Dipartimento di Economia e Management
Corso di Laurea in Economia e Management
Esame di Analisi dei Dati e Statistica – a.a. 2022/2023
Prof. Giuseppe Espa, Prof. Enrico Tundis

Per poter svolgere la prova è necessario innanzitutto caricare il file **`WorldValueSurvey.rdata`**, il quale contiene il *data-frame* **`wvs`** con le variabili da analizzare. Questo *data-frame* è relativo all'indagine *World Values Survey*, condotta in Italia nel 2005–2006. Il questionario completo si trova nel file **`wvs-questionario.pdf`**, mentre una sintesi delle variabili da analizzare si trova nel file **`wvs-variabili.pdf`**.

Per caricare il file **`WorldValueSurvey.rdata`** occorre lanciare **`R`**, cliccando poi su **`File`** e su **`Carica area di lavoro`**…, andando quindi a cercare la cartella dove è stato copiato il file in oggetto (oppure trascinando il file stesso nel *workspace* di **`R`**). Per verificare che il file sia stato caricato correttamente, digitate, dal *prompt* di **`R`**, **`ls()`**; la risposta dovrà essere:

```
[1] "wvs"
```

A questo punto è possibile passare a rispondere ai punti che seguono non prima di avervi fornito una descrizione sintetica delle variabili presenti nel *data-frame*. Le variabili da 1 a 33 sono variabili quantitative che misurano, su una scala da 1 a 10, il grado di "aderenza" alla relativa domanda. Si raccomanda di consultare il questionario per una comprensione più approfondita della risposta data dai soggetti alla domanda in esame; il nome della variabile corrisponde alla sua posizione nel questionario. Ad esempio dal file **`wvs-variabili.pdf`** si può vedere che la variabile numero 5 corrisponde alla domanda v91 del questionario (vedi pagina 8 del file **`wvs-questionario.pdf`**) che recita: "La scienza e la tecnologia stanno rendendo la nostra vita più sana, facile e più confortevole"; 1 indica un completo disaccordo, mentre 10 indica un completo accordo. Ricordate che il nome completo della variabile v91 è **`wvs$v91`**. Anche le variabili 35 (anno di nascita) e 36 (età) sono variabili quantitative. Le rimanenti (34, 37, 38, 39, 40) sono variabili qualitative le cui modalità possono essere desunte dal questionario (per le variabili 34, 37, 38) oppure da una tabulazione diretta (impiegando, ad esempio, la funzione **`table`**): la variabile 39, che per **`R`** è un "**`Ord.factor`**", è una ricodifica in classi della variabile "età". Può essere pensata e trattata come una variabile qualitativa ordinabile senza particolari problemi e questo si chiede di fare. La variabile numero 40 codifica l'area geografica come Nord-Ovest, Nord-Est, Centro, Sud, Isole.

Il testo degli esercizi (che trovate di seguito) è lo stesso per tutti, ma le variabili da analizzare sono diverse. Fa eccezione il solo esercizio numero 9, uguale per tutti quanto a domande poste e dati da analizzare.

Caricate il file **`esercizi.pdf`** e localizzate la riga contenente il vostro numero di matricola. Guardate a questo punto i valori contenuti nelle colonne successive per avere il testo completo degli esercizi da svolgere. Coloro i quali, nella pagina Moodle associata al corso di Analisi dei Dati e Statistica, si sono iscritti senza indicare il numero di matricola (magari perché in attesa di immatricolazione) dovranno localizzare le iniziali del proprio nome e cognome. Anche costoro guarderanno i valori contenuti nelle colonne successive per disporre del testo completo degli esercizi da eseguire. Infine, chi non è "utente iscritto" al gruppo dei "partecipanti" nella pagina Moodle del corso di Analisi dei Dati e Statistica è pregato di contattare il docente per avere il testo completo degli esercizi da svolgere.

Una volta completati gli esercizi, va restituito al docente un file in formato "pdf" che contiene lo *script* prodotto e l'*output* ottenuto mediante il suo utilizzo. Qualche riga di commento ai risultati è

molto gradita e, quindi, caldeggiata. Il file (completo di cognome, nome e numero di matricola da riportarsi a cura dello studente sulla prima pagina) va inviato esclusivamente al seguente indirizzo di posta elettronica e **non** all'indirizzo e-mail personale del docente:

analisidatiespa@economia.unitn.it

N.B. il nome del file deve riportare cognome e nome dello studente secondo la struttura che segue: "totti_francesco.pdf". La mail **deve** riportare nell'oggetto la dicitura: "Analisi dei Dati e Statistica".

