Esercizi di Statistica

Serena Arima, Marilena Barbieri, M.Brigida Ferraro,
Stefania Gubbiotti Brunero Liseo,
MEMOTEF
Università di Roma "La Sapienza"

February 19, 2015

Indice

Prefazione

Lista di esercizi per il corso di Statistica di base

Capitolo 1

Statistica descrittiva

1.1 Popolazione, campione e variabili

Esercizio 1.1.

Secondo un'indagine della Goldman Sachs, soltanto il 4% delle famiglie statunitensi ha un conto online. In un sondaggio della Cyber Dialogue riportato su USA Today si è cercato di indagare sui motivi per cui i clienti hanno chiuso il proprio conto online dopo un periodo di prova. Di seguito trovate le risposte degli intervistati alla domanda: "*Perchè hai chiuso il tuo conto online?*"

Perchè hai chiuso il tuo conto online?	
Troppo complicato o richiede troppo tempo	27%
Insoddisfatto dal servizio clienti	25%
Non mi necessario o non mi interessa	20%
Preoccupato per la sicurezza del conto	11%
Troppo costoso	11%
Sono preoccupato per la privacy	5%

a. Descrivere la popolazione per l'indagine della Goldman Sachs;

b. Descrivere la popolazione per l'indagine della Cyber Dialogue;

c. La risposta alla domanda considerata è qualitativa o quantitativa?

• • •

Soluzione.

a. La popolazione di riferimento per l'indagine della Goldman Sachs è costituita da tutte le famiglie statunitensi.

b. La popolazione di riferimento per l'indagine della Cyber Dialogue è costituita dalle famiglie statunitensi che avevano un conto online e hanno deciso di chiuderlo.

c. La risposta alla domanda considerata è qualitativa.

• • •

Esercizio 1.2.

In un fast food vengono venduti 3 diversi tipi di bevande: bibite, tè e caffè.

a. Spiegare perchè il tipo di bevanda venduta è un esempio di carattere qualitativo sconnesso.

b. Le bibite vengono vendute in 3 dimensioni diverse: piccola, media e grande. Di che carattere si tratta?

• • •

Soluzione.

a. Il tipo di bevanda è un carattere qualitativo sconnesso: le sue modalità sono definite mediante sostantivi e non ammettono un ordinamento tra loro (infatti date due bevande è possibile affermare soltanto se esse sono uguali o diverse tra loro).

b. La dimensione della bibita è un carattere qualitativo ordinato perchè le sue modalità sono attributi non numerici, ma logicamente ordinabili (infatti una bevanda 'piccola' è di dimensione inferiore ad una 'media', che a sua volte è di dimensione inferiore ad una 'grande').

• • •

Esercizio 1.3.

Per ognuna delle seguenti variabili dire di che tipo di variabile si tratta e la scala di misura di riferimento:

a. Numero di telefoni per famiglia;

b. Tipo di telefono usato principalmente;

c. Numero di telefonate al mese;

d. Numero medio di telefonate al mese;

e. Durata (in minuti) delle chiamate;

f. Costo mensile delle telefonate;

g. Esistenza di una linea telefonica collegata ad un modem.

• • •

Soluzione.

a. quantitativo discreto, scala proporzionale.

b. qualitativo sconnesso, scala nominale.

c. quantitativo discreto, scala proporzionale.

d. quantitativo continuo, scala proporzionale.

e. quantitativo continuo, scala proporzionale.

f. quantitativo continuo, scala proporzionale.

g. qualitativo sconnesso, scala nominale.

• • •

Esercizio 1.4. *Identificare le componenti di uno studio*
(dal libro di testo *OpenIntro Statistics* di Diez et al., es. 1.3-1.7)

Identificare (i) le unità, (ii) le variabili e la loro tipologia, (iii) l'obiettivo principale di ricerca, (iv) la popolazione di interesse e il campione negli studi descritti qui di seguito. Discutere inoltre sulla possibilità di generalizzare i risultati dello studio alla popolazione e di stabilire dei nessi causali:

a. Alcuni ricercatori hanno raccolto dei dati per esaminare la relazione tra sostanze inquinanti e nascite premature nel sud della California. Durante lo studio i livelli di inquinamento dell'aria (monossido di carbonio CO, diossido di nitrogeno, ozono, particolato PM 10) sono stati misurati in apposite stazioni di monitoraggio della qualità dell'aria. E' stata inoltre rilevata la durata della gestazione per 143196 nascite tra il 1989 e il 1993 e l'esposizione all'inquinamento dell'aria durante la gestazione è stato calcolato per ciascuna nascita. L'analisi ha mostrato che una maggiore concentrazione di PM 10 e, in misura minore, di CO possono essere associate a nascite premature.

b. Il metodo Buteyko è una tecnica di respirazione debole sviluppata dal medico russo Konstantin Buteyko nel 1952. L'evidenza empirica suggerisce che il metodo Buteyko aiuta a ridurre i sintomi dell'asma e a migliorare la qualità della vita. In uno studio clinico volto a dimostrare l'efficacia di questo metodo, i ricercatori hanno reclutato 600 pazienti malati asma di età compresa tra i 18 e i 69 anni che erano stati sottoposti ad una terapia medica contro l'asma. Questi pazienti sono stati suddivisi in due gruppi: uno sottoposto al metodo Buteyko, l'altro no. Sono stati rilevati degli indici di qualità della vita, di attività, di sintomi dell'asma e riduzione dei trattamenti medici su una scala da 0 a 10. In media, i pazienti del gruppo Buteyko

hanno sperimentato una riduzione significativa nei sintomi dell'asma e un miglioramento di qualità della vita.

• • •

Soluzione.

a. (i) Le unità sono 143196 nuovi nati registrati nel sud della California tra il 1989 e il 1993. (ii) Le variabili misurate sono tutte quantitative continue: monossido di carbonio CO, diossido di nitrogeno, ozono, particolato PM 10. (iii) L'obiettivo della ricerca è stabilire se c'è un'associazione tra l'esposizione all'inquinamento dell'aria e le nascite premature. (iv) La popolazione di interesse è quella di tutte le nascite nel sud della California. Il campione considera invece le 143196 nascite avvenute tra il 1989 e il 1993. Se le nascite in questo periodo di tempo possono essere considerate rappresentative di tutte le nascite del sud della California allora si può pensare che i risultati ottenuti siano generalizzabili all'intera popolazione. Tuttavia, poichè lo studio è di tipo osservazionale, non può essere usato per dimostrare una relazione di tipo causale.

b. (i) Le unità sono 600 pazienti adulti di età compresa tra i 18 e i 69 anni malati di asma e sotto trattamento. (ii) Le variabili misurate su una scala qualitativa ordinale da 0 a 10 (quindi trattabili come quantitative discrete) sono: indici di qualità della vita, di attività, di sintomi dell'asma e riduzione dei trattamenti medici. Inoltre viene considerata una variabile binaria che indica l'appartenenza o non appartenenza al gruppo sperimentale Buteyko. (iii) L'obiettivo della ricerca è dimostrare l'efficacia del metodo Buteyko nel miglioramento della condizione generale del malato d'asma. (iv) La popolazione di riferimento è l'insieme di tutti i pazienti di età compresa tra i 18 e i 69 anni, malati di asma e sotto trattamento. Il campione contiene 600 di questi pazienti. Se assumiamo che il campione contenga dei pazienti volontari, non possiamo pensare che sia un campione rappresentativo e quindi generalizzare i risultati all'intera popolazione. Tuttavia, la natura sperimentale dello studio consente di poter dimostrare statisticamente l'esistenza di una relazione causale.

• • •

Esercizio 1.5. *Iris di Fisher*

(dal libro di testo *OpenIntro Statistics* di Diez et al., es. 1.5)

Il Sig. Ronald Aylmer Fisher era uno statistico inglese, esperto di evoluzione, biologo e genetista, che lavorò tra l'altro su un noto dataset riguardante tre tipi di fiori iris (setosa, versicolor e virginica) per i quali erano state rilevate la lunghezza e la larghezza dei sepali e lunghezza e larghezza dei petali. I dati, relativi a 50 fiori per ciascun tipo[1], sono contenuti nel dataset `iris` (riportato in Appendice e disponibile online).

a. Quante sono le unità?

b. Quante variabili quantitative sono presenti nel dataset? Indicare quali sono, e se sono continue o discrete.

c. Quante variabili categoriche sono presenti nel dataset, e quali sono? Elenca le corrispondenti modalità.

[1]Foto di rtclauss su Flickr, Iris.; R.A Fisher. "The Use of Multiple Measurements in Taxonomic Problems". In: Annals of Eugenics 7 (1936), pp. 179-188

• • •

Soluzione.

a. Le unità sono $50 \times 3 = 150$.

b. Ci sono quattro variabili quantitative: lunghezza dei sepali, larghezza dei sepali, lunghezza dei petali e larghezza dei petali.

c. C'è una sola variabile categorica, il tipo, che presenta tre modalità: *setosa*, *versicolor* e *virginica*.

• • •

1.2 Distribuzioni di frequenza

Esercizio 1.6.
(dal libro di testo *Introduzione alla statistica* di Sheldon M. Ross, es. 1 pag.59)

I dati seguenti indicano il gruppo sanguigno di 50 donatori in un centro di raccolta del sangue.

```
O A O AB A A O O B A O A AB B O O O A B A A O A A O
B A O AB A O O A B A A A O B O O A O A B O AB A O B
```

a. Rappresentare questi dati in una tabella di frequenze.

b. Rappresentare i dati in una tabella di frequenze relative.

c. Calcolare inoltre le frequenze percentuali.

• • •

Soluzione.

	(a)	(b)	(c)
gruppo	frequenze assolute	frequenze relative	frequenze percentuali
0	19	0.38	38
A	19	0.38	38
AB	4	0.08	8
B	8	0.16	16
totale	50	1	100

• • •

Esercizio 1.7.

(dal libro di testo *Introduzione alla statistica* di Sheldon M. Ross, es. 5 pag.35)

I seguenti dati indicano la concentrazione di ozono nell'aria del centro di Los Angeles durante 25 giorni consecutivi nell'estate del 1984:

```
6.2 9.1 2.4 3.6 1.9 1.7 4.5 4.2 3.3 5.1 6.0 1.8 2.3
4.9 3.7 3.8 5.5 6.4 8.6 9.3 7.7 5.4 7.2 4.9 6.2
```

Costruire la distribuzione in classi utilizzando le seguenti classi:
$(0, 2]$,$(2, 4]$,$(4, 7]$, $(7, 10]$.

• • •

Soluzione.

concentrazione	frequenze assolute
(0,2]	3
(2,4]	6
(4,7]	11
(7,10]	5

• • •

Esercizio 1.8.

La seguente tabella riguarda la distribuzione di frequenza del costo di un pasto (espresso in euro):

Costo di un pasto	Frequenza assoluta
$[10, 15)$	1
$[15, 20)$	0
$[20, 25)$	2
$[25, 30)$	15
$[30, 35)$	5
$[25, 40)$	1
$[40, 45)$	3
> 45	15

a. Di che tipo di carattere si tratta? e di che rappresentazione tabellare si tratta?

b. Che differenza c'è rispetto a quella dell'Esercizio 1.7?

c. Costruire le frequenze relative, percentuali.

d. È possibile ricostruire la corrispondente distribuzione unitaria?

• • •

Soluzione.

a. Il carattere costo di un pasto è quantitativo continuo. La tabella precedente rappresenta la distribuzione in classi delle frequenze assolute.

b. La distribuzione data nell'esercizio precente è una distribuzione unitaria, quella che viene richiesto di ricavare è invece una distribuzione in classi: in questo caso possiamo notare che le classi sono chiuse a sinistra e aperte a destra e che l'ultima classe è aperta.

c. La seguente tabella riporta le frequenze relative e percentuali:

Costo di un pasto classi	Frequenze assolute	Frequenze relative	Frequenze percentuali
$[10, 15)$	1	0.02	2%
$[15, 20)$	0	0	0%
$[20, 25)$	2	0.05	5%
$[25, 30)$	15	0.36	36%
$[30, 35)$	5	0.12	12%
$[25, 40)$	1	0.02	2%
$[40, 45)$	3	0.07	7%
> 45	15	0.36	36%
totale	42	1	100%

d. A partire dalla distribuzione in classi non è possibile ricostruire quella unitaria, mentre è possibile il viceversa come abbiamo visto nell'esercizio precedente.

• • •

1.3 Rappresentazioni grafiche per caratteri qualitativi

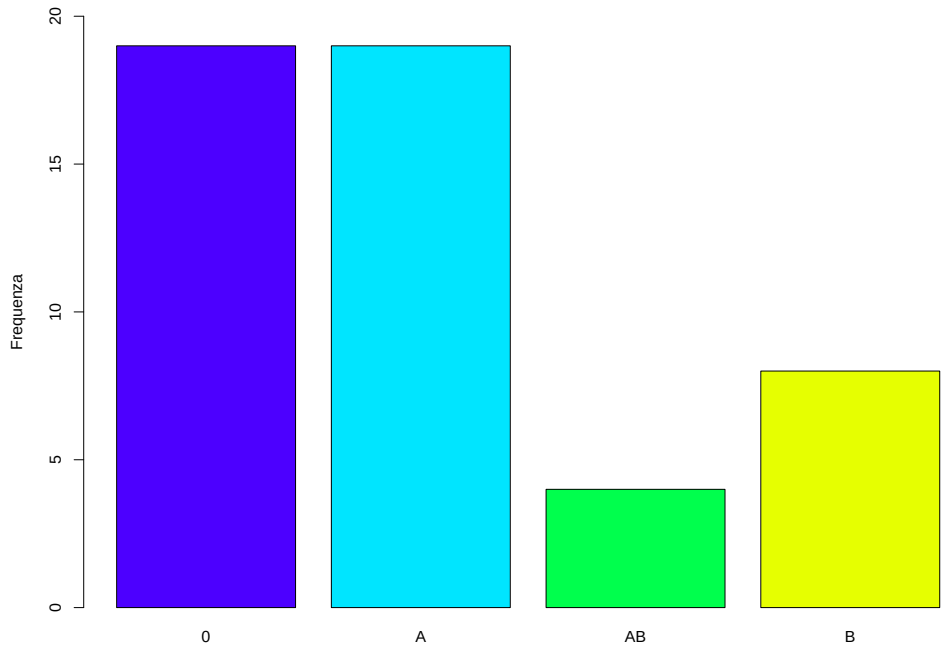
Esercizio 1.9.

(dal libro di testo *Introduzione alla statistica* di Sheldon M. Ross, es. 1 pag.59)

Riprendendo l'Esercizio 1.6, rappresentare la distribuzione mediante un diagramma a barre.

• • •

Soluzione.



• • •

Esercizio 1.10.

Un articolo del *Wall Stree Journal* del luglio 2003 discute l'influenza che Google ha avuto sul web. La tabella seguente mostra come si sono distribuite le ricerche sul web condotte nel maggio 2003 dagli utenti americani di Internet (valori percentuale).

Fonte	Percentuale
Ask Jeeves	3
AOL Time Warner	19
Google	32
MSN-Microsoft	???
Yahoo	25
Altro	6

a. Completare la tabella inserendo il valore mancante.

b. Di che tipo di carattere si tratta? Quali sono le unità statistiche di riferimento?

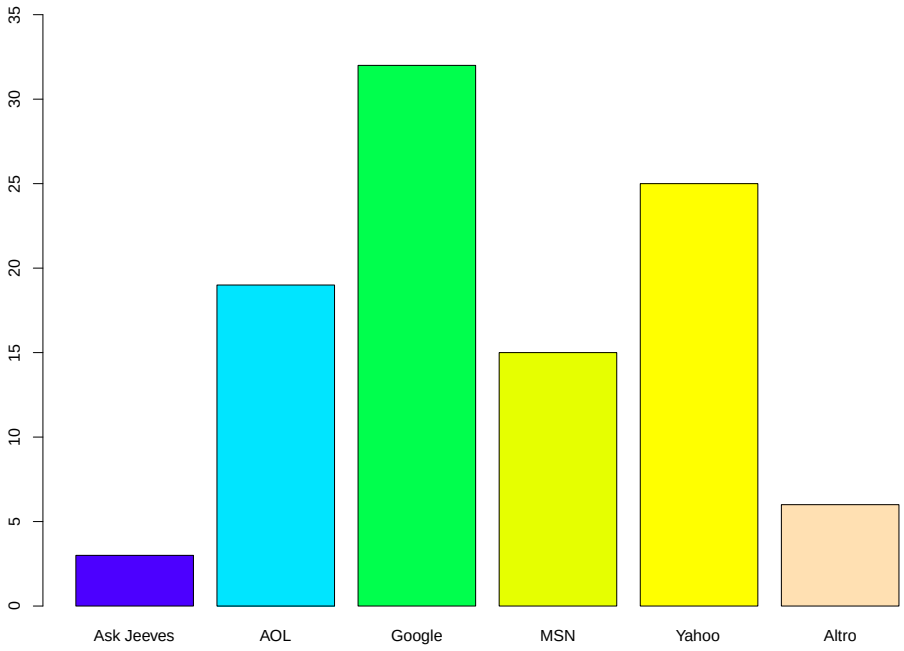
c. Rappresentare graficamente la distribuzione mediante un diagramma a barre.

• • •

Soluzione.

a. Poiché le frequenze devono sommare a 100, il valore mancante è 15.

b. Il carattere considerato è qualitativo sconnesso. Le unità statistiche di riferimento sono le ricerche sul web condotte nel maggio 2003 dagli utenti americani di Internet.



c.

• • •

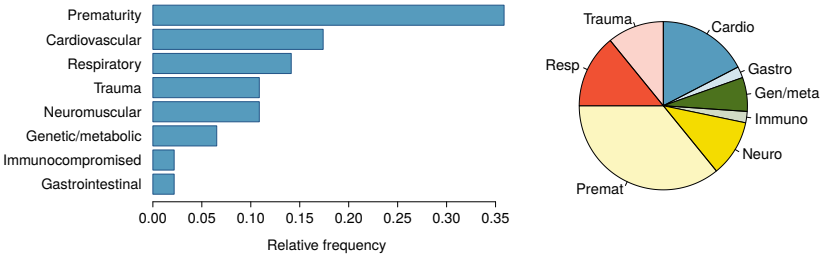
Esercizio 1.11. *Uso degli antibiotici nei bambini*
(dal libro di testo *OpenIntro Statistics* di Diez et al., es. 1.47)

Nei grafici seguenti viene rappresentata la distribuzione della condizione medica precedente di bambini arruolati in uno studio sulla durata ottimale di una terapia antibiotica per la tracheite.

a. Quali sono le caratteristiche che emergono dal diagramma a barre ma non dal diagramma a torta?

b. Quali sono le caratteristiche che emergono dal diagramma a torta ma non dal diagramma a barre?

c. Quale grafico è preferibile per rappresentare questo tipo di dati?



• • •

Soluzione.

a. Nel diagramma a barre è evidente l'ordinamento tra le categorie e vengono rappresentate le frequenze relative.

b. Il diagramma a torta non aggiunge altre informazioni utili a quanto mostrato nel diagramma a barre.

c. In genere il diagramma a barre è preferibile sia per i motivi espressi al punto a. sia perchè il confronto tra lunghezze è più immediato rispetto a quello tra aree.

• • •

1.4 Rappresentazioni grafiche e numeriche per caratteri quantitativi

Esercizio 1.12.

Con riferimento all'Esercizio 1.7

a. Costruire l'istogramma della distribuzione.

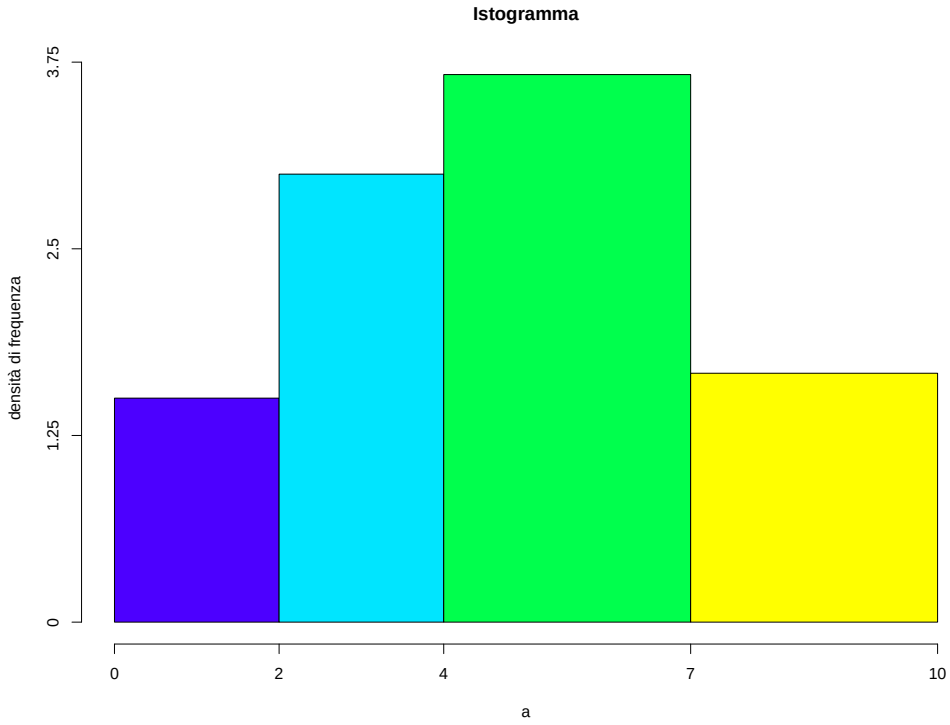
b. Calcolare la media (ovvero la concentrazione media di ozono a Los Angeles nei 25 giorni considerati).

• • •

Soluzione.

a. Calcoliamo innanzi tutto le ampiezze delle classi e le densità di frequenza.

concentrazione	frequenze assolute	ampiezze	densità di frequenza
(0,2]	3	2	1.50
(2,4]	6	2	3.00
(4,7]	11	3	3.67
(7,10]	5	3	1.67



b. Utilizzando la formula della media

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_i + \cdots + x_n}{n} = \\ = \frac{6.2 + 9.1 + 2.4 + 3.6 + 1.9 + \cdots + 6.2}{25} = \frac{125.7}{25} = 5.028$$

• • •

Esercizio 1.13.

Nella seguente tabella sono riportati i tempi di funzionamento, in mesi prima dell'esaurimento, di un campione di batterie.

Durata (mesi)	Frequenza
$[1, 3)$	10
$[3, 6)$	42
$[6, 12)$	38
$[12, 24)$	8

a. Rappresentare graficamente la distribuzione.

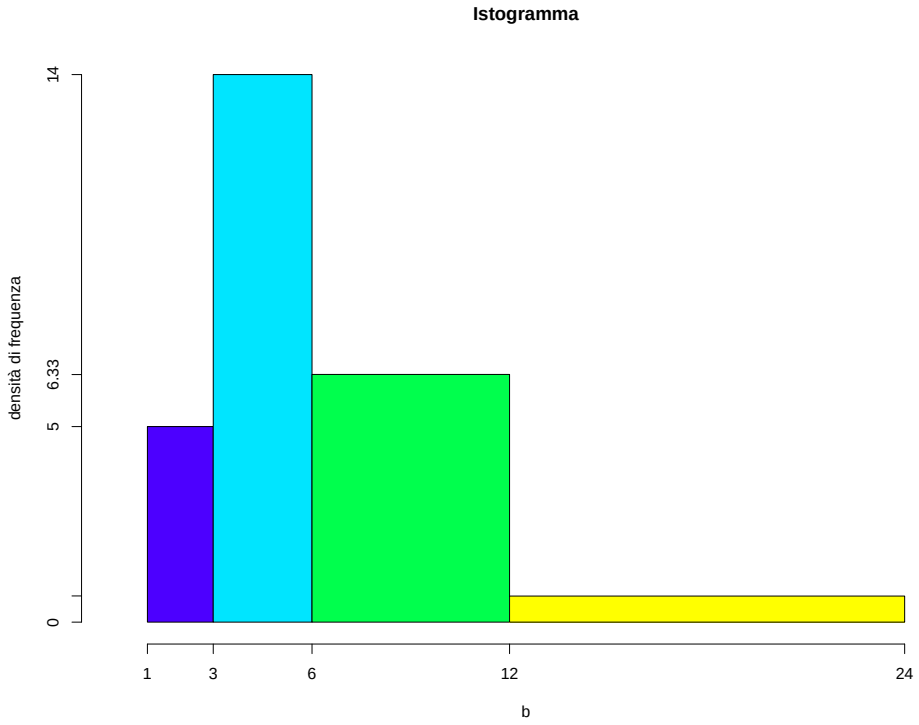
b. Definire e individuare la classe modale.

• • •

Soluzione.

a. Calcoliamo innanzi tutto le ampiezze delle classi e le densità di frequenza.

Durata (mesi)	Frequenza	Ampiezza	Densità
$[1, 3)$	10	2	5.00
$[3, 6)$	42	3	14.00
$[6, 12]$	38	6	6.33
$[12, 24)$	8	12	0.67



b. La classe modale è la classe alla quale risulta associata la massima densità di frequenza: in questo caso è la classe [3,6).

• • •

Esercizio 1.14.

In un'indagine sui consumi delle auto a benzina nei percorsi urbani è stata osservata la distribuzione del numero di litri consumati per 100 Km riportata nella seguente tabella.

Consumo (litri)	Frequenza
$[5, 10)$	15
$[10, 15)$	45
$[15, 25)$	38
$[25, 35)$	2

a. Rappresentare graficamente la distribuzione.

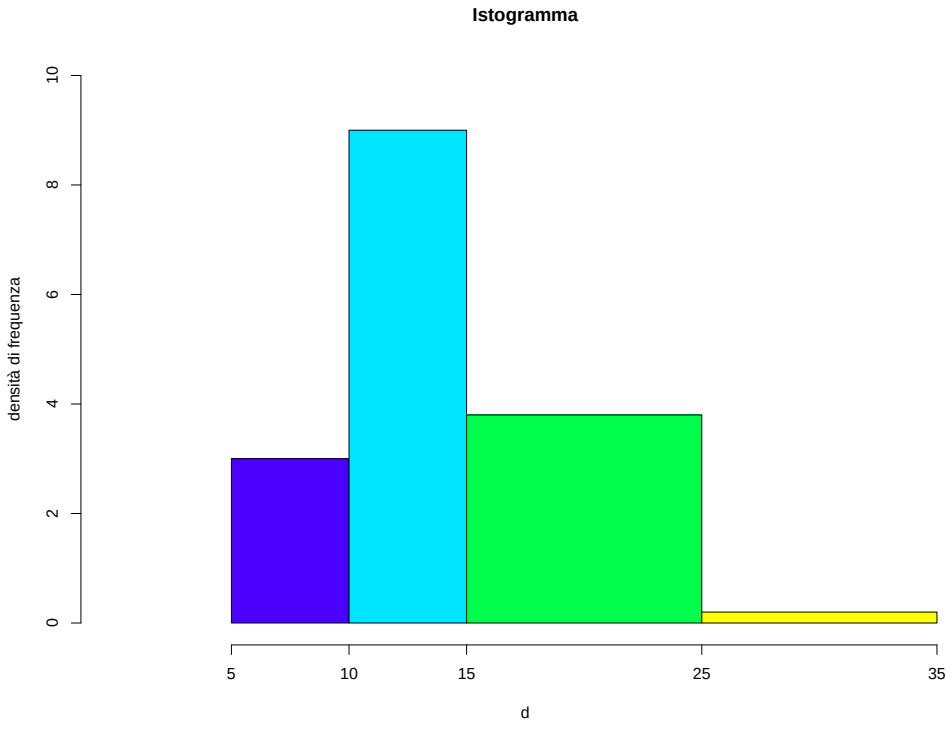
b. Definire e individuare la classe modale.

• • •

Soluzione.

a. Calcoliamo innanzi tutto le ampiezze delle classi e le densità di frequenza.

Consumo (litri)	Frequenza	Ampiezza	Densità di Frequenza
$[5, 10)$	15	5	3
$[10, 15)$	45	5	9
$[15, 25)$	38	10	3.8
$[25, 35)$	2	10	0.2

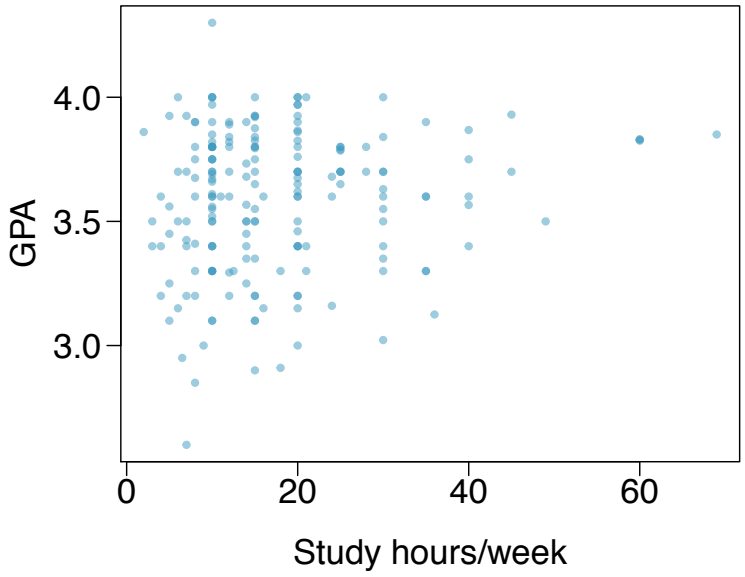


b. La classe modale è la classe alla quale risulta associata la massima densità di frequenza: in questo caso è la classe [10,15).

• • •

Esercizio 1.15. *Media dei voti e tempo di studio*
(dal libro di testo *OpenIntro Statistics* di Diez et al., es. 1.9)

È stata condotta un'indagine su 218 studenti della Duke University che hanno frequentato un corso di statistica di base nella primavera del 2012. Tra le molte altre domande, gli studenti sono stati interrogati sulla loro media dei voti (*GPA*) e sul numero di ore di studio settimanali (*Study hours/week*). Il seguente grafico a dispersione sotto mostra la relazione tra le due variabili.



a. Quale è la variabile esplicativa e quale è la variabile risposta?

b. Descrivere la relazione tra le due variabili. Mettere in evidenza osservazioni anomale, se ci sono.

c. Si tratta di un esperimento o uno studio osservazionale?

d. Possiamo concludere che all'aumentare del numero di ore di studio aumenta la media dei voti?

• • •

Soluzione.

a. La variabile esplicativa è il numero di ore di studio settimanali mentre la variabile risposta è la media dei voti.

b. C'è una relazione leggermente positiva tra le due variabili. Uno studente ha una media superiore a 4.0, quindi, si tratta di un errore. Ci sono anche alcuni studenti che riportano un numero di ore di studio settimanale inusualmente alto (60 e 70 ore/settimana). Inoltre, la variabilità della variabile media dei voti sembra essere maggiore per gli studenti che studiano meno rispetto a quelli che studiano di più. Poiché aumenta la dispersione al crescere del numero di ore di studio, è difficile valutare la forza della relazione e anche la variabilità su diversi numeri di ore di studio.

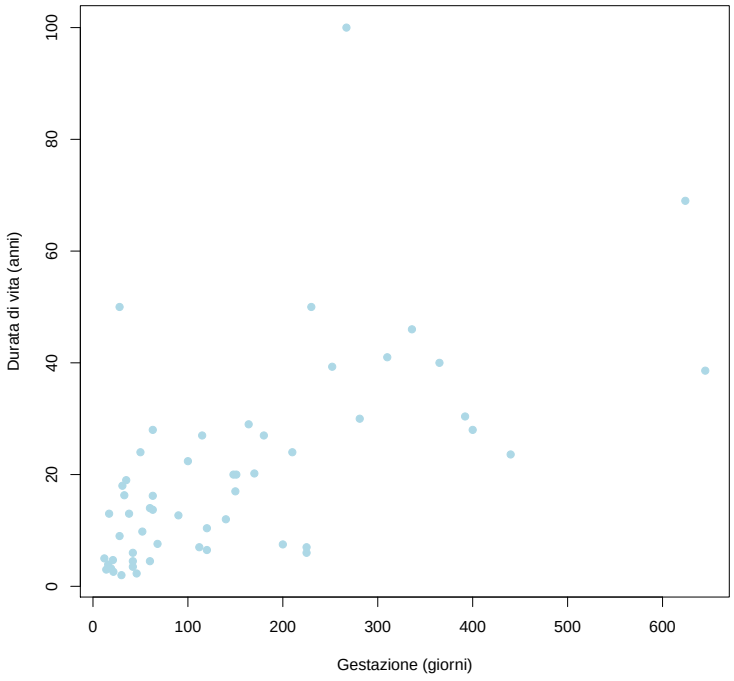
c. Si tratta di uno studio osservazionale

d. Proprio perché si tratta di uno studio osservazionale, non si può stabilire una relazione causale tra ore di studio e media dei voti.

• • •

Esercizio 1.16. *Vita dei mammiferi*
(dal libro di testo *OpenIntro Statistics* di Diez et al., es. 1.23)

Sono stati raccolti dei dati che riguardano la durata di vita (anni) e la durata della gestazione (giorni) per 62 mammiferi. Si risponda alle seguenti domande, in base al grafico a dispersione della durata di vita rispetto ai giorni di gestazione, sotto riportato:



a. Che tipo di associazione c'è tra durata della vita e durata della gestazione?

b. Che tipo di associazione ci si potrebbe aspettare se gli assi del plot fossero invertiti?

c. La durata di vita e la durata di gestazione sono indipendenti? Motivare la risposta.

• • •

Soluzione.

a. C'è un'associazione positiva: i mammiferi con periodi di gestazione più lunghi tendono a vivere più a lungo.

b. L'associazione continuerebbe ad essere positiva.

c. No, non sono indipendenti, come argomentato al punto a).

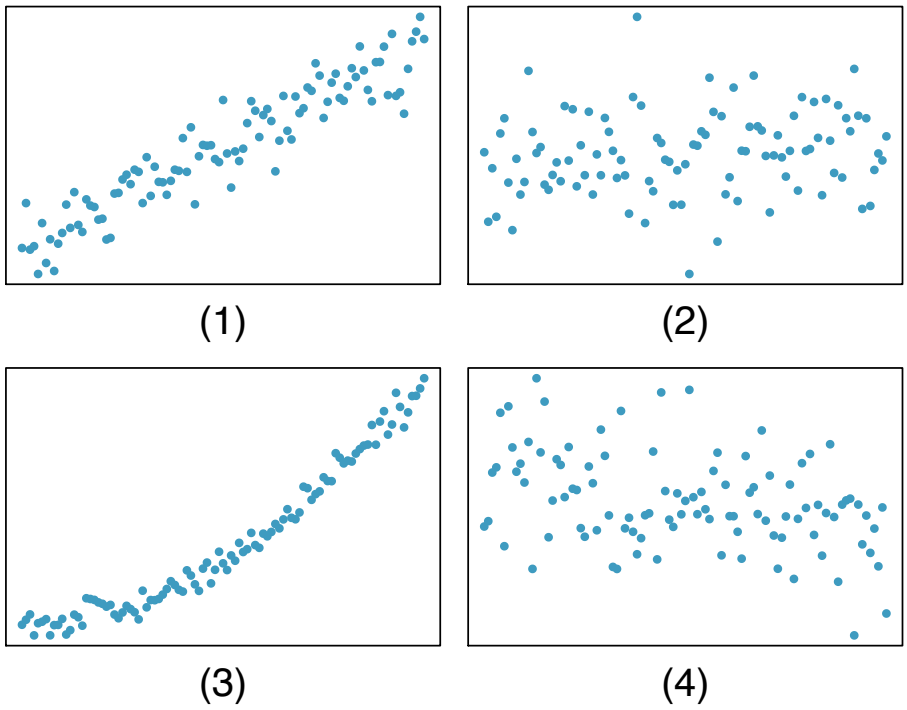
• • •

Esercizio 1.17. *Associazioni*
(dal libro di testo *OpenIntro Statistics* di Diez et al., es. 1.25)
Indicare quale dei seguenti grafici mostra

a. associazione positiva

b. associazione negativa

c. assenza di associazione



Determinare inoltre se le associazioni positive e negative sono lineari o non lineari.

• • •

Soluzione.

a. Il grafico (1) mostra associazione positiva lineare mentre il grafico (3) positiva non lineare.

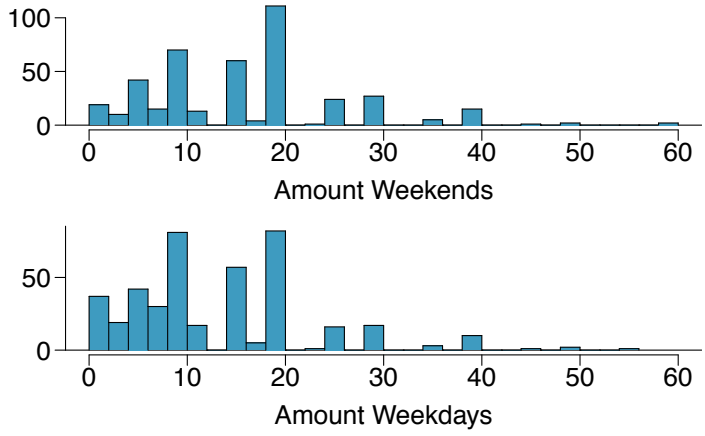
b. Il grafico (4) mostra una possibile lieve associazione negativa (non lineare) dovuta principalmente ai punti presenti nella parte destra del plot.

c. Il grafico (2) indica assenza di associazione.

• • •

Esercizio 1.18. *Abitudine al fumo tra i cittadini UK, parte I*
(dal libro di testo *OpenIntro Statistics* di Diez et al., es. 1.29)

È stata condotta un'indagine per studiare l'abitudine al fumo dei residenti UK. Di seguito sono riportati gli istogrammi relativi alle distribuzioni di numero di sigarette fumate durante i giorni della settimana (*amount weekdays*) e durante il fine settimana (*amount weekends*), escludendo i non fumatori. Descrivere le due distribuzioni e confrontarle.



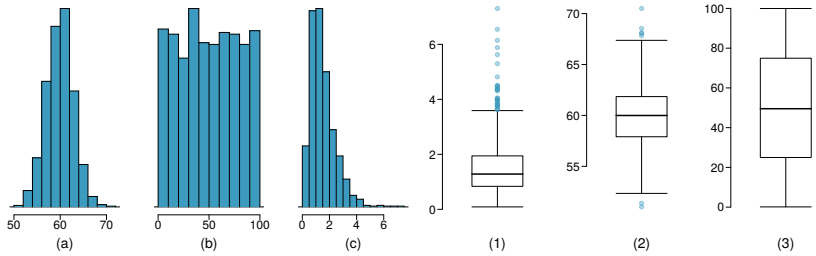
• • •

Soluzione. Entrambe le distribuzioni sono asimmetriche a destra e bimodali: una moda in corrispondenza di 10 sigarette e l'altra di 20 sigarette; ciò è dovuto al fatto che gli intervistati tendono a rispondere arrotondando a mezzo pacchetto o un pacchetto intero. La mediana di ciascuna distribuzione è tra 10 e 15 sigarette. In entrambi il range interquartile ha un'ampiezza intorno a 10-15. Ci sono delle osservazioni anomale in corrispondenza di 40 sigarette al giorno. Inoltre, sembra che coloro che fumano solo poche sigarette (da 0 a 5) fumano di più durante la settimana che durante il fine settimana.

• • •

Esercizio 1.19. *Istogrammi e boxplot*
(dal libro di testo *OpenIntro Statistics* di Diez et al., es. 1.37)

Descrivere le tre distribuzioni degli istogrammi riportati di seguito e associare ciascun istogramma al boxplot corrispondente.



• • •

Soluzione.

a. Distribuzione unimodale, simmetrica, centrata intorno al valore 60 con una standard deviation approssimativamente pari a 3. Il boxplot corrispondente è il numero 2.

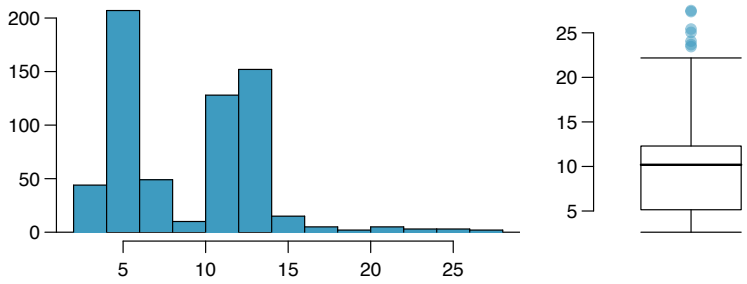
b. Distribuzione simmetrica e approssimativamente uniforme tra 0 e 100. Il boxplot corrispondente è il numero 3.

c. Distribuzione asimmetrica a destra, unimodale, centrata attorno al valore 1.5 con la maggior parte delle osservazioni tra 0 e 3 e una frazione molto piccola di osservazioni al di sopra di 5.

• • •

Esercizio 1.20. *Istogrammi e boxplot*
(dal libro di testo *OpenIntro Statistics* di Diez et al., es. 1.39)

Confrontare i due grafici riportati sotto. Quali caratteristiche della distribuzione si possono rilevare dall'istogramma e non dal boxplot? Quali caratteristiche sono evidenti nel boxplot e non nell'istogramma?



• • •

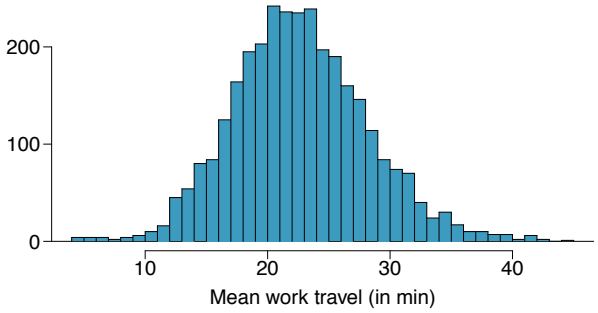
Soluzione. L'istogramma mostra che la distribuzione è bimodale, ciò non si può rilevare dal boxplot. Il boxplot invece permette di identificare in modo più preciso le osservazioni anomale.

• • •

Esercizio 1.21. *Tempi di pendolarismo, parte I.*
(dal libro di testo *OpenIntro Statistics* di Diez et al., es. 1.43)

L'istrogramma riportato sotto è relativo alla distribuzione dei tempi medi di pendolarismo (*mean work travel*) in 3,143 contee US nel 2010. Descrivere

la distribuzione e discutere se una trasformazione logaritmica può essere indicata per questi dati.



• • •

Soluzione. La distribuzione è unimodale e simmetrica con media pari a circa 25 minuti e deviazione standard pari a circa 5 minuti. Non sembra esserci nessuna contea con tempi particolarmente alti o bassi. Poichè la distribuzione è già unimodale e simmetrica, una trasformazione logaritmica non è necessaria.

• • •

1.5 Moda, media, mediana e quantili

Esercizio 1.22.

I dati seguenti riguardano il tempo impiegato per prepararsi al mattino:
52 44 43 44 40 29 31 39 35 39

a. Di che tipo di carattere si tratta?

b. Calcolare la moda di questa distribuzione;

c. Calcolare la media di questa distribuzione;

d. Calcolare la mediana;

e. Calcolare il primo e il terzo quartile di questa distribuzione.

• • •

Soluzione.

a. Il tempo impiegato per prepararsi è un carattere quantitativo continuo.

b. Costruendo la tabella di frequenza corrispondente alla distribuzione unitaria dei tempi, ci accorgiamo che le modalità 39 e 44 si presentano entrambe due volte (le altre tutte una volta), quindi la distribuzione ha due mode: 39 e 44.

c. Calcoliamo la media aritmetica:

$$\bar{x} = \frac{52 + 44 + 43 + 44 + 40 + 29 + 31 + 39 + 35 + 39}{10} = 39.6$$

d. Per calcolare la mediana, innanzi tutto ordiniamo le 10 osservazioni disponibili:

29 31 35 39 39 40 43 44 44 52

poi, dal momento che $n = 10$ è pari, consideriamo le osservazioni che occupano le posizioni $n/2$ e $n/2 + 1$, cioè rispettivamente 39 e 40 e ne calcoliamo la semisomma. La mediana è quindi 39.5.

e. Per calcolare il primo quartile, consideriamo la prima metà della distribuzione (costituita dalle prime 5 osservazioni) e ne calcoliamo la mediana:

Q1 = 35

Dopodichè ripetiamo lo stesso procedimento sulla seconda metà della distribuzione e otteniamo

Q3 = 44

• • •

Esercizio 1.23.

