STAT 210
Applied Statistics and Data Analysis
Group Session 4

Fall 2020

The questions on this list come from previous exams. At the group meetings we will only address those items that directly relate to what we have covered this week. I have posted the complete questions so you can get an idea of how the test questions went in the past.

Exercise 1

In the meeting we will focus on sections (viii), (ix) and (x).

The data for this exercise is in the file `q1.data` and come from a study of adult patients in a detoxification unit in the USA. Read the data set and store it as a data frame named `dframe.ex1`.

(i) Explore the structure of this data set. How many variables are there? How many are categorical?

(ii) In what follows you will use only the variables `pcs, mcs` and `female`.
Create a new data frame named `df.q1` with these variable, using the same names.

(iii) Variables `mcs` and `pcs` stand for 'mental component score' and 'physical component score'. The variable `female` is the gender of the subject, with code 0 for male and 1 for female. Add a new categorical variable named `gender` with values `m` and `f` for male and female, respectively, to `df.q1`.

(iv) We want now to explore the variables `mcs` and `pcs` and compare their values for both genders. Calculate means and standard deviations for both variables (`mcs` and `pcs`) according to `gender`.

(v) Draw histograms for `mcs` according to `gender`. Recall that the purpose is to make comparisons between the two populations. Repeat for `pcs`. Comment on your findings.

(vi) Draw boxplots for `mcs` according to `gender`. Both boxplots should appear in the same graph window. Repeat for `pcs`.

(vii) Draw normal quantile plots for `mcs` according to `gender` and discuss the results. Repeat for `pcs`.

(viii) Do a test to compare whether the means for `mcs` are equal in the two populations (male and female). Explain carefully the assumptions you are making. Do you think they are justified by what you have seen in your exploratory analysis? Discuss your results. Do you think a paired test would make sense in this context? (why or why not).

(ix) Do now a non-parametric test for the same variables. Compare with your previous result and comment.

(x) Repeat (viii) and (ix) for `pcs`. Comment on your findings.

Exercise 2

We will focus on sections (i), (ii), (vi) and (vii).

Use the same data set as for Exercise 1

The data come from a study of adult patients in a detoxification unit in the USA. We focus on two variables, `homeless` and `age` and we want to study the relation between them. The first variable has two values, `0` (no)

and `1` (yes) which correspond to answers to the item 'one or more nights on the street or on a shelter in past 6 months'.

(i) Use the function `cut` to divide age into four groups using the option `breaks`. Label the age categories `A1,A2,A3` and `A4`.

(ii) Using the function `table`, produce a table of the categorized `age` variable and `homeless`.

(iii) Do a bar chart for this table. The variable `homeless` should appear on the x axis. Use the option `beside = TRUE`. For each value of `homeless` there should be four bars corresponding to the four age classes. Comment on the result

(iv) Draw another bar chart with the roles of the variables interchanged. The function `t()` which transposes a matrix may come in handy for this. Comment on your results.

(v) We now wish to build a table with proportions instead of frequency counts. We want to have for each category of `homeless`, the proportion that each age group represents in that category. For each `homeless` category, these proportions should add up to one. Use the functions `apply` and `sweep` to do this.

(vi) Go back to the table you produced in (ii). We want to test whether the age distribution is the same for both `homeless` categories. Which tests do you know that can be used for this? What are their underlying assumptions? Which one would you prefer and why? Carry out the tests (all of them) and comment on the results. What are your conclusions?

(vii) Do you think the procedure we have followed to analyze the relation between `homeless` and `age` is reasonable? What are the weak/strong points of what we have done?

Exercise 3

In this exercise we focus on (v) and (vi)

Consider the data set

```
a = c(13.9, 14.5, 13.8, 16.6, 18.2, 20.2, 13.6, 16.3, 15.4, 12.3)
b = c(13.7, 14.6, 13.0, 16.2, 17.8, 21.0, 13.0, 16.9, 17.0, 12.1)
q2.df <- data.frame(before = a, after = b)
```

You can copy and paste the previous instructions to create `q2.df`.

These values represent the time taken for 10 student athletes to perform the same physical task, `a` or `before` are times before training while `b` or `after` corresponds to time after training. Training was supposed to improved the athletes performance at this task.

(i) Calculate the summary statistics for the two vectors, `before` and `after` and compare them. Do they seem different?

(ii) Calculate the differences between the times before and after training. Do a barplot of your results and comment.

(iii) Do a scatterplot of `before` in the x axis against `after` in the y axis. Add the line $y = x$. What does this graph show? Can you conclude anything about the effectiveness of the training method?

(iv) Do normal quantile plots for `before`, `after` and the difference `after-before`. Comment on your results.

(v) You are asked to compare the average time taken by the students before and after training to assess if there has been an improvement. Which test would you perform? What are the underlying assumptions? Why do you think they are satisfied? What hypotheses are you testing? Carry out the test and comment on the results.

(vi) The Wilcoxon non-parametric test can be used with paired and non-paired data. Which option would you choose for this data set? Carry out this test and comment on your results. (NB: If you want to do a Wilcoxon paired test you need to add the option `paired = TRUE` when you call the test).