STAT 210
Applied Statistics and Data Analysis
First Exam

October 22, 2022

This exam is open notes and open book but not open internet. You are not allowed to surf the internet or look for answers to the questions

You are reminded to adhere to the academic integrity code established at KAUST.

Show complete solutions to get full credit. Writing code is not enough to answer a question. Your comments are more important than the code. Label your graphs appropriately

Please identify the files you submit with your surname

Question 1 (40 points)

The file `motorskills` has information on an experiment to study the motor skills of children and youngsters. In the experiment, the speed with which subjects placed a series of cylinders into a set of holes was measured. The data set has five variables: `Age` in months, `Gender`, `DH`, which denotes the Dominant Hand, i.e., whether the dominant hand is the right or the left hand, `sp1`, which corresponds to the speed with the dominant hand and `sp2` which corresponds to the speed with the non-dominant hand. The speed is measured in cylinders per second. Use $\alpha = 0.02$ for all tests in this question.

Load the data and store it in a data frame named `df1`.

(a) (5 points) Draw boxplots for the speed with the dominant hand as a function of `DH` (dominant hand) and also for the speed with the non-dominant hand as a function of `DH`. Use a common scale. Comment on what you observe.

(b) (10 points) We want to explore if the speed with the dominant hand is the same as with the non-dominant hand. What parametric test or tests are appropriate here? What are the assumptions, and why do you think they are satisfied? Carry out this test or tests and comment on the results.

(c) (5 points) Define a new variable named `dif` in the data frame. The value for this variable is the difference between the speed with the dominant hand minus the speed with the non-dominant hand. Do a boxplot of this new variable as a function of `DH`. Comment on the graph.

(d) (10 points) We now want to test whether the average value for this difference in speed (`dif`) is the same for right-handed and left-handed subjects. What parametric test or tests are adequate here? What are the assumptions, and why do you think they are satisfied? Carry out this test or tests and comment on the results.

(e) (10 points) For the problems in (b) and (d), what non-parametric tests can be applied? What are the assumptions for these tests? Why do you think they are satisfied? Carry out these tests and comment on the results.

Question 2 (25 points)

The data set `bloodpress` has information on blood pressure for 321 males over 20 years old. The set has two variables, `Age`, the age of the subject in years, and `BP`, the blood pressure classified into three levels, `Low`, Normal, and High. Read the data and store it in a file named df2.

(a) (3 points) Check whether BP has been stored as a factor. If not, transform it into a factor. The levels should be in the order Low, Normal, and High. If BP has been stored as a factor, verify if the levels are in the correct order as stated above. If they are not, modify the variable so that they are.

(b) (4 points) Boxplot Age as a function of BP. Comment on what you observe.

(c) (5 points) Using the information in Age, add a factor fAge to df2 created according to the following rule: if the subject has less than 30 years, the value for the factor is Under30. If the subject is between 30 and 49 years old, the value is 30-49, and if the subject is 50 or more, the value is Over50. One way to do this is using the function cut.

(d) (5 points) Produce a table of fAge and BP and do a mosaic plot. The table should have fAge as columns. Comment on what you observe. Produce a second table with proportions calculated relative to the different age levels. Again, comment on what you observe.

(e) (8 points) We want to determine whether the blood pressure levels are homogeneously distributed in the age groups we created. Which test or tests do you know that can be used for this? What are the underlying assumptions? Are they satisfied in this case? Carry out all the tests you mentioned and discuss the results. What are your conclusions?

Question 3 (35 points)

A pharmaceutical company did an experiment to compare three different pain relievers for treating migraines. The data is stored in the file migraine. In the experiment, 27 volunteers participated, and nine were randomly selected for each pain reliever. The subjects were instructed to take the drug during their next migraine headache episode and to report their pain on a scale of 1 = no pain to 10 = extreme pain 30 minutes after taking the drug.

Read the data file into a data frame named df3. Make sure the data are read correctly. If Drug has character mode, transform it into a factor.

(a) (4 points) Do boxplots for Pain as a function of Drug (all the boxplots should appear on the same panel). Add the points to this graph. Comment on what you observe.

(b) (8 points) Fit an analysis of variance model to this data using the function lm and print the anova table. Use $\alpha = 0.02$ for your test. What do you conclude from this analysis?

(c) (10 points) Using the function summary, obtain a summary table for the model fitted in (b). What is the meaning of the numbers in the Estimate column? Obtain the estimate for the mean response for each treatment level from this table.

(d) (5 points) What are the estimated values for the variance and standard deviation of the errors in this experiment?

(e) (8 points) What are the assumptions on which the analysis of variance model is based? Draw diagnostic plots for checking these assumptions and discuss the results.