Di seguito viene riportata la distribuzione dei rendimenti del 2003 di 9 fondi comuni specializzati in aziende di piccole dimensioni:

37.3 39.2 44.2 44.5 53.8 56.6 59.3 62.4 66.5

a. Di che tipo di carattere si tratta?

b. Di che tipo di distribuzione si tratta?

c. Calcolare la moda di questa distribuzione;

d. Calcolare la media;

e. Calcolare la mediana.

• • •

Soluzione.

a. Si tratta di un carattere quantitativo continuo.

b. La distribuzione riportata è una distribuzione per unità.

c. In questo caso la moda della distribuzione non è definita in quanto ogni unità presenta una modalità distinta dalle altre, quindi ciascuna modalità si presenta con frequenza 1.

d. La media è pari a

$$\bar{x} = \frac{37.3 + 39.2 + 44.2 + 44.5 + 53.8 + 56.6 + 59.3 + 62.4 + 66.5}{9} = 51.53$$

e. Per calcolare la mediana innanzi tutto ordiniamo le 9 osservazioni disponibili:

37.3 39.2 44.2 44.5 53.8 56.6 59.3 62.4 66.5

poi, dal momento che $n = 9$ è dispari, la mediana è definita come l'osservazione che occupa la posizione $(n + 1)/2 = 5$, ovvero 53.8.

• • •

Esercizio 1.24. *Quanto si paga per avere accesso ad Internet?*

Di seguito sono riportate gli importi (in dollari) relativi alle bollette mensili pagate da un campione casuale di 50 utenti di provider commerciali di Internet nell'agosto del 2002:

20	40	22	22	21	21	20	10	20	20
20	13	18	50	20	18	15	8	22	26
22	10	20	22	22	21	15	23	30	12
9	20	40	22	29	19	15	20	20	20
20	15	19	21	14	22	21	35	20	22

a. Di che carattere si tratta?

b. Costruire la distribuzione in classi di questo carattere, utilizzando le seguenti classi: $(7.96, 18.5]$, $(18.5, 29]$, $(29, 39.5]$ e $(39.5, 50]$;

c. Determinare la classe modale.

• • •

Soluzione.

a. Si tratta di un carattere quantitativo continuo.

b. La distribuzione di frequenza in classi è

Bollette (dollari)	Frequenza assoluta	Ampiezza	Densità
(7.96,18.5]	13	10.54	1.23
(18.5,29]	32	10.50	3.05
(29,39.5]	2	10.50	0.19
(39.5,50]	3	10.50	0.29

c. La classe modale è la classe alla quale è associata la massima densità di frequenza (notare che le ampiezze delle classi non sono tutte uguali), ovvero la classe (18.5,29].

• • •

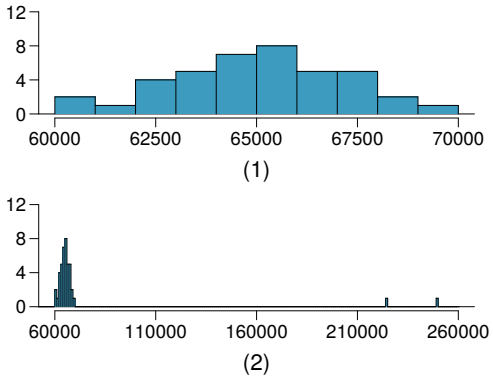
Esercizio 1.25. *Robustezza*

(dal libro di testo *OpenIntro Statistics* di Diez et al., es. 1.41)

Il primo istogramma rappresentato di seguito mostra la distribuzione dei redditi annui di 40 clienti di un bar. Due nuovi clienti hanno rispettivamente un reddito annuo di 225000 $ e 250000 $. Il secondo istogramma mostra la nuova distribuzione e la tabella riporta alcune statistiche riassuntive.

a. Quale indice rappresenta meglio il tipico reddito dei 42 clienti? La media o la mediana? Cosa rivela questa osservazione rispetto alla robustezza di queste due misure?

b. Quale indice rappresenta meglio la variabilità nella distribuzione del reddito dei 42 clienti? La deviazione standard o il range interquartilico? Cosa rivela questa osservazione rispetto alla robustezza di queste due misure?



	(1)	(2)
n	40	42
Min.	60,680	60,680
1st Qu.	63,620	63,710
Median	65,240	65,350
Mean	65,090	73,300
3rd Qu.	66,160	66,540
Max.	69,890	250,000
SD	2,122	37,321

• • •

Soluzione.

a. La mediana è l'indice più robusto; la media è fortemente influenzata dalle due osservazioni estreme.

b. Il range interquartilico è l'indice più robusto; la deviazione standard, come la media, è fortemente influenzata dalle due osservazioni estreme.

• • •

Esercizio 1.26. *Mediana e range interquartile*
(dal libro di testo *OpenIntro Statistics* di Diez et al., es. 1.33)

Per ciascuna parte, confrontare le distribuzioni (1) e (2) basandosi su mediane e range interquartile. Non è necessario calcolare queste statistiche, ma semplicemente confrontarle, spiegare il proprio ragionamento.

a. (1) 3, 5, 6, 7, 9
(2) 3, 5, 6, 7, 20

b. (1) 3, 5, 6, 7, 9
(2) 3, 5, 8, 7, 9

c. (1) 1, 2, 3, 4, 5
(2) 6, 7, 8, 9, 10

d. (1) 0, 10, 50, 60, 100
(2) 0, 100, 500, 600, 1000

• • •

Soluzione.

a. Entrambe le distribuzioni hanno la stessa mediana e stesso range interquartile

b. La seconda distribuzione ha una mediana più alta e un range interquartile più alto

c. La seconda distribuzione ha una mediana più alta e stesso range interquartile.

d. La seconda distribuzione ha una mediana più alta e un range interquartile più ampio.

• • •

1.6 Gli indici di variabilità

Esercizio 1.27.

Riprendendo l'Esercizio 1.23, consideriamo i rendimenti del 2003 per i fondi comuni ad alto rischio specializzati in aziende di piccole dimensioni.

a. Definire i **5 numeri di sintesi** della distribuzione;

b. Disegnare il boxplot della distribuzione;

c. Calcolare la varianza e la deviazione standard della distribuzione;

d. Calcolare il coefficiente di variazione.

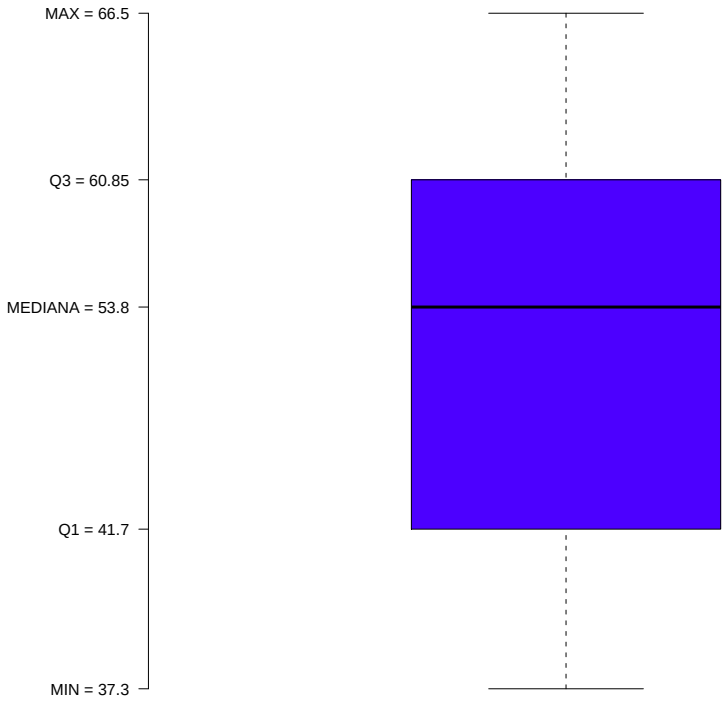
• • •

Soluzione.

a. I cinque numeri di sintesi sono:

Minimo: 37.3, Primo Quartile: 41.7, Mediana: 53.8, Terzo Quartile: 60.85, Massimo: 66.5.

b. Ecco il boxplot corrispondente:



c. Ricordando che la media è pari a $\bar{x} = 51.53$, calcoliamo la varianza, ovvero:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = 111.395$$

La deviazione standard è quindi

$$s = \sqrt{s^2} = \sqrt{111.395} = 10.554$$

d. Il coefficiente di variazione è

$$CV = \frac{s}{\bar{x}} \cdot 100 = 0.205 \cdot 100 = 20.5\%$$

• • •

Esercizio 1.28.

Il direttore operativo di un'azienda di consegna di pacchi sta pensando all'acquisto di un nuovo parco di autocarri. Quando i pacchi sono depositati negli autocarri in attesa della consegna, si deve tenere conto di 2 vincoli principali: il peso (in chilogrammi) e il volume (in metri cubi) di ciascun pacco. Si considera un campione di 200 pacchi per cui si osserva un peso medio di 9 Kg, con uno scarto quadratico medio di 1.5 Kg, e un volume medio di 2.7 metri cubi, con uno scarto quadratico medio di 0.6 metri cubi. Come è possibile confrontare la variabilità del peso e del volume?

• • •

Soluzione.

Peso e volume sono espressi in unità di misura diverse: si deve quindi prendere in considerazione la variabilità relativa delle osservazioni. Per il peso, il coefficiente di variazione è

$$CV_P = \frac{s}{\bar{x}} \cdot 100 = \frac{1.5}{9} \cdot 100 = 16.67\%$$

per il volume è pari a

$$CV_V = \frac{s}{\bar{x}} \cdot 100 = \frac{0.6}{2.7} * 100 = 22.22\%$$

Pertanto rispetto alla media, il volume dei pacchi più variabile del peso.

• • •

Esercizio 1.29.

Consideriamo la distribuzione in classi ricavata all'Esercizio 1.24

Classi	n_i
(7.96,18.5]	13
(18.5,29]	32
(29,39.5]	2
(39.5,50]	3

Calcolare varianza e deviazione standard di questo carattere.

• • •

Soluzione.

Per calcolare la varianza abbiamo bisogno delle quantità riportate nella seguente tabella ($\tilde{x}_i$ indica il valore centrale della classe i-esima):

Classi	n_i	$\tilde{x}_i$	f_i	$\tilde{x}_i^2$	$\tilde{x}_i^2 f_i$
$(7.96, 18.5]$	13	13.23	0.26	175.0329	45.51
$(18.5, 29]$	32	23.75	0.64	564.0625	361
$(29, 39.5]$	2	34.25	0.04	1173.0625	46.92
$(39.5, 50]$	3	44.75	0.06	2002.5625	120.15
totale	50		1		573.58

La media è pari a

$$\bar{x} = (13.23 \cdot 0.26) + (23.75 \cdot 0.64) + (34.25 \cdot 0.04) + (44.75 \cdot 0.06) = 22.69$$

e quindi la varianza è

$$s^2 = \frac{n}{n-1}\left(\sum_i \tilde{x}_i^2 f_i - \bar{x}^2\right) = \frac{50}{49}\left(573.58 - (22.69)^2\right) = 59.94$$

e la deviazione standard

$$s = \sqrt{s^2} = \sqrt{59.94} = 7.74$$

• • •

Esercizio 1.30.

Riprendendo dall'Esercizio **??** i dati sui tempi di funzionamento di un campione di batterie,

a. Calcolare il valore di opportuni indici di posizione e di variabilità.

b. Come variano gli indici di posizione e di variabilità se il tempo di funzionamento è espresso in settimane (assumendo, per approssimazione, che ciascun mese sia composto esattamente da quattro settimane)?

c. Se si utilizza il coefficiente di variazione per misurare la variabilità, vi è differenza se si utilizza un'unità di misura diversa (mesi o settimane)? Motivare la risposta.

• • •

Soluzione.

a. Calcoliamo innanzi tutto le quantità riportate in tabella:

Durata (mesi)	Frequenza	$\tilde{x}_i$	f_i	$\tilde{x}_i^2$	$\tilde{x}_i^2 f_i$
(1,3]	10	2	0.10	4	0.4
(3,6]	42	4.5	0.43	20.25	8.71
(6,12]	38	9	0.39	81	31.59
(12,24]	8	18	0.08	324	25.92
totale	98		1		66.62

La media è

$$\bar{x} = (2 \cdot 0.1) + (4.5 \cdot 0.43) + (9 \cdot 0.39) + (18 \cdot 0.08) = 7.085$$

e la varianza

$$s^2 = \frac{n}{n-1}\left(\sum_i \tilde{x}_i^2 f_i - \bar{x}^2\right) = \frac{98}{97}\left(66.62 - (7.085)^2\right) = 16.59$$

e la deviazione standard

$$s = \sqrt{s^2} = \sqrt{16.59} = 4.07$$

b. Esprimere il tempo in settimane anzichè in mesi significa cambiare unità di misura.

Per le proprietà della media (linearità) sappiamo che per calcolare la durata media in settimane è sufficiente moltiplicare la durata media in mesi per l'opportuno coefficiente (4), ovvero:

$$\bar{x}_{settimane} = \bar{x}_{mesi} \cdot 4 = 7.085 \cdot 4 = 28.34$$

Per quanto riguarda la varianza abbiamo invece:

$$s^2_{settimane} = s^2_{mesi} \cdot 4^2 = 16.59 \cdot 16 = 265.44$$

c. In entrambi i casi il coefficiente di variazione è pari a

$$CV = s/\bar{x} \cdot 100 = 4.07/7.085 \cdot 100 = 0.57 \cdot 100 = 57\%$$

perchè non dipende dall'unità di misura.

• • •

Esercizio 1.31. *Abitudine al fumo tra i cittadini UK, parte II*
(dal libro di testo *OpenIntro Statistics* di Diez et al., es. 1.31)

Si consideri un campione casuale di 5 fumatori per i quali sono state rilevate le variabili riportate nella seguente tabella:

Sesso	Età	Stato civile	Reddito lordo	quantità (weekend)	quantità (giorni feriali)
F	51	Coniugato/a	2.600-5.200	20	20
M	24	Celibe/Nubile	10.400-15.600	20	15
F	33	Coniugato/a	10.400-15.600	20	10
F	17	Celibe/Nubile	2.600-5.200	20	15
F	76	Vedovo/a	2.600-5.200	20	20

a. Determinare la quantità media di sigarette fumate nei giorni feriali e nei weekend dai 5 fumatori.

b. Determinare la deviazione standard della quantità di sigarette fumate nei giorni feriali e nei weekend dai 5 fumatori. La variabilità è maggiore nei weekend o nei giorni feriali?

• • •

Soluzione.

a. $\bar{x}_{weekend} = 20$; $\bar{x}_{feriali} = \frac{80}{5} = 16$.

b. $s_{weekend} = 0$; $s_{feriali} = 4.18$. La variabilità è dunque maggiore nei giorni feriali.

• • •

1.7 Proprietà delle medie

Esercizio 1.32.

A 10 studenti universitari viene chiesto il numero di esami superati in un anno. La distribuzione unitaria è la seguente:
4 0 7 1 5 5 0 2 0 12

a. Calcolare il numero medio di esami;

b. Se alle informazioni fornite dai 10 studenti si aggiungono quelle di altri 20 studenti, la media aritmetica risulta pari a 5. Determinare la media del numero di esami superati dal secondo gruppo di 20 studenti.

• • •

Soluzione.

a. Calcoliamo la media aritmetica del numero di esami:

$$\bar{x}_A = \frac{4+0+7+1+5+5+0+2+0+12}{10} = 3.6$$

b. Se indichiamo con $\bar{x}_{TOT} = 5$ il numero medio di esami del campione complessivo, con $\bar{x}_A$ il numero medio di esami nel primo gruppo di numerosità $n_A = 10$ e con $\bar{x}_B$ il numero medio di esami nel secondo gruppo di numerosità $n_B = 20$, otteniamo:

$$\bar{x}_{TOT} = \frac{n_A \cdot \bar{x}_A + n_B \cdot \bar{x}_B}{n_A + n_B} = \frac{10 \cdot 3.6 + 20 \cdot \bar{x}_B}{10 + 20} = 5$$

In questo caso però conosciamo la media complessiva e dalla formula precedente possiamo ricavare quella del secondo gruppo in questo modo:

$$\bar{x}_B = \frac{5 \cdot 30 - 10 \cdot 3.6}{20} = 5.7$$

• • •

Esercizio 1.33.

Un uomo d'affari nell'ultimo mese è andato in viaggio a Londra per 10 volte. Il costo medio del biglietto aereo è 120, con una varianza pari a 7. Se l'uomo avesse prenotato tutti i voli da Londra, sapendo che il cambio è 1 euro = 0.87 sterline e che c'è un costo fisso della commissione pari a una sterlina per ciascun cambio, quanto avrebbe speso? Calcolare il costo medio in sterline e la varianza.

• • •

Soluzione.

Per la proprietà di linearità della media otteniamo che:

$$\bar{x}_{STERLINE} = \bar{x}_{EURO} \cdot 0.87 + 1 = 105.4$$

Per quanto riguarda la varianza, sappiamo invece che:

$$s^2_{STERLINE} = 0.87^2 s^2_{EURO} = 5.298$$

• • •

Esercizio 1.34. *Esame di recupero*
(dal libro di testo *OpenIntro Statistics* di Diez et al., es. 1.27)

In una classe di 25 studenti, 24 hanno svolto un esame in classe e un solo studente è stato sottoposto a una prova di recupero il giorno successivo. Il professore ha valutato il primo blocco di esami, per i quali il punteggio medio è risultato di 74 punti con una deviazione standard di 8.9 punti. La prova di recupero dello studente del giorno dopo ha riportato un punteggio di 64 punti.

a. Il punteggio del nuovo studente fa aumentare o diminuire il punteggio medio?

b. Quale è la nuova media?

c. Il punteggio del nuovo studente fa aumentare o diminuire la deviazione standard?

• • •

Soluzione.

a. Il punteggio del nuovo studente fa diminuire il punteggio medio.

b. La media complessiva si ottiene come media ponderata della media dei 24 studenti e del nuovo punteggio con pesi pari a 24 e 1 rispettivamente: $(24 * 74 + 1 * 64)/(24 + 1) = 73.6$.

c. Il punteggio del nuovo studente fa aumentare la deviazione standard, perchè dista dalla media precendente più di una deviazione standard.

• • •

1.8 I numeri indice

Esercizio 1.35.

Nella tabella sono riportate le quantità di acciaio di prima fabbricazione prodotte in Italia negli anni del periodo 1976-1981:

Anni	Acciaio di prima fabbricazione
1976	23447
1977	23334
1978	24283
1979	24250
1980	26501
1981	24777

a. Calcolare il numero indice semplice con base 1976 per la produzione di acciaio nell'anno 1977 (ossia ${}_{1976}I_{1977}$) e intepretare tale indice.

b. Calcolare il numero indice semplice con base 1976 per la produzione di acciaio nell'anno 1980 (ossia ${}_{1976}I_{1980}$) e intepretare tale indice.

• • •

Soluzione.

a. ${}_{1976}I_{1977} = \frac{23334}{23447}100 = 0.994100 = 99.4\%$

La produzione di acciaio nel 1977 ha subito un lievissimo decremento rispetto a quella dell'anno precedente: il decremento assoluto rispetto all'anno precedente è pari a 100-99.4=0.6%.

b. ${}_{1976}I_{1980} = \frac{26501}{23447}100 = 1.129100 = 112.9\%$

La produzione di acciaio nel 1980 ha subito un incremento rispetto alla produzione di acciaio nel 1976; l'incremento è del 12.9%.

• • •

Esercizio 1.36.

L'Indice dei prezzi alla produzione dei prodotti industriali (base 2005 - Istat) a luglio e ad agosto 2009 è stato pari, rispettivamente, a 107.3 e 107.9. Qual è stato l'incremento percentuale che il fenomeno ha subito tra i due mesi?

• • •

Soluzione.

La variazione percentuale è stata pari a:

$$\frac{107.9 - 107.3}{107.3}100 = +0.6$$

cioè, rispetto al valore di luglio, ad agosto c'è stato un incremento dello 0.6%.

• • •

Esercizio 1.37.

Nel 2008 La variazione percentuale, calcolata rispetto all'anno precedente, del Prodotto Interno Lordo italiano è stata pari a -1.04%. Sapendo che nel 2008 il PIL valeva 1276439 milioni di euro, qual era il valore del PIL nel 2007?

• • •

Soluzione.

Indichiamo con x il valore del PIL nel 2007. Sappiamo che

$$\frac{1276439 - x}{x} 100 = -1.04$$

da cui

$$x = \frac{1276439}{-1.04/100 + 1} = 1289853 \text{ milioni di euro.}$$

• • •

Esercizio 1.38.

Nella tabella seguente sono riportati i tassi d'inflazione (cioè le variazioni percentuali rispetto allo stesso mese dell'anno precedente) registrati ad agosto 2009 in alcune città.

città	variazione %
Torino	0.0
Milano	-0.6
Trieste	+1.8
Roma	+0.2
Reggio Calabria	+1.3
Bologna	-0.5
Firenze	-0.5

Quali informazioni possiamo trarre dal confronto tra i dati? Quale è stata la città col più elevato livello dei prezzi?

• • •

Soluzione.

Rispetto ad agosto 2008, tra le città considerate, Trieste è quella in cui i prezzi hanno subito una maggiore accelerazione, Milano quella in cui i

prezzi sono diminuiti con più elevata velocità, mentre a Torino non è stata riscontrata alcuna variazione.

Non siamo in grado di rispondere alla seconda domanda, poichè i dati disponibili danno informazioni solo sul cambiamento che il fenomeno prezzi ha subito tra i due mesi, non sul livello.

• • •

Capitolo 2

Probabilità

Esercizio 2.1. *Vero o Falso*
Stabilire se le affermazioni seguenti sono vere o false e motivare in ciascun caso la risposta.

(a) Se una moneta regolare viene lanciata molte volte, e negli ultimi 8 lanci dà sempre T (testa), allora la probabilità che nel prossimo lancio dia T sarà leggermente inferiore a 0.5.

(b) Dato un mazzo di carte "italiano", ovvero con 40 carte e quattro semi (Spade, Denari, Coppe, Bastoni), supponiamo di estrarre una carta. Gli eventi A={ viene estratta una figura} e B={viene estratta una carta di Spade o Denari} sono mutuamente indipendenti

(c) Nella stessa situazione di prima, gli eventi A (come sopra) e C={viene estratto un Asso} sono mutuamente incompatibili.

Esercizio 2.2. *Roulette*
Il gioco della roulette consiste in una ruota con 37 slots, ovvero i numeri da 0 a 36, estremi inclusi Ad ogni giro di roulette, una pallina si sistemerà in uno dei 37 slots: gli slots sono ugualmente probabili. Lo slot 0 è verde, gli altri 36 sono 18 di colore rosso (R) e 18 di colore nero (N)

(a) Si osservano tre lanci consecutivi della pallina. Tutte e tre le volte la pallina si ferma su uno slot rosso (R). Qual è la probabilità che nel lancio successivo si ottenga ancora uno slot R?.

(b) Si osservano 300 lanci consecutivi della pallina. Tutte e tre le volte la pallina si ferma su uno slot rosso (R). Qual è la probabilità che nel lancio successivo si ottenga ancora uno slot R?.

(c) Hai risposto alle due precedenti domande con la stessa sicurezza? Perché? o perché no? Spiega.

Esercizio 2.3. *Quattro diversi giochi, un solo vincitore.*

Qui di sotto vengono riportate quattro diverse versioni dello stesso gioco. Your archnemisis gets to pick the version of the game, and then you get to choose how many times to flip a coin: 10 times or 100 times. Identify how many coin flips you should choose for each version of the game. Explain your reasoning.

(a) If the proportion of heads is larger than 0.60, you win $1.

(b) If the proportion of heads is larger than 0.40, you win $1.

(c) If the proportion of heads is between 0.40 and 0.60, you win $1.

(d) If the proportion of heads is smaller than 0.30, you win $1.

Esercizio 2.4. *Backgammon.*

Il backgammon si gioca su un tavolo da due giocatori che muovono dei pezzi a seconda dei risultati di due dadi che lanciano a turno. Il giocatore vince se riesce a togliere tutti i suoi pezzi dal tavolo, e per fare questo è di solito più favorevole che escano, dai dadi, numeri alti. Stai giocando a backgammon con un tuo amico e, nei tuoi primo e secondo turni, ottieni due volte un doppio 6. Il tuo amico ottiene invece un doppio 3, sia nel primo lancio che nel secondo. A questo punto il tuo amico si lamenta e sostiene che

stai barando perchè ottenere due volte di fila il oppio è molto improbabile e lo è molto di più del doppio 3.

Usando argomenti di probabilità, cerca di dimostrare al tuo amico che i suoi due doppi 3 hanno la stessa probabilità dei tuoi doppi 6.

Esercizio 2.5. *Lanci di monete*

Se lanci una moneta regolare 10 volte, qual è la probabilità di ottenere

(a) 10 volte Croce (C)?

(b) 10 volte Testa (T)?

(c) Almeno una T?

Esercizio 2.6. *Dadi.*

Si lancia una coppia di dadi regolari; qual è la probabilità di ottenere

(a) una somma pari a 1?

(b) una somma pari a 5?

(c) una somma pari a 12?

Esercizio 2.7. *Swing voters.*

In un'indagine demoscopica campionaria del 2014 un istituto privato a chiesto a 2373 persone scelte a caso la loro affiliazione politica. Vi erano 4 possibilità: Centro-sinistra (CS), Centro-destra (CD), Movimento Cinque Stelle (M5S) o altro (A). Inoltre si chiedeva agli intervistati se si considerassero degli "swing voters" (SV),ovvero persone che fino all'ultimo momento sarebbero state incerte su quale schieramento votare. Tra i risultati, riportiamo che il 12% dei rispondenti si dichiarava (A), mentre il 22% si dichiara (SV); inoltre il 7% si identificava in entrambe le categorie sopra citate.

(a) Gli eventi A e SV sono mutuamente incompatibili?

(b) Disegna un diagramma di Venn per riassumere le relazioni tra gli eventi.

(c) Che percentuale di elettori vota A ma non sono SV?

(d) Che percentuale di elettori vota A oppure è SV?

(e) Che percentuale di elettori non vota A E non è nemmeno SV?

(f) Gli eventi A ed SV possono essere considerati statisticamente indipendenti?

Esercizio 2.8. *Povertà e Madrelingua.*

L'indagine "American Community Survey" (ACS) è una indagine che il Bureau of Census negli USA effettua ogni anno per fornire dati utili alle varie comunità per effettuare i loro investimenti e organizzare i propri servizi. Nella edizione del 2010, l'ACS ha stimato che circa il 14.6% degli Americani vive sotto la soglia di povertà, che il 20.7% ha come lingua principale una diversa dall'inglese, e che il 4.2% degli Americani ricade in entrambe le suddette categorie.

(a) Gli eventi A=\{Vivere sotto la soglia di povertà\} e B=\{avere una madrelingua diversa dall'inglese\} sono disgiunti?

(b) Traccia un diagramma di Venn che riassuma le informazioni circa gli eventi sopra descritti e le loro probabilità.

(c) Quale percentuale di Americani vive sotto la soglia di povertà e parla inglese come madrelingua?

(d) Quale percentuale di Americani vive sotto la soglia di povertà e parla una lingua diversa dall'inglese come madrelingua?

(e) Quale percentuale di Americani vive sopra la soglia di povertà e parla inglese come madrelingua?

(f) Gli eventi A e B sono indipendenti?

Esercizio 2.9. *Indipendenza e incompatibilità.*

Nei successivi punti (a) e (b), chiarisci se gli eventi in questione sono indipendenti, incompatibili, o nessuna delle due cose.

(a) Tu e un altro studente scelto a caso nella tua classe, avete presso lo stesso voto all'esame di Matematica.

(b) Tu e lo studente con cui abitualmente prepari gli esami, avete presso lo stesso voto all'esame di Matematica.

(c) Se due eventi A e B si verificano contemporaneamente, devono per forza essere dipendenti? Spiega perché si, o perché no.

Esercizio 2.10. *"Provarci" allo scritto.*

In un esame scritto, con domande a risposta multipla, ci sono 5 domande e per ognuna ci sono 4 scelte possibili, (diciamo a, b, c, d). Cinzia non ha studiato per niente, e decide di rispondere a caso alle cinque domande. Qual è la probabilità che Cinzia:

(a) risponda bene solo alla quinta domanda?

(b) risponda correttamente a tutte le domande?

(c) risponda correttamente almeno ad una domanda?

Esercizio 2.11. *Assenze a scuola*

Nella contea di DeKalb, Georgia, USA, ogni anno vengono raccolti dati sulle assenze a scuola dei bambini delle scuole elementari. Dai dati emerge che, ogni anno, circa il 25% dei bambini si assenta un solo giorno nell'intero anno; il 15% si assenta per 2 giorni mentre il 28% dei bambini si assenta per 3 o più giorni.

(a) Qual è la probabilità che un bambino scelto a caso non abbia mancato nessun giorno di lezione?

(b) Qual è la probabilità che un bambino scelto a caso abbia mancato al più un giorno di lezione?

(c) Qual è la probabilità che un bambino scelto a caso abbia mancato almeno un giorno di lezione?

(d) Se una mamma ha due bimbi che frequentano le scuole nella contea di DeKalb, qual è la probabilità che nessuno dei due bimbi si sia mai assentato nell'anno? Sottolinea le assunzioni importanti che hai fatto per poter rispondere a questa domanda.

(e) Se una mamma ha due bimbi che frequentano le scuole nella contea di DeKalb, qual è la probabilità che entrambi i bimbi perdano almeno un giorno di scuola ciascuno? Anche in questo caso sottolinea le assunzioni necessarie per rispondere a questa domanda.

(f) Se nelle domande (d) ed (e), hai fatto delle assunzioni, pensi che queste siano ragionevoli? Discuti la questione. Se non hai fatto alcuna assunzione, ricontrolla le tue precedenti risposte.

Esercizio 2.12. *Distribuzioni dei voti*
Ogni riga della tabella che segue è una distribuzione di frequenze relative per una classe di studenti in America, dove il sistema di voti va da A ad F. Stabilisci, per ciascuna riga, se si tratta di distribuzioni ammissibili, e in ogni caso spiega le tue ragioni.

	VOTI				
(1)	0.3	0.3	0.3	0.2	0.1
(2)	0	0	1	0	0
(3)	0.3	0.3	0.3	0	0
(4)	0.3	0.5	0.2	0.1	-0.1
(5)	0.2	0.4	0.2	0.1	0.1
(6)	0	-0.1	1.1	0	0

Esercizio 2.13. *Peso e assicurazione sanitaria; I parte.*
Il Sistema di Sorveglianza sui Fattori di Rischio Comportamentali (BRFSS) è una grande indagine telefonica che viene effettuata negli USA ogni anno, per identificare i fattori di rischio nella popolazione adulta ed individuare potenziali trend nelle dinamiche sanitarie.

La tabella che segue si riferisce a 2 variabili: le condizioni di peso, misurate mediante l'indice di massa corporea, o "body mass index (BMI)" e la copertura assicurativa, che stabilisce se i vari rispondenti possedevano una assicurazione sanitaria oppure no.

		Condizioni di Peso			
		sovrappeso mai (BMI<25)	sovrappeso $25 \leq BMI < 30$	obeso BMI > 30	Totale
Copertura Sanitaria	SI	134801	141699	107301	383801
	NO	15098	15327	14412	44837
	Totale	149899	157026	121713	428638

Si estrae un individuo a caso.

(a) Qual è la probabilità che l'individuo sia sovrappeso e non abbia assicurazione sanitaria?

(b) Qual è la probabilità che l'individuo sia sovrappeso oppure non abbia assicurazione sanitaria?

2.1 Probabilità Condizionata

Esercizio 2.14. *Probabilità congiunte e condizionate.*
Siano A e B due eventi tali che

$$P(A) = 0.3; P(B) = 0.7.$$

(a) Puoi calcolare $P(A \cap B)$ conoscendo solamente $P(A)$ e $P(B)$?

(b) Assumiamo che gli eventi A e B siano indipendenti:
i) quanto vale $P(A \cap B)$?
ii) quanto vale $P(A \cup B)$?
iii) quanto vale $P(A|B)$?

(c) Se sapessimo anche che $P(A \cap B) = 0.1$, possiamo ancora dire che A e B sono eventi indipendenti?

(d) Se sapessimo anche che $P(A \cap B) = 0.1$, quanto vale $P(A|B)$?

Esercizio 2.15. *Nutella o Marmellata?*

Supponiamo che all'80% dei ragazzi italiani piaccia la marmellata, all'89% piaccia la Nutella, e al 78% piacciano entrambi. Dato che ad una persona scelta a caso piace la marmellata, qual è la probabilità che le piaccia anche la Nutella?

Esercizio 2.16. *Riscaldamento globale.*

In una indagine del 2010 effettuata negli USA dalla "Pew Research", vennero intervistati 1306 cittadini americani; la domanda era: "Da quanto hai letto, sentito, visto in TV, esiste secondo lei una solida evidenza che le temperature medie sulla Terra siano aumentate negli ultimi decenni oppure no?". La tabella che segue mostra la distribuzione delle risposte, classificate anche secondo le idee politiche degli intervistati. La tabella riporta le frequenze relative.

		Risposta			
		Terra più calda	Terra non più calda	Non so non rispondo	Totale
Partito/ Ideologia	Repubblicani	0.11	0.20	0.2	0.33
	Moderati REPUB.	0.06	0.06	0.01	0.13
	Moderati DEMOC.	0.25	0.07	0.02	0.34
	Democratici	0.18	0.01	0.01	0.20
	Totale	0.60	0.34	0.06	1.00

(a) Qual è la probabilità che una persona scelta a caso creda che la Terra si ora più calda oppure che sia un Democratico?

(b) Qual è la probabilità che una persona scelta a caso creda che la Terra si ora più calda dato che si tratta di un Democratico?

(c) Qual è la probabilità che una persona scelta a caso creda che la Terra si ora più calda dato che si tratta di un Repubblicano?

(d) Ti sembra che le risposte fornite dagli intervistati alla domanda sul riscaldamento siano collegate con le idee politiche degli stessi oppure no? Spiega perché.

(e) Qual è la probabilità che una persona scelta a caso sia un moderato Repubblicano dato che egli non crede al riscaldamento terrestre?

Esercizio 2.17. *Peso e assicurazione sanitaria; II parte.*
Nell'Esercizio 2.13 è stata introdotta una tabella di contingenza che riassume la relazione tra condizioni di peso - in termini di Body Mass Index, e possesso di assicurazione sanitaria per un campione di 428638 americani. La tabella che segue è equivalente alla precedente ma è espressa in frequenze relative.

		Condizioni di Peso			
		sovrappeso mai (BMI<25)	sovrappeso $25 \leq BMI < 30$	obeso BMI > 30	Totale
Copertura Sanitaria	SI	0.3145	0.3306	0.2503	0.8954
	NO	0.0352	0.0358	0.0336	0.1046
	Totale	0.3497	0.3664	0.2839	1.0

(a) Qual è la probabilità che un individuo scelto a caso sia obeso?

(b) Qual è la probabilità che un individuo scelto a caso sia obeso dato che egli possiede l'assicurazione sanitaria?

(c) Qual è la probabilità che un individuo scelto a caso sia obeso dato che egli NON possiede l'assicurazione sanitaria?

(d) Pensi che i due caratteri, ovvero la condizione di peso e il possesso di assicurazione sanitaria siano indipendenti oppure no? spiegare perché .

Esercizio 2.18. .
In una indagine del 2010, SurveyUSA chiese a 500 residenti a Los Angeles: *"Qual è il migliore hamburger della California de sud ?"* Le risposte possibili erano

- Five Guys Burgers
- In-N-Out Burger
- Fat Burger

- Tommy's Hamburgers
- Umami Burger
- Altro

Qui sotto viene riportata la distribuzione delle risposte, tenendo conto anche del genere dei rispondenti (Maschi=M, Femmine=F)

	Genere		
	M	F	Totale
Five Guys	5	6	11
In N Out	162	181	343
Fat	10	12	22
Tommy's	27	27	54
Umami	5	1	6
Altri	26	20	46
Non so	13	5	18
Totale	248	252	500

(a) Qual è la probabilità che un uomo scelto a caso preferisca In-N-Out?

(b) Qual è la probabilità che una donna scelta a caso preferisca In-N-Out?

(c) Qual è la probabilità che un uomo e una donna che escono insieme preferiscano entrambi In-N-Out? Annota qualunque assunzione tu faccia per rispondere a questa domanda e rifletti sul fatto che questa assunzione sia o meno ragionevole.

(d) Qual è la probabilità che una persona scelta a caso preferisca Umami oppure che quella persona sia donna?

Esercizio 2.19. *Accoppiamento ragionato (Assortative mating).*

L'accoppiamento ragionato si attua quando individui con genotipi e/o fenotipi simili si accoppiano più frequentemente di quanto farebbero mediante accoppiamenti casuali. Alcuni ricercatori hanno raccolto dati su questo tema, registrando il colore degli occhi di 204 coppie eterosessuali scandinave.

		Partner	femminile		
		Blu	Marrone	Verde	Totale
	Blue	78	23	13	114
Partner	Marrone	19	23	12	54
maschile	Verde	11	9	16	36
	Totale	108	55	41	204

(a) Qual è la probabilità che un uomo scelto a caso oppure la sua partner abbiano occhi blu?

(b) Qual è la probabilità che un uomo scelto a caso tra quelli con occhi blu abbia una partner con occhi blu?

(c) Qual è la probabilità che un uomo scelto a caso tra quelli con occhi marroni abbia una partner con occhi blu? E qual è invece la probabilità che un uomo scelto a caso tra quelli con occhi verdi abbia una partner con occhi blu?

(d) Ti sembra che i colori degli occhi dei due partner possano essere considerati caratteri indipendenti? Spiega il tuo ragionamento

Esercizio 2.20. *Saper disegnare un box-plot.*
Da indagini passate, sappiamo che, dopo aver seguito un corso di Statistica di base, l'80% degli studenti sa disegnare correttamente un box-plot. Tra questi, l'86% ha superato poi l'esame al primo appello, mentre solo il 65% degli studenti che non sa costruire un box-plotè riuscita a superare l'esame.

(a) Costruisci un diagramma ad albero oppure una tabella a doppia entrata per descrivere questo scenario.

(b) Qual è la probabilità che uno studente sia in grado di disegnare un box-plot sapendo che ha superato l'esame?

Esercizio 2.21. *Rischio di trombosi.*
Un test genetico viene usato per stabilire se un persona ha predisposizione all'insorgere di trombosi, cioè la formazione di grumi di sangue all'interno dei vasi sanguigni che ostruiscono il flusso del sangue stesso nel sistema

di circolazione. Si pensa che il 3% della popolazione mondiale ha questo tipo di predisposizione. Il test genetico è accurato al 99% sulle persone effettivamente predisposte, cioè la probabilità che il test sia positivo su un predisposto è 0.99. Lo stesso test è accurato al 98% accurate sui non predisposti. Qual è la probabilità che una persona scelta a caso nella popolazione il cui test è positivo, sia davvero una di quelle predisposte?

Esercizio 2.22. *HIV in Swaziland.*

Lo Swaziland è i paese al mondo con la più alta prevalenza di casi si HIV. Circa il 25.9% della popolazione risulta infatti sieropositiva Il test ELISA è stato uno dei primi e più accurati test per verificare la sieropositività. Per coloro effettivamente positivi all'HIV, il test ELISA ha una accuratezza[1] del 99.7%. Per i sieronegativi, il test è accurato al 92.6%. Se un cittadino dello Swaziland effettua il test e risulta positivo, qual è la probabilità che risulti effettivamente sieropositivo?

Esercizio 2.23. *Exit poll.*

Un istituto di ricerca ha effettuato delle indagini di tipo exit-poll (ovvero, interviste a a caldo fuori dai seggi) in occasione delle elezioni comunali a Roma del 2013. I ricercatori stabilirono che , secondo gli exit-poll, il 53% dei rispondenti aveva votato per il candidato del centro-sinistra Ignazio Marino. Inoltre, essi stimarono che il 37% di coloro che avevano votato per Marino, erano laureati mentre la percentuale di laureati tra coloro che NON aveva votato per Marino era del 44%. Supponiamo di selezionare casualmente una persona che ha partecipato all'exit poll e notiamo che si tratta di un laureato/a. Qual è la probabilità che abbia votato per Marino?

Esercizio 2.24. *It's never lupus.*

Il lupus eritematoso sistemico (LES o semplicemente lupus) è una malattia cronica di natura autoimmune, che può colpire diversi organi e tessuti del corpo. Si stima che il 2% della popolazione mondiale soffra di tale patologia. Esiste un test per verificare la effettiva malattia negli esseri umani. Il test ha una accuratezza del 98% tra i malati di Lupus e del 74% tra i non malati

[1] per una definizione di accuratezza si rimanda all'esercizio 2.21

La Fox Television negli USA dedica un programma ai potenziali malati di Lupus, che telefonano dopo aver effettuato un test, risultato positivo. Il titolo della trasmissione tende a sdrammatizzare ed è: "It's never lupus." Sulla base delle informazioni sopra riportate, come giudichi tale affermazione? È ragionevole? È troppo ottimista? Spiega la tua risposta.

Esercizio 2.25. *I gemelli.*

Nella specie umana, circa il 30% dei gemelli sono di tipo omozigotico (cioè identici, cresciuti nella stessa sacca uterina materna) mentre il restante 70% sono di tipo eterozigotico (gemelli diversi). I gemelli identici sono per forza dello stesso sesso, e con uguale probabilità nascono due maschi o due femmine. Al contrario, i gemelli diversi possono avere sesso diverso: infatti il 25% delle coppie di gemelli diversi è composta da due maschi, il 25% è composta da due femmine, e il restante 50% sono coppie miste. Se una coppia ha appena avuto due gemelline femmine, qual è la probabilità che siano identiche?

2.2 Estrazioni da popolazioni a bassa numerosità

Esercizio 2.26. *Palline e Urne, parte I.*

C'è un'urna con 5 palline rosse (R) 3 blu (B) e 2 arancioni (A). Effettuiamo estrazioni CON ripetizione, cioè una volta estratta,la pallina viene osservata e poi rimessa nell'urna.

(a) Qual è la probabilità che la prima pallina estratta sia di tipo B?

(b) Supponiamo che la prima pallina estratta sia B. Qual è la probabilità che la seconda pallina estratta sia ancora B?

(c) Supponiamo ora che la prima pallina estratta sia A. Sempre effettuando estrazioni senza ripetizione, qual è la probabilità che la seconda pallina estratta sia B?

(d) Qual è la probabilità di estrarre due palline B nelle prime due estrazioni?

(e) Quando effettuate con ripetizione, le varie estrazioni possono essere considerate indipendenti? Spiegare perché si oppure perché no.

Esercizio 2.27. *Calzini nel cassetto.*
Nel tuo cassetto ci sono 4 calzini blue (B), 5 grigi (G) e 3 neri (N). Ti svegli in ritardo e ti vesti di corsa prendendo due calzini a caso nel cassetto. Qual è la probabilità di ritrovarti con

(a) 2 calzini B

(b) nessun calzino G

(c) almeno 1 calzino N

(d) un calzino G

(e) calzini dello stesso colore, qualunque esso sia.

Esercizio 2.28. *Palline e Urne, parte II.*
C'è un'urna con 5 palline rosse (R) 3 blu (B) e 2 arancioni (A). Effettuiamo estrazioni SENZA ripetizione, cioè una volta estratta,la pallina viene osservata ma NON viene rimessa nell'urna.

(a) Supponiamo che alla prima estrazione venga estratta una pallina B. Qualè la probabilità che anche la seconda sia B?

(b) Supponiamo che alla prima estrazione venga estratta una pallina A. Qualè la probabilità che la seconda sia B?

(c) Qual è la probabilità di estrarre due palline B nelle prime due estrazioni?

(d) Quando effettuate senza ripetizione, le varie estrazioni possono essere considerate indipendenti? Spiegare perché si oppure perché no.

Esercizio 2.29. *Libri sullo scaffale.*

Nella tabella che segue viene riportata la distribuzione dei libri che ho comprato ma non ancora letto, in base al loro contenuto e alla loro copertina.

Formato	del libro		
	copertina rigida	copertina morbida	Totale
narrativa	13	59	72
saggistica	15	8	23
Totale	28	67	95

Prima di partire per le vacanze vogliamo portarci dietro due libri, da scegliere a caso. Effettuiamo allora delle estrazioni casuali senza ripetizione.

(a) Qual è la probabilità di scegliere per primo un libro a copertina rigida e per secondo un libro a copertina morbida di narrativa?

(b) Qual è la probabilità di estrarre per primo un libro di narrativa e poi a seguire uno a copertina rigida?

