



Dipartimento di Economia e Management  
Corso di Laurea in Economia e Management  
Esame di *Statistica, probabilità e inferenza* – a.a. 2022/2023  
Prof. Giuseppe Espa, Prof. Flavio Santi

Qui di seguito sono riportati i testi di cinque problemi da risolvere a casa. Una volta completati gli esercizi, al docente va restituito un file in formato **pdf** che contiene gli *script* prodotti, gli *output* ottenuti mediante il loro utilizzo nonché i commenti e le risposte richieste. Il file (completo di cognome, nome e numero di matricola da riportarsi a cura dello studente sulla prima pagina) va inviato al seguente indirizzo di posta elettronica:

[analisiatiespa@economia.unitn.it](mailto:analisiatiespa@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: “Statistica, probabilità e inferenza”.

**N.B. 2** Si faccia molta attenzione al fatto che l’esercizio 1), a differenza dei rimanenti 2), 3), 4) e 5), avrà risultati diversi per tutti gli studenti. Infatti, l’esecuzione della simulazione ivi richiesta prevede un numero d’innescò del generatore dei numeri pseudo-casuali uguale al numero di matricola di ciascuno studente.

1) Si sviluppino i seguenti punti:

- i) creare un vettore  $\mathbf{x}$  contenente i numeri 150, 151, 152, ..., 188, 189;
- ii) definire  $\alpha = -90$  e  $\beta = 0.8$ ;
- iii) creare un vettore  $\mathbf{ym}$  contenente i valori  $\alpha + \beta x$ ;
- iv) innescare il generatore di numeri pseudo-casuali con il proprio numero di matricola: `set.seed(numero di matricola)`. Ad esempio, immaginando che il numero di matricola dello studente Francesco Totti sia 999999, egli innescerà il generatore di numeri pseudo-casuali eseguendo l’istruzione `set.seed(999999)`;
- v) estrarre 40 valori da una normale con media 0 e deviazione standard 5 e inserirli nel vettore  $\mathbf{e}$ ;
- vi) creare un vettore  $\mathbf{y}$  (dei *valori osservati*) ottenuto sommando  $\mathbf{ym}$  ed  $\mathbf{e}$ ;
- vii) calcolare il coefficiente di correlazione lineare fra  $\mathbf{x}$  ed  $\mathbf{y}$ ;
- viii) disegnare lo *scatterplot* di  $\mathbf{x}$  (ascissa) e  $\mathbf{y}$  (ordinata);

- ix) interpolare un modello di regressione lineare considerando  $y$  come dipendente e  $x$  come indipendente; trovare le stime della pendenza e dell'intercetta; disegnare la retta di regressione;
- x) calcolare la devianza totale ( $TSS$ ), la devianza spiegata ( $SSR$ ), la devianza residua ( $SSE$ ), la varianza condizionata, il *residual standard error*  $s$ , il valore di  $R^2$ ;
- xi) calcolare l'errore standard del coefficiente di regressione lineare e l'intervallo di confidenza al 95% per la pendenza, verificando se il valore 0.8 è compreso all'interno dell'intervallo;
- xii) **centrare** le variabili  $x$  e  $y$  sottraendo da ciascun valore la rispettiva media; indichiamo, rispettivamente, con  $x^*$  e con  $y^*$  queste due nuove variabili (quindi  $x_i^* = x_i - \bar{x}$  e  $y_i^* = y_i - \bar{y}$ , dove con  $y_i$  abbiamo indicato l' $i$ -esimo *valore osservato* calcolato al punto vi); inserire i valori relativi alle variabili  $x^*$  e  $y^*$  rispettivamente nei vettori **xcen** e **ycen**;
- xiii) interpolare un modello di regressione lineare considerando **ycen** come dipendente e **xcen** come indipendente; verificare che la stima dell'intercetta è zero e che la stima della pendenza coincide con quella trovata al punto ix).

2) Caricare in **R** il file “2004 statewide crime.txt” allegato. Usando il “tasso di omicidi” (Murder) come variabile di risposta e il “tasso di povertà” (Poverty) e la “percentuale di diplomati all'*high school*” (HighSch) come predittori

- i) costruire ed interpretare i *partial regression plots*;
- ii) stimare il modello di regressione multipla. Riportare l'equazione di previsione ed interpretare le stime dei coefficienti;
- iii) ripetere l'analisi dopo aver eliminato l'osservazione relativa al distretto D.C.. Tale osservazione manifesta una pronunciata influenza sui risultati?

3) Con riferimento all'esercizio 30 del file “Esercizi (quinta parte)” disponibile *on line*, valutare la possibilità di utilizzare le variabili qualitative sconnesse *Sex* e *Race*, assieme agli anni di servizio  $X$  per prevedere il punteggio medio  $Y$  ottenuto dall'individuo nel questionario (i dati sono contenuti nel file “lavoro.txt”). In particolare:

- i) fornire una rappresentazione grafica dei dati distinguendo i punti dello *scatter* in base ai quattro livelli ottenuti combinando i livelli di sesso (**c1**) e razza (**c2**) e quindi commentare;
- ii) i termini di interazione tra  $X$  e i due predittori qualitativi **c1** e **c2** sono significativi o possono essere eliminati dal modello?
- iii) Riportare e commentare l'*output* del sottomodello eventualmente selezionato;
- iv) valutare i grafici di diagnostica del modello.

