PSP6075525 - Testing psicologico
Modelli e metodi statistici per la misurazione in psicologia

ESERCIZI - recupero prerequisiti

Versione: 6 novembre 2023

Nota: Per l'esecuzione degli esercizi che seguono si consiglia il ripasso dei contenuti (li dove opportunamente richiamati dagli esercizi) presenti nel libro di Crawley (2013)[1], in particolare le sezioni 2.1, 2.2, 2.6-2.8, 2.10, 3-5, 7.3-7.4, 8, 10.1.

Esercizio I
Si consideri il dataset `mathSchool.rda` (presente sulla pagina Moodle del corso nella cartella "Datasets") che contiene i punteggi in matematica (`mathscore`) realizzati da studenti (`subj`) all'interno del loro percorso didattico quinquennale (`year`). Le variabili (`class`) e (`teach`) indicano, rispettivamente, la classe frequentata e il docente associato a quest'ultima.

1. Dopo aver caricato in R il dataset, si esegua quanto segue:
   (a) Studiate la struttura del dataframe e riportate le statistiche di sintesi della variabile `mathscore`
   (b) Calcolate la media di `mathscore` per gli studenti frequentanti la classe `101`, `105`; la mediana per gli studenti delle classi `102`, `104`; il minimo e il massimo per quelli frequentanti la classe `103`, `102`.
   (c) Contate il numero di `teach` associati a ciascuna `class` e registrateli in una nuova variabile
   (d) Contate il numero di `subj` all'interno di ciascuna `class` e registrateli in una nuova variabile
   (e) Calcolate le statistiche di sintesi di `mathscore` per gli studenti del primo anno frequentanti la classe `101`
   (f) Calcolate le statistiche di sintesi di `mathscore` per gli studenti del terzo anno frequentanti la classe `101`
   (g) Quale classe ha ottenuto il massimo di `mathscore` nel primo anno?
   (h) Quale classe ha ottenuto il minimo di `mathscore` nel quinto anno?
   (i) Visualizzate con un opportuno grafico la distribuzione di `mathscore` per ogni anno riportato nel dataframe
   (j) Visualizzate con un grafico a scatole e baffi la distribuzione di `mathscore` condizionata ai livelli della variabile di classificazione `year`. Interpretatene il risultato
   (k) Visualizzate con un grafico a scatole e baffi la distribuzione di `mathscore` condizionata ai livelli della variabile di classificazione `teach`. Interpretatene il risultato
   (l) Visualizzate con un grafico a scatole e baffi la distribuzione di `mathscore` condizionata ai livelli della variabile di classificazione `class`. Interpretatene il risultato

[1] Il manuale è disponibile sulla pagina Moodle del corso alla voce "Materiale di supporto (online)"

2. Selezionate i punteggi di `mathscore` della classe `101` e dell'anno di corso `1` e registrateli in una opportuna matrice $\mathbf{X}$ di dimensione 13x2. Successivamente, selezionate i punteggi di `mathscore` della classe `101` e dell'anno di corso `3` e registrateli in una opportuna matrice $\mathbf{Y}$ di dimensione 13x2. Infine, eseguite la seguente operazione: $\|\text{dist}(\mathbf{X} - \mathbf{Y})\|_2$.

3. Utilizzando la funzione di R `as.vector()` sulle precedenti matrici, si definiscano due array $\mathbf{x}_{26\times 1}$ e $\mathbf{y}_{26\times 1}$ e si svolga quanto segue:

   (a) Visualizzate con un opportuno grafico la relazione tra $\mathbf{x}$ e $\mathbf{y}$

   (b) Si calcoli la nuova variabile $\hat{\mathbf{y}} = (\mathbf{x} \circ \mathbf{y}) r_{x,y}$, dove $\circ$ indica il prodotto elemento per elemento mentre $r_{x,y}$ è la correlazione tra i due vettori segnati in pedice

   (c) Create un nuovo dataframe `dataReg` con le variabili $\mathbf{x}$, $\mathbf{y}$ e $\hat{\mathbf{y}}$ ottenute dall'esercizio precedente

4. Definite una nuova variabile z contenente il risultato della seguente espressione:

$$\mathbf{z} = (\mathbf{x}^2 + \mathbf{y}^{\frac{1}{2}}) \min(\mathbf{x}, \mathbf{y}) + \epsilon$$

   dove $\epsilon \sim \mathcal{U}(\epsilon; -1, 1)$. Nota: per generare da una distribuzione uniforme si utilizzi: `runif(..)`.

5. Selezionate i punteggi di `mathscore` della classe `102` e dell'anno di corso `4` e registrateli in un opportuno array $\mathbf{x}$ di dimensione 31x1. Fate lo stesso per i punteggi della classe `102`, anno di corso `5`, registrandoli nell'array $\mathbf{y}$. Ripetete gli esercizi 2-4 utilizzando i nuovi array $\mathbf{x}$ e $\mathbf{y}$ e denominando il nuovo dataframe `dataReg2`.