(c) Qual è la probabilità dell'evento precedente se le due estrazioni vengono effettuate CON ripetizione?

(d) Le risposte ai punti (b) e (c) sono molto simili ma non uguali. Spiega perché.

Esercizio 2.30. *L'abbigliamento delle studentesse.*

In una classe composta da 24 ragazze, 7 indossano jeans, 4 indossano pantaloni corti, 8 indossano la gonna ed il resto della classe indossa dei leggings. Se scegliamo a caso 3 studentesse, senza ripetizione, qualè la probabilità che una delle tre indossi leggings e le altre due abbiano i jeans?

Esercizio 2.31. *Il problema dei compleanni.*

Scegli tre persone a caso. Rispondi alle seguenti domande, sotto l'assunzione che

- nessuno sia nato il 29 febbraio

- la distribuzione delle nascite nella popolazione può essere considerata ragionevolmente uniforme nel corso dell'anno.

(a) Qualè la probabilità che le prime due persone estratte festeggino il compleanno nello stesso giorno?

(b) Qualè la probabilità che che almeno due delle tre persone estratte festeggino il compleanno nello stesso giorno?

2.3 Variabili casuali

Esercizio 2.32. *Studenti fumatori.*
Il 13% degli studenti universitari fuma almeno 4 sigarette al giorno.

(a) Determina il numero atteso di fumatori in una classe di 250 studenti.

(b) La palestra del campus apre ogni sabato mattina alle 9 in punto. Un certo sabato, alle 8:55, ci sono 27 studenti all'ingresso principale della palestra che aspettano di entrare. È ragionevole utilizzare lo stesso approccio che hai usato al punto (a) per calcolare il numero atteso di fumatori tra i 27 studenti? spiegare la risposta.

Esercizio 2.33. *Giochi di carte.*

Consideriamo un classico mazzo di 52 carte francesi (quattro semi: cuori (C), quadri (Q), picche (P) e fiori (F), ogni seme ha 13 carte: A,2,3,4,5,6,7,8,9,10,J,Q,K). Le regole del gioco sono le seguenti:

- Se estrai a caso una carta rossa (C o Q), non vinci nulla.
- Se estrai a caso una carta di picche, vinci 5 euro.
- Se estrai una carta di fiori vinci 10 euro ma se la carta è l'asso di fiori vinci altri 20 euro.

Se X è la variabile aleatoria "vincita in una estrazione",

(a) Determina la distribuzione di probabilità di X.

(b) Determina media e deviazione standard di X.

(c) Qual è la massima cifra che ritieni giusto pagare per partecipare a questo gioco? Spiega il tuo ragionamento.

Esercizio 2.34. *Giochi di carte, 2.*
Con lo stesso mazzo di carte del gioco precedente, si consideri ora un nuovo gioco, in cui vengono estratte in blocco (in pratica, senza ripetizione) tre carte. Le regole sono le seguenti:

- Se estrai tre carte di cuori vinci 50 euro.
- Se estrai tre carte nere (P o F) vinci 25 euro.
- Con qualunque altra combinazione non si vince nulla.

Sia Y la variabile aleatoria "vincita in una mano del gioco",

(a) Determina la distribuzione di probabilità di Y.

(b) Determina media e deviazione standard di Y.

(c) Se il prezzo per partecipare ad una mano di questo gioco è 5 euro, quali saranno media e deviazione standard del ricavo aleatorio (cioè la vincita − (meno) il prezzo per partecipare)?

(d) Se il prezzo per partecipare ad una mano di questo gioco è 5 euro, decidi di giocare oppure no? spiega la tua scelta.

Esercizio 2.35. *Ne vale la pena?*
Andrea è sempre alla ricerca di modi per fare soldi velocemente e senza fatica. Negli ultimi tempi, sta provando con i giochi d'azzardo. In particolare si è concentrato sul seguente gioco: si pagano 2 euro per partecipare. Il giocatore estrae una carta dal solito mazzo di 52 carte francesi. Se il giocatore estrae un numero (le carte da 2 a 10), non vince nulla. Se estrae una figura

(J,Q,K) egli vince 3 euro. Se invece estrae un Asso, il giocatore vince 5 euro. Se poi è l'asso di fiori, allora vince altri 20 euro (per un totale di 25 euro). Sia Z la variabile aleatoria: "vincita di Andrea".

(a) Calcolare la distribuzione di Z.

(b) Te la senti di consigliare ad Andrea questo come un gioco conveniente? Spiega bene il perché.

Esercizio 2.36. *Rendimenti di un portafoglio.*

Il rendimento di un portafoglio di titoli incrementa il suo valore del 18% durante una fase di boom finanziario, mentre cresce solo del 9% in tempi normali. Durante una recessione esso decresce del 12%. Quale è il valore atteso del rendimento di questo portafoglio, se i tre scenari possibili sono considerati ugualmente probabili?

Esercizio 2.37. *Roulette, Parte I.*

Il gioco della roulette consiste in una ruota con 37 slots, ovvero i numeri da 0 a 36, estremi inclusi Ad ogni giro di roulette, una pallina si sistemerà in uno dei 37 slots: gli slots sono ugualmente probabili. Lo slot 0 è verde, gli altri 36 sono 18 di colore rosso (R) e 18 di colore nero (N)

I giocatori possono scommettere, tra le altre cose, sul colore dello slot (R o N): se la pallina si ferma in uno slot del colore da loro prescelto, essi vincono tanto denaro quanto quello giocato (in pratica, se giocano 1 euro, si riprendono l'euro giocato più un altro di vincita). Se invece la pallina si ferma su uno slot di un altro colore, perdono il denaro scommesso. Supponiamo che Tu scommetta 1 euro sul rosso (R). Qual è il valore atteso del tuo ricavo netto? qual è la deviazione standard?

Esercizio 2.38. *Roulette, Parte II.*

L'Esercizio 2.37 descrive alcuni tipi di giocate che si possono fare alla roulette.

(a) Supponiamo che Tu scommetta 3 euro sul rosso (R) in una mano di roulette. Qual è il valore atteso del tuo ricavo netto? qual è la deviazione standard?

(b) Supponiamo ora che tu decida di scommettere 1 euro in tre mani successive, giocando sempre sul rosso. Qual è il valore atteso del tuo ricavo netto? qual è la deviazione standard?

(c) Puoi fare un confronto tra le risposte che hai dato al punto (a) e al punto (b)? Che cosa ti dicono a proposito del rischio associato alle due strategie?

Esercizio 2.39. *Tariffe bagagli.*
Una compagnia aerea applica le seguenti tariffe per i bagagli.

- 25 euro per la prima valigia.
- 35 euro per la seconda valigia.

Secondo le statistiche della compagnia aerea, il 54% dei passeggeri non imbarca bagagli. il 34% imbarca un solo bagaglio,mentre il 12% dei passeggeri imbarca due bagagli. Per semplicità trascuriamo quella piccola parte di clientela che imbarca più di due bagagli. Sia X la variabile aleatoria: "Ricavo per passeggero".

(a) Determina la distribuzione di X.

(b) Calcola media e deviazione standard di X.

(c) Qual è il ricavo medio per un volo con 120 passeggeri? qual è la deviazione standard? Sottolinea ogni eventuale assunzione che hai fatto per produrre una risposta e valuta, volta per volta, se si tratta di una assunzione ragionevole.

Esercizio 2.40. *Roma - Lazio*
Tu e un tuo amico fate una scommessa relativa al risultato di Roma-Lazio, derby capitolino. Secondo le attuali statistiche, la Roma ha una probabilità di vincere pari allo 0.45; il pareggio ha probabilità 0.21, mentre la probabilità che vinca la Lazio è pari 0.34. Il tuo amico mette sul banco

5 euro e scommette sulla Roma. Assumendo che il pareggio annullerebbe la scommessa, quanto devi mettere sul banco per scommettere sulla Lazio in modo che la scommessa sia equa?

In termini di quote, sei capace di dire "a quanto vengono date" Roma e Lazio?

Esercizio 2.41. *Vendite su Ebay.*

Marzia sta monitorando le quotazioni di due articoli su Ebay:

- Un libro di testo che si vende ad una media di 110 euro con una standard deviation of 4 euro.
- Un videogame di Mario Kart per il Nintendo Wii, che si vende ad una media di 38 euro con uno standard deviation pari a 5 euro.

(a) Marzia vuole vendere il videogame e comprare il libro di testo. In media quanto ricavo (cioè entrate − uscite), si aspetta di ottenere Marzia dalle due operazioni? Qual è la deviazione standard associata a tale valore medio?

(b) Lucia sta vendendo il libro di testo su Ebay per conto di un amico, il quale le pagherà una commissione del 10%, ovvero Lucia tratterrà per sé il 10% del ricavo. Quanto denaro si aspetta in media di ottenere Lucia? Con quale deviazione standard?

Esercizio 2.42. *Quanto costa la colazione.*

Sandra mangia a colazione, ogni mattina, un cappuccino e un cornetto. Ci sono molti bar vicino alla sua abitazione e così ogni giorno ne sceglie a caso uno, indipendentemente dai giorni precedenti. Il prezzo medio di un cappuccino è di 1.40 euro con una deviazione standard di 30 centesimi; il prezzo medio del cornetto è 1 euro con una deviazione standard di 15 centesimi. I due prezzi sono considerati indipendenti.

(a) Qual è il prezzo medio della spesa che Sandra sostiene giornalmente per la colazione? qual è la deviazione standard?

(b) Qual è il prezzo medio della spesa che Sandra sostiene in una settimana (7 giorni) per la colazione? qual è la deviazione standard?

Esercizio 2.43. *Ice cream.*

Le gelaterie vendono il gelato in confezioni da 1 kg.; un cono gelato, mediamente, contiene 50 grammi di gelato. Tuttavia cè una certa variabilità nella confezione delle scatole e nella preparazione dei coni Chiamiamo X la quantità aleatoria di gelato in una scatola e Y la quantità aleatoria di gelato su un cono. Assumiamo che tali variabili aleatorie abbiano le seguenti medie, deviazioni standard e varianze, espresse in grammi.

	media	stand. deviation	varianza
X	1000	10	100
Y	50	2	4

(a) Ad un party viene servita una intera scatola di gelato più tre coni. Mediamente, quanto gelato è stato servito? con quale deviazione standard?

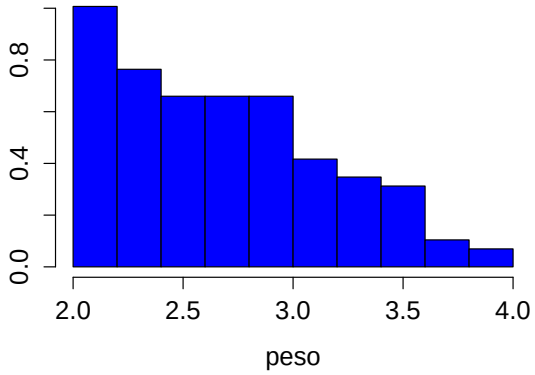
(b) Quanto gelato ti aspetti che resti in una nuova scatola, dopo che è stato riempito un cono? In termini matematici, calcola il valore atteso di $X - Y$. Qual è la deviazione standard di tale previsione?

(c) Usando come esempio il contesto di questo esercizio, spiega perché, anche quando si calcola la differenza tra due variabili aleatorie, occorre sommare le varianze.

2.4 Distribuzioni continue

Esercizio 2.44. *Peso dei gatti.*

L'istogramma di seguito riporta il peso in kg. di 47 femmine e 97 maschi di gatto.



Sulla base delle informazioni fornite dal grafico, potete dare una risposta approssimata alle seguenti domande?

(a) Quale percentuale di gatti pesa meno di 2.6 kg.?

(b) Quale percentuale di gatti pesa tra 2.4 kg. e 3 kg.?

(c) Quale percentuale di gatti pesa oltre i 3.6 kg.?

Esercizio 2.45. *Redditi e genere.*

La tabella che segue riporta le frequenze relative della distribuzione dei redditi pro-capite annui per un campione di quasi 100 milioni di cittadini americani, aggiustati in termini di inflazione e relativi al 2009. Questi dati provengono dall'American Community Survey per il periodo 2005-2009. Questo campione è formato dal 59% di uomini e 41% di donne.

Classe di di Reddito	Totale
$\leq$ 10000 euro	2.2%
fino a 15000 euro	4.7%
fino a 25000 euro	15.8%
fino a 35000 euro	18.3%
fino a 50000 euro	21.2%
fino a 65000 euro	13.9%
fino a 75000 euro	5.8%
fino a 100000 euro	8.4%
oltre 100000 euro	9.7%

(a) Calcolare il reddito mediano pro-capite, indipendentemente dal genere.

(b) Qual è la probabilità che un individuo scelto a caso guadagni meno di 50 mila dollari?

(c) Qual è la probabilità che un individuo scelto a caso guadagni meno di 50 mila dollari e sia donna? Annota le assunzioni che hai fatto per rispondere a questa domanda.

(d) La stessa fonte di dati ci dice che il 71.8% delle donne guadagna meno di 50 mila dollari all'anno. Usa questa informazione per determinare se l'assunzione che hai fatto al punto (c) possa essere considerata valida o meno.

Esercizio 2.46.

Nel censimento del 2000 ogni persona residente negli USA doveva scegliere da un lungo elenco la propria razza. La categoria "Ispanico/latino" è un caso a parte poiché in essa vi possono essere tante razze diverse. Se scegliamo un residente negli USA in modo casuale, in base ai dati del censimento del 2000 abbiamo le seguenti probabilità:

	Ispanici	Non ispanici
Asiatici	0.000	0.036
Neri	0.003	0.121
Bianchi	0.060	0.691
Altro	0.062	0.027

1. Verifica che questa tabella di probabilità sia corretta.
2. Quanto vale la probabilità che un americano scelto in modo casuale sia ispanico?
3. I bianchi di origine non ispanica rappresentano da sempre la maggioranza di residenti negli USA. Quale è la probabilità che un americano scelto in modo casuale non sia membro di questo gruppo?

• • •

Soluzione

1. Per verificare che questa tabella di probabilità sia corretta bisogna verificare che:
   - le probabilità assumono valori tra 0 e 1;
   - poiché l'evento A="il cittadino è ispanico" e l'evento B="il cittadino è non ispanico" sono complementari ed esauriscono lo spazio degli eventi S, allora la somma delle probabilità deve essere pari ad 1.

   La prima condizione è verificata in quanto le probabilità riportate assumono tutte valori tra 0 e 1; anche la seconda condizione è verificata in quanto

   $$P(S) = P(A) + P(B) = 0.125 + 0.875 = 1$$

   dove
   P(A)=P("il cittadino ispanico")= 0+0.003+0.060+0.062=0.125

   P(B)=P("il cittadino non ispanico")=0.036+0.121+0.691+0.027=0.875=
   1-P("il cittadino ispanico")

2. La probabilità che un americano scelto in modo casuale sia ispanico è pari a P(A)= 0.125;

3. La probabilità dell'evento C="il cittadino è un bianco non ispanico" è pari a P(C)=0.691; pertanto la probabilità che un americano scelto a caso non sia un bianco ispanico è la probabilità di C^c, ossia

$$P(C^c) = 1 - P(C) = 1 - 0.691 = 0.309$$

.

• • •

Esercizio 2.47.

È stato chiesto a 500 soggetti (maschi e femmine) abitanti di un'area metropolitana se amano fare shopping. 136 dei 250 uomini intervistati e 224 delle 250 donne hanno risposto affermativamente. Scelto a caso un soggetto, qual è la probabilità che:

1. ami fare lo shopping;

2. sia una donna e ami fare shopping;

3. sia una donna o ami fare shopping;

4. sia un uomo o una donna.

• • •

Soluzione

Per rispondere alle domande, può essere utile schematizzare il problema nella seguente tabella:

	UOMO	DONNA	
SI	136	224	360
NO	114	26	140
	250	250	500

Possiamo quindi rispondere alle domande:

1. Sia A l'evento A="un soggetto scelto a caso ama fare shopping".
   P(A)= # di soggetti che hanno risposto positivamente/ # di soggetti intervistati=$\frac{360}{500}$=0.72

2. Definiamo gli eventi:
   D="un soggetto scelto a caso è una donna"
   A="un soggetto scelto a caso ama fare shopping".
   La probabilità richiesta è la probabilità dell'intersezione dei 2 eventi $P(D \cap A)$ ossia la probabilità di estrarre a caso una donna che ami fare shopping. Dalla tabella, si deduce che tale probabilità è pari a
   $P(D \cap A) = \frac{224}{500} = 0.448$

3. La probabilità richiesta è la probabilità dell'unione dei 2 eventi, ossia
   $P(A \cup D) = P(A) + P(D) - P(A \cap D) = \frac{360}{500} + \frac{1}{2} - \frac{224}{500} = 0.772$

4. Definiamo gli eventi:
   D="il soggetto estratto è una donna"
   U="il soggetto estratto è un uomo"
   i 2 eventi sono complementari e disgiunti, ossia $P(D) = 1 - P(U)$. Pertanto la probabilità della loro unione è pari a
   $P(D \cup U) = P(D) + P(U) = 1$.

• • •

Esercizio 2.48.

Ogni anno vengono effettuate delle valutazioni circa le performance delle nuove automobili durante i primi 90 giorni di vita. Supponiamo che le

automobili siano classificate in base alla nazionalità della casa produttrice (americana/ non americana) e in base al fatto che la macchina abbia richiesto o meno una riparazione nel periodo di garanzia. In base ai dati raccolti si ottiene una probabilità pari a 0.04 che l'automobile richieda una riparazione durante il periodo di garanzia, una probabilità di 0.6 che l'automobile sia costruita in America e una probabilità pari a 0.025 che una macchina richieda una riparazione durante il periodo di garanzia e sia stata prodotta da una società americana. Scelta a caso un'automobile, calcolare la probabilità che:

1. richieda una riparazione durante il periodo di garanzia;
2. richieda una riparazione durante il periodo di garanzia e sia stata prodotta da una società americana;
3. richieda una riparazione durante il periodo di garanzia o sia stata prodotta da una società americana;
4. richieda una riparazione durante il periodo di garanzia o non sia stata prodotta da una società americana.

• • •

Soluzione

Formalizziamo le informazioni fornite dall'esercizio come segue:

- Sia l'evento A="un'automobile scelta a caso richiede una riparazione"; il testo dell'esercizio ci dice che P(A)=0.04;
- Sia l'evento B="un'automobile scelta a caso è costruita in America "; il testo dell'esercizio ci dice che P(B)=0.6;
- $A \cap B$ l'evento "un'automobile scelta a caso richiede una riparazione ed è costruita in America"; il testo ci dice che $P(A \cap B) = 0.025$.

Pertanto, possiamo rispondere alle domande:

1. P(A)=0.04;
2. $P(A \cap B) = 0.025$;
3. $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.04 + 0.6 - 0.025 = 0.615$;
4. L'evento "un'automobile scelta a caso non è stata prodotta da una società americana" è l'evento complementare a B; pertanto la sua probabilità è pari a $P(B^c) = 1 - P(B) = 0.4$. La probabilità richiesta è

   $$P(A \cup B^c) = P(A) + P(B^c) - P(A \cap B^c) = 0.04 + 0.4 - 0.015 = 0.425$$

   dove $P(A \cap B^c) = P(A) - P(A \cap B) = 0.04 - 0.025 = 0.015$.

• • •

Esercizio 2.49.

È stata condotta un'indagine per valutare se le aziende di grandi dimensioni sono meno propense delle aziende di medie-piccole dimensioni ad offrire azioni ai membri del proprio consiglio di amministrazione. I risultati campionari sono i seguenti: su 189 aziende di grandi dimensioni, 40 offrono le proprie azioni ai membri del consiglio di amministrazione; su 180 aziende di media-piccola dimensioni, 43 offrono azioni ai membri del proprio consiglio di amministrazione. Scelta a caso un'azienda, calcolare la probabilità che questa:

1. offra azioni ai membri del consiglio di amministrazione;
2. sia di dimensioni medio-piccole e non offra azioni ai membri del consiglio di amministrazione;

3. sia di dimensioni medio-piccole oppure offra azioni ai membri del consiglio amministrazione.

• • •

Soluzione

Per rispondere alle domande, può essere utile schematizzare il problema nella seguente tabella:

	Grandi	Medie-Piccole	
SI	40	43	83
NO	149	137	286
	189	180	369

Possiamo quindi rispondere alle domande:

1. Sia l'evento A=" un'azienda scelta a caso offre azioni ai membri del consiglio di amministrazione"; la probabilità dell'evento è pari a $P(A) = \frac{83}{369} = 0.225$
2. Sia l'evento B=" un'azienda scelta a caso è di dimensioni medio-piccole"; dobbiamo calcolare $P(A^c \cap B)$, ossia la probabilità che un'azienda di dimensioni medio-piccole offra azioni ai membri del consiglio di amministrazione. Dalla tabella si deduce che $P(A^c \cap B) = \frac{137}{369} = 0.371$
3. Dobbiamo calcolare la probabilità dell'unione, ossia $P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{180}{369} + \frac{83}{369} - \frac{43}{369} = \frac{220}{369} = 0.596$

• • •

Esercizio 2.50.

Un 4-soft nel gioco dei 2 dadi, si verifica quando si totalizza un 4, avendo 1 su un dado e 3 sull'altro.

1. Quale è la probabilità di totalizzare un 4-soft?
2. Quale è la probabilità di realizzare 4?
3. Quale è la probabilità di realizzare 5?

• • •

Soluzione

Con 2 dadi, si possono realizzare un totale di 36 possibili risultati. Con 2 dadi, si può ottenere 4 con i seguenti punteggi: $[1,3],[3,1],[2,2]$; mentre un punteggio totale pari a 5 si può ottenere come $[1,4],[4,1],[2,3],[3,2]$

1. Gli eventi favorevoli ad un 4-soft sono 2, ossia $[1,3]$ e $[3,1]$; quindi la probabilità di un 4-soft è $\frac{2}{36} = 0.055$
2. Gli eventi favorevoli ad un totale di 4 sono 3, ossia $[1,3],[3,1],[2,2]$; pertanto la probabilità di ottenere un punteggio totale di 4 è pari a $\frac{3}{36} = 0.083$
3. Gli eventi favorevoli ad un totale di 5 sono 4, ossia $[1,4],[4,1],[2,3],[3,2]$; pertanto la probabilità di ottenere un punteggio totale di 4 è pari a $\frac{4}{36} = 0.111$

• • •

Esercizio 2.51.

Sia X un numero compreso tra 0 e 1 generato casualmente. Ricava le seguenti probabilità:

1. $P(0 \leq X \leq 0.4)$;
2. $P(0.4 \leq X \leq 1)$;
3. $P(0.3 \leq X \leq 0.5)$.

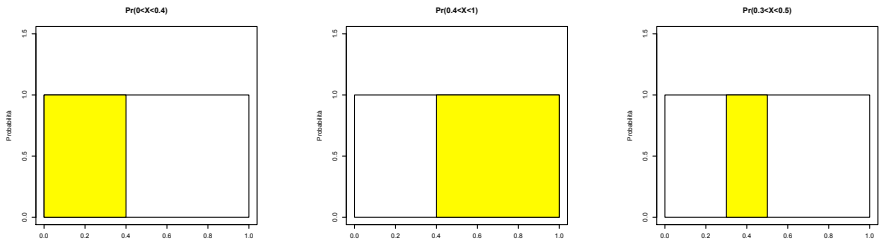
• • •

Soluzione

Le probabilità richieste sono pari rispettivamente a

1. $P(0 \leq X \leq 0.4) = 0.4$
2. $P(0.4 \leq X \leq 1) = 0.6$
3. $P(0.3 \leq X \leq 0.5) = 0.2$

Tali probabilità sono rappresentate nelle seguenti figure.



• • •

Esercizio 2.52.

Si stima che il 30% degli adulti negli Stati Uniti siano obesi, che il 3% siano diabetici e che il 2% siano sia obesi che diabetici. Determina la probabilità che un individuo scelto casualmente

1. sia diabetico se è obeso;
2. sia obeso se è diabetico.

• • •

Soluzione

Indichiamo con O e D i seguenti eventi:

O="un individuo scelto casualmente sia obeso";

D="un individuo scelto casualmente sia di diabetico".

Il testo ci dice che $P(O) = 0.30$, $P(D) = 0.03$ e $P(O \cap D) = 0.02$.

1. Il quesito chiede la probabilità che il soggetto sia diabetico *dato* che obeso, ossia $P(D|O)$; applicando la regola della probabilità condizionata si ha
$$P(D|O) = \frac{P(D \cap O)}{P(O)} = \frac{0.02}{0.3} = 0.067$$

2. Il quesito chiede la probabilità che il soggetto sia obeso *dato* che diabetico, ossia $P(O|D)$; applicando la regola della probabilità condizionata si ha
$$P(O|D) = \frac{P(D \cap O)}{P(D)} = \frac{0.02}{0.03} = 0.667$$

• • •

Esercizio 2.53.

Tra i partecipanti ad un concorso per giovani compositori il 50% suona il pianoforte, il 30% suona il violino e il 20% la chitarra. Partecipano ad un concorso per la prima volta il 10% dei pianisti, il 33% dei violinisti e il 10% dei chitarristi. Applicando i concetti di probabilità condizionata e il teorema di Bayes, rispondere alle seguenti domande.

1. Quale è la probabilità che un compositore scelto a caso sia un aspirante alla prima esperienza?

2. Sapendo che ad esibirsi per primo sarà un compositore alla prima esperienza, quale è la probabilità che sia un chitarrista?

• • •

Soluzione

Definiamo gli eventi:

A = "Un partecipante scelto a casa è un aspirante compositore alla prima esperienza"

B = "Un partecipante scelto a casa è un pianista"

C = "Un partecipante scelto a casa è un violinista"

D = "Un partecipante scelto a casa è un chitarrista"

abbiamo

$$\begin{aligned} P(A) &= P(A \cap S) = P(A \cap (B \cup C \cup D)) \\ &= P(A \cap B) + P(A \cap C) + P(A \cap D) \\ &= P(A|B)P(B) + P(A|C)P(C) + P(A|D)P(D) \\ &= 0.1 \cdot 0.5 + 0.33 \cdot 0.3 + 0.1 \cdot 0.2 = 0.17 \end{aligned}$$

Per quanto riguarda il secondo quesito abbiamo:

$$\begin{aligned} P(D|A) &= \frac{P(D \cap A)}{P(A)} \\ &= \frac{P(A|D)P(D)}{P(A)} = \frac{0.1 \cdot 0.2}{0.17} = 0.12 \end{aligned}$$

• • •

Esercizio 2.54.

Un negozio accetta sia la carta di credito American Express che la VISA. Il 22 percento dei clienti del negozio porta con sè una American Express, il 58 percento una VISA, e il 14 entrambe le carte di credito.

a. Qual è la probabilità che un cliente abbia con sè almeno una di queste carte?

b. Qual è la probabilità che un cliente abbia con sè una VISA e sicuramente non abbia con sè una American Express?

• • •

Soluzione

a. Sia A l'evento "un cliente ha una America Express", e sia B l'evento "un cliente ha una VISA". Le informazioni note sono quindi le seguenti:

1. $P(A) = 0.22$;
2. $P(B) = 0.58$;
3. $P(A \cap B) = 0.14$.

La probabilità richiesta $P(A \cup B)$ è

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.22 + 0.58 - 0.14 = 0.66$$

b. La probabilità richiesta è $P(B \cap A^c)$. Tale probabilità è pari a

$$P(B \cap A^c) = P(B) - P(A \cap B) = 0.58 - 0.14 = 0.44.$$

• • •

Esercizio 2.55.

Una scuola elementare offre due corsi opzionali di lingua straniera, uno di francese e uno di spagnolo. Questi corsi sono aperti a tutti i 120 studenti delle ultime classi della scuola. Supponiamo che 32 studenti frequentino il corso di francese, 36 il corso di spagnolo, e 60 almeno un corso. Se scegliamo casualmente uno studente delle ultime classi, qual è la probabilità che questo studente frequenti entrambi i corsi di lingue?

• • •

Soluzione

Siano A e B gli eventi che lo studente scelto sia iscritto rispettivamente al corso di francese e al corso di spagnolo. Determineremo $P(A \cap B)$, la probabilità che lo studente frequenti sia il corso di francese che quello di spagnolo, usando la seguente formula

$$P(A \cap B) = P(A) + P(B) - P(A \cup B).$$

Visto che 32 su 120 studenti sono iscritti al corso di francese, 36 su 120 frequentano il corso di spagnolo, e 60 su 120 frequentano almeno un corso, otteniamo

1. $P(A) = \frac{32}{120}$;
2. $P(B) = \frac{36}{120}$;
3. $P(A \cup B) = \frac{60}{120}$.

Quindi

$$P(A \cap B) = \frac{32}{120} + \frac{36}{120} - \frac{60}{120} = \frac{8}{120}.$$

Questo significa che la probabilità che uno studente scelto a caso frequenti entrambi i corsi di lingua è $\frac{8}{120}$.

• • •

Esercizio 2.56.

È stata condotta un'indagine per valutare se le aziende di grandi dimensioni sono meno propense delle aziende di medie-piccole dimensioni ad offrire azioni ai membri del proprio consiglio di amministrazione. I risultati campionari sono i seguenti: su 189 aziende di grandi dimensioni, 40 offrono le proprie azioni ai membri del consiglio di amministrazione; su 180 aziende di media-piccola dimensioni, 43 offrono azioni ai membri del proprio consiglio di amministrazione. Scelta a caso un'azienda, calcolare la probabilità che questa:

1. offra azioni ai membri del consiglio di amministrazione;
2. sia di dimensioni medio-piccole e non offra azioni ai membri del consiglio di amministrazione;
3. sia di dimensioni medio-piccole oppure offra azioni ai membri del consiglio amministrazione.

• • •

Soluzione

Per rispondere alle domande, può essere utile schematizzare il problema nella seguente tabella:

	Grandi	Medie-Piccole	
SI	40	43	83
NO	149	137	286
	189	180	369

Possiamo quindi rispondere alle domande:

1. Sia l'evento A=" un'azienda scelta a caso offre azioni ai membri del consiglio di amministrazione"; la probabilità dell'evento è pari a $P(A) = \frac{83}{369} = 0.225$
2. Sia l'evento B=" un'azienda scelta a caso è di dimensioni medio-piccole"; dobbiamo calcolare $P(A^c \cap B)$, ossia la probabilità che un'azienda di dimensioni medio-piccole non offra azioni ai membri del consiglio di amministrazione. Dalla tabella si deduce che $P(A^c \cap B) = \frac{137}{369} = 0.371$
3. Dobbiamo calcolare la probabilità dell'unione, ossia $P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{83}{369} + \frac{180}{369} - \frac{43}{369} = \frac{220}{369} = 0.596$

• • •

2.5 Distribuzione della media campionaria

Esercizio 2.57.

Il dottore di Anna è preoccupato che lei possa soffrire di diabete gestazionale (alto livello di glucosio nel sangue durante la gravidanza). È presente una certa variabilità sia nel reale livello di glucosio nel sangue, sia nel risultati del test che lo misura. Una paziente è affetta da diabete gestazionale se il livello di glucosio, un'ora dopo aver ingerito una bevanda zuccherata, è superiore ai 140 milligrammi per decilitro (mg/dl). Il livello di glucosio di Anna varia secondo una distribuzione Normale con media $\mu = 125$ mg/dl e $\sigma = 10$ mg/dl.

1. Se si fa una singola misurazione di glucosio, quale è la probabilità che ad Anna sia diagnosticato il diabete gestazionale?

2. Se invece le misurazioni sono fatte su 4 giorni separati e la regola dei 140 mg/dl viene applicata alla media delle 4 misurazioni, quale è la probabilità che ad Anna venga diagnosticato il diabete gestazionale?

3. Se invece si facessero 10 misurazioni, come cambierebbe in termini di media e deviazione standard la distribuzione della media?

• • •

Soluzione

1. Sia X la variabile aleatoria "livello di glucosio nel sangue"; sappiamo che $X \sim N(\mu = 125, \sigma = 10)$. Vogliamo calcolare la probabilità che

$X > 140$. Possiamo calcolare tale probabilità come segue:

$$\begin{aligned} Pr(X > 140) = 1 - Pr(X \leq 140) &= \\ 1 - Pr\left(\frac{X-\mu}{\sigma} \leq \frac{140-\mu}{\sigma}\right) &= \\ 1 - Pr\left(Z \leq \frac{140-125}{10}\right) &= \\ 1 - Pr(Z \leq 1.5) &= \\ 1 - 0.9332 = 0.0668 \end{aligned}$$

2. La media campionaria $\bar{X}$ ha distribuzione normale con media μ e deviazione standard $\frac{\sigma}{\mu}$, ossia

$$\bar{X} \sim N\left(\mu_{\bar{X}} = 125, \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = 5\right)$$

Pertanto, la probabilità richiesta :

$$\begin{aligned} Pr(\bar{X} > 140) = 1 - Pr(\bar{X} \leq 140) &= \\ 1 - Pr\left(\frac{X-\mu_{\bar{X}}}{\sigma_{\bar{X}}} \leq \frac{140-\mu_{\bar{X}}}{\sigma_{\bar{X}}}\right) &= \\ 1 - Pr\left(\frac{X-\mu}{\frac{\sigma}{\sqrt{n}}} \leq \frac{140-\mu}{\frac{\sigma}{\sqrt{n}}}\right) &= \\ 1 - Pr\left(Z \leq \frac{140-125}{\frac{10}{2}}\right) &= \\ 1 - Pr(Z \leq 3) &= \\ 1 - 0.9987 = 0.0013 \end{aligned}$$

3. Con un campione di numerosità $n = 10$ la distribuzione della media campionaria è più concentrata rispetto alla media delle singole osservazioni; poiché la deviazione standard della media campionaria è pari a $\frac{\sigma}{\sqrt{n}}$, essa diminuirà al crescere di n. Pertanto per $n = 10$, la media della distribuzione della media campionaria rimane invariata, mentre la deviazione standard si riduce e sarà pari a $\frac{10}{\sqrt{10}} = 3.162$.

• • •

Esercizio 2.58.

Negli USA, la tariffa pagata dalle famiglie ai provider di Internet è piuttosto variabile, ma la quota media mensile è di 28 dollari e la deviazione standard di 10. La distribuzione non è Normale: molte famiglie pagano circa 10 dollari per un accesso limitato oppure circa 25 dollari per un accesso illimitato, ma ve ne sono alcune che pagano molto di più per connessioni veloci. In una indagine campionaria si intervista un campione casuale di 500 famiglie con accesso a internet. Quale è la probabilità che la tariffa media pagata dal campione di famiglie sia maggiore di 29 dollari?

• • •

Soluzione

Sia X la variabile casuale "tariffa pagata dalle famiglie residenti negli USA per provider di Internet"; la distribuzione di X nella popolazione ha media $\mu = 28$ e deviazione standard $\sigma = 10$. Tale distribuzione è non Normale. Tuttavia, il teorema del limite centrale ci garantisce che, qualunque sia la distribuzione di X nella popolazione, se la dimensione del campione n è elevata, la media campionaria $\bar{X}$ ha distribuzione Normale con media μ e deviazione standard $\frac{\sigma}{n}$.
Pertanto, per il teorema del limite centrale,

$\bar{X} \sim N\left(\mu = 28, \sigma = \frac{10}{\sqrt{500}} = 0.447\right)$ e la probabilità richiesta è pari a:

$$\begin{aligned} Pr(\bar{X} > 29) = 1 - Pr(\bar{X} \leq 29) &= \\ 1 - Pr\left(\frac{X - \mu_{\bar{X}}}{\sigma_{\bar{X}}} \leq \frac{29 - \mu_{\bar{X}}}{\sigma_{\bar{X}}}\right) &= \\ 1 - Pr\left(\frac{X - \mu}{\frac{\sigma}{\sqrt{n}}} \leq \frac{29 - \mu}{\frac{\sigma}{\sqrt{n}}}\right) &= \\ 1 - Pr\left(Z \leq \frac{29 - 28}{0.447}\right) &= \\ 1 - Pr(Z \leq 2.237) &= \\ 1 - 0.9875 = 0.0125 & \end{aligned}$$

• • •

Esercizio 2.59.

La distribuzione del tempo di vita di un nuovo apparecchio telefonico (misurato in giorni) è Normale con media $\mu = 800$ e deviazione standard $\sigma = 120$.

a. Qual è la probabilità di sostituire uno qualsiasi degli apparecchi se si fissa la durata di garanzia pari a 600 giorni?

b. Quale durata deve avere la garanzia affinché solo il 9 percento degli apparecchi debba essere sostituito?

c. Dato un campione casuale di 80 apparecchi, qual è la probabilità che la media campionaria $\bar{X}$ assuma un valore al più pari a 800?

• • •

Soluzione

a. Sia X la variabile casuale "durata di un apparecchio telefonico"; la distribuzione di X nella popolazione ha media $\mu = 800$ e deviazione standard $\sigma = 120$. La probabilità richiesta è

$$\begin{aligned} Pr(X < 600) = Pr\left(\frac{X-\mu}{\sigma} < \frac{600-\mu}{\sigma}\right) &= \\ Pr\left(Z < \frac{600-800}{120}\right) &= \\ Pr(Z < -1.67) &= 0.0475 \end{aligned}$$

b. Dobbiamo individuare il valore x tale che

$$\begin{aligned} 0.09 = Pr\left(X < x\right) &= \\ Pr\left(Z < \frac{x-800}{120}\right) &= \end{aligned}$$

Poiché $\frac{x-800}{120} = -1.34$, si ha che $x = 639.2$

c. La distribuzione della media campionaria $\bar{X}$ è Normale con media 800 e deviazione standard $\frac{120}{\sqrt{80}} = 13.4$. Poiché la mediana di $\bar{X}$ è pari a 800, $P(\bar{X} \leq 800) = 0.5$

• • •

Esercizio 2.60.

Si stima che il 30% degli adulti negli Stati Uniti sia obeso, che il 3% siano diabetico e che il 2% sia obeso e diabetico. Determina la probabilità che un individuo scelto casualmente

1. sia diabetico se è obeso;
2. sia obeso se è diabetico.

• • •

Soluzione

Indichiamo con O e D i seguenti eventi:

O=”un individuo scelto casualmente è obeso”;

D=”un individuo scelto casualmente è diabetico”.

Il testo ci dice che $P(O) = 0.30$, $P(D) = 0.03$ e $P(O \cap D) = 0.02$.

1. Il quesito chiede la probabilità che il soggetto sia diabetico *dato* che obeso, ossia $P(D|O)$; applicando la regola della probabilità condizionata si ha
$$P(D|O) = \frac{P(D \cap O)}{P(O)} = \frac{0.02}{0.3} = 0.067$$

2. Il quesito chiede la probabilità che il soggetto sia obeso *dato* che diabetico, ossia $P(O|D)$; applicando la regola della probabilità condizionata si ha
$$P(O|D) = \frac{P(D \cap O)}{P(D)} = \frac{0.02}{0.03} = 0.667$$

• • •

Esercizio 2.61.

A un esame universitario si presentano sia studenti che hanno seguito il corso sia studenti che non l'hanno seguito. Il docente ritiene che il 65% degli studenti abbiano seguito il corso. La probabilità che uno studente superi l'esame dato che ha seguito il corso è 0.75, mentre la probabilità che uno studente superi l'esame dato che non ha seguito il corso è 0.40.

- Calcolare la probabilità che uno studente superi l'esame.
- Calcolare la probabilità che uno studente abbia seguito il corso dato che ha superato l'esame.

• • •

Soluzione

Indichiamo con A e B gli eventi:

A="lo studente supera l'esame";

B="lo studente ha seguito il corso"

1. Dall'informazione fornita dal docente "il 65% degli studenti hanno seguito il corso", approssimando la probabilità con la frequenza relativa, si ha P(B) = 0.65 e

$$P(B^c) = 1 - P(B) = 1 - 0.65 = 0.35$$

e inoltre $P(A|B) = 0.75$ e $P(A|B^c) = 0.40$. L'evento A può essere rappresentato come l'unione di due eventi incompatibili $A = (A \cap B) \cup (A \cap B^c)$; pertanto

$$P(A) = P(A \cap B) + P(A \cap B^c)$$

dove

- $P(A \cap B) = P(A|B) * P(B) = 0.75 * 0.65 = 0.4875$
- $P(A \cap B^c) = P(A|B^c) * P(B^c) = 0.40 * 0.35 = 0.1400$

Pertanto $P(A) = 0.4875 + 0.1400 = 0.6275$

2. La probabilità richiesta è

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{0.4875}{0.6275} = 0.7769$$

• • •

Esercizio 2.62.

Ad una conferenza, partecipano 30 psichiatri e 24 neurologi. Due di queste 54 persone vengono scelte casualmente per fare parte di una commissione. Quale è la probabilità che venga scelto almeno un neurologo?

• • •

Soluzione

Siano A e B gli eventi

A="il soggetto scelto è un neurologo"

B="il soggetto scelto è uno psichiatra"

Vogliamo calcolare la probabilità che su 2 soggetti estratti almeno uno sia un neurologo. Possiamo adottare 2 possibili strategie.

Strategia 1: l'evento "estraggo almeno 1 neurologo" è complementare all'evento "non estraggo alcun neurologo". Pertanto
P("almeno 1 sia un neurologo")=1-P("nessuno dei due è neurologo")=1-P("2 psichiatri").
Sia B_1 l'evento "seleziono uno psichiatra alla prima selezione" e B_2 l'evento "seleziono uno psichiatra alla seconda selezione". La probabilità richiesta à pertanto pari a :

$$1 - P(B_1 \cap B_2) = 1 - P(B_1)P(B_2|B_1) = 1 - \frac{30}{54}\frac{29}{53} = 0.6960$$

Strategia 2: equivalentemente, dopo aver definito gli eventi:
A_1 = "seleziono un neurologo alla prima selezione"
A_2 = "seleziono un neurologo alla seconda selezione"
questa probabilità poteva essere calcolata come probabilità dell'unione dei seguenti eventi:

$$(A_1 \cap A_2) \cup (A_1 \cap B_2) \cup (B_1 \cap A_2)$$

ossia

$$P((A_1 \cap A_2) \cup (A_1 \cap B_2) \cup (B_1 \cap A_2)) = P((A_1 \cap A_2)) + P((A_1 \cap B_2)) + P((B_1 \cap A_2)) = 0.6960$$

poiché

- $P(A_1 \cap A_2) = P(A_1)P(A_2|A_1) = \frac{24}{54}\frac{23}{53} = 0.1929$
- $P(A_1 \cap B_2) = P(A_1)P(A_2|A_1) = \frac{24}{54}\frac{30}{53} = 0.2516$
- $P(B_1 \cap A_2) = P(B_1)P(A_2|B_1) = \frac{30}{54}\frac{24}{53} = 0.2516$

• • •

Esercizio 2.63.

Su un tavolo ci sono 2 monete. Quando vengono lanciate, una moneta dà testa con probabilità 0.5 mentre l'altra dà testa con probabilità 0.6. Una moneta viene scelta a caso e lanciata.

1. Quale è la probabilità che esca testa?
2. Se esce croce, quale è la probabilità che fosse la moneta equilibrata?

• • •

Soluzione

Siano

M_1="la moneta scelta è la moneta 1"

M_2="la moneta scelta è la moneta 2"

Il testo afferma che $P(T|M_1) = 0.5$ e $P(T|M_2) = 0.6$.

1. $P(T) = P(T|M_1)P(M_1) + P(T|M_2)P(M_2) = 0.5 * 0.5 + 0.6 * 0.5 = 0.55$
2. Si vuole calcolare la probabilità che essendo uscita croce sia stata estratta la moneta 1; applicando il teorema di Bayes

$$P(M_1|C) = \frac{P(C|M_1)P(M_1)}{P(C|M_1)P(M_1) + P(C|M_2) * P(M_2)} = \frac{0.5 * 0.5}{(0.5 * 0.5) + (0.4 * 0.5)} = 0.55$$

• • •

Esercizio 2.64.

Tra i partecipanti ad un concorso per giovani compositori il 50% suona il pianoforte, il 30% suona il violino e il 20% la chitarra. Partecipano ad un concorso per la prima volta il 10% dei pianisti, il 33% dei violinisti e il 10% dei chitarristi. Applicando i concetti di probabilità condizionata e il teorema di Bayes, rispondere alle seguenti domande.

1. Quale è la probabilità che un compositore scelto a caso sia un aspirante alla prima esperienza?
2. Sapendo che ad esibirsi per primo sarà un compositore alla prima esperienza, quale è la probabilità che sia un chitarrista?

