendo $k = \epsilon/\sqrt{\text{Var}(\bar{X}_n)}$ nella disuguaglianza di Chebyshev, abbiamo pertanto

$$\mathbb{P}(|\bar{X}_n - \mathbb{E}[X_1]| > \epsilon) = \mathbb{P}\left(|\bar{X}_n - \mu| > k\sqrt{\text{Var}(\bar{X}_n)}\right) \leq \frac{1}{k^2} = \frac{\text{Var}(\bar{X}_n)}{\epsilon^2} = \frac{\text{Var}(X_1)}{n\epsilon^2}.$$

Dal momento che $\lim_{n \rightarrow \infty} \text{Var}(X_1)/(n\epsilon^2) = 0$ e poiché la probabilità è sempre un numero non negativo, passando al limite per $n \rightarrow \infty$ in entrambi i membri della disuguaglianza otteniamo $\lim_{n \rightarrow \infty} \mathbb{P}(|\bar{X}_n - \mathbb{E}[X_1]| > \epsilon) = 0$. ■

1.11 Teorema del Limite Centrale

Se X_1, X_2, \dots, X_n è un campione aleatorio *gaussiano*, cioè $X_i \sim N(\mu, \sigma^2)$ per ogni $i = 1, 2, \dots, n$ (con le stesse μ e σ per tutti gli i !), allora la Proposizione 3 ci dice che la media campionaria $\bar{X}_n = (X_1 + X_2 + \dots + X_n)/n$ è anch'essa una variabile aleatoria gaussiana, in quanto combinazione lineare di normali indipendenti. Poiché già sappiamo che $\mathbb{E}[\bar{X}_n] = \mathbb{E}[X_1] = \mu$ e $\text{Var}(\bar{X}_n) = \text{Var}(X_1)/n = \sigma^2/n$, avremo pertanto $\bar{X}_n \sim N(\mu, \sigma^2/n)$. Il seguente fondamentale teorema ci dice che questo fatto vale approssimativamente anche quando le X_i hanno densità arbitraria (e non necessariamente gaussiana), a patto però che il campione sia abbastanza numeroso.

Teorema 5 (Teorema del Limite Centrale). *Sia X_1, X_2, \dots un campione aleatorio qualsiasi. Allora*

$$\lim_{n \rightarrow \infty} \mathbb{P} \left(\frac{\bar{X}_n - \mathbb{E}[X_1]}{\sqrt{\text{Var}(X_1)}} \sqrt{n} \leq x \right) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{z^2}{2}} dz \quad \text{per ogni } x \in \mathbb{R}.$$

La dimostrazione è molto complicata e la ometteremo. Sottolineiamo però ancora una seconda volta che il significato del Teorema del Limite Centrale (TLC) è il seguente: se X_1, X_2, \dots, X_n è un campione aleatorio *qualsiasi e non necessariamente gaussiano* ma n è abbastanza grande, allora la sua media campionaria \bar{X}_n ha comunque approssimativamente densità gaussiana. Chiaramente, tale densità sarà $N(\mathbb{E}[X_1], \text{Var}(X_1)/n)$, perché $\mathbb{E}[X_1]$ e $\text{Var}(X_1)/n$ sono la media e la varianza di \bar{X}_n . Scriveremo in questo caso $\bar{X}_n \approx N(\mathbb{E}[X_1], \text{Var}(X_1)/n)$. Tipicamente, tale approssimazione vale già piuttosto bene quando $n \geq 30$.

Una forma alternativa, ma equivalente, del TLC afferma che, per un campione aleatorio X_1, X_2, \dots, X_n qualsiasi, quando n è abbastanza grande la somma $S_n = X_1 + X_2 + \dots + X_n$ ha approssimativamente densità $N(n\mathbb{E}[X_1], n\text{Var}(X_1))$. Infatti, $S_n = n\bar{X}_n$ è gaussiana perché trasformazione affine della variabile aleatoria \bar{X}_n che è gaussiana per il TLC, e

$$\mathbb{E}[S_n] = n\mathbb{E}[\bar{X}_n] = n\mathbb{E}[X_1], \quad \text{Var}(S_n) = n^2\text{Var}(\bar{X}_n) = n\text{Var}(X_1).$$

Tale forma alternativa del TLC giustifica il ruolo particolarmente importante che la densità gaussiana assume in tutta la teoria della probabilità e la sua grande rilevanza pratica. Infatti, in un modello un po' semplificato, possiamo assumere che l'errore che si commette nel misurare una fissata quantità fisica sia in realtà la somma $S_n = X_1 + X_2 + \dots + X_n$ di tanti errori più piccoli X_1, X_2, \dots, X_n , tutti indipendenti tra loro e identicamente distribuiti, ma con una densità che in generale è incognita. Il TLC ci dice allora che l'errore totale S_n è comunque approssimabile con una gaussiana, indipendentemente dalla densità incognita di ciascun contributo X_i .

Infine, sottolineiamo di nuovo che, se le X_i sono già normali per conto loro, allora la relazione $S_n \sim N(n\mathbb{E}[X_1], n\text{Var}(X_1))$ è *esatta* e non c'è bisogno del TLC per dimostrarla.

Osservazione 7. Un errore abbastanza comune (ed *enormemente grave*) quando si studia Statistica è pensare che per un campione aleatorio X_1, X_2, \dots, X_n le due variabili aleatorie $S_n = X_1 + X_2 + \dots + X_n$ e $T_n = nX_1$ siano la stessa cosa. Ciò non è assolutamente vero, in quanto:

- S_n e T_n hanno densità completamente diversa, e per convincersene basta confrontare $\text{Var}(S_n) = n\text{Var}(X_1)$ con $\text{Var}(T_n) = n^2\text{Var}(X_1)$;
- solo per S_n vale il TLC.

1.12 Approssimazione normale e poissoniana della binomiale

Supponiamo che X sia una variabile aleatoria binomiale di parametri n e p . In questa sezione vedremo come la densità di X può essere approssimata quando n è molto grande in due casi di particolare interesse pratico.

Il primo caso è quello in cui $n \rightarrow \infty$ e p resta costante e finita.

Proposizione 9 (Approssimazione normale della binomiale). *Supponiamo $X \sim B(n, p)$ con n molto grande e p non trascurabile. Allora $X \approx N(np, np(1-p))$.*

Dimostrazione. Possiamo considerare X come la somma di n bernoulliane di parametro p indipendenti. In altre parole, $X = X_1 + X_2 + \dots + X_n$, dove X_1, X_2, \dots, X_n è un campione aleatorio e ciascuna $X_i \sim B(1, p)$. Per il TLC, $X \approx N(n\mathbb{E}[X_1], n\text{Var}(X_1)) = N(np, np(1-p))$, dove p e $p(1-p)$ sono rispettivamente la media e la varianza della variabile aleatoria bernoulliana X_1 . ■

Come regola di massima per decidere se vale l'approssimazione precedente, si considera di solito che n deve essere abbastanza grande e p abbastanza diverso da 0 o da 1 aversi $np \geq 5$ e $n(1-p) \geq 5$.

Il secondo caso è invece quello in cui $n \rightarrow \infty$ e $p \rightarrow 0$, in modo però che il prodotto np resti costante e dell'ordine dell'unità.

Proposizione 10 (Approssimazione di Poisson della binomiale). *Supponiamo $X \sim B(n, p)$ con n molto grande e p infinitesimo, in modo però che il prodotto $\lambda := np$ sia confrontabile con 1. Allora X ha approssimativamente densità*

$$p_X(k) \simeq e^{-\lambda} \frac{\lambda^k}{k!} \quad \text{per ogni } k = 0, 1, 2, \dots$$

Dimostrazione. Vogliamo calcolare la densità p_X nel limite

$$n \rightarrow \infty, \quad p \rightarrow 0, \quad np = \lambda = \text{costante} \approx 1.$$

Abbiamo

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

e

$$\begin{aligned} \lim_{n \rightarrow \infty} \left(1 - \frac{\lambda}{n}\right)^n &= e^{-\lambda} & \lim_{n \rightarrow \infty} \left(1 - \frac{\lambda}{n}\right)^{-k} &= 1 \\ \lim_{n \rightarrow \infty} \frac{n!}{(n-k)! n^k} &= \lim_{n \rightarrow \infty} \frac{n(n-1)(n-2) \dots (n-k+1)}{n^k} = \lim_{n \rightarrow \infty} \frac{n^k + O(n^{k-1})}{n^k} = 1. \end{aligned}$$

Pertanto

$$\lim_{n \rightarrow \infty} p_X(k) = \frac{\lambda^k}{k!} e^{-\lambda} \quad \text{per ogni } k = 0, 1, 2, \dots$$

■

Nell'approssimazione di Poisson, dire che il prodotto np dev'essere confrontabile con 1 significa che valori come $np = 0.5$, $np = 1.4$, $np = 5$ vanno bene, ma non vanno bene valori come $np = 0.01$ o $np = 100$. La regola di massima in questo caso è che se per esempio $n \geq 20$ allora deve essere $p \leq 0.05$, oppure se $n \geq 100$ allora $np \leq 10$.

La densità $p_X(k) = e^{-\lambda} \lambda^k / k!$, definita sull'insieme dei numeri naturali $\mathbb{N} = \{0, 1, 2, \dots\}$, si chiama *densità di Poisson* di parametro λ e si indica con $\mathcal{P}(\lambda)$. Notiamo che, poiché tale densità si ottiene come limite da una variabile aleatoria binomiale $X \sim B(n, p)$, la sua media e la sua varianza sono

$$\mathbb{E}[X] = \lim_{\substack{n \rightarrow \infty, \, p \rightarrow 0 \\ np = \lambda}} np = \lambda \qquad \text{Var}(X) = \lim_{\substack{n \rightarrow \infty, \, p \rightarrow 0 \\ np = \lambda}} np(1 - p) = \lambda$$

in quanto np e $np(1 - p)$ sono rispettivamente la media e la varianza della densità $B(n, p)$.

Capitolo 2

Statistica

2.1 Intervalli di confidenza

Definizione 10. Sia X_1, \dots, X_n un campione aleatorio. Una *statistica* $T_n = t_n(X_1, \dots, X_n)$ è una funzione del vettore aleatorio $\vec{X} = (X_1, \dots, X_n)$.

Sono esempi di statistiche:

(i) la *media campionaria*

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

in cui la funzione t_n è

$$t_n(x_1, \dots, x_n) = \frac{1}{n} \sum_{i=1}^n x_i;$$

(ii) la *varianza campionaria*

$$S_n^2 := \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

in cui la funzione t_n è

$$t_n(x_1, \dots, x_n) = \frac{1}{n-1} \sum_{i=1}^n \left(x_i - \frac{1}{n} \sum_{i=1}^n x_i \right)^2;$$

(iii) la *deviazione standard campionaria* $S_n := \sqrt{S_n^2}$.

Definizione 11. Sia X_1, \dots, X_n un campione aleatorio. Supponiamo di aver fatto una misura di X_1, \dots, X_n e aver trovato i valori x_1, \dots, x_n ($x_i \in \mathbb{R}$). Allora diremo che x_1, \dots, x_n è la *realizzazione* del campione ottenuta nella nostra misura.

D'ora in poi, indicheremo ogni statistica valutata sulla nostra realizzazione del campione con la corrispondente lettera minuscola: $t_n = t_n(x_1, \dots, x_n)$, $\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i$, $s_n^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2$. Osserviamo che, mentre T_n , \bar{X}_n , S_n^2 sono tutte variabili aleatorie, le loro realizzazioni t_n , \bar{x}_n , s_n^2 sono *numeri reali*.

Esempio 34. Consideriamo l'esperimento aleatorio che consiste nel misurare l'altezza di $n = 8$ maschi adulti presi a caso dalla stessa popolazione, e indichiamo con X_i la variabile aleatoria

$$X_i = \text{altezza dell}'i\text{-esimo individuo.}$$

Supponiamo di ottenere le misure

$$x_1 = 1.82 \quad x_2 = 1.76 \quad x_3 = 1.78 \quad x_4 = 1.90 \quad x_5 = 1.85 \quad x_6 = 1.70 \quad x_7 = 1.68 \quad x_8 = 1.84.$$

Allora le nostre realizzazioni della media campionaria e della varianza campionaria sono rispettivamente

$$\bar{x}_n = \frac{1.82 + 1.76 + 1.78 + \dots + 1.84}{8} = 1.79125$$

e

$$\begin{aligned} s_n^2 &= \frac{(1.82 - 1.79125)^2 + (1.76 - 1.79125)^2 + (1.78 - 1.79125)^2 + \dots + (1.84 - 1.79125)^2}{8 - 1} \\ &= 0.005755357. \end{aligned}$$

Definizione 12. Sia X una variabile aleatoria continua con densità f_X . Dato $\alpha \in (0, 1)$, il *quantile* (destro) di ordine α della densità di X è il numero $x_\alpha \in \mathbb{R}$ t.c.

$$\mathbb{P}(X \geq x_\alpha) = \alpha.$$

Equivalentemente, il quantile x_α è il numero reale definito da

$$F_X(x_\alpha) = 1 - \alpha, \quad \int_{x_\alpha}^{+\infty} f_X(x) dx = \alpha,$$

dove F_X è la funzione di ripartizione di X . Si osservi che la funzione $\alpha \mapsto x_\alpha$ è *decescente*, in quanto al crescere di α deve aumentare l'area sottesa dalla densità alla destra di x_α , e quindi x_α deve spostarsi a sinistra.

Esempio 35. Alcuni quantili hanno nomi e significati particolari. P.es.,

- il quantile $x_{0.5}$ è la *mediana* della densità di X ;
- il quantile $x_{0.75}$ è il *primo quartile* della densità di X ;
- il quantile $x_{0.25}$ è il *terzo quartile* della densità di X .

Osservazione 8. Se la densità f_X è *simmetrica*, cioè $f_X(-x) = f_X(x) \forall x \in \mathbb{R}$, allora valgono in più le relazioni

$$x_{1-\alpha} = -x_\alpha, \quad F_X(-x_\alpha) = \alpha.$$

Infatti, dimostriamo per esempio la prima relazione:

$$\begin{aligned} \int_{-x_\alpha}^{+\infty} f_X(x) dx &= - \int_{x_\alpha}^{-\infty} f_X(-x') dx' && \text{col cambio di variabile } x' = -x \\ &= \int_{-\infty}^{x_\alpha} f_X(x') dx' && \text{proprietà dell'integrale e simmetria di } f_X \\ &= 1 - \int_{x_\alpha}^{+\infty} f_X(x') dx' && \text{normalizzazione di } f_X \\ &= 1 - \alpha. \end{aligned}$$

In particolare, le relazioni precedenti valgono se $X \sim N(0, 1)$.