6. Generate un nuovo dataframe `dataRegFin` con una opportuna concatenazione dei dataframe `dataReg` e `dataReg2` ottenuti negli esercizi precedenti.

7. Create una lista `listReg` contenente i dataframe `mathScore`, `dataReg`, `dataReg2`, `dataRegFin`. Salvate solo l'oggetto `listReg` sul vostro disco rigido. Successivamente salvate tutti gli oggetti del vostro spazio di lavoro di R ad eccezione di `listReg` sul vostro disco rigido. Salvate infine in formato .pdf il grafico ottenuto all'esercizio 3.

   Nota: Per salvare oggetti da R si veda la sezione 2.16 del manuale di Crawley (2013). Per salvare i grafici si veda invece la sezione 5.10 sempre dello stesso manuale.

Esercizio II

1. Si importi nell'ambiente di lavoro di R il dataset `dataset6.dat`, prestando attenzione al modo con cui il file è codificato. Nota: Prima di importare il file, provate ad esplorarlo con un editor di testo mentre per importare il file di dati si consulti la funzione `read.table( )` nella sezione 3.2 del manuale di Crawley (2013).

2. Esplorate il dataframe estraendone delle informazioni di sintesi con `summary( )` e `str( )`.

3. Utilizzando le variabili numeriche presenti nel dataframe, create la seguente variabile categoriale `d3` della stessa dimensione del numero di casi presenti nel dataframe caricato in input:

$$\texttt{d3} = \begin{cases} \texttt{h}, & \text{se} \quad \texttt{X1} \geq \overline{\texttt{X2}} \\ \texttt{m}, & \text{se} \quad \texttt{X4} \leq \overline{\overline{\texttt{X3}}} \\ \texttt{l}, & \text{se} \quad \texttt{X2} \leq \overline{\texttt{X1}} + \frac{\min(\texttt{X3}) - \max(\texttt{X5})}{2} \end{cases}$$

   dove $\overline{\texttt{X2}}$ indica la media della variabile `X2` mentre $\overline{\overline{\texttt{X3}}}$ indica invece la mediana della variabile `X3`. Assegnate successivamente ai valori vuoti di `d3` il carattere `vl`. Assegnate questa nuova variabile al dataframe precedente. Nota: All'inizio, generate la variabile `d3` mediante la funzione `rep( )` popolandola di `NA`.

4. Rappresentate graficamente mediante boxplot le variabili `X1`, `X2`, `X3`, `X4` in funzione (i.e., condizionato ai livelli) della variabile `d3`. Successivamente si rappresentino graficamente le stesse variabili mediante istogramma. Nota: Si divida la finestra grafica in più sottofinestre.

5. Si valutino le espressioni seguenti:

$$\texttt{Y} = \texttt{X5} \cdot \frac{s_{\texttt{X1,X2}} \cdot \overline{\texttt{X1}} \cdot \texttt{X2}}{\min(\texttt{X2}) - \max(\texttt{X1})}$$

$$\texttt{Z} = \texttt{X5} \cdot \frac{s_{\texttt{X3,X4}} \cdot \overline{\texttt{X3}} \cdot \texttt{X4}}{\min(\texttt{X4}) - \max(\texttt{X3})}$$

dove $s_{A,B}$ indica il valore della covarianza applicata alle variabili indicate in pedice mentre $\overline{\texttt{A}}$ indica la media della variabile `A`.

6. Analizzate mediante opportuno grafico la relazione tra le variabili Y e Z. Commentate ogni output grafico ottenuto.

7. La funzione di R `sample(x=Y,size=I,replace=TRUE)` permette di campionare (estrarre) a caso, secondo modello uniforme, `I` valori dalla variabile `Y`. Si scriva una funzione in R che svolga quanto segue:

   (a) riceva in input il numero `I` di campioni da estrarre e la variabile `Y`

   (b) sintetizzi le informazioni dei campioni estratti mediante *media*

   (c) restituisca in output la media ottenuta

8. Si applichi la funzione prima create sul vettore $I = \{10, 100, 1000, 5000\}$ e si salvino i risultati ottenuti in un apposito vettore.

Esercizio III

1. È noto che le altezze X dei maschi italiani di età compresa tra 19-27 anni si distribuiscono secondo la legge $X \sim (\mu, \sigma^2)$ con $\mu = 174.5$ (cm) e $\sigma^2 = 8$. Rappresentare graficamente la funzione di densità sapendo che il supporto di X può essere ristretto (per convenienza) all'intervallo chiuso [162.5, 186.9].

