Statistica - linea A

Esercizi del corso

Indice

1. Statistiche Riassuntive e Grafici

NB Il corso non richiede l'uso di software, ma usare un software, ad esempio R o Python (quest'ultimo preferibile), per lavorare con alcuni esempi concreti, vi invito perciò a fare alcune prove. Sono disponibili diverse guide online, qui richiamo giusto la documentazione per R, https://www.rdocumentation.org/ , e alcuni comandi utili sempre per R:

- read_csv (library readr) per importare file csv (il file così importato è di tipo data, da cui si può poi estrarre il vettore di interesse col comando nome_file$nome_colonna).
- hist per la rappresentazione di grafici a barre e istogrammi; opzioni: breaks per scelta intervalli; freq=TRUE per grafici a barre, freq=FALSE per istogrammi.
- mean, median, var, sd rispettivamente per media, mediana, varianza e deivazione standard campionarie (nel caso della varianza e della deviazione standard, il denominatore è quindi $n-1$).
- boxplot per la rappresentazione del boxplot; opzione range = a per disegnare come outlier i dati distanti più di $a \cdot R$ dal rettangolo centrale (dove R è lo scarto interquartile, valore di defaule 1.5; $R = 0$ estende i baffi ai dati più estremi); inserendo dati ~nome_sottogruppo vengono disegnati i boxplot per i dati divisi per sottogruppo.
- ecdf per calcolare la funzione di ripartizone empirica (plot(ecdf) per disegnarla).
- aggregate per il calcolo di indici statistici per sottogruppo (si veda la documentazione per maggiori dettagli).

Esercizio 1.1 **(consigliato 31/01)** I dati seguenti riportano il numero mensile di terremoti di magnitudo superiore a 4 in Italia, da marzo 2021 a febbraio 2022.

7, 4, 0, 2, 1, 3, 3, 3, 1, 1, 1, 3

Dire se i dati sono discreti o continui e rappresentare il corrispondente grafico a barre/istogramma (a seconda del tipo di dati). ■

Soluzione I dati sono discreti, le frequenze relative sono:

x	0	1	2	3	4	7
p	$\frac{1}{12}$	$\frac{4}{12}$	$\frac{1}{12}$	$\frac{4}{12}$	$\frac{1}{12}$	$\frac{1}{12}$

■

Esercizio 1.2 **(consigliato 31/01)** Un server viene testato 20 volte per l'esecuzione di un certo compito. I tempi di esecuzione ottenuti (in secondi) sono i seguenti (vedi file exp_numbers_data.csv):

2.42, 0.56, 0.41, 0.72, 0.38, 0.14,
1.17, 1.64, 1.13, 0.18, 0.06, 2.04,
1.6, 0.33, 0.39, 0.55, 0.61, 0.39,
2.08, 1.08

Dire se i dati sono discreti o continui e (eventualmente con l'aiuto di un software) rappresentare il corrispondente grafico a barre/istogramma (a seconda del tipo di dati). ■

Soluzione I dati sono continui. ■

Esercizio 1.3 **(consigliato 31/01)** Con l'aiuto di un software (R o Python ad esempio), rappresentare in istogrammi i dati dei seguenti files, e dire inoltre se le relative distribuzioni sono simmetriche, unimodali/bimodali/..., asimmetriche a destra o a sinistra:

- minutes_data.csv , relativo al minuto di iscrizione a un esame;
- erasmus.csv , relativo ai risultati degli esami (espressi con un numero da 0 a 100) per studenti Erasmus;
- Country-data.csv , colonna total_fer , relativo al tasso di fertilità per paese;
- Country-data.csv , colonna income , relativo al reddito medio per paese.

■

Esercizio 1.4 **(consigliato 07/02)** Per gli esercizi 1.1, 1.2 (e nel caso 1.3), calcolare media e mediana campionarie, deviazione standard campionaria, disegnare il boxplot, individuare eventuali outliers. ■

Soluzione • Esercizio 1.1: Media e varianza campionarie sono (dette p le frequenze relative dei dati)

$$\bar{x} = \sum_{k=0}^{7} k \cdot p(k) = 0 \cdot \frac{1}{12} + 1 \cdot \frac{4}{12} + 2 \cdot \frac{1}{12} + \ldots = 2.42,$$

$$\begin{aligned}\sigma(x)^2 &= \frac{12}{11}\left(\sum_{k=0}^{7} k^2 \cdot p(k) - \bar{x}^2\right) \\ &= \frac{12}{11}\left(0^2 \cdot \frac{1}{12} + 1^2 \cdot \frac{4}{12} + 2^2 \cdot \frac{1}{12} + \ldots - 2.42^2\right) = 1.87^2\end{aligned}$$

La mediana campionaria è la media del sesto e del settimo dato ordinato. Notiamo

dalle frequenze relative che il sesto dato è 2, mentre il settimo è 3. Quindi la mediana è $Q_2 = 2.5$.
Poiché $12 * 0.25$ è intero, il primo e terzo quartile sono rispettivamente la media del terzo e del quarto dato ordinato, cioè $Q_1 = 1$, e la media del nono e del decimo dato ordinato, cioè $Q_3 = 3$; lo scarto interquartile è $R = 3 - 1 = 2$.
Per gli outlier, usiamo la regola (empirica: attenzione caso per caso) dei dati che distano più di $R \cdot 1.5 = 3$ dall'intervallo $[Q_1, Q_3] = [1,3]$: il dato 7 risulta l'unico outlier, e in effeti dista in modo "significativo" dagli altri dati.

- Esercizio 1.2: Media e varianza campionarie sono

$$\bar{x} = \frac{1}{20}(2.42 + 0.56 + 0.41 + \ldots + 1.08) = 0.894,$$
$$\sigma(x)^2 = \frac{20}{19}\left(\frac{1}{20}(2.42^2 + 0.56^2 + 0.41^2 + \ldots + 1.08^2) - 0.894^2\right) = 0.511$$

Per la mediana campionaria, ordiniamo i dati:

0.06, 0.14, 0.18, 0.33, 0.38, 0.39, 0.39, 0.41, 0.55, 0.56,
0.61, 0.72, 1.08, 1.13, 1.17, 1.60, 1.64, 2.04, 2.08, 2.42

La mediana è quindi $Q_2 = (0.56 + 0.61)/2 = 0.585$. Analogamente, il primo e il terzo quartile sono $Q_1 = 0.385$ e $Q_3 = 1.385$, lo scarto interquartile è $R = Q_3 - Q_1 = 1$. Per gli outlier, usiamo la regola (empirica: attenzione caso per caso) dei dati che distano più di $R \cdot 1.5 = 1.5$ dall'intervallo $[Q_1, Q_3] = [0.385, 1.385]$: non ci sono outlier in questo caso. ■

Esercizio 1.5 **(consigliato 07/02)** Un campione di 30 persone viene diviso per titolo di studio. Nel gruppo delle 10 persone con almeno una laurea, il reddito annuo medio è di 30 mila euro. Nel gruppo delle restanti 20 persone senza laurea, il reddito annuo medio è di 25 mila euro. Qual è il reddito annuo medio dell'intero campione di 30 persone?
(più difficile) Se la deviazione standard empirica del gruppo dei laureati è di 4 mila euro, mentre quella del gruppo dei non laureati è di 3 mila euro, quanto vale la varianza empirica dell'intero campione? ■

Soluzione La media totale *non* è la media aritmetica dei due sottogruppi, perché questi hanno numerosità diversa. Piuttosto, detti $n_1 = 10$, $n_2 = 20$, $n = n_1 + n_2 = 30$, chiamiamo $x_1, \ldots x_{n_1}$ i dati del primo sottogruppo, $x_{n_1+1}, \ldots x_{n_1+n_2}$ i dati del secondo sottogruppo, $\overline{x^{(1)}}$, $\overline{x^{(a)}}$ le medie rispettivamente del primo e del secondo sottogruppo. Abbiamo (dati in migliaia di euro)

$$\begin{aligned}
\bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\
&= \frac{1}{n}\left(\sum_{i=1}^{n_1} x_i + \sum_{i=n_1+1}^{n_1+n_2} x_i\right) \\
&= \frac{1}{n}\left(n_1\overline{x^{(1)}} + n_2\overline{x_{(2)}}\right) \\
&= \frac{1}{30}(10 \cdot 30 + 20 \cdot 25) = \frac{800}{30} = 26.67.
\end{aligned}$$

Per la varianza empirica (lavoriamo con la empirica per non avere il fattore $n-1$, la campionaria si ricava da questa facilmente), ricordiamo la relazione

$$\sigma_e^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$$

Dapprima calcoliamo in modo analogo le medie totali dei quadrati dei dati a partire dalle medie dei sottogruppi (chiamiamo $\sigma_e^{(i)}$ la varianza empirica dell'i-simo gruppo):

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} x_i^2 &= \frac{1}{n}\left(n_1 \cdot \frac{1}{n_1}\sum_{i=1}^{n_1} x_i^2 + n_2 \cdot \frac{1}{n_2}\sum_{i=n_1+1}^{n_1+n_2} x_i^2\right)\\ &= \frac{1}{n}\left(n_1(\sigma_e^{(1)})^2 + n_1(\overline{x^{(1)}})^2 + n_2(\sigma_e^{(2)})^2 + n_2(\overline{x^{(2)}})^2\right).\end{aligned}$$

Da qui ricaviamo

$$\sigma_e^2 = \frac{1}{n}\left(n_1(\sigma_e^{(1)})^2 + n_1(\overline{x^{(1)}})^2 + n_2(\sigma_e^{(2)})^2 + n_2(\overline{x^{(2)}})^2\right) - \bar{x}^2.$$

Inserendo i dati si ottiene il valore numerico. ■

Esercizio 1.6 — Devore 1.48. (consigliato 07/02) Sono stati rilevati i seguenti dati sulle concentrazioni di endotossine nella polvere sedimentata per un campione di case urbane e un campione di case rurali (vedi file concentrazioni_endotossine.csv):

- Case urbane (U): 6.0, 5.0, 11.0, 33.0, 4.0, 5.0, 80.0, 18.0, 35.0, 17.0, 23.0
- Case rurali (F): 4.0, 14.0, 11.0, 9.0, 9.0, 8.0, 4.0, 20.0, 5.0, 8.9, 21.0, 9.2, 3.0, 2.0, 0.3

a. Determinare i valori di media e deviazione standard per ciascun campione, interpretare questi valori e confrontare la variabilità nei due campioni.
b. Calcolare lo scarto interquartile per ciascun campione e confrontare. Lo scarto interquartile trasmette lo stesso messaggio della deviazione standard?
c. Vengono fornite anche le concentrazioni di endotossine nella polvere delle borse per aspirapolvere:
   - Case urbane (U): 34.0, 49.0, 13.0, 33.0, 24.0, 24.0, 35.0, 104.0, 34.0, 40.0, 38.0, 1.0
   - Case rurali (F): 2.0, 64.0, 6.0, 17.0, 35.0, 11.0, 17.0, 13.0, 5.0, 27.0, 23.0, 28.0, 10.0, 13.0, 0.2

   Costruire un boxplot comparativo (cioè un grafico contenente i quattro boxplot) e confrontare i quattro campioni.

■

Soluzione Sul boxplot, notare che, ad esempio, le deviazioni standard campionarie di urbane polvere sedimentata e urbale aspirapolvere sono simili, ma i boxplot hanno forme diverse: il boxplot di urbane polvere sedimentata ha un rettangolo più ampio, segno di una dispersione più ampia nel 50% dei dati centrali, mentre ha baffi più corti, mentre il boxplot di urbane aspirapolvere ha un rettangolo più corto ma baffi più pronunciati, segno di una maggiore dispersione del 50% dei dati più estremi, e un outlier molto distante dal rettangolo. ■

NB Saper fare:

- Calcolare le frequenze campionarie relative e assolute a partire dai dati.
- Disegnare il grafico a barre/l'istogramma.

- Calcolare gli indici di posizione (media, mediana, quantili, moda) e di dispersione (varianza) a partire dai dati.
- Calcolare gli indici di posizione e di dispersione a partire dalle frequenze relative campionarie.
- Disegnare il boxplot e interpretarlo.
- Confrontare due distribuzioni tramite indici e boxplot.

2. Dati Multivariati

NB Qui alcuni comandi utili per R (si veda https://www.rdocumentation.org/ per la documentazione):

- cor per il calcolo del coefficiente di correlazione.
- modello = lm(y∼x, data = nome file) per il calcolo della retta di regressione (in realtà il comando fa molto di più), summary(lm) per la stampa, tra le altre cose, dei coefficienti della retta (la stampa restituisce anche "residuals": di che si tratta?).
- plot(data$x, data$y) per lo scatterplot; abline(modello) per il disegno della retta di regressione.

Esercizio 2.1 **(consigliato 14/02)** Sono riportati i dati (inventati) relativi alla sperimentazione di un farmaco su un campione di 12 pazienti: per ogni i, x_i è la dose del farmaco somministrata all'i-simo paziente, y_i è la variazione di un certo parametro corporeo nell'i-simo paziente in seguito alla somministrazione:

i	1	2	3	4	5	6	7	8	9	10	11	12
x_i	1.8	1.9	1.8	2.4	5.1	3.1	5.5	5.1	5.3	4.9	3.7	3.8
y_i	21.0	33.5	24.6	40.7	73.2	24.9	40.4	45.3	64.0	62.7	47.2	44.3

C'è una significativa correlazione lineare? se sì, calcolare la retta di regressione. ■

Soluzione Calcoliamo come nel capitolo 1 media e varianza campionarie dei singoli dati univariati x_i e y_i:

$$\bar{x} = 3.7, \quad \sigma(x)^2 = 2.15, \qquad \bar{y} = 43.48, \quad \sigma(y)^2 = 274.13$$

Calcoliamo ora la covarianza e il coefficiente di correlazione campionario:

$$\operatorname{cov}(x,y) = \frac{12}{11}\left(\frac{1}{12}(1.8\cdot 21.0 + 1.9\cdot 33.5 + \ldots + 3.8\cdot 44.3) - 3.7\cdot 43.48\right) = 18.72,$$
$$r(x,y) = \frac{\operatorname{cov}(x,y)}{\sigma(x)\sigma(y)} = 0.77.$$

Poiché il coefficiente è in modulo non distante da 1, c'è una buona correlazione lineare, e ha quindi senso calcolare la retta di regressione. I coefficienti di tale retta sono (b^* coefficiente angolare, a^* ordinata all'origine)

$$b^* = \frac{\operatorname{Cov}(x,y)}{\sigma(x)^2} = 8.71, \quad a^* = \bar{y} - b^*\bar{x} = 11.25.$$

■

Esercizio 2.2 **(consigliato 14/02)** Per ciascuno dei seguenti valori di r, si disegni uno o più esempi di scatterplot (grafici cartesiani che rappresentano le coppie di punti (x_i, y_i) aventi r come correlazione campionaria: $r = 0$, $r = \pm 0.6$, $r = \pm 0.9$. ■

Esercizio 2.3 **(consigliato 14/02)** Con l'aiuto di un software (R o Python ad esempio), costruire uno scatterplot dei seguenti dati bivariati, calcolare il coefficiente di correlazione e, dove significativo, disegnare la retta di regressione:

- Country-data.csv , colonne income e total_fer;
- Country-data.csv , colonne total_fer e life_expec (aspettativa di vita);
- dati_randomizzati.csv , colonne total_fer_shuffled (ottenuto riordinando in modo casuale Country-data.csv, total_fer) e life_expec.

Quanto bene la retta approssima i dati? ■

Soluzione Nel primo e nel secondo esempio, approssimazione discreta (r tra 0.5 e 0.75), nel terzo buona (r tra 0.75 e 0.9), nel quarto assente (r prossimo a 0). ■

Esercizio 2.4 Un cambiamento di unità di misura è una trasformazione affine dei dati: $\tilde{x}_i = \alpha x_i + \beta$, per opportuni α, β con $\alpha \neq 0$.

- Dimostrare che la media $\bar{\tilde{x}}$ e la varianza $\operatorname{var}(\tilde{x})$ campionarie nella nuova unità di misura soddisfano

$$\bar{\tilde{x}} = \alpha\bar{x} + \beta,$$
$$\operatorname{var}(\tilde{x}) = \alpha^2 \operatorname{var}(x).$$

- Dimostrare che il coefficiente di correlazione campionaria r è invariante per cambiamento di unità di misura: per ogni $\alpha, \beta, \gamma, \delta$ con $\alpha > 0$, $\gamma > 0$, detti $\tilde{x}_i = \alpha x_i + \beta$, $\tilde{y}_i = \gamma y_i + \delta$, il coefficiente di correlazione campionaria di $(\tilde{x}_i, \tilde{y}_i)$ è lo stesso di (x_i, y_i).
- Sia b^* il coefficiente angolare della retta di regressione. Dimostrare che

$$b^* = \frac{\sigma(y)}{\sigma(x)} r.$$

- Dati (x_i, y_i), $i = 1, \dots n$, e dato un qualunque $\tilde{b}^* \in \mathbb{R}$ non nullo, costruire dei nuovi dati bivariati $(\tilde{x}_i, \tilde{y}_i)$ con lo stesso modulo del coefficiente di correlazione di (x_i, y_i) ma coefficiente angolare della retta di regressione pari a $\tilde{b}^*$.

■

Soluzione Per il primo punto, usando le definizioni di media e varianza, abbiamo

$$\begin{aligned}
\bar{\tilde{x}} &= \frac{1}{n}\sum_{i=1}^{n}(\alpha x_i + \beta) = \frac{1}{n}\alpha\sum_{i=1}^{n}\tilde{x}_i + \frac{1}{n}\sum_{i=1}^{n}\beta = \alpha\bar{x} + \beta, \\
\operatorname{var}(x) &= \frac{1}{n-1}\sum_{i=1}^{n}(\tilde{x}_i - \bar{\tilde{x}})^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(\alpha x_i + \beta - \alpha\bar{x} - \beta)^2 \\
&= \frac{1}{n-1}\alpha^2\sum_{i=1}^{n}(x_i - \bar{x})^2 = \alpha^2 \operatorname{var}(x).
\end{aligned}$$

Per il secondo punto, usando le definizioni di covarianza e coefficiente di correlazione abbiamo

$$\begin{aligned}
\operatorname{cov}(x,y) &= \frac{1}{n-1}\sum_{i=1}^{n}(\tilde{x}_i - \bar{\tilde{x}})(\tilde{y}_i - \bar{\tilde{y}}) \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(\alpha x_i + \beta - \alpha\bar{x} - \beta)(\gamma x_i + \delta - \gamma\bar{x} - \delta) \\
&= \frac{1}{n-1}\alpha\gamma\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \alpha\gamma\operatorname{cov}(x,y)
\end{aligned}$$

e quindi

$$r(\tilde{x},\tilde{y}) = \frac{\operatorname{cov}(\tilde{x},\tilde{y})}{\sigma(\tilde{x})\sigma(\tilde{y})} = \frac{\alpha\gamma\operatorname{cov}(x,y)}{\alpha\sigma(x)\gamma\sigma(y)} = r(x,y).$$

Per il terzo punto, abbiamo

$$b^* = \frac{\operatorname{cov}(x,y)}{\sigma(x)^2} = \frac{\sigma(y)}{\sigma(x)} \cdot \frac{\operatorname{cov}(x,y)}{\sigma(x)\sigma(y)} = \frac{\sigma(y)}{\sigma(x)} r.$$

Infine, per il quarto punto, consideriamo dapprima il caso in cui $\tilde{b}^*$ ha lo stesso segno del coefficiente angolare b dei dati $(x_i, y_i$. Consideriamo un cambio di unità di misura $\tilde{x}_i = \alpha x_i$, $\tilde{y}_i = \gamma y_i$ con $\alpha > 0$, $\gamma > 0$ da determinare. Per il secondo punto, i nuovi dati $(\tilde{x}_i, \tilde{y}_i)$ avranno lo stesso coefficiente di regressione r, mentre, per il terzo punto, il nuovo coefficiente angolare sarà

$$\frac{\sigma(\tilde{y})}{\sigma(\tilde{x})} r = \frac{\alpha\sigma(x)}{\gamma\sigma(y)} r = \frac{\alpha}{\gamma} b^*.$$

Perché quest'ultimo sia il coefficiente voluto $\tilde{b}^*$, basta scegliere α e γ tali che $\alpha/\gamma = \tilde{b}^*/b^*$ (ad esempio $\gamma = 1$, $\alpha = \tilde{b}^*/b^* > 0$ funziona).
Nel caso in cui b^* e $\tilde{b}^*$ non hanno lo stesso segno, si può ripetere il ragionamento prendendo però uno tra α e γ negativo, in questo caso r cambia di segno (lasciamo i dettagli per esercizio).
■

Saper fare:

- Calcolare la covarianza e il coefficiente di correlazione campionari (per dati bivariati).
- Individuare il grado di dipendenza lineare a partire dal coefficiente di correlazione.
- Calcolare (se opportuno) la retta di regressione e darne il significato.

3. Probabilità e (In)dipendenza

Esercizio 3.1 **(consigliato 14/02)** Dati tre eventi A, B, C in uno spazio campionario, esprimere come insiemi le seguenti affermazioni:

- si verifica A ma non B;
- si verifica o A o B;
- si verifica almeno uno tra A, B, C;
- si verifiano tutti gli eventi A, B, C;
- non si verifica nessuno tra A, B, C;
- si verificano almeno due tra A, B, C.

■

Soluzione

- si verifica A ma non B: $A \setminus B$;
- si verifica o A o B: $A \Delta B := (A \setminus B) \cup (B \setminus A) = (A \cup B) \setminus (A \cap B)$;
- si verifica almeno uno tra A, B, C: $A \cup B \cup C$;
- si verifiano tutti gli eventi A, B, C: $A \cap B \cap C$;
- non si verifica nessuno tra A, B, C: $(A \cup B \cup C)^c$;
- si verificano almeno due tra A, B, C: $(A \cap B) \cup (A \cap C) \cup (B \cap C)$.

■

Esercizio 3.2 Una ditta classifica le sue consegne come in città o fuori città e come urgenti o non urgenti. Sappiamo che il 50% delle consegne è urgente, il 40% è in città e il 20% è urgente e fuori città. Calcolare le probabilità che una consegna sia rispettivamente: a) urgente e in città; b) urgente o fuori città. ■

Soluzione Anche se non è richiesto, individuiamo uno spazio di probabilità opportuno: gli esiti possibili sono tutte le coppie (tempo,spazio), dove tempo può essere urgente o non urgente

e spazio può essere in città o fuori città. In formule,

$$\Omega = \{(x,y) \mid x \in \{u,n\}, y \in \{i,f\}\} = \{u,n\} \times \{i,f\},$$

dove u,n,i,f stanno rispettivamente per urgente, non urgente, in città, fuori città. L'evento {urgente} si scrive come $\{u\} \times \{i,f\}$, l'evento {urgente e in città} come $\{(u,i)\}$.

Scriviamo ora i dati del problema (sempre consigliato scriverli): $\mathbb{P}(\text{urgente}) = 0.5$, $\mathbb{P}(\text{città}) = 0.4$, $\mathbb{P}(\text{urgente e fuori città}) = 0.2$.

- Poiché {urgente e in città} = {urgente} \ {urgente e fuori città}, abbiamo

$$\mathbb{P}(\text{urgente e in città}) = \mathbb{P}(\text{urgente}) - \mathbb{P}(\text{urgente e fuori città}) = 0.3.$$

- Abbiamo

$$\begin{aligned}\mathbb{P}(\text{urgente o fuori città}) &= \mathbb{P}(\text{urgente}) + \mathbb{P}(\text{fuori città}) - \mathbb{P}(\text{urgente e fuori città}) \\ &= \mathbb{P}(\text{urgente}) + 1 - \mathbb{P}(\text{in città}) - \mathbb{P}(\text{urgente e fuori città}) = 0.9.\end{aligned}$$

■

Esercizio 3.3 **(consigliato 21/02)** Secondo le previsioni meteo, domani a Pisa ci sarà: pioggia con probabilità del 70%, vento con probabilità del 50%, né pioggia né vento con probabilità del 20%. Calcolare le probabilità di: a) pioggia o vento; b) pioggia e vento; c) pioggia ma non vento; d) vento, sapendo che pioverà; e) assenza di vento, sapendo che pioverà; f) vento, sapendo che non pioverà. ■

Soluzione Un possibile spazio campionario è $\Omega = \{p,np\} \times \{v,nv\}$, dove p sta per pioggia, np sta per non pioggia ecc. Chiamiamo $A = \{\text{pioggia}\} = \{p\} \times \{v,nv\}$, $B = \{\text{vento}\} = \{p,np\} \times \{v\}$. Le informazioni del problema sono $\mathbb{P}(A) = 0.7$, $\mathbb{P}(B) = 0.5$, $\mathbb{P}((A \cup B)^c) = 0.2$.

a) $\mathbb{P}(A \cup B) = 1 - \mathbb{P}(A \cap B) = 0.8$.

b) $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$, da cui $\mathbb{P}(A \cap B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cup B) = 0.4$.

c) $\mathbb{P}(A \setminus B) = \mathbb{P}(A) - \mathbb{P}(A \cap B) = 0.3$.

d)

$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} = \frac{0.4}{0.7} = \frac{4}{7}.$$

e) $\mathbb{P}(B^c \mid A) = 1 - \mathbb{P}(B \mid A) = 3/7$.

f)

$$\mathbb{P}(B \mid A^c) = \frac{\mathbb{P}(A^c \cap B)}{\mathbb{P}(A^c)} = \frac{0.1}{0.3} = \frac{1}{3},$$

dove abbiamo usato $\mathbb{P}(A^c \cap B) = \mathbb{P}(B) - \mathbb{P}(A \cap B) = 0.1$ e $\mathbb{P}(A^c) = 1 - \mathbb{P}(A) = 0.3$. Notare che $\mathbb{P}(B \mid A^c) \neq 1 - \mathbb{P}(B \mid A)$.

