Esercitazioni con R

Massimiliano Pastore

(31 dicembre 2023)

Nelle pagine che seguono ci sono alcuni esercizi da svolgere utilizzando R inerenti ai temi del corso. A pag. 61 sono riportati l'elenco dei pacchetti e delle funzioni che possono essere utili per la risoluzione dei problemi proposti. Si tenga presente che, oltre a svolgere gli esercizi utilizzando R, è molto importante imparare ad interpretare i grafici ed i risultati ottenuti.

I file dati cui si fa riferimento sono elencati nell'indice analitico e possono essere scaricati alla pagina Moodle del corso: https://elearning.unipd.it/scuolapsicologia/

Per ottimizzare il tempo di scaricamento, tutti i file dati utilizzati per il corso e per gli esercizi sono compressi in formato zip. Pertanto, una volta scaricato il file è necessario decomprimere i vari file con un apposito programma (Es. winzip).

Nota bene: il presente materiale viene costantemente aggiornato, si tenga conto della data in prima pagina.

Indice

1 INTRODUZIONE A R

Esercizio 1.1

Si utilizzi la console di R per le seguenti operazioni:

1. $1/10$
2. $1/1000$
3. $1/10000$
4. $1/10^5$
5. Si scriva il numero 0.000073251 e si interpreti l'output prodotto.
6. Si scriva il numero 2955308822732 e si interpreti l'output prodotto.
7. Si arrotondi il numero 176.983345 escludendo le cifre decimali.
8. Si arrotondi il numero 176.983345 alla seconda cifra decimale.

Esercizio 1.2

Il fattoriale di un numero naturale, $n!$, è dato dal prodotto $n \times (n-1) \times (n-2) \times \ldots \times 1$. Si calcolino i fattoriali dei seguenti numeri utilizzando la console di R[1]:

1. 0
2. 5
3. $1/7$
4. 30
5. 1000

Esercizio 1.3

Si calcolino i risultati delle seguenti espressioni (tra parentesi quadra il risultato da ottenere):

1. $4 \times 5^2 + 9 \times 3^3$ [343]
2. $7 \times \sqrt{2^3 - 2}$ [17.1464]
3. $\frac{10+7}{log(2)}$ [24.53]
4. $\sqrt{(\frac{20 \times \pi}{9})^5}$ [128.8]
5. $1 - \left[\left(\frac{2}{3} - \frac{1}{6}\right) - \left(2 - \frac{3}{4}\right)\right]$ [1.75]
6. $\frac{\left(\frac{6}{5}\right)^4 \left(-\frac{10}{3}\right)^8 \left(\frac{2}{15}\right)^{-3}}{e^5 \left(\frac{3}{8}\right)^{-2}}$ [12633.65]

[1] Si può utilizzare la funzione `factorial()`

7. $\frac{\pi}{2}$ [1.5708]
8. $\frac{(9+\sqrt{7})\times\frac{1}{8\times3-3}}{8+5^3-(9^2+4^2)-36}$ [`Inf`]

Esercizio 1.4

1. Si produca la sequenza dei numeri da 1 a 90.
2. Si divida la sequenza di numeri generata per 90.
3. Si produca un oggetto `S` che contenga la sequenza dei multipli di 3 da 1 a 90.
4. Si individui il 13^o elemento della sequenza prodotta al punto precedente.
5. Si calcolino la somma, la media aritmetica e la varianza degli elementi di `S`.
6. Si crei una matrice `A` di dimensione 6×5 contenente gli elementi di `S` disposti per riga.
7. Si calcolino le medie delle colonne della matrice `A` creata al punto precedente e si aggiungano come riga sotto la stessa `A`.
8. Si calcolino le somme delle righe di `A` e si aggiungano come colonna a destra.
9. Si producano l'istogramma (con la funzione `hist()`), il boxplot (funzione `boxplot()`) ed il QQplot (funzione `qqnorm()`) degli elementi di `S`.

Esercizio 1.5

1. Si crei un vettore con nome `PI` che contenga 10 valori in sequenza e passo costante da $-\pi$ a $+\pi$ e si visualizzino i valori ottenuti.
2. Si calcolino media e varianza degli elementi contenuti in `PI`.
3. Si calcoli la somma del terzo, quinto e ottavo elemento del vettore `PI` (risultato = -0.3491).
4. Si calcolino e si visualizzino le radici quadrate degli elementi del vettore `PI`.

Esercizio 1.6

1. Si crei una matrice `A` di valori compresi tra 1 e 5 (a scelta) composta da 10 righe e 4 colonne.
2. Si calcolino le medie delle righe e delle colonne di `A`.
3. Si costruisca una matrice `B` composta da 5 righe estratte a caso della matrice `A`.
4. Si rappresenti graficamente la tabella di frequenze ottenuta dalla matrice `A` (funzione `barplot()`).

Esercizio 1.7

Si utilizzi il comando `letters`, con gli opportuni indici, per produrre le seguenti parole:

1. albero
2. cane
3. ermeneutica
4. orologio
5. patologico

Esercizio 1.8

1. Si crei un vettore numerico `X` contenente 50 valori estratti a caso dall'insieme $\{0, 1, 2, 3\}$.
2. Si produca la tabella di frequenze di `X`.
3. Si converta il vettore `X` in un vettore categoriale, `Y`, usando il comando `factor()`.
4. Si visualizzino i livelli del vettore `Y`.
5. Si converta il vettore `X` in un vettore categoriale, `Z`, assegnando ai valori $\{0, 1, 2, 3\}$ le seguenti etichette: 0 = giallo, 1 = verde, 2 = rosso, 3 = blu.
6. Si produca la tabella di frequenze di `Z`.

Esercizio 1.9

Si considerino 5 studenti (Gino, Pino, Tino, Beppe, Lino) con i rispettivi numeri di matricola (507, 535, 566, 955, 515) ed il voto preso ad un esame (19, 27, 28, 25, 25).

1. Si producano tre vettori contenenti rispettivamente i nomi, la matricola ed il voto dei tre studenti.
2. Si uniscano opportunamente i tre vettori per produrre un dataframe.
3. Si cancellino dal workspace i tre vettori prodotti al punto 1.
4. Si selezioni il voto di Beppe.
5. Si selezioni la riga con i valori relativi a Pino.
6. Si selezionino (con il comando `subset()`) gli studenti con voto pari a 25.
7. Si selezionino (con il comando `subset()`) gli studenti con voto superiore a 20.

Esercizio 1.10

Sara è una ricercatrice che ha rilevato dei dati relativi agli studenti di un corso di Statistica

della laurea magistrale. I dati sono organizzati in una apposita tabella e salvati su tre file uguali (`Sara_dataset`) ma con diversi formati: testo (`txt`), excel (`xls`) e SPSS (`sav`)[2].
Le informazioni raccolte da Sara sono le seguenti:

- genere dei soggetti (`gender`: 1 = femmine, 2= maschi)
- tipo di laurea triennale conseguita (`major`: 1 = psicologia, 2 = medicina, 3 = biologia, 4 = sociologia, 5 = economia)
- condizione sperimentale (`cond`: 1 = facile, ..., 4 = impossibile)
- autovalutazione del timore verso la matematica (`phobia`) su scala 0-10
- numero corsi di matematica frequentati (`prevmath`)
- punteggio ad un pre-test di matematica (`mathquiz`)
- punteggio ad un test di statistica (`statquiz`)
- auto-misurazione del battito cardiaco in condizioni normali (`hr_base`), prima del test (`hr_pre`) e dopo il test (`hr_post`)
- punteggi ad un test sull'ansia in condizioni normali (`anx_base`), prima del test (`anx_pre`) e dopo il test (`anx_post`)

Si prendano in considerazione i tre file, nei tre diversi formati:

1. Si importino i file in R assegnando a ciascuno un nome diverso (ad esempio `sara_txt`, `sara_xls` e `sara_sav`).
2. Si confrontino le strutture dei tre dataset importati e si valuti se sono uguali.
3. Si identifichino le proprietà e il livello di scala delle variabili del dataset.
4. Si identifichino i valori ottenuti dal 57^o soggetto.
5. Si individuino i punteggi del test `hr_base`.

Esercizio 1.11

In un reparto psichiatrico di un ospedale del nord Italia sono ricoverati 30 pazienti. Per ciascuno di essi sono state rilevate le seguenti informazioni: regione di residenza, classe sociale (definita come bassa, media e alta), punteggio su una scala di ansia (0 = poco ansioso, 7 = molto ansioso), età, tipo di disturbo presentato. I dati sono raccolti nel file `pazienti.xls`[3].

1. Si importi il file in R.
2. Si specifichi il livello di misura di ciascuna variabile del data-set.
3. Si produca la tabella di frequenze di ciascuna variabile del data-set nel modo più opportuno.
4. Si produca la tabella di frequenze cumulate (relative) per le variabili per cui abbia senso.
5. Si producano i grafici a barre o istogrammi delle variabili del data-set.
6. Si producano i grafici delle cumulate empiriche per le variabili per cui abbia senso.

[2] Dati da Welkowitz, Cohen & Ewen; `http://www.psych.nyu.edu/cohen/introwelk.html`
[3] Dati da Areni, Scalisi & Bosco, 2004

Esercizio 1.12

Utilizzando i dati dell'esercizio 1.11:

1. Si determinino i quartili delle variabili `cl.sociale`, `ansia` e `eta`.
2. Si calcoli il rango percentile[4] di 39 anni.
3. Si determini la moda delle variabili del data-set.
4. Si determini la mediana delle variabili del data-set per cui ha senso.
5. Si determini la media delle variabili del data-set per cui ha senso.
6. Si determinino deviazione standard, varianza e devianza delle variabili del data-set per cui ha senso.

Esercizio 1.13

Utilizzando i dati dell'esercizio 1.11:

1. Si produca il boxplot per le variabili del data-set per cui abbia senso.
2. Si determini, con un opportuno metodo grafico, se la distribuzione delle variabili `ansia` ed `eta` possa considerarsi approssimativamente normale.
3. Si produca un grafico opportuno per rappresentare le medie delle età dei soggetti (con un indicatore di variabilità associato) in funzione del disturbo diagnosticato.

Esercizio 1.14

Il file `healdrug.dat` contiene i dati di una ricerca sulla cura della propria salute e la propensione all'uso di farmaci (Hoffman & Fidell, 1979). Il campione è composto da 465 donne di età compresa tra 20 e 59 anni, residenti nella San Fernando Valley, Los Angeles, nel Febbraio 1975. Le variabili considerate sono: `timedrs` (numero di visite mediche), `attdrug` (propensione all'uso di farmaci), `atthouse` (propensione ai lavori domestici), `income` (livello di reddito), `emplmnt` (status lavorativo), `mstatus` (stato civile), `race` (gruppo etnico).

1. Si importi il file `healdrug.dat` in R assegnandogli nome HD.
2. Si identifichino unità statistiche e variabili del data-frame. Per ciascuna variabile si definiscano le proprietà metriche.
3. Si calcolino medie e deviazioni standard delle variabili su cui ciò è consentito.
4. Si producano le opportune statistiche riassuntive per le altre variabili.
5. Si producano gli istogrammi delle variabili `timedrs` e `attdrug` e si salvino in un file di formato jpeg.
6. Si producano i boxplot delle due variabili e si salvino in un file di formato jpeg.

[4] Dato un elemento di una distribuzione di valori x_i, il rango percentile di x_i esprime la percentuale di valori nella distribuzione minori o uguali di x_i.

7. Si esegua un test di normalità sulle due variabili utilizzando il test Kolmogorov-Smirnov (funzione `ks.test()`) oppure il test Shapiro-Wilk (funzione `shapiro.test()`).

8. Si producano i qq-plot per le due variabili e si salvino in un file di formato jpeg.

Esercizio 1.15

Il file `redditi.dat` contiene i redditi medi giornalieri (variabile reddito) di un campione rappresentativo di 1000 soggetti abitanti in Brasile e Venezuela (dati fittizi).

1. Si importi il file `redditi.dat` in R assegnandogli nome BV.

2. Si identifichino unità statistiche e variabili del data-frame. Per ciascuna variabile si definiscano le proprietà metriche.

3. Si produca un grafico in due parti (utilizzando il comando `layout()` oppure `par()`) con gli istogrammi delle distribuzioni dei redditi in Brasile e Venezuela e si salvi in un file di formato jpeg.

4. Si calcolino media e deviazione standard dei redditi separatamente per i due paesi.

5. Si esegua un test di normalità sulla variabile reddito per ciascuno dei due gruppi.

6. Si producano i qq-plot per la variabile `reddito` nei due gruppi[5] in un unico grafico diviso in due parti (utilizzando il comando `layout()` oppure `par()`) e si salvi in un file di formato jpeg.

7. Si produca un grafico (unico) che illustri le cumulate empiriche della variabile `reddito` nei due paesi[6].

8. Si calcolino i quartili della variabile `reddito` nei due gruppi e si confrontino.

Esercizio 1.16

In una ricerca sulla qualità percepita, vengono selezionate 5 stazioni sciistiche. In ciascuna stazione viene selezionato un campione di soggetti sciatori, cui viene chiesto di dare una valutazione ai costi degli impianti, le tipologie di servizi offerti, lo spessore e la qualità generale della neve. I dati sono raccolti nel file `sciatori.sav`.

1. Si importi il file `sciatori.sav` in R.

2. Si producano medie e deviazioni standard delle variabili del questionario in funzione della stazione di rilevazione (si può usare la funzione `aggregate()`).

3. Si producano la matrice di covarianza e di correlazione tra le variabili del questionario.

4. Si controlli graficamente se le distribuzioni dei punteggi nelle variabili del questionario possano considerarsi normalmente distribuite, individuando la presenza di eventuali outliers.

[5] Per uno dei due gruppi, il comando `qqline()` restituisce un messaggio di errore; si cerchi di capire il perchè.
[6] Per il calcolo della cumulata empirica si può usare la funzione `ecdf()`

Esercizio 1.17

1. Si utilizzi la funzione `rnorm()` per generare 1000 dati con media 100 e deviazione standard 15.
2. Si determini il campo di variazione dell'insieme di dati generato.
3. Si calcolino mediana, media e deviazione standard dell'insieme di dati generato.
4. Si produca il grafico di densità ad istogrammi dei dati (con la funzione `hist()`); successivamente si aggiunga al grafico la curva normale teorica utilizzando le funzioni `curve()` e `dnorm()`.
5. Si esegua un test di normalità sui valori generati.
6. Si producano in un unico grafico diviso in due parti (utilizzando il comando `layout()` oppure `par()`), il boxplot ed il qq-plot, e si salvi il grafico in un file di formato png.
7. Si produca il grafico della distribuzione cumulata empirica dei dati generati; successivamente di aggiunga al grafico la cumulata teorica utilizzando le funzioni `curve()` e `pnorm()`. Quali considerazioni è possibile trarre dalla lettura di questo grafico?

Esercizio 1.18

Si ripetano tutti i punti dell'esercizio 1.17 generando solo 10 dati anzichè 1000.

Esercizio 1.19

Si ripetano tutti i punti dell'esercizio 1.17 utilizzando la funzione `rchisq` con due gradi di libertà al posto della funzione `rnorm`.

Esercizio 1.20

Si ripetano tutti i punti dell'esercizio 1.17 utilizzando la funzione `rt` con due gradi di libertà al posto della funzione `rnorm`.

Esercizio 1.21

Si ripeta l'esercizio 1.20 generando solo 10 dati al posto di 1000. Si confrontino i risultati ottenuti con quelli degli esercizi 1.17, 1.18, 1.19 e 1.20.

Esercizio 1.22

Si costruisca un `data.frame` formato da 100 righe e 4 colonne ottenute come segue:

- Colonna 1: 100 numeri estratti con la funzione `rnorm()`, con media 4 e dev. standard 2
- Colonna 2: 100 numeri estratti con la funzione `rt()`, con 3 gradi di libertà

- Colonna 3: 100 numeri estratti con la funzione `rf()`, con 4 e 8 gradi di libertà
- Colonna 4: 100 numeri estratti con la funzione `rchisq()` con 4 gradi di libertà

Sul `data.frame` ottenuto:

1. Si producano le statistiche descrittive di base con il comando `summary()`. Dalla lettura delle statistiche ottenute si cerchi di stabilire quanto siano simili tra loro le quattro variabili.
2. Si rappresentino con istogrammi le distribuzioni univariate delle variabili.
3. Si rappresentino graficamente le distribuzioni bivariate con il comando `plot()`.

Esercizio 1.23

Il data frame `kidiq` nel pacchetto `ADati` contiene un campione di 434 coppie madri-figli sulle quali sono state rilevate le seguenti variabili: `kid_score`, punteggio del figlio ad un test cognitivo; `mom_hs`, livello di istruzione della madre (1 = diplomata, 0 = non diplomata); `mom_iq`, QI della madre; `mom_work` tipologia di lavoro della madre; `mom_age`, età della madre.

1. Si renda disponibile nel workspace il data frame `kidiq`.
2. Si producano le statistiche descrittive del data frame utilizzando la funzione `describe()` del pacchetto `psych`.
3. Si rappresentino graficamente le variabili del data frame scegliendo per ciascuna di esse un grafico appropriato.
4. Si rappresenti graficamente con un boxplot, la distribuzione di punteggi dei figli separandoli in funzione del livello di istruzione della madre.
5. Si valuti graficamente se le età delle madri siano distribuite normalmente.
6. Si determinino i seguenti valori della variabile `mom_iq`: il terzo quintile, il settimo sestile, il primo decile, l'ottavo percentile.

Esercizio 1.24

Si esplorino graficamente i dati dell'esercizio 1.11 **utilizzando le funzioni del pacchetto** `DataExplorer`; in particolare:

1. Si determini il numero di casi mancanti per ciascuna variabile.
2. Si stabilisca quante sono le variabili discrete e quante le continue.
3. Si rappresentino le distribuzioni univariate delle variabili.
4. Si rappresentino le distribuzioni con boxplot in funzione del tipo di disturbo (per le variabili in cui questo abbia senso).
5. Si rappresentino le correlazioni tra le variabili.