Definizione 13. Sia X_1, \dots, X_n un campione aleatorio, dove le X_i sono variabili aleatorie assolutamente continue con densità

$$f_{X_i}(x) \equiv f(x \mid \theta)$$

dipendente da un parametro incognito θ . Siano $A_n = a_n(X_1, \dots, X_n)$ e $B_n = b_n(X_1, \dots, X_n)$ due statistiche per le quali valga sempre $A_n < B_n$, e tali che la probabilità dell'evento $\{A_n \leq \theta \leq B_n\}$ sia

$$\mathbb{P}(A_n \leq \theta \leq B_n) = 1 - \alpha,$$

dove $\alpha \in (0, 1)$ è un numero reale *che non dipende dal parametro θ* . Supponiamo di aver fatto una misura di X_1, \dots, X_n e aver trovato la realizzazione x_1, \dots, x_n . Allora diremo che l'intervallo

$$(a_n(x_1, \dots, x_n), b_n(x_1, \dots, x_n))$$

è un *intervallo di confidenza di livello $1 - \alpha$* per il parametro incognito θ (scriveremo per brevità $IC_\theta(1 - \alpha)$). Equivalentemente, si dice che

$$\theta \in (a_n(x_1, \dots, x_n), b_n(x_1, \dots, x_n))$$

con confidenza $1 - \alpha$.

IC bilatero per la media di una popolazione normale con varianza nota. Se $X_i \sim N(\mu, \sigma_0^2)$, dove σ_0^2 è nota e μ è il parametro incognito che vogliamo stimare, allora un $IC_\mu(1 - \alpha)$ è

$$\left(\bar{x}_n - z_{\alpha/2} \frac{\sigma_0}{\sqrt{n}}, \bar{x}_n + z_{\alpha/2} \frac{\sigma_0}{\sqrt{n}} \right),$$

dove $z_{\alpha/2}$ è il quantile destro di ordine $\alpha/2$ della normale standard, che è ricavabile dalle tavole delle distribuzioni.

Infatti, scegliendo le statistiche

$$A_n = \bar{X}_n - z_{\alpha/2} \frac{\sigma_0}{\sqrt{n}}, \quad B_n = \bar{X}_n + z_{\alpha/2} \frac{\sigma_0}{\sqrt{n}}$$

nella definizione di $IC_\mu(1 - \alpha)$, abbiamo

$$\begin{aligned}\mathbb{P}\left(\bar{X}_n - z_{\alpha/2} \frac{\sigma_0}{\sqrt{n}} \leq \mu \leq \bar{X}_n + z_{\alpha/2} \frac{\sigma_0}{\sqrt{n}}\right) &= \mathbb{P}\left(-z_{\alpha/2} \frac{\sigma_0}{\sqrt{n}} \leq \mu - \bar{X}_n \leq z_{\alpha/2} \frac{\sigma_0}{\sqrt{n}}\right) \\ &= \mathbb{P}\left(-z_{\alpha/2} \frac{\sigma_0}{\sqrt{n}} \leq \bar{X}_n - \mu \leq z_{\alpha/2} \frac{\sigma_0}{\sqrt{n}}\right) \\ &= \mathbb{P}\left(-z_{\alpha/2} \leq \frac{\bar{X}_n - \mu}{\sigma_0} \sqrt{n} \leq z_{\alpha/2}\right).\end{aligned}$$

Poichè per un campione gaussiano sappiamo che $\bar{X}_n \sim N(\mu, \sigma_0^2/n)$, e quindi standardizzando $(\bar{X}_n - \mu)\sqrt{n}/\sigma_0 \sim N(0, 1)$, abbiamo

$$\begin{aligned}\mathbb{P}\left(-z_{\alpha/2} \leq \frac{\bar{X}_n - \mu}{\sigma_0} \sqrt{n} \leq z_{\alpha/2}\right) &= \Phi(z_{\alpha/2}) - \Phi(-z_{\alpha/2}) \\ &= 1 - \frac{\alpha}{2} - \frac{\alpha}{2} \quad \text{perché } N(0, 1) \text{ è simmetrica} \Rightarrow \Phi(-z_{\alpha/2}) = \frac{\alpha}{2} \\ &= 1 - \alpha,\end{aligned}$$

esattamente come richiesto dalla definizione di $IC_\mu(1 - \alpha)$.

IC bilatero per la media di un campione numeroso qualunque con varianza nota. Per il TLC, se il campione ha densità arbitraria e non necessariamente gaussiana, posto $\mu = \mathbb{E}[X_1]$ e $\sigma_0^2 = \text{Var}(X_1)$, sappiamo che vale comunque

$$\frac{\bar{X}_n - \mu}{\sigma_0} \sqrt{n} \approx N(0, 1) \quad \text{per } n \gg 1.$$

Quindi, a patto che n sia sufficientemente grande, la dimostrazione precedente continua a valere (in modo approssimato) anche in questo caso, e permette di affermare che

$$\left(\bar{x}_n - z_{\alpha/2} \frac{\sigma_0}{\sqrt{n}}, \bar{x}_n + z_{\alpha/2} \frac{\sigma_0}{\sqrt{n}}\right)$$

è un $IC_\mu(1 - \alpha)$ *approssimato* per μ quando $n \gg 1$ e le X_i hanno densità qualunque.

Esempio 36. Un astronomo misura la distanza di una galassia con uno strumento che ha una precisione di $\sigma_0 = 0.7$ anni luce. Fa $n = 8$ misure indipendenti, che si possono supporre $X_i \sim N(\mu, \sigma_0^2)$, con

$$\begin{aligned}\mu &= \text{distanza } \textit{incognita} \text{ della galassia} \\ \sigma_0 &= \text{precisione } \textit{nota} \text{ dello strumento}\end{aligned}$$

trovando i valori (espressi in anni luce)

$$x_1 = 17.3 \quad x_2 = 16.5 \quad x_3 = 18.4 \quad x_4 = 17.7 \quad x_5 = 18.2 \quad x_6 = 18.7 \quad x_7 = 16.8 \quad x_8 = 18.2.$$

Con questi dati,

$$\bar{x}_8 = \frac{17.3 + 16.5 + \dots + 18.2}{8} = 17.725,$$

e un $IC_\mu(95\%)$ per la distanza incognita μ è (con $\alpha = 0.05$, $z_{\alpha/2} = z_{0.025} = 1.96$)

$$\left(17.725 - 1.96 \frac{0.7}{\sqrt{8}}, 17.725 + 1.96 \frac{0.7}{\sqrt{8}} \right) \simeq (17.24, 18.21).$$

Da notare che, se non avessimo fatto l'ipotesi a priori che la densità del campione è gaussiana, allora l'intervallo precedente *non* sarebbe stato un $IC_\mu(95\%)$ *nemmeno approssimato*, perché un campione di $n = 8$ misure è troppo piccolo per poter applicare il TLC.

Osservazione 9. Nell'esempio precedente, non ha alcun senso affermare che

$$\mathbb{P}(\mu \in (17.24, 18.21)) = 95\%$$

in quanto $\mu \in (17.24, 18.21)$ *non* è un evento e dunque non se ne può calcolare la probabilità. Infatti, il parametro μ *non* è una *variabile aleatoria*, ma è semplicemente un numero che o sta nell'intervallo $(17.24, 18.21)$ o non ci sta. Proprio per questo motivo si dice che $\mu \in (17.24, 18.21)$ con *confidenza* del 95%, e *non* con probabilità del 95%. Ha invece senso affermare che la *procedura* in cui:

1. effettuo $n = 8$ misure X_1, \dots, X_8 , e da queste determino la statistica \bar{X}_8 ;
2. costruisco l'intervallo $I = (\bar{X}_8 - 1.96 \frac{0.7}{\sqrt{8}}, \bar{X}_8 + 1.96 \frac{0.7}{\sqrt{8}})$

ha una probabilità del 95% di farmi ottenere un intervallo I che effettivamente contenga il parametro incognito μ .

IC unilateri per la media di una popolazione normale o numerosa con varianza nota.

Se $X_i \sim N(\mu, \sigma_0^2)$, dove σ_0^2 è nota e μ è incognita, allora i seguenti sono intervalli di confidenza *unilateri* di livello $1 - \alpha$ per μ :

(a) unilatero destro:

$$\left(\bar{x}_n - z_\alpha \frac{\sigma_0}{\sqrt{n}}, +\infty \right)$$

(b) unilatero sinistro:

$$\left(-\infty, \bar{x}_n + z_\alpha \frac{\sigma_0}{\sqrt{n}} \right).$$

P.es., dimostro (a):

$$\mathbb{P} \left(\mu \geq \bar{X}_n - z_\alpha \frac{\sigma_0}{\sqrt{n}} \right) = \mathbb{P} \left(\frac{\bar{X}_n - \mu}{\sigma_0} \sqrt{n} \leq z_\alpha \right) = \Phi(z_\alpha) = 1 - \alpha.$$

In modo analogo a quanto fatto nel caso bilatero, si dimostra che i precedenti intervalli sono anche $IC_\mu(1 - \alpha)$ approssimati per la media di un campione *arbitrario* (non necessariamente gaussiano) con varianza $\sigma_0^2 = \text{Var}(X_1)$ nota, quando la numerosità n del campione è sufficientemente grande.

Fin qui abbiamo trovato degli intervalli di confidenza per la media di un campione X_1, \dots, X_n solo nel caso in cui la varianza $\sigma_0^2 = \text{Var}(X_1)$ del campione stesso è nota. Vediamo adesso come si possono costruire degli intervalli di confidenza per la media anche nel caso in cui la varianza $\sigma^2 = \text{Var}(X_1)$ non è conosciuta a priori. In questo caso, useremo la varianza campionaria S_n^2 come stimatore di σ^2 . Un risultato importante in tal senso è il prossimo teorema.

Teorema 6 (Non dimostrato). *Se X_1, \dots, X_n è un campione aleatorio gaussiano, con $X_i \sim N(\mu, \sigma^2)$, allora*

$$\frac{\bar{X}_n - \mu}{S_n} \sqrt{n} \sim t(n-1),$$

dove $t(n-1)$ è la densità t di Student con $n-1$ gradi di libertà.

Per il quantile destro di ordine α della densità di Student $t(n)$ si usa la notazione $t_{\alpha,n}$. In altre parole, $t_{\alpha,n}$ è il numero reale definito dalla relazione

$$F_{t(n)}(t_{\alpha,n}) = 1 - \alpha,$$

dove $F_{t(n)}$ è la funzione di ripartizione della densità $t(n)$. I quantili $t_{\alpha,n}$ per i valori più comuni di α (p.es., $\alpha = 0.1$, $\alpha = 0.05$ e $\alpha = 0.025$) e i valori di n non troppo grandi (p.es., $n \leq 40$) si trovano tabulati su tutti i libri di Statistica.

Osservazione 10. Come la normale standard, anche la densità $t(n)$ è simmetrica rispetto all'asse delle y , dunque i suoi quantili soddisfano la relazione

$$t_{1-\alpha,n} = -t_{\alpha,n}.$$

IC per la media di una popolazione normale con varianza incognita. Sia X_1, \dots, X_n un campione con $X_i \sim N(\mu, \sigma^2)$, dove sia μ sia σ sono entrambi parametri incogniti. Allora i seguenti sono tutti intervalli di confidenza di livello $1 - \alpha$ per μ :

- (a) $\left(\bar{x}_n - t_{\alpha/2, n-1} \frac{s_n}{\sqrt{n}}, \bar{x}_n + t_{\alpha/2, n-1} \frac{s_n}{\sqrt{n}} \right);$
- (b) $\left(\bar{x}_n - t_{\alpha, n-1} \frac{s_n}{\sqrt{n}}, +\infty \right);$
- (c) $\left(-\infty, \bar{x}_n + t_{\alpha, n-1} \frac{s_n}{\sqrt{n}} \right).$

La dimostrazione è simile al caso con σ^2 nota. La sola differenza è che per costruire gli intervalli elencati sopra ora useremo la statistica $(\bar{X}_n - \mu)\sqrt{n}/S_n$ al posto di $(\bar{X}_n - \mu)\sqrt{n}/\sigma$. Per fissare le

idee, dimostriamo p.es. (a). Dal teorema precedente sappiamo che $(\bar{X}_n - \mu)\sqrt{n}/S_n \sim t(n-1)$, e quindi

$$\begin{aligned} \mathbb{P}\left(\bar{X}_n - t_{\alpha/2, n-1} \frac{S_n}{\sqrt{n}} \leq \mu \leq \bar{X}_n + t_{\alpha/2, n-1} \frac{S_n}{\sqrt{n}}\right) &= \mathbb{P}\left(-t_{\alpha/2, n-1} \leq \underbrace{\frac{\bar{X}_n - \mu}{S_n} \sqrt{n}}_{\sim t(n-1)} \leq t_{\alpha/2, n-1}\right) \\ &= F_{t(n-1)}(t_{\alpha/2, n-1}) - F_{t(n-1)}(-t_{\alpha/2, n-1}) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} \\ &= 1 - \alpha, \end{aligned}$$

dove $F_{t(n-1)}$ è la funzione di ripartizione della densità $t(n-1)$ e abbiamo usato la simmetria della t di Student per ricavare $F_{t(n-1)}(-t_{\alpha/2, n-1}) = \alpha/2$.