■

Esercizio 3.4 Da un mazzo di 40 carte napoletane (quattro semi: bastoni, coppe, denari, spade; ciascuno seme ha carte da 1 a 10) vengono estratte quattro carte, senza rimpiazzo. a) Qual è la probabilità che le prime due siano di denaro e la quarta di coppe? b) Qual è la probabilità che

escano esattamente due carte di denaro (in qualcunque ordine)? ■

Soluzione Rappresentiamo il mazzo di carte con l'insieme

$$S = \{1_B, \dots 10_B, 1_C, \dots 10_C, 1_D, \dots 10_D, 1_S, \dots 10_S\},$$

dove a_B è la carta di valore a del seme bastoni, a_C è la carta di valore a del seme coppe e così via. Uno spazio campionario è $\Omega = \{(\omega_1, \dots \omega_4) \mid \omega_i \in S \text{ tutti distinti}\}$, con $\#\Omega = 40!/36! = 40 \cdot 39 \cdot 38 \cdot 37$. Poiché si tratta di estrazioni con ordine senza rimpiazzo, la probabilità $\mathbb{P}$ è uniforme su Ω.

a) Piuttosto che calcolare direttamente la cardinalità di $A = \{\text{prime due di denaro, quarta di coppe}\}$, facciamo questa osservazione: la cardinalità di A è la stessa di $A' = \{\text{prime due di denaro, terza di coppe}\}$, e quest'ultimo ha cardinalità

$$\begin{aligned}\#A' &= \#\text{scelte delle prime due carte di denaro} \cdot \#\text{scelte della terza di coppe} \cdot \\ &\quad \cdot \#\text{scelte della quarta} \\ &= (10 \cdot 9) \cdot 10 \cdot 37.\end{aligned}$$

La probabilità cercata è quindi

$$\mathbb{P}(A) = \frac{\#A}{\#\Omega} = \frac{10 \cdot 9 \cdot 10 \cdot 37}{40 \cdot 39 \cdot 38 \cdot 37} = 0.0152.$$

b) Poiché l'estrazione è senza rimpiazzo e in questo caso non ci interessa l'ordine, possiamo usare come nuovo spazio campionario $\Omega' = \{C \subseteq S \mid \#C = 4\}$, su cui la probabilità $\mathbb{P}'$ è ancora uniforme (trattandosi di estrazioni senza ordine senza rimpiazzo); vale inoltre $\#\Omega' = \binom{40}{4}$. L'evento $B = \{\text{esattamente due carte di denaro}\}$ ha cardinalità

$$\begin{aligned}\#B &= \#\text{scelte delle due carte di denaro} \cdot \#\text{scelte delle due carte non di denaro} \\ &= \#\text{sottoinsiemi di 2 elementi tra 10} \cdot \#\text{sottoinsiemi di 2 elementi tra 30} \\ &= \binom{10}{2} \cdot \binom{30}{2}.\end{aligned}$$

La probabilità cercata è quindi

$$\mathbb{P}'(B) = \frac{\#B}{\#\Omega'} = \frac{\binom{10}{2} \cdot \binom{30}{2}}{\binom{40}{4}} = \frac{\frac{10 \cdot 9}{2} \cdot \frac{30 \cdot 29}{2}}{\frac{40 \cdot 39 \cdot 38 \cdot 37}{4 \cdot 3 \cdot 2}} = \frac{9 \cdot 15 \cdot 29}{13 \cdot 37 \cdot 38} = 0.214.$$

■

Esercizio 3.5 **(consigliato 21/02)** Tre turisti (tra loro estranei) arrivano in un paese con cinque alberghi, avendo prenotato ciascuno un albergo in modo del tutto casuale. a) Qual è la probabilità che si trovino tutti in alberghi differenti? b) Qual è la probabilità che si trovino tutti nello stesso albergo? ■

Soluzione L'esperimento equivale a estrarre a caso, con ordine e con ripetizione, un albergo tra i 5 (poiché c'è ripetizione, conviene considerare comunque l'ordine per usare il modello uniforme). Uno spazio campionario è $\Omega = \{(a,b,c) \mid a,b,c \in \{1,2,\dots 5\}\} = \{1,2,\dots 5\}^3$, con

$\#\Omega = 5^3$, la probabilità $\mathbb{P}$ su Ω è uniforme.

a) L'evento $A = \{\text{tutti alberghi differenti}\}$ corrisponde a tutte le 3 estrazioni con ordine *senza* ripetizione, quindi ha cardinalità $5 \cdot 4 \cdot 3$ e probabilità

$$\mathbb{P}(A) = \frac{\#A}{\#\Omega} = \frac{12}{25}.$$

b) L'evento $B = \{\text{tutti nello stesso albergo}\}$ si scrive come $B = \{(a,a,a) \mid a \in \{1,2,\dots 5\}\}$ e ha quindi cardinalità $\#B = 5$ (le 5 possibili scelte dell'albergo). La probabilità di B è quindi $5/5^3 = 1/25$.

■

Esercizio 3.6 In una classe di 10 alunni, di cui 6 donne e 4 uomini, vengono estratte a sorte 5 persone. a) Qual è la probabilità che, tra le persone estratte, ci siano esattamente 3 donne? b) Qual è la probabilità che ci siano almeno 4 donne? ■

Soluzione Chiamiamo $S = \{1_D, \dots 6_D, 1_U, \dots 4_U\}$ l'insieme delle persone della classe. Siamo nel caso di estrazioni senza ordine senza rimpiazzo, uno spazio campionario è quindi $\Omega = \{C \subseteq S \mid \#C = 5\}$, con $\#\Omega = \binom{10}{5}$, e la probabilità $\mathbb{P}$ su Ω è uniforme.

a) L'evento $A = \{\text{esattamente 3 donne}\}$ ha cardinalità e probabilità

$$\begin{aligned} \#A &= \#\text{scelte delle 3 donne su 6} \cdot \#\text{scelte dei 2 uomini su 4} \\ &= \binom{6}{3} \cdot \binom{4}{2}, \\ \mathbb{P}(A) &= \frac{\#A}{\#\Omega} = \frac{\binom{6}{3} \cdot \binom{4}{2}}{\binom{10}{5}} = \frac{10}{21} = 0.476. \end{aligned}$$

b) L'evento $B = \{\text{almeno 4 donne}\}$ è l'unione disgiunta di $B_4 = \{\text{esattamente 4 donne}\}$ e $B_5 = \{\text{esattamente 5 donne}\}$, che hanno probabilità (procedendo come sopra)

$$\begin{aligned} \mathbb{P}(B_4) &= \frac{\binom{6}{4} \cdot \binom{4}{1}}{\binom{10}{5}} = \frac{5}{3 \cdot 7}, \\ \mathbb{P}(B_5) &= \frac{\binom{6}{5}}{\binom{10}{5}} = \frac{3}{9 \cdot 7 \cdot 2}, \end{aligned}$$

quindi

$$\mathbb{P}(B) = \mathbb{P}(B_4) + \mathbb{P}(B_5) = \frac{11}{42}.$$

■

NB Notiamo che, negli esercizi 3.4 e 3.6,

- c'è un gruppo di N elementi,
- con un sottogruppo di N_1 elementi, e
- in un estrazione senza rimpiazzo di n elementi del gruppo
- si vuole la probabilità dell'evento $A = \{\text{esattamente } k \text{ degli estratti siano del sottogruppo}\}$.

Lo stesso ragionamento porta alla formula per tale probabilità

$$\mathbb{P}(A) = \frac{\binom{N_1}{k}\binom{N-N_1}{n-k}}{\binom{N}{n}}.$$

Esercizio 3.7 — Ross 3.19, paradosso dei compleanni. Consideriamo un gruppo di 23 persone scelte a caso. Poiché ogni coppia di persone condivide lo stesso compleanno con probabilità $1/365$ e ci sono $\binom{25}{2} = 253$ coppie, perché la probabilità che almeno due persone abbiano lo stesso compleanno non è uguale a $253/365$? Quanto vale invece questa probabilità? ■

Soluzione Il ragionamento proposto non è corretto perché gli eventi $B_{i,j} = \{i \text{ e } j \text{ hanno lo stesso compleanno}\}$ non sono disgiunti, quindi la probabilità che almeno due persone abbiano lo stesso compleanno non è la somma delle probabilità dei $B_{i,j}$. Piuttosto, possiamo calcolare la probabilità cercata come segue: Si tratta di estrazioni di 23 giorni con ordine con ripetizione, $\Omega = \{1,2,\dots 365\}^{23}$ con probabilità $\mathbb{P}$ uniforme. L'evento $A = \{\text{almeno 2 persone con lo stesso compleanno}\}$ ha per complementare $A^c = \{\text{tutti i compleanni distinti}\}$, cioè tutte le sequenze di 23 elementi distinti, quindi

$$\mathbb{P}(A^c) = \frac{\#A^c}{\#\Omega} = \frac{365 \cdot (365-1) \cdot \ldots \cdot (365-22)}{365^{23}} = 0.4927.$$

Quindi $\mathbb{P}(A) = 1 - \mathbb{P}(A^c) = 0.5073$ (maggiore di 0.5!). ■

Esercizio 3.8 — Ross 3.45. Ci sono n tipi di coupon in un'urna, ciascun tipo $j \in \{1,\dots N\}$ è in proporzione p_j (con $p_j > 0$, $\sum_{j=1}^n p_j = 1$). Estraiamo k coupon con reimmissione. Qual è la probabilità che tra i k coupon estratti, compaia o il primo tipo o il secondo? (ad esempio, con $k = 3$, le terne $(1,3,4)$, $(2,5,4)$ soddisfano il requisito, mentre le terne $(3,4,5)$, $(1,2,4)$ non lo soddisfano). ■

Esercizio 3.9 (consigliato 21/02) Nella città di Urbopoli, si stima che, tra coloro che abitualmente utilizzano gli autobus pubblici, il 20% lo faccia sistematicamente senza pagare il biglietto. Tra coloro che non pagano il biglietto, il 70% sono individui di età inferiore a 15 anni, mentre tra coloro che lo pagano chi ha meno di 15 anni rappresenta il 40%.

a) Si sceglie a caso un individuo (nella popolazione di coloro che utilizzano gli autobus); quanto vale la probabilità che si tratti di un individuo di età inferiore a 15 anni?
b) Se l'individuo scelto ha meno di 15 anni, quanto vale la probabilità che egli paghi abitualmente il biglietto?
c) Se l'individuo scelto ha 15 anni o più, quanto vale la probabilità che egli abitualmente non paghi il biglietto?

■

Soluzione Uno spazio campionario è $\Omega = \{\text{pagato}, \text{non pagato}\} \times \{(< 15 \text{ anni}, \geq 15 \text{ anni})\}$. Con un po' di abuso di notazione indichiamo con $\{\text{pagato}\}$ l'evento "biglietto pagato", che (come evento di Ω) è $\{\text{pagato}\} \times \{(< 15 \text{ anni}, \geq 15 \text{ anni})\}$; analogamente per altri eventi.
I dati del problema sono $\mathbb{P}(\text{non pagato}) = 0.2$, $\mathbb{P}(< 15 \text{ anni} \mid \text{non pagato}) = 0.7$, $\mathbb{P}(< 15 \text{ anni} \mid$

pagato) $= 0.4$.

a) Per la formula di fattorizzazione,

$$\begin{aligned}\mathbb{P}(< 15 \text{ anni}) &= \mathbb{P}(< 15 \text{ anni} \mid \text{non pagato}) \cdot \mathbb{P}(\text{non pagato}) \\ &\quad + \mathbb{P}(< 15 \text{ anni} \mid \text{pagato}) \cdot \mathbb{P}(\text{pagato}) \\ &= 0.7 \cdot 0.2 + 0.4 \cdot 0.8 = 0.46.\end{aligned}$$

b) Per la formula di Bayes,

$$\mathbb{P}(\text{pagato} \mid < 15 \text{ anni}) = \frac{\mathbb{P}(< 15 \text{ anni} \mid \text{pagato}) \cdot \mathbb{P}(\text{pagato})}{\mathbb{P}(< 15 \text{ anni})} = 0.696.$$

c) Sempre per la formula di Bayes,

$$\begin{aligned}\mathbb{P}(\text{non pagato} \mid \geq 15 \text{ anni}) &= \frac{\mathbb{P}(\geq 15 \text{ anni} \mid \text{non pagato}) \cdot \mathbb{P}(\text{non pagato})}{\mathbb{P}(\geq 15 \text{ anni})} \\ &= \frac{\mathbb{P}(\geq 15 \text{ anni} \mid \text{non pagato}) \cdot \mathbb{P}(\text{non pagato})}{\mathbb{P}(\geq 15 \text{ anni} \mid \text{non pagato}) \cdot \mathbb{P}(\text{non pagato}) + \mathbb{P}(\geq 15 \text{ anni} \mid \text{pagato}) \cdot \mathbb{P}(\text{pagato})} \\ &= \frac{0.3 \cdot 0.2}{0.3 \cdot 0.2 + 0.4 \cdot 0.8} = 0.111.\end{aligned}$$

■

Esercizio 3.10 Un'assicurazione auto classifica i guidatori come prudenti o imprudenti. Un guidatore prudente ha almeno un incidente l'anno con probabilità dell'1%, un guidatore imprudente con probabilità del 5%. Il 60% dei guidatori assicurati è classificato come prudente.

a) Qual è la probabilità che un assicurato abbia almeno un incidente in un dato anno?
b) Se un assicurato ha avuto almeno un incidente, qual è la probabilità che sia classificato come prudente?

■

Soluzione a) Per la formula di fattorizzazione

$$\begin{aligned}\mathbb{P}(\text{incidente}) &= \mathbb{P}(\text{incidente} \mid \text{prudente}) \cdot \mathbb{P}(\text{prudente}) \\ &\quad + \mathbb{P}(\text{incidente} \mid \text{imprudente}) \cdot \mathbb{P}(\text{imprudente}) \\ &= 0.01 \cdot 0.6 + 0.05 \cdot 0.4 = 0.026.\end{aligned}$$

b) Per la formula di Bayes

$$\mathbb{P}(\text{prudente} \mid \text{incidente}) = \frac{\mathbb{P}(\text{incidente} \mid \text{prudente}) \cdot \mathbb{P}(\text{prudente})}{\mathbb{P}(\text{incidente})} = 0.231.$$

■

Esercizio 3.11 **(consigliato 21/02)** Un furto è stato commesso a Urbopoli. Sulla base degli indizi raccolti, il commissario è convinto che il sig. Rossi sia colpevole di tale furto con probabilità del 40% (cioè, su 100 casi simili, il sig. Rossi sarebbe colpevole in 40 casi). Non ci sono altri sospettati, si sa solo che il colpevole è di Urbopoli. Il commissario viene ora a conoscenza di un nuovo indizio: il colpevole ha i capelli biondi, come il sig. Rossi. Da una

statistica, emerge che il 20% degli abitanti di Urbopoli ha i capelli biondi. Come cambia la valutazione del commissario circa la probabilità di colpevolezza del sig. Rossi? ■

Soluzione Uno spazio campionario è $\Omega = \{\text{Rossi}, \text{non Rossi}\} \times \{\text{biondi}, \text{non biondi}\}$, dove l'elemento "Rossi" indica che il colpevole è il sig. Rossi, l'elemento "biondi" indica che il colpevole ha i capelli biondi. Anche in questo caso, con un piccolo abuso di notazione indichiamo ad esempio con $\{\text{Rossi}\}$ anche l'evento "il colpevole è il sig. Rossi", che è l'insieme $\{\text{Rossi}\} \times \{\text{biondi}, \text{non biondi}\}$.
I dati del problema sono: $\mathbb{P}(\text{Rossi}) = 0.4$, $\mathbb{P}(\text{biondi} \mid \text{Rossi}) = 1$ (poiché il sig. Rossi ha i capelli biondi), $\mathbb{P}(\text{biondi} \mid \text{non Rossi}) = 0.2$ (poiché, se il sig. Rossi non è colpevole e non ci sono altri sospettati, l'unica informazione che abbiamo sul colpevole è che è un individuo di Urbopoli). Quindi per Bayes, la nuova probabilità è

$$\begin{aligned}\mathbb{P}(\text{Rossi} \mid \text{biondi}) &= \frac{\mathbb{P}(\text{biondi} \mid \text{Rossi}) \cdot \mathbb{P}(\text{Rossi})}{\mathbb{P}(\text{biondi} \mid \text{Rossi}) \cdot \mathbb{P}(\text{Rossi}) + \mathbb{P}(\text{biondi} \mid \text{non Rossi}) \cdot \mathbb{P}(\text{non Rossi})} \\ &= \frac{1 \cdot 0.4}{1 \cdot 0.4 + 0.2 \cdot 0.6} = 0.769.\end{aligned}$$

■

Esercizio 3.12 Tre università a, b, c sono scelte rispettivamente dal 20%, 50%, 30% degli studenti. La precentuale di studenti che termina il proprio percorso è dell'80% per gli studenti di a, del 60% per gli studenti di b e del 90% per gli studenti di c. Si estrae uno studente a caso. Quali di queste sono partizioni dello spazio campionario?

- scelta di a, scelta di b, scelta di c;
- scelta di a, scelta di b o c;
- scelta di a, scelta di b;
- scelta di a, scelta di a o b, scelta di c.

Calcolare le probabilità che lo studente estratto: a) concluda il proprio percorso; b) scelga a o b e concluda il percorso; c) venga da a, sapendo che ha concluso il percorso. ■

Soluzione La prima e la seconda sono partizioni (accade uno e una solo degli eventi dati). La terza non è una partizione, perché può non accadere nessuno degli eventi dati (può essere scelta c). La quarta non è una partizione, perché possono accadare due eventi dati (se viene scelta a). I dati del problema sono $\mathbb{P}(a) = 0.2$, $\mathbb{P}(b) = 0.5$, $\mathbb{P}(c) = 0.3$, $\mathbb{P}(\text{termine} \mid a) = 0.8$, $\mathbb{P}(\text{termine} \mid a) = 0.6$, $\mathbb{P}(\text{termine} \mid a) = 0.9$. Si può usare una rappresentazione ad albero.

a) Per la formula di fattorizzazione,

$$\mathbb{P}(\text{termine}) = 0.8 \cdot 0.2 + 0.6 \cdot 0.5 + 0.3 \cdot 0.9 = 0.73.$$

b) Per le regole di somma e prodotto,

$$\begin{aligned}\mathbb{P}(a \text{ o } b, \text{termine}) &= \mathbb{P}(a, \text{termine}) + \mathbb{P}(b, \text{termine}) \\ &= \mathbb{P}(\text{termine} \mid a) \cdot \mathbb{P}(a) + \mathbb{P}(\text{termine} \mid b) \cdot \mathbb{P}(b) \\ &= 0.8 \cdot 0.2 + 0.6 \cdot 0.5 = 0.46.\end{aligned}$$

c) Per la formula di Bayes

$$\mathbb{P}(a \mid \text{termine}) = \frac{\mathbb{P}(\text{termine} \mid a) \cdot \mathbb{P}(a)}{\mathbb{P}(\text{termine})} = 0.219.$$

■

Esercizio 3.13 **(consigliato 28/02)** In due lanci di un dado equilibrato, gli eventi "1 al primo lancio" e "somma degli esiti dei due lanci = 3" sono indipendenti? ■

Soluzione Due lanci di un dado equilibrato sono descritti da una probabilità $\mathbb{P}$ uniforme su $\Omega = \{1,2,\dots 6\}^2$. Abbiamo

$$\mathbb{P}(1 \text{ al primo lancio}) = \frac{1}{6},$$
$$\mathbb{P}(\text{somma degli esiti dei due lanci} = 3) = \mathbb{P}\{(1,2),(2,1)\} = \frac{2}{36} = \frac{1}{18},$$
$$\mathbb{P}(1 \text{ al primo, somma} = 3) = \mathbb{P}\{(1,2)\} = \frac{1}{36} \neq \frac{1}{6} \cdot \frac{1}{18}.$$

Quindi i due eventi non sono indipendenti. ■

Esercizio 3.14 — Paradosso dei figli. **(consigliato almeno (a)-(b) 28/02)** In una famiglia ci sono due figli.

a) Se la più grande è femmina, qual è la probabilità che anche la più piccola sia femmina?
b) Se almeno una dei due è femmina, qual è la probabilità che siano entrambe femmine?
c) Vedo una dei due figli, che è femmina, ma non so se sia la più grande o la più piccola. Qual è la probabilità che siano entrambe femmine?

■

Soluzione Uno spazio campionario è $\Omega = \{F,M\}^2 = \{(F,F),(F,M),(M,F),(M,M)\}$, dove ad esempio (F,M) indica che la più grande è femmina e il più piccolo è maschio. Sulla probabilità, è ragionevole assumere che, per ciascun/a figlio/a, maschio e femmina siano equiprobabili, e che il sesso del/la più grande sia indipendente dal sesso del/la più piccola. [Si tratta di due lanci di moneta indipendenti, quindi la probabilità $\mathbb{P}$ deve essere uniforme su Ω.]

a) Per l'indipendenza abbiamo

$$\mathbb{P}(\text{seconda f} \mid \text{prima f}) = \mathbb{P}(\text{seconda f}) = \frac{1}{2}.$$

b) La probabilità cercata è

$$\mathbb{P}(\text{entrambe f} \mid \text{almeno una f}) = \frac{\mathbb{P}(\text{entrambe f e almeno una f})}{\mathbb{P}(\text{almeno una f})}$$
$$= \frac{\mathbb{P}(\text{entrambe f})}{\mathbb{P}(\text{almeno una f})} = \frac{1/4}{3/4} = \frac{1}{3},$$

dove abbiamo calcolato (con l'indipendenza o con la probabilità uniforme su Ω)

$$\mathbb{P}(\text{entrambe f}) = \mathbb{P}(\text{prima f}) \cdot \mathbb{P}(\text{seconda f}) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4},$$
$$\mathbb{P}(\text{almeno una f}) = 1 - \mathbb{P}(\text{entrambi m}) = 1 - \frac{1}{4} = \frac{3}{4}.$$

Notare che $\mathbb{P}(\text{entrambe f} \mid \text{almeno una f}) \neq \mathbb{P}(\text{seconda f} \mid \text{prima f})$!

c) Intuitivamente, il vedere uno/a dei/lle due figli/e non ci dà informazioni sull'altro, quindi ci aspettiamo che la probabilità sia ancora $1/2$. In effetti è così, mostriamo ora questo fatto in modo rigoroso. Consideriamo un nuovo spazio campionario Ω' in modo da comprendere la casualità sulla persona vista:

$$\Omega' = \{F, M\}^2 \times \{1, 2\},$$

dove 1 e 2 indicano rispettivamente aver visto il/la primo/a o il/la secondo/a figlio/a. Sulla probabilità $\mathbb{P}'$, oltre alle ipotesi precedenti, è ragionevole assumere che vedere il/la primo/a figlio/a e vedere il/la secondo/a figlio/a siano equiprobabili e indipendenti dal sesso dei/lle due figli/e. [Anche in questo caso $\mathbb{P}'$ è uniforme su Ω'.]
Chiamiamo con x_i il sesso dell'i-simo/a figlio/a e con I il/la figlio/a visto/a (aleatorio). La probabilità cercata è

$$\mathbb{P}'(\text{entrambe f} \mid x_I = \text{f}) = \frac{\mathbb{P}'(\text{entrambe f}, x_I = \text{f})}{\mathbb{P}'(x_I = \text{f})} = \frac{\mathbb{P}'(\text{entrambe f})}{\mathbb{P}'(x_I = \text{f})}.$$

Dal punto precedente $\mathbb{P}'(\text{entrambe f}) = 1/4$, mentre, per indipendenza,

$$\begin{aligned}\mathbb{P}'(x_I = \text{f}) &= \mathbb{P}'(x_1 = \text{f}, I = 1) + \mathbb{P}'(x_2 = \text{f}, I = 2) \\ &= \mathbb{P}'(x_1 = \text{f})\mathbb{P}'(I = 1) + \mathbb{P}'(x_2 = \text{f})\mathbb{P}'(I = 2) = \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2}.\end{aligned}$$

Quindi $\mathbb{P}'(\text{entrambe f} \mid x_I = \text{f}) = 1/2$. Notare che

$$\mathbb{P}'(\text{entrambe f} \mid x_I = \text{f}) \neq \mathbb{P}(\text{entrambe f} \mid \text{almeno una f}).$$

Si veda per approfondimento l'esempio 1.40 del testo Berger, Caravenna, Dai Pra, Probabilità. Un primo corso attraverso modelli e applicazioni. Seconda edizione, Springer 2021. ■

Esercizio 3.15 Lanciamo due volte una moneta equilibrata. Dimostrare che gli eventi "primo lancio testa" e "lanci concordi" sono indipendenti. ■

Soluzione Nel caso di due lanci di moneta equilibrata, la probabilità $\mathbb{P}$ su $\Omega = \{T, C\}^2$ è uniforme. L'evento "primo lancio testa" ha probabilità $1/2$, mentre

$$\mathbb{P}(\text{lanci concordi}) = \mathbb{P}\{(T,T),(C,C)\} = \frac{1}{2},$$
$$\mathbb{P}(\text{primo testa, lanci concordi}) = \mathbb{P}\{(T,T)\} = \frac{1}{4} = \mathbb{P}(\text{primo testa}) \cdot \mathbb{P}(\text{lanci concordi}).$$

Quindi i due eventi sono indipendenti. ■

Esercizio 3.16 — Forsyth 3.49. A un tempo casuale ispezioniamo i processi in esecuzione in un computer. Siano A, rispettivamente B gli eventi per cui il processo a, rispettivamente b, siano in esecuzione, sia N l'evento per cui il processo c ha un malfunzionamento. Supponiamo che $\mathbb{P}(A \cap N) = 0.07$, $\mathbb{P}(B \cap N) = 0.05$, $\mathbb{P}(A \cap B \cap N) = 0.04$, $\mathbb{P}(N) = 0.1$. Se vediamo che c ha un malfunzionamento, gli eventi A e B sono indipendenti? ■

Soluzione L'indipendenza dato N vuol dire che A e B sono indipendenti sotto $\mathbb{P}(\cdot \mid N)$, cioè

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

Verifichiamo dunque se tale condizione è vera oppure no. Abbiamo

$$\begin{aligned}
\mathbb{P}(A \mid N) &= \frac{\mathbb{P}(A \cap N)}{\mathbb{P}(N)} = \frac{0.07}{0.1} = 0.7, \\
\mathbb{P}(B \mid N) &= \frac{\mathbb{P}(B \cap N)}{\mathbb{P}(N)} = \frac{0.05}{0.1} = 0.5, \\
\mathbb{P}(A \cap B \mid N) &= \frac{\mathbb{P}(A \cap B \cap N)}{\mathbb{P}(N)} = \frac{0.04}{0.1} = 0.4 \neq 0.7 \cdot 0.5.
\end{aligned}$$

Quindi A e B non sono indipendenti dato N. ■

Esercizio 3.17 — Devore 106. Un metodo utilizzato per distinguere tra rocce granitiche (G) e basaltiche (B) consiste nell'esaminare una porzione dello spettro infrarosso dell'energia solare riflessa dalla superficie della roccia. Siano R_1, R_2 e R_3 le intensità spettrali misurate a tre diverse lunghezze d'onda. Tipicamente, per il granito si osserva $R_1 < R_2 < R_3$, mentre per il basalto $R_3 < R_1 < R_2$. Quando le misurazioni vengono effettuate a distanza (ad esempio, utilizzando un aereo), possono sorgere vari ordini dei valori R_i, sia per il granito che per il basalto. Voli effettuati su regioni di composizione nota hanno fornito le seguenti informazioni:

- Per il granito: $R_1 < R_2 < R_3$ si verifica nel 60% dei casi, $R_1 < R_3 < R_2$ nel 25% dei casi e $R_3 < R_1 < R_2$ nel restante 15% dei casi.
- Per il basalto: $R_1 < R_2 < R_3$ si verifica nel 10% dei casi, $R_1 < R_3 < R_2$ nel 20% dei casi e $R_3 < R_1 < R_2$ nel 70% dei casi.