Esercizio 1.25

Il dataframe `earlymath`[7], nel pacchetto `ADati`, contiene i dati relativi ad un campione di 120 bambini con le seguenti variabili: `gender` – genere dei soggetti–, `ENT` – misura delle abilità matematiche –, `QI` – livello di intelligenza –, `WM` – memoria di lavoro –, `STM` – memoria a breve termine –, `ANS` – misura la capacità che permette di stimare ad esempio i risultati di operazioni per via approssimata (*Approximate Number Sistem*) –.

1. Si renda disponibile il dataframe in R.
2. Si utilizzino **le funzioni del pacchetto** `DataExplorer` per:
   - Visualizzare il numero di casi mancanti per ciascuna variabile del data set.
   - Visualizzare quante sono le variabili discrete e quante le continue.
   - Rappresentare le distribuzioni univariate delle variabili.
   - Rappresentare le distribuzioni con boxplot in funzione del genere.
   - Rappresentare le correlazioni tra le variabili.

Esercizio 1.26

Il dataset `gambling`[8], nel pacchetto `ADati`, contiene i dati relativi ad un campione di 1221 studenti di età compresa tra 15 e 19 anni selezionati in 42 classi di 5 scuole. A ciascun soggetto viene somministrato un questionario che rileva le seguenti misure: `ID` – codice identificativo –, `school` – scuola frequentata –, `class` – classe frequentata –, `age` – età –, `gender` – genere –, `frequency` – frequenza di gioco –, `perc_peers` – percezione di quanto giocano i pari –, `disapproval` – grado di disapprovazione – , `risk` – percezione del rischio – e , `par_know` – percezione del controllo genitoriale –.

1. Si importi il file in R.
2. Si valuti se nel data-set ci siano casi mancanti.
3. Si producano i grafici appropriati per le distribuzioni univariate delle variabili del data set.
4. Si calcolino media, mediana e i decili della variabile `disapproval`.
5. Si rappresenti graficamente la distribuzione dei valori di `frequency` separatamente per maschi e femmine. Quindi si valuti in quale dei due gruppi presenti una frequenza media più alta.
6. Si utilizzi una rappresentazione grafica che permetta di valutare se vi sia una relazione tra la percezione di rischio e l'età dei soggetti. Si interpreti il grafico ottenuto.

[7] Fonte: Passolunghi, M.C., Cargnelutti, E., Pastore, M. (2014). The contribution of general cognitive abilities and approximate number system to early mathematics. *British Journal of Educational Psychology*, 84, 631-649.

[8] Fonte: Canale, N., Vieno, A. ter Bogt, T., Pastore, M., Siciliano, V., Molinaro, S. (2016). Adolescent gambling-oriented attitudes mediate the relantionship between parental knowledge and adolescent gambling: Implications for prevention. *Prevention Science*, 17, 970-980.

2 TEORIA DEI CAMPIONI

Esercizio 2.1

Data la popolazione $\Omega = \{0, 1, 2, 2, 5\}$

1. Si determini la distribuzione campionaria della media dei campioni ordinati di numerosità $n = 2$ **con reinserimento**.
2. Si rappresenti graficamente la distribuzione campionaria ottenuta.
3. Si calcoli la media della distribuzione campionaria delle medie ($\mu_{\overline{x}}$) e si confronti con la media della popolazione μ.
4. Si calcoli l'errore standard della media ($\sigma_{\overline{x}}$).
5. Si producano la distribuzione campionaria delle varianze con e senza la correzione e si rappresentino graficamente.
6. Si calcolino la media della distribuzione campionaria delle varianze non corrette (μ_{s^2}) e corrette ($\mu_{\hat{\sigma}^2}$) e si confrontino con la varianza della popolazione σ^2.

Esercizio 2.2

Data la popolazione $\Omega = \{0, 1, 2, 2, 5\}$

1. Si determini la distribuzione campionaria della media dei campioni ordinati di numerosità $n = 3$ **senza reinserimento**.
2. Si rappresenti graficamente la distribuzione campionaria ottenuta.
3. Si calcoli la media della distribuzione campionaria delle medie ($\mu_{\overline{x}}$) e si confronti con la media della popolazione μ.
4. Si calcoli l'errore standard della media ($\sigma_{\overline{x}}$).
5. Si produca la distribuzione campionaria delle varianze con e senza la correzione e si rappresentino graficamente.
6. Si calcolino la media della distribuzione campionaria delle varianze non corrette (μ_{s^2}) e corrette ($\mu_{\hat{\sigma}^2}$) e si confrontino con la varianza della popolazione σ^2.

Esercizio 2.3

Sia Ω una popolazione composta da 150 elementi numerici con valore compreso tra 1 e 20; in figura 1 sono rappresentate con grafico a barre le frequenze degli elementi di Ω.

1. Si ricostruisca la popolazione Ω sulla base del grafico.
2. Si calcolino i valori dei parametri della popolazione μ e σ.
3. Si estraggano da Ω 3 campioni casuali senza reinserimento di dimensione $n = 5, 50, 100$ e su ciascuno di essi si calcoli la media aritmetica (rispettivamente $\overline{x}_5$, $\overline{x}_{50}$ e $\overline{x}_{100}$).

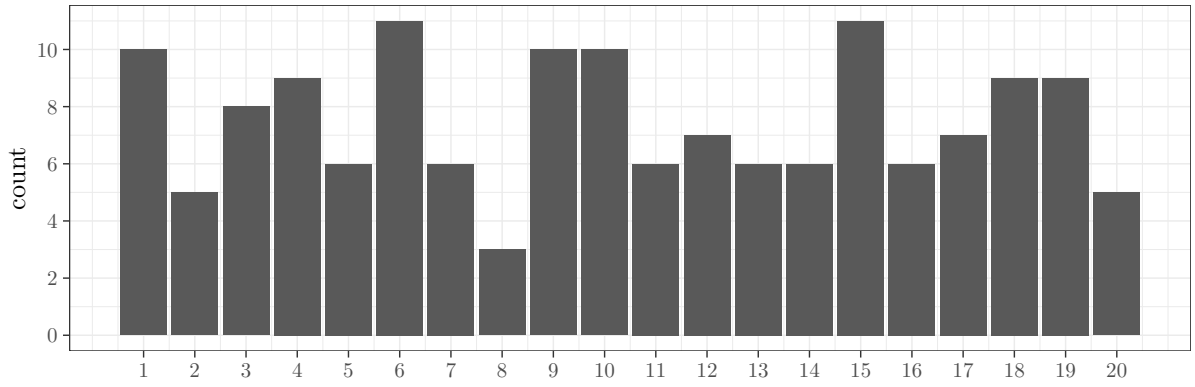

Fig. 1: Rappresentazione grafica degli elementi appartenenti alla popolazione Ω dell'esercizio 2.3.

4. Si rappresentino graficamente (con la funzione `barplot()`) le frequenze osservate dei valori nei tre campioni estratti.

5. Si producano (con la funzione `campionaria.media()`) le distribuzioni campionarie per campioni (senza reinserimento) di dimensione $n = 5, 50, 100$ e si confrontino graficamente. Suggerimento: per avere distribuzioni affidabili si producano almeno 3000 campioni.

6. Sulla base delle distribuzioni campionarie ottenute al punto precedente si stabilisca per quale numerosità campionaria sia più alta la probabilità di ottenere un campione la cui media $\overline{x}$ risulti più estrema (maggiore o minore) rispetto alla media vera della popolazione più o meno una deviazione standard.

Esercizio 2.4

Date le due popolazioni di numerosità $N = 4$: $\Omega_1 = \{3, 7, 8, 9\}$ (con $\mu = 6.75$ e $\sigma^2 = 5.1875$) e $\Omega_2 = \{2, 4, 5, 8\}$ (con $\mu = 4.75$ e $\sigma^2 = 4.6875$) si considerino le distribuzioni di tutti i possibili campioni di dimensione $n = 2$ con e senza reinserimento. Si dimostrino le seguenti uguaglianze:

- $\mu_{\overline{x}_1 - \overline{x}_2} = \mu_1 - \mu_2$
- $\sigma^2_{\overline{x}_1 - \overline{x}_2} = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n}$ nel caso con reinserimento
- $\sigma^2_{\overline{x}_1 - \overline{x}_2} = \left(\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n}\right)\left(\frac{N-n}{N-1}\right)$ nel caso senza reinserimento.

Si può seguire il seguente procedimento, prima per il caso con reinserimento e poi senza reinserimento:

1. Si determinino le distribuzioni campionarie della media dei campioni di numerosità $n = 2$ per Ω_1 e Ω_2.

2. Si determinino tutte le possibili coppie di medie date dalle due distribuzioni campionarie ottenute (si può utilizzare la funzione `expand.grid()`).

3. Si determini la distribuzione campionaria delle differenze tra le medie.

4. Si calcolino la media ($\mu_{\overline{x}_1 - \overline{x}_2}$) e la varianza ($\sigma^2_{\overline{x}_1 - \overline{x}_2}$) della distribuzione campionaria delle differenze tra medie e si confrontino con i valori ottenuti dalle formule riportate sopra.

Esercizio 2.5

Il file `esame.txt` contiene i voti di un appello di Psicometria sostenuto da 239 studenti; siano la nostra popolazione Ω.

1. Si importino i dati in R utilizzando la funzione `scan()`.
2. Si calcolino media (μ) e varianza (σ^2) della popolazione.
3. Si determini il numero di possibili campioni senza reinserimento ottenibili da Ω di dimensione $n = 3, 20, 50$.
4. Si produca la distribuzione campionaria della media per le tre numerosità e si confrontino.
5. Si stimi la probabilità che la media di un campione di dimensione $n = 10$ estratto senza reinserimento da tale popolazione abbia una media maggiore o uguale a 18.

Esercizio 2.6

Si consideri una variabile X distribuita normalmente in due distinte popolazioni caratterizzate dai seguenti parametri: $\mu_1 = 10$ e $\sigma_1 = 3$, $\mu_2 = 8$ e $\sigma_2 = 2$.

1. Utilizzando la funzione `curve()`, si rappresentino sullo stesso grafico le distribuzioni attese della variabile X nelle due popolazioni.
2. Si supponga di estrarre campioni di numerosità $n = 10$ da ciascuna distribuzione; si rappresentino graficamente le distribuzioni campionarie attese delle medie di questi campioni (sempre con la funzione `curve()`).
3. Si producano le distribuzioni campionarie delle medie per campioni estratti delle due popolazioni, sempre per $n = 10$.
4. Si rappresentino graficamente e si confrontino le distribuzioni ottenute.
5. Si utilizzino le distribuzioni ottenute per produrre la distribuzione campionaria della differenza tra le medie e si stimi il valore atteso di questa statistica.
6. Si utilizzino le distribuzioni ottenute per produrre la distribuzione campionaria della somma tra le medie e stimare il valore atteso di questa statistica.
7. Si consideri la domanda al punto 2; se la numerosità campionaria considerata fosse $n = 100$ ci aspettiamo che la sovrapposizione delle distribuzioni campionarie aumenti o diminuisca? Si verifichi graficamente cosa succede.

Esercizio 2.7

Si consideri una variabile X distribuita normalmente con: $\mu = 10$ e $\sigma = 3$. Vogliamo stimare l'errore standard della varianza campionaria (σ_{s^2}) per campioni di dimensione 50.

1. Utilizzando la funzione `curve()` si rappresenti graficamente la distribuzione (teorica) della variabile X.
2. Si estragga un campione casuale di dimensione $n = 50$ con la funzione `rnorm()` e, su questo, si calcolino media ($\overline{x}$) e varianza (s^2) campionarie.

3. Si estraggano 10000 campioni di dimensione $n = 50$ mettendoli in un oggetto chiamato `X`.
4. Si calcolino medie e varianze dei campioni estratti; per comodità si mettano le medie in un vettore chiamato `mcamp` e le varianze in un vettore chiamato `s2camp`.
5. Si rappresentino graficamente le distribuzioni campionarie ottenute della media e della varianza.
6. Si calcolino valore atteso ed errore standard della distribuzione campionaria della media e si confrontino con i valori che otteniamo con le formule.
7. Si calcolino valore atteso ed errore standard della distribuzione campionaria della varianza confrontando il valore atteso ottenuto con quello derivato dalla formula.

Esercizio 2.8

Siano date una variabile casuale X distribuita normalmente con $\mu = 5$ e $\sigma = 2$ ed una statistica $T = \sum x_i^2$ in cui x_i, $i = 1, \ldots, n$ è un campione di n valori estratti da X. Vogliamo studiare empiricamente le proprietà della distribuzione campionaria della statistica T.

1. Utilizzando la funzione `curve()` si rappresenti graficamente la distribuzione (teorica) della variabile X.
2. Si estraggano 10000 campioni di dimensione $n = 10$ e su questi si applichi la funzione T. **Suggerimento:** si può utilizzare la funzione `apply()` definendo la statistica T con la funzione `function(x){ sum(x^2) }`.
3. Si rappresenti la distribuzione campionaria ottenuta della statistica T.
4. Si calcolino media e deviazione standard della distribuzione campionaria ottenuta.
5. Se aumentiamo la numerosità campionaria cosa ci possiamo attendere rispetto alla media della distribuzione campionaria di T? Si provi a dare una risposta e quindi si verifichi empiricamente ripetendo i punti precedenti con campioni di numerosità 50 e 100.

Esercizio 2.9

Supponiamo di avere una moneta bilanciata e volere studiare la distribuzione campionaria della statistica T = numero di teste in L lanci con $L = 3, 10, 50, 200$. Possiamo ipotizzare che la popolazione di lanci sia virtualmente infinita e, con la funzione `sample()`, simulare L lanci di una moneta bilanciata, ad esempio:

```
L <- 3 # tre lanci
sample( 0:1, size = L, replace = TRUE ) # 1 = testa
```

e quindi contare il numero di teste che escono (per esempio con la funzione `sum()`) nel modo seguente:

```
sum( sample( 0:1, size = L, replace = TRUE ) )
```

Per ripetere più volte il processo di campionamento possiamo utilizzare la funzione `replicate()` come segue:

```
B <- 1000 # numero di repliche di campioni
replicate( B, sample( 0:1, size = L, replace = TRUE ) )
```

da cui si ottiene una matrice in cui ciascuna colonna rappresenta un campione. Su ogni campione prodotto (colonne della matrice) possiamo calcolare la statistica test T (contando quanti uno ci sono nei campioni prodotti).

1. Si producano 1000 campioni di L lanci della moneta.
2. Si calcoli il numero di teste per ogni campione ottenuto, ovvero la distribuzione campionaria del numero di teste per ogni valore di L.
3. Si rappresentino graficamente le quattro distribuzioni campionarie ottenute.
4. Quali considerazioni si possono fare confrontando i grafici?

Esercizio 2.10

Si ripeta l'esercizio 2.9 simulando una moneta non bilanciata in cui la probabilità di ottenere testa è 0.75. Si confrontino le distribuzioni ottenute con quelle dell'esercizio 2.9.

Esercizio 2.11

Si ripeta l'esercizio 2.9 simulando una moneta non bilanciata in cui la probabilità di ottenere testa è 0.05. Si confrontino le distribuzioni ottenute con quelle dell'esercizio 2.9.

Esercizio 2.12

Un sacchetto contiene 10 mele di cui 6 rosse e 4 gialle. Supponiamo di estrarre in sequenza (senza reinserimento) 4 mele e contare il numero di mele rosse nel campione, n_r. Vogliamo studiare la distribuzione campionaria di questa statistica.

1. Si determini, utilizzando la funzione `choose()`, il numero di possibili campioni di 4 mele che è possibile ottenere.
2. Si produca una matrice `S` contenente tutti i campioni possibili.
3. Si calcoli il numero di mele rosse in ciascuno dei campioni ottenuti e si produca la relativa tabella di frequenze (anche graficamente).
4. Si calcoli la media della distribuzione campionaria di n_r.
5. Si calcoli l'errore standard della media.
6. Si calcoli la probabilità che in un campione estratto a caso dal sacchetto ci siano esattamente 2 mele rosse.
7. Si calcoli la probabilità che in un campione estratto a caso dal sacchetto ci siano almeno 2 mele rosse.
8. Si calcoli la probabilità che in un campione estratto a caso dal sacchetto non ci siano mele rosse.

Esercizio 2.13

In una città ci sono due ospedali[9]; in quello più grande nascono circa 45 bambini la settimana, in quello più piccolo, circa 15 la settimana; il numero di maschi è approssimativamente il 50% dei nati, ovviamente tale percentuale può variare di settimana in settimana. In ciascun ospedale si conta il numero di volte in cui è stato superato il 60% dei maschi tra i nati nel corso di in un anno. In quale dei due si registrerà più volte tale evento: A) Nel più grande; B) Nel più piccolo; C) Lo stesso numero di volte?

Per rispondere alla domanda si proceda come segue:

1. Si simuli con la funzione `sample()` un campione di neonati nell'ospedale piccolo e si stimi la percentuale di maschi.
2. Si simuli un campione di neonati nell'ospedale grande e si stimi la percentuale di maschi.
3. Si utilizzi la funzione `replicate()` per generare 1000 campioni estratti da ciascun ospedale.
4. Si calcolino le percentuali di maschi osservati nei 1000 campioni generati al punto precedente.
5. Si rappresentino graficamente le distribuzioni dei valori ottenuti.
6. Si calcoli la proporzione di casi nei due ospedali in cui la percentuale di maschi supera il 60%.

Esercizio 2.14

Il dataset `vaes2015` del pacchetto `ADati` contiene un campione di 200 soggetti coinvolti in una ricerca sul pregiudizio verso gli immigrati. In particolare, la variabile `PREGIUDIZIO` misura il grado di pregiudizio su una scala tra 0 e 7 in cui più alti sono i punteggi più è alto il livello di pregiudizio.