4) Un'azienda necessita di organizzare dei controlli a campione da eseguire con regolarità sulla qualità di una materia prima acquistata dai propri fornitori. Più precisamente, si vuole verificare che il valore della merce consegnata corrisponda effettivamente a quello dichiarato. Per questo motivo, l'azienda incarica due addetti di svolgere quotidianamente dei controlli a campione sui lotti di materia prima consegnata. I due addetti, una volta selezionato il lotto, ne accertano il peso e la qualità e, in base ad alcune tavole di conversione e ai prezzi di mercato del materiale, sono in grado di dare una valutazione circa lo scostamento tra il valore effettivo e quello dichiarato. Fatti i primi 2870 controlli, i due addetti si chiedono se e come sia possibile utilizzare i dati raccolti per orientare meglio i controlli, concentrandoli sui lotti che con più probabilità presentano delle irregolarità.

Il file “dati\_controlli.csv” contiene i dati relativi ai controlli fatti finora con le informazioni relative all'identificativo del lotto di riferimento (variabile “lotto”), alla qualità dichiarata della merce (variabile binaria “prima” che assume valore 1 se la merce è di prima categoria e 0 nel caso in cui sia di seconda categoria), alla provenienza del lotto (variabile dicotomica “ExtraUE” che assume valore 1 se il lotto non proviene da un paese UE, e 0 in caso contrario), al valore dichiarato del lotto (variabile “dichiarato”), e infine all'esito del controllo (variabile dicotomica “Irr” che assume valore 1 se il valore effettivo si è scostato da quello dichiarato, e 0 in caso contrario).

- i) Mediante la regressione logistica, si studi la relazione tra la variabile dicotomica sulla regolarità dei lotti (“Irr”) e le variabili esplicative sulla qualità della merce, la provenienza, e il valore dichiarato del lotto. Ricordarsi di allegare alla relazione lo script **R** con i comandi necessari ad ottenere le stime.
- ii) Prevedere la probabilità che un lotto proveniente dall'Unione Europea di merce di prima categoria e un valore dichiarato di 200 € sia irregolare.
- iii) Estendere il modello precedente includendo un'interazione tra la variabile “prima” e la variabile “ExtraUE”. Il nuovo modello è preferibile al precedente? Confrontare i due modelli tramite il test del rapporto delle verosimiglianze.
- iv) Nel pianificare i controlli, avendo a disposizione tutte le informazioni su ciascun lotto in consegna (eccetto la sua eventuale irregolarità, naturalmente) è possibile aiutare i due addetti a indirizzare i propri controlli in modo da renderli più efficienti? (ad. es.: è meglio controllare i lotti con un elevato valore dichiarato, oppure quelli con un valore basso? Meglio i lotti provenienti dall'UE o quelli provenienti da paesi non UE?)

5) 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 esplicita 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 (M)				Femmine (F)			
Classe di età	N. deceduti	N. casi	Letalità (%)	Classe di età	N. deceduti	N. casi	Letalità (%)
0-9	7	444638	0,0016%	0-9	10	413787	0,0024%
10-19	15	667351	0,0022%	10-19	13	644080	0,0020%
20-29	61	697747	0,0087%	20-29	34	694093	0,0049%
30-39	222	675651	0,0329%	30-39	128	743312	0,0172%
40-49	948	791996	0,1197%	40-49	429	903908	0,0475%
50-59	3808	778178	0,4893%	50-59	1535	827714	0,1855%
60-69	10785	469786	2,2957%	60-69	4278	466019	0,9180%
70-79	24300	305004	7,9671%	70-79	11978	309141	3,8746%
80-89	31274	163360	19,1442%	80-89	26480	229827	11,5217%
90 e oltre	9870	32024	30,8206%	90 e oltre	18144	93094	19,4900%
Totale	81290	5025735	1,6175%	Totale	63029	5324975	1,1836%

- Si confrontino, utilizzando un idoneo test statistico, le due letalità  $p_1$  e  $p_2$  ( $1 = \text{"M"}$  e  $2 = \text{"F"}$ ). Si costruisca l'intervallo di confidenza per la differenza  $(p_1 - p_2)$ . Si commentino dettagliatamente i risultati ottenuti;
- si confrontino, utilizzando un idoneo test statistico, le due età medie al contagio  $\mu_1$  e  $\mu_2$  ( $1 = \text{"M"}$  e  $2 = \text{"F"}$ ). Si costruisca l'intervallo di confidenza per la differenza  $(\mu_1 - \mu_2)$ . Si commentino dettagliatamente i risultati ottenuti;
- si ripeta l'analisi di cui al precedente punto 2. per la variabile "età al decesso";
- eseguendo una analisi per sesso, si può concludere che il fenomeno del contagio è, a livello di popolazione, lo stesso nelle diverse classi di età? Sottoporre a verifica statistica questa affermazione con un idoneo test valutando poi la *natura* e la *forza* dell'eventuale associazione diagnosticata;
- si ripeta l'analisi di cui al precedente punto 4. Considerando, però, l'"età al decesso".

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 media campionaria e della varianza campionaria per dati organizzati in una distribuzione in classi, si faccia riferimento alle formule seguenti:

$$\bar{x} = \frac{\sum_{i=1}^k x_{ci} n_i}{\sum_{i=1}^k n_i} \text{ e } s^2 = \frac{\sum_{i=1}^k (x_{ci} - \bar{x})^2 n_i}{\sum_{i=1}^k n_i - 1}$$

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.

Nota bene 3: per risolvere il punto iv) è necessario costruire la tabella di contingenza appropriata.

Buon lavoro!

Giuseppe Espa & Flavio Santi