Si supponga che per una roccia scelta casualmente in una certa regione si abbia $P(\text{granito}) = 0.25$ e quindi $P(\text{basalto}) = 0.75$.

a) Mostrare che $P(\text{granito} \mid R_1 < R_2 < R_3) > P(\text{basalto} \mid R_1 < R_2 < R_3)$. Se le misurazioni riportano $R_1 < R_2 < R_3$, come possiamo qlassificare la roccia?
b) Come possiamo qlassificare la roccia se le misurazioni riportano $R_1 < R_3 < R_2$? E se riportano $R_3 < R_1 < R_2$?
c) Utilizzando le regole di classificazione indicate nei punti (a) e (b), qual è la probabilità di un errore di classificazione scegliendo una roccia casuale in questa regione?

■

Esercizio 3.18 La distribuzione in Italia dei gruppi sanguigni è la seguente:

	0	A	B	AB
+	40	36	7.5	2.5
-	7	6	1.5	0.5

(ad esempio, il 40% degli individui è di tipo 0+). Si estrare un individuo a caso nella popolazione italiana.

a) Qual è la probabilità che l'individuo estratto abbia Rh+?
b) Qual è la probabilità che abbia Rh+, sapendo che ha gruppo AB?
c) Gli eventi "Rh+" e "gruppo sanguigno AB" sono indipendenti?

■

Soluzione Possiamo usare come spazio campionario $\Omega = \{Rh+, Rh-\} \times \{0, A, B, AB\}$, con la probabilità $\mathbb{P}$ dei singoli esiti data dalla tabella.

a) La probabilità cercata è

$$\mathbb{P}(\text{Rh+}) = \sum_{k \in \{0,A,B,AB\}} \mathbb{P}\{(Rh+, k)\} == 0.4 + 0.36 + 0.075 + 0.025 = 0.86.$$

Più in generale, le probabilità "marginali" dei gruppi per *Rh* o per lettera sono ottenute sommando le relative probabilità in riga/colonna: abbiamo (in percentuale)

	0	A	B	AB	tot
+	40	36	7.5	2.5	86
-	7	6	1.5	0.5	14
tot	47	42	9	3	100

b) La probabilità cercata è

$$\mathbb{P}(\text{Rh+} \mid AB) = \frac{\mathbb{P}\{(Rh+, AB)\}}{\mathbb{P}(AB)} = \frac{0.025}{0.03} = \frac{5}{6} = 0.8333.$$

c) Poiché $\mathbb{P}(Rh+ \mid AB) \neq \mathbb{P}(Rh+)$, gli eventi "Rh+" e "gruppo sanguigno AB" non sono indipendenti. Si noti comunque che le due probabilità non differiscono di molto, la dipendenza è "blanda".

■

Esercizio 3.19 Siano A, B e C tre eventi indipendenti, con $B \cap C$ non trascurabile. Mostrare che $P(A \mid B \cap C) = P(A)$. ■

Soluzione Per definizione di indipendenza, ogni intersezione ha probabilità pari al prodotto degli eventi coinvolti. Abbiamo quindi

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

■

Esercizio 3.20 Vengono lanciati $n \geq 1$ dadi a sei facce: qual è la probabilità che la somma dei punti di tutti i dadi sia divisibile per 7?

Suggerimento: i casi $n = 1, 2$ sono facili, per il caso n-esimo si condizioni sul risultato dei primi $n-1$ dadi, ottenendo una formula ricorsiva. ■

Esercizio 3.21 **(consigliato 28/02)** L'urna 1 contiene 5 biglie rosse e 5 blu. L'urna 2 contiene 8 biglie rosse e 12 blu. Da ciascuna urna viene estratta una biglia, quindi viene scelta a caso una delle due biglie estratte (ciascuna con probabilità $1/2$).

- Qual è la probabilità che la biglia scelta sia rossa?
- Se la biglia scelta è rossa, qual è la probabilità che anche l'altra biglia estratta sia rossa?
- Il colore della biglia estratta è indipendente dal colore dell'altra biglia?

■

Soluzione Uno spazio campionario adatto è $\Omega = \{b,r\} \times \{b,r\} \times \{1,2\}$, dove b,r denotano il colore delle biglie estratte dalla prima e dalla seconda urna, mentre $1,2$ denotano l'urna di provenienza della biglia scelta. Le informazioni del problema sono:

- $\mathbb{P}(\text{prima r}) = 5/10$;
- $\mathbb{P}(\text{seconda r} = 8/20$;
- $\mathbb{P}(\text{urna } 1\} = 1/2$;
- gli eventi "prima r", "seconda r" e "urna 1" sono indipendenti (e analogamente sono indipendenti i tre eventi "prima r", "seconda r" e "urna 2", i tre eventi "prima r", "seconda b" e "urna 1" e così via).

a) L'evento "rossa" è l'unione disgiunta di "prima r, urna 1" e "seconda r, urna 2". Quindi, per additività e indipendenza, abbiamo

$$\begin{aligned}\mathbb{P}(\text{rossa}) &= \mathbb{P}(\text{prima rossa, urna } 1) + \mathbb{P}(\text{seconda rossa, urna } 2)\\ &= \mathbb{P}(\text{prima rossa})\mathbb{P}(\text{urna } 1) + \mathbb{P}(\text{seconda rossa})\mathbb{P}(\text{urna } 2) = \frac{9}{20}.\end{aligned}$$

b) Usando la definizione di probabilità condizionata e l'indipendenza delle due estrazioni, abbiamo

$$\begin{aligned}\mathbb{P}(\text{altra rossa} \mid \text{rossa}) &= \frac{\mathbb{P}(\text{entrambe rosse})}{\mathbb{P}(\text{rossa})}\\ &= \frac{\mathbb{P}(\text{prima rossa})\mathbb{P}(\text{seconda rossa})}{\mathbb{P}(\text{rossa})} = \frac{5/10 \cdot 8/20}{9/20} = \frac{4}{9}.\end{aligned}$$

c) In modo analogo al punto (a), o semplicemente per simmetria tra la estratta e la non estratta, otteniamo che la probabilità che la non estratta sia rossa è $9/20$. Questa probabilità non è la stessa di $\mathbb{P}(\text{altra rossa} \mid \text{rossa})$, dunque il colore della estratta non è indipendente da quello della non estratta.
Questo può sembrare paradossale, ma ha anche una ragione intuitiva: se estratto una rossa, è più probabile che questa venga dalla prima urna (che ha più rosse in proporzione), cioè che l'altra biglia venga dalla seconda urna, e quindi è più probabile che l'altra biglia sia blu.
Analoghi paradossi sono molto diffusi, si veda ad esempio il problema di Monty Hall e il paradosso delle tre carte.

■

Esercizio 3.22 — Difficile. Ci sono n candidati per un posto di lavoro. Si ammetta che i candidati possano essere ordinati dal migliore al peggiore. L'esaminatrice incontra sequenzialmente n candidati, uno dopo l'altro, in ordine casuale. L'esaminatrice deve scegliere se accettare o rifiutare ogni candidato alla fine del colloquio corrispondente, senza possibilità di tornare

indietro e cambiare la propria decisione. La strategia utilizzata (chiamata strategia k) è la seguente: si intervistano e rifiutano automaticamente k candidati e dopodiché si assume il primo candidato che è "meglio" di tutti i precedenti (inclusi i primi k). Se non c'è un tale candidato, viene assunto l'ultimo candidato. Il parametro k è scelto e fissato prima dell'inizio dei colloqui.

- Per n fisso, calcolare la probabilità che la strategia k porti all'assunzione del miglior candidato per $k \in \{1, \dots, n\}$.
- Qual è (in funzione di n) il valore k^* che massimizza tale probabilità?
- Trovare il limite k^*/n per $n \to \infty$.

■

Esercizio 3.23 — Difficile. Si elegge il rappresentante degli studenti al corso di laurea di Scienze Scientifiche. Il candidato Avogadro vince con a voti, mentre il candidato Bohr viene sconfitto ottenendo $b < a$ voti. Si mostri che la probabilità che durante lo spoglio dei voti Avogadro sia stato in testa a Bohr in ogni momento è data da $\frac{a-b}{a+b}$.

Suggerimento: Condizionare sull'ultimo voto scrutinato e procedere per ricorsione. ■

NB Saper fare:

- Individuare le informazioni di base sulla probabilità a partire dalla descrizione del problema in esame, in particolare per esperimenti su più livelli.
- Calcolare la probabilità di un evento a partire dai suoi esiti o sottocasi.
- Calcolare la probabilità di un evento tramite le proprietà di base (probabilità unione, intersezione, complementare relativo, ...), in particolare tramite complementare.
- Verificare la proprietà di indipendenza con la definizione.
- Riconoscere l'indipendenza (dove presente) a partire dalla descrizione del problema in esame.
- Calcolare la probabilità condizionata con la definizione.
- Calcolare la probabilità non condizionata a partire dalle probabilità condizionate, usando le regole di somma (partizione) e prodotto (catena), ad esempio tramite gli alberi.
- Invertire il condizionamento tramite la formula di Bayes.
- Riconoscere quando usare il modello uniforme, in particolare per estrazioni con ordine (con o senza rimpiazzo) e senza ordine senza rimpiazzo.
- Usare alcune formule di combinatoria (disposizioni semplici e con ripetizione, combinazioni semplici).
- Individuare e usare le simmetrie quando possibile.

4. Variabili aleatorie

Esercizio 4.1 **(consigliato (a) 28/02, consigliato (b) 14/03)**

a) Si introducano uno spazio di probabilità $(\Omega, \mathbb{P})$ e una variabile aleatoria $X : \Omega \to \mathbb{R}$ che descrivono "il massimo tra i valori di due lanci di un dado a sei facce equilibrato". Si calcoli la densità discreta di X.

b) Si calcolino poi valore atteso e varianza di X ($\mathrm{Var}(X) = E[X^2] - E[X]^2$).

■

Soluzione a) Per descrivere l'esperimento due lanci di un dado equilibrato, prendiamo $\Omega = \{1,2,\dots 6\}^2$ e $\mathbb{P}$ uniforme su Ω. La v.a. X che descrive il massimo dei due lanci è

$$X(\omega_1, \omega_2) = \max\{\omega_1, \omega_2\},$$

e ha valori (modalità) in $\{1,2,\dots 6\}$.
Calcoliamo la sua densità discreta, cioè la funzione $\{1,2,\dots 6\} \ni k \mapsto \mathbb{P}(X = k)$. L'evento $\{X = k\}$ corrisponde alle coppie di lanci il cui massimo è esattamente k, che sono: (k, ω_2) con $\omega_2 \leq k$ (k coppie) e (ω_1, k) con $\omega_1 < k$ ($k-1$ coppie), quindi in totale $2k-1$ coppie. Quindi, per $k = 1,2,\dots 6$,

$$\mathbb{P}(X = k) = \frac{2k-1}{36}.$$

b) Il valore atteso vale, usando la definizione,

$$E[X] = \sum_{k=1}^{6} k \cdot \frac{2k-1}{36} = \frac{161}{36} = 4.472.$$

Per la varianza, calcoliamo prima $E[X^2]$:

$$E[X^2] = \sum_{k=1}^{6} k^2 \cdot \frac{2k-1}{36} = 21.972,$$

da cui $\mathrm{Var}(X) = E[X^2] - E[X]^2 = 1.973$.

■

Esercizio 4.2 **(consigliato 21/03)** Un test a risposta multipla prevede 10 domande con 4 possibili risposte per ogni domanda. Il punteggio viene assegnato con le seguenti regole: $+1$ punto per ogni risposta esatta; -0.25 punti per ogni risposta sbagliata o non data. Supponendo di rispondere a tutte le domande del test in maniera casuale, con uguale probabilità per ogni risposta, si determini:

a) la probabilità di ottenere almeno 8 risposte esatte;
b) il valore atteso e la varianza della variabile aleatoria che descrive il punteggio ottenuto al test.

■

Soluzione a) Siamo in presenza di uno schema di Bernoulli, con singola prova la risposta alla singola domanda, 10 prove ripetute, successo risposta esatta con probabilità $1/4$. Il numero di risposte esatte X ha distribuzione $B(10,1/4)$. Quindi la probabilità cercata è

$$\begin{aligned}\mathbb{P}(X \geq 8) &= \mathbb{P}(X=8) + \mathbb{P}(X=9) + \mathbb{P}(X=10) \\ &= \binom{10}{8}\left(\frac{1}{4}\right)^8\left(\frac{3}{4}\right)^2 + \binom{10}{9}\left(\frac{1}{4}\right)^9\left(\frac{3}{4}\right)^1 + \binom{10}{10}\left(\frac{1}{4}\right)^{10}\left(\frac{3}{4}\right)^0 \\ &= \frac{10\cdot 9}{2}\frac{9}{4^{10}} + 10\frac{3}{4^{10}} + \frac{1}{4^{10}} = \frac{109}{4^9} = 4.16\cdot 10^{-4}.\end{aligned}$$

b) Per le regole date, se $X_i = 1$ per l'i-sima risposta corretta, 0 altrimenti, il punteggio Y_i per ciascuna risposta è $Y_i = (5X_i - 1)/4$. Poiché il numero di risposte esatte è $X = X_1 + \ldots X_{10}$, allora $Y = (5X - 10)/4$. In particolare, per linearità del valore atteso e scaling della varianza, abbiamo

$$E[Y] = \frac{5E[X]-10}{4}, \quad \mathrm{Var}(Y) = \frac{25}{16}\mathrm{Var}(X).$$

Poiché X è binomiale, sappiamo che $E[X] = 10\cdot 1/4 = 5/2$ e $\mathrm{Var}(X) = 10\cdot 1/4\cdot 3/4 = 15/8$, quindi otteniamo $E[Y] = 5/8$ e $\mathrm{Var}(X) = 375/128$.

■

Esercizio 4.3 **(consigliato 28/02)** In un'urna ci sono 6 biglie blu e 4 rosse. Vengono effettuate 4 estrazioni casuali con reinserimento (cioè, ad ogni estrazione, la biglia estratta viene rimessa nell'urna). Qual è la probabilità che venga estratta al massimo una biglia rossa? ■

Soluzione Poiché le estrazioni sono con reinserimento, sono prove ripetute; nella singola estrazione, la probabilità di estrarre una rossa è $4/10 = 2/5$. Quindi la v.a. X che conta il

numero di rosse estratte ha distribuzione $B(4,2/5)$. La probabilità cercata è

$$\begin{aligned}\mathbb{P}(X \leq 1) &= \mathbb{P}(X=0)+\mathbb{P}(X=1)\\ &= \binom{4}{0}\left(\frac{2}{5}\right)^0\left(\frac{3}{5}\right)^4+\binom{4}{1}\left(\frac{2}{5}\right)^1\left(\frac{3}{5}\right)^3\\ &= \frac{81}{5^4}+4\frac{54}{5^4}=0.4752.\end{aligned}$$

■

Esercizio 4.4 Sappiamo dalla letteratura scientifica che un certo farmaco è efficace contro una data malattia nell'80% dei pazienti. Presi 10 pazienti a caso, calcolare:

a) la probabilità che il farmaco sia efficace su almeno 9 di loro;
b) il valore atteso e la deviazione standard del numero di pazienti (tra i 10 scelti) su cui il farmaco è efficace.

Le risposte ai punti precedenti cambiano se sappiamo che i pazienti scelti sono parenti tra loro?

■

Soluzione a) Siamo in presenza di uno schema di Bernoulli con 10 prove ripetute (le somministrazioni del farmaco sui 10 pazienti), ciascuna con proabbilità 0.8 di successo (l'efficacia del farmaco). La v.a. $X =$ numero di pazienti su cui il farmaco è efficace ha distribuzione $B(10,0.8)$. La probabilità cercata è

$$\begin{aligned}\mathbb{P}(X \geq 9) &= \mathbb{P}(X=9)+\mathbb{P}(X=10)\\ &\quad 10\cdot 0.8^9\cdot 0.2+0.8^{10}=0.3758.\end{aligned}$$

La risposta cambia se i pazienti sono parenti: in questo caso, le prove non sono più indipendenti (o perlomeno possiamo sospettare che non lo siano): se l'azione del farmaco è influenzata da fattori genetici, una maggiore o minore efficacia su un paziente è più probabile che si ripeta sui suoi parenti.

b) Poiché X è binomiale, abbiamo $E[X]=10\cdot 0.8=8$ e $\mathrm{Var}(X)=10\cdot 0.8\cdot 0.2=1.6$.
Se i pazienti sono parenti, il valore atteso non cambia: possiamo sempre scrivere $X = X_1+\ldots X_{10}$, dove $X_i=1$ se il farmaco è efficace sull'i-simo paziente, 0 altrimenti, e le X_i sono Bernoulli di parametro 0.8, sebbene non più ndipendenti, quindi per linearità del valore atteso si ottiene la stessa risposta. Invece (come vedremo) la varianza può cambiare.

■

Esercizio 4.5 Se X ha legge $B(n,p)$, qual è la legge di $n-X$? ■

Soluzione Una v.a. $X \sim B(n,p)$ conta il numero di successi in n prove ripetute, con p probabilità di successo. Quindi $n-X$ conta il numero di insuccessi nelle n prove, e ciascun insuccesso ha probabilità $1-p$. Quindi, scambiando successo e insuccesso, $n-X$ ha distribuzione $B(n,1-p)$.

■

Esercizio 4.6 — Ross 5.2. Un canale di comunicazione trasmette le cifre 0 e 1. Tuttavia, a causa di interferenze, la cifra trasmessa viene ricevuta in modo errato con una probabilità pari a 0.2. Supponiamo di voler trasmettere un messaggio importante composto da una cifra binaria. Per ridurre la probabilità di errore, trasmettiamo 00000 al posto di 0 e 11111 al posto di 1.

a) Se il ricevitore del messaggio utilizza una decodifica basata sulla "maggioranza", qual è la probabilità che il messaggio venga decodificato in modo corretto? (Per decodifica a maggioranza si intende che il messaggio viene decodificato come "0" se ci sono almeno tre zeri nel messaggio ricevuto, e come "1" altrimenti.)
b) Quali ipotesi di indipendenza stiamo assumendo?

■

Soluzione a) Siamo in presenza di uno schema di Bernoulli, con 5 prove ripetute (l'invio delle 5 prove) e probabilità di successo (cifra corretta, in questo caso 0) 0.8. Detta X la v.a. che conta il numero di cifre corrette, X ha distribuzione $B(5,0.8)$ e la probabilità richiesta è

$$\begin{aligned}\mathbb{P}(X \geq 3) &= \mathbb{P}(X=3)+\mathbb{P}(X=4)+\mathbb{P}(X=5)\\ &= \frac{5\cdot 4}{2}0.8^3\cdot 0.2^2 + 5\cdot 0.8^4\cdot 0.2 + 0.8^5 = 0.94208.\end{aligned}$$

b) Nel punto (a), Abbiamo assunto che la ricezione corretta di ciascuna cifra fosse indipendente dalle altre, per usare lo schema di Bernoulli.

■

Esercizio 4.7 — Devore 2.94. Un trasmettitore invia un messaggio utilizzando un codice binario, ossia una sequenza di 0 e 1. Ogni bit trasmesso (0 o 1) deve passare attraverso tre ripetitori per raggiungere il ricevitore. A ogni ripetitore, la probabilità che il bit ricevuto sia diverso dal bit inviato (inversione) è 0.20. Si assume che i ripetitori operino in modo indipendente l'uno dall'altro:

Trasmettitore $\to$ Ripetitore 1 $\to$ Ripetitore 2 $\to$ Ripetitore 3 $\to$ Ricevitore.

a) Se il trasmettitore invia un 1, qual è la probabilità che un 1 venga inviato da tutti e tre i ripetitori?
b) Se il trasmettitore invia un 1, qual è la probabilità che un 1 venga ricevuto dal ricevitore?
c) Supponiamo che il 70% di tutti i bit inviati dal trasmettitore siano 1. Se il ricevitore riceve un 1, qual è la probabilità che il trasmettitore abbia inviato un 1?

■

Soluzione a) Per $i = 1,2,3$, sia $A_i = \{$no errore all'i-simo trasmettitore. Gli A_i sono eventi indipendenti e di probabilità 0.8. Quindi la probabilità che tutti trasmettano 1 correttamente è

$$\mathbb{P}(A_1 \cap A_2 \cap A_3) = 0.8^3 = 0.512.$$

b) Possiamo considerare l'esperimento come uno schema di Bernoulli con 3 prove, in cui il suggesso è l'errore all'i-simo trasmettitore (A_i^c) di probabilità 0.2. Sia $X =$ numero di errori complessivo, $X \sim B(3,0.8)$. Ora notiamo che il messaggio finale è 1 (lo stesso

di partenza) se e solo se il numero di errori commesso è pari, quindi 0 o 2. Dunque la probabililità cercata è

$$\mathbb{P}(X \text{ pari}) = \mathbb{P}(X=0) + \mathbb{P}(X=2) = 0.8^3 + 3 \cdot 0.8 \cdot 0.2^2 = 0.608.$$

■

Esercizio 4.8 **(consigliato 07/03)** Un'urna contiene 10 biglie, di cui 5 rosse e 5 blu. Effettuiamo 4 estrazioni con reimmissione.

a) Qual è la probabilità che almeno 3 biglie su 4 siano rosse?

Una seconda urna contiene 8 biglie rosse e 2 blu. Scegliamo un'urna a caso ed estraiamo da questa 4 biglie con reimmissione.

b) Qual è la probabilità che almeno 3 siano rosse?

c) Se almeno 3 biglie estratte sono rosse, qual è la probabilità che abbiamo scelto la prima urna?

d) Gli eventi "rossa alla prima estrazione" e "rossa alla seconda estrazione" sono indipendenti?