1. Si rendano disponibili i dati in R.
2. Si rappresenti graficamente la distribuzione dei punteggi di pregiudizio.
3. Si calcolino media ($\overline{x}$) e varianza (s^2) dei punteggi di pregiudizio.
4. Assumiamo che i valori campionari di $\overline{x}$ e s^2 siano una buona stima dei rispettivi parametri (μ e σ^2) di riferimento nella popolazione. Si rappresenti graficamente (con il comando `curve()`) la distribuzione teorica dei punteggi di pregiudizio nella popolazione.
5. Si calcoli la probabilità che un soggetto, selezionato a caso dalla popolazione, abbia un punteggio superiore a 4.2.
6. Si calcoli la probabilità che un campione di 8 soggetti, selezionati a caso dalla popolazione, abbia un punteggio medio superiore a 4.2. Prima di procedere si rifletta: tale probabilità sarà maggiore o minore di quella calcolata al punto precedente?

[9] tratto da: Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and Biases. *Science*, 185(4157), 1124-1131.

Esercizio 2.15

Il Body Influence Assessment Inventory (BIAI; Osman & al. 2006) è uno strumento self-report che valuta la soddisfazione per la propria immagine corporea. I punteggi possibili sono compresi tra 0 e 112 e valori più alti indicano una maggiore soddisfazione. Un ricercatore vuole sapere se vi siano differenze tra maschi e femmine nei punteggi; a tal fine recluta un campione di 1000 soggetti adolescenti, metà maschi e metà femmine, di età compresa tra 14 e 18 anni, cui somministra il questionario.

I punteggi medi dei maschi sono risultati $\overline{x}_m = 53.64$ (con $\hat{\sigma}_m = 11.5$), i punteggi delle femmine $\overline{x}_f = 50.04$ (con $\hat{\sigma}_f = 8.5$).

Vogliamo determinare la distribuzione campionaria della differenza tra le medie dei campioni.

1. Si rappresentino graficamente le distribuzioni dei punteggi nelle popolazioni di maschi e femmine, utilizzando opportunamente le stime campionarie.
2. Utilizzando opportunamente la funzione `replicate()` si estraggano B campioni[10] dalle due popolazioni (quella dei maschi e quella delle femmine) e si calcolino le B differenze tra le medie campionarie.
3. Si calcolino la media e la deviazione standard delle B differenze ottenute al punto precedente.
4. Si verifichi graficamente che la distribuzione campionaria ottenuta è normale con:
   - media uguale alla differenza tra le medie delle due popolazioni rappresentate al punto 1.
   - deviazione standard data da $\sigma_{\overline{x}_m - \overline{x}_f} = \sqrt{\frac{\sigma_m^2}{n_m} + \frac{\sigma_f^2}{n_f}}$

Esercizio 2.16

La funzione `mvrnorm()` (nel pacchetto `MASS`) permette di generare dati (a coppie) provenienti da una distribuzione normale bivariata con correlazione ρ; ad esempio, supponiamo di voler estrarre 50 coppie di valori da una popolazione in cui la correlazione è $\rho = .3$ possiamo fare come segue:

```
library(MASS)
RHO <- matrix( c( 1, .3, .3, 1 ), nrow = 2) # matrice di correlazione
X <- mvrnorm( 50, c( 0, 0 ), RHO )
```

Utilizzando la funzione `replicate()` insieme a `mvrnorm()` si producano le distribuzioni campionarie del coefficiente di correlazione r sulla base della numerosità (n) e della correlazione vera nella popolazione (ρ) indicate di seguito e si rappresentino graficamente.

1. $\rho = 0$, $n = 10, 50, 200$.
2. $\rho = 0.4$, $n = 75, 100, 400$.
3. $\rho = 0.1, 0.45, 0.8$, $n = 25$.
4. $\rho = 0.1, 0.45, 0.8$, $n = 500$.

[10] B deve essere un numero molto grande, si consiglia almeno 10000.

Esercizio 2.17

Il data frame `kidiq` nel pacchetto `ADati` contiene un campione di 434 coppie madri-figli descritto nell'esercizio 1.23.

1. Si renda disponibile nel workspace il data frame `kidiq`
2. Si calcolino media e deviazione standard dei punteggi dei figli per tutto il campione e, successivamente, suddividendo il campione tra madri con diploma e madri senza diploma.
3. Si rappresentino graficamente le densità dei punteggi di `kid_score` nei due gruppi (madri diplomate e non).
4. Supponiamo di osservare una nuova coppia in cui il bambino presenti un punteggio pari a 83. Si calcolino: 1) la probabilità di osservare a caso un punteggio inferiore a 83 nel gruppo delle madri non diplomate; 2) la probabilità di osservare a caso un punteggio inferiore a 83 nel gruppo delle madri diplomate; 3) quanto è più plausibile che la madre del bambino osservato sia diplomata rispetto a non diplomata.
5. Assumiamo che nelle popolazioni di madri diplomate e non le medie di `kid_score` corrispondano a quelle stimate sul campione. Si rappresentino graficamente (utilizzando il comando `curve()`) le distribuzioni campionarie attese delle medie nei due gruppi e si confrontino.
6. Sulla base delle formule già viste negli esercizi 2.4 e 2.15 si determini e si rappresenti graficamente la distribuzione attesa delle differenze tra le medie campionarie $\overline{x}_1 - \overline{x}_0$ dei punteggi dei figli.
7. Si calcoli la probabilità che la differenza tra le medie sia superiore a 15.

Esercizio 2.18

Sia data una variabile casuale $Y \sim \mathcal{N}(0, 1)$ e due campioni di dimensione n estratti da essa – $\mathbf{y}_1$ e $\mathbf{y}_2$ –. Vogliamo determinare la distribuzione campionaria delle seguenti statistiche: $T_1 = max(\mathbf{y}_1)$ – valore massimo nel primo campione –, $T_2 = max(\mathbf{y}_2)$ – valore massimo nel secondo campione – e $T_3 = max(\mathbf{y}_1) - max(\mathbf{y}_2)$ – differenza tra i due massimi.

1. Si consideri $n = 5$ e si produca empiricamente la distribuzione campionaria di T_1 e la si rappresenti graficamente.
2. Si calcolino media ed errore standard di T_1.
3. Si produca empiricamente la distribuzione campionaria di T_2 (sempre con $n = 5$). Ci aspettiamo che questa distribuzione abbia forma, media e deviazione standard diverse da quelle della distribuzione di T_1?
4. Si produca empiricamente la distribuzione campionaria di T_3 e se ne calcolino media e deviazione standard.
5. Si ripetano tutti i punti precedenti con campioni di dimensione $n = 1000$. Si cerchi di ipotizzare (a priori) che cosa accadrà.

3 INFERENZA

Esercizio 3.1

(Non richiede R) Date le seguenti ipotesi:

1. Si individuino le unità statistiche e le variabili.
2. Si determinino le proprietà delle variabili e la scala di misura relativa.
3. Si individui il tipo di relazione ipotizzabile tra le variabili ed i ruoli delle stesse.

- In un esperimento sulla memoria 30 studenti vengono sottoposti ad una prova di ricordo. Per ciascuno di essi viene misurato il numero di errori commessi nella rievocazione di 60 stimoli visivi. Si ipotizza che le femmine abbiano una capacità di memoria migliore rispetto a quella dei maschi.
- Uno psicologo ed uno psichiatra valutano 25 pazienti per individuare il disturbo cui sono affetti tra i seguenti: schizofrenia, nevrosi ossessiva, paranoia, fobia. Si ipotizza che non vi siano differenze tra i due nella modalità di classificare i pazienti.
- Uno psicologo scolastico sospetta che vi siano differenze nelle prestazioni in matematica tra le scuole di alcuni distretti. A tal fine confronta tra loro i voti medi delle classi di cinque scuole per ciascun distretto.
- Uno psicologo ed uno psichiatra valutano 25 pazienti per individuare la gravità di un certo disturbo. Si ipotizza che lo psicologo tenda ad attribuire un maggiore grado di gravità al disturbo considerato.
- Per valutare l'efficacia di un corso di statistica il docente effettua un test sulle competenze degli studenti prima del corso. Al termine del corso ripete il test attendendosi che i punteggi degli studenti siano aumentati.
- In una ricerca sulle preferenze musicali vengono intervistati 200 adolescenti (114 maschi e 86 femmine) cui viene chiesto quanto ascoltano musica dark. Si ipotizza che i soggetti che preferiscono la musica dark siano più propensi alla depressione e che questo effetto sia più accentuato nelle femmine.

Esercizio 3.2

Si supponga di voler testare se una moneta risulti essere bilanciata lanciandola per 12 volte. Sia T_0 il numero di teste osservato nel campione di lanci e π la probabilità di testa.

1. Si definiscano le ipotesi H_0 e H_1 del test.
2. Si rappresenti graficamente la distribuzione campionaria della statistica test sotto l'ipotesi H_0 (utilizzando la funzione `dbinom()`).
3. Si determini la regione critica del test con $\alpha = .05$ (si può utilizzare la funzione `qbinom()`).
4. Supponendo di avere osservato $T_0 = 3$ teste su 12 lanci, si prenda una decisione in merito alle ipotesi formulate al punto 1.
5. Si ripeta il test con la funzione `binom.test()`.

Esercizio 3.3

Il file `esame.txt` contiene i voti di un appello di Psicometria sostenuto da 239 studenti.

1. Si importino i dati in R utilizzando la funzione `scan()`.
2. Si valuti con una o più opportune rappresentazioni grafiche se la distribuzione dei voti possa considerarsi normale.
3. Si vuole valutare l'ipotesi che il voto medio nella vera popolazione di cui il campione fa parte sia 18: si definiscano le ipotesi H_0 e H_1, considerando, data la difficoltà della materia, che i voti risultino al di sotto della media.
4. Si calcoli la media dei voti nel campione $\overline{x}$ e l'errore standard associato alla stima $\sigma_{\overline{x}}$.
5. Assumendo che il campione estratto provenga effettivamente da una popolazione con media 18 si individui graficamente e si determini (con la funzione `pnorm()`) la probabilità di estrarre dalla popolazione un campione con media $\leq \overline{x}$.
6. Si esegua il test definito nell'ipotesi al punto 3 con la funzione `t.test()` confrontando il risultato con quello ottenuto al punto precedente.

Esercizio 3.4

Nel file `MFTP.dat` sono raccolti i punteggi ottenuti da un campione di 97 neolaureati in Psicologia nel *Major Field Test in Psychology II* (MFTP), un test per la valutazione delle competenze psicologiche. Da studi effettuati in anni precedenti è emerso che il punteggio medio ottenuto al test risulta essere 156.5. Si vuole sapere se il campione possa considerarsi compatibile con tutti quelli a cui il test è stato somministrato precedentemente.

1. Si importino i dati in R con la funzione `scan()`.
2. Si calcolino media e deviazione standard dei punteggi.
3. Si producano in un unico layout il grafico ad istogrammi, il boxplot ed il qqplot relativi ai punteggi.
4. Si valuti con un test opportuno se la distribuzione dei punteggi possa considerarsi normale formulando le ipotesi H_0 e H_1 relative al test.
5. Si valuti con il test opportuno se la media campionaria si differenzia significativamente da quella generale del test formulando le ipotesi H_0 e H_1 relative al test.

Esercizio 3.5

1. Si utilizzi la funzione `rnorm` per generare un vettore (Y) di 30 dati con media 20 e deviazione standard 5.
2. Si crei un fattore (A) con tre livelli e dieci osservazioni per livello (si possono usare la funzione `rep` e/o la funzione `factor`).
3. Si calcolino le medie e deviazioni standard del vettore Y utilizzando A come variabile indipendente (si ottengono tre gruppi da 10 osservazioni ciascuno).

4. Si rappresentino graficamente, nella maniera più opportuna, i punteggi dei tre gruppi ottenuti.
5. Si formulino le ipotesi di normalità associate alle distribuzioni di punteggi nei tre gruppi e si valutino tali ipotesi con i test più opportuni.
6. Si formuli l'ipotesi di omogeneità delle varianze nei tre gruppi, si valuti tale ipotesi con il test opportuno e si stabilisca se il risultato sia coerente con quello osservato al punto 4.
7. Si formuli l'ipotesi (nulla) di uguaglianza delle medie tra i gruppi e si valuti tale ipotesi con il test Anova.
8. Si stimi la dimensione dell'effetto con η^2 e la si interpreti.

Esercizio 3.6

1. Si aggiunga al vettore Y, creato nell'esercizio 3.5, un vettore di 10 elementi con media 12 e deviazione standard 5.
2. Si aggiunga al fattore A, creato nell'esercizio 3.5, un quarto livello con 10 osservazioni.
3. Si ripetano i punti 3-8 dell'esercizio 3.5 confrontando i risultati.

Esercizio 3.7

In figura 2 sono rappresentate rispettivamente le distribuzioni campionarie (con stessa varianza $\sigma^2 = 1$) di una statistica test T utilizzata per la stima di un parametro θ, sotto le due ipotesi H_0 e H_1.

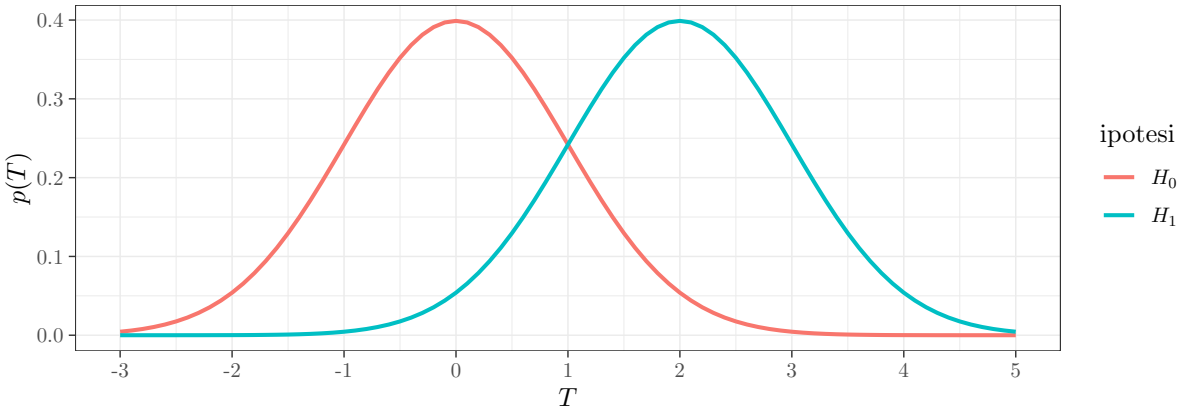

Fig. 2: Esercizio 3.7: rappresentazione grafica della distribuzione campionaria teorica di una statistica T sotto due diverse ipotesi H_0 e H_1.

1. Sulla base della rappresentazione, si formulino le ipotesi H_0 e H_1 in forma puntuale.
2. Si determini la regione critica del test per rigettare H_0 al 5%. Suggerimento: si utilizzi la funzione `qnorm()`.
3. Si calcoli la probabilità di rigettare H_0 nel caso in cui fosse vera H_1 (potenza del test). Suggerimento: si utilizzi la funzione `pnorm()`.

4. Supponendo di avere le seguenti tre stime del parametro θ ottenute su 3 diversi campioni: $\hat{\theta}_1 = 0.3$, $\hat{\theta}_2 = 2.1$ e $\hat{\theta}_3 = 3.9$; si stabilisca sotto quale delle due ipotesi sia più plausibile osservare ciascuno di tali risultati. Suggerimento: si utilizzi la funzione `dnorm()`.

Esercizio 3.8

Una borsa contiene al suo interno 5 mele gialle e rosse in proporzione ignota. Estraiamo dalla borsa 3 mele, in sequenza e senza reinserimento, ottenendo il seguente campione: [🔴🟡🟡] (una mela rossa e due gialle). Sulla base del campione osservato, quale sarà la composizione più plausibile del contenuto della borsa in termini di numero di mele rosse e gialle?

1. Si individuino le possibili composizioni del contenuto della borsa.
2. Si determinino i possibili campioni osservabili per ciascuno scenario.
3. Sulla base dei campioni ottenuti si stabilisca quali scenari sono compatibili con il campione estratto ([🔴🟡🟡]).
4. Si calcoli in ogni scenario il numero di campioni possibili che contengono lo stesso numero di mele rosse e gialle del campione osservato.
5. Si determini la plausibilità relativa di ogni scenario, ovvero la proporzione di campioni possibili compatibili con quello osservato.

Esercizio 3.9

La scala del QI è costruita in modo da avere, nella popolazione, media 100 e ds 15. Supponiamo di sapere che i soggetti affetti da una certa patologia P, che influisce sulle funzioni cognitive, ottengano, nella scala del QI, punteggi medi pari a 80 (ds = 15). Vogliamo calcolare la potenza del test statistico per rilevare la differenza tra le medie di due soggetti – uno normale ed uno patologico – estratti a caso delle due popolazioni, ipotizzando che siano entrambe normali. Suggerimento: per risolvere questo problema si veda l'esercizio 2.4.

1. Si definiscano le ipotesi H_0 e H_1 del test statistico.
2. Si rappresentino graficamente le distribuzioni teoriche delle differenze dei punteggi del QI sotto le due ipotesi.
3. Si determini il valore critico per rigettare H_0 al 5%.
4. Si determini graficamente la regione critica.
5. Si determini graficamente la potenza del test.
6. Si calcoli la potenza del test $(1 - \beta)$.
7. Si determini il numero di soggetti necessario per ottenere una potenza dell'80%. Suggerimento: si utilizzi la funzione `power.t.test()`.

Esercizio 3.10

Riconsideriamo il problema dell'esercizio 3.9 ma questa volta assumiamo che nella popolazione

di soggetti normodotati i punteggi del QI siano normali con media $\mu = 100$ e ds $\sigma = 30$, mentre nei soggetti con deficit i punteggi siano normali con media $\mu = 80$ e ds $\sigma = 30$. Ci aspettiamo che la potenza del test aumenti o diminuisca?

1. Si definiscano le ipotesi H_0 e H_1 del test statistico.
2. Si rappresentino graficamente le distribuzioni teoriche dei punteggi del QI sotto le due ipotesi.
3. Si determini il valore critico per rigettare H_0 al 5%.
4. Si determini graficamente la regione critica.
5. Si determini graficamente la potenza del test.
6. Si calcoli la potenza del test $(1 - \beta)$.
7. Si determini il numero di soggetti necessario per ottenere una potenza dell'80%. Suggerimento: si utilizzi la funzione `power.t.test()`.
8. Quali conclusioni si possono trarre confrontando questi risultati con quelli dell'esercizio 3.9?