• • •

Soluzione

Definiamo gli eventi:

A = "Un partecipante scelto a caso è un aspirante compositore alla prima esperienza"

B = "Un partecipante scelto a caso è un pianista"

C = "Un partecipante scelto a caso è un violinista"

D = "Un partecipante scelto a caso è un chitarrista"

abbiamo

$$\begin{aligned} P(A) &= P(A \cap S) = P(A \cap (B \cup C \cup D)) \\ &= P(A \cap B) + P(A \cap C) + P(A \cap D) \\ &= P(A|B)P(B) + P(A|C)P(C) + P(A|D)P(D) \\ &= 0.1 \cdot 0.5 + 0.33 \cdot 0.3 + 0.1 \cdot 0.2 = 0.17 \end{aligned}$$

Per quanto riguarda il secondo quesito abbiamo:

$$\begin{aligned} P(D|A) &= \frac{P(D \cap A)}{P(A)} \\ &= \frac{P(A|D)P(D)}{P(A)} = \frac{0.1 \cdot 0.2}{0.17} = 0.12 \end{aligned}$$

• • •

Esercizio 2.65.

Un esame del sangue riconosce una certa malattia nel 99% dei casi quando essa è in atto. Tuttavia, l'esame fornisce un *falso positivo* (esito positivo quando la malattia non è in atto) nel 2% dei pazienti. Supponiamo che 0.5% della popolazione abbia la malattia. Quale è la probabilità che una persona scelta a caso abbia effettivamente la malattia se il test è positivo?

• • •

Soluzione

Indichiamo rispettivamente con D ed E gli eventi

D = *un soggetto estratto casualmente ha la malattia*

E= *il test è positivo*

Il testo ci dice che il test è affidabile al 99%, ossia fornisce un esito positivo quando il soggetto è effettivamente malato. Ciò significa che

$$P(E|D) = 0.99$$

Tuttavia, l'esame fornisce un *falso positivo* nel 2% dei casi, ossia

$$P(E|D^c) = 0.02$$

Sapendo che P(D)=0.005, per determinare $P(D|E)$ possiamo utilizzare il teorema di Bayes come segue:

$$P(D|E) = \frac{P(E|D)P(D)}{P(E|D)P(D) + P(E|D^c)P(D^c)} = \frac{0.99 \cdot 0.005}{0.99 \cdot 0.005 + 0.02 \cdot 0.995} = 0.199$$

Risulta quindi che una persona scelta a caso che ottiene risultato positivo al test ha una probabilità del 20% di avere effettivamente la malattia.

• • •

Esercizio 2.66.

Il dottore di Anna è preoccupato che lei possa soffrire di diabete gestazionale (alto livello di glucosio nel sangue durante la gravidanza). È presente una certa variabilità sia nel reale livello di glucosio nel sangue, sia nel risultati del test che lo misura. Una paziente è affetta da diabete gestazionale se il livello di glucosio, un'ora dopo aver ingerito una bevanda zuccherata, è superiore ai 140 milligrammi per decilitro (mg/dl). Il livello di glucosio di Anna varia secondo una distribuzione Normale con media $\mu = 125$ mg/dl e $\sigma = 10$ mg/dl.

1. Se si fa una singola misurazione di glucosio, quale è la probabilitè che ad Anna sia diagnosticato il diabete gestazionale?

2. Se invece le misurazioni sono fatte su 4 giorni separati e la regola dei 140 mg/dl viene applicata alla media delle 4 misurazioni, quale è la probabilità che ad Anna venga diagnosticato il diabete gestazionale?

3. Se invece si facessero 10 misurazioni, come cambierebbe in termini di media e deviazione standard la distribuzione della media?

• • •

Soluzione

1. Sia X la variabile aleatoria "livello di= glucosio nel sangue"; sappiamo che $X \sim N(\mu = 125, \sigma = 10)$. Vogliamo calcolare la probabilità che $X > 140$. Possiamo calcolare tale probabilità come segue:

$$\begin{aligned} Pr(X > 140) = 1 - Pr(X \leq 140) &= \\ 1 - Pr\left(\frac{X-\mu}{\sigma} \leq \frac{140-\mu}{\sigma}\right) &= \\ 1 - Pr\left(Z \leq \frac{140-125}{10}\right) &= \\ 1 - Pr(Z \leq 1.5) &= \\ 1 - 0.9332 = 0.0668 \end{aligned}$$

2. Per il teorema del limite centrale, le medie campionarie $\bar{X}$ avranno anch'esse distribuzione normale con media μ e deviazione standard $\frac{\sigma}{\mu}$, ossia

$$\bar{X} \sim N\left(\mu_{\bar{X}} = 125, \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = 5\right)$$

Pertanto, la probabilità richiesta :

$$\begin{aligned} Pr(\bar{X} > 140) = 1 - Pr(\bar{X} \leq 140) &= \\ 1 - Pr\left(\frac{X - \mu_{\bar{X}}}{\sigma_{\bar{X}}} \leq \frac{140 - \mu_{\bar{X}}}{\sigma_{\bar{X}}}\right) &= \\ 1 - Pr\left(\frac{X - \mu}{\frac{\sigma}{\sqrt{n}}} \leq \frac{140 - \mu}{\frac{\sigma}{\sqrt{n}}}\right) &= \\ 1 - Pr\left(Z \leq \frac{140 - 125}{\frac{10}{2}}\right) &= \\ 1 - Pr(Z \leq 3) &= \\ 1 - 0.9987 = 0.0013 \end{aligned}$$

3. Con un campione di numerosità $n = 10$ la distribuzione delle medie campionarie è piú concentrata rispetto alla media delle singole osservazioni; poiché la deviazione standard delle medie campionarie ha deviazione standard pari a $\frac{\sigma}{\sqrt{n}}$, essa diminuirà al crescere di n. Pertanto per $n = 10$, la media della distribuzione delle medie campionarie rimane invariata, mentre la deviazione standard si riduce e sarà pari a $\frac{10}{\sqrt{10}} = 3.162$.

• • •

Esercizio 2.67.

Negli USA, la tariffa pagata dalle famiglie ai provider di Internet è piuttosto variabile, ma la quota media mensile è di 28 dollari e la deviazione standard di 10. La distribuzione non è Normale: molte famiglie pagano circa 10 dollari per un accesso limitato oppure circa 25 dollari per un accesso illimitato, ma ve ne sono alcune che pagano molto di piú per connessioni veloci. In una indagine campionaria si intervista un campione casuale di 500 famiglie con accesso a internet. Quale è la probabilità che la tariffa media pagata dal campione di famiglie sia maggiore di 29 dollari?

• • •

Soluzione

Sia X la variabile casuale "tariffa pagata dalle famiglie residenti negli USA per provider di Internet"; la distribuzione di X nella popolazione ha media $\mu = 28$ e deviazione standard $\sigma = 10$. Tale distribuzione è non Normale. Tuttavia, il teorema del limite centrale ci garantisce che qualunque sia la distribuzione di X nella popolazione, se la dimensione del campione n è elevata, la media campionaria $\bar{X}$ ha distribuzione Normale con media μ e deviazione standard $\frac{\sigma}{n}$.
Pertanto, per il teorema del limite centrale, $\bar{X} \sim N\left(\mu = 28, \sigma = \frac{10}{\sqrt{500}} = 0.447\right)$ e la probabilità richiesta è pari a:

$$\begin{aligned} Pr(\bar{X} > 29) = 1 - Pr(\bar{X} \leq 29) &= \\ 1 - Pr\left(\frac{X - \mu_{\bar{X}}}{\sigma_{\bar{X}}} \leq \frac{29 - \mu_{\bar{X}}}{\sigma_{\bar{X}}}\right) &= \\ 1 - Pr\left(\frac{X - \mu}{\frac{\sigma}{\sqrt{n}}} \leq \frac{29 - \mu}{\frac{\sigma}{\sqrt{n}}}\right) &= \\ 1 - Pr\left(Z \leq \frac{29 - 28}{0.447}\right) &= \\ 1 - Pr(Z \leq 2.237) &= \\ 1 - 0.9875 = 0.0125 & \end{aligned}$$

• • •

Esercizio 2.68.
Il 52% degli elettori di una certa città americana sono Repubblicani, e il 48% sono Democratici. Tra questi elettori, 64% dei Repubblicani e 42% dei Democratici sono contrari alle politiche di agevolazione alle assunzioni di persone svantaggiate nella città. Viene scelto un elettore a caso.

1. Quale è la probabilità che la persona scelta sia contraria alle agevolazioni?

2. Se la persona scelta è a favore delle agevolazioni, quale è la probabilità che si tratti di un Repubblicano?

• • •

Soluzione

Indichiamo con R e C i seguenti eventi:

R= *un soggetto estratto casualmente è repubblicano*

C= *un soggetto estratto casualmente è contrario alle agevolazioni*

Il testo ci dice che P(R)=0.52, $P(R^c) = 0.48$, $P(C|R) = 0.64$ e $P(C|R^c) = 0.42$.

1. Calcoliamo la probabilità $P(C)$ come segue:

$$\begin{aligned} P(C) = P(C \cap (R \cup R^c)) &= \\ P(C \cap R) + P(C \cap R^c) &= \\ P(C|R)P(R) + P(C|R^c)P(R^c) &= \\ 0.64 \cdot 0.52 + 0.42 \cdot 0.48 = 0.5344 & \end{aligned}$$

2. Dobbiamo calcolare $P(R|C^c)$; sapendo che $P(C^c) = 1 - P(C) = 0.4656$, mediante il teorema di Bayes si ha che

$$P(R|C^c) = \frac{P(C^c|R)P(R)}{P(C^c)} = \frac{(1 - P(C|R))P(R)}{P(C^c)} = \frac{0.36 \cdot 0.52}{0.4656} = 0.4020$$

• • •

Esercizio 2.69.

L'urna 1 contiene 4 biglie rosse e 3 biglie blu, e l'urna 2 contiene 2 biglie rosse e 2 blu. Una biglia viene scelta a caso dall'urna 1 e inserita nell'urna 2. Poi viene estratta una biglia dall'urna 2.

1. Quale è la probabilità che la biglia estratta dall'urna 2 sia rossa?
2. Quale è la probabilità che la biglia estratta dall'urna 1 sia rossa se la biglia estratta dall'urna 2 è blu?

• • •

Soluzione

Indichiamo con

R_1= *la biglia estratta dall'urna 1 è rossa*

B_1= *la biglia estratta dall'urna 1 è blu*

R_2= *la biglia estratta dall'urna 2 è rossa*

B_2= *la biglia estratta dall'urna 2 è blu*

1. Dobbiamo calcolare $P(R_2)$. Si noti che dopo la prima estrazione la composizione dell'urna 1 cambia: se si estrae dall'urna 1 una biglia rossa, l'urna 2 conterrà 5 palline, 3 rosse e 2 blu. Se invece la biglia estratta dall'urna 1 blu, allora l'urna 2 conterrà 5 palline, 2 rosse e 3 blu. Possiamo quindi calcolare la probabilità richiesta come segue:

$$\begin{aligned} P(R_2) = P(R_2 \cap (R_1 \cup B_1)) &= \\ P(R_2 \cap R_1) + P(R_2 \cap B_1) &= \\ P(R_2|R_1)P(R_1) + P(R_2|B_1)P(B_1) &= \\ \frac{3}{5}\frac{4}{7} + \frac{2}{5}\frac{3}{7} = 0.5143 & \end{aligned}$$

2. Applichiamo il teorema di Bayes come segue:

$$P(R_1|B_2) = \frac{P(B_2|R_1)P(R_1)}{P(B_2|R_1)P(R_1) + P(B_2|B_1)P(B_1)} = \frac{\frac{2}{5}\frac{4}{7}}{\frac{2}{5}\frac{4}{7} + \frac{3}{5}\frac{3}{7}} = 0.4706$$

• • •

Esercizio 2.70.

Vengono lanciati 4 dadi. Trovare la probabilità che:

1. Il 6 esca almeno 1 volta;
2. Il 6 esca esattamente 1 volta;
3. Il 6 esca almeno 2 volte.

• • •

Soluzione

Indichiamo con X la variabile aleatoria X = *numero di 6 in 4 prove*. La variabile X definisce il numero di successi s in n prove: X ha pertanto distribuzione binomiale con $n = 4$ e probabilità di successo (ossia probabilità di fare 6) $p = \frac{1}{6}$.

1.

$$\begin{aligned} P(X \geq 1) &= 1 - P(X < 1) = 1 - P(X = 0) \\ &= 1 - \binom{4}{0} \left(\frac{1}{6}\right)^0 \left(1 - \frac{1}{6}\right)^4 \\ &= 1 - \frac{4!}{0!\,4!} \cdot 1 \cdot \left(\frac{5}{6}\right)^4 \\ &= 1 - \left(\frac{5}{6}\right)^4 = (1 - 0.4822) = 0.5178 \end{aligned}$$

2.

$$\begin{aligned} P(X = 1) &= \binom{4}{1} \left(\frac{1}{6}\right)^1 \left(1 - \frac{1}{6}\right)^3 \\ &= \frac{4!}{1!\,3!} \cdot \frac{1}{6} \cdot \left(\frac{5}{6}\right)^3 \\ &= \frac{4 \cdot 3 \cdot 2 \cdot 1}{1 \cdot 3 \cdot 2 \cdot 1} \cdot \frac{1}{6} \cdot \left(\frac{5}{6}\right)^3 = 4 \cdot \frac{1}{6} \cdot \left(\frac{5}{6}\right)^3 = 0.3858 \end{aligned}$$

3.

$$\begin{aligned} P(X \geq 2) &= 1 - P(X \leq 1) = 1 - (P(X = 0) + P(X = 1)) \\ &= 1 - (0.4822 + 0.3858) = 0.1320 \end{aligned}$$

• • •

Esercizio 2.71.

La probabilità che un tiratore ha di centrare un bersaglio sparando un colpo è 0.23. Si indichi con X la variabile casuale che descrive il numero di tiri al bersaglio in 8 colpi sparati.

1. Qual è la probabilità che in 8 colpi sparati, nessuno centri il bersaglio?
2. Qual è la probabilità che in 8 colpi sparati, almeno 1 centri il bersaglio?
3. Determinare la media e la varianza di X.

• • •

Soluzione

La variabile casuale che descrive il numero di tiri al bersaglio in 8 colpi sparati è una variabile casuale Binomiale nella quale il numero delle prove è $n = 8$ e la probabilità di successo è $p = 0.23$, pertanto $X \sim B(8, 0.23)$.

1. $P(X = 0) = \binom{8}{0} 0.23^0 (1 - 0.23)^{8-0} = \frac{8!}{0!8!} \cdot 0.23^0 \cdot 0.77^8 = 0.1236$
2. $P(X \geq 1) = 1 - P(X < 1) = 1 - P(X = 0) = 0.8764$
3. Il valore atteso e la varianza sono rispettivamente:

$$E[X] = np = 8 \cdot 0.23 = 1.84$$

$$Var[X] = np(1 - p) = 8 \cdot 0.23 \cdot 0.77 = 1.4168$$

• • •

Esercizio 2.72.

Uno stabilimento ha 6 macchinari che usano in media energia elettrica per 20 minuti ogni ora.

a. Se i macchinari vengono usati indipendentemente, mostrare che la probabilità che 4 o più macchinari usino energia elettrica contemporaneamente è 0.1.

b. Se lo stabilimento avesse 60 macchinari, quale sarebbe la probabilità di avere al massimo 30 macchinari in funzione contemporaneamente?

c. Sempre considerando 60 macchinari trovare un numero approssimato r, tale che la probabilità che più di r macchinari usino energia elettrica allo stesso tempo sia 0.1.

• • •

Soluzione

a. Consideriamo la variabile casuale X="*numero di macchine che consumano energia*". Possiamo assumere che X abbia una distribuzione binomiale con parametri (n, p), dove

   ⋆ $n = 6$ è pari al numero di macchinari disponibili,

   ⋆ $p = \frac{20}{60} = \frac{1}{3}$ è la probabilità di successo, dove per successo intendiamo il fatto che una macchina consumi energia.

A questo punto, poiché $X \sim Bin(n,p)$ la probabilità richiesta è

$$\begin{aligned}
P(X \geq 4) &= P(X=4) + P(X=5) + P(X=6) = \\
&= \binom{6}{4}\left(\frac{1}{3}\right)^4\left(\frac{2}{3}\right)^{6-4} + \binom{6}{5}\left(\frac{1}{3}\right)^5\left(\frac{2}{3}\right)^{6-5} + \binom{6}{6}\left(\frac{1}{3}\right)^6\left(\frac{2}{3}\right)^{6-6} = \\
&= \frac{6!}{4!\,2!}\left(\frac{1}{3}\right)^4\left(\frac{2}{3}\right)^{6-4} + \frac{6!}{5!\,1!}\left(\frac{1}{3}\right)^5\left(\frac{2}{3}\right)^{6-5} + \frac{6!}{6!\,0!}\left(\frac{1}{3}\right)^6\left(\frac{2}{3}\right)^{6-6} = \\
&= \frac{6\cdot5\cdot4\cdot3\cdot2\cdot1}{4\cdot3\cdot2\cdot1\cdot2\cdot1}\left(\frac{1}{3}\right)^4\left(\frac{2}{3}\right)^2 + \frac{6\cdot5\cdot4\cdot3\cdot2\cdot1}{5\cdot4\cdot3\cdot2\cdot1\cdot1}\left(\frac{1}{3}\right)^5\left(\frac{2}{3}\right)^1 \\
&+ \frac{6\cdot5\cdot4\cdot3\cdot2\cdot1}{6\cdot5\cdot4\cdot3\cdot2\cdot1}\left(\frac{1}{3}\right)^6\left(\frac{2}{3}\right)^0 = \\
&= 15\cdot\left(\frac{1}{3}\right)^4\left(\frac{2}{3}\right)^2 + 6\cdot\left(\frac{1}{3}\right)^5\left(\frac{2}{3}\right) + \left(\frac{1}{3}\right)^6 = \\
&= 0.0823 + 0.0165 + 0.0014 = 0.1002
\end{aligned}$$

b. Se consideriamo 60 macchinari, con la stessa probabilità di successo p sappiamo che $X \sim Bin(n=60, p=\frac{1}{3})$. In questo caso, per calcolare la probabilità richiesta, possiamo ricorrere all'approssimazione normale della distribuzione binomiale, ovvero considerare X distribuita approssimativamente come $N(np, np(1-p))$, dove

   - ⋆ la media è $np = 60\frac{1}{3} = 20$
   - ⋆ la varianza è $np(1-p) = 60\frac{1}{3}\frac{2}{3} = 13.33$ e quindi la deviazione standard è 3.65.

   A questo punto la probabilità di avere al massimo 30 macchinari in funzione è

$$P(X \leq 30) \cong \Phi\left(\frac{30+0.5-20}{3.65}\right) = \Phi(2.877) = 0.998$$

   Da notare:

   - ⋆ la regola empirica per controllare la validità dell'approssimazione normale $np = 20 \geq 10$, $n(1-p) = 40 \geq 10$ è soddisfatta.

⋆ nel calcolo della probabilità di interesse è stata utilizzata la correzione di continuità.

c. Usando ancora l'approssimazione normale, dobbiamo determinare un numero r tale che

$$P(X > r) = 0.1$$

Standardizzando (e usando di nuovo la correzione di continuità)

$$P(X > r) = 1 - P(X \leq r) \cong 1 - \Phi\left(\frac{r+0.5 - 20}{3.65}\right) = 0.1 \iff$$

$$\iff \Phi\left(\frac{r - 19.5}{3.65}\right) = 0.9 \iff z_{0.9} = \frac{r - 19.5}{3.65} \iff r = 3.65 z_{0.9} + 19.5$$

Leggendo dalle tavole il valore di $z_{0.9} = 1.28$, otteniamo quindi $r = 3.65 \cdot 1.28 + 19.5 = 24.172 \approx 24$

• • •

2.6 Distribuzioni di variabili casuali

2.6.1 Distribuzione Normale

Esercizio 2.73. *Area sotto la curva normale, parte I.*
Qual è la probabilità che una v.a. normale standard assuma un valore compreso nei seguenti insiemi? Disegnare un grafico può aiutare nella risposta.

(a) $Z < -1.35$

(b) $Z > 1.48$

(c) $-0.4 < Z < 1.5$

(d) $|Z| > 2$

Esercizio 2.74. *Area sotto la curva normale, parte II.*
Qual è la probabilità che una v.a. normale standard assuma un valore compreso nei seguenti insiemi? Disegnare un grafico può aiutare nella risposta.

(a) $Z > -1.13$

(b) $Z < 0.18$

(c) $Z > 8$

(d) $|Z| < 0.5$

Esercizio 2.75. *Punteggi al test GRE, Parte I.*
Uno studente senior di college ha sostenuto l'esame Graduate Record Examination, in breve GRE, ed ha ottenuto il punteggio di 620 nella prova di *Ragionamento Verbale* e 670 nella prova di *Ragionamento Quantitativo*. Il punteggio medio per la prova di *Ragionamento Verbale* è di 462 con una deviazione standard pari a 119, mentre il punteggio medio per la prova di *Ragionamento Quantitativo* è pari a 584 con una deviazione standard pari a 151. Supponiamo che entrambe le distribuzioni siano approssimativamente normali.

(a) Scrivi in simboli le distribuzioni relative alle due grandezze.

(b) Qual è il punteggio standardizzato che lo studente ha ottenuto nella prova di *Ragionamento Verbale*? e quello ottenuto nella prova di *Ragionamento Quantitativo*? Disegna una curva normale standardizzata ed identifica i due punteggi standardizzati.

(c) Che cosa ti dicono questi due punteggi?

(d) In confronto agli altri studenti, in quale prova lo studente si è comportato meglio?

(e) Calcola le prestazioni dello studente in termini di percentili nelle due prove.

(f) Quale percentuale di esaminandi hanno fatto meglio di lui nella prova di *Ragionamento Verbale*? e in quella di *Ragionamento Quantitativo*?

(g) Spiegare perché la semplice comparazione dei punteggi originali nelle due prove farebbe dire, in modo non corretto, che lo studente si è meglio comportato nella prova di *Ragionamento Quantitativo*.

(h) Se la distribuzione dei punteggi in questi due esami non fosse approssimativamente normale, le tue risposte ai punti da (b) a (f) cambierebbero o rimarrebbero uguali? Spiega il tuo ragionamento.

Esercizio 2.76. *Prova di Triathlon, Parte I.*
Nel triathlon, è tipico per i concorrenti essere raggruppati in fasce di età e genere. Leonardo e Maria sono due amici che hanno completato la famosa prova di Hermosa Beach. Leonardo ha gareggiato nella categoria uomini di età $30--34$ anni. Maria ha gareggiato nella categoria donne di età $25--29$ anni. Leonardo ha terminato la gara con il tempo di 1 ora, 22 minuti e 28 secondi (cioè 4948 secondi), mentre Maria ha terminato la gara con il tempo di 1 ora, 31 minuti e 53 secondi (5513 secondi). Come era prevedibile, Leonardo è stato più veloce ma entrambi sono curiosi di sapere come si sono comportati relativamente ai loro concorrenti di fascia. Puoi aiutarli? Queste informazioni ti possono essere utili.

- I tempi impiegati dagli uomini di età $30--34$ hanno una media pari a 4313 secondi con una deviazione standard di 583 secondi.

- I tempi impiegati dalle donne di età $25--29$ hanno una media pari a 5261 secondi con una deviazione standard di 807 secondi.

- Per entrambe le fasce di concorrenti, i tempi di percorrenza possono essere considerati approssimativamente normali.

(a) Scrivere in simboli che distribuzione hanno le due grandezze sopra descritte.

(b) Quali sono i punteggi Z standardizzati per Leonardo e Maria? Cosa suggeriscono questi punteggi?

(c) Chi si è comportato meglio tra i due, dopo aver "aggiustato" i risultati tenendo conto di età e genere? Spiega il tuo ragionamento.

(d) Quale percentuale di triatleti ha fatto meglio di Leonardo nel suo gruppo?

(e) Quale percentuale di triatlete ha fatto meglio di Maria nel suo gruppo?

(f) Se la distribuzione dei tempi di percorrenza in queste due prove non fosse approssimativamente normale, le tue risposte ai punti da (b) a (e) cambierebbero o rimarrebbero uguali? Spiega il tuo ragionamento.

Esercizio 2.77. *Punteggi al test GRE, Parte II.*
Nell'Esercizio 2.75 abbiamo lavorato con due distribuzioni normali relative ai punteggi ottenuti nel test GRE: $N(\mu = 462, \sigma = 119)$ per la parte verbale del test e $N(\mu = 584, \sigma = 151)$ per la parte quantitativa.

Usa questa informazione per calcolare le seguenti grandezze.

(a) Il punteggio ottenuto da uno studente che si trova all'80-esimo percentile nella distribuzione relativa al test *quantitativo*.

(b) Il punteggio ottenuto da uno studente che si trova al 70-esimo percentile nella distribuzione relativa al test *verbale*.

Esercizio 2.78. *Prova di Triathlon, Parte II.*
Nell'Esercizio 2.76 abbiamo lavorato con due distribuzioni normali relative ai tempi ottenuti in due fasce di età e genere, espresse in secondi; $N(\mu =$

$4313, \sigma = 583$) per gli uomini di età $30 - -34$ e $N(\mu = 5261, \sigma = 807)$ per le donne di età $25 - -29$.

Usa questa informazione per calcolare le seguenti grandezze.

(a) Il tempo massimo necessario per entrare nel gruppo del 5% più veloce tra gli uomini di età $30 - -34$.

(b) Il tempo minimo per entrare nel gruppo del 10% delle donne di età $25 - -29$ più lente.

Esercizio 2.79. *Temperature a Los Angeles, Parte I.*
Nel mese di giugno la temperatura media giornaliera a Los Angeles è di 77 gradi Fahrenheit, in breve 77F. Ricorda che la temperatura F si ottiene da quella espressa in gradi Celsius (C) attraverso la trasformazione lineare $F = 32 + 9/5C$. La deviazione standard è pari a 5F. Supponiamo inoltre che le temperature di Giugno possano essere considerate approssimativamente normale.

(a) Qual è la probabilità che in una giornata scelta a caso di Giugno si abbia una temperatura media di 83F o maggiore a Los Angeles?

(b) Qual è il livello di temperatura media a Los Angeles Y a giugno tale per cui in 95 giorni su 100 si avrà una temperatura media più calda di Y?

Esercizio 2.80. *Rendimenti di Portafoglio.*
Il modello CAPM (Capital Asset Pricing Model) è un modello usato in finanza in cui si assume che i rendimenti di un portafoglio sono distribuiti in modo normale. Supponiamo che un certo portafoglio abbia un rendimento medio annuo stimato pari al 14.7% con una deviazione standard del 33%. Tieni conto che un rendimento pari allo 0% implica che il valore di un portafoglio non cambia, che un rendimento negativo implica che il portafoglio *perde* denaro, e un rendimento positivo significa che il portafoglio *guadagna* denaro.

(a) Qual è la frequenza relativa di anni in cui questo portafoglio perde denaro, ovvero ha un rendimento negativo?

(b) Qual è il il punto di cut-off relativo al 15% più elevato dei rendimenti? Ovvero, in una classifica dei rendimenti, qual è il rendimento che si piazzerebbe al quindicesimo posto su 100?

Esercizio 2.81. *Temperature a Los Angeles, Parte II.*
Nell'Esercizio 2.79 si diceva che la temperatura media a Los Angeles nel mese di giugno è pari a $77F$ con una deviazione standard di $5F$, e si può assumere che la distribuzione delle temperature medie è approssimativamente normale. Ricordiamo che la formula di trasformazione delle temperature da F (Fahrenheit) a C (Celsius) è:

$$C = (F - 32) \times 5/9.$$

(a) Qual è la distribuzione delle temperature medie a giugno a Los Angeles espressa in gradi Celsius?

(b) Qual è la probabilità che in un giorno a caso di giugno, a Los Angeles, si abbia una temperatura di $28C$ o più alta (tieni conto che $28C \approx 83F$)? Per calcolarla, usa l'espressione ottenuta al punto (a).

(c) Le risposte fornite al punto (b) precedente e nella parte (a) dell'Esercizio 2.79 sono uguali oppure no? Spiegare perché debbono (o non debbono) essere uguali.

Esercizio 2.82. *Le altezze dei bambini a 10 anni.*
A dieci anni, indipendentemente dal genere, la distribuzione delle altezze segue una distribuzione normale di media 139.7 centimetri con deviazione standard di 15 centimetri.

(a) Qual è la probabilità che un bambino di 10 anni scelto a caso sia alto meno di 122 cm.?

(b) Qual è la probabilità che un bambino di 10 anni scelto a caso abbia un'altezza compresa tra 152 e 165 cm.?

(c) Il 10% dei bambini più alti viene classificato come "molto alto". Qual è il punto di cut-off per entrare in questa categoria? In altri termini qual è l'altezza minima necessaria per essere classificato come "molto alto"?

(d) Al Luna Park, per essere ammessi alle montagne russe, bisogna essere alti almeno 137 cm.; quale percentuale di bambini di 10 anni viene esclusa dalle montagne russe?

Esercizio 2.83. *Premi assicurativi.*
Sul quotidiano di ieri un articolo sosteneva che la distribuzione dei premi di assicurazione auto per i residenti della California è all'incirca normale con media pari a \$1650. L'articolo sostiene anche che il 25% dei residenti della California paga più di \$1800.

(a) Qual è il punteggio standardizzato Z che corrisponde al 75-esimo percentile della distribuzione?

(b) Determina la deviazione standard della distribuzione dei premi.

Esercizio 2.84. *Velocità sulle autostrade, Parte I.*
In un dato tratto della autostrada A1, sono state registrate le velocità medie di un gran numero di automobili. Si è ottenuta una media di 140 km/h. con una deviazione standard di 13 km/h.

(a) Quale percentuale di autoveicoli ha una velocità media inferiore a 110 km/h?

(b) Quale percentuale di autoveicoli ha una velocità media compresa tra 1 120 e i 130 km/h?

(c) Qual è la velocità media necessaria per essere classificato tra il 5% delle auto più veloci?

(d) Il limite di velocità sulla A1 è di 130 km/h. In termini approssimati, qual è la percentuale di automobili che viola tale limite?

Esercizio 2.85. *Bagagli troppo pesanti.*
La distribuzione dei pesi dei bagagli dei passeggeri su un certo volo di linea è approssimativamente normale con media 20.5 kg.e deviazione standard 1.45 kg. La compagnia aerea impone un sovrapprezzo per i bagagli che pesano più di 23 kg. Determina quale percentuale di passeggeri incorrerà nel sovrapprezzo.

Esercizio 2.86. *Deviazioni standard.*
Calcola la deviazione standard delle seguenti distribuzioni

(a) Il MENSA è un'organizzazione i cui membri hanno un quoziente di intelligenza (QI) che va oltre il 98-esimo percentile della distribuzione del QI nell'intera popolazione. I QI sono misurati secondo un punteggio convenzionale e sono distribuiti normalmente nella popolazione, con media pari a 100. Il minimo QI richiesto per essere ammesso al MENSA è pari a 132.

(b) Il livello di colesterolo tra le donne di età nella fascia $20 - -34$ segue una distribuzione normale con media di 185 milligrammi per decilitro (mg/dl). Le donne con un livello di colesterolo superiore alla soglia di 220 mg/dl sono considerate a rischio di complicazioni sanitarie e circa il 18.5% delle donne supera tale soglia.

Esercizio 2.87. *Acquisti su Ebay*
Devi acquistare il libro di testo XXX per un certo corso. Poiché il testo costa molto in libreria, stai considerando l'ipotesi di acquistarlo su Ebay. Osservando alcune precedenti transazioni relative al testo XXX, puoi assumere che la distribuzione dei prezzi di vendita di XXX su Ebay è approssimativamente normale con media pari a 89 euro e deviazione standard pari a 15 euro.

(a) Qual è la probabilità che in una transazione scelta a caso il prezzo del libro venga fissato sopra i 100 euro?

(b) Il sistema automatico di offerte di eBay ti consente di inserire un'offerta massima. Il sistema aumenterà automaticamente la tua offerta solo di quanto è necessario per consentirti di restare il miglior offerente, ma fino alla soglia che tu hai stabilito. Se tu sei impegnato in un'unica asta, quali sono i vantaggi e gli svantaggi di fissare un'offerta massima troppo alta o troppo bassa? Cosa cambia se tu stai seguendo più aste sullo stesso oggetto, ovvero il testo XXX?

(c) Supponiamo che tu, prima di entrare nell'asta, abbia seguito 10 aste in precedenza, sulle quali hai basato le assunzioni distributive. Ragionando in modo approssimato quale percentile potresti usare per la tua offerta massima, per essere praticamente certo di vincere l'asta? È possibile in pratica determinare un'offerta massima che ti dia la pratica certezza di vincere l'asta?

Esercizio 2.88. *Punteggi al test SAT.*
I punteggi al test SAT (Standard Assessment Test), necessario per essere ammessi in alcuni college americani, seguono una distribuzione normale con media pari a 1500 e deviazione standard di 300. L'assunzione di normalità è un po' forzata poiché il punteggio massimo raggiungibile è 2400 (e non infinito, come una distribuzione normale supporrebbe). Supponiamo che un comitato di college fornisca un certificato di eccellenza a quegli studenti che realizzano un punteggio superiore a 1900 nel test SAT.

(a) Prendendo uno studente a caso, qual è la probabilità che sia uno di quelli col certificato di eccellenza?

(b) Qual è la probabilità che quello stesso studente ottenga un punteggio di almeno 2100?

Esercizio 2.89. *Voti all'esame di statistica, Parte I.*
I voti di seguito riportati si riferiscono al voto finale dei 20 studenti che hanno superato l'esame di Statistica nell'ultimo appello .

$$18, 20, 21, 21, 22, 22, 22, 23, 23, 23$$

$$24, 24, 24, 24, 25, 25, 25, 26, 27, 28.$$

Il voto medio è di 23.35, con una deviazione standard di 2.4 punti. Usa questa informazione per determinare se i punteggi seguono effettivamente la regola del 68-95-99.7%.

Esercizio 2.90. *Altezza delle studentesse al college, Parte I.*

Qui di sotto ci sono le altezze in pollici di 25 studentesse di un college americano (1 pollice = 2.54 centimetri):

$$54, 55, 56, 56, 57, 58, 58, 59, 60, 60, 60, 61, 61,$$

$$62, 62, 63, 63, 63, 64, 65, 65, 67, 67, 69, 73.$$

L'altezza media è di 61.52 pollici con una deviazione standard di 4.58 pollici. Usa questa informazione per determinare se le altezze seguono effettivamente la regola del 68-95-99.7%.

2.6.2 La distribuzione geometrica

Esercizio 2.91. *Prove bernoulliane?*

Stabilire se le varie situazioni descritte di seguito possono essere considerate esempi di prove bernoulliane.

(a) Le cinque carte ottenute in una mano di poker.

(b) I risultati di cinque lanci consecutivi di un dado.

(c) I cinque tiri liberi consecutivi eseguiti da un giocatore di basket

(d) Gli ultimi 5 calci di rigore calciati da un certo calciatore.

Esercizio 2.92. *Con e senza ripetizione.*

Nelle situazioni che seguono si assume che metà dei componenti del collettivo siano uomini e metà donne.

(a) Supponiamo di dover scegliere a caso due persone da un gruppo di 10. Qual è la probabilità di scegliere due donne se si sceglie senza ripetizione? Qual è la stessa probabilità nel caso di campionamento con ripetizione?

(b) Adesso ci troviamo in uno stadio con 10 mila persone. Qual è la probabilità di scegliere due donne se si sceglie senza ripetizione? Qual è la stessa probabilità nel caso di campionamento con ripetizione?

- (c)] Spesso si considerano gli individui estratti da una popolazione come se fossero estratti con ripetizione anche se in realtà sono stati campionati senza ripetizione. Usa i risultati dei due punti precedenti per stabilire se tale assunzione può essere considerato ragionevole.

Esercizio 2.93. *Donne sposate.*
Nella edizione del 2010 dell'*American Community Survey* si è stimato che il 47.1% delle donne con età superiore a 15 anni risulta coniugata.

(a) Scegliamo a caso tre donne con più di 15 anni. Qual è la probabilità che solo la terza donna selezionata sia sposata?

(b) Qual è la probabilità che siano tutte e tre sposate?

(c) In media, quante donne ci aspettiamo di estrarre prima di trovarne una sposata? Qual è la deviazione standard associata a tale media?

(d) Se la proporzione di donne sposate fosse in realtà del 30%, quante donne ci aspettiamo in media di dover selezionare prima di trovarne una sposata?

(e) Con riferimento alle risposte fornite nei punti (c) e (d), quanto la diminuzione della frequenza di un evento (in questo caso la frequenza di osservare una donna sposata) modifica la media e la deviazione standard del numero di donne selezionate prima di ottenere un *successo*, in questo caso l'osservare una donna sposata?

Esercizio 2.94. *Tassi di errori.*
Una macchina produce uno speciale transistor, componente necessario nella costruzione dei computer. La macchina ha un tasso di errori del 2%, ovvero produce 2 pezzi difettosi ogni 100, in media. Il processo di produzione è tale che i vari pezzi possono essere considerati mutuamente indipendenti.

(a) Qual è la probabilità che il decimo transistor prodotto sia il primo a essere difettoso?

(b) Qual è la probabilità che nei primi 100 pezzi prodotti non vi sia alcun pezzo difettoso?

(c) In media, quanti transistor ti aspetti che vengano prodotti prima di osservare il primo difettoso? qual è la deviazione standard associata?

(d) Un'altra macchina, che produce anch'essa transistor, ha un tasso di errori del 5% e anche in questo caso i pezzi prodotti possono essere considerati indipendenti. In media, quanti transistor ti aspetti che vengano prodotti prima di osservare il primo difettoso? qual è la deviazione standard associata?

(e) Tenendo conto delle risposte fornite ai punti (c) e (d), quanto la diminuzione della probabilità di un evento influenza la media e la deviazione standard del numero di pezzi necessari a osservare il primo *successo* (la produzione di un pezzo difettoso)?

Esercizio 2.95. *Il colore degli occhi, Parte I.*
Marito e moglie hanno entrambi occhi castani ma il loro corredo genetico è tale che, potenzialmente, i loro figli potranno avere occhi di diversi colori. In particolare, ogni loro figlio avrà occhi castani con probabilità 0.75, occhi blu con probabilità 0.125, e occhi verdi con probabilità 0.125. Si assume che il colore degli occhi di ciascun figlio sia indipendente dal colore degli occhi degli altri figli.

(a) Qual è la probabilità che il loro primo figlio con occhi verdi sia il terzo?

(b) In media, quanti figli deve avere questa coppia prima di avere un figlio con occhi verdi? Qual è la deviazione standard della v.a. *numero di figli necessari per averne uno con occhi verdi*?

Esercizio 2.96. *Velocità sulle autostrade, Parte II.*
Nell'Esercizio 2.84 abbiamo visto come la distribuzione delle velocità medie in un certo tratto di autostrada fosse normale con una media di 140 km/h e con una deviazione standard di 13 km/h. La velocità limite in quel tratto di strada è di 130 km/h. Le velocità delle singole auto sono mutuamente indipendenti.

(a) Una macchina della polizia stradale è nascosta su un lato della A1. Qual è la probabilità che delle prime 5 macchine che passano, nessuna superi i limiti di velocità?

(b) In media, quante auto devono passare prima di osservare la prima che supera il limite di velocità? Qualè la deviazione standard del numero di auto necessarie a osservare la prima auto che supera il limite di velocità?

2.6.3 Distribuzione Binomiale

Esercizio 2.97.
La probabilità che un giocatore di basket segni un tiro libero è 0.43. Supponiamo che tiri 8 volte e che gli 8 lanci possano essere considerati mutuamente indipendenti.

(a) Qual è la probabilità che non segni mai?

(b) Qual è la probabilità che segni almeno una volta?

(c) Qual è la probabilità che segni 8 volte?

(d) Quanti tiri liberi segnerà in media?

(e) Supponiamo ora che il giocatore abbia a disposizione 50 tentativi. Qual è la probabilità che segni almeno 20 volte?

• • •

Soluzione

Definiamo una variabile aleatoria X che rappresenta il "numero di tiri liberi in 8 prove". X ha distribuzione Binomiale con parametri $n = 8$ $p = 0.43$, cioè $X \sim Binom(8, 0.43)$. Le probabilità richieste sono quindi:

(a) $P(X = 0) = \binom{n}{0} p^0 (1-p)^{n-0} = \binom{8}{0} 0.43^0 (1 - 0.43)^{8-0} = 0.011$

(b) $P(X \geq 1) = 1 - P(X < 1) = 1 - P(X = 0) = 0.989$

(c) $P(X = 8) = \binom{8}{8} 0.43^8 (1 - 0.43)^{8-8} = 0.001$

(d) il numero di tiri liberi segnati in media è pari al valore atteso

$$E[X] = np = 8 \cdot 0.43 = 3.44$$

(e) Se consideriamo 50 tentativi, con la stessa probabilità di segnare p sappiamo che $X \sim Binom(n = 50, p = 0.43)$. In questo caso, per calcolare la probabilità richiesta, possiamo ricorrere (è verificata la regola empirica $np > 10$ e $n(1-p) > 10$) all'approssimazione normale della distribuzione binomiale, ovvero considerare X distribuita approssimativamente come $N(np, np(1-p))$, dove $E(X) = np = 50 \cdot 0.43 = 21.5$ e $Var(X) = np(1-p) = 12.255$.
A questo punto la probabilità di segnare almeno 20 volte è

$$P(X \geq 20) = 1 - P(X < 20) \cong 1 - \Phi\left(\frac{20 - 0.5 - 21.5}{\sqrt{12.255}}\right) =$$

$$= 1 - \Phi\left(\frac{-2}{3.5}\right) = 1 - \Phi(-0.57) = 1 - 0.284 = 0.716$$

• • •

Esercizio 2.98. *Minorenni e alcool, Parte I.*
L'agenzia federale americana per il *Monitoraggio degli abusi di sostanze e la salute mentale* ha stimato che il 70% dei giovani nella fascia di età 16-18 anni ha consumato bevande alcoliche nel 2008.

(a) Supponiamo di estrarre a caso un campione di 10 minorenni tra i 16 e i 18 anni. Possiamo usare la distribuzione binomiale per calcolare la probabilità che esattamente sei di loro abbiano consumato alcool? Spiegare il perché.

(b) Calcolare la probabilità che esattamente sei di loro abbiano consumato alcool.

(c) Calcolare la probabilità che esattamente quattro di loro NON abbiano consumato alcool.

(d) Calcolare la probabilità che al più due minorenni, su un campione di 5, abbia consumato alcool.

(e) Calcolare la probabilità che almeno un minorenne, su un campione di 5, abbia consumato alcool.

Esercizio 2.99. *Varicella, Parte I.*
Il Centro Nazionale per i Vaccini informa che il 90% dei residenti adulti in Italia ha contratto la varicella prima dei 18 anni

(a) Supponiamo di considerare un campione di 100 residenti adulti in Italia. L'uso della distribuzione binomiale per calcolare la probabilità che esattamente 97 sui 100 selezionati abbia contratto la varicella prima dei 18 anni è appropriata? Spiegare.

(b) Calcolare la probabilità che esattamente 97 sui 100 selezionati abbia contratto la varicella prima dei 18 anni.

(c) Calcolare la probabilità che esattamente 3 residenti sui 100 di un nuovo campione estratto non abbiano avuto la varicella prima dei 18 anni.

(d) Calcolare la probabilità che almeno 1 residente sui 10 selezionati in un nuovo campione abbia contratto la varicella prima dei 18 anni.

(e) Calcolare la probabilità che al più 3 residenti sui 10 selezionati in un nuovo campione non abbiano contratto la varicella prima dei 18 anni.

Esercizio 2.100. *Minorenni e alcool, Parte II.*
Nell'Esercizio 2.98 si è visto come circa il 70% dei ragazzi nella fascia di età $16 - -18$ ha consumato bevande alcoliche nel 2008. Consideriamo ora un campione di 50 ragazzi in quella fascia di età

(a) Quante persone ti aspetti ci siano nel campione che hanno consumato alcool? Qual è la deviazione standard?

(b) Saresti sorpreso se nel campione ci fossero 45 o più persone che hanno consumato alcool?

(c) Qual è la probabilità che 45 o più persone nel campione hanno consumato alcool? Come si lega questa risposta a quella fornita al punto (b)?

Esercizio 2.101. *Varicella, Parte II.*
Nell'Esercizio 2.99 si è visto come circa il 90% degli adulti ha contratto la varicella prima dei 18 anni. Prendiamo ora un campione casuale di 120 adulti residenti.

(a) Quante persone ti aspetti ci siano nel campione che hanno contratto la varicella prima dei 18 anni? Qual è la deviazione standard?

(b) Saresti sorpreso se nel campione ci fossero 105 persone che hanno contratto la varicella prima dei 18 anni?