■

Soluzione a) Poiché le estrazioni sono con reimmissione (quindi sono prove ripetute), siamo in presenza di uno schema di Bernoulli di 4 prove con probabilità di successo, cioè rossa, $1/2$. Il numero X di biglie rosse estratte ha quindi distribuzione $B(4,1/2)$ e la probabilità cercata è

$$\mathbb{P}((X \geq 3) = \mathbb{P}(X=3) + \mathbb{P}(X=4) = \binom{4}{3} \cdot 0.5^4 + \binom{4}{4} \cdot 0.5^4 = \frac{5}{16} = 0.3125.$$

b) Anche se non è richiesto, scriviamo uno spazio campionario: $\Omega = \{1,2\} \times \{r,b\}^4$, dove $1,2$ indica l'urna scelta e r,b il colore della i-sima biglia estratta. Sia X la v.a. che conta il numero di rosse estratte. Le informazioni del problema sono:

- $\mathbb{P}(\text{urna } 1) = \mathbb{P}(\text{urna } 2) = 1/2$;
- se scegliamo l'urna 1, cioè sotto la probabilità $\mathbb{P}(\cdot \mid \text{urna } 1)$, abbiamo uno schema di Bernoulli con probabilità di successo (rossa) $1/2$, in particolare $X \sim B(4,1/2)$ sotto $\mathbb{P}(\cdot \mid \text{urna } 1)$;
- se scegliamo l'urna 2, cioè sotto la probabilità $\mathbb{P}(\cdot \mid \text{urna } 2)$, abbiamo uno schema di Bernoulli con probabilità di successo (rossa) $4/5$, in particolare $X \sim B(4,4/5)$ sotto $\mathbb{P}(\cdot \mid \text{urna } 2)$.

Ne segue che

$$\mathbb{P}(X \geq 3 \mid \text{urna } 1) = 0.3125,$$
$$\mathbb{P}(X \geq 3 \mid \text{urna } 2) = \binom{4}{3} \cdot 0.8^3 \cdot 0.2 + \binom{4}{4} \cdot 0.8^4 = 0.8192.$$

Per la formula di partizione, la probabilità cercata è

$$\mathbb{P}(X \geq 3) = \mathbb{P}(X \geq 3 \mid \text{urna } 1)\mathbb{P}(\text{urna } 1) + \mathbb{P}(X \geq 3 \mid \text{urna } 2)\mathbb{P}(\text{urna } 2) = 0.489.$$

c) Per Bayes, la probabilità cercata è

$$\mathbb{P}(\text{urna 1} \mid X \geq 3) = \frac{\mathbb{P}(X \geq 3 \mid \text{urna 1})\mathbb{P}(\text{urna 1})}{\mathbb{P}(X \geq 3)} = 0.32.$$

d) Gli eventi "rossa alla prima estrazione" e "rossa alla seconda estrazione" non sono indipendenti: l'idea è che, se viene estratta una rossa alla prima estrazione, allora è più probabile che abbiamo scelto l'urna 2 (che contiene più rosse), quindi è più probabile che venga estratta una rossa anche alla seconda estrazione. Rigorosamente, per partizione abbiamo

$$\begin{aligned}\mathbb{P}(\text{prima rossa}) &= \mathbb{P}(\text{prima rossa} \mid \text{urna 1})\mathbb{P}(\text{urna 1}) \\ &\quad + \mathbb{P}(\text{prima rossa} \mid \text{urna 2})\mathbb{P}(\text{urna 2}) \\ &= 0.5 \cdot 0.5 + 0.8 \cdot 0.5 = 0.65,\end{aligned}$$

e per simmetria $\mathbb{P}(\text{seconda rossa}) = 0.65$, mentre

$$\begin{aligned}\mathbb{P}(\text{prima e seconda rossa}) &= \mathbb{P}(\text{prima e seconda rossa} \mid \text{urna 1})\mathbb{P}(\text{urna 1}) \\ &\quad + \mathbb{P}(\text{prima e seconda rossa} \mid \text{urna 2})\mathbb{P}(\text{urna 2}) \\ &= 0.5^2 \cdot 0.5 + 0.8^2 \cdot 0.5 = 0.445 \neq 0.65^2,\end{aligned}$$

dove abbiamo usato che "prima rossa" e "seconda rossa" sono indipendenti sapendo "urna 1" (e analogamente sapendo "urna 2").

■

NB L'esercizio 4.8 è un tipico esempio di esperimento su più livelli (o formato da differenti sottoesperimenti): al livello 1, ho una scelta tra le due urne, al livello 2 ho uno schema di Bernoulli. Individuare correttamente la struttura dell'esperimento è fondamentale per risolvere gli esercizi.

NB Come l'esercizio 4.8 mostra, l'indipendenza dipende dalla probabilità usata, e quindi dall'informazione a disposizione: l'indipendenza condizionata all'urna non implica l'indipendenza non condizionata.

Esercizio 4.9 Un sacchetto contiene 10 monete, di cui 8 "oneste" (ossia una faccia è Testa e una faccia Croce, e nel lancio della moneta le due facce sono equiprobabili) e 2 con entrambe le facce uguali a Testa. L'esperimento consiste nell'estrarre una moneta dal sacchetto e lanciarla: se esce Testa l'esperimento è concluso, altrimenti si reinserisce la moneta e si ripete l'operazione.

a) Determinare la probabilità che durante l'esperimento vengano eseguiti k lanci, con $k \in \mathbb{N}$.
b) Sapendo che l'esito Testa non si è avuto nei primi 10 lanci, determinare la probabilità di successo in meno di 13 lanci.
c) Supponiamo che al primo lancio sia uscita Testa, qual è la probabilità che sia stata estratta dal sacchetto una moneta onesta?

■

Soluzione a) Anche se non è richiesto, scriviamo uno spazio campionario: $\Omega = \{\{o,t\} \times \{0,1\}\}^{\mathbb{N}_+} = \{(\omega_1, \omega_2, \ldots) \mid \omega_i \in \{o,t\} \times \{0,1\} \text{ per ogni } i\}$, dove o,t indicano il tipo di

moneta (onesta o truccata) e $0,1$ indicano l'esito del lancio ($0 =$ croce, $1 =$ testa) (ammettiamo qui che l'esperimento prosegua all'infinito, anche se poi ci interessa solo il primo lancio testa). Le informazioni del problema sono:

- le prove "estrazione moneta e lancio" sono ripetute, quindi indipendenti;
- per ciascuna prova i, $\mathbb{P}(i\text{-sima onesta}\} = 8/10$ (qui, con un piccolo abuso di notazione, indichiamo con "i-sima testa" l'evento "testa all'i-sima prova");
- per ciascuna prova i, $\mathbb{P}(i\text{-sima testa} \mid i\text{-sima onesta}) = 1/2$ e $\mathbb{P}(i\text{-sima testa} \mid i\text{-sima truccata}) = 1$.

In ciascuna prova, la probabilità che esca testa (successo) è

$$\begin{aligned}\mathbb{P}(i\text{-sima testa}) &= \mathbb{P}(i\text{-sima testa} \mid i\text{-sima onesta})\mathbb{P}(i\text{-sima onesta}) \\ &\quad + \mathbb{P}(i\text{-sima testa} \mid i\text{-sima truccata})\mathbb{P}(i\text{-sima truccata}) \\ &= 0.5 \cdot 0.8 + 0.2 \cdot 1 = 0.6.\end{aligned}$$

Poiché le prove sono ripetute, la v.a. T che conta l'istante del primo successo è geometrica di parametro 0.6. La probabilità cercata è quindi

$$\mathbb{P}(T = k) = 0.6 \cdot 0.4^{k-1}, \quad k \in \mathbb{N}_+.$$

b) Per la proprietà di assenza di memoria della v.a. geometrica, la probabilità cercata è

$$\mathbb{P}(T < 13 \mid T > 10) = \mathbb{P}(T < 3) = \mathbb{P}(T = 1) + \mathbb{P}(T = 2) = 0.84.$$

c) Per Bayes, abbiamo (ometto "prima" dalla notazione)

$$\mathbb{P}(\text{onesta} \mid \text{testa}) = \frac{\mathbb{P}(\text{testa} \mid \text{onesta})\mathbb{P}(\text{onesta})}{\mathbb{P}(\text{testa})} = 2/3 = 0.667.$$

■

Esercizio 4.10 **(consigliato (a) 07/03, consigliato (b) 21/03)** Sia $c \in \mathbb{R}$ un parametro e sia $p : \mathbb{N} \to \mathbb{R}$ data da $p(k) = c2^{-k}$.

a) Determinare l'unico valore di c per cui p sia una funzione di massa.

Prendiamo ora c pari a tale valore e sia X una v.a. discreta con funzione di massa p.

b) Dire per quali valori di $\alpha \in \mathbb{R}$ la v.a. $e^{\alpha X}$ ammette momento primo.

■

Soluzione a) p è funzione di massa se e solo se $p(k) \geq 0$ per ogni k e $\sum_k p(k) = 1$. La prima condizione è verificata se e solo se $c \geq 0$. Poiché

$$\sum_{k=0}^{\infty} c2^{-k} = c\frac{1}{1/2} = 2c,$$

la seconda condizione è verificata se e solo se $c = 1/2$. Quindi l'unico valore ammissibile è $c = 1/2$.

b) La v.a. $e^{\alpha X}$ è sempre non-negativa, quindi esiste il valore atteso in $[0, +\infty]$ e vale

(scrivendo $2^{-k} = e^{-k\log 2}$)

$$\mathbb{E}[e^{\alpha X}] = \sum_{k=0}^{\infty} e^{\alpha k} p(k) = \frac{1}{2}\sum_{k=0}^{\infty} e^{\alpha k} 2^{-k} = \frac{1}{2}\sum_{k=0}^{\infty} e^{(\alpha-\log 2)k}$$

La serie sopra è una serie geometrica di ragione $r = e^{\alpha-\log 2}$, che quindi converge se e solo se $r < 1$, cioè se e solo se $\alpha < \log 2$, e in questo caso vale $1/(1-e^{\alpha-\log 2})$. Quindi $e^{\alpha X}$ ammette valore atteso se e solo se $\alpha < \log 2$, e in questo caso

$$\mathbb{E}[e^{\alpha X}] = \frac{1}{2-e^{\alpha}}.$$

■

Esercizio 4.11 I 42 studenti di un corso di calligrafia, dei quali 13 sono mancini, sono divisi per le esercitazioni in due gruppi di eguale numero per sorteggio. Si indichino rispettivamente con X e Y le v.a. il cui valore è il numero di mancini presenti nel primo e nel secondo gruppo.

a) Calcolare la funzione di massa (o densità discreta) della v.a. X.
b) Provare che le v.a. X e Y sono equidistribuite, ma non sono indipendenti.
c) È possibile determinare i valori attesi $E[X]$ e $E[Y]$ senza usare le leggi di X e Y ?

■

Soluzione a) La v.a. X può assumere valori $0,1,2,\dots 13$, dunque la sua funzione di massa è $\mathbb{P}(X=k)$ per $0,1,2,\dots 13$. Poiché si tratta di un'estrazione di 21 elementi su 42 senza ordine senza rimpiazzo, la probabilità è uniforme sulle combinazioni semplici di 21 oggetti su 42. L'evento $\{X=k\} = \{\text{esattamente } k \text{ mancini}\}$ ha cardinalità

$$\begin{aligned}\#\{X=k\} &= \#\text{scelte dei } k \text{ mancini} \cdot \#\text{scelte dei } 21-k \text{ non mancini} \\ &= \binom{13}{k}\cdot\binom{29}{13-k},\end{aligned}$$

e dunque la probabilità cercata è

$$\mathbb{P}(X=k) = \frac{\binom{13}{k}\cdot\binom{29}{13-k}}{\binom{42}{21}}.$$

Si vedano anche gli esercizi 3.4 e 3.6.

b) La v.a. Y rappresenta ancora il numero di mancini in un'estrazione di 21 persone su 42 (senza rimpiazzo), quindi Y ha la stessa distribuzione di X. Tuttavia X e Y non sono indipendenti, poiché $X+Y=13$ e quindi informazioni su X ci danno informazioni su Y: ad esempio

$$\mathbb{P}(X=0, Y=0) = 0 \neq \mathbb{P}(X=0)\cdot\mathbb{P}(Y=0).$$

c) Possiamo usare un ragionamento di simmetria per calcolare i valori attesi. Anzitutto notiamo che X e Y hanno lo stesso valore atteso, perché hanno la stessa legge (il valore atteso dipende solo dalla legge). Inoltre $X+Y=13$. Quindi, per linearità del valore atteso, $E[X]+\mathbb{E}[Y]=13$. Ne segue che $\mathbb{E}[X] = 13/2 = 6.5$.

■

Esercizio 4.12 — Ross 5.4. **(consigliato 07/03)** Supponiamo che una particolare caratteristica (come il colore degli occhi o l'essere mancino) di una persona venga classificata sulla base di una coppia di geni, e supponiamo che d rappresenti un gene dominante e r un gene recessivo. Pertanto:

- Una persona con geni dd è a dominanza pura,
- Una persona con geni rr è a recessività pura,
- Una persona con geni rd è ibrida.

La dominanza pura (dd) e l'ibrido (rd) appaiono uguali dal punto di vista del fenotipo. I figli ricevono un gene da ciascun genitore. Se, rispetto a una particolare caratteristica, due genitori ibridi (rd) hanno un totale di 4 figli, qual è la probabilità che 3 dei 4 figli abbiano il fenotipo dominante? ■

Soluzione Anche se non richiesto, scriviamo uno spazio campionario: $\Omega = \{ab \mid a,b \in \{d,r\}\}^4$, dove a,b indicano rispettivamente i geni ereditati dal genitore 1 e dal genitore 2. Le informazioni del problema sono

- i geni dei singoli figli sono indipendenti (in quanto prove ripetute);
- per ciascun figlio, ciascun genitore passa al figlio il genere d o r con probabilità $1/2$, indipendentemente dall'altro genitore; in particolare, ciascuna coppia dd, dr, rd e rr ha probabilità $1/4$.

La v.a. X che conta il numero di figli con fenotipo dominante (successo), cioè con dd, dr o rd, ha distribuzione $B(4,3/4)$. La probabilità cercata è

$$\mathbb{P}(X=3) = \binom{4}{3} \cdot \left(\frac{3}{4}\right)^3 \cdot \frac{1}{4} = \frac{27}{64}.$$

■

Esercizio 4.13 — Ross 5.14. Circa 80.000 matrimoni si sono celebrati nello stato di New York in un dato anno. Stimare la probabilità che per almeno una di queste coppie:

a) Entrambi i partner siano nati il 30 aprile.
b) Entrambi i partner abbiano festeggiato il compleanno lo stesso giorno dell'anno.

Indicare le ipotesi assunte. ■

Soluzione a) Per semplicità assumiamo l'anno non bisestile (365 giorni). Possiamo ragionevolamente assumere che la data di nascita di una persona sia indipendente dalle date di nascita delle altre persone (sapere che una persona è nata in un dato giorno non ci dà informazioni sulle date delle altre persone); in particolare, le date di nascita di una coppia sono indipendenti delle date di nascita delle altre coppie. Inoltre, possiamo assumere che ogni giorno sia equiprobabile. Quindi, per due dati individui, la probabilità che entrambi siano nati il 30 aprile (successo) è

$$\frac{1}{365} \cdot \frac{1}{365} = \frac{1}{365^2}.$$

Siamo quindi in presenza di uno schema di Bernoulli con $n = 80000$ prove e probabilità di successo $p = 1/365^2$. Poiché p è molto piccola, n è molto grande e $np = 0.6$, possiamo usare l'approssimazione di Poisson: il numero X di coppie con entrambi nati il 30 aprile

ha distribuzione (circa) di Poisson di parametro 0.6. In particolare, la probabilità cercata è

$$\mathbb{P}(X \geq 1) = 1 - \mathbb{P}(X = 0) = 1 - e^{-0.6} = 0.451.$$

b) ...

■

Esercizio 4.14 — Devore 3.105. L'acquirente di un'unità di generazione di energia richiede c avviamenti consecutivi riusciti prima che l'unità venga accettata. Supponiamo che gli esiti dei singoli avviamenti siano indipendenti tra loro. Sia p la probabilità che un particolare avviamento abbia successo. La variabile aleatoria di interesse è X, ossia il numero totale di avviamenti richiesti affinché l'unità venga accettata quando $c = 2$.

a) Determinare la funzione di massa di X per il caso $c = 2$.
b) Se $x = 5$ e $p = 0.9$, calcolare $P(X \leq 8)$.

Suggerimento: Per il calcolo della $p(x)$, esprimerla in modo "ricorsivo" in termini della pmf valutata per valori più piccoli, ossia $x-3$, $x-4$, ecc. ■

Esercizio 4.15 Un'auto va in panne in un punto a caso di una strada di 60km. Non avendo altre informazioni, con quale probabilità l'auto si è fermata tra il km 20 e il km 40? ■

Soluzione Non avendo altre informazioni, possiamo assumere che il punto X (in km) dove l'auto va in panne abbia distribuzione uniforme sull'intervallo $(0,60)$. La probabilità cercata è

$$\mathbb{P}(X \in [20,40]) = \frac{40-20}{60-0} = \frac{1}{3}.$$

■

Esercizio 4.16 (consigliato 07/03) Il tempo (in minuti) in cui un cliente viene servito a uno sportello ha distribuzione esponenziale di parametro 0.5.

a) Qual è la probabilità che un cliente venga servito entro 2 minuti?
b) Se il cliente non è stato servito nei precedenti 3 minuti, qual è la probabilità che venga servito entro i successivi 2 minuti?

Supponiamo che i tempi per servire clienti differenti siano indipendenti.

c) Qual è la probabilità che, su 5 clienti, almeno 4 vengano serviti entro 2 minuti?

■

Soluzione a) Sia T il tempo (in minuti) in cui un cliente viene servito, $T \sim \text{Exp}(\lambda = 0.5)$ e ha quindi densità $f(x) = \lambda e^{-\lambda x} 1_{(0,\infty)}(x)$. La probabilità cercata è

$$\mathbb{P}(T \leq 2) = \int_{-\infty}^{2} f(x)\,dx = \int_{0}^{2} \lambda e^{-\lambda x}\,dx = -e^{-\lambda x}\,|_{x=0}^{2} = 1 - e^{-1} = 0.632.$$

b) Per la proprietà di assenza di memoria, la probabilità cercata è

$$\mathbb{P}(T \leq 5 \mid T > 3) = \mathbb{P}(T \leq 2) = 0.632.$$

c) Siamo in presenza di uno schema di Bernoulli, con 5 prove ripetute "servizio del singolo cliente", successo "cliente servito entro 2 minuti" di probabilità $p = 0.632$ per il punto (a). Detta X la v.a. che conta il numero di clienti serviti entro 2 minuti, X ha distribuzione $B(5,p)$ e la probabilità cercata è

$$\mathbb{P}(X \geq 4) = \binom{5}{4} p^4 (1-p) + \binom{5}{5} p^5 = 0.3944.$$

■

Esercizio 4.17 Sia $p \in [0,1]$, e si consideri la funzione

$$f(x) = \begin{cases} p, & 0 < x \leq 1 \\ (1-p), & 1 < x \leq 2 \\ 0, & x \leq 0, x > 2. \end{cases}$$

a) Verificare che la funzione sopra scritta è una densità di probabilità
b) Scrivere l'espressione della funzione di ripartizione per una v.a. X che abbia quella densità.
c) Scrivere la formula per il β-quantile (con $0 < \beta < 1$) per una v.a. X che abbia quella densità.
d) Calcolare i momenti primo e secondo di X e la varianza di X. Esaminare quando, al variare di p in $[0,1]$, la varianza è massima e quando è minima.

■

Soluzione a) La funzione f è densità di probabilità se e solo se $f(x) \geq 0$ per ogni x e $\int_{-\infty}^{+\infty} f(x)\,dx = 1$. La prima condizione è banalmente verificata, e anche la seconda è vera perché

$$\int_{-\infty}^{+\infty} f(x)\,dx = \int_0^1 p\,dx + \int_1^2 (1-p)\,dx = p + (1-p) = 1.$$

b) La funzione di ripartizione (FdR) verifica, per una v.a. con densità,

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)\,dy, \quad x \in \mathbb{R}.$$

Quindi

$$F(x) = \begin{cases} 0 & \text{per } x < 0, \\ \int_0^x p\,dx = px & \text{per } 0 \leq x < 1, \\ \int_0^1 p\,dx + \int_1^x (1-p)\,dx = p + (1-p)(x-1) & \text{per } 1 \leq x < 2, \\ 1 & \text{per } x \geq 2. \end{cases}$$

c) Notiamo che F è strettamente crescente, quindi invertibile, su $(0,2)$. Inoltre $F(1) = p$. Dunque:
- per $0 < \beta < p$, il β-quantile r_β sta in $(0,1)$ e verifica $pr_\beta = \beta$, da cui $r_\beta = \beta/p$;
- per $p \leq \beta < 1$, il β-quantile r_β sta in $[1,2)$ e verifica $p + (1-p)(r_\beta - 1) = \beta$, da cui $r_\beta = 1 + (\beta - p)/(1-p)$.

(Disegnare F può essere di aiuto.)

■

Esercizio 4.18 Lanciamo ripetutamente un dado equilibrato. Qual è la probabilità che il "6" appaia per la prima volta al quarto lancio? Se il "6" non è apparso nei primi due lanci, qual è la probabilità che appaia per la prima volta al sesto lancio? ■

Soluzione Siamo in presenza di uno schema di Bernoulli (con possibilmente infinite prove), con probabilità di successo, cioè "6", $1/6$. La v.a. T che conta l'istante del primo successo ha quindi distribuzione $G(1/6)$. La probabilità che il "6" appaia per la prima volta al quarto lancio è

$$\mathbb{P}(T=4) = \frac{1}{6}\left(\frac{5}{6}\right)^3 = \frac{5^3}{6^4} = 0.096.$$

Per la proprietà di assenza di memoria, la probabilità che il "6" appaia per la prima volta al sesto lancio, sapendo che non è apparso nei primi due lanci, è

$$\mathbb{P}(T=6 \mid T>2) = \mathbb{P}(T=4) = 0.096.$$

■

Esercizio 4.19 **(consigliato 14/03)** Il numero di clienti delle poste di Urbopoli, in una data ora, è descritto da una v.a. di Poisson di parametro 2.3. Qual è la probabilità che, in quell'ora, vi siano al massimo 2 clienti? ■

Soluzione Sia X il numero di clienti, $X \sim P(2.3)$, abbiamo

$$\mathbb{P}(X \leq 2) = \mathbb{P}(X=0) + \mathbb{P}(X=1) + \mathbb{P}(X=2) = e^{-2.3}\left(\frac{2.3^0}{0!} + \frac{2.3^1}{1!} + \frac{2.3^2}{2!}\right) = 0.596.$$

■

Esercizio 4.20 Il tempo di vita (in giorni) di un certo macchinario (cioè il tempo di funzionamento prima del primo guasto) è descritto da una v.a. esponenziale di parametro $1/8$. Qual è la probabilità che il primo guasto si verifichi dopo 6 giorni? ■

Soluzione Sia $T \sim \text{Exp}(\lambda = 1/8)$ il tempo di vita, la probabilità cercata è

$$\mathbb{P}(T>6) = \int_6^\infty \lambda e^{-\lambda x} dx = -e^{-\lambda x}\,|_{x=6}^{\infty} = e^{-3/4}.$$

■

Esercizio 4.21 Lanciamo quattro volte un dato equo. Supponiamo di ricevere, per ogni lancio, due euro se esce 6, un euro se esce 4 o 5, zero altrimenti. Qual è il valore atteso del denaro che riceviamo? ■

Soluzione Sia X_i il denaro ricevuto per l'i-simo lancio, X_i ha valore atteso

$$E[X_i] = 2 \cdot \mathbb{P}(X=2) + 1 \cdot \mathbb{P}(X=1) + 0 \cdot \mathbb{P}(X=0) = 2 \cdot \frac{1}{6} + 1 \cdot \frac{2}{6} = \frac{2}{3}.$$

Per linearità, il valore atteso della vincita $S = X_1 + \ldots X_4$ è $E[S] = E[X_1] + \ldots + E[X_4] = 8/3$. ■

Esercizio 4.22 — Devore 3.103. Considera una malattia la cui presenza può essere identificata tramite un esame del sangue. Sia p la probabilità che un individuo selezionato casualmente abbia la malattia. Supponiamo di selezionare n individui in modo indipendente per il test. Un modo per procedere consiste nell'effettuare un test separato per ciascuno dei n campioni di sangue. Un approccio potenzialmente più economico, noto come *test di gruppo*, è stato introdotto durante la Seconda Guerra Mondiale per identificare uomini affetti da sifilide tra le reclute dell'esercito. Questo approccio prevede i seguenti passi:

- Si preleva una parte di ciascun campione di sangue, si combinano i campioni e si effettua un unico test sul campione combinato.
- Se nessuno degli individui è malato, il risultato del test sarà negativo, e sarà necessario un solo test.
- Se almeno un individuo è malato, il test sul campione combinato darà un risultato positivo; in tal caso, sarà necessario effettuare i n test individuali.