Esercizio 3.11

Riconsideriamo l'esercizio 2.15; abbiamo due campioni di 500 soggetti adolescenti maschi e femmine, cui è stato somministrato il BIAI. I punteggi medi dei maschi sono risultati $\overline{x}_m = 53.64$ (con $\hat{\sigma}_m = 11.5$), i punteggi delle femmine $\overline{x}_f = 50.04$ (con $\hat{\sigma}_f = 8.5$). Vogliamo sapere se la differenza osservata tra le medie dei due campioni di soggetti sia statisticamente significativa.

1. Si formulino le ipotesi H_0 e H_1:
2. Si calcoli il valore della statistica test $(T0)$.
3. Sulla base del risultato ottenuto nell'esercizio 2.15, si rappresenti la distribuzione campionaria della statistica test sotto l'ipotesi nulla e la posizione del valore ottenuto $T0$.
4. Si calcoli il p-value del test, ovvero la probabilità di ottenere sotto H_0 un valore più estremo della statistica test.
5. Supponendo che, nella popolazione, la vera differenza tra le medie sia $\delta_{\mu_m - \mu_f} = 1.5$, si calcoli la potenza associata al test, fissando $\alpha = 0.05$.

Esercizio 3.12

Si ripeta l'esercizio 3.11 considerando che le stesse medie e deviazioni standard siano state ottenute su due campioni di numerosità 20.

1. Si calcoli il valore della statistica test $(T0)$ ed il suo errore standard.
2. Si rappresenti la distribuzione campionaria della statistica test sotto l'ipotesi nulla e la posizione del valore ottenuto $T0$.
3. Si calcoli il p-value del test, ovvero la probabilità di ottenere sotto H_0 un valore più estremo della statistica test.

4. Supponendo che, nella popolazione, la vera differenza tra le medie sia $\delta_{\mu_m - \mu_f} = 1.5$, si calcoli la potenza associata al test, fissando $\alpha = 0.05$.

Esercizio 3.13

Il data frame `kidiq` nel pacchetto `ADati` contiene un campione di 434 coppie madri-figli descritto negli esercizi 1.23 e 2.17. In particolare, 341 madri del campione sono diplomate mentre le altre 93 no. Vogliamo sapere se ci sono differenze dei punteggi dei figli in funzione del livello di istruzione delle madri.

1. Si renda disponibile nel workspace il data frame `kidiq`
2. Si calcolino media e deviazione standard dei punteggi dei figli e dei QI delle madri per tutto il campione e, successivamente, le stesse statistiche suddividendo il campione tra madri con diploma e madri senza diploma.
3. Si rappresentino graficamente con boxplot le distribuzioni dei punteggi di `kid_score` e `mom_iq` nei due gruppi (madri diplomate e non).
4. Si consideri la seguente contrapposizione di ipotesi per il test che valuti se ci siano differenze medie nei punteggi dei figli di madri con e senza diploma:
$$\begin{cases} H_0 : \mu_1 - \mu_0 = 0 \\ H_1 : \mu_1 - \mu_0 = 5 \end{cases}$$
in cui μ_0 indica il punteggio medio nella popolazione dei figli di madri non diplomate e μ_1 il punteggio medio nella popolazione dei figli di madri diplomate.
Si rappresenti graficamente la distribuzione campionaria della statistica test $T = \overline{x}_0 - \overline{x}_1$ sotto le ipotesi H_0 e H_1. Suggerimento: si tengano presenti i risultati ottenuti nell'esercizio 2.17.
5. Si calcoli il valore della statistica test e si rappresenti opportunamente nel grafico prodotto al punto precedente.
6. Si calcoli il p-value (unidirezionale) associato alla statistica test.
7. Si calcoli la potenza del test.

Esercizio 3.14

Riconsideriamo ancora lo strumento utilizzato nell'esercizio 2.15 e i risultati ottenuti degli esercizi 3.11 e 3.12. Supponendo che la differenza plausibile tra le medie dei punteggi di maschi e femmine, espressa con la statistica d di Cohen, sia pari a 0.3, vogliamo stimare la potenza e gli errori di tipo S e M nelle due situazioni, con campione grande (es. 3.11) e piccolo (es 3.12). Suggerimento: si utilizzi la funzione `retrospecitve()` del pacchetto `PRDA`.

I punteggi medi dei maschi sono risultati $\overline{x}_m = 53.64$ (con $\hat{\sigma}_m = 11.5$), i punteggi delle femmine $\overline{x}_f = 50.04$ (con $\hat{\sigma}_f = 8.5$). In pratica, medie e deviazioni standard dei punteggi sono uguali, ciò che cambia è la numerosità campionaria, 500 per gruppo nel primo caso e 20 per gruppo nel secondo.

1. Si calcoli il valore osservato di d utilizzando la seguente formula
$$d = \frac{\overline{x}_m - \overline{x}_f}{\sqrt{(\hat{\sigma}_m^2 + \hat{\sigma}_f^2)/2}}$$

2. Si calcolino e si interpretino la potenza e gli errori di tipo M e S nel campione di 500 soggetti.
3. Si calcolino e si interpretino la potenza e gli errori di tipo M e S nel campione di 20 soggetti.

Esercizio 3.15

Uno psicologo vuole effettuare uno studio per valutare l'efficacia di una tecnica per la riduzione dell'ansia negli adolescenti. A tal fine predispone un semplice studio pre/post: ad un campione di soggetti ansiosi verrà misurato, con un apposito strumento, il livello di ansia prima e dopo il training con la tecnica per la riduzione dell'ansia. Per stabilire quanti soggetti servano effettua una analisi di potenza definendo come effetto plausibile un valore di Cohen's $d = 0.2$. In altre parole, si attende che, se la tecnica funziona, i punteggi medi dei soggetti tra prima e dopo si riducano in media di 0.2 punti standard.

1. Si determini la statistica test che sarà utilizzata e si definiscano le ipotesi H_0 e H_1 del relativo test statistico.
2. Si determini, utilizzando la funzione `power.t.test()`, quanti soggetti servirebbero per avere una potenza dell'80%.
3. Si rappresenti graficamente la distribuzione campionaria della statistica test sotto H_0 e si determini il valore critico per rigettare tale ipotesi con $\alpha = 0.05$.
4. Si stimino gli errori di tipo S e M; suggerimento: si può utilizzare la funzione `prospective()` del pacchetto `PRDA`.
5. Si supponga che il campione effettivo utilizzato dallo psicologo sia di soli 30 soggetti. Quanto sarà in tal caso la potenza del test? E gli errori di tipo M e S?

Esercizio 3.16

Si supponga di estrarre un campione di 5 osservazioni – y_i, con $i = 1, \ldots, 5$ – da una certa distribuzione normale $\mathcal{N}(\mu, \sigma)$ e di calcolare su tale campione la media aritmetica delle osservazioni ($\overline{y}$). Vogliamo costruire la funzione di verosimiglianza della media associata al risultato osservato.

1. Sia il campione $y = \{1.4, 2.2, 1.2, 3.6, 2.3\}$; si calcoli la media aritmetica del campione.
2. Si calcoli la deviazione standard del campione.
3. Assumendo che la media della popolazione μ sia un valore compreso tra -5 e 5 e che la deviazione standard σ sia 1, si calcolino e si rappresentino graficamente i valori di verosimiglianza per il valore osservato della media.
4. Si determini il valore di massima verosimiglianza.
5. Si consideri ora un secondo campione di osservazioni $z = \{0.6, 1.9, 2.1, 2, 1.1\}$. Quale dei due campioni, tra y e z, è più probabile che provenga da una popolazione in cui la media $\mu = 2$ e la deviazione standard $\sigma = 1$? E di quanto?
6. Da quale delle seguenti popolazioni è più probabile che provengano i due campioni: $\mathcal{N}(0, 1)$, $\mathcal{N}(2, 10)$, $\mathcal{N}(1, 5)$.

4 REGRESSIONE LINEARE SEMPLICE

Esercizio 4.1

Per ciascuna coppia di variabili (x, y) generate come specificato di seguito:

a. si rappresentino le distribuzioni univariate (con boxplot) e un grafico a dispersione;

b. si calcoli la correlazione tra esse;

c. si valuti con un test opportuno se la differenza tra le media sia significativamente diversa da zero.

Al termine si confrontino i risultati e si traggano le conclusioni.

1. Siano: `y <- rnorm(50)` e `x1 <- y`.
2. Siano: `y` (generato al punto 1) e `x2 <- y+10`.
3. Siano: `y` (generato al punto 1) e `x3 <- rnorm(50)`.
4. Siano: `y` (generato al punto 1) e `x4 <- x3+10`.
5. Sulla base di questi risultati, quali conclusioni possiamo trarre, in particolare per quanto riguarda la relazione tra correlazione e differenza tra medie?

Esercizio 4.2

Analisi di dati simulati.

1. Si simulino 100 valori dal modello $y = \beta_0 + \beta_1 x + \epsilon$ con $\beta_0 = 4$ e $\beta_1 = 6$ campionando i valori di x da una distribuzione uniforme nel range $[0, 50]$ – si usi la funzione `runif()` – e gli errori da una normale $\mathcal{N}(0, 3)$.
2. Si stimino i parametri del modello sui dati simulati ed i relativi intervalli di credibilità al 90%.
3. Si rappresentino graficamente i dati e la retta di regressione.
4. Si valutino graficamente gli assunti del modello.

Esercizio 4.3

Stime dei parametri con un modello sbagliato.

1. Si simulino 100 valori dal modello $y = \beta_0 + \beta_1 x + \beta_2 x^2 \epsilon$ con $\beta_0 = 4$, $\beta_1 = 3$ e $\beta_2 = 7$ campionando i valori di x da una distribuzione uniforme nel range $[0, 50]$ – si usi la funzione `runif()` – e gli errori da una normale $\mathcal{N}(0, 3)$.
2. Si stimino i parametri del modello sui dati simulati ed i relativi intervalli di credibilità al 90% con il modello lineare `y` $\sim$ `x`.
3. Si rappresentino graficamente i dati e la retta di regressione.
4. Si valutino graficamente gli assunti del modello.

Esercizio 4.4

Nel file `autor.dat` sono riportati i punteggi di 12 soggetti su una scala di autoritarismo ed una di pregiudizi sociali (dati: Siegel & Castellan, 1992).
Ci si chiede se vi sia una relazione significativa tra il livello di autoritarismo e il livello di pregiudizio dei soggetti interpellati.

1. Si importi il file `autor.dat` in R assegnandogli nome `AP`.
2. Si identifichino unità statistiche e variabili del data-frame. Per ciascuna variabile si definiscano le proprietà metriche.
3. Si calcoli il campo di variazione dei punteggi di `autoritarismo` e `pregiudizio`.
4. Si determini l'80° percentile della variabile `autoritarismo`.
5. Si rappresenti graficamente la distribuzione dei punteggi di autoritarismo (in ascissa) e pregiudizio (in ordinata); sulla base del grafico ottenuto si ipotizzi un valore plausibile di correlazione.
6. Si calcolino covarianza e correlazione tra autoritarismo e pregiudizio e si interpretino.

Esercizio 4.5

Al campione di soggetti dell'esercizio 4.4 si aggiunge un nuovo soggetto che ottiene i seguenti punteggi: 55 nell'autoritarismo e 85 nel pregiudizio.

1. Si aggiunga il soggetto al data-frame `AP` (ottenuto al punto 1 dell'esercizio 4.4)
2. Si rappresenti nuovamente la distribuzione dei punteggi di autoritarismo e pregiudizio; sulla base del grafico ottenuto si ipotizzi un valore plausibile di correlazione.
3. Si calcolino covarianza e correlazione tra autoritarismo e pregiudizio e si interpretino.
4. Quali considerazioni si possono fare su questo ultimo soggetto, confrontando le stime di correlazione ottenute con quelle dell'esercizio precedente?

Esercizio 4.6

Il file `Hooker.dat` contiene dati raccolti da J.Hooker sulle montagne dell'Himalaya (cfr. Weisberg, 1985). Tali dati rappresentano le temperature in gradi Fahrenheit (variabile `temp`) di ebollizione dell'acqua a diversi valori di pressione atmosferica (mmhg; variabile `press`).

1. Si importi il file `Hooker.dat` in R.
2. Si identifichino unità statistiche e variabili del data-frame. Per ciascuna variabile si definiscano le proprietà metriche.
3. Si calcolino moda, mediana e media della variabile `temp`. Sulla base del risultato si valuti se la distribuzione di tale variabile possa considerarsi simmetrica.
4. Si valuti con un grafico opportuno la simmetria della distribuzione.
5. Si produca un grafico in quattro parti (utilizzando il comando `layout()` oppure `par()`) con i boxplot e qqplot per le variabili `press` e `temp`.

6. Si produca il diagramma di dispersione relativo alle variabili `press` e `temp` valutando se sia ipotizzabile una relazione lineare.
7. Si calcolino covarianza e correlazione tra le variabili `press` e `temp`.
8. Si stimino i parametri della retta di regressione.
9. Si valutino graficamente gli assunti del modello utilizzando i grafici dei residui ed individuando la presenza di eventuali casi anomali o influenti.
10. Si stimi la probabilità a posteriori che il valore di β_1 sia maggiore di 2.3.
11. Si aggiunga al grafico ottenuto al punto 6 la retta di regressione teorica.
12. Si stimi, sulla base dei parametri calcolati, il valore atteso di temperatura con una pressione di 27.

Esercizio 4.7

Supponiamo di avere misurato la statura di 10 bambini di età compresa tra 6 e 12 anni e di riportare i dati in tabella 1. Vogliamo studiare la relazione tra età e statura.

	Età (anni)	Statura (cm)
1	6	115
2	6	120
3	7	122
4	8	130
5	8	128
6	9	134
7	10	136
8	10	140
9	11	147
10	12	151

Tab. 1: Età e statura di un campione di 10 bambini.

1. Si identifichino unità statistiche e variabili del problema. Per ciascuna variabile si definiscano le proprietà metriche.
2. Si costruisca un data-frame con i dati della tabella 1 in R.
3. Si produca il diagramma di dispersione relativo alle due variabili valutando se sia ipotizzabile una relazione lineare.
4. Si calcolino covarianza e correlazione tra le variabili età e statura.
5. Si stimino i parametri della retta di regressione.
6. Si valutino graficamente gli assunti del modello utilizzando i grafici dei residui ed individuando la presenza di eventuali casi anomali o influenti.
7. Si determinino gli intervalli di credibilità del parametro β_1 al 70%, 80% e 90%.

8. Si aggiunga al grafico la retta di regressione teorica.

Esercizio 4.8

Ad un campione di 50 adolescenti affetti da una sindrome metabolica vengono misurate la concentrazione di globuli bianchi nel sangue (*white blood cell count*, `wbcc`) e l'indice di massa corporea (*body mass index*, `bmi`). I dati sono nel file `MetS.dat`. Si vuole sapere se la concentrazione di globuli bianchi sia predittiva del `bmi`.

1. Si importino i dati in R.
2. Si calcolino media, varianza e deviazione standard delle variabili `wbcc` e `bmi`.
3. Si produca un grafico a dispersione che rappresenti la distribuzione congiunta delle due variabili.
4. Si stimino i parametri del modello di regressione in cui la variabile `wbcc` è il predittore della variabile `bmi`.
5. Si aggiunga al grafico ottenuto al punto 3 la retta di regressione attesa.
6. Si stimi il valore atteso di indice di massa corporea quando la concentrazione di globuli bianchi è 8 e lo si rappresenti graficamente.
7. Si calcoli la dimensione dell'effetto (R^2).

Esercizio 4.9

Un campione di 13 soggetti con disturbi di memoria viene sottoposto ad una terapia per il recupero delle funzioni cognitive compromesse. Prima della terapia i soggetti vengono sottoposti al Mini Mental State Examination Test (variabile `MMSE1`). Al termine della terapia, ai soggetti viene risomministrato il test (variabile `MMSE3`). I dati sono riportati nel file `mmse.dat`. Si vuole sapere se la prima somministrazione possa essere considerata predittiva della terza.

1. Si importi il file `mmse.dat` in R.
2. Si identifichino unità statistiche e variabili del data-frame. Per ciascuna variabile si definiscano le proprietà metriche.
3. Si calcolino il campo di variazione e la semidifferenza interquartilica[11] per le variabili del data-frame.
4. Si produca la matrice di correlazione tra le variabili del data-frame.
5. Si produca un opportuno grafico che permetta di confrontare le distribuzioni dei punteggi delle tre variabili.
6. Si produca un grafico in tre parti (utilizzando il comando `layout()` oppure `par()`) con i qqplot per le tre variabili del data-frame e si valuti se ci sono potenziali outliers.
7. Si produca il diagramma di dispersione relativo alle variabili `MMSE1` e `MMSE3` valutando se sia ipotizzabile tra di esse una relazione lineare.

[11] Data una variabile casuale X: il campo di variazione è la differenza tra $max(X)$ e $min(X)$, la semidifferenza interquartilica è $(Q3 - Q1)/2$ in cui $Q1$ è il primo quartile e $Q3$ il terzo.

8. Si stimino i parametri della retta di regressione.
9. Si valutino graficamente gli assunti del modello utilizzando i grafici dei residui ed individuando la presenza di eventuali casi anomali o influenti.
10. Si stimi la probabilità a posteriori che il parametro β_1 sia compreso nell'intervallo $[0, 0.2]$.
11. Si aggiunga al grafico la retta di regressione teorica.
12. Si ripetano i punti 7, 8, 9, 10 e 11 eliminando il secondo soggetto (Per quale ragione potrebbe essere eliminato?).
13. Si confrontino i risultati della regressione con tutti i soggetti e della regressione senza il secondo soggetto.
14. Si determini un idoneo indicatore di effect size per ciascuno dei due modelli.
15. Quali considerazioni si possono trarre dal risultato ottenuto?