1. Carica il file `WorldValueSurvey.rdata` e costruisci la distribuzione di frequenza della variabile la cui posizione nel *data-frame* `wvs` è riportata nella colonna (1). Calcola inoltre media, mediana, varianza e deviazione standard della stessa variabile (attenzione ai dati mancanti; usa l'argomento `na.rm=TRUE`).
2. Carica il file `WorldValueSurvey.rdata` e calcola la media della variabile la cui posizione nel *data-frame* `wvs` è riportata nella colonna (1) all'interno delle modalità della variabile la cui posizione nel *data-frame* `wvs` è riportata nella colonna (2). Suggerimento: puoi usare la funzione `by` oppure la funzione `tapply`.
3. Per la variabile la cui posizione nel *data-frame* `wvs` è riportata nella colonna (1) disegna i *boxplot* all'interno delle modalità della variabile la cui posizione nel *data-frame* `wvs` è riportata nella colonna (2). Suggerimento: puoi usare, come fatto a lezione, una *formula* (con la "tilde" `~`) come argomento della funzione `boxplot`.
4. Carica il file `WorldValueSurvey.rdata` e, utilizzando le variabili le cui posizioni nel *data-frame* `wvs` sono riportate nella colonna (3) e nella colonna (4), calcolate la covarianza e il coefficiente di correlazione lineare (attenzione ai dati mancanti; in questo caso usa l'argomento `use=complete.obs`).
5. Immagina di avere un dado non truccato con un numero di facce pari a quelle indicate nella colonna (5) e di lanciarlo un numero di volte pari al valore indicato nella colonna (6) registrando il valore più alto ottenuto. Ripetendo questo esperimento per 100000 volte, stima con il metodo Monte Carlo la probabilità che il massimo raggiunga o superi il valore riportato nella colonna (7).
6. Un portafoglio di crediti raccoglie le posizioni a elevato rischio di insolvenza di un gruppo bancario. Dalle stime svolte, ciascuno dei crediti inclusi nel portafoglio ha una probabilità di insolvenza a due anni pari al valore indicato nella colonna (8), mentre il numero complessivo di posizioni (crediti) è pari al valore riportato nella colonna (9). Calcola il numero di insolvenze attese, la probabilità di superare tale valore, e infine la probabilità di registrare un numero di insolvenze superiore al valore indicato nella colonna (10). Suggerimento: si può usare la funzione `R pbinom` oppure la funzione `Binomiale` del pacchetto `rmf`.
7. Considera una distribuzione normale, i cui parametri μ e σ assumono i valori riportati, rispettivamente, nelle colonne (11) e (12). Calcola la probabilità di osservare un valore compreso tra μ-σ e il valore riportato nella colonna (13). Suggerimento: si può usare la funzione `pnorm` di `R` oppure la funzione `ProbNorm` del pacchetto `rmf`.
8. Genera un campione `x` di 300 osservazioni mediante l'istruzione:
   ```
   set.seed(987654); x<- sample(-3:5,size=300,replace=TRUE); set.seed(NULL)
   ```
   sostituendo al numero `987654` il tuo numero di matricola (ossia il numero riportato nella colonna intestata "innesco"). Considerando ora la distribuzione per dati grezzi per la variabile X (ossia gli elementi del vettore `x`)

a) calcola la mediana utilizzando la distribuzione di frequenza e, successivamente ed in maniera più diretta, lavorando con i dati grezzi;
b) verifica che la media degli scarti in valore assoluto dalla mediana è minore della media degli scarti in valore assoluto da un qualsiasi altro valore (per es. 3.14) Suggerimento: per calcolare il valore assoluto puoi usare la funzione **abs** di **R**.