Supponendo che $p = 0.1$:

a) Qual è il numero atteso di test utilizzando questa procedura se $n = 3$?
b) Qual è il numero atteso di test utilizzando questa procedura se $n = 5$?

■

Esercizio 4.23 (consigliato 14/03)

a) Mostrare che la seguente funzione è la densità di probabilità di una variabile aleatoria:

$$f(x) = \begin{cases} 3x^{-4}, & \text{se } x > 1 \\ 0, & \text{se } x \leq 1 \end{cases}.$$

b) Se X ha densità $f_X = f$, si determini quali momenti possiede X e li si calcoli.
c) Si determini se la variabile $Y = \log(X)$ ha densità e momento primo (ed in tal caso li si calcoli).

■

Soluzione a) Una funzione f è densità se e solo se $f(x) \geq 0$ per ogni x e $\int_{\mathbb{R}} f(x)\,dx = 1$. In questo caso la prima condizione è banalmente verificata, mentre per la seconda abbiamo

$$\int_{\mathbb{R}} f(x)\,dx = \int_1^{\infty} 3x^{-4}\,dx = -x^{-3} \mid x = 1^{\infty} = 1.$$

b) Poiché X è positiva ($f = 0$ fuori da $(1,\infty)$), esiste $E[X^n] \in [0,\infty]$ per ogni n intero positivo.

Vale

$$E[X^n] = \int_1^\infty x^n 3x^{-4}\,dx$$
$$= \begin{cases} \frac{3}{n-3}x^{n-3}\,|_{x=1}^{\infty} = \frac{3}{3-n} & \text{per } n<3 \\ 3\log x\,|_{x=1}^{\infty} = +\infty & \text{per } n=3 \\ \frac{3}{n-3}x^{n-3}\,|_{x=1}^{\infty} = +\infty & \text{per } n>3 \end{cases}$$

In particolare X possiede momento n-simo se e solo se $n<3$.

c) f_X è supportata su $(1,\infty)$, la funzione $h:(1,\infty)\to(0,\infty)$, $h(x)=\log x$ è C^1, invertibile con inversa $h^{-1}(y)=e^y$ a sua volta C^1, con derivata e^y. Quindi, per il teorema di cambio variabile, $Y=h(X)$ ammette densità

$$f_Y(y) = f_X(h^{-1}(y))\left|\frac{d}{dy}h^{-1}(y)\right| 1_{(0,\infty)}(y)$$
$$= 3e^{-4y}e^y 1_{(0,\infty)}(y) = 3e^{-3y}1_{(0,\infty)}(y).$$

In particolare, Y è v.a. esponenziale di parametro 3. Come tale, ha momento primo $1/3$.

■

Esercizio 4.24 Sia $\beta>0$ un parametro e sia X una v.a. avente densità $f(x)=\beta x^{-\beta-1}$ per $x\in(1,+\infty)$, $f(x)=0$ per $x\notin(1,+\infty)$. Determinare, in funzione di β e n, se X ammette momento di ordine n e, se sì, calcolarlo. ■

Soluzione Poiché X è positiva ($f=0$ fuori da $(1,\infty)$), esiste $E[X^n]\in[0,\infty]$ per ogni n intero positivo. Vale

$$E[X^n] = \int_1^\infty x^n \beta x^{-\beta-1}\,dx$$
$$= \begin{cases} \frac{\beta}{n-\beta}x^{n-\beta}\,|_{x=1}^{\infty} = \frac{\beta}{n-\beta} & \text{per } n<\beta \\ (\beta)\log x\,|_{x=1}^{\infty} = +\infty & \text{per } n=\beta \\ \frac{\beta}{n-\beta}x^{n-\beta}\,|_{x=1}^{\infty} = +\infty & \text{per } n>\beta \end{cases}$$

In particolare X possiede momento n-simo se e solo se $n<\beta$. ■

Esercizio 4.25 Sia $c\in\mathbb{R}$ un parametro, si consideri la funzione

$$f(x) = \begin{cases} cx^2 & -1<x<1 \\ 0 & |x|\geq 1 \end{cases}.$$

a) Dimostrare che f è densità se e solo se $c=3/2$.

D'ora in poi consideriamo $c=3/2$, sia X una v.a. con densità f.

b) Determinare i momenti, se esistono, di X e la varianza, se esiste, di X.

c) Determinare la densità, se esiste, di $Y=X^2$ e riconoscerla.

■

Soluzione a) f è densità se e solo se $f \geq 0$ e $\int_{\mathbb{R}} f(x)\,dx = 1$. La prima condizione è soddisfatta se e solo se $c \geq 0$, la seconda se e solo se

$$1 = c\int_{-1}^{1} x^2\,dx = \frac{2c}{3},$$

quindi $c = 3/2$ è l'unico valore che rende f densità.

b) Poiché $f = 0$ fuori da un intervallo limitato, X ammette tutti i momenti. Poichè f è pari, per n dispari $x^n f(x)$ è dispari, quindi

$$E[X^n] = \int_{\mathbb{R}} x^n f(x)\,dx = 0.$$

Invece per n pari abbiamo

$$E[X^n] = \int_{-1}^{1} \frac{3}{2}x^{2+n}\,dx = \frac{3}{3+n}.$$

In particolare la varianza è $\mathrm{Var}(X) = E[X^2] = 3/5$.

c) Notiamo che la trasformazione $[-1,1] \ni x \mapsto x^2 \in [0,1]$ non è invertibile ($x^2 = (-x)^2$), quindi non possiamo usare direttamente la formula di cambio variabile. Usiamo allora un altro metodo: calcoliamo la funzione di ripartizione F_Y di $Y = X^2$. Poiché $Y \in [0,1]$ q.c., abbiamo $F_Y(y) = 0$ per $y < 0$ e $F(y) = 1$ per $y \geq 1$, mentre per $0 \leq y < 1$,

$$F_Y(y) = \mathbb{P}(X^2 \leq y) = \mathbb{P}(-\sqrt{y} \leq X \leq \sqrt{y}) = \frac{1}{2}x^3\,|_{x=-\sqrt{y}}^{\sqrt{y}} = y^{3/2}.$$

Quindi F_Y è una funzione continua su $\mathbb{R}$ (come si verifica in particolare nei punti 0 e 1) e C^1 sui tratti $(-\infty,0)$, $(0,1)$, $(1,\infty)$. Quindi Y ammette densità

$$f_Y(y) = F_Y'(y) = \frac{3}{2}\sqrt{y}1_{(0,1)}(y).$$

■

Esercizio 4.26 Dati $a, b \in \mathbb{R}$ parametri reali, si consideri la funzione

$$f(x) = \begin{cases} -x & -1 \leq x < 0 \\ ax+b & 0 \leq x \leq 1 \\ 0 & |x| > 1 \end{cases},$$

- Trovare tutti i valori di a e b per i quali f è una densità di probabilità.
- Sia X una v.a. avente densità f (con a e b che soddisfano alle condizioni trovate nel punto precedente): calcolare a e b in modo che si abbia $E[X] = 0$.
- In funzione di a e b, calcolare $\mathbb{P}\left(|X| \geq \frac{1}{2} \mid X \leq 0\right)$.

■

Esercizio 4.27 Consideriamo la seguente funzione di ripartizione di v.a. X con densità:

$$F(x) = \begin{cases} 0 & x < 0 \\ x^2 & 0 \leq x < 1 \\ 1 & x \geq 1, \end{cases}$$

Calcolare la densità f corrispondente. ■

Soluzione Poiché F è continua su $\mathbb{R}$ e C^1 a tratti, la densità f è data da $f = F'$ (dove esiste la derivata), cioè

$$f(x) = F'(x) = 2x1_{(0,1)}(x).$$

■

Esercizio 4.28 Siano X, Y due v.a. discrete con Y Bernoulli di parametro q. Supponiamo che, se $Y = 1$ (cioè sotto $P(\cdot \mid Y = 1)$), X abbia distribuzione $p^{(1)}$ e che, se $Y = 0$ (cioè sotto $P(\cdot \mid Y = 0)$), X abbia distribuzione $p^{(0)}$.

a) Dimostrare che la funzione di massa p di X è

$$p(k) = qp^{(1)}(k) + (1-q)p^{(0)}(k).$$

b) Dimostrare che

$$\mathbb{E}[X] = \mathbb{E}[X \mid Y = 1]q + \mathbb{E}[X \mid Y = 0](1-q),$$

dove $\mathbb{E}[X \mid Y = 1]$ è il valore atteso di X sapendo che $Y = 1$.

c) Mostrare con un esempio che non vale l'analoga formula per la varianza, cioè trovare X tale che

$$\mathrm{Var}[X] \neq \mathrm{Var}[X \mid Y = 1]q + \mathrm{Var}[X \mid Y = 0](1-q).$$

dove $\mathrm{Var}[X \mid Y = 1]$ è la varianza di X sapendo che $Y = 1$.

■

Esercizio 4.29 — Devore 4.110. Sia t l'importo delle tasse di vendita che un rivenditore deve al governo per un determinato periodo. L'articolo *"Statistical Sampling in Tax Audits"* (*Statistics and the Law*, 2008: 320–343) propone di modellare l'incertezza di t come una variabile casuale normale con valore medio μ e deviazione standard σ. Sia a l'importo stimato che il rivenditore deve. Diciamo che si verifica una **sotto-stima** (*under-assessment*) se $t > a$, una **sovrastima** (*over-assessment*) se $a > t$. La funzione di perdita proposta è:

$$L(a,t) = \begin{cases} t-a & \text{se } t > a, \\ k(a-t) & \text{se } t \leq a, \end{cases}$$

dove $k > 1$ per incorporare l'idea che una sovrastima è più grave di una sotto-stima. Trovare il

valore di a che minimizza la perdita attesa:

$$\mathbb{E}[L(a,t)].$$

■

Esercizio 4.30 — Difficile. Si mostri che la seguente è una densità di probabilità,

$$f(x) = \begin{cases} \frac{1}{1+\pi^{\sin(x)}} & x \in [-1,1] \\ 0 & x \notin [-1,1]. \end{cases}$$

e si calcolino tutti i momenti pari di una variabile aleatoria con tale densità. ■

Esercizio 4.31 (consigliato 14/03) Sia X una v.a. gaussiana di media 2 e varianza 9.

a) Si calcolino $\mathbb{P}\{1 \le X \le 3\}$, $\mathbb{P}\{X \ge 2\}$, $\mathbb{P}\{X \ge 3\}$, $\mathbb{P}\{|X-2| \le 3\}$.
b) Si trovino valori a, b tali che $\mathbb{P}\{X \le a\} = 0.7$, $\mathbb{P}(2 < X < b) = 0.3$.
c) Si trovi il valore c tale che $\mathbb{P}\{2-c < X < 2+c\} = 0.95$.

■

Soluzione a) Usiamo la standardizzazione: $Z = (X-2)/\sqrt{9}$ è v.a. gaussiana standard. Quindi, usando le tavole e le proprietà di simmetria della FdR normale standard Φ,

$$\begin{aligned} \mathbb{P}(1 \le X \le 3) &= \mathbb{P}(-1/3 \le Z \le 1/3) = \Phi(1/3) - \Phi(-1/3) \\ &= 2\Phi(1/3) - 1 = 0.2586, \\ \mathbb{P}(X \ge 2) &= \mathbb{P}(Z \ge 0) = 0.5, \\ \mathbb{P}(X \ge 3) &= \mathbb{P}(\ge 1/3) = 1 - \Phi(1/3) = 0.5707, \\ \mathbb{P}(|X-2| \le 3) &= \mathbb{P}(-1 \le Z \le 1) = 2\Phi(1) - 1 = 0.682. \end{aligned}$$

b) Chiamando q_β il β-quantile di $N(0,1)$, abbiamo

$$\mathbb{P}(X \le a) = \Phi(Z \le (a-2)/3) = 1 - \Phi((a-2)/3)$$

da cui, detta $\tilde{a} = (a-2)/3$, $\mathbb{P}(X \le a) = 0.7$ se e solo se $\tilde{a} = q_{0.3} = -q_{0.7} = -0.53$, cioè $a = 0.41$. Analogamente abbiamo

$$\mathbb{P}(2 < X < b) = \Phi(0 < Z < (b-2)/3) = \Phi((b-2)/3) - 0.5$$

da cui, detta $\tilde{b} = (b-2)/3$, $\mathbb{P}(2 < X < b) = 0.3$ se e solo se $\tilde{b} = q_{0,8} = 0.84$, cioè $b = 4.52$.

c) Abbiamo

$$\mathbb{P}(2-c < X < 2+c) = \Phi(-c/3 < Z < c/3) = 2\Phi(c/3) - 1,$$

da cui $\mathbb{P}(2-c < X < 2+c) = 0.95$ se e solo se $c/3 = q_{0.975} = 1.96$, cioè $c = 5.88$.

■

Esercizio 4.32 Una fabbrica produce viti la cui lunghezza ha distribuzione gaussiana di media 1.7cm e deviazione standard 0.6cm.

a) Calcolare la probabilità che una vite abbia lunghezza inferiore a 1.6cm.

b) Quanto deve valere x (in cm) in modo che il 95% delle viti abbia lunghezza almeno x?
c) Qual è la probabilità che, estratte 3 viti a caso, esattamente 2 abbiano lunghezza inferiore a 1.6cm?

■

Soluzione a) Sia X la lunghezza di una vite, $X \sim N(1.7, 0.6^2)$. Per standardizzazione, $Z = (X - 1.7)/0.6$ è normale standard. La probabilità cercata è

$$\mathbb{P}(X < 1.6) = \mathbb{P}(Z < (1.6 - 1.7)/0.6) = \Phi(-1/6) = 1 - \Phi(1/6) = 0.433.$$

b) Detto $\tilde{x} = (x - 1.7)/0.6$, abbiamo

$$\mathbb{P}(X \geq x) = \mathbb{P}(Z \geq \tilde{x}) = 1 - \Phi(\tilde{x}),$$

quindi $\mathbb{P}(X \geq x) = 0.95$ se e solo se $\tilde{x} = q_{0.05} = -q_{0.95} = -1.64$, cioè $x = 0.713$.
c) Siamo in presenza di 3 prove ripetute (l'estrazione delle viti), ci interessa la v.a. S che conta il numero di viti con lunghezza inferiore a 1.6cm (successo, con probabilità $p = 0.433$). Quindi S ha distribuzione $\mathrm{Bin}(3, 0.433)$. La probabilità cercata è

$$\mathbb{P}(S \geq 2) = \mathbb{P}(S = 2) + \mathbb{P}(S = 3) = 3 \cdot 0.433^2 \cdot 0.567 + 0.433^3 = 0.319.$$

■

Esercizio 4.33 L'esplosione di un'autocisterna in autostrada ha provocato l'emissione di un inquinante. Si stima che la quantità relativa di inquinante riversato sull'autostrada segua una distribuione gaussiana centrata nel punto dell'esplosione e di deviazione standard 0.5km. Le autorità vogliono chiudere l'autostrada per un tratto assicurandosi che la quantità di inquinante riversato al di fuori di quel tratto sia inferiore allo 0.5% di tutto l'inquinante riversato. Quale tratto bisogna chiudere? ■

Soluzione Sia X la distanza di una generica particella inquinante dal punto dell'esplosione, per ipotesi $X \sim N(0, 0.5^2)$. Vogliamo trovare x tale che $\mathbb{P}(|X| \leq x) = 0.995$. Per standardizzazione $Z = X/0.5 = 2X$ è normale standard, quindi

$$\mathbb{P}(|X| \leq x) = \mathbb{P}(|Z| \leq 2x) = 2\Phi(2x) - 1.$$

Quindi $\mathbb{P}(|X| \leq x) = 0.995$ se e solo se $2x = q_{0.9975} = 2.81$, cioè $x = 1.405$. ■

Esercizio 4.34 (consigliato 21/03) Diciamo che una v.a. X positiva soddisfa il principio di Pareto, o la regola 80:20, se il 20% più ricco della popolazione detiene l'80% della ricchezza, dove interpretiamo X come ricchezza di un individuo della popolazione. Rigorosamente (per una v.a. X con densità f):

- il gruppo che detiene la frazione γ più ricca ha una ricchezza X pari almeno a $r_{1-\gamma}$, dove $r_{1-\gamma}$ è il quantile di ordine $1 - \gamma$ di X (cioè $P(X \geq r_{1-\gamma}) = \gamma$);
- la percentuale di ricchezza detenuta da questo gruppo (sul totale della ricchezza della

popolazione) è il rapporto

$$\frac{\mathbb{E}[X1_{[r_{1-\gamma},\infty)}(X)]}{\mathbb{E}[X]}$$

(dove $X1_{[a,\infty)}(X)$ vale X per $X \geq a$, 0 per $X < 0$).
Una v.a. X ha distribuzione di Pareto di parametro $\beta > 0$ se ammette densità

$$f(x) = \frac{\beta}{x^{\beta+1}} 1_{(1,\infty)}(x).$$

- Trovare il parametro $\beta > 1$ tale che la distribuzione di Pareto soddisfi il principio di Pareto.
- Per il parametro trovato, determinare quali momenti ammette X; X ammette varianza?

■

Soluzione a) Dato γ, il valore $r_{1-\gamma}$ si trova imponendo $\mathbb{P}(X \geq r_{1-\gamma}) = \gamma$, cioè

$$1-\gamma = \int_{r_{1-\gamma}}^{\infty} \frac{\beta}{x^{\beta+1}}\,dx = -x^{-\beta}\,|_{x=r_{1-\gamma}}^{\infty} = \frac{1}{r_{1-\gamma}^{\beta}},$$

da cui $r_{1-\gamma} = (1-\gamma)^{-1/\beta}$. Per $\beta > 1$, $E[X] < \infty$ e quindi possiamo calcolare i valori attesi. I valori attesi di interesse sono

$$E[X1_{[r_{1-\gamma},\infty)}(X)] = \int_{r_{1-\gamma}}^{\infty} x\frac{\beta}{x^{\beta+1}}\,dx = -\frac{\beta}{\beta-1}x^{1-\beta}\,|_{x=r_{1-\gamma}}^{\infty} = \frac{\beta}{(\beta-1)}r_{1-\gamma}^{1-\beta},$$
$$E[X] = \int_{r_{1-\gamma}}^{\infty} x\frac{\beta}{x^{\beta+1}}\,dx = \frac{\beta}{(\beta-1)}.$$

Imponiamo ora il principio di Pareto, ponendo $\gamma = 1/5$, $E[X1_{[r_{1-\gamma},\infty)}(X)]/E[X] = 4/5$, otteniamo

$$\frac{4}{5} = \left(\frac{4}{5}\right)^{1-1/\beta},$$

da cui $\beta = \log 5/\log 4 \approx 1.16$.

b) Per l'esercizio 4.24, $E[X^n] < \infty$ se e solo se $n < \beta = \log 5/\log 4$. In particolare, X non ammette momento secondo finito e quindi non ammette varianza.

■

Esercizio 4.35 — Integrale Gaussiano, metodo di Laplace. Si consideri il quadrato dell'integrale Gaussiano,

$$\left(\int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}}dx\right)^2 = \left(\int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}}dx\right)\cdot\left(\int_{-\infty}^{+\infty} e^{-\frac{y^2}{2}}dy\right) = \iint_{\mathbb{R}^2} e^{-\frac{x^2+y^2}{2}}dxdy,$$

e si mostri passando in coordinate polari $x = r\cos(\theta)$, $y = r\sin(\theta)$ che tale integrale doppio ha valore 2π. ■

Esercizio 4.36 — Integrale Gaussiano, metodo di Feynman. Sia $G = \int_0^\infty e^{-x^2} dx$. Si consideri la funzione di $a \geq 0$ data da

$$g(a) = \int_0^\infty \frac{e^{-a^2(x^2+1)}}{x^2+1} dx; \quad g(0) = \frac{\pi}{2}, \quad g(\infty) = 0,$$

(il valore in $a = 0$ ricordando la derivata dell'arcotangente), e differenziando sotto il segno di integrale,

$$g'(a) = -2Ge^{-a^2},$$

da cui, integrando ambo i membri in da si ottiene una equazione chiusa per G. ∎

Il problema di generare (con una macchina) numeri *veramente* casuali è complesso, esula dallo scopo di queste note e richiederebbe una discussione approfondita sul significato di "esito casuale" di un esperimento. Assumendo però di poter generare un numero casuale con distribuzione uniforme sull'intervallo $[0,1]$ (compito implementato come funzione di base in praticamente ogni software per il calcolo numerico), possiamo mostrare come generare un numero casuale la cui legge di probabilità è data da una funzione di ripartizione F assegnata. Per semplicità, ci restringiamo al caso in cui la distribuzione ha densità f continua a tratti e strettamente positiva.

Esercizio 4.37 Sia data una densità f continua a tratti e strettamente positiva e sia $F : \mathbb{R} \to \mathbb{R}$ la corrispondente funzione di ripartizione (quindi $F' = f$ e F è invertibile). Sia poi $U : \Omega \to [0,1]$ una variabile con densità uniforme definita su uno spazio di probabilità $(\Omega, \mathscr{F}, \mathbb{P})$. Si mostri che

$$X : \Omega \to \mathbb{R}, \quad X = F^{-1}(U),$$

è una variabile aleatoria avente densità f e funzione di ripartizione F. ∎

NB Saper fare:

- Individuare le informazioni di base sulle v.a. e sulle loro distribuzioni a partire dalla descrizione del problema in esame, in particolare esperimenti su più livelli.
- Calcolare le funzioni di massa/densità a partire dalla descrizione del problema in esame.
- Verificare se una funzione è funzione di massa/densità.
- Calcolare probabilità del tipo $\mathbb{P}(X \in A)$ a partire dalla funzione di massa/densità di X.
- Riconoscere gli esempi notevoli: Bernoulli, binomaile, geometrica, Poisson, uniforme (su intervallo), esponenziale, gaussiana.
- Usare l'approssimazione di Poisson per la binomiale.
- Usare la proprietà di assenza di memoria per geometrica ed esponenziale.
- C alcolare valori attesi di $E[\varphi(X)]$, inclusi momenti e varianza, a partire dalla funzione di massa/densità di X.
- Calcolare il valore atteso di una combinazione lineare di v.a. (per linearità).
- Usare la proprietà di scaling per la varianza.
- Calcolare la densità (se esiste) di una funzione di v.a. (cambio variabili).
- Calcolare la densità dalla funzione di ripartizione e viceversa.
- Calcolare le probabilità relative a una v.a. gaussiana tramite funzione di ripartizione.
- Usare i quantili della v.a. gaussiana.

5. Variabili Aleatorie Multivariate

Esercizio 5.1 **(consigliato 21/03)** Siano X_1, X_2 gli esiti di due lanci di un dado equilibrato. Trovare la funzione di massa congiunta delle v.a. $U = \min\{X_1, X_2\}$ e $V = \max\{X_1, X_2\}$. ■

Soluzione Notiamo dapprima che (X,Y) ha valori in $\{1,2,\dots 6\}^2$ e legge uniforme (la distribuzione dei due lanci). La v.a. doppia (U,V) ha valori in $\{(h,k) \in \{1,2,\dots 6\}^2 \mid h \leq k\}$ e distribuzione congiunta

$$\mathbb{P}(U=h, V=k) = \begin{cases} \mathbb{P}(X=Y=h) = \frac{1}{36} & \text{se } h=k, \\ \mathbb{P}((X,Y)=(h,k) \text{ o } (X,Y)=(k,h)) = \frac{2}{36} = \frac{1}{18} & \text{se } h<k. \end{cases}$$

■

Esercizio 5.2 Uno sportello ha due linee; la prima linea serve al massimo 2 clienti, mentre la seconda al massimo 3. Siano X e Y il numero di clienti rispettivamente alla prima e alla seconda linea. La funzione di massa congiunta di X e Y è data da

X \Y	0	1	2	3
0	0.05	0.15	0.1	0.05
1	0.1	0.2	0.05	0.03
2	0.1	0.1	0.05	0.02

(ad esempio, la probabilità di $X=0$ e $Y=1$ è 0.15).

a) Determinare le funzioni di massa marginali di X e di Y.
b) Determinare la funzione di massa del numero totale di clienti.
c) Determinare la probabilità che la seconda linea abbia (strettamente) più clienti della prima.
d) Le v.a. X e Y sono indipendenti?