Esempio 37. Riprendiamo l'Esempio 36, e supponiamo sempre che il campione delle 8 misure X_1, \dots, X_8 sia gaussiano, con $X_i \sim N(\mu, \sigma^2)$, ma adesso rimuoviamo l'ipotesi che l'astronomo conosca a priori la precisione σ del suo strumento. Se ora vogliamo dare una stima di μ basandoci sui dati x_1, \dots, x_8 dell'esperimento, dovremo pertanto usare gli intervalli di confidenza appena ricavati. Per esempio, se desideriamo costruire un $IC_\mu(95\%)$ bilatero come nell'Esempio 36, dovremo ora utilizzare l'intervallo

$$\left(\bar{x}_n - t_{\frac{\alpha}{2}, n-1} \frac{s_n}{\sqrt{n}}, \bar{x}_n + t_{\frac{\alpha}{2}, n-1} \frac{s_n}{\sqrt{n}}\right).$$

Abbiamo già ricavato nell'esempio precedente $\bar{x}_8 = 17.725$. Resta da calcolare dai dati

$$\begin{aligned} s_n^2 &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2 = \frac{(17.3 - 17.725)^2 + (16.5 - 17.725)^2 + \dots + (18.2 - 17.725)^2}{8-1} \\ &\simeq 0.62786 \end{aligned}$$

e dalle tavole della t di Student

$$t_{\frac{\alpha}{2}, n-1} = t_{\frac{0.05}{2}, 8-1} = t_{0.025, 7} = 2.365.$$

Al 95% di confidenza avremo pertanto

$$\begin{aligned} \mu &\in \left(17.725 - 2.365 \frac{\sqrt{0.62786}}{\sqrt{8}}, 17.725 + 2.365 \frac{\sqrt{0.62786}}{\sqrt{8}}\right) \\ &\simeq (17.06, 18.39). \end{aligned}$$

Osservazione 11. Nell'esempio precedente, il calcolo della varianza campionaria S_n^2 poteva essere leggermente semplificato utilizzando la seguente formula alternativa

$$S_n^2 = \frac{1}{n-1} \left(\sum_{i=1}^n X_i^2 - n\bar{X}_n^2 \right) \quad (\text{formula alternativa per la varianza campionaria}).$$

Tale formula si ricava dalla definizione della varianza campionaria come segue:

$$\begin{aligned}
 S_n^2 &= \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \\
 &= \frac{1}{n-1} \left(\sum_{i=1}^n X_i^2 - 2 \sum_{i=1}^n X_i \bar{X}_n + \sum_{i=1}^n \bar{X}_n^2 \right) \\
 &= \frac{1}{n-1} \left(\sum_{i=1}^n X_i^2 - 2n\bar{X}_n \frac{1}{n} \sum_{i=1}^n X_i + \bar{X}_n^2 \sum_{i=1}^n 1 \right) \\
 &= \frac{1}{n-1} \left(\sum_{i=1}^n X_i^2 - 2n\bar{X}_n \cdot \bar{X}_n + \bar{X}_n^2 \cdot n \right) \\
 &= \frac{1}{n-1} \left(\sum_{i=1}^n X_i^2 - n\bar{X}_n^2 \right).
 \end{aligned}$$

Usando la formula alternativa della varianza campionaria, si dimostra facilmente la prossima proposizione.

Proposizione 11. *Sia X_1, \dots, X_n un campione qualsiasi, e sia $\sigma^2 := \text{Var}(X_i)$ la sua varianza. Allora $\mathbb{E}[S_n^2] = \sigma^2$. In altre parole, la varianza campionaria S_n^2 è uno stimatore non distorto della varianza vera σ^2 del campione.*

Dimostrazione. Utilizzando la formula alternativa ricavata nell'osservazione precedente, per la linearità della media

$$\begin{aligned}
 \mathbb{E}[S_n^2] &= \frac{1}{n-1} \left(\sum_{i=1}^n \mathbb{E}[X_i^2] - n\mathbb{E}[\bar{X}_n^2] \right) = \frac{1}{n-1} (n\mathbb{E}[X_1^2] - n\mathbb{E}[\bar{X}_n^2]) \\
 &= \frac{n}{n-1} (\mathbb{E}[X_1^2] - \mathbb{E}[\bar{X}_n^2]).
 \end{aligned}$$

Ricordiamo che, dalla formula alternativa per la varianza di una variabile aleatoria Y qualunque,

$$\mathbb{E}[Y^2] = \text{Var}(Y) + \mathbb{E}[Y]^2.$$

Inserendo questa relazione nell'espressione precedente,

$$\begin{aligned}
 \mathbb{E}[S_n^2] &= \frac{n}{n-1} \left[\text{Var}(X_1) + \mathbb{E}[X_1]^2 - \left(\text{Var}(\bar{X}_n) + \mathbb{E}[\bar{X}_n]^2 \right) \right] \\
 &= \frac{n}{n-1} \left[\sigma^2 + \mathbb{E}[X_1]^2 - \left(\frac{1}{n}\sigma^2 + \mathbb{E}[X_1]^2 \right) \right] \quad \text{perché } \text{Var}(\bar{X}_n) = \frac{\text{Var}(X_1)}{n} \text{ e } \mathbb{E}[\bar{X}_n] = \mathbb{E}[X_1] \\
 &= \frac{n}{n-1} \left(1 - \frac{1}{n} \right) \sigma^2 \\
 &= \sigma^2.
 \end{aligned}$$

□

Come conseguenza della Legge dei Grandi Numeri (LGN), la variabile aleatoria S_n^2 approssima bene la varianza vera del campione $\sigma^2 = \text{Var}(X_1)$ quando n è sufficientemente grande. In altre parole, se $n \gg 1$, la varianza campionaria S_n^2 è con buona approssimazione la variabile aleatoria costante e identicamente uguale a σ^2 . Ciò può essere euristicamente dimostrato nel modo che segue. Per la LGN, se n è grande abbiamo $\bar{X}_n \simeq \mathbb{E}[X_1] =: \mu$ (cioè, \bar{X}_n è approssimativamente la variabile aleatoria costante e uguale a μ). Perciò, la formula alternativa della varianza può essere riscritta

$$S_n^2 \simeq \frac{1}{n-1} \left(\sum_{i=1}^n X_i^2 - n\mu^2 \right) \quad (\text{se } n \gg 1).$$

D'altra parte, posto $Y_i = X_i^2$, l'equazione precedente diventa

$$S_n^2 \simeq \frac{1}{n-1} (n\bar{Y}_n - n\mu^2) \quad (\text{se } n \gg 1),$$

dove $\bar{Y}_n = \frac{1}{n} \sum_{i=1}^n X_i^2$ è la media campionaria di Y_1, \dots, Y_n . Applicando ancora una volta la LGN, ricaviamo che

$$\bar{Y}_n \simeq \mathbb{E}[Y_1] = \mathbb{E}[X_1^2] = \text{Var}(X_1) + \mathbb{E}[X_1]^2 = \sigma^2 + \mu^2 \quad (\text{se } n \gg 1).$$

In conclusione,

$$S_n^2 \simeq \frac{1}{n-1} [n(\sigma^2 + \mu^2) - n\mu^2] = \frac{n}{n-1} \sigma^2 \simeq \sigma^2 \quad (\text{se } n \gg 1).$$

Questo fatto può essere utilizzato per costruire intervalli di confidenza per la media di un campione arbitrario, purché sufficientemente numeroso, quando anche la sua varianza è incognita.

IC per la media di un campione numeroso con varianza incognita. Sia X_1, \dots, X_n un campione con

$$X_i \sim \text{qualunque} \quad \mu := \mathbb{E}[X_i] \text{ e } \sigma^2 := \text{Var}(X_i) \text{ entrambe incognite.}$$

Supponiamo $n \gg 1$. Allora i seguenti sono tutti intervalli di confidenza *approssimati* di livello $1 - \alpha$ per μ :

- (a) $\left(\bar{x}_n - z_{\alpha/2} \frac{s_n}{\sqrt{n}}, \bar{x}_n + z_{\alpha/2} \frac{s_n}{\sqrt{n}} \right);$
- (b) $\left(\bar{x}_n - z_{\alpha} \frac{s_n}{\sqrt{n}}, +\infty \right);$
- (c) $\left(-\infty, \bar{x}_n + z_{\alpha} \frac{s_n}{\sqrt{n}} \right).$

Infatti, per quello che abbiamo appena visto $S_n \simeq \sigma$ al limite per $n \gg 1$ e quindi

$$\frac{\bar{X}_n - \mu}{S_n} \sqrt{n} \simeq \frac{\bar{X}_n - \mu}{\sigma} \sqrt{n}.$$

Ma per il TLC

$$\frac{\bar{X}_n - \mu}{\sigma} \sqrt{n} \approx N(0, 1) \quad \Rightarrow \quad \frac{\bar{X}_n - \mu}{S_n} \sqrt{n} \approx N(0, 1).$$

Pertanto, per mostrare p.es. (b),

$$\mathbb{P}\left(\mu \geq \bar{X}_n - z_\alpha \frac{S_n}{\sqrt{n}}\right) = \mathbb{P}\left(\underbrace{\frac{\bar{X}_n - \mu}{S_n} \sqrt{n}}_{\approx N(0,1)} \leq z_\alpha\right) \simeq \Phi(z_\alpha) = 1 - \alpha.$$

Le dimostrazioni di (a) e (c) sono del tutto simili.

Osservazione 12. Abbiamo visto che, se X_1, \dots, X_n è un campione con densità qualsiasi, per $n \gg 1$ vale l'approssimazione

$$\frac{\bar{X}_n - \mu}{S_n} \sqrt{n} \underset{LGN}{\simeq} \frac{\bar{X}_n - \mu}{\sigma} \sqrt{n} \underset{TLC}{\approx} N(0, 1).$$

D'altra parte, se in più sappiamo che il campione è *gaussiano*, con $X_i \sim N(\mu, \sigma^2)$, allora $(\bar{X}_n - \mu)\sqrt{n}/S_n \sim t(n-1)$ *sempre* (anche per n piccolo!). Confrontando le due cose, questo significa che quando $n \gg 1$ la densità $t(n)$ di Student è approssimativamente una normale standard. Segue che in tal caso anche i quantili della t di Student sono approssimativamente quelli della normale standard:

$$t_{\alpha, n} \simeq z_\alpha \quad \text{per } n \gg 1.$$

Si può verificare guardando direttamente sulle tavole che questa approssimazione vale già abbastanza bene quando $n \geq 40$.

IC per la frequenza di una popolazione bernoulliana numerosa. Sia X_1, \dots, X_n un campione con

$$X_i \sim B(1, p) \quad p \text{ incognita.}$$

Supponiamo $n \gg 1$. Allora i tre intervalli (a), (b), (c) del caso precedente sono tutti intervalli di confidenza approssimati di livello $1 - \alpha$ per la *frequenza* $p = \mathbb{E}[X_i]$ del campione. Notiamo che in questo caso la media campionaria \bar{X}_n assume un significato particolare. Infatti, le variabili aleatorie X_i possono prendere solo i valori 0 o 1, dunque

$$\bar{X}_n = \frac{X_1 + \dots + X_n}{n} = \frac{|\{i \in \{1, \dots, n\} : X_i = 1\}|}{n}.$$

Se interpretiamo la somma $X_1 + \dots + X_n$ come il numero di successi in una serie di n prove di Bernoulli, allora la media campionaria \bar{X}_n non è altro che la *frequenza empirica* dei successi. Mostriamo invece che per la deviazione standard campionaria S_n vale la relazione

$$S_n^2 \simeq \bar{X}_n(1 - \bar{X}_n).$$

Infatti, ricordando la formula alternativa

$$S_n^2 = \frac{1}{n-1} \left(\sum_{i=1}^n X_i^2 - n\bar{X}_n^2 \right)$$

e osservando che nel caso presente $X_i^2 = X_i$ (perchè X può prendere solo i due valori 0 o 1, e $0^2 = 0$ e $1^2 = 1$), abbiamo

$$\begin{aligned} S_n^2 &= \frac{1}{n-1} \left(\sum_{i=1}^n X_i - n\bar{X}_n^2 \right) = \frac{1}{n-1} \left(n \frac{1}{n} \sum_{i=1}^n X_i - n\bar{X}_n^2 \right) = \frac{n}{n-1} (\bar{X}_n - \bar{X}_n^2) \\ &= \frac{n}{n-1} \bar{X}_n (1 - \bar{X}_n) \simeq \bar{X}_n (1 - \bar{X}_n), \end{aligned}$$

in quanto $n/(n-1) \simeq 1$ per $n \gg 1$. Perciò, i tre intervalli di confidenza (a), (b), (c) per la media di un campione numeroso con varianza incognita si riscrivono

$$(a) \left(\bar{x}_n - z_{\alpha/2} \sqrt{\frac{\bar{x}_n(1-\bar{x}_n)}{n}}, \bar{x}_n + z_{\alpha/2} \sqrt{\frac{\bar{x}_n(1-\bar{x}_n)}{n}} \right);$$

$$(b) \left(\bar{x}_n - z_{\alpha} \sqrt{\frac{\bar{x}_n(1-\bar{x}_n)}{n}}, +\infty \right);$$

$$(c) \left(0, \bar{x}_n + z_{\alpha} \sqrt{\frac{\bar{x}_n(1-\bar{x}_n)}{n}} \right),$$

e sono tutti intervalli di confidenza approssimati di livello $1 - \alpha$ per p (notare che in (c) il primo estremo è 0 anzichè $-\infty$, in quanto $p > 0$ per definizione).

Teorema 7. Se X_1, \dots, X_n e Y_1, \dots, Y_m sono due campioni normali e indipendenti, con

$$X_i \sim N(\mu_X, \sigma_X^2) \quad Y_i \sim N(\mu_Y, \sigma_Y^2)$$

allora:

(i) si ha sempre

$$\frac{\bar{X}_n - \bar{Y}_m - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim N(0, 1);$$

(ii) (non dimostrato) se vale in più la condizione $\sigma_X \equiv \sigma_Y$, definita la varianza pooled

$$S_p^2 := \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2},$$

in cui S_X^2 e S_Y^2 sono la varianza campionaria del primo e del secondo campione, rispettivamente, si ha

$$\frac{\bar{X}_n - \bar{Y}_m - (\mu_X - \mu_Y)}{S_p \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t(n+m-2).$$

Dimostrazione. (i) La variabile aleatoria $Z = \bar{X}_n - \bar{Y}_m$ ha densità gaussiana, in quanto combinazione lineare di variabili aleatorie normali indipendenti. La sua media e la sua varianza sono

$$\begin{aligned}\mathbb{E}[Z] &= \mathbb{E}[\bar{X}_n] - \mathbb{E}[\bar{Y}_m] = \mu_X - \mu_Y \\ \text{Var}(Z) &= \text{Var}(\bar{X}_n) + \text{Var}(\bar{Y}_m) \quad \text{per l'indipendenza di } \bar{X}_n \text{ e } \bar{Y}_m \\ &= \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}.\end{aligned}$$

Pertanto la variabile aleatoria $Z_0 = \frac{\bar{X}_n - \bar{Y}_m - (\mu_X - \mu_Y)}{\sqrt{\sigma_X^2/n + \sigma_Y^2/m}}$ ha densità $N(0, 1)$, perché è la standardizzazione di Z .