Esercizio 4.10

Il Sensation Seeking (SS) è un tratto di personalità che indica la tendenza alla ricerca di sensazioni intense e di esperienze rischiose. Un ricercatore utilizza, per misurare tale tratto, un test da cui si ottengono punteggi che variano da zero a trenta; maggiore è il punteggio di un soggetto, maggiore la sua proprensione al Sensation Seeking. Il test viene somministrato ad un campione di 8 soggetti di varie età, ottenendo i risultati riportati in tabella 2 (pag. 32).

	Età	SS
1	31	24
2	27	15
3	34	20
4	38	8
5	38	9
6	39	12
7	25	28
8	31	11

Tab. 2: Propensione al Sensation Seeking per età.

Si vuole sapere se esista una relazione credibile tra età e Sensation Seeking.

1. Si riportino opportunamente i dati della tabella 2 in R.
2. Si calcoli la correlazione tra le due variabili utilizzando il coefficiente di Pearson e quello di Spearman.
3. Si produca il diagramma di dispersione relativo alle due variabili valutando se sia ipotizzabile una relazione lineare.
4. Si stabilisca, con il modello opportuno, se l'età sia un predittore credibile (al 90%) del SS.
5. Si aggiunga al dataset un nuovo soggetto con età $= 37$ e SS $= 25$, e si ripeta il punto precedente. Quali differenze ci sono tra i due modelli?

Esercizio 4.11

Un ricercatore vuole sapere se vi siano differenze nell'atteggiamento verso l'attività extradomestica tra le donne sposate con figli e quelle senza figli. Allo scopo somministra una scala di atteggiamento ad un campione casuale di 72 donne coniugate, metà delle quali con figli e metà senza figli. I punteggi di atteggiamento ottenuti sono riportati nel file `extra.dat` (variabile `attegg`).

1. Si importi il file `extra.dat` in R.
2. Si produca un grafico in due parti (utilizzando il comando `layout()` oppure `par()`) con gli istogrammi delle distribuzioni dei punteggi di atteggiamento per i due gruppi di donne.
3. Si valuti, con un opportuno test grafico, se le distribuzioni dei punteggi nei due gruppi sono normali.
4. Si valuti se le varianze dei due gruppi possono considerarsi omogenee.
5. Si valuti, servendosi degli opportuni grafici, se vi siano potenziali outliers nei due gruppi.
6. Si stabilisca, con il modello opportuno, se vi siano differenze di atteggiamento credibili al 90% tra le donne con figli e quelle senza figli.
7. Si stimi la dimensione dell'effetto e la si interpreti.

Esercizio 4.12

Nel periodo dal 1900 al 2000, negli Stati Uniti, si sono registrati i valori di popolazione (in decine di milioni di abitanti) riportati in tabella 3.

Anno	Popolazione
1900	7.6
1910	9.2
1920	10.57
1930	12.28
1940	13.17
1950	15.11
1960	17.93
1970	20.32
1990	24.87
2000	28.14

Tab. 3: Popolazione degli Stati Uniti, in decine di milioni di abitanti, nel periodo 1900-2000.

Si vuole stimare, con le dovute assunzioni, la popolazione attuale e quella mancante del 1980.

1. Si riportino opportunamente i dati della tabella 3 in R.
2. Si produca il diagramma di dispersione relativo alle due variabili valutando se sia ipotizzabile una relazione lineare.
3. Si stimino i parametri della retta di regressione.
4. Si calcoli il valore di R^2 del modello.

5. Si aggiunga al grafico la retta di regressione teorica.

6. Si stimi, sulla base dei parametri calcolati, il valore atteso di popolazione nel 1980 e nel 2020 con relativi intervalli di incertezza al 68% ed al 95%.

Esercizio 4.13

Due campioni di 300 studenti di due diverse facoltà si sottopongono ad un test di memoria. I punteggi di tutti i soggetti (600) sono riportati nel file `studenti.dat` (variabile `punt`) con la relativa facoltà di appartenenza (variabile `fac`). Si vuole sapere se esista una differenza significativa i punteggi nei due gruppi.

1. Si importi il file `studenti.dat` in R.

2. Si calcolino i percentili dei punteggi al test separatamente nelle due facoltà; quindi si produca un grafico che abbia in ascissa i valori perentili calcolati ed in ordinata i punteggi al test (Suggerimento: per una migliore lettura del grafico si usino colori diversi per le due facoltà).

3. Si calcolino moda, mediana e media dei punteggi nelle due diverse facoltà. Quali considerazioni è possibile fare dal confronto tra tali statistiche?

4. Si calcolino semi differenza interquartilica, deviazione standard e varianza dei punteggi nelle due diverse facoltà. Quali considerazioni è possibile fare dal confronto tra tali statistiche?

5. Si produca un grafico in due parti (utilizzando il comando `layout()` oppure `par()`) con gli istogrammi delle distribuzioni dei punteggi al test di memoria dei due gruppi.

6. Si valuti, con opportuni grafici, se le distribuzioni dei punteggi nei due gruppi sono approssimativamente normali.

7. Si valuti se le varianze dei due gruppi sono omogenee.

8. Si stabilisca, con il modello opportuno, se vi siano differenze credibili al 90% tra i due gruppi nei punteggi al test di memoria.

9. Si determini la dimensione dell'effetto e la si interpreti.

10. Si stimi il *weight* del modello confrontandolo con il modello nullo.

Esercizio 4.14

Un ricercatore ritiene che possa esservi una differenza di genere nell'attitudine al pronto soccorso. Per valutare tale ipotesi somministra un questionario apposito ad un campione di 562 studenti (165 maschi e 397 femmine). I dati (punteggi di attitudine e genere dei soggetti) sono raccolti nel file `firstaid.dat`.

1. Si importino i dati in R.

2. Si calcolino media e deviazione standard dei punteggi separatamente per maschi e femmine.

3. Si produca un grafico delle medie dei punteggi con i relativi intervalli di confidenza (suggerimento: si può utilizzare la funzione `plotmeans()` del pacchetto `gplots`).

4. Si valuti l'assunto di normalità utilizzando gli opportuni grafici.
5. Si esegua il modello opportuno per valutare se vi siano differenze credibili al 70% tra maschi e femmine nei punteggi del questionario.
6. Si valuti graficamente l'assunto di omoschedasticità dei residui (si può utilizzare la funzione `residuals()` che calcola i residui di un modello lineare).
7. Si calcoli R^2 associato al genere.
8. Si stimi l'evidenza relativa rispetto al modello nullo utilizzando i model weights.
9. Sulla base dei risultati ottenuti si traggano le conclusioni.

Esercizio 4.15

Nel file `gestazioni.dat` sono raccolti i valori di gestazione in settimane (variabile `weeks`) di un campione di 328 donne che hanno partorito presso un ospedale. Per ciascuna donna è indicata l'età in anni compiuti al momento del parto (variabile `years`). Le stesse età sono poi raggruppate in due classi (variabile `age`): `25-30` e `40-47`. Si ipotizza che vi siano differenze nel numero medio di settimane di gestazione in relazione all'età.

1. Si importino i dati in R.
2. Si calcolino media e deviazione standard delle settimane di gestazione per le due fasce di età.
3. Si producano in un unico layout un grafico a boxplot che rappresenti le distribuzioni delle settimane di gestazione nelle due fasce di età e un grafico con le medie ed i relativi intervalli di confidenza.
4. Si valuti se le settimane di gestazione abbiano una distribuzione normale nei due gruppi di età utilizzando gli opportuni grafici.
5. Si esegua il modello opportuno per valutare se vi siano differenze credibili al 90% tra le due fasce di età nelle settimane di gestazione.
6. Si rappresenti graficamente la distribuzione a posteriori del parametro β_1 e si determini la probabilità che il valore del parametro si collochi nell'intervallo $[-0.5, 0]$.
7. Si calcoli il valore di R^2 per il test eseguito.
8. Si rappresenti graficamente la distribuzione delle età (in anni compiuti) e delle settimane di gestazione in modo da evidenziare l'eventuale relazione tra queste variabili.
9. Si valuti, con il modello opportuno, se la variabile `years` sia predittiva delle settimane di gestazione con credibilità del 90%.
10. Si calcoli il valore di R^2 del modello definito al punto 9 e lo si confronti con il valore di R^2 relativo al modello del punto 7. Quali conclusioni si possono trarre?
11. Si confrontino i modelli dei punti 5 e 9 utilizzando i *model weights*. Quali conclusioni si possono trarre da queste statistiche?

Esercizio 4.16

Si vuole studiare l'effetto della deprivazione da sonno in un compito attentivo. Tale compito consiste nell' individuazione di un oggetto in movimento su uno schermo radar. 16 soggetti vengono suddivisi in quattro gruppi, ciascuno dei quali viene privato del sonno per un certo numero di ore (4, 12, 20 e 28; variabile `hr`). I punteggi dei soggetti nel compito (variabile `score`) sono espressi nel numero di mancate individuazioni durante un periodo di 30 minuti e sono riportati nel file `anova1.dat` (dati Keppel, 1991). Si vuole sapere se esista una differenza tra i punteggi nel compito in relazione alle ore di deprivazione da sonno.

1. Si importi il file `anova1.dat` in R.
2. Si produca un grafico con le medie e le barre di errore dei punteggi per i quattro gruppi (a scelta si usi la funzione `errbarr` del pacchetto `Hmisc` oppure `plotmeans` del pacchetto `gplots`).
3. Si valuti, con uno o più grafici opportuni, se le distribuzioni dei punteggi nei quattro gruppi sono normali.
4. Si valuti se le varianze dei gruppi sono omogenee.
5. Si stabilisca, con il modello opportuno, se le ore di deprivazione da sonno siano predittive dei punteggi al compito con credibilità 95%.
6. Si valuti la bontà del modello lineare controllando graficamente gli assunti sui residui.
7. Si stimi l'intervallo di credibilità al 95% della dimensione dell'effetto (con R^2) e lo si interpreti.
8. Si stimi il weight del modello e lo si interpreti in relazione ai risultati ottenuti ai punti 5 e 7.

Esercizio 4.17

Ad un campione di 102 studenti vengono somministrate, a distanza di una settimana, due versioni del *Minnesota Multiphasic Personality Inventory* (MMPI): una versione cartacea (*paper-and-pencil*, `PAP`) ed una al computer (*computer-administered*, `CA`). Si vuole sapere se vi siano differenze tra i due tipi di somministrazione. I dati sono nel file `MMPI.dat`.

1. Si importino i dati in R.
2. Si calcolino media e deviazione standard dei punteggi al test con i due metodi di somministrazione.
3. Si utilizzi l'opportuna rappresentazione grafica per valutare se i punteggi nelle due modalità di somministrazione risultino associati.
4. Si calcolino covarianza e correlazione tra i punteggi PAP e CA.
5. Si stabilisca con l'opportuno modello se i punteggi in versione cartacea siano predittivi di quelli in versione computerizzata utilizzando gli intervalli di credibilità a 80% e 90%.
6. Si determini l'evidenza relativa di questo rispetto al modello nullo.
7. Si utilizzi una opportuna rappresentazione grafica per confrontare le distribuzioni dei punteggi nelle due somministrazioni.

8. Si stimi la probabilità (a posteriori) che la differenza tra le medie dei punteggi PAP e CA sia maggiore di 2.

Esercizio 4.18

Per valutare se bambini con esperienze sociali più ricche siano anche più capaci di parlare delle proprie esperienze relazionali viene predisposto un esperimento che coinvolge un campione di bambini di età intorno ai 5 anni. Durante le sessioni sperimentali i bambini vengono osservati ed al termine viene attribuito loro un livello di esperienza sociale in una scala a 5 categorie, definite sulla base delle azioni sociali agite e ricevute: presenti (alto numero di agite e ricevute), assenti (basso numero di agite e ricevute), cercatori (alto numero di agite e basso di ricevute), cercati (basso numero di agite e alto di ricevute) e medi. Per la valutazione della capacità di descrivere le esperienze relazionali viene calcolato un punteggio di competenza linguistica. I dati sono disponibili nel file `relbambini.dat`. Si vuole sapere se vi sia una relazione tra il comportamento sociale e le competenze linguistiche.

1. Si importino i dati in R.
2. Si calcolino le numerosità, medie e deviazioni standard dei punteggi di competenza linguistica (variabile `comp.ling`) per ciascuno dei 5 gruppi definiti dalla scala di comportamento sociale (variabile `cat.comp`).
3. Si produca il grafico delle medie di `comp.ling` con relativi intervalli di confidenza (suggerimento: si può utilizzare la funzione `plotmeans()` del pacchetto `gplots`).
4. Si utilizzino gli opportuni grafici per valutare se le distribuzioni di punteggi in ciascun gruppo di bambini siano approssimativamente normali.
5. Si individui il modello appropriato per valutare se esistano differenze tra le medie dei 5 gruppi e se ne stimino i parametri (quanti sono?).
6. Si utilizzino gli opportuni test grafici per valutare l'assunto di indipendenza e di omoschedasticità dei residui.
7. Si rappresentino graficamente le distribuzioni a posteriori dei parametri del modello.
8. Si determinino e si rappresentino graficamente le distribuzioni a posteriori dei valori attesi dei punteggi di competenze per ogni gruppo di bambini.
9. Si determini se è credibile una differenza al 90% tra il gruppo dei medi ed il gruppo dei presenti ed una differenza allo stesso livello di credibilità tra il gruppo dei cercatori e quello dei presenti.

Esercizio 4.19

Ad un campione di 15 bambini viene presentata una serie di problemi di ragionamento. I bambini sono divisi in tre gruppi di cinque soggetti ciascuno: ai bambini del primo gruppo viene dato un rinforzo positivo (`rinforzo` = +) ad ogni risposta corretta, ai soggetti del secondo gruppo viene dato un rinforzo negativo (`rinforzo` = -) ad ogni risposta errata, il terzo gruppo, di controllo, non riceve nessun tipo di rinforzo (`rinforzo` = 0). Nel file `anova4.dat` sono riportati i punteggi ottenuti dai bambini. Si vuole sapere se i diversi tipi di rinforzo abbiano effetto sulle prove di ragionamento.

1. Si importi il file `anova4.dat` in R.
2. Si rappresentino graficamente le distribuzioni dei punteggi dei tre gruppi separatamente (si può usare la funzione `boxplot`).
3. Si rappresentino graficamente le medie dei tre gruppi con i relativi intervalli di confidenza.
4. Si stimino i parametri del modello per valutare se il tipo di rinforzo sia predittivo dei punteggi nelle prove di ragionamento.
5. Si valuti graficamente se i residui del modello abbiano una distribuzione approssimativamente normale.
6. Si determini se c'è una differenza credibile al 90% tra la condizione di controllo e le altre due.
7. Si valuti la qualità delle previsioni utilizzando il *Posterior Predictive Check*; il modello funziona bene?

Esercizio 4.20

Nel dataset `radar` nel pacchetto `ADati` sono presenti gli stessi dati dell'esperimento descritto nell'esercizio 4.16 con l'aggiunta di otto soggetti.

1. Si renda disponibile il dataset `radar` in R.
2. Utilizzando la funzione `aggregate()`, si produca la tabella con le medie dei punteggi nei quattro gruppi.
3. Si rappresenti graficamente la distribuzione delle medie (con relativo intervallo di confidenza) in funzione delle ore di deprivazione da sonno.
4. Si valuti, con una o più rappresentazioni opportune, se le distribuzioni dei punteggi nei quattro gruppi siano approssimativamente normali.
5. Si valuti se le varianze dei gruppi sono omogenee.
6. Si stabilisca, con il modello opportuno, se vi sia una relazione credibile al 90% tra le ore di deprivazione ed il punteggio al compito.
7. Si valuti la bontà del modello analizzando opportunamente i residui.
8. Si stimi la dimensione dell'effetto e la si interpreti.
9. Si confrontino i risultati con quelli ottenuti nell'esercizio 4.16.

Esercizio 4.21

Un ricercatore vuole sapere se vi siano differenze nell'atteggiamento verso l'attività extradomestica tra le donne sposate con figli e quelle senza figli. Allo scopo somministra una scala di atteggiamento a due campioni casuali di donne coniugate, di cui $n_1 = 45$ con figli e $n_2 = 36$ senza figli (i dati sono riportati nel file `donne.dat`). Si vuole sapere se vi siano differenze di atteggiamento tra le donne con figli e quelle senza.

1. Si importi il file `donne.dat` in R.

2. Si calcolino i quartili della variabile atteggiamento separatamente per i due gruppi.
3. Si produca un grafico a scatole con le distribuzioni dei punteggi di atteggiamento per le donne con figli e senza separatamente.
4. Si valuti se le varianze dei punteggi nei due gruppi siano omogenee.
5. Si valuti, con il modello opportuno, se vi siano differenze credibili al 90% tra i due gruppi in relazione ai punteggi di atteggiamento.
6. Si stimi l'Effect Size.
7. Si calcoli l'evidenza relativa del modello rispetto al modello nullo interpretandolo e confrontando il risultato con quello ottenuto al punto 5.
8. Si stimi la probabilità a posteriori che il punteggio tra due donne estratte a caso da ciascuno dei due gruppi non differisca di più di cinque punti.

Esercizio 4.22

Nel file `orevoti.csv` sono riportate le ore di frequenza ed il voto di 37 studenti partecipanti ad un corso di statistica. Vogliamo sapere se le ore di frequenza siano predittive del voto.

1. Si importi il file `orevoti.csv` in R.
2. Si produca il grafico a dispersione dei voti in funzione delle ore di frequenza.
3. Si calcolino covarianza e correlazione tra le due variabili.
4. Si determini, con il modello statistico opportuno, se le ore di frequenza sono predittive del voto con credibilità del 90%.
5. Si individui un appropriato indice di effect size e lo si interpreti.
6. Si aggiunga al grafico prodotto al punto 2 la retta di regressione teorica.
7. Si determinino devianza di regressione e devianza residua del modello.
8. Si rappresenti graficamente, la distribuzione a posteriori del parametro β_1 del modello definito al punto 4 e si stimi la probabilità che tale parametro sia inferiore a 0.25.

Esercizio 4.23

Ad un campione di 20 soggetti di varie età con un particolare deficit di memoria viene somministrato un test per la valutazione delle capacità di memoria, più sono alti i punteggi migliori sono le prestazioni (i dati sono raccolti nel file `memory.dat`). Vogliamo sapere se l'età sia predittiva del deficit in questione.