■

Soluzione a) Le funzioni di massa marginali si determinano, per la X, sommando sulle righe, mentre per la Y, sommando sulle colonne:

X \Y	0	1	2	3	totale
0	0.05	0.15	0.1	0.05	0.35
1	0.1	0.2	0.05	0.03	0.38
2	0.1	0.1	0.05	0.02	0.27
totale	0.25	0.45	0.2	0.1	1

(ad esempio, $\mathbb{P}(X=0)=0.35$, $\mathbb{P}(Y=0)=0.25$).

b) Il numero totale di clienti è $S=X+Y$, a valori in $0,1,\dots 5$, la sua funzione di massa si calcola a partire dalla funzione di massa congiunta sommando sulle "diagonali da sinistra in basso a destra in alto":

$$\begin{aligned}
&\mathbb{P}(S=0)=\mathbb{P}(X=0,Y=0)=0.05,\\
&\mathbb{P}(S=1)=\mathbb{P}(X=0,Y=1)+\mathbb{P}(X=1,Y=0)=0.25,\\
&\mathbb{P}(S=2)=\mathbb{P}(X=0,Y=2)+\mathbb{P}(X=1,Y=1)+\mathbb{P}(X=1,Y=0)=0.4,\\
&\mathbb{P}(S=3)=\mathbb{P}(X=0,Y=3)+\mathbb{P}(X=1,Y=2)+\mathbb{P}(X=2,Y=1)=0.2,\\
&\mathbb{P}(S=4)=\mathbb{P}(X=1,Y=3)+\mathbb{P}(X=2,Y=2)=0.08,\\
&\mathbb{P}(S=5)=\mathbb{P}(X=2,Y=3)=0.02.
\end{aligned}$$

c) La probabilità cercata è la somma dei numeri nel triangolo "sopra la diagonale $X=Y$", cioè

$$\mathbb{P}(Y>X)=0.15+0.1+0.05+0.05+0.03+0.02=0.4.$$

d) Le v.a. X e Y non sono indipendenti, perché $\mathbb{P}(X=a,Y=b)$ non coincide in generale con $\mathbb{P}(X=a)\mathbb{P}(Y=b)$, come si può verificare ad esempio per $a=b=0$.

■

Esercizio 5.3 Lanciamo due volte un dado equilibrato. Calcolare il valore atteso della somma degli esiti e il valore atteso del prodotto degli esiti. ■

Soluzione Siano X e Y gli esiti dei due lanci, essi sono uniformi su $\{1,2,\dots 6\}$ e indipendenti, inoltre $E[X]=E[Y]=7/2$. Quindi abbiamo

$$E[X+Y]=E[X]+E[Y]=7,$$

e, per l'indipendenza,

$$E[XY]=E[X]E[Y]=\frac{49}{4}.$$

■

Esercizio 5.4 **(consigliato 28/03)** In una fabbrica, il numero di guasti in un giorno ha distribuzione di Poisson di parametro 1.5; inoltre supponiamo che il numero di guasti in un giorno non influenzi il numero di guasti in altri giorni.

a) Qual è la probabilità che, in un dato giorno, ci siano almeno 2 guasti?
b) Qual è la probabilità che, in 2 giorni consecutivi, ci siano almeno 3 guasti?
c) Qual è la probabilità che, in una data settimana lavorativa (cioè 5 giorni), in almeno 4 giorni si verifichi al massimo un guasto?

■

Soluzione a) Sia X il numero di guasti nel dato giorno, $X \sim P(1.5)$. La probabilità cercata è

$$\mathbb{P}(X \geq 2) = 1 - \mathbb{P}(X = 0) - \mathbb{P}(X = 1) = 1 - e^{-1.5}(1 + 1.5) = 0.4422.$$

b) Siano X_1, X_2 il numero di guasti nel primo e nel secondo giorno rispettivamente, per ipotesi X_1 e X_2 sono indipendenti e ciascuna con legge $P(1.5)$. Per riproducibilità, il numero totale di guasti $Y = X_1 + X_2$ ha legge $P(1.5 + 1.5 = 3)$, quindi la probabilità cercata è

$$\begin{aligned}\mathbb{P}(Y \geq 3) &= 1 - \mathbb{P}(Y = 0) - \mathbb{P}(Y = 1) - \mathbb{P}(Y = 2) \\ &= 1 - e^{-3}(1 + 3 + 9/2) = 0.5768.\end{aligned}$$

c) Siamo in presenza di uno schema di Bernoulli, con 5 prove indipendenti, in cui ciascun successo $X_i \leq 1$ (dove X_i è il numero di guasti nel giorno i-simo) ha probabilità $p := \mathbb{P}(X_i \leq 1) = 1 - \mathbb{P}(X_i \geq 2) = 0.5578$ per il punto (a). Quindi il numero di giorni N in cui si verifica al massimo un guasto è una v.a. binomiale di parametri 5 e p. La probabilità cercata è

$$\mathbb{P}(N \geq 4) = \binom{5}{4} p^4 (1-p) + \binom{5}{5} p^5 = 0.268.$$

■

Esercizio 5.5 **(consigliato 28/03, tranne (d))** L'altezza degli italiani adulti segue una distribuzione Gaussiana di media 177 cm e deviazione standard 10.6 cm (dati inventati).
a) Qual è la probabilità che l'altezza di un italiano (adulto) sia maggiore di 180 cm?
b) Qual è la probabilità che, scelti due italiani a caso, la media delle loro altezze sia maggiore di 180 cm?
c) Qual è la probabilità che, scelti 100 italiani a caso, la media delle loro altezze sia maggiore di 180 cm?
d) Come cambia la risposta al punto precedente, se la distribuzione dell'altezza ha sempre media 1.77 e deviazione standard 10.6 ma non è necessariamente Gaussiana?

■

Soluzione a) Sia X l'altezza (in cm) di un italiano, $X \sim N(177, 10.6^2)$. La probabilità cercata è (qui e altrove $Z \sim N(0,1)$)

$$\mathbb{P}(X \geq 180) = \mathbb{P}(Z \geq (180 - 177)/10.6) = 1 - \Phi((180 - 177)/10.6) = 0.4768.$$

b) Siano X_1, X_2 le altezze dei due italiani scelti, per ipotesi X_1 e X_2 sono indipendenti e ciascuna con legge $N(177, 10.6^2)$. Per riproducibilità, la media delle due altezze

$\bar{X}_2 = (X_1 + X_2)/2$ ha distribuzione $N(177, 10.6^2/2)$, quindi la probabilità cercata è

$$\begin{aligned}\mathbb{P}(\bar{X}_2 \geq 180) &= \mathbb{P}(Z \geq (180-177)/(10.6/\sqrt{2})) \\ &= 1 - \Phi((180-177)/(10.6/\sqrt{2})) = 0.3859.\end{aligned}$$

c) Come per il punto (b), la media $\bar{X}_{100}$ delle altezze ha distribuzione $N(177, 10.6^2/100)$, quindi la probabilità cercata è

$$\begin{aligned}\mathbb{P}(\bar{X}_{100} \geq 180) &= \mathbb{P}(Z \geq (180-177)/(10.6/10)) \\ &= 1 - \Phi((180-177)/(10.6/10)) = 0.0188.\end{aligned}$$

d) Per il teorema centrale del limite, (assumendo che $n = 100$ sia abbastanza grande), $\sqrt{100} \cdot (\bar{X}_{100} - 177)/10.6$ ha distribuzione approsimativamente gaussiana standard, o equivalentemente $\bar{X}_{100}$ ha distribuzione approsimativamente $N(177, 10.6^2/100)$. Quindi la risposta al punto (c) non cambia almeno approsimativamente.

■

NB Ricordiamo che, per riproducibilità e invarianza per traformazioni affini, se $X_1 \sim N(m_1, \sigma_1^2), \ldots X_n \sim N(m_n, \sigma_n^2)$ e $a_1, \ldots a_n, b \in \mathbb{R}$, allora

$$a_1 X_1 + \ldots a_n X_n + b \sim N(a_1 m_1 + \ldots a_n m_n + b, a_1^2 \sigma_1^2 + \ldots a_n^2 \sigma_n^2).$$

In particolare, se X_i sono i.i.d. gaussiane di media m e varianza σ^2, $\bar{X}_n = (X_1 + \ldots X_n)/n$ è gaussiana di media m e varianza σ^2/n.

Esercizio 5.6 La temperatura corporea in una persona sana ha distribuzione gaussiana di media 98.2 gradi Fahrenheit (F) e deviazione standard 0.62 gradi F. Una certa direttiva indica 100.6 come temperatura minima per la febbre.

a) Quale percentuale di persone sane sono considerate con la febbre dalla direttiva?
b) In quale intervallo centrale si colloca il 60% delle temperature delle persone sane?
c) Se misuriamo la temperatura a 10 persone scelte a caso, con quale probabilità la loro temperatura media sarà superiore a 98.7 gradi F?
d) Se misuriamo la temperatura a 10 persone scelte a caso, con quale probabilità almeno 6 di loro avranno temperatura superiore ai 98.5 gradi F?

■

Soluzione a) Sia $X \sim N(98.2, 0.62^2)$ la tempreratura corporea (in una persona sana). La proporzione di persone sane considerate con febbre secondo la direttiva è

$$\mathbb{P}(X \geq 100.6) = \mathbb{P}(\frac{X - 98.2}{0.62} \geq \frac{100.6 - 98.2}{0.62} = 3.87) = 1 - \Phi(3.87) \approx 0.$$

b) Dobbiamo trovare un intervallo della forma $I = [98.2 - d, 98 - 2 + d]$ tale che $\mathbb{P}(X \in I) = 0.6$. Abbiamo

$$\begin{aligned}0.6 = \mathbb{P}(X \in I) &= \mathbb{P}(|X - 98.2| \leq d) \\ &= \mathbb{P}(Z \leq \frac{d}{0.62}) = 1 - 2\Phi(d/0.62),\end{aligned}$$

da cui $\Phi(d/0.62) = 0.2$, cioè $d/0.62 = q_{0.2} = -q_{0.8} = 0.84$, cioè $d = 0.52$.

c) Siano X_i le temperature delle 10 persone, le X_i sono i.i.d. di distribuzione $N(98.2, 0.62^2)$. Per riproducibilità delle gaussiane, la temperatura media $\bar{X}$ delle 10 persone ha distribuzione $N(98.2, 0.62^2/10)$. Quindi la probabilità cercata è

$$\mathbb{P}(\bar{X} \geq 98.7) = \mathbb{P}(Z \geq \frac{98.7 - 98.2}{0.62/\sqrt{10}} = 2.55) 1 - \Phi(2.55) = 0.0054.$$

d) Siamo in presenza di uno schema di Bernoulli, con 10 prove ripetute (le temperature dei 10 individui), con probabilità di successo (temperatura superiore a 98.5) $p = \mathbb{P}(X > 98.5) = 0..3156$ (si calcola come al punto (a)). In particolare, il numero N di persone con temperatura superiore a 98.5 ha distribuzione binomiale di parametri 10 e p. La probabilità cercata è

$$\mathbb{P}(N \geq 6) = \sum_{k=6}^{10} \mathbb{P}(N = k) = 0.06$$

(facendo i conti con la densità binomiale).

■

Esercizio 5.7 Il numero giornaliero di clienti a uno sportello segue una distribuzione di Poisson di parametro 2.5. Supponiamo che il numero di clienti in un dato giorno sia indipendenti dal numero di clienti in altri giorni.

a) Qual è la distribuzione del numero di clienti in una settimana lavorativa (= 5 giorni)?
b) Stimare, tramite la disuguaglianza di Chebyshev, la probabilità che in una settimana ci siano almeno 20 clienti.

■

Soluzione a) Sia X_i il numero di clienti nel giorno i-simo della settimana, X_i sono indipendenti con distribuzione $P(2.5)$. Per riproducibilità, il numero totale di clienti $Y = X_1 + \ldots + X_5$ ha distribuzione $P(5 \cdot 2.5 = 12.5)$.

b) Ricordando che $E[Y] = 12.5$, $\mathrm{Var}(Y) = 12.5$, per la disuguaglianza di Chebyshev abbiamo

$$\mathbb{P}(Y \geq 20) = \mathbb{P}(Y - E[Y] \geq 20 - 12.5 = 7.5) \leq \frac{1}{7.5^2} \cdot \mathrm{Var}(Y) = 0.222.$$

■

Esercizio 5.8 Il tempo di vita (in anni) di un certo componente ha distribuzione esponenziale di parametro 0.5.

a) Calcolare la probabilità di guasto del componente entro un anno.

Se il componente si guasta, viene immediatamente sostituito da un altro, il cui tempo di vita ha ancora distribuzione esponenziale di parametro 0.5. Supponiamo che i tempi di vita dei due componenti siano indipendenti.

b) Calcolare il valore atteso e la varianza del tempo di vita totale dei due componenti.

■

Soluzione a) Sia $X \sim \text{Exp}(1/2)$ il tempo di vita del componente. La probabilità cercata è

$$\mathbb{P}(X \leq 1) = \int_0^1 \frac{1}{2} e^{-x/2}\, dx = -e^{-x/2}\,|_{x=0}^{1} = 1 - e^{-1/2}.$$

b) Siano X_1, X_2 i tempi di vita dei due componenti, X_1 e X_2 sono indipendenti di legge $\text{Exp}(1/2)$; ricordiamo che $E[X_i] = 2$, $\text{Var}(X_i) = 4$. Il tempo di vita totale $Y = X_1 + X_2$ ha media $E[Y] = E[X_1] + E[X_2] = 4$ e, per indipendenza delle X_i, ha varianza

$$\text{Var}(Y) = \text{Var}(X_1) + \text{Var}(X_2) = 8.$$

■

Esercizio 5.9 Lanciamo un dado equilibrato, chiamiamo N l'esito, quindi lanciamo una moneta equilibrata N volte, sia X il numero di teste.

a) Calcolare la funzione di massa congiunta di X e N.
b) Calcolare la funzione di massa marginale di X.
c) Calcolare il coefficiente di correlazione tra X e N.

■

Soluzione a) Le informazioni del problema sono le seguenti:

- N ha distribuzione uniforme su $\{1, \dots 6\}$;
- "X ha distribuzione $B(N, 1/2)$": rigorosamente, per ogni $n \in \{1, \dots 6\}$, dato $N = n$ (cioè sotto $\mathbb{P}(\cdot \mid N = n)$), X ha distribuzione $B(n, 1/2)$.

La v.a. doppia (X, N) ha valori in $\{(k,n) \mid n \in \{1, \dots 6\}, k \in \{0, 1, \dots n\}\}$ e la distribuzione di massa congiunta è

$$\mathbb{P}(X = k, N = n) = \mathbb{P}(X = k \mid N = n)\mathbb{P}(N = n) = \binom{n}{k} \left(\frac{1}{2}\right)^n \cdot \frac{1}{6}, \quad k \leq n.$$

b) La v.a. X ha valori in $\{0, 1, \dots 6\}$ e la sua funzione di massa è

$$\mathbb{P}(X = k) = \sum_n \mathbb{P}(X = k, N = n) = \sum_{n=k}^{6} \frac{1}{6} \binom{n}{k} \left(\frac{1}{2}\right)^n$$

(calcolabile per ogni $k = 0, 1, \dots 6$).

■

Esercizio 5.10 **(consigliato (c) 28/03)** Sia $\alpha \in [0, 1]$ un parametro fissato. Lanciamo una moneta equilibrata; se esce testa, lanciamo una moneta con probabilità di testa pari ad α; se invece esce croce, lanciamo una moneta con probabilità di testa pari ad $1 - \alpha$. Siano X_1, X_2 gli esiti rispettivamente del primo e del secondo lancio, indicando con 1 l'esito testa e con 0 l'esito croce.

a) Calcolare la legge congiunta di (X_1, X_2).
b) Calcolare le leggi marginali di X_1 e di X_2.
c) Calcolare il coefficiente di correlazione tra X_1 e X_2.
d) Per quali valori di α, X_1 e X_2 sono indipendenti?

■

Soluzione a-b) Come si vede ad esempio da una rappresentazione ad albero, la legge congiunta di (X_1, X_2) e le leggi marginali sono

$X_1 \backslash X_2$	0	1	totale
0	$\alpha/2$	$(1-\alpha)/2$	$1/2$
1	$(1-\alpha)/2$	$\alpha/2$	$1/2$
totale	$1/2$	$1/2$	1

c) X_1 e X_2 hanno distribuzione Bernoulli di parametro $1/2$ e quindi valore atteso $1/2$ e varianza $1/4$. Abbiamo

$$E[X_1X_2] = \sum_{a,b=0}^{1} ab\mathbb{P}(X_1 = a, X_2 = b) = 1 \cdot 1 \cdot \mathbb{P}(X_1 = X_2 = 1) = \alpha/2$$

e quindi la covarianza è

$$\mathrm{Cov}(X_1, X_2) = E[X_1X_2] - E[X_1]E[X_2] = \alpha/2 - 1/4.$$

Il coefficiente di correlazione è

$$\rho(X_1, X_2) = \frac{\mathrm{Cov}(X_1, X_2)}{\sqrt{\mathrm{Var}(X_1)\,\mathrm{Var}(X_2)}} = \frac{\alpha/2 - 1/4}{1/4} = 2\alpha - 1.$$

d) Se X_1 e X_2 sono indipendenti, allora $\mathrm{Cov}(X, Y) = 0$, e quindi $\alpha = 1/2$. Dunque per $\alpha \neq 1/2$, X_1 e X_2 non sono indipendenti. Per $\alpha = 1/2$, si verifica facilmente che $\mathbb{P}(X_1 = a, X_2 = b) = \mathbb{P}(X_1 = a)\mathbb{P}(X_2 = b)$ per ogni $a, b \in \{0, 1\}$, quindi X_1 e X_2 sono indipendenti. ■

Esercizio 5.11 Calcolare in modo approssimato la probabilità che, su 1000 lanci di moneta equilibrata, ci siano almeno 480 teste. Calcolare poi, sempre in modo approssimato, il valore k tale che, con probabilità del 95%, testa compaia almeno k volte (su 1000). ■

Soluzione a) Sia X il numero di teste nei 1000 lanci, X è binomiale di media 500 e varianza 250. In particolare, $np(1-p) = 250 \geq 10$, quindi "n è sufficientemente grande" per usare il TCL. Usando il TCL, abbiamo

$$\begin{aligned}\mathbb{P}(X \geq 480) &= \mathbb{P}((X - 480)/\sqrt{250} \geq (480 - 500)/\sqrt{250} = -1.26) \\ &\approx \mathbb{P}(Z \geq -1.26) = 1 - \Phi(-1.26) = \Phi(1.26) = 0.896.\end{aligned}$$

b) Dobbiamo cercare k tale che $\mathbb{P}(X \geq k) = 0.95$. Usando ancora il TCL, abbiamo

$$\begin{aligned}\mathbb{P}(X \geq k &= \mathbb{P}((X - 480)/\sqrt{250} \geq (k - 500)/\sqrt{250}) \\ &\approx \mathbb{P}(Z \geq (k - 500)/\sqrt{250}) = 1 - \Phi((k - 500)/\sqrt{250}),\end{aligned}$$

da cui otteniamo

$$(k - 500)/\sqrt{250} \approx q_{0.05} = -q_{0.95} = -1.64$$

e quindi $k \approx 473.91$. Il valore intero $k = 474$ soddisfa il requisito.

■

Esercizio 5.12 (consigliato 31/03) Una ditta produce certi componenti elettronici dei quali circa il 20% sono difettosi: questi componenti sono esportati in scatole da 400 pezzi e la ditta si impegna a sostituire integralmente la scatola se il numero di pezzi difettosi è superiore a 90.

a) Qual è (approssimativamente) la probabilità che la ditta debba sostituire una scatola?
b) Se si vuole che la probabilità di dover sostituire una scatola sia inferiore a 0.05, come deve migliorare la produzione (cioè di quanto deve -approssimativamente- scendere la percentuale di pezzi difettosi)?

■

Soluzione a) Sia X il numero di pezzi difettosi in una scatola, X ha distribuzione binomiale di parametri 400 e 0.2, quindi con media $400 \cdot 0.2 = 80$ e varianza $400 \cdot 0.2 \cdot 0.8 = 64 (= 8^2)$. In particolare, $np(1-p) = 64 \geq 10$, quindi n è sufficientemente grande per usare il TCL. La probabilità cercata è approssimativamente

$$\begin{aligned}\mathbb{P}(X \geq 90) &= \mathbb{P}((X-80)/8 \geq (90-80)/8 = 1.25) \\ &\approx \mathbb{P}(Z \geq 1.25) = 1 - \Phi(1.25) = 0.1056.\end{aligned}$$

b) Sia p la probabilità di un pezzo difettoso. Assumiamo per ora $n = 400$ grande tale che $np(1-p) \geq 10$, a posteriori verificheremo se il p trovato soddisfa o no questa condizione (e quindi se l'approssimazione gaussiana che usiamo è valida o meno). Per il TCL, abbiamo (chiamando $z(p) = (90-400p)/\sqrt{400p(1-p)}$)

$$\begin{aligned}\mathbb{P}(X \geq 90) &= \mathbb{P}((X-400p)/\sqrt{400p(1-p)} \geq (90-400p)/\sqrt{400p(1-p)}) \\ &\approx \mathbb{P}(Z \geq z(p)) = 1 - \Phi(z(p)).\end{aligned}$$

Imponiamo che tale probabilità sia 0.05, troviamo $z(p) = q_{0.95} = 1.64$. Elevando al quadrato, otteniamo

$$(90-400p)^2 = q_{0.95}^2 400p(1-p).$$

Risolvendo l'equazione di secondo grado, otteniamo $p = 0.193$ e $p = 0.261$. Il secondo valore però non risolve il problema originario, perché la media $400 \cdot 0.261$ è addirittura > 90 (in effetti $p = 0.261$ è la soluzione di $z(p) = -1.64$). Quindi il valore p cercato è 0.193. Notiamo che, per questo p, $np(1-p) \geq 10$ e quindi l'approssimazione normale usata è giustificata.

■

Esercizio 5.13 (consigliato 31/03) Un server riceve email la cui dimensione ha valore atteso 4.73 MB e deviazione standard 0.53 MB. In un giorno il server riceve 70 email.

a) Qual è la probabilità che la dimensione totale delle 70 email superi 340 MB?
b) Quanto deve valere y in MB affinché, con probabilità del 95%, la dimensione totale delle email sia inferiore a y?

■

Soluzione a) Sia X_i la dimensione (in MB) dell'i-sima mail ricevuta. In assenza di altre informazioni, è ragionevole supporre che le X_i siano indipendenti (se conosco la dimensione di una mail, questo non mi dà informazioni sulle altre mail) e identicamente distribuite (la distribuzione della dimensione dell'i-sima mail non dipende da i) [notare però che alcune informazioni aggiuntive potrebbero richiedere un modello non i.i.d.: ad esempio, se sappiamo che ogni giorno viene inviata una mail promemoria molto pesante]. Sia $\bar{X}_{70}$ la dimensione media delle mail, e quindi $70\bar{X}_{70}$ è la dimensione totale. Poiché $n = 70$ è grande, possiamo usare il TCL e ottenere

$$\begin{aligned}\mathbb{P}(70\bar{X}_{70} \geq 340) &= \mathbb{P}(\sqrt{70}(\bar{X}_{70} - 4.73)/0.53 \geq \sqrt{70}(340/70 - 4.73)/0.53 = 2.01) \\ &\approx \mathbb{P}(Z \geq 2.01) = 1 - \Phi(2.01) = 0.0222.\end{aligned}$$

b) Cerchiamo y tale che $\mathbb{P}(70\bar{X}_{70} \leq y) = 0.95$. Sempre per il TCL abbiamo

$$\begin{aligned}\mathbb{P}(70\bar{X}_{70} \leq y) &= \mathbb{P}(\sqrt{70}(\bar{X}_{70} - 4.73)/0.53 \leq \sqrt{70}(y/70 - 4.73)/0.53 =: z(y)) \\ &\approx \mathbb{P}(Z \leq z(y)) = \Phi(z(y)),\end{aligned}$$

da cui $z(y) = q_{0.95} = 1.64$. Risolvendo in y otteniamo $y = 338.37$.

■

Esercizio 5.14 Il numero giornaliero di errori commessi da un server segue una distribuzione di Poisson di parametro 2.5. Supponiamo che il numero di errori commessi in un dato giorno sia indipendenti dal numero di errori in altri giorni.

a) Qual è la probabilità che, in un giorno, avvengano almeno 3 errori?
b) Qual è la probabilità che, in due dati giorni consecutivi, avvengano almeno 3 errori?
c) Qual è la probabilità che, in un anno (= 365 giorni), avvengano almeno 900 errori?
d) Qual è la probabilità che, in almeno 2 giorni di una data settimana, il server commetta almeno 3 errori?