(ii) Non dimostrato. □

IC per la differenza delle medie di due popolazioni normali. Sotto le ipotesi del teorema precedente,

(i) se σ_X e σ_Y sono entrambe note, allora gli intervalli

$$\begin{aligned}\text{(a)} & \left(\bar{x}_n - \bar{y}_m - z_{\alpha/2} \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}, \bar{x}_n - \bar{y}_m + z_{\alpha/2} \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} \right) \\ \text{(b)} & \left(\bar{x}_n - \bar{y}_m - z_{\alpha} \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}, +\infty \right) \\ \text{(c)} & \left(-\infty, \bar{x}_n - \bar{y}_m + z_{\alpha} \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} \right)\end{aligned}$$

sono tutti intervalli di confidenza di livello $1 - \alpha$ per la differenza delle medie $\mu_X - \mu_Y$;

(ii) se né σ_X né σ_Y sono note, ma è noto che $\sigma_X = \sigma_Y$, allora gli intervalli

$$\begin{aligned}\text{(a)} & \left(\bar{x}_n - \bar{y}_m - t_{\alpha/2, n+m-2} s_p \sqrt{\frac{1}{n} + \frac{1}{m}}, \bar{x}_n - \bar{y}_m + t_{\alpha/2, n+m-2} s_p \sqrt{\frac{1}{n} + \frac{1}{m}} \right) \\ \text{(b)} & \left(\bar{x}_n - \bar{y}_m - t_{\alpha, n+m-2} s_p \sqrt{\frac{1}{n} + \frac{1}{m}}, +\infty \right) \\ \text{(c)} & \left(-\infty, \bar{x}_n - \bar{y}_m + t_{\alpha, n+m-2} s_p \sqrt{\frac{1}{n} + \frac{1}{m}} \right)\end{aligned}$$

sono tutti intervalli di confidenza di livello $1 - \alpha$ per la differenza delle medie $\mu_X - \mu_Y$.

La dimostrazione è un'applicazione immediata del Teorema 7 e dei soliti passaggi.

IC per la differenza delle medie di due popolazioni numerose a varianze incognite. Supponiamo che X_1, \dots, X_n e Y_1, \dots, Y_m siano due campioni *arbitrari* (non necessariamente gaussiani) e *indipendenti*, in cui i parametri

$$\mu_X := \mathbb{E}[X_i] \quad \sigma_X^2 := \text{Var}(X_i) \quad \mu_Y := \mathbb{E}[Y_i] \quad \sigma_Y^2 := \text{Var}(Y_i)$$

sono tutti incogniti. Supponiamo inoltre che sia $n \gg 1$ sia $m \gg 1$, cioè che *entrambi* i campioni siano numerosi. Allora i seguenti intervalli sono tutti $IC_{\mu_X - \mu_Y}(1 - \alpha)$ approssimati

- (a) $\left(\bar{x}_n - \bar{y}_m - z_{\alpha/2} \sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}, \bar{x}_n - \bar{y}_m + z_{\alpha/2} \sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}} \right)$
- (b) $\left(\bar{x}_n - \bar{y}_m - z_{\alpha} \sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}, +\infty \right)$
- (c) $\left(-\infty, \bar{x}_n - \bar{y}_m + z_{\alpha} \sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}} \right)$

Infatti, per la LGN $S_X^2 \simeq \sigma_X^2$ e $S_Y^2 \simeq \sigma_Y^2$ in quanto sia $n \gg 1$ sia $m \gg 1$. Pertanto

$$\frac{\bar{X}_n - \bar{Y}_m - (\mu_X - \mu_Y)}{\sqrt{S_X^2/n + S_Y^2/m}} \underset{LGN}{\simeq} \frac{\bar{X}_n - \bar{Y}_m - (\mu_X - \mu_Y)}{\sqrt{\sigma_X^2/n + \sigma_Y^2/m}} \underset{TLC}{\approx} N(0, 1)$$

Il resto della dimostrazione è il solito.

2.2 Test d'ipotesi

Supponiamo di avere a che fare con i due seguenti problemi pratici.

Esempio 38. Problema: Un amico mi propone una partita a testa o croce, e decidiamo che vince 2 euro chi ottiene più uscite favorevoli in una serie di 6 lanci. Il mio amico vuole puntare su testa, e pretende che per giocare usiamo una sua moneta. Io sono d'accordo. Perciò lanciamo la sua moneta 6 volte, e tutte e 6 le volte esce sempre testa. A questo punto comincio a sospettare che la moneta sia truccata, e voglio trovare un modo per dimostrarlo.

Soluzione: Se il mio amico avesse usato una moneta equa, cioè se fosse vera l'affermazione

$$H_0 : \text{la probabilità che esca testa è } p = 1/2,$$

allora la probabilità che escano 6 teste in 6 lanci consecutivi sarebbe

$$\mathbb{P}_{H_0}(T_1 \cap T_2 \cap \dots \cap T_6) = \mathbb{P}_{H_0}(T_1) \mathbb{P}_{H_0}(T_2) \dots \mathbb{P}_{H_0}(T_6) = \left(\frac{1}{2}\right)^6 = 1.5625\%$$

dove come al solito T_i è l'evento “esce testa all' i -esimo lancio”, mentre con \mathbb{P}_{H_0} denotiamo la probabilità calcolata supponendo che l'affermazione H_0 sia vera. In tal caso, infatti, gli eventi T_1, \dots, T_6 sono indipendenti e $\mathbb{P}_{H_0}(T_i) = 1/2$.

Una probabilità dell'1.5625% è molto piccola, e rende pertanto poco plausibile l'affermazione H_0 . In altre parole, se la moneta fosse equa, ciò che è successo nella nostra partita sarebbe estremamente poco probabile. Concludiamo pertanto che ci sono buoni motivi per ritenere che H_0 sia falsa, cioè siamo abbastanza sicuri che la moneta sia truccata.

Esempio 39. Problema: Un produttore di tondini di ferro afferma che i suoi tondini hanno un diametro medio di 17.2mm, con una precisione di fabbrica di $\sigma_0 = 0.7\text{mm}$. Volendo mettere alla prova tale affermazione, misuriamo un campione di $n = 8$ tondini, ottenendo per i loro diametri i seguenti valori (in mm)

$$\begin{array}{cccc} x_1 = 17.3 & x_2 = 16.5 & x_3 = 18.4 & x_4 = 17.7 \\ x_5 = 18.2 & x_6 = 18.7 & x_7 = 16.8 & x_8 = 18.2. \end{array} \quad (2.1)$$

Possiamo supporre che le misure provengano da un campione X_1, \dots, X_8 , con

$$X_i \sim N(\mu, \sigma_0^2) \quad \mu = \text{diametro incognito} \quad \sigma_0 = 0.7\text{mm} = \text{precisione nota}.$$

Ci poniamo allora la seguente domanda: *Questi dati sono compatibili con il diametro medio dichiarato dal produttore oppure no?* In effetti, vediamo che ci sono misure che si discostano anche in modo significativo dal valore di 17.2mm. Tali scostamenti possono essere giustificati da semplici fluttuazioni statistiche oppure abbiamo fondati motivi per ritenere che il produttore ci stia dicendo il falso?

Soluzione: Vogliamo decidere se i dati della nostra misura sono compatibili con l'affermazione del produttore

$$H_0 : \text{il diametro medio è } \mu_0 = 17.2\text{mm},$$

oppure se al contrario dimostrano che il produttore ci sta mentendo. In altre parole, vogliamo verificare se lo scostamento del nostro campione dal valore dichiarato $\mu_0 = 17.2\text{mm}$ può essere giustificato dall'imprecisione della macchina produttrice, cioè dall'errore statistico intrinseco nel processo di fabbricazione, oppure se la differenza tra i valori osservati e quello dichiarato è talmente grande da rendere H_0 molto poco plausibile.

A questo scopo, fissiamo un *valore di soglia* c per lo scostamento $|\bar{X}_n - \mu_0|$ tra la media empirica \bar{X}_n del nostro campione e il diametro medio $\mu_0 = 17.2\text{mm}$ dichiarato dal produttore. Una volta deciso il valore di c , se coi nostri dati troviamo $|\bar{x}_n - \mu_0| > c$, allora rigettiamo l'affermazione del produttore, altrimenti la accettiamo. Come soglia, fissiamo p.es. $c \equiv 0.5\text{mm}$. Con tale scelta, se fosse vera H_0 , allora sappiamo che

$$\bar{X}_n \sim N\left(\mu_0, \frac{\sigma_0^2}{n}\right) \Rightarrow Z_0 := \frac{\bar{X}_n - \mu_0}{\sigma_0} \sqrt{n} \sim N(0, 1) \quad \text{se è vera } H_0$$

e quindi la probabilità di ottenere uno scostamento $|\bar{X}_n - \mu_0| > 0.5\text{mm}$ sarebbe

$$\begin{aligned} \mathbb{P}_{H_0}(|\bar{X}_n - \mu_0| \geq 0.5) &= \mathbb{P}_{H_0}\left(\left|\frac{\bar{X}_n - \mu_0}{\sigma_0} \sqrt{n}\right| \geq \frac{0.5}{0.7} \sqrt{8}\right) \\ &= \mathbb{P}_{H_0}(|Z_0| \geq 2.02) = \mathbb{P}_{H_0}(Z_0 \geq 2.02) + \mathbb{P}_{H_0}(Z_0 \leq -2.02) \\ &= 1 - \Phi(2.02) + \Phi(-2.02) = 1 - \Phi(2.02) + 1 - \Phi(2.02) \\ &= 2(1 - \Phi(2.02)) \simeq 2(1 - 0.9783) \\ &= 4.34\%. \end{aligned}$$

Nell'espressione precedente, \mathbb{P}_{H_0} è la probabilità calcolata supponendo che $X_i \sim N(\mu_0, \sigma_0^2)$, come affermato in H_0 . Si vede quindi che, se fosse vera H_0 , la probabilità di ottenere uno scostamento pari o superiore a 0.5mm sarebbe $\simeq 4.34\%$, che è un numero piuttosto piccolo. Perciò, se coi nostri dati trovassimo *effettivamente* uno scostamento $|\bar{x}_n - \mu_0| > 0.5\text{mm}$, l'affermazione H_0 diventerebbe molto poco plausibile, e saremmo di conseguenza molto propensi a ritenere che il produttore ci stia mentendo.

Se calcoliamo la media empirica delle nostre 8 misure, troviamo

$$\bar{x}_n = 17.725\text{mm}$$

che si discosta dal diametro $\mu_0 = 17.2\text{mm}$ dichiarato dal produttore di

$$|\bar{x}_n - \mu_0| = |17.725 - 17.2|\text{mm} = 0.525\text{mm} > 0.5\text{mm}.$$

cioè più della soglia fissata. C'è pertanto una buona evidenza che il produttore ci stia mentendo.

Nella definizione seguente, fissiamo la caratteristica principale di ciascuna delle due affermazioni H_0 messe alla prova nei due esempi precedenti.

Definizione 14. Sia X_1, \dots, X_n un campione. Un'*ipotesi statistica* è un'affermazione su uno o più parametri incogniti della densità delle X_i .

Esempio 40. Nei due problemi precedenti:

- (a) H_0 è l'*ipotesi nulla*, cioè l'ipotesi di partenza che siamo disposti a rigettare solo se i dati empirici la rendono estremamente poco plausibile (vogliamo essere *molto sicuri* che il nostro amico sia un baro, oppure che il produttore di tondini menta, prima di dar loro del disonesto e del bugiardo!). Nel primo esempio, $H_0 : p = 1/2$ è un'ipotesi sul parametro p di una popolazione bernoulliana $X_i \sim B(1, p)$. Nel secondo, $H_0 : \mu = 17.2\text{mm}$ è un'ipotesi sul parametro μ della densità di $X_i \sim N(\mu, \sigma_0^2)$.
- (b) $H_1 : p \neq 1/2$ (primo esempio) o $H_1 : \mu \neq 17.2\text{mm}$ (secondo esempio) è l'*ipotesi alternativa*, incompatibile con H_0 : siamo disposti a rigettare H_0 e accettare H_1 solo quando H_0 spiega estremamente poco i dati empirici.

Definizione 15. Un *test d'ipotesi* è una regola di decisione tra H_0 e H_1 , e consiste nella seguente procedura:

1. fisso una *statistica test* $T = t(X_1, \dots, X_n)$;
2. stabilisco una *regione di rifiuto* (o *regione critica*) $C \subset \mathbb{R}$;
3. *rifiuto* H_0 se con le mie misure x_1, \dots, x_n trovo $t(x_1, \dots, x_n) \in C$; in caso contrario, la accetto.

Definizione 16. In un test d'ipotesi, commetto un errore:

- (a) *di I specie*, se rifiuto H_0 quando H_0 in realtà è vera;
- (b) *di II specie*, se accetto H_0 quando H_0 in realtà è falsa.

La definizione dei due tipi di errore è chiarita dalla seguente tabella.

	Accetto H_0	Rifiuto H_0
H_0 è vera	OK	Errore I specie
H_0 è falsa	Errore II specie	OK

Definizione 17. Il *livello di significatività* α di un test è la probabilità di commettere l'errore di I specie, cioè

$$\alpha := \mathbb{P}_{H_0}(\text{"rifiuto } H_0\text{"}) = \mathbb{P}_{H_0}(T \in C).$$

Esempio 41. Nell'Esempio 39, il livello di significatività del test dato dalla regola

$$\text{rifiuto } H_0 \text{ se } |\bar{X}_n - \mu_0| > 0.5\text{mm}$$

è $\alpha = 4.34\%$.

In un test, si vuole che la probabilità di errore di I specie sia piccola. P.es., se faccio un test per decidere se il produttore di tondini dell'Esempio 39 è sincero o meno, voglio che sia piccola la probabilità di dargli del bugiardo (cioè rifiutare H_0) quando in realtà egli è in perfetta buona fede (cioè H_0 è vera). Perciò, i tipici livelli di significatività di un test sono molto piccoli: $\alpha = 5\%$, $\alpha = 2.5\%$, o anche meno.

Al contrario, la probabilità di errore di II specie può essere anche molto alta (vedi Osservazione 14 seguente): nel nostro esempio, ciò significa che per noi accettare l'affermazione del produttore di tondini quando in realtà egli sta mentendo è un rischio molto meno grave di dargli ingiustamente del bugiardo.

Esempio 42. Nell'Esempio 39 del produttore di tondini, scelgo

$$T \equiv Z_0 = \frac{\bar{X}_n - \mu_0}{\sigma_0} \sqrt{n}$$

$$C \equiv (-\infty, -z_{\alpha/2}) \cup (z_{\alpha/2}, +\infty).$$

Allora so che

$$Z_0 \sim N(0, 1) \quad \text{se è vera } H_0$$

mentre vale l'uguaglianza di eventi

$$\text{"rifiuto } H_0"} = \{Z_0 \in C\} = \left\{ \left| \frac{\bar{X}_n - \mu_0}{\sigma_0} \sqrt{n} \right| \geq z_{\alpha/2} \right\}.$$

Il livello di significatività è pertanto

$$\begin{aligned} \mathbb{P}_{H_0}(\text{"rifiuto } H_0") &= \mathbb{P}_{H_0}(|Z_0| \geq z_{\alpha/2}) = \mathbb{P}_{H_0}(Z_0 \leq -z_{\alpha/2}) + \mathbb{P}_{H_0}(Z_0 \geq z_{\alpha/2}) \\ &= \Phi(-z_{\alpha/2}) + [1 - \Phi(z_{\alpha/2})] = \frac{\alpha}{2} + \left[1 - \left(1 - \frac{\alpha}{2}\right)\right] \\ &= \alpha. \end{aligned}$$

Con i nostri dati (2.1)

$$|z_0| = \left| \frac{\bar{x}_n - \mu_0}{\sigma_0} \sqrt{n} \right| = \left| \frac{17.725 - 17.2}{0.7} \sqrt{8} \right| \simeq 2.12.$$

Pertanto:

(i) al livello di significatività $\alpha = 5\%$, ho $z_{\alpha/2} = z_{0.025} = 1.96$, e quindi

$$2.12 > 1.96 \quad \Rightarrow \quad \text{rifiuto } H_0;$$

(ii) al livello di significatività $\alpha = 2\%$, ho $z_{\alpha/2} = z_{0.01} = 2.325$, e quindi

$$2.12 < 2.325 \quad \Rightarrow \quad \text{accetto } H_0.$$

Il test dell'esempio precedente è un caso particolare della famiglia degli Z -test per una popolazione gaussiana, che ora andremo a descrivere in generale.

