Politecnico di Milano
II Scuola - Ingegneria dei Sistemi (MI)
Appello di Statistica Applicata
13 settembre 2012

Nome e cognome:
Numero di matricola:

Problema 1

Nel file `library.txt` sono riportate per le 23 biblioteche comunali milanesi le spese (riferite all'anno 2011) relative all'acquisto libri, alla retribuzione del personale, al consumo di energia elettrica e infine ad altri costi. Assumendo *iid* normali i dati relativi alle 23 biblioteche:

a) si forniscano 5 intervalli di confidenza globale 90% per la media delle quattro voci di spesa e per la loro somma;

b) si forniscano 5 intervalli di confidenza globale 90% per la deviazione standard delle quattro voci di spesa e per la loro somma;

c) si confermi/smentisca l'ipotesi secondo la quale la spesa media in libri copre la metà della spesa totale.

Problema 2

Nel file `genes.txt` sono riportati i livelli di attivazione dei geni X19-A e X19-C registrati per 2000 individui arruolati in un esperimento clinico controllato.

a) Utilizzando un algoritmo k-medie si individuino eventuali cluster presenti nei dati (in particolare, a seguito di un'ispezione grafica dei cluster, si scelga il valore di k che ritenete più opportuno e si riportino le numerosità dei cluster individuati).

Utilizzando l'appartenenza al cluster come fattore di raggruppamento:

b) si esegua una MANOVA per verificare l'effettiva presenza dei cluster;

c) si eseguano due ANOVA (una per ciascun gene) per caratterizzare l'appartenenza ai cluster;

d) si commentino brevemente i risultati delle analisi.

Problema 3

Nel file `exams.txt` sono riportati i voti degli scritti di 32 studenti che hanno sostenuto tutti e quattro gli ultimi appelli di statistica applicata.

a) Vi è evidenza statistica al 5% per affermare che il voto medio sia cambiato nel tempo?

b) Alcuni studenti sostengono che il quarto appello fosse più difficile del terzo. Utilizzando i voti come un indicatore di difficoltà dell'appello, vi è evidenza statitica al 5% per provare quanto affermato?

c) Si esplori e commenti la struttura di covarianza del dataset tramite un'analisi delle componenti principali.

Problema 4

Nel sistema fognario di Mouseville vive da sempre il ratto grigio. Nel 2001 è stato introdotto accidentalmente il ratto rosso. A seguito della sua diffusione, nel 2004 è stata intrapresa una campagna mirata all'eliminazione della specie aliena. Nel file `mouse.txt` sono riportate le date degli avvistamenti relativi al periodo 2000-2011 di ratti rossi e grigi ($t = 0$ indica l'inizio dell'anno 2000 e $t = 12$ l'inizio dell'anno 2012). Si indichino con p_R e p_G rispettivamente la proporzione di ratti rossi e grigi rispetto all'intera popolazione di ratti e si assuma un'evoluzione nel tempo delle proporzioni del tipo: $\log(p_R/p_G) = at^2 + bt + c$ con $t \in (0; 12)$.

a) Utilizzando un modello di regressione logistica, si stimino i 3 parametri del modello. Si riporti inoltre su di un grafico l'andamento stimato di p_R e p_G.

b) Secondo il modello stimato al punto (a), qual è la proporzione stimata di ratti rossi all'inizio del 2004?

c) Stimate che la proporzione di ratti rossi abbia mai superato il 50% della popolazione dei ratti? Se sì, in che periodo?

d) Quando stimate che il ratto rosso abbia raggiunto la sua massima diffusione in termini percentuali? Con che valore percentuale?