1. Si importi il file `memory.dat` in R.
2. Si produca un grafico in due parti con gli istogrammi della distribuzione dei punteggi e delle età dei soggetti.
3. Si valuti, con un grafico opportuno, se la distribuzione dei punteggi si possa considerare normale.

4. Si produca un grafico a dispersione con i punteggi in funzione dell'età. Dalla lettura di questo grafico è ragionevole ipotizzare una relazione lineare tra le due variabili?

5. Si produca un grafico che riporti in ascissa i percentili della variabile età ed in ordinata i percentili dei punteggi. Dalla lettura di questo grafico è ragionevole ipotizzare una relazione lineare tra le due variabili?

6. Si valuti, con un modello opportuno, se l'età sia predittiva dei punteggi al test con credibilità del 95%.

Esercizio 4.24

Il dataset `impiegati` del pacchetto `ADati` contiene dati relativi ad un campione di 474 impiegati di banca. Le variabili considerate sono le seguenti: `id` = codice identificativo dell'impiegato; `sesso` = sesso dell'impiegato; `istruz` = livello di istruzione (in anni di studio); `catlav` = categoria lavorativa; `stipatt` = stipendio percepito; `stipiniz` = stipendio iniziale percepito al momento dell'assunzione.

1. Si renda disponibile in R il dataset `impiegati`.

2. Si producano le opportune statistiche per descrivere le variabili presenti nel data-set.

3. Si valuti con il modello opportuno, se è credibile al 90% che il campione si possa considerare estratto da una popolazione (di impiegati di banca) che percepisce come stipendio medio una somma pari a 37.000.

4. Si valuti, con il modello opportuno, se vi siano differenze credibili al 90% tra maschi e femmine in relazione allo stipendio percepito.

5. Si calcoli la dimensione dell'effetto nel test eseguito al punto precedente e la si interpreti.

6. Si valuti, con il modello opportuno, se lo stipendio iniziale al momento dell'assunzione sia predittivo di quello percepito a livello di credibilità del 90%.

7. Si valuti se esistano differenze credibili al 90% tra le medie degli stipendi percepiti in funzione della categoria lavorativa, controllando opportunamente gli assunti per il modello.

8. Si calcoli un indice di effect size per il modello eseguito al punto precedente e lo si interpreti.

9. Si confrontino i modellidei punti 4, 6 e 7 utilizzando Δ_{LOO} e *model weigths*. Quale tra i predittori considerati dello stipendio attuale è il più plausibile?

Esercizio 4.25

Il data set `Gini`, nel pacchetto `ADati` contiene alcuni indicatori socio-economici relativi alle 20 regioni italiane, rilevati per uno studio sulle disuguaglianze sociali in Italia. In particolare, per ciascuna regione, è riportato l'indice di Gini (variabile `gini`), che misura la disuguaglianza nella distribuzione del reddito[12], e la speranza di vita alla nascita (variabile `life`). Si vuole sapere se l'indice di Gini sia predittivo della speranza di vita.

1. Si carichi il data set in R.

[12] L'indice varia tra 0, massima omogeneità, e 1, massima concentrazione.

2. Si rappresentino graficamente le distribuzioni univariate delle variabili `gini` e `life`.
3. Si rappresenti la distribuzione della variabile `life` in funzione della variabile `gini`. Sulla base del grafico è possibile ipotizzare una relazione lineare tra le due variabili?
4. Si stimino i parametri del modello lineare che prevede la speranza di vita alla nascita in funzione dell'indice di Gini.
5. Si valutino graficamente gli assunti del modello.
6. Si determini la regione che presenta lo scarto più ampio dal valore atteso.
7. Si calcolino la devianza totale, la devianza residua e la devianza di regressione del modello.
8. Si confrontino graficamente le previsioni del modello con i valori osservati realmente. Quale valutazione possiamo dare del modello lineare?

Esercizio 4.26

Il data frame `kidiq` nel pacchetto `ADati` contiene un campione di 434 coppie madri-figli descritto nell'esercizio 1.23. Vogliamo individuare, tra le variabili delle madri, quale sia il migliore predittore dei punteggi dei figli (variabile `kid_score`).

1. Si renda disponibile nel workspace il data frame `kidiq`
2. Si rappresentino graficamente, nei modi opportuni, i punteggi dei figli in funzione di ciascuna delle altre variabili del data set (`mom_hs, mom_iq, mom_work, mom_age`). Nota: si devono ottenere in pratica 4 grafici.
3. Si rappresentino graficamente le previsioni del modello nullo che stima i punteggi dei figli nella popolazione e si confrontino con la distribuzione attesa teorica.
4. Si stimino i parametri dei 4 modelli lineari associati ai grafici ottenuti al punto 2.
5. Si calcolino i LOO ed i weights relativi ai modelli ottenuti nel punto precedente, includendo anche il modello nullo.
6. Quale, tra le quattro variabili delle madri, è la migliore per prevedere i punteggi dei figli?

Esercizio 4.27

Nel dataset `inibition` sono raccolti i dati di una ricerca sui meccanismi di inibizione cognitiva rilevati su un campione di 150 soggetti di età compresa tra 8 e 81 anni. L'età è stata codificata in due modi diversi: in anni compiuti (variabile `Age`) ed in classi di età (variabile `Ageclass`); i punteggi di inibizione sono riportati nella variabile `INI`. Si vuole studiare la relazione tra età ed inibizione.

1. Si renda disponibile il dataset `inibition`
2. Si producano le statistiche descrittive di base del data frame.
3. Si rappresentino graficamente, in forma univariata, le due variabili dell'età ed i punteggi di inibizione.

4. Si scriva il modello nullo per la previsione dei punteggi di inibizione e se ne stimino i parametri.

5. Si simulino le previsioni del modello e si confrontino graficamente con i valori osservati dei punteggi di inibizione.

6. Si aggiunga al modello, come predittore, la variabile `Age` (età in anni compiuti) e si stimino ed interpretino i parametri .

7. Si valutino gli assunti del modello al punto precedente.

8. Si stimino e si interpretino i parametri del modello di regressione che prevede i punteggi di inibizione in funzione dell'età in classi (variabile `Ageclass`). Quali sono le differenze rispetto al modello del punto 6?

9. Si valutino gli assunti di questo modello e si confrontino con quelli del modello precedente, visti al punto 7.

Esercizio 4.28

Si ripeta l'esercizio 4.2 utilizzando le seguenti dimensioni campionarie: 10, 20, 30, 100, 300, 500, 1000, 2000, 5000. In particolare, per ciascuna dimensione si ricavino la stima di β_1 con la relativa incertezza al 90% e si rappresentino graficamente in funzione di N.

Esercizio 4.29

Analisi di dati simulati.

1. Si simulino 100 valori dal modello $y = 2 + 4x + \epsilon$ campionando i valori di x da una distribuzione uniforme nel range $[0, 30]$ – si usi la funzione `runif()` – e gli errori da una normale $\mathcal{N}(0, 5)$.

2. Si stimino i parametri del modello sui dati simulati.

3. Si rappresentino graficamente i dati e la retta di regressione.

4. Si valuti se le stime ottenute siano ragionevolmente vicine ai valori veri.

5. Si ripetano i punti 1 e 2 per 1000 volte e si valuti: 1) se le deviazioni standard delle distribuzioni campionarie dei parametri sono approssimativamente uguali agli errori standard stimati al punto 1; 2) se approssimativamente il 95% degli intervalli di credibilità contiene il valore vero dei parametri.

5 REGRESSIONE LINEARE MULTIPLA

Esercizio 5.1

Consideriamo un esperimento sulle capacità di apprendimento delle scimmie in relazione al tempo di deprivazione da cibo e l'assunzione di determinati farmaci. 24 animali vengono sottoposti ad una serie di problemi di identificazione di oggetti e vengono ricompensati con del cibo quando rispondono correttamente. La V.D. è rappresentata dal numero di risposte corrette su 20 tentativi (variabile `score`). I predittori sono il tipo di farmaco (x, y e nessuno (c); variabile `drug`) e il tempo di deprivazione da cibo (1 e 24 ore; variabile `fdep`). I dati sono contenuti nel dataset `monkeys` nel pacchetto `ADati` (Keppel, 1991). Si vuole sapere se esista una differenza nel numero di risposte corrette in funzione del farmaco assunto, del periodo di deprivazione da cibo e se vi sia un'interazione tra le due variabili.

1. Si renda disponibile il data frame `monkeys`.
2. Si calcolino media e deviazione standard del numero di risposte corrette in relazione alla variabile `drug`, alla variabile `fdep` e all'interazione tra le due (si può usare la funzione `aggregate`).
3. Si rappresentino graficamente le medie delle risposte corrette in funzione del tipo di farmaco (`drug`) e tempo di deprivazione (`fdep`) inserendo anche le barre di errore (si può utilizzare la funzione `errbarr()`, pacchetto `Hmisc`).
4. Si valuti se le varianze dei 6 gruppi sono omogenee (quali sono i sei gruppi?).
5. Si stimino, con il modello opportuno, i coefficienti delle variabili `drug`, `fdep` e dell'interazione tra di esse e si interpretino.
6. Si determini, utilizzando il LOO, se il modello con interazione sia meglio di quello additivo (senza interazione).
7. Si stimino le evidenze relative di ciascun effetto – drug e fdep separati, in forma additiva e interazione – utilizzando opportunamente i *model weights*.

Esercizio 5.2

Il file `testingresso.sav` contiene dei dati relativi ad un campione di studenti universitari che hanno effettuato un test di ingresso. Per ciascuno studente sono riportati la matricola (variabile `matricol`), l'anno di corso (`an_corso`), il sesso (`sex`), il tipo di diploma di scuola superiore (`codmat`), il voto di diploma(`votodip`), il voto medio ottenuto negli esami sostenuto (`media`), il numero di esami sostenuti (`n_es`) ed il punteggio al test di ingresso (`tottest`). Si vuole sapere se il test di ingresso possa essere considerato predittivo della prestazione degli studenti in termini di numero e media degli esami.

1. Si importi il file `testingresso.sav` in R.
2. Si eliminino dal data-frame i soggetti con zero esami.
3. Si produca un grafico in quattro parti con la rappresentazione delle medie, con relativo intervallo di confidenza, suddivise per anno di corso e genere, delle seguenti variabili: `votodip`, `media`, `n_es` e `tottest` (Suggerimento: si può utilizzare la funzione `plotmeans()` del pacchetto `gplots`).

4. Si crei una variabile `prestazione` ottenuta con la seguente formula

$$\texttt{prestazione} = \frac{\texttt{media} \times \texttt{n\_es}}{30 \times max(\texttt{n\_es})}$$

5. Si calcolino i decili della variabile creata al punto precedente.
6. Si produca un grafico in tre parti con istogramma, boxplot e qqplot della variabile `prestazione`.
7. Si producano la matrice di covarianza e di correlazione tra le variabili `votodip`, `tottest` e `prestazione`.
8. Si produca un grafico in due parti con la distribuzione dei punteggi della variabile `prestazione` in funzione delle variabili `votodip` e `tottest`.
9. Si determini, con un opportuno modello, se le variabili `votodip` e `tottest` siano predittive della variabile `prestazione`.
10. Si valuti la bontà del modello analizzando opportunamente i residui.

Esercizio 5.3

Nel file `metodo.sav` sono riportati i punteggi ottenuti da 20 studenti in una prova sostenuta alla fine di un corso. Gli studenti sono stati suddivisi in quattro gruppi, in ciascuno dei quali il corso è stato impostato diversamente nel metodo (lezione o discussione) e nella durata delle sessioni (30 o 50 minuti). Si vuole sapere se il metodo e la durata hanno influito sui punteggi ottenuti dagli studenti nella prova di fine corso.

1. Si importi il file `metodo.sav` in R.
2. Si calcolino la media generale e la varianza della variabile `punteggi`.
3. Si calcolino le quattro medie e relative varianze della variabile `punteggi` in funzione delle variabili `metodo` e `durata` e si riportino opportunamente in una tabella.
4. Si produca il grafico delle medie, con relativi intervalli di confidenza, della variabile `punteggi` in funzione delle variabili `metodo` e `durata`.
5. Determinare, con un opportuno modello, se vi sia una influenza credibile delle variabili `metodo` e `durata` sulla variabile `punteggi`.
6. Si determini l'evidenza relativa dell'interazione tra le variabili `metodo` e `durata`.

Esercizio 5.4

In uno studio sull'utilizzo di un particolare servizio sociale disponibile su internet in relazione al grado di estroversione sono selezionati 100 studenti, 50 dei quali sono membri di un gruppo on-line e 50 no (variabile `member`). Ai soggetti viene somministrato l'EPQ (*Eysenck personality questionnaire*) e, sulla base dei punteggi ottenuti, essi vengono suddivisi in due gruppi: introversi ed estroversi (variabile `trait`). Il grado di utilizzo del servizio sociale (variabile `use`) viene valutato con un questionario composto di 10 item, il cui punteggio finale varia da 0 (nessun uso del servizio) a 7 (massimo uso del servizio). I dati sono disponibili nel file `eysenck.dat`. Si vuole sapere se il fare parte del gruppo ed il grado di estroversione influenzano l'uso del servizio sociale.

1. Si importino i dati in R.
2. Per ciascuno dei quattro gruppi, definiti dall'incrocio tra le variabili `trait` e `member` si calcolino, medie, deviazioni standard, numerosità ed intervallo di confidenza delle medie al 95%.
3. Si valuti graficamente se sia ipotizzabile una interazione delle variabili `member` e `trait` sulla variabile `use`.
4. Si definisca il modello appropriato per valutare l'interazione tra le variabili `trait` e `member`.
5. Si calcoli l'evidenza relativa dell'interazione rispetto al modello senza interazione.
6. Sulla base dei risultati ottenuti si traggano le conclusioni.

Esercizio 5.5

Un campione di 405 modelli di auto, prodotte tra il 1970 ed il 1982, viene analizzato per una valutazione comparativa. Per ciascun modello vengono riportate alcune informazioni relative al *consumo in km al litro*, *cilindrata*, *potenza in cavalli*, *peso* e *accelerazione*, più il *paese* e l'*anno di produzione*. I dati sono nel file `auto.sav`.

1. Si importi il file `auto.sav` in R.
2. Si calcoli il consumo medio (variabile `consumo`), con la relativa deviazione standard, in funzione dell'anno di produzione (`anno`) e della nazione di produzione (`origine`).
3. Si rappresentino graficamente i consumi medi, in funzione dell'anno di produzione, differenziandoli per nazione di produzione.
4. Si valuti, con gli opportuni grafici, se le distribuzioni dei consumi nelle nazioni di produzione possono essere considerate normali.
5. Si produca un grafico in quattro parti che presenti i grafici a dispersione della variabile `consumo` in funzione delle variabili cilindrata (variabile `motore`), potenza (`cv`), peso (`peso`) e numero di cilindri (`cilindri`).
6. Si determini quali tra le variabili cilindrata (variabile `motore`), potenza (`cv`), peso (`peso`) e numero di cilindri (`cilindri`) siano predittive della variabile `consumo` (ovvero quali β relativi siano credibili al 90%).
7. Si determini con il confronto tra modelli quale sia il predittore più evidente della variabile `consumo`.
8. Si utilizzi la variabile individuata al punto precedente per stimare i consumi dei modelli di auto in cui il dato è mancante.

Esercizio 5.6

Nel dataset `SNA1` del pacchetto `ADati` sono contenuti i dati relativi ad un esperimento condotto su 76 bambini di età compresa tra i 3 ed i 6 anni. Ciascun bambino è stato classificato in base alla modalità con cui utilizza l'associazione numeri-spazio (Spatial-Numeric Association, SNA). Con questo criterio sono stati identificati tre gruppi (definiti nella variabile `SNA`): soggetti SNA1, ovvero i bambini che mostrano un'associazione numeri-spazio compatibile con la direzione di

lettura-scrittura, SNA2, ovvero i bambini con associazione inversa e non-SNA, ovvero bambini che non mostrano un'associazione stabile. Ai bambini viene mostrata una sequenza di numeri target da 1 a 9 ed essi devono riportare su una linea di 10 cm la posizione del numero. Nel file sono presenti le variabili che indicano la differenza in cm. tra la posizione indicata e la vera posizione dei numeri, e la media di tali scarti (variabile `media.scarti`). La variabile `gruppo.eta` contiene le età dei bambini raggruppate in 6 fasce.

1. Si rendano disponibili i dati in R.
2. Si producano le statistiche descrittive per le variabili del data-frame.
3. Volendo sapere se l'età dei soggetti e il tipo di SNA siano predittive degli scarti commessi dai bambini si definisca il modello opportuno e si stimino i relativi parametri.
4. Si valuti se vi sia un'interazione credibile tra i predittori del modello definito al punto precedente.
5. Si considerino ora le singole valutazioni ottenute per i nove numeri target e si individui un grafico opportuno per visualizzare le distribuzioni degli scarti indicati dai bambini rispetto a ciascun numero target.
6. Si valuti, sulla base del grafico prodotto al punto 5, se sia ipotizzabile una relazione lineare tra la grandezza del numero target e lo scarto indicato. In caso positivo, si stimino i parametri del modello che predice lo scarto sulla base del numero target.

Esercizio 5.7

Il dataset `mathschool` nel pacchetto `ADati` contiene i punteggi di fine anno al test di matematica rilevati in 7 classi di una scuola nella successione dei 5 anni scolastici utilizzati per valutare l'apprendimento della matematica. Nel set di dati ci sono le seguenti variabili: `subj`, codice identificativo dello studente; `classe`, codice identificativo della classe; `docente`, sigla docente della classe; `anno`, anno di rilevazione; `math`, punteggio finale di matematica.