Z -test per la media di un campione normale con varianza nota. Supponiamo che X_1, \dots, X_n sia un campione normale, con

$$X_i \sim N(\mu, \sigma_0^2) \quad \mu \text{ incognita} \quad \sigma_0 \text{ nota.}$$

Allora, posto

$$z_0 = \frac{\bar{x}_n - \mu_0}{\sigma_0} \sqrt{n}$$

si ha che

	rispetto alle ipotesi	la regola
(a)	$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$	rifiuto H_0 se $ z_0 > z_{\alpha/2}$
(b)	$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$	rifiuto H_0 se $z_0 > z_\alpha$
(c)	$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu < \mu_0 \end{cases}$	rifiuto H_0 se $z_0 < -z_\alpha$

sono tutti test di livello di significatività α .

Infatti, dimostriamolo in ciascuno dei tre casi:

(a) Già visto nell'esempio precedente.

(b) $\mathbb{P}_{H_0}(\text{"rifiuto } H_0") = \mathbb{P}_{H_0}(Z_0 > z_\alpha) \equiv \alpha$ perchè la statistica $Z_0 = (\bar{X}_n - \mu_0)\sqrt{n}/\sigma_0$ ha densità $N(0, 1)$ se H_0 è vera.

(c) Simile a (b) (provare per esercizio!).

Osservazione 13. Si osservi che nello Z -test l'ipotesi alternativa H_1 determina la *forma* della regione di rifiuto, cioè stabilisce se la regione di rifiuto è l'unione di due intervalli (caso (a)) oppure un intervallo unilatero destro (caso (b)) oppure unilatero sinistro (caso (c)). P.es., nel test (b) ciò si spiega perchè solo quando trovo

$$\bar{x}_n \gg \mu_0$$

posso dire che l'evidenza sperimentale a favore di H_1 è talmente grande da permettermi di rifiutare H_0 con un elevato margine di sicurezza. La condizione $\bar{x}_n \gg \mu_0$ è equivalente a

$$z_0 = \frac{\bar{x}_n - \mu_0}{\sigma_0} \sqrt{n} > c,$$

dove $c > 0$ è una soglia che deve essere abbastanza alta. Il livello di significatività α del test quantifica tale soglia di rifiuto esattamente in $c = z_\alpha$.

Osservazione 14. In uno Z -test, la probabilità di errore di II specie, cioè la probabilità di accettare H_0 quando in realtà H_0 è falsa, dipende naturalmente da qual è il vero valore del parametro incognito μ . Infatti, se H_0 è falsa, non è più vero che $\mu = \mu_0$, ma μ potrà invece assumere un

qualunque valore $\neq \mu_0$ (caso (a)) o $> \mu_0$ (caso (b)) oppure ancora $< \mu_0$ (caso (c)). In tal caso, *non è più vero* che la statistica test $Z_0 = (\bar{X}_n - \mu_0)\sqrt{n}/\sigma_0$ ha densità $N(0, 1)$, perchè quest'espressione non è più la standardizzazione di \bar{X}_n . La standardizzazione corretta è invece

$$\frac{\bar{X}_n - \mu}{\sigma_0}\sqrt{n} \sim N(0, 1).$$

Per fissare le idee, supponiamo di fare uno Z -test bilatero (tipo (a)), per il quale si ha l'uguaglianza di eventi

$$\text{"accetto } H_0"} = \{|Z_0| < z_{\alpha/2}\} = \left\{ \left| \frac{\bar{X}_n - \mu_0}{\sigma_0}\sqrt{n} \right| < z_{\alpha/2} \right\}.$$

Se $X_i \sim N(\mu, \sigma_0^2)$ e $\mu \neq \mu_0$, la probabilità di errore di II specie sarà dunque

$$\begin{aligned} \mathbb{P}_\mu(\text{"accetto } H_0") &= \mathbb{P}_\mu(|Z_0| < z_{\alpha/2}) \\ &= \mathbb{P}_\mu\left(-z_{\alpha/2} < \frac{\bar{X}_n - \mu_0}{\sigma_0}\sqrt{n} < z_{\alpha/2}\right) \\ &= \mathbb{P}_\mu\left(-z_{\alpha/2} < \frac{\bar{X}_n - \mu}{\sigma_0}\sqrt{n} + \frac{\mu - \mu_0}{\sigma_0}\sqrt{n} < z_{\alpha/2}\right) \\ &= \mathbb{P}_\mu\left(-z_{\alpha/2} - \frac{\mu - \mu_0}{\sigma_0}\sqrt{n} < \underbrace{\frac{\bar{X}_n - \mu}{\sigma_0}\sqrt{n}}_{\sim N(0,1)} < z_{\alpha/2} - \frac{\mu - \mu_0}{\sigma_0}\sqrt{n}\right) \\ &= \Phi\left(z_{\alpha/2} - \frac{\mu - \mu_0}{\sigma_0}\sqrt{n}\right) - \Phi\left(-z_{\alpha/2} - \frac{\mu - \mu_0}{\sigma_0}\sqrt{n}\right). \end{aligned}$$

In questo calcolo per evidenziare che $\mu \neq \mu_0$ abbiamo indicato con \mathbb{P}_μ la probabilità calcolata quando $X_i \sim N(\mu, \sigma_0^2)$.

P.es., supponiamo che nell'Esempio 39 il produttore di tondini ci stia dichiarando il falso, e che il diametro medio reale sia invece $\mu = 17.5\text{mm}$ anzichè i $\mu_0 = 17.2\text{mm}$ da lui dichiarati. Allora, se facciamo uno Z -test bilatero con livello di significatività $\alpha = 5\%$, abbiamo una probabilità pari a

$$\begin{aligned} &\Phi\left(1.96 - \frac{17.5 - 17.2}{0.7}\sqrt{8}\right) - \Phi\left(-1.96 - \frac{17.5 - 17.2}{0.7}\sqrt{8}\right) = \\ &= \Phi(0.75) - \Phi(-3.17) = 0.7734 - 0.0008 = 77.26\%. \end{aligned}$$

di non accorgerci che il produttore ci sta mentendo. Tale probabilità è molto alta!

Naturalmente, quando $\mu = \mu_0$ l'espressione precedente per la probabilità di errore di II specie assume il suo massimo valore $1 - \alpha$, in quanto

$$\mathbb{P}_{\mu_0}(\text{"accetto } H_0") = \mathbb{P}_{H_0}(\text{"accetto } H_0") = 1 - \mathbb{P}_{H_0}(\text{"rifiuto } H_0") = 1 - \alpha.$$

Per gli altri valori di μ , invece, si può mostrare che otteniamo una funzione *decescente* della distanza $|\mu - \mu_0|$: tanto più il vero valore di μ sarà lontano da μ_0 , tanto meno probabile sarà non accorgersi della menzogna del produttore.

Z-test con ipotesi nulla composta. Sia X_1, \dots, X_n il solito campione normale, con

$$X_i \sim N(\mu, \sigma_0^2) \quad \mu \text{ incognita} \quad \sigma_0 \text{ nota.}$$

Supponiamo di voler testare

$$H_0 : \mu \leq \mu_0 \quad \text{contro} \quad H_1 : \mu > \mu_0.$$

In questo caso, H_0 non determina più completamente la densità della statistica test $Z_0 = (\bar{X}_n - \mu_0)\sqrt{n}/\sigma_0$. Per distinguere questo tipo di test da quelli visti in precedenza, in cui H_0 invece fissava univocamente la densità della statistica test, si dice che ora l'ipotesi nulla è *composta*, mentre prima era *semplice*.

Tuttavia, si può ancora definire il livello di significatività α come la *probabilità massima* di errore di I specie quando il parametro incognito μ soddisfa H_0 :

$$\alpha := \max_{\mu \leq \mu_0} \mathbb{P}_\mu (\text{“rifiuto } H_0\text{”}).$$

Qui abbiamo di nuovo indicato con \mathbb{P}_μ la probabilità calcolata supponendo $X_i \sim N(\mu, \sigma_0^2)$, e pertanto $\mathbb{P}_\mu (\text{“rifiuto } H_0\text{”})$ è una funzione di μ ; il livello di significatività α è definito come il massimo di tale funzione sull'intervallo $(-\infty, \mu_0)$. Nel nostro caso, fissando la stessa regione di rifiuto dello Z-test (b) (si ricordi che la forma della regione di rifiuto è determinata dall'ipotesi alternativa H_1) abbiamo

$$\text{“rifiuto } H_0\text{”} = \{Z_0 \geq z_\alpha\} = \left\{ \frac{\bar{X}_n - \mu_0}{\sigma_0} \sqrt{n} \geq z_\alpha \right\}$$

e quindi

$$\begin{aligned} \mathbb{P}_\mu (\text{“rifiuto } H_0\text{”}) &= \mathbb{P}_\mu \left(\frac{\bar{X}_n - \mu_0}{\sigma_0} \sqrt{n} > z_\alpha \right) \\ &= \mathbb{P}_\mu \left(\frac{\bar{X}_n - \mu}{\sigma_0} \sqrt{n} + \frac{\mu - \mu_0}{\sigma_0} \sqrt{n} > z_\alpha \right) \\ &= \mathbb{P}_\mu \left(\underbrace{\frac{\bar{X}_n - \mu}{\sigma_0} \sqrt{n}}_{\sim N(0,1)} > z_\alpha - \frac{\mu - \mu_0}{\sigma_0} \sqrt{n} \right) \\ &= 1 - \Phi \left(z_\alpha - \frac{\mu - \mu_0}{\sigma_0} \sqrt{n} \right) \\ &= \Phi \left(\frac{\mu - \mu_0}{\sigma_0} \sqrt{n} - z_\alpha \right). \end{aligned}$$

In questo calcolo abbiamo usato il fatto che, se $X_i \sim N(\mu, \sigma_0^2)$, allora $(\bar{X}_n - \mu)\sqrt{n}/\sigma_0 \sim N(0, 1)$, ma *non* $(\bar{X}_n - \mu_0)\sqrt{n}/\sigma_0 \sqrt{n}$!!! La funzione di ripartizione Φ è crescente, dunque il $\max_{\mu \leq \mu_0}$ è raggiunto per $\mu \equiv \mu_0$, e vale

$$\Phi \left(\frac{\mu_0 - \mu_0}{\sigma_0} \sqrt{n} - z_\alpha \right) = \Phi(-z_\alpha) = \alpha.$$

In altre parole, il livello di significatività del test

$$\text{rifiuto } H_0 \text{ se } \frac{\bar{x}_n - \mu_0}{\sigma_0} \sqrt{n} > z_\alpha$$

è ancora α anche per questo tipo di test con ipotesi composta H_0 .

Similmente, si dimostra che se le ipotesi sono

$$H_0 : \mu \geq \mu_0 \quad \text{contro} \quad H_1 : \mu < \mu_0$$

allora la regola

$$\text{rifiuto } H_0 \text{ se } \frac{\bar{x}_n - \mu_0}{\sigma_0} \sqrt{n} < -z_\alpha$$

è ancora un test di livello di significatività α .

Riassumendo, i seguenti sono tutti test di livello di significatività α quando l'ipotesi nulla H_0 è composta:

	ipotesi	corrispondente test di livello α
(d)	$\begin{cases} H_0 : \mu \leq \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$	$\begin{array}{l} \text{rifiuto } H_0 \\ \text{se } z_0 > z_\alpha \end{array}$
(e)	$\begin{cases} H_0 : \mu \geq \mu_0 \\ H_1 : \mu < \mu_0 \end{cases}$	$\begin{array}{l} \text{rifiuto } H_0 \\ \text{se } z_0 < -z_\alpha \end{array}$

Osservazione 15. Se un test si conclude accettando l'ipotesi nulla H_0 , tale conclusione è in generale *debole*: significa solo che, in base al campione che abbiamo misurato, non abbiamo un margine di sicurezza sufficiente per rigettarla. Se al contrario il test si conclude rifiutando H_0 , ciò vuol dire che abbiamo una *forte* evidenza sperimentale a favore dell'ipotesi alternativa H_1 .

Ciò è esemplificato dal fatto che gli *stessi dati* (cioè lo stesso valore numerico della statistica test z_0) possono indurre ad accettare H_0 sia nello Z-test (d) sia in (e). Questo infatti succede ogniqualevolta troviamo per la statistica test un valore compreso nell'intervallo $-z_\alpha < z_0 < z_\alpha$.

Esempio 43. Facciamo un altro esempio di come scegliere le ipotesi H_0 e H_1 in un caso concreto. Supponiamo di dover costruire un ponte, e di aver bisogno per questo di piloni con un carico di rottura di almeno 10t. Supponiamo di avere a disposizione dei piloni con un carico di rottura μ incognito, e di voler determinare se essi sono sicuri per il nostro ponte oppure no. Allora, il test corretto a cui sottoporre un campione di n dei nostri piloni avrà le ipotesi

$$H_0 : \mu \leq 10t \quad \text{contro} \quad H_1 : \mu > 10t.$$

Infatti, notiamo anzitutto che utilizzeremo i piloni se e solo se il test precedente si concluderà rifiutando H_0 . In altre parole, abbiamo l'uguaglianza di eventi

$$\text{“rifiuto } H_0\text{”} = \text{“uso i piloni per costruire il ponte”}.$$

La probabilità di errore di I specie diventa dunque la probabilità di usare i piloni per costruire il ponte quando H_0 è vera, cioè quando il carico di rottura μ è più piccolo di 10t. Siccome non

vogliamo causare disastri, è *questa* (e cioè la significatività α) la probabilità che desideriamo rendere piccola. Purchè α sia piccola, siamo disposti a tollerare anche un'elevata probabilità di gettar via i piloni quando in realtà essi sono adatti, cioè un'alta probabilità di errore di II specie. Questo è infatti il male di gran lunga minore.