(c) Qual è la probabilità che nel campione ci siano AL PIU' 105 persone che hanno contratto la varicella prima dei 18 anni? Come si lega questa risposta a quella fornita al punto (b)?

Esercizio 2.102. *Ammissioni all'Università*
Una certa Università americana, ogni anno, ammette 2500 nuove matricole. I posti letto a disposizione sono soltanto 1786. Tuttavia, non tutti gli studenti ammessi decidono di accettare il posto letto: si ritiene che circa il 70% degli studenti ammessi utilizzerà il posto letto fornito dall'università.

(a) Qual è la probabilità (approssimata) che l'Università, all'inizio delle lezioni, non abbia sufficienti posti letto?

(b) Prendendo per buona la stima del 70% utilizzata al punto (a), quanti posti dovrebbe avere a disposizione l'Università per avere una probabilità del 95% di coprire tutte le richieste?

(c) Spiega perché questo problema è, da un punto di vista astratto, identico a quello dell'*overbooking* delle compagnie aeree.

Esercizio 2.103. *Tassi di risposta ad un'indagine.*
Una agenzia demoscopica ha riportato che, nel 2012, il tasso tipico di risposta degli intervistati alle indagini da loro svolte è stato circa del 9%. Se, per una particolare indagine, vengono contattate 15 mila famiglie, qual è la probabilità che rispondano almeno in 1500?
Esercizio 2.104. *Il dreidel.*
Il dreidel è una sorta di dado a quattro facce, con su scritte quattro lettere dell'alfabeto ebraico: `nun, gimel, hei, shin`, una su ogni lato. Il dreidel è regolare, ovvero ogni faccia ha la stessa probabilità e i lanci effettuati possono essere considerati mutuamente indipendenti. Lanciamo il dreidel 3 volte. Calcolare la probabilità di ottenere

(a) almeno un nun;

(b) esattamente 2 nun;

(c) esattamente 1 hei;

(d) al più 2 gimels.

Esercizio 2.105. *Aracnofobia.*
Una indagine del 2005 della Gallup, ha evidenziato come il 7% dei *teenager* (età compresa tra i 13 e i 17 anni) soffra di una qualche forma di aracnofobia ed sia particolarmente spaventata dai ragni. In un campo estivo ci sono 10 teenager che dormono in ogni tenda; possiamo assumere che i comportamenti dei vari ragazzi nei confronti dei ragni siano mutuamente indipendenti.

(a) Qual è la probabilità che esattamente 2 di loro soffrano di aracnofobia in una certa tenda?

(b) Qual è la probabilità che almeno 1 di loro soffra di aracnofobia in una certa tenda?

(c) Qual è la probabilità che al più 1 di loro soffra di aracnofobia in una certa tenda?

(d) Il responsabile del campo vuole essere sicuro che non ci sia più di un ragazzo con problemi di aracnofobia in ciascuna tenda: è ragionevole, allora, assegnare i posti in modo casuale oppure occorrerebbe una strategia diversa?

Esercizio 2.106. *Il colore degli occhi, Parte II.*
L'Esercizio 2.95 considerava una coppia, un uomo e una donna, entrambi con occhi castani. Ad ogni parto, essi hanno una probabilità pari a 0.75 di avere un bambino con occhi castani, pari a 0.125 di avere un bambino con occhi blu e 0.125 di avere un bambino con occhi verdi. I parti possono essere considerati indipendenti.

(a) Qual è la probabilità che il loro primo bambino abbia occhi verdi e il secondo non verdi?

(b) Qual è la probabilità che esattamente uno dei loro primi due figli abbia occhi verdi?

(c) Se la coppia ha sei figli, qual è la probabilità che esattamente due dei loro figli abbia occhi verdi?

(d) Se la coppia ha sei figli, qual è la probabilità che almeno 1 dei loro figli abbia occhi verdi?

(e) Qual è la probabilità che il loro primo figlio con occhi verdi sia il quarto?

(f) Se solo 2 dei loro 6 figli hanno occhi castani, riterresti questo evento improbabile?

Esercizio 2.107. *Anemia falciforme.*

L'anemia falciforme è una malattia genetica del sangue, che provoca un irrigidimento dei globuli rossi, che assumono una forma simile a quella della falce. Questa patologia può provocare diverse complicazioni. Se entrambi i genitori sono portatori della malattia, ogni loro figlio ha una probabilità di 0.25 di contrarre la malattia, di 0.50 di essere un portatore sano, e di 0.25 di essere sano e non portatore. Supponiamo che due genitori portatori abbiano tre figli. Calcolare la probabilità che

(a) due di loro abbiano la malattia;

(b) nessuno abbia la malattia;

(c) almeno uno sia sano e non portatore;

(d) il primo figlio a contrarre la malattia sia il terzo.

Esercizio 2.108. *La roulette.*

Nella roulette, si può scommettere su diversi eventi; tra questi, sul colore del numero che uscirà. Su 37 slots (i numeri da 0 a 36) ce ne sono 18 rossi. Se si scommette sul rosso ed esce un numero rosso, per ogni euro scommesso se ne vince un altro. Se non esce il rosso, si perde la quota giocata. Supponiamo che Tu decida di giocare tre partite, mutuamente indipendenti, ed ogni volta scommetti 1 euro sul rosso. Sia Y la v.a. che rappresenta l'ammontare totale della vincita, che potrà ovviamente essere sia positivo che negativo.

(a) Quali valori può assumere Y e con quali probabilità?

(b) Calcola il valore medio di Y.

Esercizio 2.109. *Quiz a risposta multipla.*
In un quiz a risposta multipla ci sono 5 domande. Per ogni domanda ci 4 possibili risposte, diciamo (a, b, c, d). Roberta non ha studiato per niente, e decide di rispondere a caso ad ogni domanda. Calcolare la probabilità che

(a) la prima domanda a cui risponde correttamente sia la terza;

(b) risponda correttamente a tre domande su cinque;

(c) risponda correttamente almeno a tre domande.

Esercizio 2.110. *Le combinazioni.*
Il numero di modi in cui si possono ordinare n oggetti distinguibili, ad esempio le 52 carte di un mazzo, è dato dalla semplice formula

$$n! = n \times (n-1) \times \cdots \times 2 \times 1.$$

In questo esercizio considereremo alcuni casi speciali e molto semplici di questa formula.

Una piccola compagnia ha 5 impiegati: Anna, Bice, Carlo, Dario, e Emma. Nel cortile ci sono 5 slot per parcheggiare, uno a fianco dell'altro. Nessuno ha un parcheggio personale, ed ogni mattina i cinque impiegati parcheggiano a caso in uno degli slot disponibili. In pratica tutti i possibili ordinamenti delle 5 auto sono possibili ed ugualmente probabili.

(a) In quanti modi diversi si possono arrangiare le 5 auto?

(b) In un dato giorno, qual è la probabilità che gli impiegati parcheggino in ordine alfabetico?

(c) Se gli impiegati fossero 8, con 8 posti auto, in quanti modi diversi potremmo ordinare le 8 auto.

(d) Se gli impiegati fossero 8, ma i posti auto solo 5, tre auto dovrebbero essere parcheggiate all'esterno del cortile. Quante diverse cinquine di auto si possono osservare nel cortile, senza tener conto dell'ordine in cui si trovano?

(d) Nella stessa situazione precedente, quante diverse cinquine di auto si possono osservare nel cortile, tenendo conto ANCHE dell'ordine in cui si trovano?

Esercizio 2.111. *Figli maschi.*
In genere si crede che, per ogni bimbo che nasce, ci sia una uguale probabilità che sia maschio (M) o femmina (F); in realtà non è così. Quasi in tutto il mondo la probabilità di avere un maschio è $P(M) = 0.51$. Una coppia pianifica di avere 3 figli.

(a) Usa il modello binomiale per calcolare la probabilità che esattamente due dei tre siano maschi.

(b) Elenca esplicitamente tutte le possibili sequenze di M e F potenzialmente osservabili nei tre parti, in cui compaiono due M. Usa questi calcoli per riottenere, utilizzando la regola delle probabilità totali, lo stesso risultato del punto (a).

(c) Se volessimo calcolare la probabilità che una coppia che pianifica di avere 8 figli, abbia tre maschi, descrivi brevemente perché l'approccio usato al punto (b) sarebbe in questo caso molto meno conveniente rispetto al metodo usato nel punto (a).

2.6.4 Altre distribuzioni

Esercizio 2.112. *Trova la distribuzione.*
Si lancia un dado regolare 5 volte. Qual è la probabilità che

(a) il primo 6 esca al quinto lancio?

(b) si ottengano esattamente tre 6?

(c) si abbiano tre sei e che il terzo sei esca al quinto lancio?

Esercizio 2.113. *Il gioco delle freccette.*
Un bravo giocatore di freccette è in grado di colpire il *bull's eye*, (il cerchio rosso al centro del bersaglio) nel 65% dei suoi lanci. Qual è la probabilità che egli

(a) colpisca il cerchio rosso per la decima volta al 15-esimo lancio?

(b) colpisca il cerchio rosso 10 volte nei primi 15 lanci?

(c) colpisca il cerchio rosso per la prima volta al terzo lancio?

(d) Ad ognuna delle tre domande precedenti si può rispondere utilizzando una distribuzione di probabilità nota: sapresti indicare quale nei tre casi?

Esercizio 2.114. *Campionamento a scuola.*
Devi effettuare un'indagine campionaria, che consiste nell'intervistare 20 studenti della tua università. Adotti la seguente strategia: ti sistemi all'ingresso della mensa e intervisti 20 persone a caso che escono dopo il pasto. Sappiamo che gli studenti che frequentano la mensa sono per il 45% maschi e per il 55% femmine.

(a) Quale modello probabilistico ritieni più adatto per calcolare la probabilità che la quarta persona intervistata sia la seconda femmina? Spiegare perché.

(b) Calcola la probabilità del punto (a).

(c) I tre possibili scenari che consentono alla 4^a persona intervistata di essere la 2^a donna sono

$$\{M, M, F, F\}, \{M, F, M, F\}, \{F, M, M, F\}.$$

Una caratteristica comune dei tre scenari è che l'ultima lettera è sempre F. Nei primi tre posti della sequenza ci sono invece, sempre, 2 M e 1 F, sia pure in ordine diverso. Che legame cè tra questa constatazione e il coefficiente binomiale?

(d) Utilizza le considerazioni fatte al punto precedente per spiegare perché nella formula della distribuzione binomiale negativa appaia il simbolo $\binom{n-1}{k-1}$, laddove nella formula della distribuzione binomiale appaia il simbolo $\binom{n}{k}$.

Esercizio 2.115. *Il servizio a pallavolo.*

Un non eccezionale giocatore di pallavolo ha una percentuale di servizi vincenti del 15%. Questo significa che, quando è alla battuta, 15 volte su 100 riesce a ottenere il punto direttamente. Consideriamo le sue battute mutuamente indipendenti.

(a) Qual è la probabilità che, alla decima prova, egli ottenga il 3^o servizio vincente?

(b) Supponiamo che abbia ottenuto due battute vincenti nei primi 9 servizi. Qual è la probabilità che ottenga il suo decimo servizio sia vincente?

(c) Anche se le parti (a) e (b) considerano lo stesso scenario, le probabilità calcolate non dovrebbero essere uguali. Puoi spiegare la ragione di questa differenza?

Esercizio 2.116. *Clienti al bar, Parte I.*

Un bar dell'ateneo serve, in media, 75 clienti l'ora, nell'ora di punta mattutina.

(a) Tra le distribuzioni che conosci, quale ti sembra più appropriata per calcolare la probabilità che un certo numero di clienti arrivi in un'ora in questo periodo di punta?

(b) Quali sono la media e la deviazione standard del numero di clienti serviti in un'ora nel periodo di punta?

(c) Riterresti inusuale osservare solo 60 clienti serviti in un'ora di punta di un certo giorno? Perché?

Esercizio 2.117. *Errori stenografici, Parte I.*
Un'ottima stenografa commette mediamente un errore tipografico ogni ora di lavoro.

(a) Tra le distribuzioni che hai studiato quale ti sembra più adatta per calcolare la probabilità di un dato numero di errori commessi dalla stenografa in un'ora?

(b) Quali sono la media e la deviazione standard del numero di errori commessi dalla stenografa?

(c) Riterresti inusuale rilevare 4 errori della stenografa in una certa ora?

Esercizio 2.118. *Clienti al bar, Parte II.*
Nell'Esercizio 2.116 avevamo stabilito che il numero medio di clienti serviti in un'ora in un bar dell'ateneo era pari a 75. Qual è la probabilità che il bar serva esattamente 70 clienti un una certa ora di punta? Qual è la probabilit' che ne serva ALMENO 70? Nel rispondere a queste domande puoi utilizzare, eventualmente, delle approssimazioni.

Esercizio 2.119. *Errori stenografici, Parte II.*
Nell'Esercizio 2.117 abbiamo visto come il numero medio di errori commessi in un'ora da una certa stenografa era pari a 1.

(a) Calcola la probabilità che la stenografa commetta al più due errori in una certa ora.

(b) Calcola la probabilità che la stenografa commetta almeno 5 errori in una certa ora.

Esercizio 2.120. *Il poker, parte I*
Nel Poker ad ogni giocatore vengono distribuite 5 carte scelte a caso da un mazzo di 52 carte da gioco. Qual è la probabilità di avere "poker d'assi" servito in una mano di 5 carte?

Soluzione

Indichiamo con A l'evento che si verifica se il giocatore ha in mano un poker d'assi servito, cioè tra le sue cinque carte ci sono i quattro assi e una qualsiasi altra carta del mazzo.

Dal momento che possiamo assumere che tutti gli insiemi di 5 carte estratti a caso dal mazzo di 52 siano ugualmente possibili, applichiamo la definizione classica e calcoliamo la probabilità richiesta come:

$$P(A) = \frac{\text{n. casi favorevoli all'evento}}{\text{n. casi possibili}}.$$

Il numero di casi possibili è dato dal numero di modi in cui posso scegliere 5 carte da un mazzo di 52, indipendentemente dall'ordine di estrazione, cioè

$$\left(\begin{array}{c} 52 \\ 5 \end{array} \right) = \frac{52!}{5!\,(52-5)!} = \frac{52 \cdot 51 \cdot 50 \cdot 49 \cdot 48}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}$$

in cui $n!$ indica il prodotto dei primi n numeri naturali.

Infatti, possiamo scegliere la prima carta in 52 modi diversi, per ognuno di questi abbiamo 51 modi di scegliere la seconda carta tra quelle restanti nel mazzo, 50 per scegliere la terza, 49 per la quarta, 48 per la quinta. Quindi in tutto da un mazzo di 52 possiamo scegliere $52 \cdot 51 \cdot 50 \cdot 49 \cdot 48$ mani diverse di 5 carte. Ma queste non sono tutte diverse, perché al al loro interno sono contenute anche mani composte dalle stesse carte estratte in ordine differente. Quindi le possibili mani composte da 5 carte diverse sono meno di $52 \cdot 51 \cdot 50 \cdot 49 \cdot 48$. Per trovare il loro numero basta contare quanti sono i modi di ordinare 5 carte fissate e dividere $52 \cdot 51 \cdot 50 \cdot 49 \cdot 48$ per tale numero. Fissate 5 carte, abbiamo 5 possibilità di scelta per la prima carta, 4 per la seconda (tra le rimanenti), 3 per la terza, 2 per la quarta e 1 solo per la quinta. In tutto il numero di modi di ordinare 5 carte fissate è $5 \cdot 4 \cdot 3 \cdot 2 \cdot 1$. Quindi il numero delle possibili mani di 5 carte diverse è

$$\frac{52 \cdot 51 \cdot 50 \cdot 49 \cdot 48}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}.$$

Per individuare il numero dei casi favorevoli all'evento A dobbiamo contare quante sono le possibili mani di 5 carte di cui 4 sono assi e 1 è una

qualsiasi delle 48 carte restanti. Per i 4 assi abbiamo $\begin{pmatrix} 4 \\ 4 \end{pmatrix} = 1$ possibilità, cioè di fatto c'è un solo modo di sceglierli; mentre per l'ultima carta abbiamo $\begin{pmatrix} 48 \\ 1 \end{pmatrix} = 48$ possibilità. Quindi il numero di casi favorevoli è $1 \cdot 48$.

Allora

$$P(A) = \frac{1 \cdot 48}{\begin{pmatrix} 52 \\ 5 \end{pmatrix}} = 48 \cdot \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{52 \cdot 51 \cdot 50 \cdot 49 \cdot 48} = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{52 \cdot 51 \cdot 50 \cdot 49} = 0.000018$$

Esercizio 2.121. *Il poker texano*
Due giocatori, Alessandro e Federico, stanno giocando a Poker texano. In questo gioco a ciascun giocatore vengono distribuite due carte ed altre 5 carte vengono messe coperte sul tavolo. Ogni carta viene scelta a caso da un mazzo di 52 carte da gioco. Quando le 5 carte sul tavolo vengono scoperte, ciascun giocatore può comporre la sua mano scegliendo 5 carte tra le due che ha ricevuto e le cinque comuni.

(a) Se Federico ha 2 assi, qual è la probabilità che faccia "poker d'assi"?

(b) Se Federico ha 1 asso, qual è la probabilità che faccia "poker d'assi"?

(c) Prima che Federico guardi le sue carte, qual è la probabilità che faccia "poker d'assi"?

Soluzione In linea di principio, Federico fa poker d'assi anche se tra le cinque carte comuni ci sono i quattro assi, ma in questo caso anche Alessandro consegue lo stesso risultato. Non teniamo conto di questa evenienza e ci riferiamo al caso in cui solo Federico fa poker d'assi.

Consideriamo gli eventi:

$F = \{$Federico fa Poker d'Assi$\}$
$F_1 = \{$Federico ha un asso$\}$
$F_2 = \{$Federico ha due assi$\}$
$A = \{$Alessandro non ha in mano nessun Asso$\}$

$B_2 = \{2 \text{ delle } 5 \text{ carte sul tavolo sono Assi}\}$
$B_3 = \{3 \text{ delle } 5 \text{ carte sul tavolo sono Assi}\}$

1. Se Federico ha 2 assi, fa "poker d'assi" se Alessandro non ha assi in mano e tra le 5 carte comuni ci sono i restanti 2 assi. Quindi:

$$P(F|F_2) = P(A \cap B_2|F_2) = P(B_2|A \cap F_2)P(A|F_2)$$

Dal momento che

$$P(A|F_2) = P(A) = \frac{\binom{48}{2}}{\binom{50}{2}} = \frac{48 \cdot 47}{50 \cdot 49} = 0.9208$$

e

$$P(B_2|A \cap F_2) = \frac{\binom{2}{2}\binom{46}{3}}{\binom{48}{5}} = \frac{5 \cdot 4}{48 \cdot 47} = 0.0089$$

si ha

$$P(F|F_2) = P(B_2|A \cap F_2)P(A) = \frac{5 \cdot 4}{48 \cdot 47} \cdot \frac{48 \cdot 47}{50 \cdot 49} = \frac{5 \cdot 4}{50 \cdot 49} = 0.0082$$

2. Se Federico ha 1 asso, fa "poker d'assi" se Alessandro non ha assi in mano e tra le 5 carte comuni ci sono i restanti 3 assi. Quindi:

$$P(F|F_1) = P(A \cap B_3|F_1) = P(B_3|A \cap F_1)P(A|F_1)$$

Ma

$$P(B_3|A \cap F_1) = \frac{\binom{3}{3}\binom{45}{2}}{\binom{48}{5}} = \frac{5 \cdot 4 \cdot 3}{48 \cdot 47 \cdot 46} = 0.00058$$

e $P(A|F_1) = P(A)$. Allora

$$P(F|F_1) = P(B_3|A \cap F_1)P(A) = \frac{5 \cdot 4 \cdot 3}{48 \cdot 47 \cdot 46} \cdot \frac{48 \cdot 47}{50 \cdot 49} = \frac{5 \cdot 4 \cdot 3}{50 \cdot 49 \cdot 46} = 0.00053$$

3. Prima che Federico guardi le sue carte, sappiamo che può fare "poker d'assi" se ha in mano due assi oppure uno e che i due eventi sono incompatibili. Abbiamo detto che escludiamo il caso in cui fa poker d'assi senza avere assi in mano, poiché in questo caso anche Alessandro otterrebbe lo stesso risultato.

   Possiamo scrivere

   $$F = (F \cap F_2) \cup (F \cap F_1)$$

   e

   $$P(F) = P(F \cap F_2) + P(F \cap F_1) = P(F|F_2)P(F_2) + P(F|F_1)P(F_1).$$

   Ma

   $$P(F_2) = \frac{\binom{4}{2}}{\binom{52}{2}} = \frac{6 \cdot 2}{52 \cdot 51} = 0.00453$$

   e

   $$P(F_1) = \frac{\binom{4}{1}\binom{48}{1}}{\binom{52}{2}} = \frac{4 \cdot 48 \cdot 2}{52 \cdot 51} = 0.144796.$$

   Utilizzando le risposte alle domande precedenti:

   $$\begin{aligned} P(F) &= P(F|F_2)P(F_2) + P(F|F_1)P(F_1) = \\ &= \frac{5 \cdot 4}{50 \cdot 49} \cdot \frac{6 \cdot 2}{52 \cdot 51} + \frac{5 \cdot 4 \cdot 3}{50 \cdot 49 \cdot 46} \cdot \frac{48 \cdot 4 \cdot 2}{52 \cdot 51} = 0.000114 \end{aligned}$$

Capitolo 3

Inferenza

3.1 Intervalli di confidenza e test per campioni estratti da una popolazione Normale

Esercizio 3.1.

Per un certo prodotto, il prezzo di vendita al dettaglio si distribuisce secondo una Normale, con varianza pari a 144. Al fine di costruire una stima intervallare al livello $1 - \alpha = 0.90$ per il prezzo medio nella popolazione di riferimento,

1. determinare gli estremi dell'intervallo di confidenza, sulla base di un campione casuale di 36 unità con media pari a 15;

2. determinare gli estremi dell'intervallo di confidenza a livello $1 - \alpha = 0.98$;

3. determinare la numerosità campionaria necessaria affinché l'ampiezza dell'intervallo al livello $1 - \alpha = 0.90$ sia al massimo pari a 4.

• • •

Soluzione

Indichiamo con X la variabile "prezzo di vendita al dettaglio". Il testo ci dice che $X \sim N(\mu, \sigma^2 = 144)$.

1. L'intervallo di confidenza per la variabile X si ottiene come:

$$\left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] \tag{3.1}$$

   - $\bar{x}$ è il prezzo medio del campione osservato: $\bar{x} = 15$;
   - σ è la deviazione standard della popolazione di riferimento: $\sigma = \sqrt{144} = 12$;
   - n è la numerosità del campione osservato: $n = 36$
   - $z_{\alpha/2}$ è il quantile a livello $\alpha/2 = 1 - 0.90/2 = 0.05$ di una distribuzione Normale standardizzata: $z_{\alpha/2} = 1.64$

   Sostituendo nell'equazione 3.1 si ottiene il seguente intervallo di confidenza:

$$\left[15 - 1.64\frac{12}{\sqrt{36}}, 15 + 1.64\frac{12}{\sqrt{36}}\right] = [11.72, 18.28]$$

2. Applichiamo la stessa formula del punto precedente modificando solo il livello di confidenza e quindi $z_{\alpha/2} = z_{0.01} = 2.33$:

$$\left[15 - 2.33\frac{12}{\sqrt{36}}, 15 + 2.33\frac{12}{\sqrt{36}}\right] = [10.34, 19.66]$$

3. Poiché la varianza è nota, la numerosità minima per un'ampiezza $a = 4$, il margine di errore è $m = a/2 = 2$, si ottiene mediante la seguente formula

$$n = \left(\frac{z_{\alpha/2}\sigma}{a/2}\right)^2. \tag{3.2}$$

   In questo caso, $n = \left(\frac{1.64 \cdot \sqrt{144}}{2}\right)^2 = 96.83 \cong 97$, arrotondando per eccesso.

• • •

Esercizio 3.2.

Da una sorgente di acque minerali è stato prelevato un campione casuale di 81 provette di acqua. Il contenuto medio di sali minerali disciolti in acqua è risultato pari a 600 mg/l. Supponendo che il contenuto di sali minerali sia distribuito come una variabile casuale Normale con deviazione standard uguale a 50 mg/l, verificare, al livello di significatività α=0.001, l'ipotesi che l'acqua della sorgente contenga mediamente 500 mg/l di sali minerali, contro l'alternativa che ne contenga piú di 500.

• • •

Soluzione

Indichiamo con X la variabile "contenuto di sali minerali". Il testo ci dice che $X \sim N(\mu, \sigma = 50)$.
Dobbiamo verificare il seguente si sistema di ipotesi:

$$H_0 : \mu = 500 \qquad\qquad H_1 : \mu > 500$$

Per effettuare la verifica di ipotesi sulla media di una popolazione Normale con varianza nota, la statistica test è

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

che sotto l'ipotesi nulla ha distribuzione Normale standardizzata.
Calcoliamo quindi il valore p come $p = P(Z > z)$ dove z il valore della statistica test nel campione osservato. Pertanto si ha:

$$\begin{aligned} p = P(Z > z) &= P\left(Z > \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}\right) = \\ &= P\left(Z > \frac{600 - 500}{50/\sqrt{81}}\right) = P(Z > 18) = 0 \end{aligned}$$

Poiché $p < \alpha = 0.001$, si ha abbastanza evidenza sperimentale per rifiutare l'ipotesi nulla.

• • •

Esercizio 3.3.

Dalla popolazione dei docenti universitari è stato estratto un campione casuale di 100 docenti di sesso femminile rilevandone l'età.

1. Determinare l'intervallo di confidenza a livello 95% per l'età media, sapendo che l'età media del campione delle 100 donne osservate è pari a 42.2 e che nella popolazione dei docenti di sesso femminile la variabile età presenta distribuzione Normale con varianza pari a 49;

2. Si vuole verificare l'ipotesi che l'età media sia pari a 44 anni contro l'ipotesi alternativa bilaterale. Cosa possiamo concludere a livello di significatività 0.05? E se il livello di significatività fosse 0.1?

3. Supponendo che per il complesso dei docenti la variabile età si distribuisca secondo una Normale con varianza pari a 100, determinare il numero minimo di docenti per i quali il margine di errore dell'intervallo di confidenza a livello 95% per la media sia pari al 10%.

• • •

Soluzione

1. Definiamo X la variabile "età dei docenti universitari di sesso femminile". Il testo ci dice che $X \sim N(\mu, \sigma^2 = 49)$. Possiamo definire l'intervallo di confidenza per X come segue

$$\left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] \tag{3.3}$$

dove

- $\bar{x}$ è l'età media delle donne del campione osservato: $\bar{x} = 42.2$;
- σ è la deviazione standard della popolazione di riferimento: $\sigma = \sqrt{49} = 7$;
- n è la numerosità del campione osservato: $n = 100$
- $z_{\alpha/2}$ è il quantile a livello $\alpha/2 = 1 - 0.95/2 = 0.025$ di una distribuzione Normale standardizzata: $z_{\alpha/2} = 1.96$

Sostituendo si ottiene il seguente intervallo di confidenza:

$$\left[42.2 - 1.96\frac{7}{\sqrt{100}}, 42.2 + 1.96\frac{7}{\sqrt{100}}\right] = [40.828, 43.572]$$

2. Per concludere il test è sufficiente osservare che il livello di significatività α corrisponde al livello di confidenza $1 - \alpha$ dell'intervallo che abbiamo determinato al punto precedente. Poiché il valore 44 non è contenuto nell'intervallo osservato $[40.828, 43.572]$, possiamo concludere che c'è abbastanza evidenza sperimentale per rifiutare l'ipotesi nulla.
   Se il livello di significatività fosse 0.1, avremmo corrispondentemente un intervallo di confidenza a livello 90% che risulterebbe più stretto di quello precedente e quindi a maggior ragione non conterrebbe il valore 44, inducendoci a rifiutare l'ipotesi nulla.

3. Definiamo Y la variabile "et dei docenti universitari di sesso maschile". Il testo ci dice che $Y \sim N(\mu, \sigma^2)$. A differenza di quanto accade nel quesito precedente, la varianza di tale distribuzione incognita. Pertanto, l'intervallo di confidenza per la media definito come segue:

$$\left[\bar{y} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}, \bar{y} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}\right] \tag{3.4}$$

   dove

   - $\bar{y}$ il valore della media campionaria;
   - s il valore della deviazione standard campionaria corretta;

- n la numerosit del campione osservato: n=80;
- $t_{n-1,\alpha/2}$ il quantile a livello $\alpha/2$ della distribuzione T di Student con $n-1$ gradi di libertà.
  Dalle tavole della distribuzione T di Student si ha $t_{0.025,79} = 1.990$ (avendo approssimato n=80).

A partire quindi dalla tabella di frequenze, definiamo le quantit necessarie per il calcolo dell'intervallo di confidenza:

- $\bar{y} = \frac{1}{n}\sum_{j=1}^{k} \bar{x}_j n_j = \frac{1}{80}(30 \cdot 15 + 40 \cdot 10 + ... + 60 \cdot 28) = 48.5$
- $s^2 = \frac{1}{n-1}\sum_{j=1}^{k} \bar{x}_j^2 n_j - \frac{n}{n-1}\bar{x}^2 =$
  $\frac{1}{79}(30^2 \cdot 15 + 40^2 \cdot 10 + ... + 60^2 \cdot 28) - \frac{80}{79}(48.5^2) =$
  $2503.797 - 2382.025 = 121.772$
  da cui si ottiene $s = \sqrt{121.772} = 11.035$

Inserendo queste quantit, possiamo quindi calcolare l'intervallo di confidenza richiesto:

$$[48.5 - 1.990\frac{11.035}{\sqrt{80}}, 48.5 + 1.990\frac{11.035}{\sqrt{80}}] = [46.045, 50.955]$$

4. Poiché la varianza è nota, la numerosità minima per un margine di errore $m = 0.1$ si ottiene mediante la seguente formula

$$n = \left(\frac{z_{\alpha/2}\sigma}{m}\right)^2$$

dove $z_{\alpha/2}$ il quantile a livello $\alpha/2$ di una distribuzione Normale standard. In questo caso, $n = \left(\frac{1.96 \cdot 10}{0.1}\right)^2 = 38416$.

• • •

Esercizio 3.4.

Supponiamo che in questo momento 10 persone siano collegate ad un sito per l'acquisto di articoli su internet. Sapendo che la probabilità che ciascuno dei 10 soggetti acquisti effettivamente un articolo è pari a 0.2, calcolate:

1. la probabilità che nessun soggetto acquisti un articolo;

2. la probabilità che 2 soggetti acquistino un articolo;

3. la probabilità che al massimo 2 soggetti acquistino un articolo;

4. il numero medio di articoli acquistati;

Un esperto in comunicazioni ritiene che più della metà della popolazione effettua acquisti su internet. Sapendo che dei 10 soggetti intervistati, 4 hanno effettuato un acquisto su internet, cosa si può concludere sull'affermazione dell'esperto (utilizzare un livello di significatività del 95%)?

Soluzione

Definiamo X la variabile X="numero di acquisti su internet". Questa variabile ha distribuzione binomiale con parametri $n = 10$ e $p = 0.2$, ossia

$$X \sim Binomiale(n = 10, p = 0.2)$$

Ricordando che

$$Pr(X = k) = \binom{n}{x} p^k (1-p)^{n-k}$$

possiamo calcolare le probabilità richieste come segue:

1. $Pr(X = 0) = \binom{10}{0} 0.2^0 (1 - 0.2)^{10-0} = 0.1074$

2. $Pr(X = 2) = \binom{10}{2} 0.2^2 (1 - 0.2)^{10-2} = 0.3020$

3. $Pr(X \leq 1) = Pr(X = 0) + Pr(X = 1) + Pr(X = 0) = 0.1074 + \binom{10}{1} 0.2^1 (1 - 0.2)^{10-1} + 0.3020 = 0.6778$

4. Poiché X ha una distribuzione binomiale, allora il numero medio di articoli acquistati è $E[X] = np = 10 \cdot 0.2 = 2$

Per validare o smentire l'affermazione dell'esperto, dobbiamo valutare il seguente sistema di ipotesi:

$$H_0 : p = 0.5 \quad H_1 : p > 0.5$$

Sotto l'ipotesi nulla, sappiamo

$$T = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

dove $\hat{p} = \frac{40}{100} = 0.4$; sappiamo inoltre che sotto H_0 T ha distribuzione T di Student con $n-1$ gradi di libertà. Possiamo quindi calcolare il p-value come segue:

$$\begin{aligned} Pr(T > t) &= Pr\left(T > \frac{0.4 - 0.5}{\sqrt{\frac{0.5(1-0.5)}{10}}}\right) = \\ &= Pr(T > -0.6324) = 1 - Pr(T \leq -0.6324) = 0.7357 \end{aligned}$$

Poiché $0.7357 >> 0.05$, allora non ho abbstanza evidenza sperimentale per rifiutare l'ipotesi nulla.

3.2 Intervalli di confidenza e test per campioni estratti da popolazioni Normali con media e varianza incognite

Esercizio 3.5.

Il numero medio di ore di sonno per notte ha una distribuzione normale. In un campione di 20 individui sottoposto ad un trattamento farmacologico ipotensivo, il numero medio di ore di sonno risulta pari a 6.5 con uno scarto quadratico medio di 2 ore. Sulla base dei dati disponibili:

1. Si costruisca un intervallo di confidenza al 95% per il numero medio di ore di sonno.

2. Si consideri l'ipotesi nulla $H_0 : \mu = 7$ di un test bidirezionale al livello di significatività del 5%. Sulla base del risultato del punto precedente l'ipotesi nulla può essere respinta?

• • •

Soluzione

Sia X la variabile aleatoria "ore di sonno per notte". Il testo dice che $X \sim N(\mu, \sigma^2)$, con μ e σ^2 entrambi incogniti.

1. L'intervallo di confidenza a livello $1 - \alpha = 0.95$ per la media μ della variabile X è definito come

$$\left[\bar{x} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}\right] \tag{3.5}$$

dove

- $\bar{x}$ è il valore della media campionaria: $\bar{x} = 6.5$;
- s è il valore della deviazione standard campionaria corretta: $s = 2$;
- n è la numerosità del campione osservato: $n = 20$;
- $t_{n-1,\alpha/2}$ è il quantile a livello $\alpha/2$ della distribuzione T di Student con $n - 1$ gradi di libertà.
  Dalle tavole della distribuzione T di Student si ha $t_{0.025,19} = 2.093$.

Pertanto, sostituendo queste quantità nella formula 3.5, si ha che l'intervallo di confidenza al 95% per la media μ è

$$\left[6.5 - 2.093\frac{2}{\sqrt{20}}, 6.5 + 2.093\frac{2}{\sqrt{20}}\right] = [5.564, 7.436]$$

2. Circa il test di ipotesi, è possibile fornire una risposta al quesito senza fare nessun calcolo. Infatti sfruttando le informazioni fornite dal precedente punto, si può osservare che il valore del numero medio di ore di sonno ipotizzato sotto H_0 appartiene all'intervallo di confidenza appena individuato. Tale informazione è sufficiente per decidere che l'ipotesi nulla non può essere respinta.

• • •

Esercizio 3.6.

Un professore è interessato a conoscere la spesa media annuale in libri di testo degli studenti universitari. La spesa ha una distribuzione normale. In un campione di 26 studenti, la spesa media è risultata 180 euro con uno scarto quadratico medio di 30 euro.

1. Costruire un intervallo di confidenza al 95% per la spesa media;

2. Come varia l'intervallo di confidenza quando aumenta la numerosità campionaria?

3. Un collega sostiene che la spesa media è 185 euro. Sulla base dei risultati del punto 1. è possibile sostenere questa affermazione al livello di significatività del 5%?

• • •

Soluzione

Indichiamo con X =Spesa dei libri di testo.
Sappiamo che $X \sim N(\mu, \sigma^2)$ con i parametri entrambi incogniti. Abbiamo anche: $n = 26$, $\bar{x} = 180$ e $s = 30$.

1. Considerando che $1-\alpha = 0.95; \alpha = 0.05; \alpha/2 = 0.025$, l'intervallo di confidenza sarà:
$$\left[\bar{x} - t_{n-1;\frac{\alpha}{2}}\frac{s}{\sqrt{n}}; \bar{x} + t_{n-1;\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right]$$
quindi
$$\left[180 - t_{25;0.025}\frac{30}{\sqrt{26}}; 180 + t_{25;0.025}\frac{30}{\sqrt{26}}\right]$$
e
$$\left[180 - 2.06\frac{30}{\sqrt{26}}; 180 + 2.06\frac{30}{\sqrt{26}}\right]$$
dove 2.06 è il quantile a livello 0.025 (dalla tavola C) di una distribuzione t con 25 gradi di libertà. Otteniamo
$$[167.88; 192.12]$$

2. All'aumentare di n l'intervallo di confidenza si restringe. Per $n \to \infty$ collassa sulla media $\bar{x} = 180$.

3. È possibile sostenerla in quanto 185 cade all'interno dell'intervallo appena calcolato. Questa considerazione può essere fatta in quanto intervallo di confidenza e test di ipotesi sono definiti allo stesso livello di significatività del 5%.

• • •

Esercizio 3.7.

In una clinica un gruppo di medici che si occupa della ricerca su un nuovo farmaco per il colesterolo ritiene che una variazione media del colesterolo pari a 1.2 dopo la somministrazione di tale farmaco sia sufficiente per poter mettere il farmaco sul mercato. Si effettua un test di significatività al 5% per la verifica di

$$H_0 : \mu = \mu_0 = 0 \qquad H_a : \mu = \mu_a = 1.2$$

basato su un campione di 41 volontari, a cui è stato somministrato il farmaco per 60 giorni, con deviazione standard pari a 2. Qual è la potenza del test?

• • •

Soluzione

Riassumiamo i dati: $n = 41$, X =Variazione di colesterolo, $s = 2$. La statistica test sarà:

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$$

La potenza del test rispetto all'alternativa puntuale $\mu = \mu_a = 1.2$ è definita come la probabilità con cui il test rifiuta H_0 quando $\mu = 1.2$ è vera.

Passo 1: Scriviamo la formula per rifiutare H_0 in termini di $\bar{x}$. Il test rifiuta H_0 ad un livello $\alpha = 0.05$ quando

$$t = \frac{\bar{x} - 0}{2/\sqrt{41}} \geq 1.684$$

ovvero quando

$$\bar{x} \geq 0 + 1.684\frac{2}{\sqrt{41}}$$

per cui si rifiuta H_0 quando $\bar{x} \geq 0.526$. In questo modo abbiamo riformulato il test in termini di $\bar{x}$. Osserviamo che la regola che ci dice quando rifiutare H_0 non dipende dal valore specifico dell'alternativa.

Passo 2: La *potenza* è la probabilità che si verifichi l'evento $\bar{X} \geq 0.526$ quando l'alternativa $\mu = 1.2$ è vera. Per calcolare questa probabilità, occorre standardizzare $\bar{x}$ utilizzando $\mu = 1.2$:

$$\begin{aligned} potenza &= P(\bar{X} \geq 0.526 \quad quando \quad \mu = 1.2) \\ &= P\left(\frac{\bar{X} - 1.2}{2/\sqrt{41}} \geq \frac{0.526 - 1.2}{2/\sqrt{41}}\right) \\ &= P\left(T \geq -2.1579\right) \approx 0.98 \end{aligned}$$

Il test dichiarerà che i pazienti presentano una variazione significativa del colesterolo dopo la somministrazione del farmaco soltanto il 5% delle volte,

quando tale variazione non si verifica (quando H_0 è vera) e circa il 98% delle volte quando la variazione effettiva è pari a $\mu = 1.2$ (quando H_a è vera).

• • •

Esercizio 3.8.

I dati storici indicano che l'acidità media della pioggia in una certa zona del West Virginia è 5.2. Per vedere se recentemente ci sono state delle variazioni, viene misurata l'acidità dell'acqua durante 12 rovesci nell'ultimo anno, con media e deviazione standard pari rispettivamente a 5.667 e 0.921. Ritieni che, con un livello di significatività del 5%, si possa concludere che l'acidità della pioggia sia cambiata rispetto al valore storico?

• • •

Soluzione

Riassumiamo i dati: x =acidità della pioggia, $\bar{x} = 5.667$, $s = 0.921$. Il sistema di ipotesi sarà:

$$H_0 : \mu = 5.2; \qquad H_a : \mu \neq 5.2$$

I parametri sono incogniti quindi dovremo utilizzare un test di tipo t. Calcoliamo il p-value:

$$p = P(T > t) + P(T < -t)$$

dove

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$$

è la statistica test che si distribuisce come una $t_{n-1} = t_{11}$ e t è il valore della statistica test calcolato sul campione in esame, cioè $t = (5.667 - 5.2)/(0.921/\sqrt{12}) = 1.76$. Otteniamo:

$$p = P(T > 1.76) + P(T < -1.76) = 2P(T > 1.76)$$

Dalle tavole della distribuzione T, si ha che

$$0.05 < P(T > 1.76) < 0.1$$

da cui

$$0.1 < p = 2P(T > 1.76) < 0.2$$

Il p-value è maggiore di 0.05 (il livello di significatività), quindi non c'è abbastanza evidenza sperimentale per poter rifiutare l'ipotesi nulla.

• • •

3.3 Test t per campioni appaiati

Esercizio 3.9.

In uno studio è stato chiesto a 25 persone "destre" di girare completamente due manopole (con la loro mano destra). La prima (progettata per destri) andava girata in senso orario. La seconda (progettata per mancini) andava girata in senso antiorario. Si vuole mostrare che persone destre hanno piú facilità ad usare oggetti per destri. La seguente tabella riporta i tempi medi per girare completamente una manopola.

Soggetto	Manopola DX	Manopola SX
1	113	137
2	105	105
3	130	133
4	101	108
5	138	115
6	118	170
7	87	103
8	116	145
9	75	78
10	96	107
11	122	84
12	103	148
13	116	147
14	107	87
15	118	166
16	103	146
17	111	123
18	104	135
19	111	112
20	89	93
21	78	76
22	100	116
23	89	78
24	85	101
25	88	123

• • •

Soluzione

Va subito notato come a ciascun soggetto siano state fatte girare le manopole in un ordine casuale per evitare una sorta di "apprendimento".

Il parametro che si vuole sottoporre a verifica è la media μ delle differenze tra il tempo impiegato a girare la manopola per destri e quella per mancini. Vogliamo quindi effettuare il seguente test di ipotesi:

$$H_0 : \mu = 0 \qquad H_a : \mu < 0$$

La prima cosa da fare è calcolare le differenze tra i tempi impiegati da ciascun soggetto (cioè riga per riga nella tabella precedente):

Soggetto	Manopola DX	Manopola SX	Differenze
1	113	137	-24
2	105	105	0
3	130	133	-3
4	101	108	-7
5	138	115	23
6	118	170	-52
7	87	103	-16
8	116	145	-29
9	75	78	-3
10	96	107	-11
11	122	84	38
12	103	148	-45
13	116	147	-31
14	107	87	20
15	118	166	-48
16	103	146	-43
17	111	123	-12
18	104	135	-31
19	111	112	-1
20	89	93	-4
21	78	76	2
22	100	116	-16
23	89	78	11
24	85	101	-16
25	88	123	-35

Dopo aver calcolato le differenze possiamo determinare la media campionaria $\bar{x} = -13.32$ e la deviazione standard campionaria $s = 22.94$

La statistica t ha $n - 1 = 24$ gradi di libertà ed il valore osservato è

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{-13.32 - 0}{22.94/\sqrt{25}} = -2.90$$

Dalla riga corrispondente a 24 gradi di libertà delle tavole ricaviamo che il valore -2.90 della statistica è compreso tra i valori critici -3.091 e -2.797, corrispondenti ai livelli 0.0025 e 0.005. Quindi il valore P è $0.0025 < P < 0.005$

Concludiamo che il test risulta significativo ad un livello inferiore al 5 per mille, ovvero i dati forniscono forte evidenza contro l'ipotesi nulla che il tempo necessario a compiere le due operazioni sia mediamente lo stesso.

Attenzione: non è corretto analizzare dati appaiati come se avessimo due campioni, uno riferito alla prima "circostanza" ed uno alla seconda. Le procedure inferenziali per confrontare due campioni presuppongono che i campioni selezionati siano indipendenti. Questa assunzione non è vera quando gli stessi soggetti sono misurati due volte o le misurazioni sono effettuate su soggetti diversi, ma scelti a coppie. La procedura di analisi da applicare dipende sempre dal modo in cui sono stati ricavati i dati.

