STAT 210
Applied Statistics and Data Analysis:
Homework 4

Due on Oct. 2/2022

Question 1

The data for this question are stored in the file `hw4q1` and correspond to an experiment to measure the effect of a new drug in the memory of patients in a nursing home. The patients were tested for memory before the treatment started and again after one month taking the drug. The dataset has two variables, `mem`, the score in the test, and `type` with two values, `before` for the initial score and `after` for the final score

(a) Load the dataset and check whether `type` is stored as a factor. If it is not, transform it into a factor.

(b) Draw boxplots for `mem` according to `type` and comment.

(c) Draw a scatterplot of the memory score `after` versus the memory score `before` and comment on what you observe. Do you think the two scores are independent?

(d) We want to determine whether the treatment had an impact on the memory score of the patients. State clearly the statistical hypothesis that you want to test. What test or tests would you consider adequate in this situation and why? What are the assumptions? Are they satisfied in this case? Carry out all appropriate tests for this problem and comment on your results.

Question 2

We will use the data set `Pima.te` in the `MASS` package for this question. Open the help file for this data set and get acquainted with it. We are going to focus on two variables, `bp` and `type`.

(a) Divide the plotting window into two regions, one single column with two rows, and plot histograms for `bp` for types `Yes` and `No`. Since you want to use these graphs for comparing the two populations, use the same scales in both cases. Use reasonable labels for the axes and a title indicating the corresponding type. Make sure that the area for the figure is large enough so that the histograms are clearly seen. Compare the two graphs and comment on similarities and differences.

(b) Boxplot blood pressure as a function of `type` and comment on the graph. Make sure you have a single plotting window with both boxplot.

(c) Calculate mean and standard deviation for both types and find how many subjects of each type are there in the dataset.

(d) We want to determine if the pulse rate for diabetic women is significantly different from a reference value of 70 mm Hg. What (parametric) statistical test do you think is appropriate in this case? Carry this test out and discuss your results. Describe the assumptions you need for this test to be valid and check whether they are satisfied by the data set.

(e) We now want to compare the two populations (Pima women with and without diabetes) to see if there is a difference in the average blood pressure. What (parametric) test would you perform in this case? What assumptions are needed? Do they look reasonable in this case? Carry out this test and discuss your results.

(f) What non-parametric tests would be adequate for parts (d) and (e)? Carry this test out and compare your results with the tests in (d) and (e).