In altre parole, è quando accettiamo l'ipotesi alternativa H_1 che siamo molto sicuri di non sbagliare (conclusione forte), non quando accettiamo H_0 (conclusione debole).

Fin qui, abbiamo sempre supposto che il livello di significatività di un test fosse fissato a priori una volta per tutte (p.es., tipicamente $\alpha = 5\%$), e in base al suo valore abbiamo stabilito la soglia della regione di rifiuto. È spesso però altrettanto utile considerare il problema inverso: in base ai dati che ho misurato, a quali livelli di significatività posso accettare H_0 e a quali devo invece rifiutarla? Ciò giustifica la definizione seguente.

Definizione 18. In un test d'ipotesi, il *p-value* è il più piccolo livello di significatività α che impone di rifiutare H_0 in base ai dati x_1, \dots, x_n che abbiamo misurato:

$$\begin{cases} \alpha > p\text{-value} & \Rightarrow & \text{rifiuto } H_0 \\ \alpha < p\text{-value} & \Rightarrow & \text{accetto } H_0 \end{cases}.$$

Quanto più il *p-value* è piccolo, tanto più l'ipotesi nulla H_0 è poco verosimile. Infatti, H_0 può essere accettata solo per valori di significatività ancora più piccoli del *p-value*.

Esempio 44 (*p-value* dello *Z-test*). Di seguito è calcolato il *p-value* dello *Z-test* in ciascuno dei casi (a) - (e).

(a) Il *p-value* è il valore di α per cui $|z_0| \equiv z_{\alpha/2}$, cioè, per la definizione di quantile,

$$\frac{p\text{-value}}{2} = 1 - \Phi(|z_0|) \quad \Rightarrow \quad p\text{-value} = 2[1 - \Phi(|z_0|)].$$

(b,d) Il *p-value* è il valore di α per cui $z_0 \equiv z_\alpha$, cioè, sempre per la definizione di quantile,

$$p\text{-value} = 1 - \Phi(z_0).$$

(c,e) Il *p-value* è il valore di α per cui $z_0 \equiv -z_\alpha$, o, equivalentemente, $-z_0 \equiv z_\alpha$, cioè

$$p\text{-value} = 1 - \Phi(-z_0) = \Phi(z_0).$$

P.es., nello *Z-test* bilatero per la media del diametro dei tondini di ferro (vedi Esempi 39 e 42) ritroviamo

$$p\text{-value} = 2[1 - \Phi(2.12)] = 3.4\%.$$

Segue che al $5\% > 3.4\%$ non accettiamo H_0 , ma la accettiamo al $2\% < 3.4\%$ (vedi conclusioni (i) e (ii) dell'Esempio 42).

Come nel caso degli intervalli di confidenza, quando abbiamo a che fare con un campione gaussiano in cui sia la media sia la varianza sono incognite, la statistica che si usa non ha più densità normale standard, ma è un'opportuna t di Student. Infatti, in tal caso si applica la seguente famiglia di test.

T -test per la media di un campione normale con varianza incognita. Supponiamo che X_1, \dots, X_n sia un campione normale, con

$$X_i \sim N(\mu, \sigma^2) \quad \mu \text{ e } \sigma \text{ entrambe incognite.}$$

Sia

$$t_0 = \frac{\bar{x}_n - \mu_0}{s_n} \sqrt{n}.$$

Allora i seguenti sono tutti test di livello di significatività α :

ipotesi	corrispondente test di livello α
$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$	rifiuto H_0 se $ t_0 > t_{\alpha/2, n-1}$
$\begin{cases} H_0 : \mu \leq \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$ o $\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$	rifiuto H_0 se $t_0 > t_{\alpha, n-1}$
$\begin{cases} H_0 : \mu \geq \mu_0 \\ H_1 : \mu < \mu_0 \end{cases}$ o $\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu < \mu_0 \end{cases}$	rifiuto H_0 se $t_0 < -t_{\alpha, n-1}$

Le dimostrazioni sono analoghe al caso dello Z -test, con la sola differenza che in questo caso la statistica test $T_0 = (\bar{X}_n - \mu_0)\sqrt{n}/S_n$ ha densità $t(n-1)$ quando H_0 è vera (nei casi con ipotesi nulla semplice).

Esempio 45. Supponiamo che nell'Esempio 39 dei tondini di ferro il produttore non ci abbia rivelato la sua precisione di fabbrica σ_0 , ma ci abbia solo dichiarato il diametro medio $\mu_0 = 17.2\text{mm}$. Allora, se con i nostri dati (2.1) vogliamo testare

$$H_0 : \mu = \mu_0 = 17.2 \quad \text{contro} \quad H_1 : \mu \neq \mu_0$$

dobbiamo fare un T -test bilatero. Per calcolare $t_0 = (\bar{x}_n - \mu_0)\sqrt{n}/s_n$, già abbiamo trovato $\bar{x}_8 = 17.725$. Ci resta da calcolare

$$s_8 = \sqrt{\frac{1}{8-1} \left(\sum_{i=1}^8 x_i^2 - 8 \bar{x}_8^2 \right)} = \sqrt{\frac{17.3^2 + \dots + 18.2^2 - 8 \cdot 17.725^2}{7}} = 0.7924.$$

(dove abbiamo usato la formula alternativa della varianza campionaria). Pertanto,

$$t_0 = \frac{17.725 - 17.2}{0.7924} \sqrt{8} = 1.874.$$

Guardando le tavole dei quantili della t di Student troviamo

$$t_{0.20/2, 8-1} = 1.415 < |t_0| = 1.874 < t_{0.10/2, 8-1} = 1.895 \\ \Rightarrow 0.10 < p\text{-value} < 0.20.$$

Perciò, accettiamo l'affermazione del produttore a tutti i livelli di significatività $\leq 10\%$, e la rifiutiamo ai livelli $\geq 20\%$. Notare che, senza sapere la precisione di fabbrica, l'affermazione del produttore sul diametro medio diventa molto più plausibile.

Z-test per la media di un campione numeroso con varianza incognita. Supponiamo che X_1, \dots, X_n sia un campione arbitrario, con

$$X_i \sim \text{qualunque} \quad \mu := \mathbb{E}[X_i] \text{ e } \sigma^2 := \text{Var}(X_i) \text{ entrambe incognite.}$$

Supponiamo inoltre che $n \gg 1$. Definiamo

$$z_0 = \frac{\bar{x}_n - \mu_0}{s_n} \sqrt{n}.$$

Allora i seguenti sono tutti test di livello di significatività *approssimativamente* α :

ipotesi	corrispondente test di livello α
$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$	rifiuto H_0 se $ z_0 > z_{\alpha/2}$
$\begin{cases} H_0 : \mu \leq \mu_0 \\ H_1 : \mu > \mu_0 \end{cases} \quad \text{o} \quad \begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$	rifiuto H_0 se $z_0 > z_\alpha$
$\begin{cases} H_0 : \mu \geq \mu_0 \\ H_1 : \mu < \mu_0 \end{cases} \quad \text{o} \quad \begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu < \mu_0 \end{cases}$	rifiuto H_0 se $z_0 < -z_\alpha$

Per dimostrarlo, come già fatto con gli analoghi intervalli di confidenza, osserviamo che per la statistica $Z_0 = (\bar{X}_n - \mu_0)\sqrt{n}/S_n$ valgono le approssimazioni

$$Z_0 = \frac{\bar{X}_n - \mu_0}{S_n} \sqrt{n} \underset{LGN}{\approx} \frac{\bar{X}_n - \mu_0}{\sigma} \sqrt{n} \underset{TLC}{\approx} N(0, 1) \quad \text{se } H_0 \text{ è vera (nei casi con } H_0 \text{ semplice).}$$

Da questo punto, il resto della dimostrazione è identico al caso degli analoghi Z -test.

Z-test per la frequenza di una popolazione bernoulliana numerosa. Sia X_1, \dots, X_n un campione, con

$$X_i \sim B(1, p) \quad p \text{ incognita.}$$

Supponiamo $n \gg 1$. Allora i seguenti sono tutti test di livello di significatività *approssimativamente* α :

ipotesi	corrispondente test di livello α
$\begin{cases} H_0 : p = p_0 \\ H_1 : p \neq p_0 \end{cases}$	rifiuto H_0 se $\left \frac{\bar{x}_n - p_0}{\sqrt{p_0(1-p_0)}} \sqrt{n} \right > z_{\alpha/2}$
$\begin{cases} H_0 : p \leq p_0 \\ H_1 : p > p_0 \end{cases} \quad \text{oppure} \quad \begin{cases} H_0 : p = p_0 \\ H_1 : p > p_0 \end{cases}$	rifiuto H_0 se $\frac{\bar{x}_n - p_0}{\sqrt{p_0(1-p_0)}} \sqrt{n} > z_\alpha$
$\begin{cases} H_0 : p \geq p_0 \\ H_1 : p < p_0 \end{cases} \quad \text{oppure} \quad \begin{cases} H_0 : p = p_0 \\ H_1 : p < p_0 \end{cases}$	rifiuto H_0 se $\frac{\bar{x}_n - p_0}{\sqrt{p_0(1-p_0)}} \sqrt{n} < -z_\alpha$

Infatti, se $n \gg 1$, sappiamo che la statistica

$$\frac{\bar{X}_n - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n} \approx N(0,1) \quad \text{se è vera } H_0$$

come conseguenza del TLC. Le dimostrazioni del fatto che i test precedenti hanno significatività α sono dunque analoghe al caso dello Z -test, sostituendo a Z_0 la statistica precedente.

Z e T -test per la differenza delle medie di due campioni normali. Siano X_1, \dots, X_n e Y_1, \dots, Y_m due campioni normali *indipendenti*, con

$$X_i \sim N(\mu_X, \sigma_X^2) \quad Y_i \sim N(\mu_Y, \sigma_Y^2).$$

(a) Supponiamo che σ_X e σ_Y siano entrambe note. Allora i seguenti sono tutti test di livello di significatività α per la differenza delle medie:

ipotesi	corrispondente test di livello α
$\begin{cases} H_0 : \mu_X = \mu_Y \\ H_1 : \mu_X \neq \mu_Y \end{cases}$	rifiuto H_0 se $\left \frac{\bar{x}_n - \bar{y}_m}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \right > z_{\alpha/2}$
$\begin{cases} H_0 : \mu_X \leq \mu_Y \\ H_1 : \mu_X > \mu_Y \end{cases}$ oppure $\begin{cases} H_0 : \mu_X = \mu_Y \\ H_1 : \mu_X > \mu_Y \end{cases}$	rifiuto H_0 se $\frac{\bar{x}_n - \bar{y}_m}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} > z_{\alpha}$
$\begin{cases} H_0 : \mu_X \geq \mu_Y \\ H_1 : \mu_X < \mu_Y \end{cases}$ oppure $\begin{cases} H_0 : \mu_X = \mu_Y \\ H_1 : \mu_X < \mu_Y \end{cases}$	rifiuto H_0 se $\frac{\bar{x}_n - \bar{y}_m}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} < -z_{\alpha}$

(b) Supponiamo che σ_X e σ_Y siano entrambe incognite, ma che sia noto che $\sigma_X = \sigma_Y$. Allora i seguenti sono tutti test di livello di significatività α :

ipotesi	corrispondente test di livello α
$\begin{cases} H_0 : \mu_X = \mu_Y \\ H_1 : \mu_X \neq \mu_Y \end{cases}$	rifiuto H_0 se $\left \frac{\bar{x}_n - \bar{y}_m}{S_p \sqrt{\frac{1}{n} + \frac{1}{m}}} \right > t_{\alpha/2, n+m-2}$
$\begin{cases} H_0 : \mu_X \leq \mu_Y \\ H_1 : \mu_X > \mu_Y \end{cases}$ oppure $\begin{cases} H_0 : \mu_X = \mu_Y \\ H_1 : \mu_X > \mu_Y \end{cases}$	rifiuto H_0 se $\frac{\bar{x}_n - \bar{y}_m}{S_p \sqrt{\frac{1}{n} + \frac{1}{m}}} > t_{\alpha, n+m-2}$
$\begin{cases} H_0 : \mu_X \geq \mu_Y \\ H_1 : \mu_X < \mu_Y \end{cases}$ oppure $\begin{cases} H_0 : \mu_X = \mu_Y \\ H_1 : \mu_X < \mu_Y \end{cases}$	rifiuto H_0 se $\frac{\bar{x}_n - \bar{y}_m}{S_p \sqrt{\frac{1}{n} + \frac{1}{m}}} < -t_{\alpha, n+m-2}$

Le dimostrazioni sono analoghe a quelle degli Z -test e T -test corrispondenti, usando le statistiche del Teorema 7.

Z -test per la differenza delle medie di due campioni numerosi. Supponiamo che X_1, \dots, X_n e Y_1, \dots, Y_m siano due campioni *arbitrari* (non necessariamente gaussiani) e *indipendenti*, in cui i parametri

$$\mu_X := \mathbb{E}[X_i] \quad \sigma_X^2 := \text{Var}(X_i) \quad \mu_Y := \mathbb{E}[Y_i] \quad \sigma_Y^2 := \text{Var}(Y_i)$$

sono tutti incogniti. Supponiamo inoltre che sia $n \gg 1$ sia $m \gg 1$, cioè che *entrambi* i campioni siano numerosi. Allora i seguenti sono tutti test di livello di significatività approssimativamente α :

ipotesi	corrispondente test di livello α
$\begin{cases} H_0 : \mu_X = \mu_Y \\ H_1 : \mu_X \neq \mu_Y \end{cases}$	rifiuto H_0 se $\left \frac{\bar{x}_n - \bar{y}_m}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}} \right > z_{\alpha/2}$
$\begin{cases} H_0 : \mu_X \leq \mu_Y \\ H_1 : \mu_X > \mu_Y \end{cases} \text{ oppure } \begin{cases} H_0 : \mu_X = \mu_Y \\ H_1 : \mu_X > \mu_Y \end{cases}$	rifiuto H_0 se $\frac{\bar{x}_n - \bar{y}_m}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}} > z_\alpha$
$\begin{cases} H_0 : \mu_X \geq \mu_Y \\ H_1 : \mu_X < \mu_Y \end{cases} \text{ oppure } \begin{cases} H_0 : \mu_X = \mu_Y \\ H_1 : \mu_X < \mu_Y \end{cases}$	rifiuto H_0 se $\frac{\bar{x}_n - \bar{y}_m}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}} < -z_\alpha$

Infatti, come nei corrispondenti intervalli di confidenza, la statistica

$$\frac{\bar{X}_n - \bar{Y}_m - (\mu_X - \mu_Y)}{\sqrt{S_X^2/n + S_Y^2/m}} \approx N(0, 1)$$

come conseguenza della LGN e del TLC. La dimostrazione che i tre test hanno tutti significatività α è simile a quella per lo Z -test.