1. Si rendano disponibili i dati in R.
2. Si determini il numero di classi assegnate a ciascun docente.
3. Si determini il numero di studenti assegnati a ciascun docente.
4. Si rappresentino graficamente i punteggi al test di matematica in funzione dell'anno scolastico e si interpreti il risultato ottenuto.
5. Si formuli un modello (lineare) per valutare i punteggi nel test di matematica durante i cinque anni di scuola e se ne stimino i parametri.
6. Si aggiunga al modello precedente l'effetto del docente e si stabilisca se tale effetto risulti credibile al 90%.
7. Si aggiunga al modello precedente anche l'effetto di interazione e lo si valuti utilizzando l'evidenza relativa.
8. Si utilizzi il *model weight* per individuare il migliore tra i modelli definiti ai punti 5, 6 e 7.
9. Si rappresentino graficamente gli effetti del modello migliore tra i tre valutati e si interpretino i risultati.

Esercizio 5.8

Il dataset `vaes2015` nel pacchetto `ADati` contiene un campione di 200 soggetti coinvolti in una ricerca sul pregiudizio verso gli immigrati. Nel dataset sono presenti le seguenti variabili:

- `PREGIUDIZIO` = livello di pregiudizio verso gli extracomunitari, punteggi su una scala tra 0 e 7 in cui più alti sono i punteggi più è alto il livello di pregiudizio
- `NORME` = livello di accordo espresso dai partecipanti in relazione alla presenza di norme anti discriminazione (es. Le persone che esprimono atteggiamenti offensivi verso gli immigrati devono essere perseguite legalmente),
- `PAURACRI` = livello di paura della criminalità percepita,
- `STIME` = stima soggettiva della percentuale di immigrati che commettono crimini,
- `CONTATTO` = livello di contatto percepito con gli immigrati,
- `AGENTI_SOC` = livello di pregiudizio delle persone importanti (partner, genitori, amici stretti, etc.),
- `Or_Politico` = orientamento politico, espresso su una scala da 1 (estrema sinistra) a 16 (estrema destra),
- `GIORNALIsx` = frequenza di lettura giornali con orientamento di sinistra,
- `GIORNALIdx` = frequenza di lettura giornali con orientamento di destra,
- `TG` = frequenza di esposizione ai telegiornali.

Si vuole individuare quali siano i predittori del livello di pregiudizio.

1. Si importino i dati in R.
2. Si rappresentino graficamente le distribuzioni univariate delle variabili del dataset.
3. Si rappresentino le distribuzioni bivariate delle variabili (si può utilizzare la funzione `pairs()`).
4. Si definisca un modello lineare per individuare quali siano i predittori del livello di pregiudizio e se ne stimino i parametri (quanti sono?).
5. Si valutino con le opportune analisi grafiche gli assunti del modello.
6. Si individuino i predittori credibili al 90%.
7. Si calcolino i *model weight* associati ai predittori del modello.
8. Considerando i risultati dei punti 6 e 7 si formulino i seguenti tre modelli alternativi: 1) togliendo i predittori non credibili, 2) togliendo i predittori con $\beta < 0$, 3) togliendo i quattro predittori con weight più basso. Si stimino i parametri di questi modelli e si confrontino utilizzando un criterio di informazione.
9. Si stimi l'evidenza relativa del modello migliore individuato al punto precedente rispetto al modello con tutti i predittori.

Esercizio 5.9

In una ricerca sulla depressione nei preadolescenti viene selezionato un campione di 676 studenti di scuola media e primi anni di superiore. A ciascun soggetto viene chiesto di esprimere su una scala a 5 punti il proprio grado di preferenza verso la musica gotica (variabile `goth`, più è alto il valore maggiore la preferenza), e viene somministrato un test per valutare il livello di depressione con una misura che varia tra -2 e 2 (variabile `dep`, più è alto il valore maggiore il livello di depressione). Il dataframe `gothic` nel pacchetto `ADati` contiene i dati rilevati su queste variabili con l'aggiunta delle informationi relative a genere (variabile `gender`) ed età (variabile `age`) dei soggetti.

Adottando un approccio di confronto tra i modelli, si vogliono individuare i migliori predittori della depressione.

1. Si rendano disponibili i dati in R.
2. Si producano le statistiche descrittive delle variabili nel dataset utilizzando la funzione `describe()` del pacchetto `psych`.
3. Si rappresentino graficamente le distribuzioni univariate delle variabili del dataset, ciascuna nel modo più opportuno.
4. Si definisca il modello nullo della depressione e se ne stimino i parametri.
5. Si definiscano i modelli per la previsione della depressione con un singolo predittore alla volta – `goth`, `gender` e `age` –, e se ne stimino i parametri.
6. Si definiscano i modelli per la previsione della depressione con due predittori alla volta, senza interazioni, e se ne stimino i parametri.
7. Si aggiungano ai modelli del punto precedente le relative interazioni.
8. Si calcolino i valori di LOO di tutti i modelli dei modelli precedenti e si confrontino per individuare il modello migliore.
9. Si individui il modello migliore utilizzando i *model weights*.
10. Si analizzino e si interpretino i parametri del modello migliore ottenuto
11. Si rappresentino graficamente e si interpretino gli effetti del modello migliore.

Esercizio 5.10

In una ricerca sul ruolo dell'attaccamento e dell'alleanza tra genitori come predittori dello stress in genitori adottivi, viene intervistato un gruppo di 40 coppie con un figlio adottato. Nel dataset `parenting` del pacchetto `ADati` sono raccolti i dati relativi a 5 variabili che misurano: lo stress dei genitori relativo a tre domini (*Parent Distress*, PD; *Parent/Child Disfunctional interaction*, PCD; *Difficulty Child*, CD), l'attaccamento irrisolto (U, in cui il valore 1 indica questa caratteristica del genitore) e l'alleanza tra genitori (*Parenting Alliance Measure*, PAM). Per ciascuna variabile è presente il valore ottenuto sulle madri (`.m`) e sui padri (`.p`).

Si vuole individuare il modello che spiega meglio il livello di stress.

1. Si importino i dati in R.
2. Si producano le statistiche descrittive delle variabili del dataset utilizzando la funzione `describe()` del pacchetto `psych`.

3. Si modifichi il data frame in modo da ottenere una riga per ogni soggetto e cinque colonne (relative alle variabili PD, PCD, CD, U, PAM) più una nuova colonna che indichi il genitore. Suggerimento: si creino prima due data frame separati per madri e padri e poi li si unisca con `rbind()`.
4. Si calcoli un punteggio totale di stress (`tot.stress`) sommando le variabili `PD`, `PCD` e `CD`.
5. Si ispezionino con gli opportuni grafici le variabili `U`, `PAM`, `genitore` (creata al punto 3) ed il punteggio totale di stress appena calcolato.
6. Si ispezionino graficamente, con le opportune rappresentazioni, le relazioni bivariate tra le quattro variabili del punto precedente.
7. Si definiscano i modelli (senza effetti di interazione) per la previsione dello stress in funzione dei tre predittori `U`, `PAM`, `genitore` e si stimino i parametri.
8. Si individui il modello migliore tra quelli valutati utilizzando il LOO e se ne calcoli l'evidenza relativa verso il modello nullo.
9. Si stimino i parametri anche dei modelli con le interazioni a due e tre vie.
10. Si utilizzi il LOO per confrontare questi nuovi modelli con il modello migliore ottenuto prima e si valuti quantitativamente l'impatto delle interazioni su questo modello.
11. Si ripeta il confronto tra modelli utilizzando il *model weight*. Il modello migliore finale è lo stesso?
12. Si analizzino i residui del modello che include solo PAM come predittore.
13. Si rappresentino graficamente le previsioni del modello e si interpretino.

Esercizio 5.11

Il data set `trust` nel pacchetto `ADati` contiene i dati relativi ad una ricerca condotta in ambito europeo sui comportamenti sociali in funzione della fiducia che gli individui pongono sulle proprie istituzioni, sulla scienza etc. Il campione si compone di 5000 soggetti, di età media 34.3 anni (sd 7.9) reclutati in 23 paesi europei. Le variabili `X1`, `X2`, `X3`, `X4` sono relative a misure di fiducia espresse verso le proprie istituzioni politiche, sociali, scolastiche e scientifiche, punteggi più alti indicano un maggiore grado di fiducia. La variabile `Y` indica il grado in cui si è disponibili a rispettare le regole, anche in questo caso maggiore è il punteggio è maggiore l'adesione al rispetto.

1. Si importino i dati in R
2. Si producano le statistiche descrittive delle variabili nel dataset.
3. Si rappresentino graficamente, nel modo opportuno, le distribuzioni univariate delle variabili del dataset.
4. Si produca la matrice di correlazione tra le variabili del dataset per cui abbia senso.
5. Si rappresentino le distribuzioni dei punteggi nelle variabili `X1`, `X2`, `X3`, `X4` e `Y` suddivise nei 23 paesi.
6. Si valuti, separatamente per le variabili `X1`, `X2`, `X3`, `X4` e `Y`, l'ipotesi che la variabilità dei punteggi sia la stessa nei 23 paesi e si valutino i risultati confrontandoli con i grafici prodotti al punto precedente.

7. Si individui, con LOO e model weights, quale, tra i predittori `X1`, `X2`, `X3`, `X4`, abbia la maggiore evidenza nella capacità predittiva delle dipendente `Y`. Tale modello sarà utilizzato nei punti successivi.
8. Si analizzino i residui del modello migliore tra quelli definiti al punto precedente.
9. Si simuli un campione di osservazioni predette dallo stesso modello e si confrontino con la distribuzione dei valori osservati. Suggerimento: si può utilizzare la funzione `model.predictions()` del pacchetto `ADati`.
10. Si calcoli R^2 con intervallo di credibilità al 90% del modello e si commenti.

Esercizio 5.12

Si riprenda il data set `gambling` descritto nell'esercizio 1.26. Si vuole studiare, utilizzando le variabili a disposizione, quali siano i predittori della frequenza di gioco.

1. Si renda disponibile il data set in R.
2. Si produca la matrice di correlazione tra le variabili quantitative del dataset.
3. Si valuti graficamente se ci siano differenze di genere nelle distribuzioni delle variabili quantitative.
4. Si valuti graficamente se ci siano differenze tra le scuole nelle distribuzioni delle variabili quantitative.
5. Si determini, quale tra le variabili disponibili nel dataset presa singolarmente abbia la migliore capacità predittiva della frequenza di gioco.
6. Si stimino i parametri del modello che valuta se la relazione tra grado di disapprovazione (`disapproval`) e frequenza di gioco (`frequency`) sia diversa tra maschi e femmine.
7. Si interpreti graficamente il modello.
8. Si stimi e si interpreti l'R^2 del modello con relativo intervallo di credibilità al 90%.
9. Si produca il grafico con le previsioni del modello a confronto con i dati osservati.
10. Si confronti il modello con i due modelli che includono i due predittori separatamente ottenuti al punto 5.

6 REGRESSIONE LINEARE MULTILIVELLO

Esercizio 6.1

In relazione ai dati dell'esercizio 1.16 si vuole valutare la plausibilità del modello in figura 3, ovvero che i costi degli impianti (variabile `costo`), le tipologie di servizi offerti (variabile `tipo`) e lo spessore della neve siano predittori della qualità generale della neve. Inoltre si vuole valutare se vi siano eventuali differenze tra le stazioni sciistiche (variabile `stazione`).

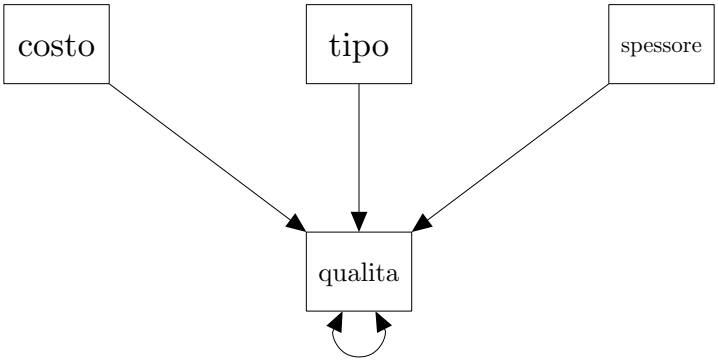

Fig. 3: Modello dell'esercizio 6.1.

1. Si importino i dati in R.
2. Si stimino i parametri del modello di regressione rappresentato in figura 3.
3. Si valuti, utilizzando gli intervalli di credibilità, se esistano differenze tra le stazioni sciistiche nei punteggi medi di qualità aggiungendo al modello del punto precedente il predittore `stazione`.
4. Si determini con il modello opportuno, se sia plausibile ipotizzare l'invarianza delle intercette rispetto alle 5 stazioni sciistiche.

Esercizio 6.2

Il radon è un elemento chimico radioattivo che si ottiene naturalmente dal decadimento dell'uranio ed è cancerogeno. Per controllare la presenza nelle case di radon l'agenzia USA di protezione ambientale ha raccolto dati in più di 80000 case. Nel dataset `radon` del pacchetto `ADati` ci sono dati relativi a 12777 rilevazioni in 8 stati americani. In particolare sono rilevate le seguenti variabili: `state`: stato della rilevazione, `county`: contea della rilevazione, `floor`: piano dell'abitazione (0 = piano terra), `radon`: concentrazione di radon, `log.radon`: logaritmo della concentrazione di radon, `u`: concentrazione di uranio, `log.u`: logaritmo della concentrazione di uranio.

1. Si rendano disponibili i dati in R.
2. Si confrontino graficamente le distribuzioni di radon (sia effettive che trasformate in logaritmo) negli otto stati per valutare se siano tra loro omogenee.
3. Si valuti se il piano delle abitazioni sia un predittore credibile del livello di radon utilizzando come variabile dipendente quella più opportuna tra le due confrontate al punto precedente.

4. Si valuti, con il modello opportuno ed una appropriata rappresentazione grafica, se vi siano differenze credibili tra gli otto stati.

5. Si modifichi il modello testato al punto 4 considerando gli stati come effetto random (intercette variabili).

6. Si rappresenti graficamente la distribuzione degli effetti random del modello.

Esercizio 6.3

Nel data set `school` del pacchetto `ADati` sono presenti dei dati relativi a 260 studenti di scuola superiore. Le variabili presenti nel data set sono le seguenti: `schid`, codice identificativo della scuola; `public`, status della scuola, pubblica `yes` o privata `no`; `ratio`, rapporto studenti-docenti della scuola; `percmin`, percentuale di studenti appartenenti a minoranze etniche; `classid`, codice identificativo della classe; `stuid`, codice identificativo dello studente; `sex`, genere dello studente; `ethn`, gruppo etnico dello studente; `ses`, status socio-economico della famiglia dello studente; `homework`, ore dedicate alla settimana ai compiti di matematica; `math`, punteggio al test di matematica.

1. Si carichino i dati in R.

2. Si producano le statistiche descrittive ed i grafici idonei per valutare le distribuzioni delle variabili del set di dati.

3. Si rappresenti graficamente la relazione tra le variabili `homework` e `math` differenziando le varie scuole valutando se sia ipotizzabile un modello di regressione unico e valido per tutte.

4. Si definisca un modello di regressione semplice per valutare se `homework` sia un predittore di `math`, si stimino i parametri e si valutino.

5. Si aggiunga al modello testato al punto precedente l'effetto casuale delle scuole considerando le intercette variabili, si stimino nuovamente i parametri confrontandoli con il risultato del punto 4.

6. Si consideri un nuovo modello, aggiungendo al precedente le pendenze variabili rispetto al fattore casuale scuole. Quindi si confrontino i due modelli per valutare quale risulti essere migliore.

7. Si rappresentino graficamente gli effetti random del modello giudicato migliore dal confronto eseguito al punto 6.

8. Si produca il grafico delle distribuzioni a posteriori per i parametri del modello giudicato migliore dal confronto eseguito al punto 6.

Esercizio 6.4

In uno studio sull'effetto dello stato di umore sull'attenzione vengono selezionati 24 soggetti (i dati sono disponibili nel file `mood.dat`). Il compito dei soggetti consiste nell'individuazione di alcuni stimoli ed il punteggio è dato dal numero di risposte corrette. Ciascun soggetto esegue il compito per tre volte in sequenza. Nella prima prova, di controllo, non viene indotta alcuna alterazione. Prima dell'esecuzione del compito per la seconda prova, al soggetto viene indotto uno stato di umore triste. Prima della terza prova al soggetto viene indotto uno stato di umore allegro. Si vuole sapere se le prestazioni dei soggetti variano in funzione dell'umore.

1. Si importino i dati in R.
2. Si rappresenti graficamente la distribuzione dei punteggi nelle tre condizioni sperimentali (`t0, t1, t2`) con tre boxplot.
3. Si aggiunga al grafico ottenuto al punto 2 la linea di tendenza delle medie. Quali considerazioni è possibile trarre dalla lettura di questo grafico?
4. Si calcolino la matrice di covarianza e di correlazione tra i punteggi nelle tre condizioni.
5. Si stimino i parametri del modello lineare considerando la condizione nei tre momenti come predittore del punteggio.
6. Si aggiunga al modello definito nel punto precedente la variabile `subj` come effetto random per valutare se le intercette dei soggetti siano equivalenti.
7. Si rappresentino graficamente le informazioni relative alle intercette dei soggetti.
8. Si valuti la plausibilità relativa dell'ipotesi di uguaglianza delle intercette. Quale sarà l'ipotesi nulla del test?

Esercizio 6.5

In relazione ai dati dell'esercizio 5.6 si considerino le seguenti ipotesi: 1) la grandezza dei numeri target influisce sulla grandezza degli scarti indicati dai bambini; 2) la relazione definita nell'ipotesi 1 varia anche in funzione dell'età dei bambini; 3) la relazione definita nell'ipotesi 1 varia anche in funzione del tipo di SNA.

1. Si individuino le variabili delle tre ipotesi e si identifichino i modelli opportuni per valutarle.
2. Si ristrutturi il file `SNA1.rda` nella forma lunga (osservazioni $\times$ variabili) più opportuna per poter testare le ipotesi deifinite al punto 1. Suggerimento: per ottenere la forma lunga si può utilizzare la funzione `stack()`.
3. Si stimi la plausibilità relativa dell'ipotesi 1 considerando i soggetti come effetto random.
4. Si formulino due modelli, aggiungendo al modello testato al punto 3 gli effetti di età e tipo di SNA, e si valuti (graficamente) la credibilità al 90% delle ipotesi 2 e 3 sulla base di questi modelli.
5. Si determini quale dei tre modelli considerati sia il più plausibile.
6. Si analizzino i residui del modello migliore per valutare se abbiano una distribuzione approssimativamente normale.