■

Soluzione a) Sia $X \sim P(2.5)$ il numero di errori nel giorno considerato, la probabilità cercata è

$$\mathbb{P}(X \geq 3) = 1 - e^{-2.5}\left(1 + \frac{2.5}{2} + \frac{2.5^2}{6}\right) = 0.4562.$$

b) Siano X_1, X_2 il numero di errori nei due giorni considerati, X_1 e X_2 sono i.i.d. $\sim P(2.5)$. Quindi per riproducibilità delle Poisson, il numero totale di errori ha distribuzione $P(2.5 + 2.5 = 5)$. La probabilità cercata è

$$\mathbb{P}(X \geq 3) = 1 - e^{-5}\left(1 + \frac{5}{2} + \frac{5^2}{6}\right) = 0.8753.$$

c) Siano $X_1 \ldots X_{365}$ il numero di errori in ciascun giorno dell'anno, le $X_i$sono i.i.d. $\sim P(2.5)$. Poiché $n = 365$ è "grande". possiamo usare il TCL e ottenere

$$\begin{aligned}\mathbb{P}(X_1 + \ldots X_{365} \geq 900) &= \mathbb{P}((X_1 + \ldots X_{365} - 465 \cdot 2.5)/\sqrt{365 \cdot 2.5} \geq (900 - 365 \cdot 2.5)/\sqrt{365 \cdot 2.5}) \\ &\approx \mathbb{P}(Z \geq -0.41) = \Phi(0.41) = 0.6591.\end{aligned}$$

d) Siamo in presenza di uno schema di Bernoulli, con 7 prove (numero di giorni della settimana), successo "almeno 3 errori" con probabilità $p = 0.4562$ dal punto (a). In particolare il numero di giorni N in cui il server commette almeno 3 errori ha distribuzione $B(7,p)$. La probabilità cercata è quindi

$$\mathbb{P}(N \geq 2) = 0.9034.$$

■

Esercizio 5.15 Un test diagnostico per una certa malattia ha specificità del 97% (cioè è negativo su una data persona sana con probabilità del 97%) e sesnibilità del 99% (cioè è positivo su una data persona malata con probabilità del 99%). Supponiamo che l'1% della popolazione soffra di questa malattia. Supponiamo inoltre che gli esiti di ripetizioni del test su una data persona siano indipendenti. Estraiamo una persona a caso ed eseguiamo su questa il test 2 volte.

a) Se la persona è sana, qual è la probabilità che il test dia esito positivo in entrambe le esecuzioni?
b) Se il test dà esito positivo in entrambe le esecuzioni, qual è la probabilità che la persona sia sana?
c) Gli esiti delle due esecuzioni sono indipendenti?

■

Soluzione Non forniamo la soluzione completa, ma solo le soluzioni numeriche, rimandiamo all'esercizio 4.8, che è simile per struttura (perché?). Soluzioni: a) 0.0009; b) 0.0833; c) no. ■

Esercizio 5.16 Un dado equilibrato viene lanciato ripetutamente.

a) Calcolare la probabilità di ottenere un numero pari almeno 2 volte nei primi 5 lanci.
b) Calcolare (almeno in modo approssimato) la probabilità che, nei primi 100 lanci, esca un numero pari almeno 55 volte.
c) Calcolare (almeno in modo approssimato) la probabilità che esca un numero pari almeno 55 volte sia nei primi 100 lanci, sia nei successivi 100 lanci.
d) Un giocatore scommette per un numero pari ad ogni lancio dei primi 100, e per un numero non superiore a 2 ad ogni lancio dei successivi 100 lanci. Calcolare (almeno in modo approssimato) la probabilità che il giocatore vinca almeno 90 volte.

■

Soluzione a) Sia Y_n il numero di lanci, nei primi n, con esito pari, Y_n ha distribuzione binomiale di parametri n e $3/6 = 1/2$. La probabilità cercata è

$$\mathbb{P}(Y_5 \geq 2) = 1 - \mathbb{P}(Y_5 = 0) - \mathbb{P}(Y_5 = 1) = 1 - \frac{6}{2^5} = 0.8125.$$

b) Poiché il numero $n = 100$ di lanci è grande, e in particolare $100 \cdot 1/2 \cdot 1/2 = 25 \geq 10$, possiamo applicare il TCL: la probabilità cercata è circa

$$\begin{aligned}\mathbb{P}(Y_{100} \geq 55) &= \mathbb{P}((Y_{100} - 100 \cdot 1/2)/\sqrt{100 \cdot 1/2 \cdot 1/2} \geq (55 - 100 \cdot 1/2)/\sqrt{100 \cdot 1/2 \cdot 1/2} = 1) \\ &\approx \mathbb{P}(Z \geq 1) = 1 - \Phi(1) = 0.1587.\end{aligned}$$

c) Sia Y_{100} il numero di lanci con esito pari tra i primi 100 lanci e sia V_{100} il numero di lanci con esito pari tra i successivi 100 lanci. La probabilità cercata è $\mathbb{P}(Y_{100} \geq 55, V_{100} \geq 55)$. Ora le v.a. Y_{100} e V_{100}, in quanto riferite a gruppi disgiunti di prove ripetute, sono indipendenti. Inoltre V_{100} ha la stessa distribuzione di Y_{100} (si tratta sempre di 100 lanci di dato). Quindi

$$\mathbb{P}(Y_{100} \geq 55, V_{100} \geq 55) = \mathbb{P}(Y_{100} \geq 55)\mathbb{P}(V_{100} \geq 55) = \mathbb{P}(Y_{100} \geq 55)^2 = 0.0252.$$

d) Sia Y_{100} (come sopra) il numero di lanci con esito pari tra i primi 100 lanci e W_{100} il numero di lanci con esito non superiore a 2 nei successivi 100 lanci. La probabilità cercata è $\mathbb{P}(Y_{100} + W_{100} \geq 90)$, in particolare ci interessa la distribuzione di $\mathbb{P}(Y_{100} + W_{100}$.
Come prima, le v.a. Y_{100} e W_{100}, in quanto riferite a gruppi disgiunti di prove ripetute, sono indipendenti. Inoltre, per il TCL, $Y_{100} \sim Bin(100, 1/2)$ ha distribuzione circa gaussiana di media $100 \cdot 1/2 = 50$ e varianza $100 \cdot 1/2 \cdot 1/2 = 25$, e analogamente W_{100} ha distribuzione circa gaussiana di media $100 \cdot 1/3 = 33.33$ e varianza $100 \cdot 1/3 \cdot 2/3 = 22.22$. Ora, per riproducibilità delle v.a. gaussiane (indipendenti), $Y_{100} + W_{100}$ ha distribuzione circa gaussiana di media $50 + 33.33 = 83.33$ e varianza $25 + 22.22 = 47.22$. Quindi

$$\begin{aligned}\mathbb{P}(Y_{100} + W_{100} \geq 90) &= \mathbb{P}((Y_{100} + W_{100} - 83.33)/\sqrt{47.22} \geq (90 - 83.33)/\sqrt{47.22} = 0.97) \\ &\approx \mathbb{P}(Z \geq 0.97) = 0.166.\end{aligned}$$

■

Esercizio 5.17 — Devore 5.56. Un canale di comunicazione binario trasmette una sequenza di "bit" (0 e 1). Supponiamo che, per ogni bit trasmesso, ci sia una probabilità del 10% di un errore di trasmissione (cioè un 0 che diventa un 1 o un 1 che diventa un 0). Supponiamo inoltre che gli errori di bit si verifichino indipendentemente l'uno dall'altro.

a. Consideriamo la trasmissione di 1000 bit. Qual è la probabilità approssimativa che si verifichino al massimo 125 errori di trasmissione?
b. Supponiamo che lo stesso messaggio di 1000 bit venga inviato due volte in modo indipendente. Qual è la probabilità approssimativa che il numero di errori nella prima trasmissione sia meno della metà del numero di errori nella seconda?

■

Esercizio 5.18 — Difficile. Si considerino due variabili X, Y indipendenti equidistribuite con legge uniforme su $[0,1]$. Ricordando la formula di Leibniz per π,

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \cdots = \sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1},$$

si mostri che la probabilità che l'intero più vicino a X/Y sia pari è data da $\frac{5-\pi}{4}$.

Suggerimento: Si rappresenti Ω come il quadrato $[0,1]^2$, X, Y come le due coordinate, e si osservi che l'evento considerato è dato da una successione di sottoinsiemi triangolari di $[0,1]^2$, di cui è sufficiente sommare le aree. ■

NB Saper fare:

- Individuare le informazioni di base sulle v.a. e sulle loro distribuzioni congiunte a partire dalla descrizione del problema in esame, in particolare esperimenti su più livelli.

- Calcolare probabilità del tipo $\mathbb{P}((X,Y) \in A)$ a partire dalla funzione di massa congiunta di (X,Y).
- Verificare l'indipendenza di v.a. con la definizione o con la funzione di massa congiunta.
- Riconoscere l'indipendenza di v.a. (dove presente) a partire dalla descrizione del problema in esame.
- Usare la stabilità dell'indipendenza per composizione e raggruppamenti.
- Calcolare valori attesi del tipo $E[\varphi(X,Y)]$, inclusa la covarianza, a partire dalla funzione di massa congiunta.
- Calcolare il valore atteso del prodotto di v.a. indipendenti.
- Usare la proprietà di riproducibilità per v.a. binomiali/Poisson/gaussiane indipendenti.
- Applicare la LGN per v.a. indipendenti (o scorrelate).
- Applicare il TCL/TLC per calcolo $\mathbb{P}(\text{funzione di somma o media campionaria in } A)$, con n grande e v.a. i.i.d.
- Applicare LGN e TCL per v.a. Bernoulli (frequenza relativa, approssimazione normale della binomiale).

6. Campioni e stimatori

Esercizio 6.1 **(consigliato 04/04)** Sia X una v.a. Poisson di parametro $\lambda > 0$ (non noto) e sia $(X_1, \dots X_n)$ un campione i.i.d. di X. Trovare uno stimatore di massima verosimiglianza (se esiste) per λ. Lo stimatore trovato è corretto? è consistente? ■

Soluzione Sia $p_\lambda(k) = \frac{\lambda^k}{k!} e^{-\lambda}$ la funzione di massa Poisson di parametro λ. La funzione di verosimiglianza è

$$L(\lambda; x_1, \dots x_n) = \prod_{i=1}^{n} p_\lambda(x_i) = e^{-n\lambda} \frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}$$

Notiamo che massimizzare L equivale a massimizzare $\log L$, che ha una forma più semplice (come funzione di λ):

$$\log L = -n\lambda + \log(\lambda) \sum_{i=1}^{n} x_i - \log(\prod_{i=1}^{n} x_i!),$$

$$\frac{d}{d\lambda} \log L = -n + \frac{1}{\lambda} \sum_{i=1}^{n} x_i.$$

In particolare, $\frac{d}{d\lambda} \log L \geq 0$ se e solo se $\lambda \leq \bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i$. Quindi

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

è l'unico stimatore di massima verosimiglianza. In quanto (in questo caso) $\hat{\lambda}$ è la media campionaria, $\hat{\lambda}$ è uno stimatore corretto e consistente di $E[X_1] = \lambda$. ■

Esercizio 6.2 Sia X una v.a. con densità

$$f_\theta(x) = \theta x^{-\theta-1} 1_{(1,+\infty)}(x),$$

con $\theta > 0$ parametro non noto, e sia $(X_1, \dots X_n)$ un campione i.i.d. di X.

a) Verificare che f_θ è una densità (per ogni $\theta > 0$).
b) Trovare uno stimatore di massima verosimiglianza (se esiste) di θ.
c) Nel caso $n = 1$, dire se lo stimatore trovato è corretto.
d) Dire se lo stimatore trovato è consistente. (Si può usare il seguente fatto: se Y_n converge a Y in probabilità, con $P(Y \in G) = 1$ per qualche aperto $G \subseteq \mathbb{R}$, e $g : G \to \mathbb{R}$ è continua, allora $g(Y_n)$ converge a $g(Y)$ in probabilità.)

■

Soluzione a) f_θ è densità se e solo se $f_\theta \geq 0$, banalmente vero qui, e con integrale 1 su $\mathbb{R}$: in questo caso

$$\int_{-\infty}^{\infty} f_\theta(x)\,dx = \int_1^\infty \theta x^{-\theta-1}\,dx = -x^{-\theta}\,|_{x=1}^{\infty} = 1.$$

b) Poiché $f_\theta(x) = 0$ per $x \leq 0$ per ogni θ, possiamo assumere $x_i > 0$ per ogni i (altrimenti il modello è sbagliato). La funzione di verosimiglianza e il suo logaritmo sono

$$L(\theta; x_1, \dots x_n) = \prod_{i=1}^n f_\theta(x_i) = \theta^n \prod_{i=1}^n x_i^{-\theta-1},$$

$$L(\theta; x_1, \dots x_n) = n \log\theta - (\theta+1) \sum_{i=1}^n \log x_i.$$

Derivando, otteniamo

$$\frac{d}{d\theta} \log L = \frac{n}{\theta} - \sum_{i=1}^n \log x_i,$$

in particolare $\frac{d}{d\theta}\log L \geq 0$ se e solo se $\theta \leq \left(\frac{1}{n}\sum_{i=1}^n \log x_i\right)^{-1}$. Dunque

$$\hat\theta = \frac{1}{\frac{1}{n}\sum_{i=1}^n \log X_i}$$

è l'unico stimatore di massima verosimiglianza per θ.

c) Per $n = 1$, per sostituzione $\log x = t$,

$$\begin{aligned}\mathbb{E}[1/\log X_1] &= \int_1^\infty \frac{1}{\log x}\,\theta x^{-\theta-1}\,dx \\ &= \int_0^\infty \frac{1}{t}\,\theta e^{-\theta t}\,dx = +\infty\end{aligned}$$

(poiché $\int_0^1 1/t\,dt = +\infty$). In particolare per $n = 1$ $\hat\theta$ non è corretto.

d) Le variabili $\log X_i$ sono a loro volta i.i.d., in quanto funzione di variabili i.i.d.; inoltre si può verificare che $E[(\log X_1)^2] < \infty$. Quindi per la LGN, $\frac{1}{n}\sum_{i=1}^n \log X_i$ tende in probabilità

a $\mathbb{E}[\log X_1]$, che è (integrando per parti)

$$\begin{aligned}\mathbb{E}[\log X_1] &= \int_1^\infty \log x\, \theta x^{-\theta-1}\, dx \\ &= -x^{-\theta}\log x\,|_{x=1}^{\infty} + \int_1^\infty x^{-\theta-1}\, dx \\ &= 0 - \frac{1}{\theta}x^{-\theta}\,|_{x=1}^{\infty} = \frac{1}{\theta}.\end{aligned}$$

Per il fatto enunciato nel testo, usando come funzione $g : (0,\infty) \to \mathbb{R}$ data da $g(y) = 1/y$, otteniamo che $\hat{\theta} = g\left(\frac{1}{n}\sum_{i=1}^n \log X_i\right)$ tende in probabilità a $g(1/\theta) = \theta$. Quindi $\hat{\theta}$ è consistente.

■

Esercizio 6.3 Sia X una v.a. con densità

$$f_\theta(x) = \frac{1}{12\theta^5}x^9 e^{-x^2/\theta}1_{(0,+\infty)}(x),$$

con $\theta > 0$ parametro non noto, e sia $(X_1,\dots X_n)$ un campione i.i.d. di X. Trovare uno stimatore di massima verosimiglianza (se esiste) di θ. ■

Esercizio 6.4 **(consigliato 04/04)** Sia X una v.a. con densità

$$f_\theta(x) = \frac{\theta}{x^2}1_{(\theta,+\infty)}(x),$$

con $\theta > 0$ parametro non noto, e sia $(X_1,\dots X_n)$ un campione i.i.d. di X.

a) Trovare uno stimatore di massima verosimiglianza (se esiste) di θ.
b) Dire se lo stimatore trovato è consistente.

■

Soluzione a) Poichè $f_\theta(x) = 0$ per $x \le 0$ per ogni $\theta > 0$, possiamo supporre che i dati x_i siano tutti positivi. La funzione di verosimiglianza è

$$L(\theta; x_1,\dots x_n) = \theta^n \prod_{i=1}^n x_i^{-2} 1_{\min_i x_i > \theta}.$$

In particolare, come funzione della sola θ (fissati gli $x_i > 0$), L è crescente su $(0, \min_i x_i)$, mentre è nulla su $(\min_i x_i, +\infty)$. Di conseguenza, L ha massimo (o meglio sup) in $\theta = \min_i x_i$. Dunque

$$\hat{\theta} = \min_i X_i$$

è l'unico stimatore di massima verosimiglianza per θ.

b) L'idea è che, per n grande, con alta probabilità ci sia almeno un X_i basso, e quindi $\min X_i$ sia vicino a θ, cioè la consistenza. Rigorosamente, scriviamo, per $\varepsilon > 0$ arbitrario ma fissato,

$$\mathbb{P}(|\hat{\theta} - \theta| > \varepsilon) = \mathbb{P}(\min_{1\le i\le n} X_i > \theta + \varepsilon) = \mathbb{P}(X_i > \theta + \varepsilon\, \forall i = 1,\dots n)$$

(dove abbbiamo usato che $\hat{\theta} = \min_i X_i$ è sempre $> \theta$). Oiché le v.a. X_i sono i.i.d. abbiamo

$$\mathbb{P}(|\hat{\theta} - \theta| > \varepsilon) = \mathbb{P}(X_1 > \theta + \varepsilon)^n.$$

Ora $\mathbb{P}(X_1 > \theta + \varepsilon) < 1$ (come si può facilmente verificare), perciò $\mathbb{P}(X_1 > \theta + \varepsilon)^n \to 0$ per $n \to \infty$. Segue che $\mathbb{P}(|\hat{\theta} - \theta| > \varepsilon) \to 0$, cioè la consistenza.

■

NB Saper fare:

- Verificare la correttezza di uno stimatore con il calcolo del valore atteso.
- Verificare la consistenza di uno stimatore con la definizione.
- Applicare correttezza e consistenza di media e varianza campionarie.
- Calcolare uno stimatore di massima verosimiglianza, sia tramite derivata sia osservando eventuali discontinuità della funzione di verosimiglianza.

7. Intervalli di fiducia

Esercizio 7.1 **(consigliato 11/04)** Un certo metodo per misurare il pH di una soluzione fornisce un risultato distribuito come una Gaussiana, di media il valore autentico del pH della soluzione e deviazione standard 0.1. Vengono effettuate 50 misurazioni di una soluzione; la media degli esiti di tali misurazioni risulta 8.19.

a) Fornire un intervallo di fiducia di livello 0.95 per il valore autentico del pH della soluzione.
b) Quante misurazioni è necessario eseguire affinché, con livello di fiducia del 95%, la semiampiezza dell'intervallo sia 0.01?

■

Soluzione a) Dobbiamo calcolare un intervallo di fiducia (IF) per la media di una popolazione gaussiana con varianza nota. Precisamente, sia $X \sim N(m, \sigma^2)$ l'esito di una misurazione del pH, con $\sigma = 0.1$, $X_1, \dots X_n$ il campione i.i.d. di X, con $n = 50$. L'IF per m di livello $1 - \alpha = 0.95$ è

$$I = [\bar{X}_n \pm \frac{\sigma}{\sqrt{n}} q_{1-\alpha/2} = [\bar{X}_n \pm 0.0277].$$

Inserendo il valore $\bar{x}_n = 8.19$ fornito dai dati, otteniamo l'IF numerico $[8.19 \pm 0.0277]$.

b) Dobbiamo trovare n tale che

$$0.01 = \frac{\sigma}{\sqrt{n}} q_{1-\alpha/2} = \frac{0.1}{\sqrt{n}} \cdot 1.96.$$

Risolvendo, otteniamo $n = 385$.

■

Esercizio 7.2 **(consigliato 11/04)** Un server che elabora un certo tipo di richiesta può commettere un errore. Su 1000 richieste, sono stati osservati 50 errori.

a) Fornire un intervallo di fiducia di livello 0.99 per la probabilità di errore su una singola richiesta.
b) Quante richieste dobbiamo analizzare, per ridurre la semiampiezza a 0.003 (mantenendo lo stesso livello di fiducia)?

■

Soluzione a) Dobbiamo calcolare un intervallo di fiducia (IF) per la media di una popolazione Bernoulli. Precisamente, sia $X \sim B(p)$ la presenza o meno di errore in una data richiesta (= 1 se presente errore), $X_1, \dots X_n$ il campione i.i.d. di X, con $n = 1000$. Poiché $n\bar{x}_n(1-\bar{x}_n) = 47.5$ è (ben) più grande di 10, possiamo ritenere n grande per applicare il TCL. L'IF per p di livello $1-\alpha = 0.99$ è

$$I = [\bar{X}_n \pm \sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} q_{1-\alpha/2} = [\bar{X}_n \pm \sqrt{\bar{X}_n(1-\bar{X}_n)} \cdot 0.0816].$$

Inserendo il valore $\bar{x}_n = 50/1000 = 0.05$ fornito dai dati, otteniamo l'IF numerico $[0.05 \pm 0.0178]$.

b) In questo caso, la semiampiezza dipende anche dai dati, non noti a priori se cambiamo il numero di misurazioni. Un primo modo di procedere, sicuro ma conservativo, è rimpiazzare $\bar{X}_n(1-\bar{X}_n)$ con il massimo di $x(1-x)$ su $x \in [0,1]$: se n soddisfa

$$0.003 = \sqrt{\frac{\max_{x\in[0,1]} x(1-x)}{n}} \cdot 2.58,$$

allora

$$0.003 \geq \sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} \cdot 2.58,$$

Poichè $\max_{x\in[0,1]} x(1-x) = 1/4$, risolvendo la prima equazione otteniamo $n = \frac{2.58^2}{=} 4 \cdot 0.003^2 = 184900$. A posteriori, è comunque necessario verificare che l'applicazione del TCL era lecita, ad esempio se $n\bar{x}_n(1-\bar{x}_n \geq 10$.
Un secondo modo di procedere è stimare $\bar{X}_n(1-\bar{X}_n)$, o meglio la varianza teorica $p(1-p)$, usando i dati della precedente rilevazione. In particolare, sappiamo che, con fiducia del 99%, p sta in $[0.05 \pm 0.0178]$. Quindi possiamo stimare la varianza teorica $p(1-p)$ di $B(p)$ con il massimo di $x(1-x)$ su $x \in [0.05 \pm 0.0178]$, tale massimo si ha in $X = 0.0678$ e vale 0.0632. Imponendo quindi

$$0.003 = \sqrt{\frac{0.0678}{n}} \cdot 2.58$$

troviamo $n = 46319$, un valore decisamente più basso del precedente. A posteriori, oltre a verificare che potevamo applicare il TCL, dobbiamo verificare anche che l'intervallo abbia effettivamente l'ampiezza richiesta (anche se con bassa probabilità, $\bar{x}_n$ potrebbe cadere al di fuori di $[0.05 \pm 0.0178]$ e rendere la nostra stima non corretta).

■

Esercizio 7.3 (consigliato 11/04) Si misurano i livelli di emoglobina di 80 persone adulte di sesso femminile, ottenendo una media di 14.1 g/dl e una deviazione standard di 0.7 g/dl. Fornire

un intervallo di fiducia del 99% per il livello medio di emoglobina (in una persona adulta di sesso femminile). ■

Soluzione Dobbiamo calcolare un intervallo di fiducia (IF) per la media di una popolazione di cui non conosciamo la distribuzione, caso grandi campioni. Precisamente, sia X il livello medio di emoglobina in una data persona adulta di sesso femminile, $X_1, \dots X_n$ il campione i.i.d. di X, con $n = 80$. L'IF per m di livello $1 - \alpha = 0.99$ è

$$I = [\bar{X}_n \pm \frac{S_n}{\sqrt{n}} q_{1-\alpha/2} = [\bar{X}_n \pm S_n \cdot 0.2885].$$

Inserendo i valori $\bar{x}_n = 14.1$, $s_n = 0.7$ forniti dai dati, otteniamo l'IF numerico $[14.1 \pm 0.202]$. ■

Esercizio 7.4 Su un campione di 80 individui di una certa popolazione, 48 presentano un certo gene.

a) Trovare un intervallo di fiducia di livello 95% per la frequenza del gene nella popolazione.
b) Quanto grande deve essere la taglia del campione, affinché la semiampiezza della stima sia inferiore al 5%, mantenendo lo stesso livello 95%?

■

Soluzione a) Dobbiamo calcolare un intervallo di fiducia (IF) per la media di una popolazione Bernoulli. Precisamente, sia $X \sim B(p)$ la presenza o meno del gene in una data persona (= 1 se presente), $X_1, \dots X_n$ il campione i.i.d. di X, con $n = 80$. Poiché $n\bar{x}_n(1 - \bar{x}_n) = 19.2$ è più grande di 10, possiamo ritenere n grande per applicare il TCL. L'IF per p di livello $1 - \alpha = 0.95$ è

$$I = [\bar{X}_n \pm \sqrt{\frac{\bar{X}_n(1 - \bar{X}_n)}{n}} q_{1-\alpha/2} = [\bar{X}_n \pm \sqrt{\bar{X}_n(1 - \bar{X}_n)} \cdot 0.2191].$$

Inserendo il valore $\bar{x}_n = 48/80 = 0.6$ fornito dai dati, otteniamo l'IF numerico $[0.6 \pm 0.1073]$.

b) Procediamo come nel punto (b) dell'esercizio 7.2. La stima conservativa, rimpiazzando $\bar{X}_n(1 - \bar{X}_n)$ con il massimo di $x(1 - x)$ su $x \in [0, 1]$, cioè $1/4$ (realizzato per $x = 1/2$), fornisce

$$0.05 = \sqrt{\frac{1/4}{n}} \cdot 1.96,$$

fornisce $n = 385$. La stima usando il precedente IF non differisce da questa, poiché l'IF trovato contiene il punto di massimo $x = 1/2$ di $x(1 - x)$. A posteriori, verifichiamo che potevamo applicare il TCL.

■

Esercizio 7.5 **(consigliato (a) 11/04, consigliato (b) 24/04)** Per ogni anno dal 2016 al 2021 viene misurata la temperatura media (in gradi centigradi) a Milano nel mese di settembre; la media e la deviazione standard di tali dati risultano rispettivamente 20.83 e 1.35. Supponiamo che la temperatura media di settembre sia distribuita come una gaussiana.

a) Fornire un intervallo di fiducia a livello 95% per la media delle temperature medie di

settembre a Milano.

b) Per quanti anni dobbiamo misurare la temperatura media (a Milano a settembre) per ottenere una semiampiezza inferiore a 1 grado (con lo stesso livello di fiducia)?