2.3 Test non parametrici

Finora abbiamo considerato solo test su ipotesi che si riferivano a uno o più *parametri* della densità del campione, tuttavia la *forma* della densità si supponeva sempre fissata a priori. Per esempio, nello Z -test per la media di una popolazione normale a varianza σ_0^2 nota, avevamo a che fare con un campione X_1, \dots, X_n in cui le X_i avevano media μ incognita, ma sapevamo comunque a priori che la loro densità era la funzione gaussiana $f_{X_i}(x) = e^{(x-\mu)^2/(2\sigma_0^2)} / \sqrt{2\pi\sigma_0^2}$. Nei due nuovi test che considereremo adesso, invece, le ipotesi riguarderanno proprio il tipo di densità della popolazione, cioè la forma della funzione f_{X_i} stessa.

Il primo test confronta come la densità empirica ricavata dal campione si adatta a una densità teorica prefissata, e per farlo usa il seguente teorema.

Teorema 8 (Non dimostrato). *Sia X_1, \dots, X_n un campione aleatorio, con X_i variabili aleatorie discrete a valori nell'insieme $\{1, \dots, k\}$ (cioè $X_i : \Omega \rightarrow \{1, \dots, k\}$ per ogni i). Sia $p : \{1, \dots, k\} \rightarrow [0, 1]$ la densità di una qualsiasi delle X_i . Per ogni $l \in \{1, \dots, k\}$, definiamo le variabili aleatorie*

$$O_l = |\{i \in \{1, \dots, n\} : X_i = l\}|.$$

Sia inoltre T la statistica

$$T = \sum_{l=1}^k \frac{O_l^2}{np(l)} - n.$$

Allora, se $n \gg 1$, la statistica T ha approssimativamente densità chi-quadrato con $k - 1$ gradi di libertà:

$$T \approx \chi^2(k - 1).$$

Per fissare le idee, possiamo pensare che ogni $l \in \{1, \dots, k\}$ sia una particolare *caratteristica* (o *classe*), e che ogni individuo X_i del campione possa avere una e una sola di tali caratteristiche (o, equivalentemente, appartenere a una e una sola classe l). La variabile aleatoria O_l è dunque la *numerosità empirica della caratteristica l* all'interno del campione, cioè il numero di individui del campione che presentano tale caratteristica. Analogamente, la densità p rappresenta la distribuzione delle caratteristiche $\{1, \dots, k\}$ all'interno dell'intera popolazione.

Osservazione 16. Una regola empirica per valutare quando n è sufficientemente grande da far valere l'approssimazione $T \approx \chi^2(k - 1)$ è che almeno l'80% delle numerosità teoriche $np(l)$ dovrebbero essere maggiori di 5, e il restante 20% dovrebbero essere tutte maggiori di 1. Se questo non è verificato, le classi vanno raggruppate in modo opportuno.

Test di buon adattamento a una densità completamente nota. Supponiamo che $p_0 : \{1, \dots, k\} \rightarrow [0, 1]$ sia una densità discreta assegnata. Sia X_1, \dots, X_n un campione, con X_i variabile aleatoria discreta con densità $p : \{1, \dots, k\} \rightarrow [0, 1]$ incognita. Supponiamo inoltre $n \gg 1$. Allora, rispetto alle ipotesi

$$H_0 : p = p_0 \quad \text{contro} \quad H_1 : p \neq p_0,$$

la regola

$$\text{rifiuto } H_0 \text{ se } t = \sum_{l=1}^k \frac{o_l^2}{np_0(l)} - n \geq \chi_{\alpha, k-1}^2$$

è un test di livello di significatività α .

La dimostrazione segue direttamente dal teorema precedente, che ci dice che

$$T \approx \chi^2(k-1) \quad \text{se è vera } H_0$$

e quindi

$$\mathbb{P}_{H_0}(\text{"rifiuto } H_0\text{"}) = \mathbb{P}_{H_0}(T \geq \chi_{\alpha, k-1}^2) = \alpha.$$

Esempio 46. *Problema:* La roulette ha 37 numeri: 18 neri, 18 rossi e 1 verde. Osservo $n = 300$ giocate, registrando le uscite dei seguenti colori:

	nero	rosso	verde
numero di uscite	135	159	6

Sulla base di questi dati, posso mettere in discussione il fatto che la roulette sia equilibrata?

Soluzione: Definiamo le classi $1 \equiv \text{nero}$, $2 \equiv \text{rosso}$, $3 \equiv \text{verde}$. Sia $X_i : \Omega \rightarrow \{1, 2, 3\}$ il colore uscito all' i -esima giocata, e sia p la sua densità incognita. Se la roulette fosse equilibrata, X_i avrebbe densità

$$p_0(1) = \frac{18}{37} \quad p_0(2) = \frac{18}{37} \quad p_0(3) = \frac{1}{37}.$$

Testiamo quindi

$$H_0 : p = p_0 \quad \text{contro} \quad H_1 : p \neq p_0.$$

La tabella delle numerosità empiriche a confronto con quelle teoriche è

	$l = 1$	$l = 2$	$l = 3$
o_l	135	159	6
$np_0(l)$	$300 \cdot 18/37$	$300 \cdot 18/37$	$300 \cdot 1/37$

Vediamo in particolare che tutte le numerosità teoriche $np_0(l)$ sono ben maggiori di 5, dunque possiamo applicare il test di adattamento appena descritto. Coi nostri dati, la statistica test vale

$$t = \frac{135^2}{300 \cdot 18/37} + \frac{159^2}{300 \cdot 18/37} + \frac{6^2}{300 \cdot 1/37} - 300 \simeq 2.537$$

Si ha

$$2.537 < \chi_{0.05, 3-1}^2 = 5.991 \quad \Rightarrow \quad \text{accetto } H_0 \text{ al } 5\% \text{ di significatività.}$$

Il secondo test non parametrico che descriveremo considera invece due diverse proprietà x e y del campione e cerca di stabilire se tali proprietà sono tra loro indipendenti. Per esempio, in un campione di n persone scelte a caso le proprietà x e y potrebbero essere

$$x = \text{sex}, \quad y = \text{colore degli occhi} \quad \text{di ciascun individuo.}$$

Questo test si basa sul seguente teorema.

Teorema 9 (Non dimostrato). Sia $(X_1, Y_1), \dots, (X_n, Y_n)$ un campione costituito da vettori aleatori discreti a due componenti, con

$$(X_i, Y_i) : \Omega \rightarrow \{1, \dots, r\} \times \{1, \dots, s\}.$$

Definiamo le statistiche

$$O_{l,m} = |\{i \in \{1, \dots, n\} : (X_i, Y_i) = (l, m)\}|$$

$$O_{l\cdot} = \sum_{m=1}^s O_{l,m} \quad O_{\cdot m} = \sum_{l=1}^r O_{l,m}.$$

Allora, se vale l'ipotesi nulla

$$H_0 : X_i \text{ e } Y_i \text{ sono indipendenti per ogni } i,$$

la statistica test

$$T = n \left(\sum_{l=1}^r \sum_{m=1}^s \frac{O_{l,m}^2}{O_{l\cdot} O_{\cdot m}} - 1 \right)$$

per $n \gg 1$ ha approssimativamente densità chi-quadrato con $(r-1)(s-1)$ gradi di libertà:

$$T \approx \chi^2((r-1)(s-1)) \quad \text{se è vera } H_0.$$

Test d'indipendenza. Se $(X_1, Y_1), \dots, (X_n, Y_n)$ è il campione del teorema precedente e $n \gg 1$, allora, rispetto alle ipotesi

$$H_0 : X_i \text{ e } Y_i \text{ sono indipendenti} \quad \text{contro} \quad H_1 : X_i \text{ e } Y_i \text{ non sono indipendenti},$$

la regola

$$\text{rifiuto } H_0 \text{ se } t = n \left(\sum_{l=1}^r \sum_{m=1}^s \frac{o_{l,m}^2}{o_{l\cdot} o_{\cdot m}} - 1 \right) \geq \chi_{\alpha, k-1}^2$$

è un test di livello di significatività α .

La dimostrazione è identica a quella del test di buon adattamento.

Osservazione 17. Anche in questo caso, la regola empirica per vutare quando n è sufficientemente grande da far valere l'approssimazione $T \approx \chi^2((r-1)(s-1))$ è che almeno l'80% delle numerosità teoriche $np_{l,m}$ dovrebbero essere maggiori di 5, e le restanti dovrebbero esser tutte maggiori di 1. Tuttavia, poichè in questo caso le numerosità teoriche non sono note, si usano quelle stimate dai dati $n\hat{p}_{l,m} = o_{l\cdot} o_{\cdot m} / n$. Pertanto, n deve essere abbastanza grande e ciascuna classe sufficientemente numerosa da aversi almeno l'80% delle numerosità $o_{l\cdot} o_{\cdot m} / n$ maggiori o uguali a 5, e tutte le restanti maggiori o uguali a 1.

Esempio 47. Problema: Il corso di statistica viene insegnato dallo stesso docente in 3 diversi corsi di laurea: fisica, matematica e ingegneria. I risultati dei 3 corsi sono riassunti nella seguente tabella:

	fisica	matematica	ingegneria
promossi	50	15	30
bocciati	37	8	40

Sulla base di questi dati, possiamo affermare che ci sia dipendenza tra il corso di laurea e il livello di preparazione degli studenti?

Soluzione: Definiamo le classi nel modo seguente:

	1	2	3
1	$o_{1,1} = 50$	$o_{1,2} = 15$	$o_{1,3} = 30$
2	$o_{2,1} = 37$	$o_{2,2} = 8$	$o_{2,3} = 40$

Notiamo che ogni casella contiene più di 5 individui, dunque sono soddisfatte le condizioni dell'osservazione precedente e non è necessario raggruppare più classi in una sola. Completiamo la tabella aggiungendo sotto e a destra le rispettive numerosità marginali, ottenute sommando sulle relative colonne e righe:

	1	2	3	
1	$o_{1,1} = 50$	$o_{1,2} = 15$	$o_{1,3} = 30$	$o_{1.} = 95$
2	$o_{2,1} = 37$	$o_{2,2} = 8$	$o_{2,3} = 40$	$o_{2.} = 85$
	$o_{.1} = 87$	$o_{.2} = 23$	$o_{.3} = 70$	

La numerosità del campione è

$$n = o_{1.} + o_{2.} = o_{.1} + o_{.2} + o_{.3} = 180.$$

La statistica test vale

$$\begin{aligned}
 t &= n \left(\sum_{l=1}^r \sum_{m=1}^s \frac{o_{l,m}^2}{o_{l.} o_{.m}} - 1 \right) \\
 &= 180 \left(\frac{50^2}{87 \cdot 95} + \frac{15^2}{23 \cdot 95} + \frac{30^2}{70 \cdot 95} + \frac{37^2}{87 \cdot 85} + \frac{8^2}{23 \cdot 85} + \frac{40^2}{70 \cdot 85} - 1 \right) \\
 &\simeq 4.961
 \end{aligned}$$

(N.B.: Nel calcolo precedente, attenzione a non approssimare ciascun termine della somma in modo troppo grossolano, altrimenti si rischia di trovare un valore molto diverso per t !!!). Questo valore va confrontato con i quantili della $\chi^2((3-1)(2-1)) = \chi^2(2)$, e si vede che

$$4.961 < 5.991 = \chi_{0.05,2}^2.$$

Perciò, il p -value dei dati è $> 5\% \Rightarrow$ accetto H_0 al 5% di significatività.

Capitolo 3

Regressione lineare

3.1 Il metodo dei minimi quadrati

Supponiamo che x e y siano due grandezze legate da una certa relazione $y = y(x)$, cioè che sussista una qualche relazione funzionale che assegna ad ogni valore di x (*input*, o valore in ingresso) un determinato valore di y (*output*, o risposta).

Esempio 48. Esempi di relazioni funzionali tra due grandezze diverse sono:

- (i) x è tensione variabile applicata ai capi di una resistenza e $y(x)$ è la corrente che passa attraverso la resistenza a quella tensione; in tal caso, la teoria prevede una relazione funzionale del tipo $y(x) = Rx$, dove R è il valore della resistenza;
- (ii) x è il pressione di una mole di gas a temperatura ambiente e $y(x)$ è il corrispondente volume; la legge dei gas perfetti ci dice allora che deve essere $y(x) = C/x$, dove C è un'opportuna costante.

In generale, tuttavia, la funzione y non è nota a priori, oppure è nota solo a meno di un certo numero di parametri incogniti (p.es., la resistenza R o la costante C nei due esempi precedenti). Il nostro scopo sarà pertanto di determinare la funzione y a partire dalla misura delle risposte y_1, y_2, \dots, y_n a una serie di valori in ingresso x_1, x_2, \dots, x_n , in modo che $y_i = y(x_i)$ per ogni x_i . Non solo: vorremmo anche che la funzione y così determinata sia in grado di predire la risposta a un *qualunque* altro ingresso x diverso da x_1, x_2, \dots, x_n .

Naturalmente, essendo la misura di ogni risposta y_i soggetta a un certo grado di imprecisione (o *rumore*), ci aspettiamo che la nostra funzione y non soddisfi esattamente la relazione $y_i = y(x_i)$ per ogni $i = 1, \dots, n$, ma piuttosto che

$$y_i = y(x_i) + e_i,$$

dove e_i è lo scostamento dovuto al rumore.

Poiché l'insieme di tutte le relazioni possibili tra x e y è enorme (è l'insieme di tutte le funzioni $y : \mathbb{R} \rightarrow \mathbb{R}$), occorre fare qualche ipotesi *a priori* sulla funzione y che renda trattabile il problema di determinare y dai dati. Non solo: un'ipotesi a priori su y è necessaria per il fatto stesso di

dar senso al nostro problema, in quanto per ogni serie finita di dati $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ esisteranno sempre infinite funzioni (anche molto diverse tra loro) che passano per ciascun punto (x_i, y_i) del piano.

L'ipotesi più semplice possibile è che y sia una *relazione lineare*

$$y(x) = a + bx.$$

In tal caso, la funzione y è univocamente determinata dai due parametri a e b . Il problema di determinare a e b dai dati prende il nome di *regressione lineare*.