• • •

Esercizio 3.10. *Emicrania e agopuntura.*

L'emicrania è un tipo di mal di testa particolarmente doloroso. A volte i pazienti provano a curarsi con l'agopuntura. Per stabilire se l'agopuntura allevia il mal di testa, alcuni ricercatori hanno effettuato iuno studio controllato randomizzato in cui 89 donne con emicrania sono state assegnate casualmente ad uno dei due gruppi: trattamento (agopuntura) o controllo (ovvero nulla o cura tradizionale). I 43 pazienti nel gruppo di trattamento hanno ricevuto l'agopuntura specifica per l'emicrania. I 46 pazienti nel gruppo di controllo hanno invece ricevuto un'agopuntura placebo, ovvero punture in punti non sensibili. Dopo 24 ore, ai pazienti è stato chiesto se avevano ancora dolore oppure no. I risultati sono riassunti nella tabella di contingenza di seguito.

	Miglioramento		
	SI	NO	Totale
Controllo	10	33	43
Trattamento	2	44	46
Totale	12	77	89

- Quale percentuale di pazienti a cui è stato somministrato il trattamento ha avuto un miglioramento? Quale percentuale nel gruppo di controllo?
- A colpo d'occhio, quale trattamento appare migliore per la cura dell'emicrania?
- I dati forniscono evidenza statistica convincente che i due trattamenti differiscono oppure pensi che le differenze possano essere dovute soltanto al caso?

Esercizio 3.11. *Gocce di cioccolato*
Ad un gruppo di studenti viene chiesto di contare il numero di gocce di cioccolato contenute in 22 biscotti. Gli studenti hanno trovato che i biscotti contengono in media 14.77 gocce di cioccolato con una deviazione standard di 4.37 gocce di cioccolato.

(a) Utilizzando queste informazioni, quanta variabilità si dovrebbero attendere di vedere nel numero medio di gocce di cioccolato in un campione casuale di 22 biscotti?

(b) Sulla confezione è dichiarato che ogni biscotto contiene almeno 20 gocce di cioccolato. Uno studente trova che questo numero sia irragionevolmente alto in quanto il numero medio di gocce di cioccolato che hanno contato è molto più basso. Un altro studente sostiene che la differenza puó essere solo effetto del caso. Cosa ne pensi?

Esercizio 3.12. *Una indagine statistica*
La General Social Survey (GSS) è una indagine sociologica utilizzata negli

Stati Uniti per collezionare dati circa le caratteristiche demografiche e le attitudini dei residenti. Nel 2012, i residenti intervistati sono stati 1154. Gli intervistati vengono estratti casualmente da un campione di adulti e sono intervistati personalmente. Una delle domande dell'indagine è: "Dopo un normale giorno di lavoro, quante ore circa hai a disposizione per rilassarti o dedicarti ai tuoi hobby?". Dalla GSS del 2010 è risultato un intervallo di confidenza al 95% pari a [3.53; 3.83].

(a) Come si puó interpretare questo intervallo?

(b) Cosa rappresenta un intervallo di confidenza al 95% in questo specifico contesto?

(c) Supponiamo che alcuni ricercatori sostengono che un intervallo al 90% sia più appropriato per questo tipo di dati. Assumendo che la deviazione standard rimanga costante dal 2010, questo intervallo sarà più ampio o meno ampio dell'intervallo al 95%?

Esercizio 3.13. *Salute mentale*
Un'altra domanda dell'indagine GSS (vedi esercizio 3.12) è la seguente: " Definendo "salute mentale" lo stato di stress, depressione, problemi personali, per quanti giorni nel mese precedente (30 giorni) la tua salute mentale non è stata in buone condizioni?" Utilizzando le risposte di 1151 residenti, si è ottenuto il seguente intervallo di confidenza $[3.40; 4.24]$ (livello 95%).

(a) Interpreta questo intervallo.

(b) Cosa rappresenta un intervallo di confidenza al 95% in questo specifico contesto?

(c) Supponiamo che alcuni ricercatori sostengono che un intervallo al 99% sia più appropriato per questo tipo di dati. Assumendo che la deviazione standard rimanga costante dal 2010, questo intervallo sarà più ampio o meno ampio dell'intervallo al 95%?

(d) Se si conducesse una nuova intervista e se la stessa domanda fosse sottoposta a 500 residenti, l'errore standard della stima sarebbe più alto, più basso o rimarrebbe uguale? Assumiamo anche in questo caso che la deviazione standard rimanga costante dal 2010.

Esercizio 3.14. *Intervalli di confidenza*

Ampiezza di un intervallo di confidenza. Con riferimento al capitolo 4, calcolammo l'intervallo di confidenza a livello 99% per il numero medio di corridori della corsa Cherry Blossom: utilizzando un campione di 100 corridori, l'intervallo di confidenza risulta pari a $[32.7; 37.4]$. Come possiamo diminuire l'ampiezza di questo intervallo senza diminuire il livello di confidenza?

Esercizio 3.15. *Livelli di confidenza*

Se un livello di confidenza pi'u elevato significa che noi siamo più fiduciosi circa i numeri che stiamo riportando, perchè non utilizziamo sempre intervalli di confidenza con il più elevato livello di confidenza?

Esercizio 3.16. *Pronto soccorso*

Il dirigente di un ospedale al fine di migliorare il tempo di attesa, decide di stimare il tempo medio di attesa al pronto soccorso del suo ospedale. Il dirigente collezione un campione semplice casuale di 64 pazienti e calcola il tempo (in minuti) trascorso dall'ingresso al pronto soccorso alla prima visita con un dottore. Un intervallo di confidenza a livello 95% è pari a $[126, 146]$ minuti. Tale intervallo è stato costruito assumendo un modello Normale per la media.

Stabilire se le seguenti affermazioni sono vere o false e giustificare la propria risposta:

(a) Questo intervallo di confidenza non è valido in quanto non sappiamo se la distribuzione del tempo di attesa al pronto soccorso sia veramente Normale;

(b) Siamo confidenti al 95% che il tempo medio di attesa al pronto soccorso di questi 64 pazienti sia tra 128 e 147 minuti;

(c) Siamo confidenti al 95% che il tempo medio di attesa al pronto soccorso di tutti i pazienti dell'ospedale sia tra 128 e 147 minuti;

(d) Supponendo di poter estrarre altri campioni casuali, il 95% di questi campioni casuali potrebbe avere la media campionaria tra 128 e 147 minuti.

(e) Poiché vogliamo essere più sicuri delle nostre stime, è meglio utilizzare un intervallo di confidenza a livello 99% che è più stretto rispetto all'intervallo al 95%

(f) Il margine di errore è 9.5 e la media campionaria è 137.5

(g) Al fine di ridurre il margine di errore dell'intervallo di confidenza al 95% della metà, dobbiamo aumentare la numerosità campionaria.

Esercizio 3.17.

Per determinare l'età media dei suoi acquirenti, un negozio di abbigliamento intervista un campione di 50 acquirenti e determina che $\bar{X} = 36$. Sapendo che l'età degli acquirenti si distribuisce normalmente e che $\sigma = 12$:

1. si determini l'intervallo di confidenza al 95% per l'età media μ di tutti gli acquirenti;

2. si supponga di voler ridurre l'ampiezza dell'intervallo di confidenza al 95%, in modo tale che gli estremi distino dal valore centrale dell'intervallo $\pm$ 2 anni. Quanto deve essere grande il campione?

Soluzione

Si tratta di determinare l'intervallo di confidenza per un campione estratto da una popolazione Normale con media μ incognita e deviazione standard nota ($\sigma = 12$).

1. L'intervallo di confidenza è definito come

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

dove

- $\bar{x} = 36$
- $z^* = 1.96$

Pertanto, l'intervallo di confidenza è pari a:

$$\left[36 - 1.96\frac{12}{\sqrt{50}}; 36 + 1.96\frac{12}{\sqrt{50}}\right] = [32.673; 39.326]$$

2. L'esercizio richiede la numerosità campionaria tale che l'ampiezza dell'intervallo $A = 4$, o equivalentemente, il margine di errore $m = 2$. Tale numerosità campionaria si ottiene come

$$n^* = \left(\frac{z^*\sigma}{m}\right)^2 = \left(\frac{1.96 \cdot 12}{36}\right)^2 = 138.2976$$

Quindi si può concludere che la numerosità minima per avere un'ampiezza dell'intervallo pari a 4 è 139.

Esercizio 3.18.

Una agenzia immobiliare vuole stimare il prezzo medio di vendita degli appartamenti di una zona di Roma. Considera un campione di 25 vendite e calcola il prezzo medio $\bar{X} = 148000$ Euro, con deviazione standard campionaria $s = 62000$ Euro. Si calcoli l'intervallo di confidenza al 95% per il prezzo medio delle vendite.

Soluzione

Si tratta di determinare l'intervallo di confidenza per un campione estratto da una popolazione Normale con media μ incognita e deviazione standard incognita. L'intervallo di confidenza è definito come

$$\bar{x} \pm t^*\frac{s}{\sqrt{n}}$$

dove

- $\bar{x} = 148000$

- $s = 62000$
- $t^* = 2.064$ (quantile a livello 0.025 di una distribuzione T con $n-1 = 24$ gradi di libertà).

Quindi l'intervallo di confidenza richiesto è

$$\left[148000 - 2.064\frac{62000}{\sqrt{25}}; 148000 + 2.064\frac{62000}{\sqrt{25}}\right] = [122406.4; 173593.5]$$

Esercizio 3.19.
Il direttore del personale di una grande società intende stimare le assenze del personale dipendente dell'ufficio centrale della società nel corso di 1 anno. Si estrae un campione casuale di 25 dipendenti e si osservano i seguenti risultati:

- $\bar{X} = 9.7$ giorni, $S = 4$ giorni;
- 12 dipendenti sono stati assenti più di 10 giorni.

1. Costruire un intervallo di confidenza al 95% per il numero medio di giorni di assenza dei dipendenti nello scorso anno;
2. Costruire un intervallo di confidenza al 95% per stimare la proporzione di dipendenti che lo scorso anno sono stati assenti più di 10 giorni.

Soluzione
La variabile X="giorni di assenza" ha distribuzione Normale con media μ e varianza σ^2 entrambe incognite. Dovendo fare inferenza sulla media μ ed essendo la varianza incognita, la statistica test da utilizzare è

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim T_{n-1}$$

che ha distribuzione T con $n - 1$ gradi di libertà.

1. L'intervallo di confidenza per μ è definito come

$$\left[\bar{X} - t^*\frac{S}{\sqrt{n}}, \bar{X} + t^*\frac{S}{\sqrt{n}}\right]$$

dove $t^* = 2.064$. Sostituendo i valori, si ottiene il seguente intervallo di confidenza:

$$[9.7 - 2.064\frac{4}{\sqrt{25}}, 9.7 + 2.064\frac{4}{\sqrt{25}}] = [8.0488, 11.3512]$$

2. Circa la proporzione, l'intervallo di confidenza è definito come

$$\left[\hat{p} - z^*\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}; \hat{p} + z^*\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] = [0.284; 0.676]$$

dove $\hat{p} = \frac{12}{25} = 0.48$ e $z^* = 1.96$

Capitolo 4

Dati categorici

4.1 Tabelle di contingenza

Esercizio 4.1. *Depressione e stato civile, I parte*

Consideriamo i dati di un medico su 159 pazienti depressi. La seguente tabella riporta il livello depressivo osservato rispetto allo stato civile:

	stato civile			
livello depressivo	sposato	celibe	vedovo	Totale
grave	22	16	19	57
medio	33	29	14	76
leggero	14	9	3	26
Totale	69	54	36	159

a. Di che tipo di variabili si tratta?

b. Determinare la distribuzione marginale di frequenza dello stato civile.

c. Quale é la moda per il carattere livello depressivo?

d. Quale è la percentuale di pazienti che risultano vedovi e con un livello depressivo grave?

e. Quale è la percentuale di vedovi con livello depressivo grave?

f. Quale è la percentuale di pazienti con livello depressivo almeno pari a un livello 'medio'?

g. Determinare la distribuzione condizionata, di frequenze assolute e di frequenze percentuali, del livello depressivo allo stato civile vedovo.

h. Determinare la distribuzione marginale del livello depressivo e confrontarla con la distribuzione condizionata ricavata al punto precedente. Cosa si può dire sull'associazione tra i due caratteri?

• • •

Soluzione.

a. Il livello depressivo è un carattere qualitativo ordinato, lo stato civile è un carattere qualitativo sconnesso.

b. La distribuzione marginale dello stato civile è la seguente:

sposato	celibe	vedovo	Totale
69	54	36	159

c. La moda è il livello depressivo 'medio' ovvero la modalità del carattere alla quale è associata la massima frequenza.

d. La percentuale di pazienti che sono simultaneamente vedovi e con livello depressivo grave è data da:

$$\frac{19}{159}100 = 11.9\%$$

e. La percentuale di vedovi che presentano un livello depressivo grave è data da

$$\frac{19}{36}100 = 52.8\%$$

f. Il numero di pazienti con un livello depressivo pari almeno ad un livello medio è dato dalla somma tra il numero di pazienti con livello depressivo medio e quello con livello depressivo grave, $76 + 57 = 133$. La percentuale richiesta è quindi:

$$\frac{133}{159}100 = 84\%$$

g. la distribuzione condizionata è riportata nella seguente tabella:

stato civile=vedovo

livello depressivo	freq. assolute	freq.percentuali
grave	19	$19/36 \cdot 100 = 52.8\%$
medio	14	$14/36 \cdot 100 = 38.9\%$
leggero	3	$3/36 \cdot 100 = 8.3\%$
Totale	36	100

h. La distribuzione marginale del livello depressivo è la seguente:

livello depressivo	freq. assolute	freq.percentuali
grave	57	$57/159 \cdot 100 = 35.8\%$
medio	76	$76/159 \cdot 100 = 47.8\%$
leggero	26	$26/159 \cdot 100 = 16.4\%$
Totale	159	100

Analizzando la distribuzione condizionata allo stato civile vedovo, possiamo notare che la proporzione di vedovi con livello depressivo grave è superiore rispetto a quella calcolata sul totale. Dal confronto tra distribuzione condizionata e distribuzione marginale si può notare che le frequenze percentuali sono diverse, il che indica la presenza di un'associazione tra i due caratteri.

• • •

Esercizio 4.2. *Effetti collaterali di Avandia, I parte.*
(dal libro di testo *OpenIntro Statistics* di Diez et al., es. 1.51)

Il rosiglitazone è il principio attivo presente nella medicina per il diabete di tipo 2 Avandia ed è stato considerato causa dell'insorgenza di seri problemi cardiovascolari come infarto, insufficienza cardiaca e morte. Un trattamento alternativo è il pioglitazone, principio attivo presente in un'altra medicina per il diabete, Actos. Nella seguente tabella sono riportati i dati relativi ad uno studio osservazionale retrospettivo su 22,571 beneficiari di assistenza pubblica di età pari a 65 anni o più.

	Problemi cardiovascolari		
Trattamento	Si	No	Totale
Rosiglitazone	2,593	65,000	67,593
Pioglitazone	5,386	154,592	159,978
Totale	7,979	219,592	227,571

Determinare se ciascuna delle seguenti affermazioni è vera o falsa. Se falsa, spiegare perché. Attenzione: il ragionamento può essere sbagliato anche se la conclusione dell'affermazione è corretta. I questi casi, l'affermazione dovrebbe essere considerata falsa.

a. Poiché più pazienti con trattamento pioglitazone hanno avuto problemi cardiovascolari (5,386 vs. 2,593), possiamo concludere che il tasso di problemi cardiovascolari per quelli a cui è stato somministrato questo trattamento è più alto.

b. I dati suggeriscono che i pazienti diabetici a cui è stato somministrato rosiglitazone sono più inclini ad avere problemi cardiovascolari poiché il tasso di incidenza è $(2,593/67,593 = 0.038)$ 3.8% per pazienti con questo trattamento, mentre solo $(5,386/159,978 = 0.034)$ 3.4% per pazienti a cui è stato somministrato l'altro trattamento (pioglitazone).

c. Il fatto che il tasso di incidenza sia più alto per il gruppo rosiglitazone dimostra che il rosiglitazone causa seri problemi cardiovascolari.

d. Sulla base delle informazioni a disposizione, non possiamo dire se la differenza tra i tassi di incidenza è dovuta alla relazione tra le due variabili o al caso.

• • •

Soluzione.

a. Falso. Invece di confrontare le frequenze assolute, bisognerebbe confrontare le percentuali.

b. Vero.

c. Falso. Non possiamo dedurre una relazione causale da una associazione in uno studio osservazionale. Comunque, possiamo dire che il trattamento a cui uno è sottoposto ha un impatto sul rischio in questo caso, perché il paziente ha scelto quel trattamento e la sua scelta può essere associata ad altre variabili, che è il motivo per cui il punto b. è vero. La differenza in queste affermazioni è sottile ma importante.

d. Vero.

• • •

4.2 Inferenza su una singola proporzione

Esercizio 4.3. *Studenti fumatori*

Su un campione casuale di 100 studenti di un'università, 82 hanno dichiarato di non essere fumatori. Sulla base di questo, costruisci un

intervallo di confidenza a livello $1 - \alpha = 0.99$ per p, la proporzione di tutti gli studenti dell'università che non fumano.

• • •

Soluzione

Dobbiamo costruire un intervallo di confidenza per la proporzione p di tutti gli studenti dell'università che non fumano. Tale intervallo di confidenza ha la seguente forma

$$\left[\hat{p} - z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

dove

- $\hat{p}$ è la proporzione di non fumatori nel campione osservato: $\hat{p} = \frac{82}{100} = 0.82$;
- z^* è il quantile a livello $\alpha/2 = (1 - 0.99)/2 = 0.005$ di una distribuzione Normale standard; dalla tavola C (oppure dalla tavola della distribuzione Normale standard) si ha $z^* = 2.576$;
- $n = 100$ è la numerosità campionaria.

Sostituendo questi valori, si ottiene il seguente intervallo di confidenza:

$$[0.721, 0.919]$$

• • •

Esercizio 4.4. *Legalizzazione delle droghe leggere*

Un'indagine Gallup studia periodicamente un campione casuale di 1500 americani. La percentuale di individui nel campione che è a favore della legalizzazione del possesso di marijuana è scesa dal 52% nel 1980 al 46% nel 1985.

1. Si costruisca un intervallo di confidenza al 95% per la percentuale della popolazione a favore della legalizzazione nel 1980;

2. Si costruisca un intervallo di confidenza al 95% per la percentuale della popolazione a favore della legalizzazione nel 1985.

• • •

Soluzione

Dobbiamo costruire un intervallo di confidenza per una proporzione:

$$\left[\hat{p} - z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}; \hat{p} + z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

Quindi per ciascun anno si ha:

1. Per l'anno 1980:

$$\left[0.52 - 1.96\sqrt{\frac{0.52(1-0.52)}{1500}}; 0.52 + 1.96\sqrt{\frac{0.52(1-0.52)}{1500}}\right] = [0.495; 0.545]$$

2. Per l'anno 1985:

$$\left[0.46 - 1.96\sqrt{\frac{0.46(1-0.46)}{1500}}; 0.46 + 1.96\sqrt{\frac{0.46(1-0.46)}{1500}}\right] = [0.435; 0.485]$$

• • •

Esercizio 4.5. *Effetti collaterali di Avandia, II parte.*
(dal libro di testo *OpenIntro Statistics* di Diez et al., es. 1.53)

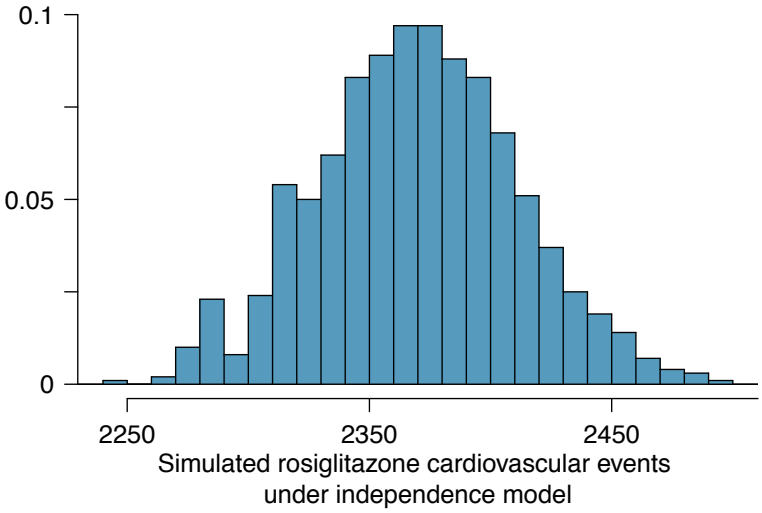
Nell'Esercizio 4.2 è stato introdotto uno studio in cui vengono confrontati i tassi di incidenza di alcune gravi patologie cardiovascolari per pazienti affetti da diabete, trattati con rosiglitazone e pioglitazone. La seguente tabella rappresenta i dati raccolti:

	Problemi cardiovascolari		
Trattamento	Si	No	Totale
Rosiglitazone	2,593	65,000	67,593
Pioglitazone	5,386	154,592	159,978
Totale	7,979	219,592	227,571

a. Quale è la proporzione di pazienti sul totale che hanno avuto problemi cardiovascolari?

b. Se il tipo di trattamento e l'insorgenza di problemi cardiovascolari fossero indipendenti, quanti pazienti con problemi cardiovascolari ci dovremmo aspettare nel gruppo rosiglitazone?

c. La relazione tra trattamento e outcome in questo studio può essere analizzata adottando una tecnica di randomizzazione. L'istogramma seguente rappresenta la simulazione dei conteggi degli eventi cardiovascolari nel gruppo rosiglitazione assumendo il modello di indipendenza. (i) Quali sono le ipotesi sottoposte a verifica? (ii) Rispetto al numero calcolato al punto b., fornirebbe maggiore supporto all'ipotesi alternativa un numero maggiore o un numero minore di pazienti con problemi cardiovascolari nel gruppo rosiglitazione? (iii) Cosa suggeriscono i risultati della simulazione rispetto alla relazione tra il trattamento rosiglitazione e l'insorgenza di problemi cardiovascolari nei pazienti diabetici?



• • •

Soluzione.

a. La proporzione di pazienti che hanno avuto problemi cardiovascolari è pari a $\frac{7979}{227571} \approx 0.035$.

b. Il numero atteso di problemi cardivascolari nel gruppo rosiglitazione sotto l'ipotesi di indipendenza, può essere calcolato moltiplicando il numero di pazienti di quel gruppo per il tasso complessivo di problemi cardiovascolari osservato nello studio, ovvero: $67593 \cdot \frac{7979}{227571} = 2730$

c. (i) L'ipotesi nulla H_0 corrisponde al modello di indipendenza: il trattamento e l'insorgenza di problemi cardiovascolari sono indipendenti, ovvero non c'è relazione tra loro, quindi la differenza riscontrata nei tassi di incidenza nei due gruppi di trattamento è dovuta al caso. L'ipotesi alternativa H_A corrisponde invece alla negazione del modello di indipendenza: il trattamento e l'insorgenza di problemi cardiovascolari non sono indipendenti, ovvero la differenza

riscontrata nei tassi di incidenza nei due gruppi di trattamento non è dovuta al caso, ma il rosiglitazione è associato con un maggior rischio di sviluppare problemi cardiovascolari. (ii) Un numero di pazienti con problemi cardiovascolari nel gruppo rosiglitazione piú elevato rispetto a quello atteso sotto l'ipotesi di indipendenza fornirebbe un maggiore supporto all'ipotesi alternativa. Questo suggerirebbe che il rosiglitazione comporta un incremento del rischio di problemi cardiovascolari. (iii) In questo studio sono stati effettivamente osservati 2593 eventi cardiovascolari nel gruppo rosiglitazione. Nelle 1000 simulazioni effettuate sotto il modello di indipendenza sono stati osservati praticamente sempre meno di 2593 eventi, il che induce a concludere che i dati osservati non siano compatibili con il modello di indipendenza. In altre parole, l'analisi fornisce forte evidenza contro l'ipotesi nulla ovvero a supporto dell'ipotesi che il rosiglitazione sia associato significativamente con un maggiore rischio di problemi cardiovascolari.

• • •

Esercizio 4.6. *Studenti vegetariani*
(dal libro di testo *OpenIntro Statistics* di Diez et al., es. 6.1)

Supponendo che l' 8% degli studenti siano vegetariani, determinare se le seguenti affermazioni sono vere o false, fornendo motivazioni appropriate.

a. La distribuzione della proporzione campionaria di vegetariani in un campione casuale di dimensione 60 è approssimativamente normale dal momento che $n \geq 30$.

b. La distribuzione della proporzione campionaria di vegetariani in un campione casuale di dimensione 50 è asimmetrica a destra.

c. Un campione casuale di 125 studenti di cui il 12% sono vegetariani può essere considerato anomalo.

d. Un campione casuale di 250 studenti di cui il 12% sono vegetariani può essere considerato anomalo.

e. L'errore standard si dimezzerebbe se la dimensione campionaria aumentasse da 125 a 250.

• • •

Soluzione.

a. Falso. Infatti non è soddisfatta la condizione: $np \geq 10$ e $n(1-p) \geq 10$.

b. Vero. Infatti non è soddisfatta la condizione: $np \geq 10$ e $n(1-p) \geq 10$. Inoltre, nella maggior parte dei campioni ci si può aspettare che $\hat{p}$ sia vicina a 0.08 che rappresenta la proporzione vera di vegetariani nella popolazione. Mentre $\hat{p}$ può assumere valori anche di molto superiori a 0.08, sarà certamente limitata dal valore 0; ciò implica che la forma della distribuzione tenderà ad essere asimmetrica a destra.

c. Falso. L'errore standard è pari a $SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = 0.0243$ e $\hat{p} = 0.12$ è distante soltanto $\frac{0.12-0.08}{0.0243} = 1.65SE$ dalla media, cosa che non può essere considerata anomala.

d. Vero. L'errore standard è pari a $SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = 0.0172$ e $\hat{p} = 0.12$ è distante soltanto $\frac{0.12-0.08}{0.0172} = 2.32SE$ dalla media e rappresenta quindi un valore anomalo.

e. Falso. L'errore standard si ridurrebbe di un fattore $\frac{1}{\sqrt{2}}$.

• • •

Esercizio 4.7. *Gatti rossi tigrati*
(dal libro di testo *OpenIntro Statistics* di Diez et al., es. 6.3)

Supponendo che il 90% dei gatti rossi tigrati sia maschio, determinare se le seguenti affermazioni sono vere o false, fornendo motivazioni appropriate.

a. La distribuzione della proporzione campionaria di un campione casuale di dimensione 30 è asimmetrica a sinistra.

b. Adottando una dimensione campionaria 4 volte maggiore, lo standard error della proporzione campionaria si dimezza.

c. La distribuzione della proporzione campionaria di un campione casuale di dimensione 140 è approssimativamente normale.

d. La distribuzione della proporzione campionaria di un campione casuale di dimensione 280 è approssimativamente normale.

• • •

Soluzione.

a. Vero. Infatti non è soddisfatta la condizione: $np \geq 10$ e $n(1-p) \geq 10$. Inoltre, nella maggior parte dei campioni ci si può aspettare che $\hat{p}$ sia vicina a 0.90 che rappresenta la proporzione vera di maschi nella popolazione. Mentre $\hat{p}$ può assumere valori anche di molto inferiori a 0.90, sarà certamente limitata dal valore 1; ciò implica che la forma della distribuzione tenderà ad essere asimmetrica a sinistra.

b. Vero. Nella formula dell'errore standard compare infatti la radice quadrata della numerosità campionaria.

c. Vero. Sono rispettate sia la condizione di indipendenza sia la condizione: $np \geq 10$ e $n(1-p) \geq 10$.

d. Vero. Sono rispettate sia la condizione di indipendenza sia la condizione: $np \geq 10$ e $n(1-p) \geq 10$.

• • •

Esercizio 4.8. *Prop 19 in California*
(dal libro di testo *OpenIntro Statistics* di Diez et al., es. 6.5)

In un'indagine condotta da Survey USA nel 2010, il 70% dei 119 rispondenti di età compresa tra i 18 e i 34 anni ha affermato che avrebbe votato a favore della cosiddetta *Prop 19*, un provvedimento per legalizzare la marjiuana in California modificando la legge vigente e adottando una opportuna regolamentazione e tassazione. Ad un livello di confidenza del 95%, il margine di errore per questo campione è pari all'8%. Sulla base di queste informazioni, determinare se le seguenti affermazioni sono vere o false, fornendo motivazioni appropriate.

a. Si può essere fiduciosi, con un livello di confidenza del 95%, che una percentuale di elettori compresa tra il 62% e il 78% in questo campione voterà a favore della Prop 19.

b. Si può essere fiduciosi, con un livello di confidenza del 95%, che tra gli elettori di età tra i 18 e i 34 anni, una percentuale compresa tra il 62% e il 78% voterà a favore della Prop 19.

c. Se si considerassero molti campioni casuali di 119 elettori di età tra i 18 e i 34 anni, e si calcolassero i corrispondenti intervalli di confidenza, il 95% di essi includerebbe il valore vero della proporzione di elettori favorevoli alla Prop 19 nella popolazione.

d. Per ridurre il margine di errore al 4%, è necessario moltiplicare la dimensione campionaria per 4.

e. In base a questo intervallo di confidenza, c'è evidenza sufficiente per concludere che la maggioranza degli elettori Californiani di età tra i 18 e i 34 anni, supportano la Prop 19.

• • •

Soluzione.

a. Falso. Un intervallo di confidenza viene costruito per stimare la proporzione nella popolazione, non nel campione.

a. Vero. L'intervallo di confidenza al 95% è $70\% \pm 8\%$.

c. Vero, per la definizione di intervallo di confidenza.

d. Vero. Moltiplicando la dimensione campionaria per 4, l'errore standard e il margine di errore si riducono di un fattore $\frac{1}{\sqrt{4}}$.

e. Vero. L'intervallo di confidenza al 95% è tutto al di sopra del 50%.

• • •

Esercizio 4.9. *Fuochi d'artificio il 4 Luglio*
(dal libro di testo *OpenIntro Statistics* di Diez et al., es. 6.7)

Alla fine di Giugno 2012, Survey USA ha pubblicato i risultati di un'indagine in cui si diceva che il 56% di 600 residenti del Kansas scelti in modo casuale aveva programmato di fare i fuochi d'artificio il 4 Luglio. Determinare il margine di errore per la stima puntuale per un livello di confidenza del 95%.

• • •

Soluzione.

Dal momento che il campione considerato è inferiore al 10% della popolazione, la condizione di indipendenza è soddisfatta. Anche la condizione $np \geq 10$ e $n(1-p) \geq 10$ è verificata. Il margine di errore è dunque: $ME = z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1.96 \sqrt{\frac{0.56 \cdot 0.44}{600}} = 0.0397 = 4\%$.

• • •

Esercizio 4.10. *Vita dopo il college*
(dal libro di testo *OpenIntro Statistics* di Diez et al., es. 6.9)

L'obiettivo di questa indagine è stimare la proporzione di laureati in una classe di piú di 4500 studenti che hanno trovato un lavoro entro un anno dalla laurea. Supponendo che 348 su 400 studenti estratti casualmente abbiano dichiarato di avere un lavoro, rispondere ai seguenti quesiti.

a. Descrivere il parametro di interesse della popolazione. Quale è la stima puntuale di questo parametro?

b. Controllare se le condizioni per la costruzione di un intervallo di confidenza sono soddisfatte dai dati a disposizione.

c. Calcolare un intervallo di confidenza al 95% per la proporzione di laureati che ha trovato lavoro entro un anno dalla laurea e fornire una sua interpretazione.

d. Cosa significa confidenza al 95%?

e. Calcolare ora un intervallo di confidenza al 99% per la proporzione di laureati che ha trovato lavoro entro un anno dalla laurea e fornire una sua interpretazione.

f. Confrontare le ampiezze dei due intervalli al livello 95% e 99%. Quale è piú ampio? Spiegare il perché.

• • •

Soluzione.

a. Il parametro di interesse della popolazione è la proporzione di laureati che ha trovato lavoro a un anno dalla laurea. La stima puntuale è $\hat{p} = 348/400 = 0.87$.

b. Il campione considerato è inferiore al 10% della popolazione, quindi la condizione di indipendenza è soddisfatta. Anche la condizione $np \geq 10$ e $n(1-p) \geq 10$ è verificata.

c. L'intervallo è: $(0.8371, 0.9029)$. Si può essere fiduciosi al 95% che approssimativamente una percentuale compresa tra l'84% e il 90% dei laureati abbia trovato lavoro entro un anno dalla laurea.

d. Significa che estraendo un gran numero di campioni e calcolando gli intervalli corrispondenti, si otterrebbe nel 95% dei casi un intervallo contenente il valore vero del parametro.

e. L'intervallo è $(0.8267, 0.9133)$. Si può essere fiduciosi al 99% che approssimativamente una percentuale compresa tra l'83% e il 91% dei laureati abbia trovato lavoro entro un anno dalla laurea.

f. L'intervallo a livello 99% è piú ampio, perché richiede un livello di fiducia maggiore che la proporzione vera sia contenuta all'interno dell'intervallo e quindi deve coprire un range maggiore.

• • •

Esercizio 4.11. *Studiare all'estero*
(dal libro di testo *OpenIntro Statistics* di Diez et al., es. 6.11)

Un'indagine su 1509 studenti liceali che hanno compilato un questionario online non obbligatorio tra il 25 e il 30 aprile 2007 mostra che il 55% degli intervistati è piuttosto sicuro che in seguito parteciperà ad un programma di studio all'estero.

a. Questo campione è rappresentativo della popolazione di tutti i liceali degli Stati Uniti? Spiegare il perché.

b. Supponendo che le condizioni per fare inferenza siano soddisfatte, anche se la risposta al punto a. indicasse che questo approccio non è

affidabile, questa analisi potrebbe essere ancora interessante. Costruire un intervallo di confidenza al 90% per la proporzione di studenti che è piuttosto sicura che in seguito parteciperà ad un programma di studio all'estero e fornire una sua interpretazione.

c. Cosa significa confidenza al 90%?

d. In base a questo intervallo, sarebbe corretto afferemare che la mggior parte degli studenti è piuttosto sicura che in seguito parteciperà ad un programma di studio all'estero?

• • •

Soluzione.

a. No. Si tratta di un campione di volontari, quindi un campione non casuale.

b. $(0.5289, 0.5711)$. Si può essere fiduciosi al 90% che una percentuale di studenti compresa tra il 53% e il 57% sia piuttosto sicura che in seguito parteciperà ad un programma di studio all'estero.

c. Significa che estraendo un gran numero di campioni e calcolando gli intervalli corrispondenti, si otterrebbe nel 90% dei casi un intervallo contenente il valore vero del parametro.

d. Si perché l'intervallo cade al di sopra del valore 0.5.

• • •

Esercizio 4.12. *Sistema sanitario pubblico, I parte*
(dal libro di testo *OpenIntro Statistics* di Diez et al., es. 6.13)

Un articolo del Washington Post del 2009 ha riportato che 7 democratici su 10 sostengono la riforma della sanità, mentre quasi 9 su 10 repubblicani

sono contrari a questa riforma. Il 52% degli indipendenti sono contro e il 42% a favore (il 6% rispondono 'altro'). Complessivamente erano stati intervistati 819 democratici, 566 repubblicani e 783 indipendneti.

a. Una trasmissione televisiva riportando questa notizia, ha affermato che la maggior parte degli Indipendenti è contraria alla riforma. Questi dati forniscono una forte evidenza a supporto di questa affermazione?

b. Ci si può aspettare che un intervallo di confidenza per la proporzione di indipendenti contrari alla riforma includa il valore 0.5? Motivare la risposta.

• • •

Soluzione.

a. In questo caso si può impostare il seguente sistema di ipotesi: H_0 : $p = 0.50$ vs $H_A : p > 0.50$. Le due condizioni (indipendenza e $np \geq 10$ e $n(1-p) \geq 10$) sono entrambe verificate. In questo caso si ottiene un valore osservato della statistica test $z = 1.12$ che corrisponde ad un p-value pari a 0.1314. Dal momento che il p-value supera la soglia 0.05, non è possibile rifiutare H_0, ovvero i dati non forniscono forte evidenza a favore dell'affermazione di interesse.

b. Si, da quanto affermato al punto precedente segue che l'intervallo conterrà il valore 0.5.

• • •

Esercizio 4.13. *Internet su dispositivi mobili*
(dal libro di testo *OpenIntro Statistics* di Diez et al., es. 6.15)

Un'indagine del 2012 su 2254 americani adulti ha mostrato che il 17% di coloro che possiedono un telefono cellulare naviga su internet utilizzando il suo dispositivo mobile anziché un personal computer.

a. Secondo un articolo pubblicato online, una ricerca condotta da una compagnia telefonica ha mostrato che il 38% degli utenti Cinesi accede a internet solo attraverso i cellulari. Verificare mediante un test di ipotesi se questi dati supportano l'ipotesi che la proporzione di Americani che utilizza il telefono cellulare per navigare su internet è differente dalla medesima proporzione nella popolazione cinese.

b. Interpretare il p-value ottenuto al punto a.

c. Calcolare un intervallo al 95% per la proporzione di americani che utilizza il telefono cellulare per navigare su internet e fornire un'interpretazione.

• • •

Soluzione.

a. In questo caso si può impostare il seguente sistema di ipotesi: H_0 : $p = 0.38$ vs $H_A : p \neq 0.38$. Le due condizioni (indipendenza e $np \geq 10$ e $n(1-p) \geq 10$) sono entrambe verificate. Il valore osservato della statistica test è $z = 20.5$ e il corrispondente p-value ≈ 0, Dal momento che il p-value è trascurabile si può rifiutare l'ipotesi nulla, quindi i dati forniscono forte evidenza che la proporzione di Americani che utilizza il telefono cellulare per navigare su internet è diversa dalla (in particolare, inferiore alla) proporzione corrispondente nella popolazione Cinese.

b. Se il 38% degli americani usasse il cellulare come mezzo per accedere a internet, la probabilità di ottenere un campione casuale di 2254 americani in cui una percentuale inferiore o uguale al 17% o superiore o uguale al 59% di utenti di internet via cellulare sarebbe pressoché trascurabile.

c. L'intervallo è $(0.1545, 0.1855)$. Si può avere fiducia a livello 95% che approssimativamente una percentuale compresa tra il 15% e il 18.6% degli americani utilizza il proprio cellulare per navigare su internet.

• • •

Esercizio 4.14. *Test sul gusto*
(dal libro di testo *OpenIntro Statistics* di Diez et al., es. 6.17)

Alcune persone sostengono di poter individuare la differenza tra una bevanda dietetica e una normale al primo sorso. Un ricercatore volendo sottoporre a verifica questa affermazione, ha estratto un campione casuale tra queste persone di numerosità pari a 80. Ha poi predisposto 40 bicchieri della bevanda dietetica e 40 di quella normale in modo casuale e infine ha chiesto a ciascun assaggiatore di provare le bevande e classificarle come dietetiche o regolari. 53 partecipanti hanno classificato correttamente le bevande.

a. Questi dati forniscono forte evidenza del fatto che queste persone sono capaci di individuare la differenza tra la bevanda dietetica e quella normale? In altre parole, i risultati sono significativamente migliori rispetto a un'assegnazione casuale alle due tipologie?

b. Interpretare il p-value ottenuto al punto a.

• • •

Soluzione.

a. In questo caso si può impostare il seguente sistema di ipotesi: H_0 : $p = 0.5$ vs H_A : $p > 0.5$. Le due condizioni (indipendenza e $np \geq 10$ e $n(1-p) \geq 10$) sono entrambe verificate. Il valore osservato della statistica test è $z = 2.91$ e il corrispondente p-value è pari a 0.0018. Poiché il p-value è inferiore a 0.05, rifiutiamo l'ipotesi nulla. I dati forniscono forte evidenza che il tasso di corretta identificazione della tipologia di bevanda di queste persone è significativamente migliore rispetto a un'assegnazione casuale.

b. Se le persone assegnassero casualmente la tipologia di bevanda, la probabilità di ottenere una campione casuale in cui 53 persone su 80 identificassero correttamente la bevanda sarebbe pari a 0.0018.

• • •

Esercizio 4.15. *Fumatori universitari*
(dal libro di testo *OpenIntro Statistics* di Diez et al., es. 6.19)

Si vuole stimare la proporzione di studenti universitari fumatori. In un campione casuale di 200 studenti universitari, 40 sono fumatori.

a. Calcolare un intervallo di confidenza al 95% per la proporzione di studenti universitari fumatori, e fornire un'interpretazione dell'intervallo.

b. Se si volesse un margine di errore non superiore al 2% per l'intervallo di confidenza al 95%, quale dovrebbe essere la dimensione campionaria?

• • •

Soluzione.

a. Le due condizioni (indipendenza e $np \geq 10$ e $n(1-p) \geq 10$) sono entrambe verificate. L'intervallo di confidenza a livello 95% risulta essere $(0.145, 0.255)$. Si può avere un livello di fiducia del 95% che una percentuale compresa tra il 14% e il 25.5% degli studenti universitari fumi.

b. z^*SE non deve eccedere il valore 0.02. Dato che $z^* = 1.96$, sostituendo la stima puntuale di p, $\hat{p} = 0.2$ nella formula dell'errore standard si ha $1.96\sqrt{0.2(1-0.2)/n} \leq 0.02$, da cui segue che la numerosità campionaria dovrà essere almeno pari a 1537.

• • •

Esercizio 4.16. *Sistema sanitario pubblico, II parte*
(dal libro di testo *OpenIntro Statistics* di Diez et al., es. 6.21)

Nell'Esercizio 4.12 si è detto che il 52% degli Indipendenti intervistati a proposito della riforma del sistema sanitario, si è dichiarato contrario alla riforma stessa. Se volessimo stimare questo numero con un margine di errore dell'1% con un livello di confidenza del 90%, quale dimensione campionaria risulterebbe adeguata a tale obiettivo?

• • •

Soluzione.

Il margine di errore z^*SE deve essere minore di 0.01. Poiché vogliamo un livello di confidenza del 90% avremo $z^* = 1.65$ e sostiuiremo la stima puntuale $\hat{p} = 0.52$ nella formula $1.96\sqrt{0.52(1-0.52)/n} \leq 0.01$, ottenendo una numerosità campionaria maggiore o uguale a 6796.

• • •

Esercizio 4.17. *Attività in fallimento*

In un campione di 400 proprietari di negozi e piccole imprese, che hanno dichiarato fallimento, 88 non hanno alcuna esperienza professionale precedente.

1. Sottoporre a test l'ipotesi nulla che il 25% di coloro che vanno in fallimento non hanno esperienze precedenti al livello di significatività del 5% contro l'ipotesi alternativa che la percentuale sia inferiore;

2. Definire il $p-value$ del test e calcolarlo;

3. Se il livello di significatività fosse stato il 10% l'ipotesi nulla sarebbe stata respinta?

• • •

Soluzione

È un test sulla proporzione di successi. Qui il successo è "il proprietario ha dichiarato fallimento". La proporzione stimata è $\hat{p} = 88/400 = 0.22$. Useremo l'approssimazione normale della statistica test, avendo cura di sostituire la deviazione standard con l'errore standard. Le ipotesi sono:

$$H_0 : \ p = p_0 = 0.25; \qquad H_1 : \ p < 0.25$$

Determiniamo il $p - value$:

$$p - value = P(Z < z)$$

La statistica test Z si distribuisce come una normale standard. Il valore z è il valore della statistica test osservato nel campione

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.22 - 0.25}{\sqrt{\frac{0.25 \cdot (1-0.25)}{400}}} = -1.39$$

Il $p - value$ risulta quindi

$$p - value = P(Z < z) = 0.0823$$

Non possiamo rifiutare l'ipotesi H_0 a livello di significatività $\alpha = 0.05$ perché ovviamente il $p - value$ è > 0.05.

Se consideriamo invece un livello di significatività $\alpha = 0.1$, poiché $p - value < \alpha$, possiamo rifiutare l'ipotesi nulla.

• • •

Esercizio 4.18. *Educazione in TV*

Un famoso educatore dichiara che piú della metà della popolazione adulta degli USA è preoccupata dalla carenza di programmi educativi

in televisione. Per raccogliere dati sulla questione, nell'ambito di un sondaggio nazionale vengono scelti e intervistati 920 individui. Se 478 (52%) degli intervistati dichiarano di essere preoccupati, abbiamo dimostrato la dichiarazione dell'educatore?

