Politecnico di Milano
Scuola di Ingegneria Industriale e dell'Informazione
Applied Statistics
September 12th, 2019

First Name and Family Name:
ID Number:

Problem n.1

Along the centuries, the Yosemite National Park (US) has been regularly affected by vast fires, with strong impact on its forest. The file `sequoia.txt` reports the heights [m] and diameters [m] of 294 giant sequoias of the park.

a) Perform a cluster analysis of the trees by using a hierarchical clustering method (Euclidean distance and Ward linkage). Report the number of clusters you deem appropriate to represent the data, the centroids of the clusters, and their size. Suggest a possible interpretation in terms of the number of vastest fires that dramatically affected the growth of the forest.

b) Provide Bonferroni intervals (global level 90%) for the mean and the variances of the diameter of the trees, within each of the clusters identified at point (a). Introduce and verify the appropriate assumptions. Comment the results.

Problem n.2

To possibly patent a modified version of their last car engine, the racing car team PoliFast&Furious is performing a trial made of a series of identical and independent stress tests on 10 modified and 5 unmodified car engines. For each stress test, 8 performance indicators are recorded in a discrete scale from 0 to 10 (with 0 indicating the worst possible outcome and 10 the best possible one). The recorded data are reported in the file `stress.txt`.

a) For each performance indicator, perform a permutation one-sided test to look for possible statistical improvements of one or more performance indicators. In detail, for each indicator, use the difference of the sample medians as test statistic and use 10000 random permutations with random seed equal to 123 to estimate the permutational distribution. Report the value of the eight test statistics and their corresponding p-values.

b) To limit investment risks, the management company policy is to limit the false discovery rate of each trial to a maximum value of 25%. According to this policy, which are the performance indicators that can be considered statistically improved?

Problem n.3

The last American space mission measured the position in space of several debris of two collapsing asteroids. In the file `debris.txt`, the coordinates (x, y) of 300 debris are reported. Debris were classified by the experts in two types $\{L, H\}$ based on the risks of interaction with Earth observation satellites (L = low risk, H = high risk).

a) Build a classifier for the type of debris based on its spatial coordinates. Report the model for the data, the estimates of its parameters (means and covariances), the priors within the groups and verify the model assumptions. Report a qualitative plot of the classification regions.

b) Analyse the possible weaknesses of the model. Estimate the AER of the classifier through leave-one-out cross-validation.

c) Build a k-nearest neighbor classifier for the type of debris based on its spatial coordinates. Choose parameter k in the range $[10, 30]$ as to optimize the misclassification error, assessed via leave-one-out cross-validation (set the random seed equal to 321 prior to perform cross-validation). Report the error rate associated with the optimal classifier, and a qualitative plot of the classification regions.

d) Using the best classifier among those at points (a) and (c), how would you classify a new debris observed at position (1,-4)?

Problem n.4

The file `mickey.txt` reports the daily average waiting times [min] when queueing for taking a picture with Mickey Mouse at a Disneyland Park in the United States, recorded during 245 days of 2018. Consider for the daily average waiting time y the following model

$$y = \alpha_g + \beta_g \left(1 + \cos\left(\frac{4\pi}{365} t\right)\right) + \epsilon,$$

where t is the day of the year, g denotes the day of the week ($g = 1$ for weekends, $g = 0$ for weekdays), and $\epsilon \sim N(0, \sigma^2)$. Identify the first term α_g with the contribution to the queue due to the residents, and the second term with the contribution to the queue due to the tourists.

a) Estimate the 5 parameters of the model (α_0, α_1, β_0, β_1, σ). State and verify the model assumption and interpret the parameters.

b) Perform a statistical test of level 95% to verify whether the variable *day of the week* has a significant effect on the daily average waiting times.

c) Perform a model reduction, by using appropriate statistical tests of level 95%. Update the estimates of the parameters.

d) Perform a statistical test of level 95% to verify if the maximum of the mean waiting times is 60 minutes. If needed, update the estimates of the parameters.

e) Provide a prediction interval of level 95% for the mean waiting time expected for Monday 26 August 2019 ($t = 238$). Comment the results and highlight the assumptions used to obtain the result.