■

Soluzione a) Dobbiamo calcolare un intervallo di fiducia (IF) per la media di una popolazione gaussiana con varianza non nota. Precisamente, sia $X \sim N(m, \sigma^2)$ la temperatura media (a Milano a settembre) di un dato anno, $X_1, \dots X_n$ il campione i.i.d. di X, con $n = 6$ (gli anni da 2016 al 2021). L'IF per m di livello $1 - \alpha = 0.95$ è

$$I = [\bar{X}_n \pm \frac{S_n}{\sqrt{n}} \tau_{1-\alpha/2,n-1} = [\bar{X}_n \pm S_n \cdot 1.0492].$$

Inserendo i valori $\bar{x}_n = 20.83$, $s_n = 1.35$ forniti dai dati, otteniamo l'IF numerico $[20.83 \pm 1.4164]$.

b) La semiampiezza dipende anche dai dati, non noti a priori se cambiamo il numero di misurazioni. In questo caso, non disponiamo di una stima sicura per S_n o per σ, perciò stimiamo σ, e implicitamente S_n, con un intervallo di fiducia a partire dai dati della precedente rilevazione. Un intervallo di fiducia per σ^2 di livello 95% (il livello non deve essere per forza lo stesso) è

$$[0, \frac{(n-1)S_n^2}{\chi^2_{\alpha,n-1}}],$$

che con il valore $s_n = 1.35$ fornito dai dati diventa $[0, 5 \cdot 1.35^2/1.1455 = 7.955]$. Quindi stimare la varianza teorica σ^2 e, con un po' di abuso, la varianza campionaria S_n^2, con 7.955. Un ulteriore problema è dato dal fatto che anche il quantile $\tau_{1-\alpha/2,n-1}$ dipende da n. Notiamo però che $\tau_{1-\alpha/2,n-1}$ è decrescente in n. Quindi avremo $\tau_{1-\alpha/2,n-1} \leq \tau_{1-\alpha/2,5} \leq 2.5706$ (questa stima in verità è piuttosto conservativa: come vedremo n sarà parecchio più grande di 6). Imponiamo ora

$$1 = \sqrt{\frac{7.955}{n}} \cdot 2.5706$$

e troviamo $n = 53.57$, quindi $n = 53$ per eccesso. A posteriori, dobbiamo verificare che l'intervallo abbia effettivamente l'ampiezza richiesta (anche se con bassa probabilità, $\bar{S}_n^2$ potrebbe cadere al di fuori di $[0, 7.955]$ e rendere la nostra stima non corretta).

■

Esercizio 7.6 — Ross 7.20. Una compagnia autoassicura la propria vasta flotta di automobili contro le collisioni. Per determinare il costo medio di riparazione per collisione, è stato scelto casualmente un campione di 16 incidenti. Supponiamo che il costo di una riparazione per collisione segua una distribuzione gaussiana. Se il costo medio di riparazione per questi incidenti è di 2200 euro, con una deviazione standard campionaria di 800 euro, determinare un intervallo di confidenza al 90% per il costo medio per collisione. ■

Soluzione Dobbiamo calcolare un intervallo di fiducia (IF) per la media di una popolazione gaussiana con varianza non nota. Precisamente, sia $X \sim N(m, \sigma^2)$ il costo per una collisione,

$X_1, \dots X_n$ il campione i.i.d. di X, con $n = 16$. L'IF per m di livello $1 - \alpha = 0.90$ è

$$I = [\bar{X}_n \pm \frac{S_n}{\sqrt{n}} \tau_{1-\alpha/2,n-1} = [\bar{X}_n \pm S_n \cdot 0.4383].$$

Inserendo i valori $\bar{x}_n = 2200$, $s_n = 800$ forniti dai dati, otteniamo l'IF numerico $[2200 \pm 350.64]$. ■

Esercizio 7.7 — Ross 7.49. Per stimare la proporzione p di tutti i neonati che sono maschi, è stato registrato il sesso di 10000 neonati. Se 5106 di questi erano maschi, determinare un intervallo di fiducia di livello 95% per la percentuale di maschi. ■

Soluzione Dobbiamo calcolare un intervallo di fiducia (IF) per la media di una popolazione Bernoulli. Precisamente, sia $X \sim B(p)$ il sesso del neonato ($= 1$ se maschio), $X_1, \dots X_n$ il campione i.i.d. di X, con $n = 10000$. Poiché $n\bar{x}_n(1 - \bar{x}_n) = 2499$ è molto più grande di 10, siamo ampiamente nelle condizioni di applicare il TCL. L'IF per p di livello $1 - \alpha = 0.95$ è

$$I = [\bar{X}_n \pm \sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} q_{1-\alpha/2} = [\bar{X}_n \pm \sqrt{\bar{X}_n(1-\bar{X}_n)} \cdot 0.0196].$$

Inserendo il valore $\bar{x}_n = 5106/10000 = 0.5106$ fornito dai dati, otteniamo l'IF numerico $[0.5106 \pm 0.0196] = [0.4910, 0.5302]$. ■

Esercizio 7.8 — Ross 7.51. Un recente sondaggio pubblicato da un quotidiano ha indicato che il Candidato A è preferito al Candidato B con una percentuale di 53 a 47, con un margine di errore di ± 4 percento. Il giornale ha poi affermato che, poiché il divario di 6 punti è maggiore del margine di errore, i lettori possono essere certi che il Candidato A risulti eletto. Questo ragionamento è corretto? [Supporre che gli intervistati diano risposte oneste.] ■

Soluzione Il ragionamento è sbagliato. Infatti, in termini matematici, il sondaggio fornisce un IF per la percentuale di voti per il candidato A di $[0.53 \pm 0.04] = [0.49, 0.57]$ (peraltro non viene fornito il livello di fiducia). In particolare, non possiamo escludere, nemmeno con alta fiducia, che A prenda il 49% dei voti, e che quindi vinca B. ■

Esercizio 7.9 — Devore 7.56. L'esposizione cronica alle fibre di amianto è un noto rischio per la salute. L'articolo *"The Acute Effects of Chrysotile Asbestos Exposure on Lung Function"* (Environ. Research, 1978: 360–372) riporta i risultati di uno studio basato su un campione di operai edili che erano stati esposti all'amianto per un periodo prolungato.

Tra i dati forniti nell'articolo vi sono i seguenti valori (ordinati) di compliance polmonare ($cm^3/cm\ H_2O$) per ciascuno dei 16 soggetti, 8 mesi dopo il periodo di esposizione (la compliance polmonare è una misura dell'elasticità polmonare, ovvero dell'efficienza con cui i polmoni riescono a inspirare ed espirare):

167.9, 180.8, 184.8, 189.8, 194.8, 200.2, 201.9, 206.9,
207.2, 208.4, 226.3, 227.7, 228.5, 232.4, 239.8, 258.6

(a) È plausibile che la distribuzione della popolazione sia normale? Costruire un boxplot dei

dati e confrontarlo con un boxplot di una distribuzione gaussiana.

(b) Calcolare un intervallo di confidenza al 95% per il valore medio vero della compliance polmonare dopo tale esposizione.

(c) Calcolare un intervallo che, con un livello di confidenza del 95%, includa almeno il 95% dei valori di compliance polmonare nella distribuzione della popolazione.

■

Esercizio 7.10 Sia X una v.a. di Poisson di parametro $\lambda > 0$ (non noto) e sia $X_1, \dots X_n$ un campione i.i.d. di X. Determinare un intervallo di fiducia per λ usando solo la media campionaria dei dati. [Soluzione: $[\bar{X}_n \pm \sqrt{\frac{\bar{X}_n}{n}} q_{1-\alpha/2}]$ ■

8. Test statistici

Esercizio 8.1 Una ditta produce aste e dichiara una lunghezza media di 2.3m. Supponiamo che la lunghezza delle aste segua una distribuzione gaussiana, di deviazione standard 0.1m. Viene rilevata la lunghezza per 100 aste, ottenendo una media campionaria di 2.317m.

a) L'affermazione della ditta sulla lunghezza media è plausibile o no? Formulare un test di livello 0.05 e applicarlo ai dati. Calcolare inoltre il p-value dei dati.
b) Se accettiamo H_0, c'è evidenza che la media non discosti da 2.3m di almeno ± 0.02m?
c) Come cambia la risposta al punto (a) se non conosiamo la distribuzione della v.a. né la sua varianza, ma sappiamo che la deviazione standard campionaria è 0.1m?

■

Soluzione a) Sia $X \sim N(m, \sigma^2)$ la lunghezza (in metri) di un'asta, con $\sigma = 0.1$, $X_1, \dots X_n$ il campione i.i.d. di X, con $n = 100$. Siamo in presenza di un test Z per la media m della popolazione gaussiana X, con varianza nota ($\sigma = 1$), ipotesi nulla $H_0 : m = m_0 := 2.3$ (test bilatero). La statistica di test è

$$Z = \sqrt{n}\frac{\bar{X}_n - m_0}{\sigma} = 100 \cdot (\bar{X}_n - 2.3)$$

con distribuzione $N(0,1)$ per $m = m_0$. La regione critica di livello $\alpha = 0.05$ è

$$C = \{|Z| < q_{1-\alpha/2} = 1.96\},$$

quindi rifiutiamo H_0 se e solo se i dati cadono in C. Applichiamo il test a $\bar{x}_n = 2.317$ fornito dai dati: il valore z assunto dalla statistica di test è $100 \cdot (2.317 - 2.3) = 1.7$ con modulo minore di 1.96, che non cade nella regione critica, quindi accettiamo H_0 con questo livello e potenza.

Il p-value corrispondente a $\bar{x}_n = 2.317$ è

$$\bar{\alpha}(\bar{x}_n = 2.317) = \mathbb{P}(|Z| < |z| = 1.7) = 2(1 - \Phi(1.7)) = 0.0892.$$

Si rifiuta per qualunque $\alpha > 0.0892$.

b) Dobbiamo calcolare la potenza del test in 2.3 ± 0.02. La potenza del test in $m = 2.3 + 0.02$ è

$$\begin{aligned}
&\mathbb{P}_{2.32}\left(\left|\sqrt{n}\frac{\bar{X}_n - m_0}{\sigma}\right| \leq q_{1-\alpha/2}\right) \\
&= \mathbb{P}_{2.32}\left(\left|\sqrt{n}\frac{\bar{X}_n - m}{\sigma} + \sqrt{n}\frac{m_0 - m}{\sigma}\right| \leq q_{1-\alpha/2}\right) \\
&= \mathbb{P}_{2.32}\left(-q_{1-\alpha/2} + \sqrt{n}\frac{m - m_0}{\sigma} \leq \sqrt{n}\frac{\bar{X}_n - m}{\sigma} + \sqrt{n}\frac{m_n - m}{\sigma} \leq q_{1-\alpha/2} + \sqrt{n}\frac{m - m_0}{\sigma}\right) \\
&= \Phi\left(q_{1-\alpha/2} + \sqrt{n}\frac{m - m_0}{\sigma}\right) - \Phi\left(-q_{1-\alpha/2} + \sqrt{n}\frac{m - m_0}{\sigma}\right) \\
&= \Phi(3.96) - \Phi(0.04) = 1 - 0.484
\end{aligned}$$

(per simmetria del test bilatero, la potenza in $m = 2.3 - 0.02$ è la stessa). Si tratta di un valore piuttosto basso: se $m = 2.32$, abbiamo comunque probabilità 0.484 di accettare H_0: quindi non c'è evidenza contro 2.30 ± 0.02.

c) Se non conosciamo la distribuzione né la sua varianza, poiché il campione è (relativamente) grande ($n = 100$), possiamo comunque usare un test Z approssimato, con la statistica di test

$$TS = \sqrt{n}\frac{\bar{X}_n - m_0}{S_n}.$$

Se $s_n = 0.1$, il risultato del test sarà identico a quello del punto (a).

■

Esercizio 8.2 **(consigliato 17/04)** Un certo tipo di misurazioni del pH di una sostanza produce misure distribuite in modo gaussiano, con media pari al valore autentico del pH e deviazione standard $\sigma = 0.02$. Dieci misurazioni forniscono una media campionaria di 8.179.

a) Formulare un test di livello 0.05 per decidere se l'ipotesi "valore del pH pari o superiore a 8.2" sia plausibile o no e applicarlo al valore 8.179 assunto dalla media campionaria.
b) Qual è il p-value corrispondente?
c) Qual è la potenza del test in 8.1?

■

Soluzione a) Sia $X \sim N(m, \sigma^2)$ l'esito di una misurazione del pH, con $\sigma = 0.02$, $X_1, \ldots X_n$ il campione i.i.d. di X, con $n = 10$. Siamo in presenza di un test Z per la media m della popolazione gaussiana X, con varianza nota ($\sigma = 0.02$), ipotesi nulla $H_0 : m \geq m_0 := 8.2$ (test unilatero). La statistica di test è

$$Z = \sqrt{n}\frac{\bar{X}_n - m_0}{\sigma} = 158.11 \cdot (\bar{X}_n - 8.2)$$

con distribuzione $N(0,1)$ per $m = m_0$. La regione critica di livello $\alpha = 0.05$ è

$$C = \{Z < q_\alpha = -1.64\},$$

quindi rifiutiamo H_0 se e solo se i dati cadono in C. Applichiamo il test a $\bar{x}_n = 8.179$ fornito dai dati: il valore z assunto dalla statistica di test è $158.11 \cdot (8.179 - 8.2) = -3.32 < -1.64$, che cade nella regione critica, quindi rifiutiamo H_0 con questo livello e potenza.

b) Il p-value corrispondente a $\bar{x}_n = 8.179$ è

$$\bar{\alpha}(\bar{x}_n = 8.179) = \mathbb{P}(Z < z = -3.32) = 1 - \Phi(3.32) = 0.00045,$$

molto vicino a 0: si rifiuta per qualunque $\alpha > 0.00045$, di fatto per ogni ragionevole livello.

c) La potenza del test in 8.1 è

$$\begin{aligned}\mathbb{P}_{8.1}(C) &= \mathbb{P}_{8.1}\left(\sqrt{n}\frac{\bar{X}_n - 8.1}{\sigma} + \sqrt{n}\frac{8.1 - 8.2}{\sigma} < -1.64\right) \\ &= \mathbb{P}(Z < -1.64 - \sqrt{n}\frac{8.1 - 8.2}{\sigma} < 14.17) \approx 1.\end{aligned}$$

Quindi, se il valore vero di m fosse 8.1, il test produrrebbe rifiuto di H_0 con probabilità prossima a 1.

■

Esercizio 8.3 **(consigliato 17/04)** Si hanno a disposizione 16 osservazioni indipendenti di una v.a. gaussiana con media m sconosciuta e varianza nota eguale a 36 ; con queste oservazioni si vuole effettuare il test dell'ipotesi $H_0) m = 30$ contro $H_1) m \neq 30$. Si decide di respingere l'ipotesi $H_0)$ se la media campionaria delle 16 osservazioni cade al di fuori dell'intervallo $(26.91, 33.09)$.

a) A quale livello viene effettuato il test?

b) Se le osservazioni fossero 25 e il test venisse effettuato ancora allo stesso livello del punto (a), quale sarebbe la regione critica?

■

Soluzione a) Siamo nel caso di test Z bilatero per la media m di una popolazione gaussiana, con varianza nota ($\sigma^2 = 36$). In questo caso, la regione critica C a livello α è tale che $\alpha = \mathbb{P}_{m_0}(C)$, quindi, usando la standardizzazione,

$$\begin{aligned}\alpha &= 1 - \mathbb{P}_{30}(|\bar{X}_n - 30| \leq 3.09) = 1 - \mathbb{P}(Z \leq \sqrt{n}\frac{3.09}{\sigma} = 2.06) \\ &= 2\Phi(2.06) - 1 = 0.04\end{aligned}$$

b) In questo caso, la regione critica sarebbe

$$C = \left\{\sqrt{n}\frac{|\bar{X}_n - m_0|}{\sigma} > q_{1-\alpha/2} = 2.06\right\} = \{|\bar{X}_n - 30| > 6/5 \cdot 2.06 = 2.47\}.$$

■

Esercizio 8.4 Una ditta produce chiodi di lunghezza dichiarata uguale a 5 cm, e il proprietario della ditta afferma che la deviazione standard delle lunghezze dei chiodi prodotti non supera 0.2 cm. Analizzando la lunghezza di un campione di 16 pezzi si trova media campionaria di 4.935 cm e varianza campionaria 0.06 cm^2.

a) Supponendo che le lunghezze dei chiodi possano essere rappresentate con variabili aleatorie Gaussiane, si può accettare al livello 0.05 l'affermazione del proprietario della ditta sulla deviazione standard della lunghezza dei chiodi prodotti?
b) Descrivere un test da utilizzare per verificare l'ipotesi che la lunghezza media dei chiodi prodotti sia di 5 cm, e usare i dati del campione analizzato per determinare la plausibilità di questa ipotesi.

■

Esercizio 8.5 **(consigliato 24/04)** Una ditta produce una lozione per la ricrescita dei capelli ed afferma che si nota una ricrescita in almeno il 60% dei casi: tuttavia l'unione consumatori ha effettuato un'indagine ed ha rilevato che su 137 persone che hanno usato quella lozione solo 70 hanno notato una ricrescita dei capelli ed afferma che questa indagine contraddice l'affermazione della ditta. In termini di un modello statistico, se p è la percentuale (sconosciuta) delle persone sulle quali la lozione ha effetti positivi, l'affermazione della ditta si traduce nell'ipotesi

$$H_0)\, p \geq 0.6 \text{ contro } H_1)\, p < 0.6.$$

a) Si può accettare, al livello 0.05, l'ipotesi sopra indicata (cioè l'affermazione della ditta)?
b) A quale livello (approssimativamente) si può accettare l'affermazione della ditta?

■

Soluzione a) Sia $X \sim B(p)$ la v.a. Bernoulli che indica se una data persona (che ha usato la lozione) ha notato una ricrescita ($X = 1$ se sì, $= 0$ altrimenti), $X_1, \dots X_n$ il campione i.i.d. di X con $n = 139$. Siamo in presenza di un test per la media p (probabilità di ricrescita) di una popolazione di Bernoulli, ipotesi nulla $H_0 : p \geq p_0 := 0.6$ (test unilatero). Poiché $np_0(1-p_0) = 33.36 \geq 10$, possiamo ritenere n grande per applicare il TCL, il test quindi è un test Z approssimato. La statistica di test è

$$Z = \sqrt{\frac{n}{p_0(1-p_0)}}(\bar{X}_n - p_0) = 24.07 \cdot (\bar{X}_n - 0.6)$$

con distribuzione approsimativamente $N(0,1)$ sotto $p = p_0$. La regione critica di livello $\alpha = 0.05$ è

$$C = \{Z < q_\alpha = -1.64\}.$$

Applichiamo il test a $\bar{x}_n = 70/139 = 0.5036$ fornito dai dati. Il valore assunto dalla statistica di test è $24.07 \cdot (0.5036 - 0.6) = -2.13 < -1.64$, che cade nella regione critica, quindi rifiutiamo H_0 con questo livello e potenza.

b) Il p-value corrispondente a $\bar{x}_n = 70/139 = 0.5036$ è

$$\bar{\alpha}(\bar{x}_n = 0.5036) = \mathbb{P}(Z < z = -2.13) = 0.0166,$$

quindi rifiutiamo H_0 per ogni $\alpha > 0.0166$ (di fatto, per α molto vicini a 0.0166, il risultato è più incerto data l'approssimazione del TCL).

■

Esercizio 8.6 **(consigliato 24/04)** Il responsabile di una ditta petrolifera afferma che il contenuto medio di zolfo per litro, nella benzina prodotta da quella ditta, non supera 0.15mg/l; tuttavia l'unione consumatori contesta questa affermazione perché sono stati prelevati 41 campioni che hanno dato valori $x_1, \dots, x_{41}$ dai quali si ottiene

$$\bar{x} = \frac{x_1 + \cdots + x_{41}}{41} = 0.2.$$

Il responsabile afferma che questo dato non è significativo poiché la variabilità era alta: si è infatti ottenuto il valore $\sum_{i \le 41} (x_i - \bar{x})^2 = 1$. Si interpretino i valori delle misurazioni come variabili Gaussiane con media e varianza ignote. Si può accettare l'affermazione del responsabile della ditta (cioè l'ipotesi che il contenuto medio di zolfo non superi 0.15mg/l)? Impostare un opportuno test e calcolare il relativo p-value. ■

Soluzione Sia $X \sim N(m, \sigma^2)$ la concentrazione di zolfo nella benzina, con σ non noto, $X_1, \dots X_n$ il campione i.i.d. di X, con $n = 41$. Siamo in presenza di un test t per la media m della popolazione gaussiana X, con varianza non nota, ipotesi nulla $H_0 : m \le m_0 := 0.15$ (test unilatero). La statistica di test è

$$TS = \sqrt{n}\frac{\bar{X}_n - m_0}{S_n} = \sqrt{41}\frac{\bar{X}_n - 0.15}{S_n}$$

con distribuzione $t_{n-1} = t_{40}$ per $m = m_0$. La regione critica di livello $\alpha = 0.05$ è

$$C = \{TS > \tau_{1-\alpha,n-1} = 1.6839\},$$

quindi rifiutiamo H_0 se e solo se i dati cadono in C. Applichiamo il test a $\bar{x}_n = 0.2$, $s_n^2 = (\sum_i (x_i - \bar{x}_n)^2)/(n-1) = 0.025$ forniti dai dati: il valore ts assunto dalla statistica di test è $\sqrt{41} \cdot (0.2 - 0.15)/\sqrt{0.025} = 2.02 > 1.6839$, che cade nella regione critica, quindi rifiutiamo H_0 con questo livello e potenza.
Il p-value corrispondente a $\bar{x}_n = 0.2$ e $s_n^2 = 0.025$ è (dove T_{n-1} ha distribuzione t-di-Student a $n-1$ gradi di libertà)

$$\bar{\alpha}(\bar{x}_n = 0.2, s_n^2 = 0.025) = \mathbb{P}(T_{n-1} > ts = 2.02) = 1 - F_{T_{n-1}}(2.02).$$

Non sappiamo quanto vale esattamente $F_{T_{n-1}}(2.02)$, ma dalle tavole vediamo che il quantile di T_{40} di ordine 0.975 è 2.0211, molto vicino a 2.02, quindi $F_{T_{n-1}}(2.02)$ è circa 0.975 e il p-value è circa 0.025. Quindi rifiutiamo H_0 per ogni $\alpha > 0.025$. ■

Esercizio 8.7 Vengono effettuate 25 misurazioni del peso di un piccolo mammifero africano, e si ottengono i risultati $(x_1, \dots, x_{25})$ (espressi in grammi) dai quali si ricava una media campionaria eguale a 648.6 ed una varanza campionaria eguale a 457.66.

Si vuole verificare l'ipotesi che il peso medio di questo mammifero sia di 640 grammi, e a tal fine si suppone che i dati possano essere rappresentati con 25 variabili Gaussiane indipendenti.
(i) Effettuare il test dell'ipotesi

$$H_0)\, m = 640 \quad \text{contro} \quad H_1)\, m \neq 640$$

supponendo che la varianza sia nota ed eguale a 400: dopo aver calcolato il p-value, che cosa si conclude? (ii) Effettuare lo stesso test sopra scritto, supponendo però che la varianza sia sconosciuta: che cosa si conclude? ■

Esercizio 8.8 Si consideri un monitoraggio sulla presenza di sostanze tossiche nell'aria, effettuato in 10 stazioni di monitoraggio vicine. I valori ottenuti restituiscono una concentrazione media di $4.8\text{mg}/\text{dm}^3$ con una varianza campionaria di $0.49\text{mg}/\text{dm}^3$. Supponiamo che la distribuzione della concentrazione sia Gaussiana. (i) Fornire una stima della concentrazione delle sostanze tossiche con una fiducia del 90% mediante un intervallo bilatero. Con quale fiducia si ottiene una semiampiezza di $5 \cdot 10^{-2}$? (ii) Dire se l'ipotesi che la concentrazione non sia superiore a $4.3\text{mg}/\text{dm}^3$ è plausibile. ■

Esercizio 8.9 Un certo farmaco viene testato per controllare possibili effetti collaterali. Su 863 pazienti cui viene somministrato il farmaco, 19 manifestano sintomi influenzali. Sappiamo (da precedenti studi) che il tasso di insorgenza di sintomi influenzali, in pazienti non trattati con il farmaco in esame, è 1.9%. (i) Vi è evidenza a livello 0.05 che il farmaco induca sintomi influenzali? (ii) Calcolare il p-value dei dati. ■

Esercizio 8.10 Uno studio afferma che la pressione sanguigna (sistolica) nell'uomo sia distribuita come una normale di media $\mu = 120$ e deviazione standard $\sigma = 7$ (dati inventati). Da una rilevazione su 100 individui, risulta una deviazione standard campionaria pari a 8.5. L'ipotesi "deviazione standard inferiore a 7" è plausibile a livello 0.05? ■