• • •

Soluzione

Dobbiamo valutare il seguente sistema di ipotesi per la proporzione p di popolazione americana preoccupata per la carenza di programmi educativi in televisione:

$$H_0 : p = 0.50 \qquad H_1 : p > 0.50$$

Si tratta di un test per proporzioni; in questo caso la statistica test, Z, è definita come

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

dove

- $\hat{p}$ è la proporzione di individui preoccupati nel campione osservato ($\hat{p} = 0.52$);
- n è la numerosità campionaria ($n = 920$).

Si dimostra che sotto l'ipotesi nulla H_0, questa statistica test ha distribuzione Normale standardizzata, ossia

$$Z \sim N(0, 1)$$

Al fine di valutare questo sistema di ipotesi possiamo calcolare il p-value corrispondente, ossia dobbiamo valutare $Pr(Z > z)$ dove $z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$ con

p_0 valore della proporzione assunta sotto l'ipotesi nulla ($p_0 = 0.5$). Possiamo quindi calcolare il $p - value$ come segue:

$$\begin{aligned} p - value = Pr(Z > z) &= Pr\left(Z > \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}\right) = \\ &= Pr\left(Z > \frac{0.52 - 0.50}{\sqrt{\frac{0.50(1-0.50)}{920}}}\right) = \\ &= Pr(Z > 1.21) = 1 - Pr(Z \leq 1.21) = 1 - 0.887 = 0.113 \end{aligned}$$

dove il valore 0.877 è stato controllato sulle tavole della distribuzione Normale standardizzata.
Concludendo, poiché il $p - value = 0.113$, possiamo rifiutare l'ipotesi nulla e quindi validiamo l'ipotesi dell'educatore, solo se consideriamo un livello di significatività $\alpha > 0.113$ (con $\alpha = 0.05$ o $\alpha = 0.10$ l'ipotesi nulla non viene rifiutata).

• • •

Esercizio 4.19. *Pubblicazioni su riviste internazionali*

Un professore ritiene che la percentuale di ricercatori che, nel suo settore scientifico disciplinare, pubblicano su riviste internazionali è pari al 70%. In un campione di 160 ricercatori, 108 hanno pubblicazioni internazionali. Verificare il seguente sistema di ipotesi:

$$H_0 : p = 0.7 \quad H_1 : p < 0.7$$

• • •

Soluzione

Anche in questo caso possiamo procedere come fatto nel precedente esercizio calcolando il p-value come segue ($\hat{p} = \frac{108}{160} = 0.675$):

$$\begin{aligned} p - value &= Pr(Z < z) = Pr\left(Z < \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}\right) = \\ &= Pr\left(Z < \frac{0.675 - 0.70}{\sqrt{\frac{0.70(1-0.70)}{160}}}\right) = Pr(Z < -0.69) = 0.2451 \end{aligned}$$

Pertanto, il valore minimo di α per rifiutare l'ipotesi nulla è 0.25.

• • •

Esercizio 4.20. *Bullismo nelle scuole*
(dal libro di testo *OpenIntro Statistics* di Diez et al., es. 6.47)

In un'indagine campionaria USA del 2012 i residenti in Florida sono stati interrogati su quanto grande fosse secondo loro il problema del bullismo nelle scuole locali. 9 su 191 tra i 18 e i 34 anni hanno risposto che il bullismo non è affatto un problema. Usando questi dati, è possibile costruire un intervallo di confidenza utilizzando la formula $\widehat{p} \pm z^* \sqrt{\widehat{p}(1-\widehat{p})/n}$ per la proporzione vera di residenti in Florida di età 18-34 che pensano che il bullismo non sia per niente un problema? Se si ritiene appropriato, costruire l'intervallo di confidenza, altrimenti, spiegare il perché.

• • •

Soluzione. Non è appropriato. Ci sono solo 9 successi nel campione, quindi, la condizione $np \geq 10$ e $n(1-p) \geq 10$ non è verificata.

• • •

4.3 Inferenza sulla differenza tra due proporzioni

Esercizio 4.21. *Esperimento sociologico, I parte*
(dal libro di testo *OpenIntro Statistics* di Diez et al., es. 6.23)

Un esperimento sociologico condotto da un programma televisivo ha analizzato il comportamento di alcune persone quando assistono a un litigio di coppia in cui la donna viene palesemente offesa dall'uomo, in due differenti occasioni allo stesso ristorante. Nel primo caso la donna è vestita in modo provocante e nell'altro caso è invece abbligliata in modo castigato. La seguente tabella riassume i dati raccolti su quante persone hanno deciso di intervenire o meno:

	Provocante	Castigato	Totale
Intervenuti	5	15	20
Non intervenuti	15	10	25
Totale	20	25	45

Spiegare perché la distribuzione campionaria della differenza tra le proporzioni di intervento sotto i due scenari non segue una distribuzione approssimativamente normale.

• • •

Soluzione.

Si tratta di un esperimento non randomizzato e non è chiaro se le persone possono essere influenzato dal comportamento degli altri avventori del ristorante. In questo caso non è quindi possibile assumere l'indipendenza. In piú ci sono solo 5 persone intervenute nel caso dello scenario 'Provocante', quindi non vale neanche la condizione $np \geq 10$ e $n(1-p) \geq 10$. Anche se considerassimo un test di ipotesi basato su una media delle proporzioni, tale condizione non potrebbe essere soddisfatta. Per questi motivi non è possibile

assumere che la distribuzione campionaria della differenza delle proporzioni sia approssimativamente normale.

• • •

Esercizio 4.22. *Sesso e preferenze sui colori*
(dal libro di testo *OpenIntro Statistics* di Diez et al., es. 6.25)

In uno studio del 2001, 1924 maschi e 3666 femmine tra gli studenti del college sono stati intervestati in merito al loro colore preferito. Un intervallo al 95% per la differenza tra le proporzioni di maschi e femmine il cui colore preferito è il nero $(p_{male} - p_{female})$ è risultato essere $(0.02, 0.06)$. Sulla base di questa informazione, determinare se le seguenti affermazioni sono vere o false, motivando la risposta.

a. Si può essere fiduciosi al 95% che la proporzione vera di maschi il cui colore preferito è nero sia il 2% inferiore e il 6% superiore rispetto alla proporzione vera di femmine.

b. Si può essere fiduciosi al 95% che la proporzione vera di maschi il cui colore preferito è nero sia tra il 2% e il 6% superiore rispetto alla proporzione vera di femmine.

c. Il 95% dei campioni casuali produrrà intervalli di confidenza che includono la differenza vera tra le proporzioni di maschi e femmine il cui colore preferito è nero.

d. Possiamo concludere che c'è una differenza significativa tra le proporzioni di maschi e femmine il cui colore preferito è nero e che la grandezza della differenza tra le due proporzioni campionarie sia plausibilmente imputabile al caso.

e. L'intervallo di confidenza al 95% per $(p_{female} - p_{male})$ non può essere calcolato sulla base delle informazioni disponibili in questo esercizio

• • •

Soluzione.

a. Falso. L'intero intervallo di confidenza supera lo 0.

b. Vero.

c. Vero.

d. Vero.

e. Falso. Si può ottenere banalmente cambiando i segni all'intervallo di cui sopra, ovvero $(-0.06, -0.02)$.

• • •

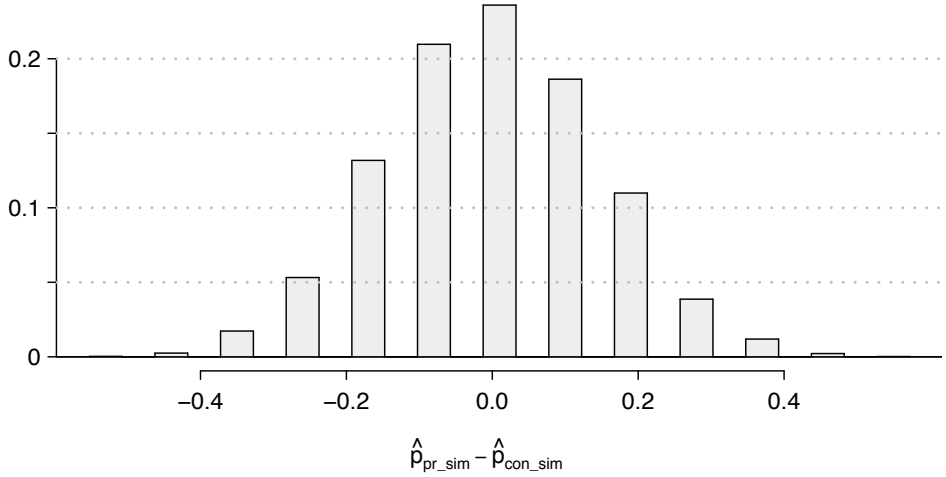
Esercizio 4.23. *Esperimento sociologico, II parte*
(dal libro di testo *OpenIntro Statistics* di Diez et al., es. 6.51)

Nell'esercizio 4.21 è stato introdotto un esperimento sociologico condotto da un programma televisivo in cui è stato analizzato il comportamento di alcune persone quando assistono a un litigio di coppia in cui la donna viene palesemente offesa dall'uomo, in due differenti occasioni allo stesso ristorante. Nel primo caso la donna è vestita in modo provocante e nell'altro caso è invece in modo castigato. La seguente tabella riassume i dati raccolti su quante persone hanno deciso di intervenire o meno:

	Provocante	Castigato	Totale
Intervenuti	5	15	20
Non intervenuti	15	10	25
Totale	20	25	45

Una simulazione è stata effettuata per verificare se le persone reagiscono in maniera diversa nelle due situazioni. 10000 differenze simulate sono state

generate per costruire la distribuzione sotto l'ipotesi nulla. Il valore $\widehat{p}_{pr,sim}$ rappresenta la proporzione di clienti che è intervenuta nella simulazione per difendere una donna vestita in modo provocante e $\widehat{p}_{con,sim}$ la proporzione che è intervenuta per una donna vestita in modo castigato.



(a) Quali sono le ipotesi? Per gli scopi di questo esercizio, si può assumere che ogni persona osservata al ristorante si comporti in modo indipendente, anche se tale assunzione dovrebbe essere verificata in modo rigoroso se volessimo riportare ufficialmente i risultati dei questo esperimento.

(b) Calcolare la differenza osservata tra i tassi di intervento nelle due situazioni: $p_{pr} - p_{con}$.

(c) Stimare il p-value usando il grafico riportato sopra. Cosa si può dedurre?

• • •

Soluzione. Il suffisso $_{pr}$ corrisponde a provocante e $_{con}$ a castigato.

(a) H_0: $p_{pr} = p_{con}$. H_A: $p_{pr} \neq p_{con}$.

(b) -0.35.

(c) La coda sinistra per il p-value è calcolata sommando 0.005 e 0.015. Raddoppiando tale valore (0.02) si ottiene che il p-value è pari a 0.04. (Gli studenti possono ottenere risultati approssimati, e un piccolo numero di studenti può ottenere un p-value pari a 0.05.) Poiché il p-value è piccolo, rifiutiamo H_0. I dati forniscono una forte evidenza empirica che le persone reagiscono in modo diverso nelle due situazioni.

• • •

Esercizio 4.24. *Sistema sanitario pubblico, III parte*
(dal libro di testo *OpenIntro Statistics* di Diez et al., es. 6.27)

L'esercizio 4.12 presenta i risultati di un sondaggio per valutare il sostegno alla riforma del sistema sanitario nel 2009. Il 70% dei democratici e il 42% degli indipendenti supporta tale riforma.

(a) Costruire l'intervallo di confidenza al 95% per la differenza tra p_D e p_I, $(p_D - p_I)$, e commentare i risultati. Già sono state verificate le condizioni.

(b) Vero o falso: se estraiamo casualmente un democratico e un indipendente, contemporaneamente, dal campione preso in esame, è più probabile che un democratico sostenga la riforma del sistema sanitario piuttosto che un indipendente.

• • •

Soluzione.

(a) L'intervallo di confidenza al 95% è $(0.23, 0.33)$. Quindi, siamo sicuri al 95% che la proporzione di Democratici che sostiene il sistema sanitario pubblico è da 23% a 33% più alta della proporzione di Indipendentisti.

(b) Vero.

• • •

Esercizio 4.25. *Trivellazione in mare aperto, I parte*
(dal libro di testo *OpenIntro Statistics* di Diez et al., es. 6.29)

Un'indagine del 2010 ha chiesto a 827 elettori scelti casualmente in California "Sei a favore o contro la trivellazione per estrarre petrolio e gas naturale al largo delle coste della California? Oppure non ne sai abbastanza per esprimerti?" Di seguito è riportata la distribuzione delle risposte, in cui gli elettori sono stati divisi tra laureati e non laureati.

	Laureati	Non laureati
Favorevoli	154	132
Contrari	180	126
Non sanno	104	131
Totale	438	389

(a) Qual è la percentuale di laureati e quale la percentuale di non laureati in questo campione che non ne sa abbastanza per avere un'opinione sull'estrazione di petrolio e gas naturale al largo delle coste della California?

(b) Usare un test d'ipotesi per determinare se vi è una forte evidenza empirica per cui la proporzione di laureati che non ha un'opinione sull'argomento è diversa dalla proporzione di non laureati.

• • •

Soluzione.

(a) Laureati: $23,7\%$. Non laureati: $33,7\%$.

(b) Siano p_L e p_{NL}, rispettivamente, la proporzione di laureati e la proporzione di non laureati che hanno risposto "non so". $H_0 : p_L = p_{NL}$ e $H_A : p_L \neq p_{NL}$. Le due condizioni (indipendenza e $np \geq 10$ e $n(1-p) \geq 10$) sono entrambe verificate. Per la seconda condizione si usa la proporzione empirica/stimata ($\widehat{p} = 235/827 = 0,284$). $Z = -3.18 \rightarrow p$-value$= 0.0014$. Poiché il p-value è molto piccolo, si rifiuta H_0. C'è abbastanza evidenza sperimentale per poter rifiutare l'ipotesi nulla, in altre parole la differenza tra laureati e non laureati che non hanno un'opinione sull'argomento è statisticamente significativa. I dati indicano anche che meno laureati che non laureati hanno risposto "non so" (cioè i dati indicano la direzione dopo il rifiuto di H_0).

• • •

Esercizio 4.26. *Trivellazione in mare aperto, II parte*
(dal libro di testo *OpenIntro Statistics* di Diez et al., es. 6.31)

I risultati di un'indagine sulla trivellazione per l'estrazione di petrolio e gas naturale al largo della costa della California sono stati introdotti nell'Esercizio **??**.

(a) Qual è la percentuale di laureati e quale la percentuale di non laureati in questo campione che è a favore dell'estrazione di petrolio e gas naturale al largo delle coste della California?

(b) Usare un test d'ipotesi per determinare se vi è una forte evidenza empirica per cui la proporzione di laureati che è favorevole è diversa dalla proporzione di non laureati.

• • •

Soluzione.

(a) Laureati: $35,2\%$. Non laureati: $33,9\%$.

(b) Siano p_L e p_{NL}, rispettivamente, la proporzione di laureati e la proporzione di non laureati che sono favorevoli. $H_0 : p_L = p_{NL}$ e $H_A : p_L \neq p_{NL}$. Le due condizioni (indipendenza e $np \geq 10$ e $n(1-p) \geq 10$) sono entrambe verificate. Per la seconda si usa la proporzione empirica/stimata ($\widehat{p} = 286/827 = 0.346$). $Z = 0.39 \rightarrow$ p-value$= 0.6966$. Poiché il p-value è maggiore di α (0.05), non si può rifiutare H_0. Non c'è abbastanza evidenza sperimentale per poter rifiutare l'ipotesi nulla, in altre parole la differenza tra laureati e non laureati che sono favorevoli alla trivellazione in California non è statisticamente significativa.

• • •

Esercizio 4.27. *Carenza di sonno dei lavoratori del mondo dei trasporti* **(dal libro di testo *OpenIntro Statistics* di Diez et al., es. 6.33)**

La Fondazione del sonno statunitense ha condotto un'indagine sulle abitudini di un campione aleatorio di lavoratori dei trasporti e un campione controllo di persone che non lavorano nel mondo dei trasporti. I risultati dell'indagine sono riportati sotto.

		Lavoratori dei Trasporti			
Ore di sonno	*Controllo*	Piloti	Camionisti	Operatori treno	Autisti bus/taxi/limo
< 6	35	19	35	29	21
$[6, 8]$	193	132	117	119	131
> 8	64	51	51	32	58
Totale	292	202	203	180	210

Usare un test d'ipotesi per valutare se i dati forniscono una forte evidenza sulla significatività della differenza tra la proporzione di camionisti e il gruppo

controllo che dormono meno di 6 ore al giorno, cioè che sono considerati carenti di sonno.

• • •

Soluzione. Indichiamo con NT il gruppo controllo e con C i camionisti. $H_0 : p_{NT} = p_C$ e $H_A : p_{NT} \neq p_C$. Le due condizioni (indipendenza e $np \geq 10$ e $n(1-p) \geq 10$) sono entrambe verificate. Per la seconda si usa la proporzione empirica/stimata ($\widehat{p} = 70/495 = 0,141$). $Z = -1,58 \rightarrow p$-value$= 0,1164$. Poiché il p-value è maggiore di α (0,05), non si può rifiutare H_0. Non c'è una forte evidenza sperimentale per poter rifiutare l'ipotesi nulla. La differenza tra i tassi di carenza di sonno del gruppo controllo e il gruppo dei camionisti non è statisticamente significativa.

• • •

Esercizio 4.28. *HIV in Africa sub-Sahariana*
(dal libro di testo *OpenIntro Statistics* di Diez et al., es. 6.35)

A luglio 2008 gli istituti nazionali statunitensi di sanità hanno annunciato che era stato interrotto uno studio clinico a causa di risultati inaspettati. La popolazione oggetto di studio era formata da donne dell'Africa sub-Sahariana a cui era stata somministrata una singola dose di Nevaripine (un trattamento per l'HIV) durante il parto, per prevenire la trasmissione dell'HIV al neonato. Con questo studio ci si chiedeva se proseguire il trattamento dopo il parto con Nevaripine oppure con il trattamento alternativo Lopinavir. Allo studio hanno partecipato 240 donne; 120, aleatoriamente scelte, per ogni trattamento. Ventiquattro settimane dopo aver iniziato lo studio sul trattamento, ogni donna è stata analizzata per vedere se la situazione era peggiorata (un risultato chiamato fallimento virologico). Un fallimento virologico è stato riscontrato su 26 delle 120 donne trattate con Nevaripine e 10 delle 120 donne a cui era stato somministrato l'altro trattamento.

(a) Costruire una tabella a due vie in cui vengono riportati i risultati dello studio.

(b) Definire un appropriato test d'ipotesi per verificare l'indipendenza tra trattamento e fallimento virologico.

(c) Analizzare i risultati del test d'ipotesi e trarre le conclusioni. (N.B: verificare tutte le condizioni necessarie per il test.)

• • •

Soluzione.

(a) Sintesi dello studio

	Fallimento Virale		
	Si	No	Totale
Nevaripine	26	94	120
Lopinavir	10	110	120
Totale	36	204	240

(b) $H_0 : p_N = p_L$. Non c'è differenza tra i tassi di fallimento virologico nei due gruppi (Nevaripine e Lopinavir). $H_A : p_N \neq p_L$. C'è differenza tra i tassi di fallimento virologico nei due gruppi.

(c) È stata usata un'assegnazione aleatoria, quindi, le osservazioni in ciascun gruppo sono indipendenti. Se i pazienti in uno studio sono rappresentativi di quelli dell'intera popolazione (impossibile da verificare con le informazioni a disposizione), allora possiamo anche generalizzare i risultati alla popolazione. La condizione $np \geq 10$ e $n(1-p) \geq 10$, che si verifica usando la proporzione campionaria ($\widehat{p} = 36/240 = 0.15$), è soddisfatta. $Z = 3.04 \rightarrow p$-value$= 0.0024$. Poiché il p-value è piccolo, si può rifiutare H_0. C'è una forte evidenza

sperimentale per poter rifiutare l'ipotesi nulla. La differenza tra i tassi di fallimento virologico del gruppo Nevaripine e del gruppo Lopinavir è statisticamente significativa.

• • •

4.4 Verifica della bontà di adattamento

Esercizio 4.29. *Vero o Falso, I Parte*
(dal libro di testo *OpenIntro Statistics* di Diez et al., es. 6.37)

Indicare se le seguenti affermazioni sono vere o false. Per ogni affermazione falsa, suggerire un modo alternativo di scriverla per renderla vera.

(a) La distribuzione χ^2, così come la distribuzione Normale, ha due parametri, la media e la deviazione standard.

(b) La distribuzione χ^2 è sempre asimmetrica a destra, qualsiasi sia il valore del parametro "gradi di libertà".

(c) La statistica Chi quadrato (X^2) è sempre positiva.

(d) All'aumentare dei gradi di libertà, la forma della distribuzione χ^2 diventa più asimmetrica.

• • •

Soluzione.

(a) Falso. La distribuzione χ^2 ha un parametro chiamato "gradi di libertà".

(b) Vero.

(c) Vero.

(d) Falso. All'aumentare dei gradi di libertà, la forma della distribuzione χ^2 diventa più simmetrica.

• • •

Esercizio 4.30. *Libro di testo open-source*
(dal libro di testo *OpenIntro Statistics* di Diez et al., es. 6.39)

Un professore che usa un libro di introduzione alla statistica open-source prevede che il 60% degli studenti comprerà una copia cartacea del libro, il 25% stamperà il libro dal web e il 15% lo leggerà online. Alla fine del semestre chiede ai suoi studenti di rispondere a un questionario dove dovranno indicare qual è il formato del libro che hanno usato. Dei 126 studenti, 71 hanno detto di aver comprato una copia cartacea del libro, 30 lo hanno stampato dal web e 25 lo hanno letto online.

(a) Definire le ipotesi per verificare se le previsioni del professore erano accurate.

(b) Quanti studenti il professore si aspettava che comprassero il libro, lo stampassero e lo leggessero esclusivamente online?

(c) Si tratta di una situazione appropriata per usare un test Chi quadrato. Elencare le condizioni richieste per tale test e verificare che siano soddisfatte.

(d) Calcolare la statistica Chi quadrato, i gradi di libertà associati e il p-value.

(e) Sulla base del p-value calcolato, cosa possiamo concludere? Commentare i risultati ottenuti

• • •

Soluzione.

(a) H_0 : La distribuzione del formato del libro usato dagli studenti è quella prevista dal professore. H_A : La distribuzione del formato del libro usato dagli studenti non è quella prevista dal professore

(b) $E_{copia\ cartacea} = 126 \times 0.60 = 75.6$. $E_{stampa} = 126 \times 0.25 = 31.5$. $E_{online} = 126 \times 0.15 = 18.9$

(c) Indipendenza: il campione non è aleatorio. Comunque, se il professore ritiene che le proporzioni siano stabili da un periodo (semestre) al successivo e che le abitudini degli studenti non siano influenzate da quelle degli altri, allora l'indipendenza è probabilmente ragionevole.

(d) $X^2 = 2.32$, i gradi di libertà sono 2 e il p-value è maggiore di 0.3.

(e) Poiché il p-value è grande, non possiamo rifiutare H_0. I dati non forniscono una forte evidenza che indichi che le previsioni del professore siano statisticamente non significative.

• • •

4.5 Test di indipendenza

Esercizio 4.31. *Depressione e stato civile, II parte*

Torniamo a considerare i dati relativi all'Esercizio 4.1. La seguente tabella riporta il livello depressivo osservato rispetto allo stato civile:

	stato civile			
livello depressivo	sposato	celibe	vedovo	Totale
grave	22	16	19	57
medio	33	29	14	76
leggero	14	9	3	26
Totale	69	54	36	159

1. Determinare la distribuzione marginale e la distribuzione condizionata del livello depressivo allo stato civile vedovo e confrontarle. Cosa si può dire sull'associazione tra i due caratteri?

2. Verificare con un opportuno test l'ipotesi che ci sia associazione tra i due caratteri fissando il livello di significatività a 0.05.

• • •

Soluzione.

1. La distribuzione marginale del livello depressivo è la seguente:

livello depressivo	freq. assolute	freq.percentuali
grave	57	$57/159 \cdot 100 = 35.8\%$
medio	76	$76/159 \cdot 100 = 47.8\%$
leggero	26	$26/159 \cdot 100 = 16.3\%$
Totale	159	100

La distribuzione condizionata del livello depressivo allo stato civile vedovo è riportata nella seguente tabella:

stato civile=vedovo

livello depressivo	freq. assolute	freq.percentuali
grave	19	$19/36 \cdot 100 = 52.8\%$
medio	14	$14/36 \cdot 100 = 38.9\%$
leggero	3	$3/36 \cdot 100 = 8.3\%$
Totale	36	100

Analizzando la distribuzione condizionata allo stato civile vedovo, possiamo notare che la proporzione di vedovi con livello depressivo grave è superiore rispetto a quella calcolata sul totale. Dal confronto tra distribuzione condizionata e distribuzione marginale si può notare che le frequenze percentuali sono diverse, il che indica la presenza di un'associazione tra i due caratteri. **Non possiamo dire però se questa associazione sia significativa o no.**

2. Il test Chi quadrato ci consente di verificare il seguente test di ipotesi:

   H_0 : il livello depressivo non è associato allo stato civile

   H_A : c'è un'associazione significativa tra livello depressivo e stato civile

   La statistica test

$$X^2 = \sum_i \sum_j \frac{\left(n_{ij} - \frac{n_{i0}n_{0j}}{n}\right)^2}{\frac{n_{i0}n_{0j}}{n}} = n\left(\sum_i \sum_j \frac{n_{ij}^2}{n_{i0}n_{0j}} - 1\right)$$

   si distribuisce, sotto l'ipotesi nulla come una v.a. Chi quadrato con $(r-1)(c-1) = 4$ gradi di libertà (dove r e c indicano rispettivamente il numero di righe e il numero di colonne della tabella di contingenza).

   Calcoliamo innanzi tutto il valore osservato della statistica test Chi-quadrato che esprime una misura della distanza tra frequenze osservate e frequenze teoriche. Si noti che le due formule sono equivalenti, ma la seconda ci consente di abbreviare i calcoli, ottenendo:

$$\chi^2 = 159\left(\frac{22^2}{57\cdot 69} + \frac{16^2}{57\cdot 54} + \frac{19^2}{57\cdot 36} + \frac{33^2}{76\cdot 69} + \right.$$
$$\left. + \frac{29^2}{76\cdot 54} + \frac{14^2}{76\cdot 36} + \frac{14^2}{26\cdot 69} + \frac{9^2}{54\cdot 26} + \frac{3^2}{36\cdot 26} - 1\right) = 6.828$$

   Possiamo ora calcolare il p-value, ovvero la probabilità di osservare un valore della statistica test Chi-quadrato piú estremo di quello effettivamente osservato:

$$P\left(X^2_{(r-1)(c-1)} > \chi^2\right) = P\left(X^2_4 > 6.828\right) = 0.14$$

Poiché il p-value è pari a $0.14 > 0.05 = \alpha$, possiamo concludere che non c'è abbastanza evidenza sperimentale per poter rifiutare l'ipotesi nulla, in altre parole l'associazione tra livello depressivo e stato civile non è statisticamente significativa.

• • •

Esercizio 4.32. *Smettere di fumare*
(dal libro di testo *OpenIntro Statistics* di Diez et al., es. 6.41)

Far parte di un gruppo di sostegno influenza la capacità delle persone di smettere di fumare? Il ministero della salute ha coinvolto 400 fumatori in un esperimento aleatorio. 150 partecipanti hanno usato un cerotto alla nicotina e hanno incontrato settimanalmente un gruppo di sostegno.; gli altri 150 hanno usato il cerotto ma non hanno incontrato il gruppo di sostegno. Alla fine dell'esperimento, 40 partecipanti del primo gruppo ha smesso di fumare mentre solo 30 fumatori del secondo gruppo ha smesso di fumare.

(a) Creare una tabella a doppia entrata riportando i risultati di questo studio.

(b) Rispondere a ciascuna delle seguenti domande sotto l'ipotesi nulla che essere parte di un gruppo di sostegno non influenza la capacità di smettere di fumare, ed indicare se i valori attesi sono più alti o più bassi di quelli osservati.

   i. Quanti soggetti del primo gruppo ti aspetti che smettano di fumare?

   ii. Quanti soggetti del secondo gruppo ti aspetti che smettano di fumare?

• • •

Soluzione.

(a) La tabella a doppia entrata è la seguente

	Smettere di fumare		
Trattamento	Si	No	Totale
Cerotto + gruppo di sostegno	40	110	150
Solo cerotto	30	120	150
Totale	70	230	300

(b-i) $E_{rig1,\ col1} = \dfrac{(totale\ riga\ 1) \times (totale\ colonna\ 1)}{totale\ tabella} = \dfrac{150 \times 70}{300} = 35.$
Questo valore è più piccolo di quello osservato.

(b-ii) $E_{rig2,\ col2} = \dfrac{(totale\ riga\ 2) \times (totale\ colonna\ 2)}{totale\ tabella} = \dfrac{150 \times 230}{300} = 115.$
Questo valore è più piccolo di quello osservato.

• • •

Esercizio 4.33. *Trivellazione in mare aperto, III parte*
(dal libro di testo *OpenIntro Statistics* di Diez et al., es. 6.43)
La tabella sotto sintetizza il dataset analizzato nell'Esercizio **??** dove sono riportate le risposte di un campione aleatorio di laureati e non laureati sul tema della trivellazione. Usare un test Chi quadrato per verificare se c'è una differenza statisticamente significativa tra le risposte dei laureati e quelle dei non laureati.

	Laureati	Non laureati
Favorevoli	154	132
Contrari	180	126
Non sanno	104	131
Totale	438	389

• • •

Soluzione. H_0: L'opinione dei laureati e dei non laureati è differente sul tema della trivellazione per estrarre petrolio e gas naturale al largo delle coste della California. H_A: L'opinione riguardante la trivellazione per estrarre petrolio e gas naturale al largo delle coste della California ha un'associazione con l'essere laureati oppure no.

$$E_{rig1,\ col1} = 151.5 \qquad E_{rig1,\ col2} = 134.5$$

$$E_{rig2,\ col1} = 162.1 \qquad E_{rig2,\ col2} = 143.9$$

$$E_{rig3,\ col1} = 124.5 \qquad E_{rig3,\ col2} = 110.5$$

Indipendenza: i campioni sono entrambi aleatori, non collegati ed estratti da meno del 10% della popolazione, perciò l'ipotesi di indipendenza tra le osservazioni è ragionevole. Campione: tutti le frequenze osservate sono almeno pari a 5. Gradi di libertà: $(R-1)\times(C-1) = (3-1)\times(2-1) = 2$, che è più grande di 1. $X^2 = 11.47$ e p-value compreso tra 0.001 e 0.005. Quindi c'è una forte evidenza empirica sull'associazione tra supportare la trivellazione e l'essere laureati.

• • •

Esercizio 4.34. *Privacy su Facebook*
(dal libro di testo *OpenIntro Statistics* di Diez et al., es. 6.45)

In un'indagine del 2011 806 utenti Facebook adulti, scelti aleatoriamente, sono stati interrogati sulle loro impostazioni sulla privacy di Facebook. Una delle domande era "Sai come cambiare le impostazioni riguardanti privacy di Facebook per controllare le persone che possono e non possono vederti?". Le risposte sono riportate nella seguente tabella divise per genere.

	Genere		
	Maschile	Femminile	Totale
Si	288	378	666
No	61	62	123
Non so	10	7	17
Totale	359	447	806

(a) Definire un test d'ipotesi per verificare l'indipendenza tra genere e la capacità degli utenti di Facebook di modificare le impostazioni sulla privacy.

(b) Verificare tutte le condizioni necessarie per il test e determinare se è possibile utilizzare un test Chi quadrato.

• • •

Soluzione.

(a) H_0: Non c'è relazione tra genere e la capacità degli utenti di Facebook di modificare le impostazioni sulla privacy. H_A: C'è una relazione tra genere e la capacità degli utenti di Facebook di modificare le impostazioni sulla privacy.

(b) I valori attesi sono:

$$E_{rig1,\ col1} = 296.6 \qquad E_{rig1,\ col2} = 369.3$$
$$E_{rig2,\ col1} = 54.8 \qquad E_{rig2,\ col2} = 68.2$$
$$E_{rig3,\ col1} = 7.6 \qquad E_{rig3,\ col2} = 9.4$$

Il campione è aleatorio, tutti i valori attesi sono più grandi di 5 e i gradi di libertà sono pari a $(3-1)\times(2-1) = 2 > 1$, quindi è possibile effettuare il test.

• • •

Capitolo 5

Regressione lineare

5.1 Regressione lineare semplice

Esercizio 5.1.

I dati nella seguente tabella mostrano l'indice di produttivita' X e lo stipendio mensile Y di un campione di dipendenti di un'azienda:

X	Y
1.6	10
2	15
3.5	20
3	21
3.2	24
4	30

1. Calcolare i coefficienti del modello di regressione lineare e calcolare r^2.
2. Stabilire di quanto varia in media il reddito mensile se l'indice di produttività cresce di una unità.

3. Prevedere in base al modello l'ammontare dello stipendio mensile per un'indice di produttività pari a 2.8.

• • •

Soluzione

1. Dato un modello di regressione lineare $y = \beta x + \alpha$, si devono calcolare i due coefficienti β e α usando le note formule:

$$\hat{\beta} = r\frac{S_y}{S_x} \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

dove r è il coefficiente di correlazione, S_x e S_y le due deviazioni standard e $\bar{x}$ e $\bar{y}$ le due medie. Calcoliamo tutto ciò che' è necessario:

$$\bar{x} = \frac{1.6 + 2 + 3.5 + 3 + 3.2 + 4}{6} = 2.88$$

$$\bar{y} = \frac{10 + 15 + 20 + 21 + 24 + 30}{6} = 20$$

$$S_x = \sqrt{S_x^2}$$
$$= \sqrt{\frac{(1.6 - 2.88)^2 + (2 - 2.88)^2 + (3.5 - 2.88)^2 + (3 - 2.88)^2 + (3.2 - 2.88)^2 + (4 - 2.88)^2}{5}}$$
$$= \sqrt{0.83} = 0.913$$

$$S_y = \sqrt{S_y^2}$$
$$= \sqrt{\frac{(10 - 20)^2 + (15 - 20)^2 + (20 - 20)^2 + (21 - 20)^2 + (24 - 20)^2 + (30 - 20)^2}{5}}$$
$$= \sqrt{48.4} = 6.96$$

$$\sum_{i=1}^{n} x_i y_i = 1.6 \cdot 10 + 2 \cdot 15 + 3.5 \cdot 20 + 3 \cdot 21 + 3.2 \cdot 24 + 4 \cdot 30 = 375.8$$

Possiamo cosi' calcolare il valore del coefficiente di correlazione usando la formula semplificata:

$$r = \frac{1}{S_x S_y}\left[\frac{1}{n-1}\sum_{i=1}^{n} x_i y_i - \frac{n}{n-1}\bar{x}\bar{y}\right]$$
$$= \frac{1}{0.913 \cdot 6.96}\left[\frac{1}{5} \cdot 375.8 - \frac{6}{5} \cdot 2.88 \cdot 20\right] = 0.95$$

L'associazione fra le due variabili è positiva ed è molto forte. Il valore di r^2 è $0.95^2 = 0.9$, cioè circa il 90% della variabilità totale della variabile y (il reddito) è spiegato dal modello. Troviamo ora i valori dei coefficienti di regressione:

$$\hat{\beta} = 0.95 \cdot \frac{6.96}{0.913} = 7.24$$
$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 20 - 7.24 \cdot 2.88 = -0.85$$

Il modello di regressione lineare che rappresenta meglio i dati è quindi

$$\hat{y} = \hat{\beta}x + \hat{\alpha} = 7.24 \cdot x - 0.85$$

2. Quando l'indice di produzione cresce di 1, il reddito aumenta in media di $\hat{\beta} = 7.24$.
3. È sufficiente valutare il modello in $x = 2.8$, cioè basta calcolare

$$\hat{y} = 7.24 \cdot 2.8 - 0.85 = 19.422$$

Lo stipendio corrispondente è 19.422.

```
sol: medX = 2.88, medY = 20, sd X = 0.913, sd Y = 6.96,
r = 0.95, r^2 = 0.9, beta = 7.24, alfa = -0.85
```

• • •

Esercizio 5.2.

Consideriamo un campione di 10 esemplari di fiore di Codolina per ciascuno dei quali si misura in cm la lunghezza della foglia superiore (indicata con X)

```
X | 23.4  22  25  18.1  18.9  20.5  19.1  27.5  21.6  15
```

1. Classificare il carattere nelle classi [14,20], (20,23], (23,26], (26,30].

2. Rappresentare adeguatamente la distribuzione del carattere X.

3. Determinare la moda e la classe modale.

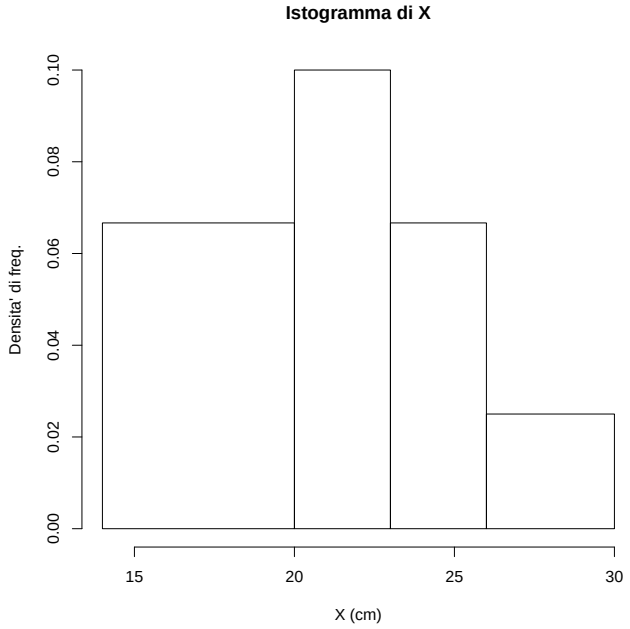
4. Calcolare la media, la varianza, la deviazione standard, la mediana e i quartili.

5. Calcolare il valor medio usando solo la distribuzione in classi.

Soluzione 1)

X	freq.ass.	freq.rel.	amp.class.	dens.freq.
$[14, 20]$	4	0.4	6	0.067
(20,23]	3	0.3	3	0.1
(23,26]	2	0.2	3	0.067
(26,30]	1	0.1	4	0.025
10	1			

2)



3)

La moda è indeterminata. La classe modale è (20,23].

4)

$$\bar{x} = \frac{23.4 + 22 + \cdots + 15}{10} = 21.11$$

$$\begin{aligned} S^2 &= \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{1}{n-1} \sum_{i=1}^n (x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\ &= \frac{\sum_{i=1}^n x_i^2}{n-1} - \frac{2\bar{x}\sum_{i=1}^n x_i}{n-1} + \frac{n}{n-1}\bar{x}^2 \\ &= \frac{1}{n-1}\left[\sum_{i=1}^n x_i^2 - 2n\bar{x}^2 + n\bar{x}^2\right] \\ &= \frac{1}{n-1}\left[\sum_{i=1}^n x_i^2 - n\bar{x}^2\right] \end{aligned}$$

Calcoliamo ora

$$\sum_{i=1}^{n} x_i^2 = 4574.25$$

quindi

$$S^2 = \frac{1}{9}\left[4574.25 - 10(21.11)^2\right] = 13.10322$$

$$S = \sqrt{S^2} = \sqrt{13.10322} = 3.6198\text{cm}$$

Determiniamo ora la mediana. Riordiniamo i dati:

```
X | 15  18.1  18.9  19.1  20.5  21.6  22  23.4  25  27.5
```

$$\text{Mediana} = \frac{20.5 + 21.6}{2} = 21.05$$

$$\text{Primo quartile} = 18.9$$

$$\text{Terzo quartile} = 23.4$$

5)
Calcoliamo prima di tutto i valori centrali:

```
[14,20]  |  17
(20,23]  |  21.5
(23,26]  |  24.5
(26,30]  |  28
```

E poi si può calcolare il valor medio approssimato:

$$\bar{x} = \frac{17 \cdot 4 + 21.5 \cdot 3 + 24.5 \cdot 2 + 28}{10} = 20.95$$

Esercizio 5.3.

I dati nella seguente tabella mostrano l'indice di produttività X e lo stipendio mensile Y di un campione di dipendenti di un'azienda:

X	Y
1.6	10
2	15
3.5	20
3	21
3.2	24
4	30

1. Calcolare i coefficienti del modello di regressione lineare e calcolare l'r^2.
2. Stabilire di quanto varia in media il reddito mensile se l'indice di produttività cresce di una unità.
3. Prevedere in base al modello l'ammontare dello stipendio mensile per un indice di produttività pari a 2.8.

• • •

Soluzione

1. Dato un modello di regressione lineare $y = \beta x + \alpha$, si devono calcolare i due coefficienti β e α usando le note formule:

$$\hat{\beta} = r\frac{S_y}{S_x} \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

dove r è il coefficiente di correlazione, S_x e S_y le due deviazioni standard e $\bar{x}$ e $\bar{y}$ le due medie. Calcoliamo tutto ciò che è necessario:

$$\bar{x} = \frac{1.6 + 2 + 3.5 + 3 + 3.2 + 4}{6} = 2.88$$

$$\bar{y} = \frac{10 + 15 + 20 + 21 + 24 + 30}{6} = 20$$

$$
\begin{aligned}
S_x &= \sqrt{S_x^2} \\
&= \sqrt{\frac{(1.6-2.88)^2+(2-2.88)^2+(3.5-2.88)^2+(3-2.88)^2+(3.2-2.88)^2+(4-2.88)^2}{5}} \\
&= \sqrt{0.83} = 0.913
\end{aligned}
$$

$$
\begin{aligned}
S_y &= \sqrt{S_y^2} \\
&= \sqrt{\frac{(10-20)^2+(15-20)^2+(20-20)^2+(21-20)^2+(24-20)^2+(30-20)^2}{5}} \\
&= \sqrt{48.4} = 6.96
\end{aligned}
$$

$$
\sum_{i=1}^{n} x_i y_i = 1.6 \cdot 10 + 2 \cdot 15 + 3.5 \cdot 20 + 3 \cdot 21 + 3.2 \cdot 24 + 4 \cdot 30 = 375.8
$$

Possiamo così calcolare il valore del coefficiente di correlazione usando la formula semplificata:

$$
\begin{aligned}
r &= \frac{1}{S_x S_y}\left[\frac{1}{n-1}\sum_{i=1}^{n} x_i y_i - \frac{n}{n-1}\bar{x}\bar{y}\right] \\
&= \frac{1}{0.913 \cdot 6.96}\left[\frac{1}{5} \cdot 375.8 - \frac{6}{5} \cdot 2.88 \cdot 20\right] = 0.95
\end{aligned}
$$

L'associazione fra le due variabili è positiva ed è molto forte. Il valore di r^2 è $0.95^2 = 0.9$, cioe circa il 90% della variabilità totale della variabile y (il reddito) è spiegato dal modello. Troviamo ora i valori dei coefficienti di regressione:

$$
\begin{aligned}
\hat{\beta} &= 0.95 \cdot \frac{6.96}{0.913} = 7.24 \\
\hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x} = 20 - 7.24 \cdot 2.88 = -0.85
\end{aligned}
$$

Il modello di regressione lineare che rappresenta meglio i dati è quindi

$$
\hat{y} = \hat{\beta}x + \hat{\alpha} = 7.24 \cdot x - 0.85
$$

2. Quando l'indice di produzione cresce di 1, il reddito aumenta in media di $\hat{\beta} = 7.24$.

3. È sufficiente valutare il modello in $x = 2.8$, cioè basta calcolare

$$\hat{y} = 7.24 \cdot 2.8 - 0.85 = 19.422$$

Lo stipendio corrispondente è 19.422.

```
sol: medX = 2.88, medY = 20, sd X = 0.913, sd Y = 6.96,
r = 0.95, r^2 = 0.9, beta = 7.24, alfa = -0.85
```

• • •

Esercizio 5.4.

Si vuole verificare se il consumo (Y) delle automobili (in litri di carburante per un dato chilometraggio) dipende dal loro peso (X) (in tonnellate). In un campione di n=12 automobili sono stati ottenuti i seguenti risultati:

$$\sum_{i=1}^{n} x_i = 24.62, \qquad \sum_{i=1}^{n} y_i = 279.76, \qquad \sum_{i=1}^{n} x_i^2 = 51.49,$$

$$\sum_{i=1}^{n} y_i^2 = 6525.47, \qquad \sum_{i=1}^{n} x_i y_i = 575.53$$

a. Determinare la retta di regressione.

b. Spiegare il significato del valore assunto dal coefficiente b.

c. Calcolare l'indice di determinazione e commentare il risultato.