2. Sulla base del modello precedente, calcolate le seguenti probabilità: $P[X \geq 180]$, $P[X \geq 185.9]$, $P[X < 180]$, $P[X \geq 180]$, $P[170 \leq X \leq 180]$.

3. L'azienda *ArchiMeteC* è nota per la produzione di bulloni a testa esagonale la cui resistenza al calore X è nota seguire la legge $X \sim \chi^2(; g)$. Dopo approfonditi studi, l'azienda ha determinato che $g = 3$ mentre (per convenienza) il supporto può essere ristretto su $S(X) = [0, 16.5]$ (0 indica resistenza nulla, 16.5 indica massima resistenza). Si produca il grafico della funzione di densità della v.c. in oggetto. Cosa possiamo notare dal grafico? Inoltre, l'azienda vorrebbe determinare le seguenti quantità di interesse: $P[X \leq 1]$, $P[X \leq 3]$, $P[X > 8.1]$, $P[2.5 \leq X \leq 7.1]$, $P[2.5 \leq X \geq 7.1]$. Infine, considerate le attuali risorse tecniche di cui dispone, qual è la probabilità che l'azienda ha di produrre bulloni oltre la resistenza massima? Sulla base di questo risultato, consiglierete all'azienda di aggiornare la strumentazione per la produzione di bulloni?

4. L'azienda farmaceutica *NoxArtis* vorrebbe introdurre un nuovo farmaco la cui letalità Y è nota seguire la legge $Y \sim \mathcal{E}xp(; \lambda)$ con $\lambda = 1.9$. Disegnate il grafico per la funzione di densità indicata. Inoltre, si vogliono determinare le seguenti probabilità: $P[Y \leq 1]$, $P[Y \geq 1.7]$, $P[Y < 0.5]$, $P[0.05 \leq Y \leq 0.15]$. Sapendo che i valori di Y possono essere ristretti all'intervallo [0, 3.1] (per convenzione), dove 0 indica letalità nulla mentre 3.1 indica letalità massima, qual è la probabilità che il nuovo farmaco abbia letalità oltre quella nota?

5. La *ConfiTer*, azienda leader nella produzione di spaccalegna a vite, deve decidere se dismettere o meno un suo macchinario. Tale decisione si basa sugli errori X che questo compie nel processo produttivo dei cilindri delle spaccalegna a vite. L'azienda suppone che in media la macchina sbagli i diametri dei cilindri per un valore di 0.24, un valore

troppo alto per continuare ad utilizzare il macchinario. Per valutare tale ipotesi, decide di condurre un'indagine campionaria estraendo a caso il seguente campione:

$$\mathbf{x} = \{-0.01, -0.04, 0.24, -0.13, 0.17, -0.17, 0.30, 0.23, 0.21, 0.23, 0.17, 0.28, 0.09, 0.12, -0.11\}$$

Fissando $\alpha = 5\%$ e sapendo che $X \sim \mathcal{N}(\mu, \sigma)$, $\sigma = 0.25$, si valuti l'ipotesi $\mathcal{H}_0 : \mu = 0.24$ vs. $\mathcal{H}_1 : \mu > 0.24$ (ipotesi alternativa unidirezionale). Sulla base del campione estratto, l'azienda può dismettere il macchinario?

6. Luigi e Mario sono amici di vecchia data ed hanno un hobby in comune, osservare il numero di *Branta canadensis* che ogni giorno stanziano sul fiume della loro cittadina. Luigi e Mario sanno che il numero totale di stanziamenti n è pari a 28 ma sono in disaccordo se vi siano più esemplari di *Branta canadensis maxima* (A) o *Branta canadensis inferior* (B). A tal fine conducono un'indagine campionaria osservativa: Luigi conta 15 esemplari di A ($n_A = 15$) mentre Mario conta 23 esemplari di B ($n_B = 23$). Scegliendo un opportuno test statistico, con un $\alpha = 5\%$ e i dati osservativi a disposizione, possiamo rigettare l'ipotesi di Luigi che il numero di stanziamenti nella popolazione della *Branta canadensis* non differisce per le due sottospecie?

7. Un paese europeo si avvicina alle elezioni per il rinnovo del suo parlamento. In particolare, il noto partito delle *cugine di Malta* sta costruendo da tempo la sua campagna elettorale sull'affermazione che non vi sia associazione tra impiego dei giovani a 2 anni dalla laurea ($X = \{\texttt{impiego si}, \texttt{impiego no}\}$) e l'aver frequentato un corso di laurea generico ($Y = \{\texttt{laurea si}, \texttt{laurea no}\}$). Sulla base di tale affermazione vorrebbe abolire le università. Lo zio di Marta fervido sostenitore del partito e delle sue iniziative, cerca di convincere la nipote della verità di tale affermazione. Marta, dopo aver superato l'esame di statistica inferenziale all'università, decide di rispondere allo zio importuno in maniera razionale. A tal fine, conduce un'indagine campionaria e dopo aver estratto un campione casuale dalla popolazione dei giovani laureati, registra i seguenti dati:

	`laurea no`	`laurea si`
`impiego si`	18	24
`impiego no`	45	26

Dopo aver condotto un test basato sulla statistica χ^2 di Pearson, possiamo affermare ad un $\alpha = 5\%$ che vi sia indipendenza tra X e Y? Lo zio di Marta, dopo aver consultato un amico statistico anche lui fervido sostenitore delle *cugine di Malta*, suggerisce alla nipote che la statistica test utilizzata non è sensibile a sufficienza e le indica di utilizzare invece la c.d. statistica *deviance*:

$$G = 2 \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} \ln \left(\frac{n_{ij}}{\hat{n}_{ij}} \right)$$

dove n_{ij} indica le frequenze osservate mentre $\hat{n}_{ij}$ le frequenze attese sotto l'ipotesi nulla di indipendenza. Dopo aver utilizzato una statistica test differente, è possibile ancora affermare che vi sia indipendenza tra X e Y?
Nota: Per il calcolo della statistica G si faccia riferimento alla sezione 8.8.2 del manuale di Crawley (2013).

Esercizio IV
Il dataset `CFC.csv` contiene dati provenienti da un questionario sugli stili di attaccamento in età adulta (si consulti il file `CFC_info.txt` per ulteriori dettagli). Dopo aver importato il dataset nell'ambiente di lavoro di R, si svolga quanto segue. Nota: È possibile importare il file anche mediante la funzione apposita `read.csv( )`.

1. Si analizzino le relazioni tra le variabili `Q3`-`Q4` mediante opportuni grafici e misure di associazione. Cosa possiamo affermare circa la relazione tra le due variabili?

2. Definite un modello lineare in cui la variabile `Q3` è predetta da `Q2`. Commentatene i risultati.

3. Estendente il modello lineare precedente aggiungendo come predittori le variabili `Q2` e `Q1`. Commentate i risultati. Possiamo rigettare, ad un livello di significatività $\alpha = 5\%$, che $\beta_{Q_1} = 0$? Cosa possiamo concludere sui residui del modello?

4. Eliminate il predittore `Q1` dal modello precedente, aggiungendovi la variabile di interazione `Q11`×`Q5`. Cosa possiamo concludere?

5. Dopo aver definito la variabile:

$$\texttt{Y} = (Q_1 + Q_2) r_{Q_1,Q_2} + (Q_3 + Q_{11}) r_{Q_3,Q_{11}}$$

   dove in generale $r_{X,Y}$ indica il coefficiente di correlazione calcolato sulle variabili indicate in pedice, definite un modello lineare in cui Y è predetto dalla variabile categoriale `gender`. Possiamo concludere, ad un livello di significatività $\alpha = 5\%$, che la variabile `gender` produce una differenza in media su Y? Nota: Ci si assicuri che la variabile `gender` sia correttamente codificata come `factor`.

6. La variabile `country` ha 28 livelli. Definite una nuova variabile `country_new` a due livelli, come segue:

$$\text{country\_new} = \begin{cases} 0 & \text{quando } \texttt{country} = \text{US} \\ 1 & \text{quando } \texttt{country} \neq \text{US} \end{cases}$$

   Definite un modello in cui la variabile precedente Y è predetta da `country_new`. Commentate i risultati.

7. Flavia sta conducendo le analisi dei dati della sua ricerca utilizzando `R`. In particolare vuole studiare la relazione tra le variabili `age` e `accuracy` presenti nel dataset `CFS.csv`. Dopo aver stimato i parametri del modello mediante la funzione `lm( )` di `R`, riporta con grande soddisfazione che il parametro $\beta_{\text{accuracy}} = 0.146$ e il valore $p = .031$ è minore del livello di significatività $\alpha = .05$. Mentre si accinge a scrivere finalmente i risultati significativi della sua ricerca, Sofia, studentessa di dottorato al primo anno, le comunica con insistenza che `R` ha un *bug* interno e calcola in maniera errata i p-value del modello di regressione lineare. La invita dunque ad utilizzare SPSS, più accurato e preciso. Flavia, rattristata, vorrebbe fare un ultimo tentativo prima di abbandonare `R`. A tal fine prova a verificare manualmente che l'ipotesi $\text{H}_0 : \beta_{\text{accuracy}} = 0$ sia rigettata ad un $\alpha = 5\%$ (come suggerisce l'output di `R`). Sapendo che la statistica-test da utilizzare è $T_n = \frac{\beta}{s_\beta} \sim t(n-2)$ (s_β indica lo *standard error* della stima di β mentre n la numerosità campionaria), suggerireste a Flavia di continuare ad usare `R`, rigettando l'ipotesi di Sofia ad un $\alpha = 0.05$?