Esercizio 6.6

Si riprendano in esame i dati ed il modello finale dell'esercizio 5.7 al fine di considerare anche le eventuali differenze tra le classi.

1. Si modifichi il modello aggiungendo le intercette variabili per le classi. Si stimino e si individuino i valori dei parametri stimati.
2. Si valutino graficamente le differenze tra intercette nelle classi, interpretando il risultato.

3. Si modifichi il modello del punto 1 aggiungendo anche le pendenze variabili. Si stimino e si individuino i valori dei parametri stimati.

4. Si valutino graficamente le differenze tra pendenze nelle classi, interpretando il risultato.

5. Si individui con un opportuno confronto, quale tra i due modelli definiti ai punti 1 e 3 risulti più plausibile.

Esercizio 6.7

Si ripeta l'esercizio 6.6 utilizzando i soggetti come effetto random al posto della classe, per valutare le differenze individuali.

Esercizio 6.8

Si riconsiderino i dati dell'esercizio 1.10; in particolare si vuole sapere se le conoscenze di matematica (rilevate nella variabile `mathquiz`) siano predittive dei punteggi al test di statistica (variabile `statquiz`) tenendo conto di eventuali differenze legate al tipo di laurea triennale conseguita (variabile `major`).

1. Si importino i dati in R.

2. Si rappresentino graficamente i punteggi del test di statistica in funzione del punteggio in matematica.

3. Si stimino i parametri del modello di regressione in cui il punteggio di matematica è predittore del punteggio di statistica.

4. Si modifichi il modello del punto precedente per determinare se vi sia un effetto del tipo di laurea, valutando graficamente se tale effetto risulti credibile.

5. Si calcolino le evidenze relative (rispetto al modello nullo) dei modelli ai punti 3 e 4 e si determini quale tra i due risulti più evidente.

6. Si modifichi il modello del punto 4 considerando la variabile `major` come effetto casuale, per valutare la variabilità delle intercette e delle pendenze. Si stimino i parametri del modello.

7. Si produca un grafico che abbia in ascissa le intercette ed in ordinata le pendenze random e si interpreti.

8. Si rappresentino gli effetti fissi del modello e si interpretino.

Esercizio 6.9

Nel data frame `crimi` del pacchetto `ADati` ci sono i dati relativi ad uno studio sulla paura della criminalità e le sue determinanti, raccolti in 4 diversi quartieri di una grande città. Le variabili contenute nel file sono relative a 7 aspetti presi in considerazione nella ricerca: grado di pregiudizio verso extracomunitari (`Preg`), `Paura` della criminalità, strategie per fronteggiare la paura (`Strat`), percezione della presenza di extracomunitari (`Extra`), percezione di disordine fisico (`DisFis`) e sociale (`DisSoc`), grado di vittimizzazione (`Vitt`).

1. Si rendano disponibili i dati in R
2. Si analizzino le distribuzioni univariate delle variabli nel data-set.
3. Si analizzino le relazioni bivariate delle variabili nel data-set per cui ha senso.
4. Si utilizzino i *model weights* per determinare quali siano (escluse le variabili strategia e quartiere) i migliori predittori della paura della criminalità.
5. Si formuli un nuovo modello che includa i due migliori predittori della paura individuati al punto precedente con l'aggiunta dell'interazione. Si determini se l'interazione sia credibile al 70%, al 90% ed al 95%
6. Si valutino gli assunti del modello analizzandone i residui.
7. Si aggiunga al modello la variabile `Quartiere` per stimare la variabilità delle intercette rispetto a quest'ultima.
8. Si valutino gli assunti di questo modello analizzandone i residui (suggerimento: si può utilizzare la funzione `check_model()` del pacchetto `performance`).
9. Si stimino gli intervalli di credibilità dei parametri del modello considerato al punto precedente.
10. Si modifichi il modello definito al punto 7 aggiungendo la stima del parametro relativo alla variabilità delle pendenze. Si confronti questo nuovo modello con quello del punto 7 utilizzando i LOO.
11. Si valutino gli assunti di quest'ultimo modello analizzandone i residui.
12. Si rappresentino graficamente le previsioni del modello migliore tra quelli confrontati al punto 10.
13. Si rappresentino graficamente gli effetti random del modello migliore tra quelli confrontati al punto 10.

Esercizio 6.10

Il data set `sherifdat`[13] nel pacchetto `multilevel` contiene i dati relativi a 24 soggetti che hanno partecipato ad un esperimento in cui il compito era di stimare il movimento di una luce in una stanza buia per tre volte. I soggetti sono suddivisi in 8 gruppi (variabile `group`) da tre soggetti ciascuno. I soggetti dei gruppi 1, 2, 3, 4 hanno prima effettuato una stima individuale e poi hanno fornito una stima basata sulla valutazione di gruppo, i soggetti dei gruppi 5, 6, 7, 8 hanno fornito solo una valutazione come gruppo. Nella variabile `condition` sono identificati con 1 i soggetti dei primi quattro gruppi e con 0 quelli dei secondi quattro. La variabile `person` contiene l'identificativo dei partecipanti entro i gruppi (attenzione, non è un codice univoco per soggetto), la variabile `time` indica le ripetizioni dell'esperimento e la variabile `y` la stima fornita del movimento (in pollici).

L'obiettivo principale dello studio è di valutare se la stima del movimento sia predetta dal tipo di compito eseguito (se in gruppo o meno) e se vari nelle ripetizioni del compito.

1. Si acceda al data set `sherifdat` e si controlli la sua struttura.

[13] Fonte: Sherif, M. (1935). A study of some social factors in perception: Chapter 3. *Archives of Psychology*, 27, 23- 46.

2. Si individui la scala di misura delle variabili contenute nel data set. É coerente con il significato empirico delle stesse?
3. Si producano le statistiche descrittive univariate delle variabili del data set.
4. Si rappresenti graficamente la distribuzione univariata delle stime.
5. Si producano tre grafici che rappresentino graficamente la distribuzione delle stime in funzione di `time`, `group` e `condition`. Cosa emerge dall'ispezione dei grafici ottenuti?
6. Si valuti, con un modello di regressione lineare, se le variabili `time`, `group` e `condition` siano predittori credibili delle stime (variabile `y`).
7. Si valutino graficamente gli assunti del modello prodotto al punto precedente.
8. Si modifichi opportunamente il modello del punto 6 utilizzando degli effetti random che tengano conto del disegno sperimentale e dei dati. Si interpretino i parametri ottenuti.
9. Si valutino graficamente gli assunti del modello.
10. Si rappresentino graficamente le previsioni del modello e si commentino.

Esercizio 6.11

Consideriamo un campione di 51 pazienti con disturbi ossessivi, seguiti con una terapia in uno studio longitudinale per 5 anni. Per misurare il livello di disturbo viene utilizzato il test OBQ (Obsessive Beliefs Questionnaire) in cui valori alti indicano livelli di ossessione alti. I dati sono nel dataframe `OBQ` del pacchetto `ADati`.

A ciascun soggetto il test è stato somministrato 4 volte, rispettivamente all'inizio della terapia e dopo uno, tre e cinque anni dall'inizio; quindi abbiamo un totale di 51×4 osservazioni.

Si vuole valutare l'efficacia della terapia.

1. Si rendano disponibili i dati in R
2. Si producano delle opportune statistiche descrittive per le variabili del data-set.
3. Si rappresenti graficamente la distribuzione dei punteggi.
4. Si rappresenti graficamente la distribuzione dei punteggi in funzione del tempo di rilevazione.
5. Si stimino i parametri del modello di regressione lineare.
6. Si rappresentino graficamente i punteggi dei soggetti in modo da evidenziare se ci siano differenze individuali.
7. Si definisca un modello nullo multilivello utilizzando la variabile `subj` come effetto random e si stimino i parametri (quanti sono?).
8. Si aggiunga al modello del punto precedente il predittore `time` e si stimino i parametri (quanti sono?).
9. Si aggiunga al modello del punto precedente l'effetto random sulle pendenze e si stimino i parametri (quanti sono?).
10. Si confrontino i tre modelli ottenuti per individuare il migliore.

11. Si rappresentino graficamente le previsioni e gli effetti random del modello migliore.

Esercizio 6.12

Per valutare l'efficacia di una nuova terapia il cui obiettivo è di migliorare la Qualità della Vita in pazienti Alzheimer vengono presi in esame 140 soggetti di età compresa tra 61 e 99 anni ricoverati in 16 diversi centri specializzati. A 72 pazienti (scelti a caso) è stata applicata la nuova terapia (*treatment group*) mentre ai restanti 68 (*control group*) sono state applicate le terapie tradizionali. Lo strumento utilizzato per valutare la Qualità della Vita è il *Quality Of Life in Alzheimer's Disease*, QOL-AD, somministrato a ciascun paziente per 3 volte, rispettivamente prima dell'inizio della terapia e dopo 9 e 23 settimane.

L'obiettivo della ricerca è di valutare se vi siano differenze tra i due gruppi di soggetti nel corso delle settimane.

I dati sono nel dataset `QOLAD` del pacchetto `ADati`, organizzati in forma lunga; quindi ci aspettiamo un totale di $140 \times 3 = 420$ osservazioni. Il file contiene le seguenti variabili: `subj` – codice del soggetto, `age` – età, `gender` – genere, `group` – gruppo di appartenenza (treatment o control), `center` – centro di ricovero, `week` – settimana di rilevazione, `score` – punteggio al QOL-AD (valori più alti indicano una migliore qualità di vita).

1. Si rendano disponibili i dati in R
2. Si ispezionino opportunamente i dati e si valuti se il numero di osservazioni riportate nel dataset coincidano con le attese.
3. Si valuti se i due gruppi siano omogenei per genere ed età.
4. Si scriva e si stimino i parametri del modello nullo includendo come effetti random le intercette dei soggetti e dei centri.
5. Si rappresentino graficamente con dei boxplot gli effetti random del modello nullo e si interpretino.
6. Si determini, con un opportuno confronto tra modelli, quale sia il miglior predittore della qualità di vita (in termini di plausibilità) tra tipo di terapia, età e genere controllando il fattore temporale.
7. Si produca il grafico delle distribuzioni a posteriori dei parametri del modello più plausibile tra quelli appena confrontati.
8. Si stimino le seguenti probabilità a posteriori cercando di prevederle utilizzando il grafico prodotto al punto precedente: 1) probabilità che l'intercetta sia minore di 30; 2) probabilità che il parametro β_{week} sia compreso nell'intervallo $[-0.04, 0.04]$; 3) probabilità che il parametro β_{age} sia positivo.
9. Si rappresentino e si interpretino le previsioni del modello.

Esercizio 6.13

Si consideri il dataset `trust` descritto nell'esercizio 5.11. Si vuole sapere se il grado in cui si è disponibili a rispettare le regole (variabile `Y`) dipenda dal grado di fiducia (misurato dalle variabili `X1`, `X2`, `X3`, `X4`) tenendo conto dei diversi paesi in cui sono stati raccolti i dati.

1. Si rendano disponibili i dati in R

2. Si stimino i parametri del modello *pooled* in cui `Y` è la dipendente e `X1`, `X2`, `X3`, `X4` i predittori, considerati solo in forma additiva. Si ipotizzi il numero di parametri (prima di vedere il risultato).

3. Si stimino i parametri del modello *non-pooled* aggiungendo l'effetto dei paesi (variabile `country`). Si ipotizzi il numero di parametri (prima di vedere il risultato).

4. Si definisca il modello multilivello che include le intercette variabili rispetto ai paesi e se ne stimino i parametri. Si ipotizzi il numero di parametri (prima di vedere il risultato).

5. Si analizzino i residui del modello per valutare se siano accettabili.

6. Si rappresentino graficamente i valori attesi del modello.

7. Si confrontino le stime dei parametri ottenute nel modello multilivello con quelle dei modelli *pooled* e *non-pooled*.

8. Si valuti, utilizzando R^2 e LOO, se l'inclusione delle intercette variabili sia utile nel modello.

Esercizio 6.14

Si consideri ancora il dataset `trust` descritto nell'esercizio 5.11 ed i risultati ottenuti nell'esercizio 6.13.

1. Si rappresentino graficamente le relazioni bivariate tra la dipendente `Y` ed i quattro predittori `X1`, `X2`, `X3`, `X4` assegnando colori diversi ai vari paesi. Suggerimento: si può semplicemente usare la funzione `plot()` includendo l'opzione `col = country`.

2. Si scrivano quattro modelli in ciascuno dei quali vengano introdotte le pendenze variabili rispetto ad uno dei predittori (`X1`, `X2`, `X3`, `X4`), si stimino i parametri di questi modelli.

3. Si confrontino con il LOO i modelli appena adattati, aggiungendo anche il modello con intercette variabili dell'esercizio 6.13 (modello nullo) per individuare quello più plausibile.

4. Si valutino i residui del modello migliore ottenuto per stabilire se sia accettabile.

5. Si rappresentino e si interpretino i valori attesi della parte fissa del modello.

6. Si rappresentino i BLUPS del modello.

Esercizio 6.15

Nel dataset `attivamente` del pacchetto `ADati` sono raccolti i dati di una ricerca inclusa in un progetto per la prevenzione degli abusi di tecnologie in studenti di scuola primaria. Tali dati, rilevati prima e dopo un intervento a scuola della durata di quattro settimane, sono organizzati in forma lunga nel dataframe `attiva.long`; ogni riga del data set è un'osservazione, ogni colonna una variabile, nell'ordine: `ID` = codice identificativo dei soggetti, `genere`, `eta` = età in anni, `scuola` frequentata, `fase` = momento della raccolta dati – pre e post intervento –, `IU` = punteggio ad un questionario sull'uso di internet (più è alto il punteggio maggiore il tempo dedicato all'utilizzo di internet), `CG` = punteggio ad un questionario sul controllo dei genitori (maggiore il punteggio, maggiore il controllo).

L'obiettivo dello studio è determinare se ci siano stati dei cambiamenti nei punteggi di IU tra prima e dopo l'intervento.

1. Si rendano disponibili i dati in R.
2. Si producano le statistiche descrittive delle variabili.
3. Si determini il numero di soggetti inclusi in totale nel dataset e quanti sono per ciascuna scuola.
4. Si producano delle opportune rappresentazioni grafiche univariate per descrivere le variabili del data set.
5. Si rappresenti graficamente la distribuzione dei punteggi di `IU` prima e dopo l'intervento.
6. Si definisca il modello per valutare se ci siano differenze nei punteggi di `IU` prima e dopo l'intervento, tenendo conto delle differenze individuali, e se ne stimino i parametri.
7. Si determini, sulla base del modello ottenuto, se ci sia una differenza credibile nei punteggi di `IU` tra prima e dopo l'intervento.
8. Si rappresentino graficamente le distribuzioni dei punteggi di `IU` in funzione del controllo dei genitori, prima e dopo l'intervento.
9. Si scriva un modello per valutare se il controllo dei genitori sia predittivo dei punteggi di `IU` prima e dopo l'intervento; si stimino i parametri di questo modello.
10. Si calcolino i weight dei due modelli ottenuti ai punti 6 e 9 e si interpretino.
11. Si modifichi il modello opportunamente per tenere conto delle differenze tra le scuole e si valuti quanto cambiano i parametri relativi alla `fase` ed al controllo dei genitori.
12. Si confrontino i tre modelli per valutare quale sia il più plausibile.

Esercizio 6.16

Si consideri il modello migliore individuato nell'esercizio 6.15 per valutarne le diagnostiche.

1. Si analizzino i residui del modello.
2. Si calcoli il LOO del modello.
3. Si rappresentino graficamente i Pareto k per visualizzare i casi problematici.
4. Si selezionino le osservazioni del dataset con $k > 1$.
5. Si rappresentino in due grafici separati i punteggi di `IU` in funzione di `fase` e `CG` evidenziando i casi problematici.
6. Si valuti se escludendo dal dataset queste osservazioni si risolve il problema.

7 Appendice

Pacchetti utili:

```
 [1] "BayesFactor"  "DataExplorer" "foreign"      "gdata"        "ggplot2"      "gplots"
 [7] "gtools"       "Hmisc"        "labstatR"     "lattice"      "loo"          "MASS"
[13] "performance"  "PRDA"         "psych"        "Rcpp"         "readxl"       "rstanarm"
[19] "sjPlot"       "yarrr"
```

Elenco funzioni suggerite per le soluzioni:

Il seguente elenco contiene alcune tra le molte funzioni di R; in particolare si tratta di comandi che possono risultare utili nella soluzione degli esercizi proposti. Per vedere i dettagli sull'uso di queste funzioni si consulti il relativo `help`.

```
 [1] "aggregate"          "apply"              "as.character"       "barplot"
 [5] "boxplot"            "c"                  "campionaria.media" "cbind"
 [9] "choose"             "colnames"           "combinations"       "cor"
[13] "cov"                "cumsum"             "curve"              "cut"
[17] "data"               "data.frame"         "density"            "describe"
[21] "dnorm"              "ecdf"               "exp"                "expand.grid"
[25] "factor"             "factorial"          "file.choose"        "geom_density"
[29] "hist"               "ks.test"            "lapply"             "legend"
[33] "length"             "levels"             "library"            "matrix"
[37] "max"                "mean"               "median"             "min"
[41] "mvrnorm"            "nrow"               "ordered"            "par"
[45] "permutations"       "plot"               "plot_bar"           "plot_boxplot"
[49] "plot_correlation"   "plot_intro"         "plot_missing"       "plotmeans"
[53] "pnorm"              "qqline"             "qqnorm"             "quantile"
[57] "range"              "rbind"              "read_excel"         "read.spss"
[61] "read.table"         "read.xls"           "rep"                "rm"
[65] "rnorm"              "round"              "sample"             "scan"
[69] "sd"                 "seq"                "shapiro.test"       "sigma2"
[73] "sort"               "sqrt"               "stack"              "str"
[77] "subset"             "sum"                "summary"            "table"
[81] "tapply"             "unique"             "var"                "with"
```

Indice dei data-set

Elapsed time 26.27 mins