• • •

Soluzione

a. La numerosità è $n = 12$ e le medie risultano quindi

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{24.62}{12} = 2.05 \qquad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{279.76}{12} = 23.31$$

Per determinare la retta di regressione, ovvero i coefficienti

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

e

$$a = \bar{y} - b\bar{x}$$

sfruttando i dati a disposizione, è necessario calcolare:

$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} = 575.53 - 12 \cdot 2.05 \cdot 23.31 = 2.10$

$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = 51.49 - 12 \cdot 2.05^2 = 1.06$

$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = 6525.47 - 12 \cdot 23.31^2 = 5.20.$

Otteniamo quindi:

$$b = \frac{2.10}{1.06} = 1.98$$

e

$$a = 23.31 - 1.98 \cdot 2.05 = 19.25$$

Quindi la retta di regressione è:

$$\widehat{y}_i = 19.25 + 1.98 x_i$$

b. Il valore del coefficiente $b = 1.98$ si può interpretare come il consumo medio di carburante a fronte di un aumento di peso dell'auto di 1 tonnellata.

c. L'indice di determinazione è pari

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

e si può calcolare anche elevando al quadrato il coefficiente di correlazione, ovvero:

$$R^2 = r^2 = \frac{\left(\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})\right)^2}{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot \sum_{i=1}^n (y_i - \bar{y})^2} = \frac{2.10^2}{1.06 \cdot 5.20} = 0.8.$$

Poiché il valore dell'indice R^2 è risultato pari a 0.8 possiamo concludere che l' 80% della variabilità del consumo delle automobili è spiegata tramite la relazione lineare con il peso delle automobili.

• • •

Esercizio 5.5.

In uno studio sulle cause dell'inquinamento sono stati rilevati, in 41 città americane, la concentrazione di anidride solforosa (microgrammi per metro cubo) e il numero di aziende manifatturiere con oltre 20 addetti. Indicando con x_i il numero di aziende e con y_i le osservazioni sulla concentrazione di anidride solforosa, sono stati ottenuti i valori seguenti:

$$\sum_{i=1}^{41} x_i = 18987,\ \sum_{i=1}^{41} y_i = 1232,\ \sum_{i=1}^{41} x_i^2 = 21492949,\ \sum_{i=1}^{41} y_i^2 = 59058,$$
$$\sum_{i=1}^{41} x_i y_i = 911645,\ s^2 = 327.7$$

1. Stimare un modello di regressione che spieghi la concentrazione di anidride solforosa in funzione del numero di aziende manifatturiere;

2. Sottoporre a test l'ipotesi nulla $\beta = 0$ verso l?ipotesi alternativa $\beta > 0$ al livello di significatività del 5% e commentare il risultato.

Soluzione

Vogliamo stimare la retta di regressione

$$Y_i = \alpha + \beta x_i + \epsilon_i$$

1. La numerosità campionaria è $n = 41$, la media del numero di aziende manifatturiere è $\bar{x} = \frac{18987}{41} = 463.1$ e il valore osservato della media campionaria della concentrazione di anidride solforosa è $\bar{y} = \frac{1232}{41} = 30.1$
   Aplicando le formule ridotte si ottiene la stima del coefficiente angolare

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{911645 - 41 \cdot 463.1 \cdot 30.1}{21492949 - 41 \cdot 463.1^1} = 0.027$$

   e l'intercetta risulta

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 17.6$$

2. Le ipotesi da sottoporre a test sono

$$H_0 : \beta = 0 \quad H_1 : \beta > 0$$

   La statistica test è

$$T = \frac{B}{\sqrt{\hat{\sigma}^2 / \sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

   e ha una distribuzione t con $n - 2 = 39$ gradi di libertà. Le tavole della t di Student non riportano i percentili in corrispondenza di 39 gradi di libertà, ma questi ultimi si possono approssimare con quelli di una variabile casuale t_{40}.
   Il valore empirico della statistica test è:

$$t = \frac{\hat{\beta}}{SE(\hat{\beta})} = \frac{0.027}{0.005} = 5.4$$

   dove

$$SE(\hat{\beta}) = \sqrt{s^2 / \sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{s^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}} = \sqrt{\frac{327.7}{21492949 - 41 \cdot 463.1^2}} = 0.005$$

   Il valore di t deve essere quindi confrontato con $t^* = 1.684$.
   Poiché $t > t^*$, possiamo rifiutare l'ipotesi nulla: infatti, il p-value

$$Pr(T > t) < 0.0005$$

5.2 Inference for linear regression

In the following exercises, visually check the conditions for fitting a least squares regression line, but you do not need to report these conditions in your solutions.

Esercizio 5.6. *Beer and blood alcohol content.*

Many people believe that gender, weight, drinking habits, and many other factors are much more important in predicting blood alcohol content (BAC) than simply considering the number of drinks a person consumed. Here we examine data from sixteen student volunteers at Ohio State University who each drank a randomly assigned number of cans of beer. These students were evenly divided between men and women, and they differed in weight and drinking habits. Thirty minutes later, a police officer measured their blood alcohol content (BAC) in grams of alcohol per deciliter of blood.23 The scatterplot and regression table summarize the findings.

	Stima	SE	t-value	p-value
Intercetta	-0.0127	0.0126	-1.00	0.332
Birra	0.0180	0.0024	7.48	0.0000

(a) Describe the relationship between the number of cans of beer and BAC.

(b) Write the equation of the regression line. Interpret the slope and intercept in context.

(c) Do the data provide strong evidence that drinking more cans of beer is associated with an increase in blood alcohol? State the null and alternative hypotheses, report the p-value, and state your conclusion.

(d) The correlation coefficient for number of cans of beer and BAC is 0.89. Calculate R2 and interpret it in context.

(e) Suppose we visit a bar, ask people how many drinks they have had, and also take their BAC. Do you think the relationship between number of

drinks and BAC would be as strong as the relationship found in the Ohio State study?

Esercizio 5.7. *Misure corporee, IV parte.*

The scatterplot and least squares summary below show the relationship between weight measured in kilograms and height measured in centimeters of 507 physically active individuals.

	Stima	SE	t-value	p-value
Intercetta	-105.011	7.539	-13.93	0.000
Altezza	1.0176	0.0440	23.13	0.0000

(a) Describe the relationship between height and weight.

(b) Write the equation of the regression line. Interpret the slope and intercept in context.

(c) Do the data provide strong evidence that an increase in height is associated with an increase in weight? State the null and alternative hypotheses, report the p-value, and state your conclusion.

(d) The correlation coefficient for height and weight is 0.72. Calculate R^2 and interpret it in context.

Esercizio 5.8. *Husbands and wives, Part II.*

L'Esercizio 5.26 presents a scatterplot displaying the relationship between husbands' and wives' ages in a random sample of 170 married couples in Britain, where both partners' ages are below 65 years. Given below is summary output of the least squares fit for predicting wife's age from husband's age.

(a) We might wonder, is the age difference between husbands and wives constant over time? If this were the case, then the slope parameter

	Stima	SE	t-value	p-value
Intercetta	1.5740	1.1501	1.37	0.173
eta-marito	0.9112	0.0259	35.25	0.000

would be $\beta_1 = 1$. Use the information above to evaluate if there is strong evidence that the difference in husband and wife ages actually has changed.

(b) Write the equation of the regression line for predicting wife's age from husband's age.

(c) Interpret the slope and intercept in context.

(d) Given that $R^2 = 0.88$, what is the correlation of ages in this data set?

(e) You meet a married man from Britain who is 55 years old. What would you predict his wife's age to be? How reliable is this prediction?

(f) You meet another married man from Britain who is 85 years old. Would it be wise to use the same linear model to predict his wife's age? Explain

Esercizio 5.9. *Husbands and wives, Part III*
. The scatterplot below summarizes husbands' and wives' heights in a random sample of 170 married couples in Britain, where both partners' ages are below 65 years. Summary output of the least squares fit for predicting wife's height from husband's height is also provided in the table.

	Stima	SE	t-value	p-value
Intercetta	43.575	4.6842	9.30	0.000
altezza-marito	0.2863	0.0686	4.17	0.000

(a) Is there strong evidence that taller men marry taller women? State the hypotheses and include any information used to conduct the test.

(b) Write the equation of the regression line for predicting wife's height from husband's height.

(c) Interpret the slope and intercept in the context of the application.

(d) Given that R2 = 0.09, what is the correlation of heights in this data set?

(e) You meet a married man from Britain who is 5'9" (69 inches). What would you predict his wife's height to be? How reliable is this prediction?

(f) You meet another married man from Britain who is 6'7" (79 inches). Would it be wise to use the same linear model to predict his wife's height? Why or why not?

Esercizio 5.10. *Urban homeowners, Part II.*
L'Esercizio 5.47 gives a scatterplot displaying the relationship between the percent of families that own their home and the percent of the population living in urban areas. Below is a similar scatterplot, excluding District of Columbia, as well as the residuals plot. There were 51 cases.

(a) For these data, $R^2 = 0.28$. What is the correlation? How can you tell if it is positive or negative?

(b) Examine the residual plot. What do you observe? Is a simple least squares fit appropriate for these data?

Esercizio 5.11. *Babies.*
Is the gestational age (time between conception and birth) of a low birth-weight baby useful in predicting head circumference at birth? Twenty-five low birth-weight babies were studied at a Harvard teaching hospital; the investigators calculated the regression of head circum- ference (measured in centimeters) against gestational age (measured in weeks). The estimated regression line is

$$\text{circonferenza del cranio} = 3.91 + 0.78 \times \text{età di gestazione}$$

(a) What is the predicted head circumference for a baby whose gestational age is 28 weeks?

(b) The standard error for the coefficient of gestational age is 0.35, which is associated with df = 23. Does the model provide strong evidence that gestational age is significantly associated with head circumference?

Esercizio 5.12. *Rate my professor.*

Some college students critique professors' teaching at RateMyProfessors.com, a web page where students anonymously rate their professors on quality, easiness, and attractiveness. Using the self-selected data from this public forum, researchers examine the relations between quality, easiness, and attractiveness for professors at various universities. In this exercise we will work with a portion of these data that the researchers made publicly available[1].

The scatterplot on the right shows the relationship between teaching evaluation score (higher score means better) and standardized beauty score (a score of 0 means average, negative score means below average, and a positive score means above average) for a sample of 463 professors. Given below are associated diagnostic plots. Also given is a regression output for pre- dicting teaching evaluation score from beauty score.

	Stima	SE	t-value	p-value
Intercetta	4.010	0.0255	157.21	0.000
	bellezza	0.0322	4.13	0.000

(a) Given that the average standardized beauty score is -0.0883 and average teaching evaluation score is 3.9983, calculate the slope. Alternatively, the slope may be computed using just the information provided in the model summary table.

[1] J. Felton et al. "Web-based student evaluations of professors: the relations between perceived quality, easiness and sexiness". In: Assessment and Evaluation in Higher Education 29.1 (2004), pp. 91–108.

(b) Do these data provide convincing evidence that the slope of the relationship between teaching evaluation and beauty is positive? Explain your reasoning.

(c) List the conditions required for linear regression and check if each one is satisfied for this model.

5.3 Correlazione e Regressione

Esercizio 5.13. *Correlazione*
I dati relativi alle lunghezze di femore e omero di 5 reperti fossili sono riportati nella seguente tabella (valori espressi in cm):

Femore	Omero
38	41
56	63
59	70
64	72
74	84

- Si realizzi un grafico a dispersione;
- Si calcoli il coefficiente di correlazione e si commenti il risultato ottenuto.

Soluzione
Le lunghezze di femore e omero sono rappresentate tramite grafico a dispersione (figura 5.1).

Una nota sul calcolo del coefficiente di correlazione
Supponiamo di aver osservato le variabili x e y su un sottoinsieme di n unità dalla popolazione di riferimento. I valori per la prima unità

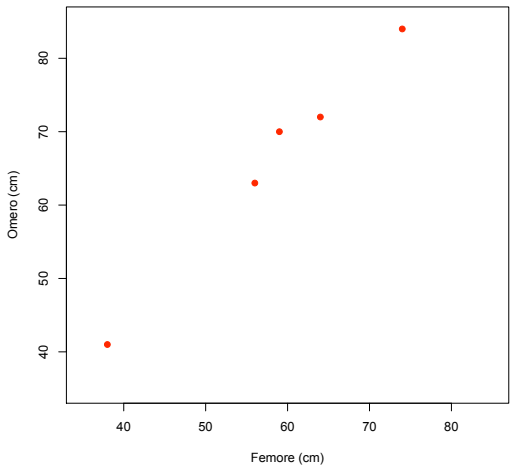


Figura 5.1: Grafico a dispersione

sono x_1 e y_1, i valori per la seconda unità sono x_2 e y_2 e così via. Le medie e le deviazioni standard delle due variabili sono

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{e} \qquad s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} x_i^2 - \bar{x}^2}$$

per i valori x, e

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad \text{e} \qquad s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} y_i^2 - \bar{y}^2}$$

per i valori y. Il coefficiente di correlazione r fra x e y è dato da

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) = \frac{1}{s_x\, s_y}\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Notiamo che

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

allora

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \frac{1}{n-1}\sum_{i=1}^{n}x_iy_i - \frac{n}{n-1}\bar{x}\bar{y}.$$

In analogia a quanto già visto per la varianza, quando non si dispone di un supporto informatico, anche per il calcolo del coefficiente di correlazione è utile ricorrere alla seguente formula alternativa:

$$r = \frac{1}{s_x\ s_y}\left[\frac{1}{n-1}\sum_{i=1}^{n}x_iy_i - \frac{n}{n-1}\bar{x}\bar{y}\right]$$

Quest'ultima formula è più conveniente in termini di economia di calcolo, è del tutto equivalente alla formula che definisce il coefficiente di correlazione (si ottiene da questa tramite alcuni passaggi algebrici) e dà lo stesso risultato numerico (a meno di approssimazioni dovute ad arrotondamenti nei calcoli intermedi).

Calcoliamo tale coefficiente come segue:

	Femore (x_i)	Omero (y_i)	x_iy_i	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$
	38	41	$38 \cdot 41 = 1558$	408.04	625
	56	63	$56 \cdot 63 = 3528$	4.84	9
	59	70	$59 \cdot 70 = 4130$	0.64	16
	64	72	$64 \cdot 72 = 4608$	33.64	36
	74	84	$74 \cdot 84 = 6216$	249.64	324
Tot	291	330	20040	696.8	1010

Da cui ricaviamo le seguenti quantitÃ :

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i = \frac{1}{5}291 = 58.2$

$\bar{y} = \frac{1}{n}\sum_{i=1}^{n}y_i = \frac{1}{5}330 = 66$

$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2} = \sqrt{174.2} = 13.20$

$s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2} = \sqrt{252.5} = 15.88$

Pertanto il coefficiente di correlazione Ã¨ pari a

$$r = \frac{1}{13.20 \cdot 15.88}\left[14 \cdot 20040 - \frac{5}{4}58.2 \cdot 66\right] = 0.99$$

Ciò significa tra lunghezza del femore e dell'omero esiste una forte associazione lineare positiva.

Esercizio 5.14.

Si consideri la seguente variabile X:

X: 1 2 3 4 5 -1 -2 -3 -4 -5;

- si costruisca la variabile $Y = X^2$ e se ne calcoli media e varianza;
- si costruisca il diagramma a dispersione tra le variabili X e Y;
- si calcoli il coefficiente di correlazione tra le 2 variabili e si commenti il risultato.

Soluzione

La variabile Y assume i seguenti valori:

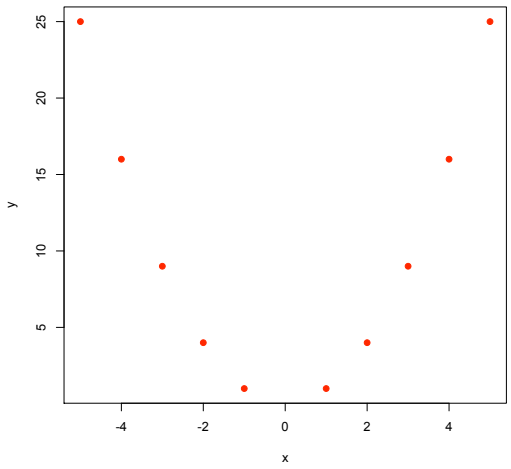
Y: 1 4 9 16 25 1 4 9 16 25;

La media e la varianza di Y sono pari rispettivamente a :

- $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 11$
- $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = 83.11$

Il diagramma a dispersione per le variabili X e Y Ã¨

Calcoliamo quindi il coefficiente di correlazione tra X e Y:



	x_i	y_i	$x_i y_i$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
	1	1	$1 \cdot 1 = 1$	1	100
	2	4	$2 \cdot 4 = 8$	4	49
	3	9	$3 \cdot 9 = 27$	9	4
	4	16	$4 \cdot 16 = 64$	16	25
	5	25	$5 \cdot 25 = 125$	25	196
	-1	1	$-1 \cdot 1 = -1$	1	100
	-2	4	$-2 \cdot 4 = -8$	4	49
	-3	9	$-3 \cdot 9 = -27$	9	4
	-4	16	$-4 \cdot 16 = -64$	16	25
	-5	25	$-5 \cdot 25 = 125$	25	196
Tot	0	110	0	110	748

$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = 0$

$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i = 11$

$s_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} = 3.50$

$s_y = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2} = 9.12$

Pertanto il coefficiente di correlazione è pari a

$$r = \frac{1}{3.50 \cdot 9.12}\left[\frac{1}{9}0 - \frac{10}{9}0 \cdot 11\right] = 0$$

Il coefficiente di correlazione è pari a 0: ciò significa che non c'Ã¨ relazione lineare tra le 2 variabili.

Esercizio 5.15.
Si trovi l'errore contenuto in ognuna delle seguenti affermazioni:

- C'Ã¨ una forte correlazione tra il sesso dei lavoratori americani e il loro reddito;
- E' stata trovata un'alta correlazione($r = 1.09$) fra i voti che gli studenti ottengono all'esame di statistica e i voti presi all'esame di matematica;
- La correlazione fra l'altezza e il peso calcolata su 50 studenti corrisponde a $r = 0.25$Kg.

Soluzione

- Errore: il sesso Ã¨ un carattere qualitativo sconnesso per il quale il coefficiente di correlazione non Ã¨ calcolabile;
- Errore: la correlazione assume valori tra -1 e 1. Pertanto un valore $r = 1.09$ non Ã¨ accettabile;
- Errore: l'indice di correlazione Ã¨ un numero puro, ossia non dipende dall'unitÃ di misura. Pertanto $r = 0.25$ Kg non Ã¨ un valore accettabile.

Esercizio 5.16.

I dati seguenti mostrano i quozienti intellettivi (QI) di 10 madri e figlie primogenite.

x=QI madre	y= QI figlia
135	121
127	131
124	112
120	115
115	99
112	118
104	106
96	89
94	92
85	90

1. Disegnare il diagramma a dispersione;

2. Calcolare il coefficiente di correlazione;

3. Calcolare la retta di regressione dei minimi quadrati e rappresentarla sul grafico a dispersione.

• • •

Soluzione

1. Il diagramma a dispersione del QI delle figlie rispetto al QI delle madri è rappresentato in figura 1.

2. Calcoliamo nella seguente tabella gli elementi necessari per il calcolo del coefficiente di correlazione r:

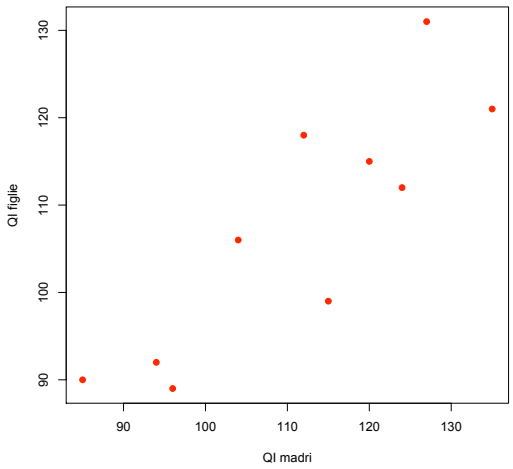


Figura 5.2: Diagramma a dispersione per le variabili x=QI delle madri e y=QI delle figlie.

	x	y	x_i^2	y_i^2	$x_i y_i$
	135	121	18225	14641	16335
	127	131	16129	17161	16637
	124	112	15376	12544	13888
	120	115	14400	13225	13800
	115	99	13225	9801	11385
	112	118	12544	13924	13216
	104	106	10816	11236	11024
	96	89	9216	7921	8544
	94	92	8836	8464	8648
	85	90	7225	8100	7650
Tot	1112	1073	125992	117017	121127

e deriviamo le quantità necessarie per il calcolo di r:

- $\bar{x} = \frac{1}{10}1112 = 111.2$
- $\bar{y} = \frac{1}{10}1073 = 107.3$

- $s_x = \sqrt{\frac{1}{9}125992 - \frac{10}{9}(111.2^2)} = 16.1162$
- $s_y = \sqrt{\frac{1}{9}117017 - \frac{10}{9}(107.3^2)} = 14.4687$

da cui segue che:

$$r = \frac{1}{16.1162 \cdot 14.4687}\left[\frac{1}{9}121127 - \frac{10}{9}111.2 \cdot 107.3\right] = 0.8621$$

3. La retta di regressione $y = a + b \cdot x$ dove $b = r\frac{s_y}{s_x} = 0.8621 \cdot \frac{14.4687}{16.1162} = 0.774$ e $a = \bar{y} - b\bar{x} = 107.3 - 0.774 \cdot 111.2 = 21.23$.
   Ciò significa che al crescere di 1 unità del quoziente intellettivo della madre, quello delle figlie aumenta di 0.774. Rappresentiamo la retta di regressione sul grafico 4

4. correlazione spuria

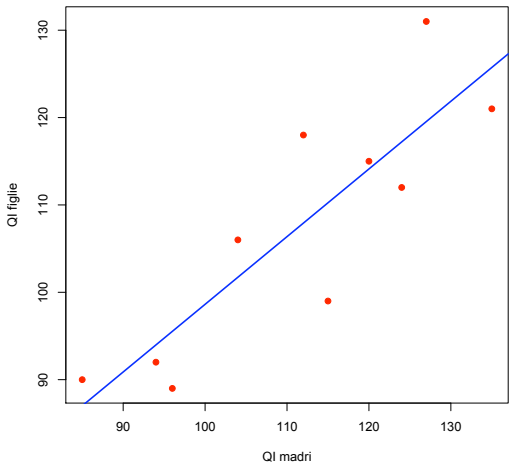


Figura 5.3: Diagramma a dispersione per le variabili x=QI delle madri e y=QI delle figlie: la retta blu corrisponde alla retta di regressione ottenuta mediante il metodo dei minimi quadrati.

• • •

Esercizio 5.17.

I dolcificanti causano un aumento di peso? Le persone che utilizzano dolcificanti al posto dello zucchero tendono ad essere piú grasse rispetto a quelle che usano lo zucchero. Dare una spiegazione plausibile per questa associazione.

• • •

Soluzione

È un esempio di **correlazione spuria**: non esiste un legame causale tra il peso e l'uso di dolcificante, ma esiste una variabile nascosta influente sia sul peso che sull'uso di dolcificante, che induce un'alta correlazione tra di esse. È lecito, infatti, pensare che il consumo di dolcificante sia suggerito a persone sovrappeso o che seguono una dieta alimentare.

• • •

Esercizio 5.18.

Un recente studio ha rilevato una forte correlazione positiva tra il livello di colesterolo dei giovani adulti e il tempo speso a guardare la televisione.

1. Ti saresti aspettato questo risultato? Perché?
2. Ritieni che guardare la tv causi un aumento del livello di colesterolo?

• • •

Soluzione

1. Il risultato può essere giustificato alla luce del fatto che chi spende molto tempo a guardare la televisione tende a non fare molta attività fisica con un conseguente incremento della massa grassa corporea e del colesterolo.
2. L'affermazione non è esatta in quanto la correlazione **non** dimostra causalità.

• • •

5.4 Analisi dei residui

Esercizio 5.19.

Si ritiene che più alcool c'è in circolo, più lento sia il tempo di reazione di una persona. Per verificare questa affermazione, 7 volontari assumono ciascuno una diversa quantità di alcool. La concentrazione di alcool nel sangue viene determinata come percentuale del peso corporeo. In seguito viene misurato il tempo di reazione di ciascuno a un certo stimolo, ottenendo i seguenti dati.

x=concentrazione di alcool nel sangue (%)	y= tempo di reazione (secondi)
0.08	0.32
0.10	0.38
0.12	0.44
0.14	0.42
0.15	0.47
0.16	0.70
0.18	0.63

a. Disegnare il grafico a dispersione dei dati.

b. Disegnare la retta di regressione.

c. Usare la retta di regressione per predire il tempo di reazione di un individuo con una concentrazione di alcool nel sangue di x=0.15.

d. Disegnare il grafico dei residui. Cosa ci dice?

e. Determinare l'indice di determinazione. Come lo interpretiamo?

• • •

Soluzione

a. Il diagramma a dispersione per le variabili x=concentrazione di alcool nel sangue e y=tempo di reazione rappresentato nella figura 5.4.

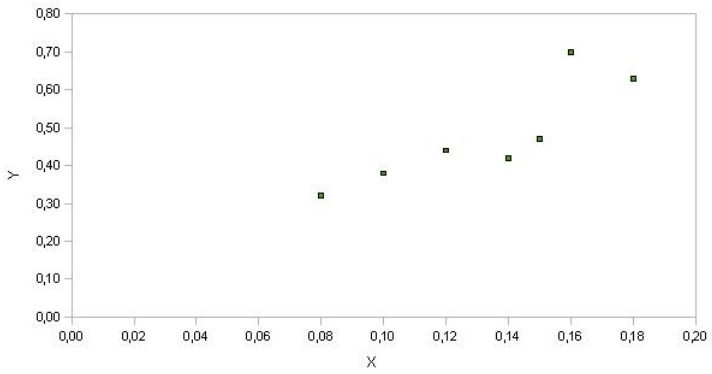


Figura 5.4: Diagramma a dispersione per le variabili x=concentrazione di alcool nel sangue e y=tempo di reazione.

b. Dobbiamo calcolare la retta di regressione $\hat{y} = a + bx$ dove a e b sono:

$$b = r\frac{s_y}{s_x}$$
$$a = \bar{y} - b\bar{x}$$

Per il calcolo quindi di a e b abbiamo bisogno delle seguenti quantità:

- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- $s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} x_i^2 - \frac{n}{n-1}\bar{x}^2}$
- $s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} y_i^2 - \frac{n}{n-1}\bar{y}^2}$
- $r = \frac{1}{s_x s_y}\left[\frac{1}{n-1}\sum_{i=1}^{n} x_i y_i - \frac{n}{n-1}\bar{x}\bar{y}\right]$

Calcoliamo:

	x	y	x_i^2	y_i^2	$x_i y_i$
	0.08	0.32	0.0064	0.1024	0.0256
	0.10	0.38	0.0100	0.1444	0.0380
	0.12	0.44	0.0144	0.1936	0.0528
	0.14	0.42	0.0196	0.1764	0.0588
	0.15	0.47	0.0225	0.2209	0.0705
	0.16	0.70	0.0256	0.4900	0.1120
	0.18	0.63	0.0324	0.3969	0.1134
Tot	0.93	3.36	0.1309	1.7246	0.4711

da cui si ha che:

- $\bar{x} = \frac{1}{7}0.93 = 0.1329$
- $\bar{y} = \frac{1}{7}3.36 = 0.48$
- $s_x = \sqrt{\frac{1}{6}0.1309 - \frac{7}{6}(0.1329^2)} = 0.0348$
- $s_y = \sqrt{\frac{1}{6}1.7246 - \frac{7}{6}(0.48^2)} = 0.14$
- $r = \frac{1}{0.0348 \cdot 0.14}\left[\frac{1}{6}0.4711 - \frac{7}{6}0.1329 \cdot 0.48\right] = 0.87$

e quindi $b = 0.87 \cdot \frac{0.14}{0.0348} = 3.4$ e $a = 0.48 - 3.4 \cdot 0.1329 = 0.03$.

La retta di regressione è pertanto

$$\hat{y} = 0.03 + 3.4x$$

e possiamo rappresentarla sul diagramma a dispersione nella figura 5.4. Tale retta ci dice che al crescere di 1 unità percentuale di alcool, il tempo di reazione cresce in media di 3.4 secondi.

c. La retta di regressione può essere utilizzata anche per fare previsioni: il valore previsto dalla retta per $x = 0.15$ è $\hat{y}_{x=0.15} = 0.03 + 3.4 \cdot 0.15 = 0.5383$.

d. I residui si possono ottenere calcolando le differenze tra valori osservati e valori predetti dal modello:

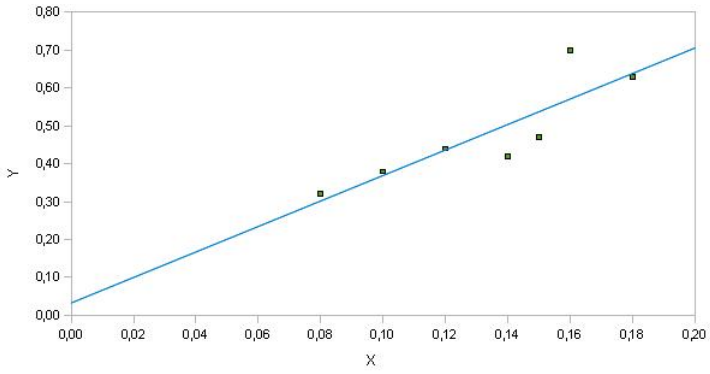
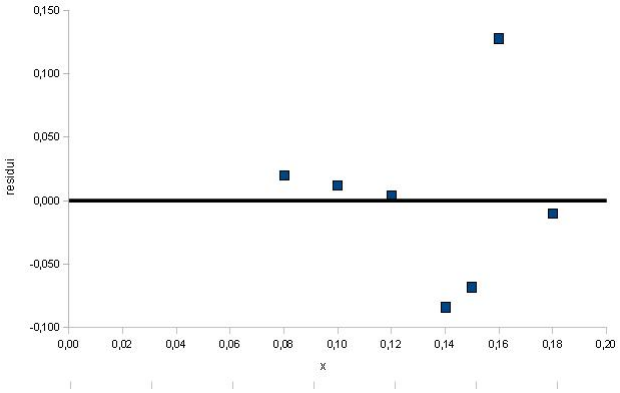


Figura 5.5: Diagramma a dispersione per le variabili x=concentrazione di alcool nel sangue e y=tempo di reazione: la retta blu è la retta di regressione

i	x_i	y_i	$\hat{y}_i$	$y_i - \hat{y}_i$
1	0.08	0.32	0.3003	0.020
2	0.10	0.38	0.3683	0.012
3	0.12	0.44	0.4363	0.004
4	0.14	0.42	0.5043	-0.084
5	0.15	0.47	0.5383	-0.068
6	0.16	0.70	0.5723	0.128
7	0.18	0.63	0.6403	-0.010

Il grafico dei residui rispetto ai valori della variabile esplicativa, è il seguente



Si può osservare che il residuo corrispondente all'osservazione $x = 0.16$ è un valore anomalo che rappresenta una deviazione dal modello.

e. L'indice $R^2 = r^2$ misura quanta parte della variabilità totale di y spiegata dalla x: nel nostro esempio, $r^2 = 0.87^2 = 0.75$, quindi circa il 75% della variabilità del tempo di reazione è spiegata dalla concentrazione di alcool.

• • •

Esercizio 5.20. *Visualize the residuals.*

The scatterplots shown below each have a superimposed regression line. If we were to construct a residual plot (residuals versus x) for each, describe what those plots would look like.

Esercizio 5.21. *Trends in the residuals.*

Shown below are two plots of residuals remaining after fitting a linear model to two different sets of data. Describe important features and determine if a linear model would be appropriate for these data. Explain your reasoning.

Esercizio 5.22. *Identify relationships, Part I.*

For each of the six plots, identify the strength of the relationship (e.g. weak, moderate, or strong) in the data and whether fitting a linear model would be reasonable.

Esercizio 5.23. *Identify relationships, Part II.*

For each of the six plots, identify the strength of the relationship (e.g. weak, moderate, or strong) in the data and whether fitting a linear model would be reasonable.

Esercizio 5.24. *Scatterplots.*

The two scatterplots below show the relationship between final and mid-semester exam grades recorded during several years for a Statistics course at a university.

(a) Based on these graphs, which of the two exams has the strongest correlation with the final exam grade? Explain.

(b) Can you think of a reason why the correlation between the exam you chose in part (a) and the final exam is higher?

Esercizio 5.25.

Volendo costruire un modello che spieghi il Peso (espressa in funzione dell'Altezza (espressa in cm) si è osservato u n = 10 studenti della facoltà di Economia; i dati riportati nella tabella seguente:

Altezza	Peso
165	71
172	75
159	81
168	76
166	88
158	72
157	98
177	89
164	83
172	81

Sia la variabile Altezza la variabile esplicativa X e la variabile Peso la variabile dipendente Y.

1. Stimare la retta di regressione;

2. costruire un intervallo di confidenza per il coeffiente angolare a livello di significatività $\alpha = 0.05$;

3. sulla base delle osservazioni campionarie verificare l'ipotesi nulla di assenza di legame lineare tra le due variabili.

Soluzione

1. Stimiamo i parametri della retta di regressione

$$y = \hat{\alpha} + \hat{\beta}x$$

mediante il metodo dei minimi quadrati:

$$\hat{\beta} = \frac{\sum_{i=1}^{10}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{10}(x_i - \bar{x})^2} = -0.0813$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 94.8952$$

dove $\bar{x} = 165.8$ e $\bar{y} = 81.4$

2. L'intervallo di confidenza per il parametro β è

$$[\hat{\beta} - t^* SE(\hat{\beta}); \hat{\beta} + t^* SE(\hat{\beta})]$$

dove

$$SE[\beta] = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n (x_i - x)^2}} = 0.4510$$

e

$$\hat{\sigma}^2 = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \bar{y})^2 - \hat{\beta}^2 (x_i - \bar{x})^2 = 8.97064$$

e $t^* = 2.306$ (quantile a livello 0.025 di una distribuzione T con $n - 2 = 8$ gradi di libertà).
Pertanto l'intervallo di confidenza è pari a

$$[-1.12144, 0.95865]$$

3. Dobbiamo valutare il seguente sistema di ipotesi:

$$H_0 : \beta = 0 \qquad\qquad H_1 : \beta \neq 0$$

Poiché il valore 0 cade all'interno dell'intervallo di confidenza, non si ha abbastanza evidenza sperimentale per rifiutare l'ipotesi nulla.

Esercizio 5.26. *Husbands and wives, Part I.*
The Great Britain Office of Population Census and Surveys once collected data on a random sample of 170 married couples in Britain, recording the age (in years) and heights (converted here to inches) of the husbands and wives. The scatterplot on the left shows the wife's age plotted against her husband's age, and the plot on the right shows wife's height plotted against husband's height.

(a) Describe the relationship between husbands' and wives' ages.

(b) Describe the relationship between husbands' and wives' heights.

(c) Which plot shows a stronger correlation? Explain your reasoning.

(d) Data on heights were originally collected in centimeters, and then converted to inches. Does this conversion affect the correlation between husbands' and wives' heights?

Esercizio 5.27. *Correlazione, Part I.*

Match the calculated correlations to the corresponding scatterplot.

R = -0.7

R = 0.45

R = 0.06

R = 0.92

Esercizio 5.28. *Correlazione, Part II.*

Match the calculated correlations to the corresponding scatterplot.

R = 0.49

R = -0.48

R = -0.03

R = -0.85

Esercizio 5.29. *Speed and height.*

1302 UCLA students were asked to fill out a survey where they were asked about their height, fastest speed they have ever driven, and gender. The scatterplot on the left displays the relationship between height and fastest speed, and the scatterplot on the right displays the breakdown by gender in this relationship.

(a) Describe the relationship between height and fastest speed.

(b) Why do you think these variables are positively associated?

(c) What role does gender play in the relationship between height and fastest driving speed?

Esercizio 5.30. *Trees.*

The scatterplots below show the relationship between height, diameter, and volume of timber in 31 felled black cherry trees. The diameter of the tree is measured 4.5 feet above the ground.

(a) Describe the relationship between volume and height of these trees.

(b) Describe the relationship between volume and diameter of these trees.

(c) Suppose you have height and diameter measurements for another black cherry tree. Which of these variables would be preferable to use to predict

the volume of timber in this tree using a simple linear regression model? Explain your reasoning.

Esercizio 5.31. *Un treno costiero*
Part I. The Coast Starlight Amtrak train runs from Seattle to Los Angeles. The scatterplot below displays the distance between each stop (in miles) and the amount of time it takes to travel from one stop to another (in minutes).

(a) Describe the relationship between distance and travel time.

(b) How would the relationship change if travel time was instead measured in hours, and distance was instead measured in kilometers?

(c) Correlation between travel time (in miles) and distance (in minutes) is R = 0.636. What is the correlation between travel time (in kilometers) and distance (in hours)?

Esercizio 5.32. *Crawling babies, Part I.*
A study conducted at the University of Denver investigated whether babies take longer to learn to crawl in cold months, when they are often bundled in clothes that restrict their movement, than in warmer months.18 Infants born during the study year were split into twelve groups, one for each birth month. We consider the average crawling age of babies in each group against the average temperature when the babies are six months old (that is when babies often begin trying to crawl). Temperature is measured in degrees Fahrenheit (F) and age is measured in weeks.

(a) Describe the relationship between temperature and crawling age.

(b) How would the relationship change if temperature was measured in de- grees Celsius (C) and age was measured in months?

(c) The correlation between temperature in F and age in weeks was R = -0.70. If we converted the temperature to C and age to months, what would the correlation be?

Esercizio 5.33. *Misure corporee*
Part I. Researchers studying anthropometry collected body girth measurements

and skeletal diameter measurements, as well as age, weight, height and gender for 507 physically active individuals.19 The scatterplot below shows the relationship between height and shoulder girth (over deltoid muscles), both measured in centimeters.

(a) Describe the relationship between shoulder girth and height.

(b) How would the relationship change if shoulder girth was measured in inches while the units of height remained in centimeters?

Esercizio 5.34. *Misure corporee, II parte.*
The scatterplot below shows the relationship between weight measured in kilograms and hip girth measured in centimeters from the data described nell'Esercizio 5.33.

(a) Describe the relationship between hip girth and weight.

(b) How would the relationship change if weight was measured in pounds while the units for hip girth remained in centimeters?

Esercizio 5.35. *Correlation, Part I.*
What would be the correlation between the ages of husbands and wives if men always married woman who were

(a) 3 years younger than themselves?

(b) 2 years older than themselves?

(c) half as old as themselves?

Esercizio 5.36. *Correlation, Part II.*
What would be the correlation between the annual salaries of males and females at a company if for a certain type of position men always made

(a) $5,000$ USD more than women?

(b) 25% more than women?

(c) 15% less than women?

Esercizio 5.37. *Tourism spending.*
The Association of Turkish Travel Agencies reports the number of foreign tourists visiting Turkey and tourist spending by year.20 The scatterplot below shows the relationship between these two variables along with the least squares fit.

(a) Describe the relationship between number of tourists and spending.

(b) What are the explanatory and response variables?

(c) Why might we want to fit a regression line to these data?

(d) Do the data meet the conditions required for fitting a least squares line?

(e) In addition to the scatterplot, use the residual plot and histogram to answer this question.

Esercizio 5.38. *Nutrition at Starbucks, Part I.*
The scatterplot below shows the relationship between the number of calories and amount of carbohydrates (in grams) Starbucks food menu items con- tain.21 Since Starbucks only lists the number of calories on the display items, we are interested in predicting the amount of carbs a menu item has based on its calorie content.

(a) Describe the relationship between number of calories and amount of carbohydrates (in grams) that Starbucks food menu items contain.

(b) In this scenario, what are the explanatory and response variables?

(c) Why might we want to fit a regression line to these data?

(d) Do these data meet the conditions required for fitting a least squares line?

Esercizio 5.39. *Treno Costiero, parte II*
Exercise 7.11 introduces data on the Coast Starlight Amtrak train that runs from Seattle to Los Angeles. The mean travel time from one stop to the next on the Coast Starlight is 129 mins, with a standard deviation of 113 minutes. The mean distance traveled from one stop to the next is 107 miles with a standard deviation of 99 miles. The correlation between travel time and distance is 0.636.

(a) Write the equation of the regression line for predicting travel time.

(b) Interpret the slope and the intercept in this context.

(c) Calculate R2 of the regression line for predicting travel time from distance traveled for the Coast Starlight, and interpret R2 in the context of the application.

(d) The distance between Santa Barbara and Los Angeles is 103 miles. Use the model to estimate the time it takes for the Starlight to travel between these two cities.

(e) It actually takes the the Coast Starlight about 168 mins to travel from Santa Barbara to Los Angeles. Calculate the residual and explain the meaning of this residual value.

(f) Suppose Amtrak is considering adding a stop to the Coast Starlight 500 miles away from Los Angeles. Would it be appropriate to use this linear model to predict the travel time from Los Angeles to this point?

Esercizio 5.40. *Misure corporee, III parte.*

L'Esercizio 5.33 introduces data on shoulder girth and height of a group of individuals. The mean shoulder girth is 108.20 cm with a standard deviation of 10.37 cm. The mean height is 171.14 cm with a standard deviation of 9.41 cm. The correlation between height and shoulder girth is 0.67.

(a) Write the equation of the regression line for predicting height.

(b) Interpret the slope and the intercept in this context.

(c) Calculate R^2 of the regression line for predicting height from shoulder girth, and interpret it in the context of the application.

(d) A randomly selected student from your class has a shoulder girth of 100 cm. Predict the height of this student using the model.

(e) The student from part (d) is 160 cm tall. Calculate the residual, and explain what this residual means.

(f) A one year old has a shoulder girth of 56 cm. Would it be appropriate to use this linear model to predict the height of this child?

Esercizio 5.41. *Grades and TV.*

Data were collected on the number of hours per week students watch TV and the grade they earned in a biology class on a 100 point scale. Based on the scatterplot and the residual plot provided, describe the relationship between the two variables, and determine if a simple linear model is appropriate to predict a student's grade from the number of hours per week the student watches TV.

Esercizio 5.42. *Nutrition at Starbucks, Part II.*

Nell'Esercizio 5.38 abbiamo introduced a data set on nutrition information on Starbucks food menu items. Based on the scatterplot and the residual plot provided, describe the relationship between the protein content and calories of these menu items, and determine if a simple linear model is appropriate to predict amount of protein from the number of calories.

Esercizio 5.43. *Helmets and lunches.*

The scatterplot shows the relationship between socioeconomic status measured as the percentage of children in a neighborhood receiving reduced-fee lunches at school (lunch) and the percentage of bike riders in the neighborhood wearing helmets (helmet). The average percentage of children receiving reduced-fee lunches is 30.8of 26.7% and the average percentage of bike riders wearing helmets is 38.8deviation of 16.9%.

(a) If the R^2 for the least-squares regression line for these data is 72%, what is the correlation between lunch and helmet?

(a) Calculate the slope and intercept for the least- squares regression line for these data.

(b) Interpret the intercept of the least-squares regression line in the context of the application.

(c) Interpret the slope of the least-squares regression line in the context of the application.

(d) What would the value of the residual be for a neighborhood where 40% of the children receive reduced-fee lunches and 40% of the bike riders receiving

reduced-fee lunch wear helmets? Interpret the meaning of this residual in the context of the application

Esercizio 5.44. *Outliers, Part I.*
Identify the outliers in the scatterplots shown below, and determine what type of outliers they are. Explain your reasoning.
Esercizio 5.45. *Outliers, Part II.*
Identify the outliers in the scatterplots shown below, and determine what type of outliers they are. Explain your reasoning.
Esercizio 5.46. *Crawling babies, Part II.*
L'Esercizio 5.32 introduces data on the average monthly tem- perature during the month babies first try to crawl (about 6 months after birth) and the average first crawling age for babies born in a given month. A scatterplot of these two variables reveals a potential outlying month when the average temperature is about 53 F and average crawling age is about 28.5 weeks. Does this point have high leverage? Is it an influential point?
Esercizio 5.47. *Urban homeowners, Part I.*
The scatterplot below shows the percent of families who own their home vs. the percent of the population living in urban areas in 2010.22 There are 52 observations, each corresponding to a state in the US. Puerto Rico and District of Columbia are also included.

(a) Describe the relationship between the per- cent of families who own their home and the percent of the population living in urban ar- eas in 2010.

(b) The outlier at the bottom right corner is Dis- trict of Columbia, where 100% of the population is considered urban. What type of outlier is this observation?

Esercizio 5.48.

Esercizio 5.49.

Esercizio 5.50.

Esercizio 5.51.