



## **M.Sc. in Data Science**

**Course:** Probability and Statistics for Data Analysis

**Semester:** Fall 2022

**Instructor:** Ioannis Vrontos (vrontos@aueb.gr)

**Grader:** Konstantinos Bourazas (kbourazas@aueb.gr)

## **Assignment 2**

**Deadline: 3 January 2023**

**Note:** Use R in this assignment and submit your .R code that was used to answer the questions, along with a small report where you will present plots and results for each question of this assignment.

1. A study of the effect of two drugs on the reduce of cholesterol used 100 volunteers who tested the drugs. 50 of them were randomly selected to take the first drug (A), while the rest 50 took the second one (B). We measured the cholesterol after the received of the drugs, while we tested the presence of Myalgia symptoms and we measured the Glucose, in order to check the side effects of the drugs. The observations are in the file

“cholesterol.txt” (available on e-class assignments site).

- (a) Provide a 99% confidence interval for Cholesterol values.
- (b) Provide a 95% confidence interval for Cholesterol values, after receiving the drug A and B respectively.
- (c) Provide a 90% confidence interval for mean difference of the Cholesterol values, after receiving drug A and drug B respectively.
- (d) Examine the following hypothesis test:

$$H_0 : \mu_A = \mu_B \quad H_1 : \mu_A < \mu_B$$

where  $\mu_A$  and  $\mu_B$  are the mean Cholesterol values after receiving drug A and B respectively. The level of significance is  $\alpha = 0.05$ .

- (e) Provide a hypothesis test ( $\alpha = 0.01$ ) for the equality of variances of the Glucose after receiving drug A and drug B respectively.
- (f) In a significance level of 5%, test if there is statistical significant side effect for the Glucose.
- (g) Provide a 95% confidence interval for the proportion of the volunteers who had myalgia symptoms.
- (h) Test if the proportion of the volunteers who had myalgia symptoms is statistically greater than 5% in level of significance 5%.
- (i) Test if the drug and the myalgia symptoms are independent ( $\alpha = 0.05$ )
- (j) Provide a 95% confidence interval for the mean difference  $\mu_1 - \mu_2$ , where  $\mu_1$  and  $\mu_2$  are the mean of Glucose with or without myalgia symptoms respectively.

**2.** In file “data2.txt” (available on e-class assignments site) you will find

the recorded variables Y, X1, X2, X3, X4 (continuous) and W (categorical with three levels) on 150 cases. Using these data answer the following questions:

- (a) Run the parametric one-way ANOVA of each of the continuous variables (Y, X1, X2, X3, X4) on the categorical variable (W). Specifically,
  - (i) provide a graphical representation of each of the continuous versus the categorical variable
  - (ii) provide the ANOVA output
  - (iii) check the assumptions.
- (b) Provide a scatter-plot matrix of Y, X1, X2, X3, X4, annotating the different levels of W in each plot using a different color.
- (c) Run the regression model of Y on X4
- (d) Run the regression model of Y on all the remaining variables (X1, X2, X3, X4, W), including the non-additive terms (i.e. interactions of the continuous predictors with the categorical).
- (e) Examine the regression assumptions and provide alternatives if any of them fails.
- (f) Use the “stepwise regression” approach to examine whether you can reduce the dimension of the model.
- (g) Using the model found in (f), provide a point estimate and a 95% confidence interval for the prediction of Y when: (X1, X2, X3, X4, W) = (120, 30, 10, 90, B)
- (h) Using the cut() function, create a categorical variable (named Z) with 3 levels based on the quantiles of X4. Provide the contingency table of X4 and W.

(i) Run the parametric two-way ANOVA of  $Y$  on the categorical variables  $W$  and  $Z$  (including the interaction term). Provide the fit, examine the assumptions and comment on the significance of the terms.