9. Le tabelle che seguono riportano alcuni dati relativi all'epidemia COVID-19 di fonte Istituto Superiore di Sanità, Roma (https://www.epicentro.iss.it/coronavirus/sars-cov-2-dashboard). I dati sono contenuti nell'aggiornamento nazionale dell'1 febbraio 2022 e si riferiscono al 31 gennaio 2022. Si tratta di informazioni raccolte attraverso una piattaforma web dedicata ed include tutti i casi di COVID-19 diagnosticati dai laboratori di riferimento regionali e i decessi segnalati per classi di età e sesso. Si precisa che, nonostante la fonte consultata ne faccia menzione, sono stati esclusi dalle tabelle riportate qui di seguito i dati relativi al numero di casi ed al numero di decessi di quegli individui (maschi e femmine) per i quali non è stato possibile stabilire, neppure approssimativamente, l'età.

Maschi			
Classe di età	N. deceduti	N. casi	Letalità (%)
0-9	7	444638	0,0016%
10-19	15	667351	0,0022%
20-29	61	697747	0,0087%
30-39	222	675651	0,0329%
40-49	948	791996	0,1197%
50-59	3808	778178	0,4893%
60-69	10785	469786	2,2957%
70-79	24300	305004	7,9671%
80-89	31274	163360	19,1442%
90 e oltre	9870	32024	30,8206%
Totale	81290	5025735	1,6175%

Femmine			
Classe di età	N. deceduti	N. casi	Letalità (%)
0-9	10	413787	0,0024%
10-19	13	644080	0,0020%
20-29	34	694093	0,0049%
30-39	128	743312	0,0172%
40-49	429	903908	0,0475%
50-59	1535	827714	0,1855%
60-69	4278	466019	0,9180%
70-79	11978	309141	3,8746%
80-89	26480	229827	11,5217%
90 e oltre	18144	93094	19,4900%
Totale	63029	5324975	1,1836%

i) Distintamente per i due sessi si costruisca, avvalendosi del software **R**, l'istogramma di frequenza per la variabile "età al contagio";
ii) distintamente per i due sessi si calcoli, sempre avvalendosi del software **R**, mediana, media e scarto quadratico medio della variabile "età al contagio". Si scelga una misura idonea di tendenza centrale fra le due calcolate e si motivi la scelta effettuata;
iii) mediante l'uso del software **R**, si ripetano le analisi di cui ai precedenti punti i) e ii) per la totalità dei casi segnalati;
iv) nel bollettino da cui provengono i dati riprodotti in tabella si afferma che (per maschi e femmine insieme) "l'età mediana dei casi è di 47 anni". Perché l'età mediana al contagio ottenuta al precedente punto iii) è diversa da quella fornita dall'Istituto Superiore di Sanità?
v) Si ripetano, sempre avvalendosi del software **R**, le analisi di cui ai precedenti punti i), ii) e iii) per la variabile "età al decesso" e si commentino i risultati ottenuti.

Nota bene 1: si chiuda l'ultima classe a 100 cioè si ipotizzi che non siano presenti nella popolazione italiana contagiati o deceduti per COVID-19 con età superiore ai 100 anni.
Nota bene 2: per il calcolo della mediana per dati organizzati in una distribuzione per classi, si usi la formula seguente:

$$me \simeq I_m + \left(\frac{0.5 - F_{m-1}}{F_m - F_{m-1}} \right) \Delta_m ,$$

dove

$I_m =$ estremo inferiore della classe mediana;

$F_m =$ frequenza relativa cumulata fino alla classe mediana

$F_{m-1} =$ frequenza relativa cumulata fino alla classe precedente a quella mediana

$\Delta_m =$ ampiezza della classe mediana.

Nota bene 3: per il calcolo della media e della varianza per dati organizzati in una distribuzione in classi, si faccia riferimento alle formule seguenti:

$$\mu = \frac{\sum_{i=1}^{k} x_{ci} n_i}{\sum_{i=1}^{k} n_i} \quad \text{e} \quad \sigma^2 = \frac{\sum_{i=1}^{k} \left(x_{ci} - \mu \right)^2 n_i}{\sum_{i=1}^{k} n_i} ,$$

dove

$x_{ci} =$ valore centrale della classe *i–ma*; le classi sono tutte di ampiezza 10 e i valori centrali sono 5, 15, 25, etc.

Buon Lavoro!

Prof. Giuseppe Espa & Prof. Enrico Tundis