Supponiamo quindi che

$$y_i = a + bx_i + e_i \quad \forall i = 1, \dots, n. \quad (3.1)$$

La retta che meglio approssima i nostri dati sarà dunque quella che minimizza (in qualche senso) gli scostamenti e_i . Definito il funzionale non negativo

$$L(a, b) := \sum_{i=1}^n (y_i - a - bx_i)^2 = \sum_{i=1}^n e_i^2,$$

la *retta dei minimi quadrati* è quella corrispondente ai valori (a_0, b_0) che minimizzano L . Per ricavare (a_0, b_0) , troviamo i punti stazionari di L :

$$\begin{aligned} \frac{\partial L}{\partial a} &= -2 \sum_{i=1}^n (y_i - a - bx_i) = -2(n\bar{y} - na - nb\bar{x}) \equiv 0 \\ \frac{\partial L}{\partial b} &= -2 \sum_{i=1}^n x_i (y_i - a - bx_i) = -2 \left(\sum_{i=1}^n x_i y_i - na\bar{x} - b \sum_{i=1}^n x_i^2 \right) \equiv 0, \end{aligned}$$

dove abbiamo posto

$$\bar{x} := \frac{1}{n} \sum_{i=1}^n x_i \quad \bar{y} := \frac{1}{n} \sum_{i=1}^n y_i.$$

Dalla prima equazione ricaviamo

$$a_0 = \bar{y} - b_0 \bar{x}$$

che inserita nella seconda dà

$$\sum_{i=1}^n x_i y_i - n\bar{x}(\bar{y} - b_0 \bar{x}) - b_0 \sum_{i=1}^n x_i^2 \Rightarrow b_0 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}.$$

Quest'ultima espressione per b_0 può risciversi

$$b_0 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2},$$

in quanto

$$\begin{aligned}\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} &= \frac{\sum x_i y_i - \bar{x} \sum y_i - \bar{y} \sum x_i + n\bar{x}\bar{y}}{\sum x_i^2 - 2\bar{x} \sum x_i + n\bar{x}^2} \\ &= \frac{\sum x_i y_i - n\bar{x}\bar{y} - n\bar{y}\bar{x} + n\bar{x}\bar{y}}{\sum x_i^2 - 2n\bar{x}^2 + n\bar{x}^2} \\ &= \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}\end{aligned}$$

D'ora in poi definiremo per semplicità

$$s_{xx} := \sum_{i=1}^n (x_i - \bar{x})^2 \quad s_{xy} := \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad s_{yy} := \sum_{i=1}^n (y_i - \bar{y})^2.$$

Con questa definizione,

$$a_0 = \bar{y} - \frac{s_{xy}}{s_{xx}} \bar{x} \quad b_0 = \frac{s_{xy}}{s_{xx}}. \quad (3.2)$$

Il funzionale L ha (a_0, b_0) come unico punto stazionario, e inoltre il $\lim_{(a,b) \rightarrow \infty} L(a, b) = +\infty$. Segue pertanto che (a_0, b_0) è effettivamente *l'unico punto di minimo* di L .

Chiameremo

$$ss_e := L(a_0, b_0) = \sum_{i=1}^n (y_i - a_0 - b_0 x_i)^2$$

la *varianza residua* del nostro modello. Notiamo che

$$ss_e = 0 \quad \Leftrightarrow \quad \text{i dati } (x_1, y_1), \dots, (x_n, y_n) \text{ giacciono tutti su una retta.}$$

Inoltre, in tal caso la retta è proprio $y = a_0 + b_0 x$. Infatti:

\Rightarrow $ss_e = 0$ implica che $(y_i - a_0 - b_0 x_i)^2 = 0$ per ogni $i = 1, \dots, n$ (in quanto tutti i quadrati nella somma sono positivi o al più nulli), o, in altre parole, $y_i = a_0 + b_0 x_i$ per ogni i ;

\Leftarrow se i dati $(x_1, y_1), \dots, (x_n, y_n)$ giacciono tutti su una stessa retta, diciamo $y_i = a + b x_i$ per ogni i , allora $L(a, b) = \sum (y_i - a - b x_i)^2 = 0$, e quindi $(a, b) = (a_0, b_0)$ per l'unicità del minimo. Quindi, $ss_e = L(a_0, b_0) = 0$.

Notiamo pertanto che il nome di *varianza residua* è pienamente giustificato dal fatto che il valore di ss_e quantifica la variabilità dei dati che non è giustificata dal modello lineare.

La *varianza spiegata* è invece la quantità

$$ss_t := \sum_{i=1}^n (\bar{y} - a_0 - b_0 x_i)^2 = \sum_{i=1}^n (b_0 \bar{x} - b_0 x_i)^2 = b_0^2 s_{xx} = \frac{s_{xy}^2}{s_{xx}}. \quad (3.3)$$

Dimostreremo ora la relazione

$$s_{yy} = ss_t + ss_e, \quad (3.4)$$

che giustifica il nome di varianza spiegata per ss_t . Infatti, più i nostri dati sono in accordo col modello lineare, cioè più ss_e è prossima a 0, tantopiù ss_t è prossima alla varianza s_{yy} dei dati in uscita y_1, \dots, y_n .

Per mostrare (3.4), calcoliamo

$$\begin{aligned}
 ss_t + ss_e &= \sum (\bar{y} - a_0 - b_0 x_i)^2 + \sum (y_i - a_0 - b_0 x_i)^2 \\
 &= \frac{s_{xy}^2}{s_{xx}} + \sum (y_i - a_0 - b_0 x_i)^2 && \text{per (3.3)} \\
 &= \frac{s_{xy}^2}{s_{xx}} + \sum \left[y_i - \bar{y} + \frac{s_{xy}}{s_{xx}}(\bar{x} - x_i) \right]^2 && \text{per (3.2)} \\
 &= \frac{s_{xy}^2}{s_{xx}} + \sum \left[(y_i - \bar{y})^2 + \frac{s_{xy}^2}{s_{xx}^2}(\bar{x} - x_i)^2 + 2 \frac{s_{xy}}{s_{xx}}(y_i - \bar{y})(\bar{x} - x_i) \right] \\
 &= \frac{s_{xy}^2}{s_{xx}} + \sum (y_i - \bar{y})^2 + \frac{s_{xy}^2}{s_{xx}^2} \sum (\bar{x} - x_i)^2 + 2 \frac{s_{xy}}{s_{xx}} \sum (y_i - \bar{y})(\bar{x} - x_i) \\
 &= \frac{s_{xy}^2}{s_{xx}} + s_{yy} + \frac{s_{xy}^2}{s_{xx}^2} s_{xx} - 2 \frac{s_{xy}}{s_{xx}} s_{xy} \\
 &= s_{yy}.
 \end{aligned}$$

Notiamo che da (3.4) segue anche la seguente formula, che fornisce esplicitamente ss_e in funzione dei dati:

$$ss_e = ss_{yy} - ss_t = ss_{yy} - \frac{s_{xy}^2}{s_{xx}}. \quad (3.5)$$

Il *coefficiente di determinazione* r^2 del modello lineare esprime il rapporto tra la varianza spiegata e la varianza totale

$$r^2 = \frac{ss_t}{s_{yy}} = 1 - \frac{ss_e}{s_{yy}}.$$

Dall'equazione (3.4) e dalla positività di ss_t e ss_e si vede che $0 \leq ss_t \leq ss_{yy}$, da cui $0 \leq r^2 \leq 1$. Inoltre, r^2 è tanto più prossimo a 1 quanto più la varianza residua ss_e è prossima a 0, cioè tanto meglio il modello lineare è in grado di spiegare i dati. Valori di $r^2 \geq 0.9$ superiori a 0.9 indicano che il modello lineare spiega bene i nostri dati.

Notiamo che r^2 si può esprimere in modo molto semplice come funzione dei dati:

$$r^2 = \frac{ss_t}{s_{yy}} = \frac{b_0^2 s_{xx}}{s_{yy}} = \frac{s_{xy}^2}{s_{xx} s_{yy}}.$$

La sua radice quadrata $r = \frac{s_{xy}}{\sqrt{s_{xx} s_{yy}}}$ si chiama *coefficiente di correlazione campionario*.

3.2 Il modello statistico

Supponiamo adesso che nell'equazione (3.1), che definisce il modello lineare, il rumore e_i sia di tipo probabilistico, cioè che e_i sia una v.a. per ogni i . Allora anche la risposta y_i sarà una v.a.. Per sottolineare questo fatto, indicheremo il rumore e la risposta con le corrispondenti lettere maiuscole:

$$Y_i = a + bx_i + E_i.$$

Supporremo inoltre che il rumore sia gaussiano a media nulla, e che i rumori di ciascuna risposta siano tutti indipendenti tra loro:

$$E_1, E_2, \dots, E_n \text{ i.i.d. con } E_i \sim N(0, \sigma^2).$$

Tuttavia, supporremo che l'ampiezza del rumore, cioè la sua deviazione standard σ , sia incognita.

Avremo pertanto

$$Y_1, Y_2, \dots, Y_n \text{ i.i.d. con } Y_i \sim N(a + bx_i, \sigma^2).$$

Inoltre, i parametri della retta dei minimi quadrati passante per i punti $(x_1, Y_1), \dots, (x_n, Y_n)$ saranno le v.a.

$$A_0 = \bar{Y} - \frac{S_{xY}}{s_{xx}} \bar{x} \quad B_0 = \frac{S_{xY}}{s_{xx}},$$

dove

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i \quad S_{xY} = \sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y}).$$

Analogamente, denoteremo con la corrispondente lettera maiuscola le v.a. associate a ciascuna delle quantità s_{yy} e ss_e :

$$S_{YY} = \sum_{i=1}^n (Y_i - \bar{Y})^2, \quad SS_E = S_{YY} - \frac{S_{xY}^2}{s_{xx}}.$$

Notiamo che A_0 e B_0 sono stimatori non distorti dei veri parametri a e b , cioè

$$\mathbb{E}[A_0] = a \quad \mathbb{E}[B_0] = b.$$

Infatti, ciò segue immediatamente dal fatto che

$$\begin{aligned} \mathbb{E}[\bar{Y}] &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}[Y_i] = \frac{1}{n} \sum_{i=1}^n (a + bx_i) = a + b\bar{x} \\ \mathbb{E}[S_{xY}] &= \sum_{i=1}^n (x_i - \bar{x})(\mathbb{E}[Y_i] - \mathbb{E}[\bar{Y}]) \\ &= \sum_{i=1}^n (x_i - \bar{x})[a + bx_i - (a + b\bar{x})] \\ &= b \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= bs_{xx}, \end{aligned}$$

e pertanto

$$\begin{aligned}\mathbb{E}[A_0] &= \mathbb{E}\left[\bar{Y} - \frac{S_{xY}}{s_{xx}}\bar{x}\right] = \mathbb{E}[\bar{Y}] - \frac{\bar{x}}{s_{xx}}\mathbb{E}[S_{xY}] = a + b\bar{x} - \frac{\bar{x}}{s_{xx}}bs_{xx} = a \\ \mathbb{E}[B_0] &= \mathbb{E}\left[\frac{S_{xY}}{s_{xx}}\right] = \frac{bs_{xx}}{s_{xx}} = b.\end{aligned}$$

Inoltre, sia A_0 sia B_0 sono esprimibili come combinazioni lineari delle v.a. normali indipendenti Y_1, \dots, Y_n , in quanto ciò vale sia per \bar{Y} sia per S_{xY} . Segue da questo che pure le v.a. A_0 e B_0 sono normali. Resta da calcolare la loro varianza (calcolo che ometteremo), e si trova

$$A_0 \sim N\left(a, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right)\right) \quad B_0 \sim N\left(b, \frac{\sigma^2}{s_{xx}}\right).$$

Notiamo però che le statistiche A_0 e B_0 non ci permettono ancora di fare inferenza sui parametri incogniti a e b , in quanto la varianza σ^2 del rumore è anch'essa incognita.

Si può tuttavia dimostrare che

$$\sqrt{\frac{n(n-2)s_{xx}}{(s_{xx} + n\bar{x}^2)SS_E}}(A_0 - a) \sim t(n-2) \quad (3.6)$$

$$\sqrt{\frac{(n-2)s_{xx}}{SS_E}}(B_0 - b) \sim t(n-2). \quad (3.7)$$

Queste due statistiche permettono facilmente di costruire i seguenti $IC(1 - \alpha)$ per i parametri incogniti a e b (verifichetelo!):

$$\begin{aligned}a &\in \left(a_0 - t_{\frac{\alpha}{2}, n-2} \sqrt{\frac{(s_{xx} + n\bar{x}^2)ss_e}{n(n-2)s_{xx}}}, a_0 + t_{\frac{\alpha}{2}, n-2} \sqrt{\frac{(s_{xx} + n\bar{x}^2)ss_e}{n(n-2)s_{xx}}}\right) \\ b &\in \left(b_0 - t_{\frac{\alpha}{2}, n-2} \sqrt{\frac{ss_e}{(n-2)s_{xx}}}, b_0 + t_{\frac{\alpha}{2}, n-2} \sqrt{\frac{ss_e}{(n-2)s_{xx}}}\right).\end{aligned}$$

Nei due intervalli precedenti, le realizzazioni a_0 , b_0 e ss_e delle statistiche A_0 , B_0 e SS_E sulle misure y_1, \dots, y_n si ottengono dalle equazioni (3.2) e (3.5), che esprimono a_0 , b_0 e ss_e in termini delle quantità \bar{x} , \bar{y} , s_{xx} , s_{yy} e s_{xy} note dalle misure.

Inoltre, da (3.7) ricaviamo che, se vogliamo testare le ipotesi

$$H_0 : b = 0 \quad \text{contro} \quad H_1 : b \neq 0,$$

una regione di rifiuto al livello di significatività α è

$$\sqrt{\frac{(n-2)s_{xx}}{ss_e}} |b_0| \geq t_{\frac{\alpha}{2}, n-2}.$$

Questo test ha il significato seguente: se accettiamo H_0 , ciò significa che non c'è evidenza che il valor medio delle risposte Y_1, \dots, Y_n dipenda dai dati in ingresso x_1, \dots, x_n .

Infine, se x^* è un nuovo valore, diverso dai dati in ingresso x_1, \dots, x_n , possiamo determinare un intervallo di confidenza per valor medio della corrispondente risposta Y^* . Ricordiamo che, nel nostro modello,

$$Y^* = a + bx^* + E^* \quad E^* \sim N(0, \sigma^2) \text{ e indipendente da } E_1, \dots, E_n,$$

e quindi

$$\mathbb{E}[Y^*] = a + bx^*.$$

Si può dimostrare che

$$\frac{A_0 + B_0 x^* - (a + bx^*)}{\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{s_{xx}}} \sqrt{\frac{SS_E}{n-2}}} \sim t(n-2),$$

che dà il seguente $IC(1 - \alpha)$ per il parametro incognito $a + bx^*$:

$$a + bx^* \in \left(a_0 + b_0 x^* - t_{\frac{\alpha}{2}, n-2} \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{s_{xx}}} \sqrt{\frac{ss_e}{n-2}}, \right. \\ \left. a_0 + b_0 x^* + t_{\frac{\alpha}{2}, n-2} \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{s_{xx}}} \sqrt{\frac{ss_e}{n-2}} \right).$$