MSc in Statistical Science
Week 2 Practical, MT 2020

This practical sheet contains two exercises. Any queries you have about Exercises 1 and 2 may be directed to the Teaching Assistant. Exercise 3 is meant to be for independent study, so the TA will not answer questions about it, with the sole exception of questions relating to a limited number of programming issues. Although you shouldn't hand anything in, it would be beneficial if you wrote a brief report with your findings on exercise 3, to get practice for the upcoming practicals.

Model solutions will be provided during the report writing session in wk 2 and you will have a chance to ask any questions.

1 Ozone Data (NOT ASSESSED)

The data 'Ozone.dat' (Cole and Katz, 1966) contain oxidant content of dew water in part per million ozone. The samples of dew were collected during the period August 25-September 13, 1960, at Port Burwell, Ontario. The file contains the resulting oxidant content.

1. Import the data by using

```
> ozone = scan("ozone.dat")
```

2. Start by a preliminary analysis of the data, use the functions `min()`, `max()`, `mean()`, `sd()`, `quantile()`. Then use `summary()` to obtain directly some of this information.

3. Look at the boxplot of the data to get an idea of the centre of location and variability of the data. Use `boxplot()`.

4. Check the normality of the data. Relevant graphical displays are the normal probability plot (use the functions `qqnorm()` and `qqline()`), and a comparison of the empirical CDF (use `ecdf()`) to the CDF of a Normal distribution. Try this

```
> qqnorm(ozone)
> qqline(ozone)
```

and also this

```
> plot(ecdf(ozone), cex = 0.5)
> x = seq(0, 4, 0.01)
> lines(x, pnorm(x, mean(ozone), sd(ozone)), col = "red")
```

What can you say about the normality of the data? What are the things you notice?

5. You can also write your own function to obtain a normal probability plot. Recall that we need to plot the ordered observations against the normal quantiles.

```
> my.qqnorm = function(data.vec){
+
+   n = length(data.vec)
+   norm.quantiles = qnorm((1:n) / (n + 1))
+   plot(norm.quantiles, sort(data.vec), main = "normal quantile plot",
+     xlab = "Theoretical quantiles", ylab = "")
+
+ }#MY.QQNORM
```

6. Assuming the normality of the data, test the hypothesis that the mean oxidant content of dew water is 0.25 against the alternative that it is greater than 0.25. Compute first the value of t_{obs} and then compute the p-value p_{obs}. Recall that `pt(`t_{obs}`,n-1)` computes $P(T \leq t_{\text{obs}})$, where T is a Student t distribution with $n-1$ degrees of freedom.

7. Have a look at the plot of the Student t distribution with $n-1$ degrees of freedom. Try

   ```
   > curve(dt(x, 11), -5, 5)
   ```

   Place t_{obs} in the graph as

   ```
   > points(tobs, 0, cex = 1, bg = 7, pch = 4)
   ```

8. Compute the lower a limit of the confidence interval $(a, +\infty)$ for the mean at the 0.95 level.

9. Now check your results against the results of `t.test()`. In this case you need to specify `mu` and the `alternative`.

10. Perform now a Wilcoxon test and compare its results with the previous test. Use `wilcox.test()` and remember to specify `mu` and the `alternative`. Why the test cannot compute the exact value of the p-value?

11. As further exercise, you can test the hypothesis that the mean oxidant content of dew water is 0.25 against the alternative that it is not 0.25, at the 0.90 level.

2 Bleeding Time Data (NOT ASSESSED)

The dataset 'bleeding.dat' contains a subset of the data of Adams and Schmalhorst (1976) who studied the reactions of normal subjects to aspirin. The X observation for each subject is the bleeding time (in seconds) before ingestion of 600 mg aspirin and the Y observation is the bleeding time (again in seconds) 2h after administration of aspirin.

1. Import the data by using `read.table("bleeding.dat",h=T)`.

2. Start as before with a preliminary analysis of the data. Look at the boxplot, probability plots, histogram, empirical cumulative distribution function to have a sense of the data. What do you think of the centre of location, variation of the data and normality assumption of the data?

3. Use also `qqplot()` to obtain the q-q plot of X and Y.

4. Test the hypothesis that a 600-mg dose of aspirin has no effect on bleeding time versus the alternative that it leads to an increase in bleeding time. Assuming the normality of the data, try first to compute t_{obs}, p_{obs} and a confidence interval with level 0.95 writing your own function and then compare your results with the results of `t.test()`. Which `alternative` do you have to set in this case? Which value do you use for `mu`?

5. Use now the Wilcoxon test. Do you find any difference with the results of the t-test? Which one would you prefer and why?

6. Use the command `power.t.test` to compute the power of a t-test with true difference in mean of 2 minutes (120 seconds) bleeding time for a significance level $\alpha = 0.05$ with 14 subjects in he study. Here we assume that the standard deviation of differences within pairs is 175. Then, compute the power for different number of subjects and determine the smallest number of subjects we would need to recruit for the study in order to obtain a power at least equal to 0.95.

3 Finger Tapping Data (MOCK ASSESSED)

Draper and Smith (1981) reported data from a double-blind experiment carried out to investigate the effect of caffeine on performance on a simple physical task. Thirty college students were trained in finger tapping and then divided at random into three groups of 10. Each group received different dose of caffeine: 0, 100 or 200ml. Two hours after treatment, each student was required to do finger tapping and the number of taps per minute was recorded. The question of interest is whether caffeine affects performance on this task. You are asked to analyse the Finger Tapping Dataset and discuss your findings in your report. The points below are here to guide you through the statistical analysis.

Suggestions for the analysis:

Get the data into a form which is useful for exploring the question of interest. For example:

```
> finger = read.csv("caffeine.csv", header=T)
```

1. Start by looking at some summaries of the data such as the minimum, maximum, mean, median, mode and standard deviations of the speed of finger tapping in each group to see if there may be a basis to the claim that caffeine impacts finger tapping.

2. Create a quality histogram display and/or box plot that you can use to compare the distributions of the data in each group. Some things to keep in mind are

   (a) Is it easy to compare the distributions for different groups?
   (b) Is the comparison fair (unbiased)?
   (c) Are the axes appropriately labelled and scaled?

   Remarks: you may want to use the function `par(mfrow=c(3,1))` to place histograms one under the other. The scale of the x-axis (resp. y-axis) can be adjusted using the `xlim` (resp. `ylim`) option. In addition, you can improve these plot by adding appropriate labels.

3. Suppose now we want to test if there is a significant difference in mean speed of finger tapping between the group who received no caffeine and the one who received 100ml of caffeine.

    - Before using a t-test, we may may want to first assess whether the samples are consistent with a normal distribution and to test the hypothesis that the groups have equal variance. This can be done using relevant graphical displays and/or hypothesis testing.
    - Use the `t.test()` function to carry out an appropriate test of the null hypothesis that the mean speed of finger tapping is not increased by caffeine. Make sure to specify the options correctly. Should you use a test assuming equal variances? Does this decision affect your conclusions for this example?
    - Perform a Wilcoxon (Mann-Whitney) test of equal location, using `wilcox.test()`. Compare the results of this test to those of the t-tests and critically assess the validity of each.
    - Assuming that the standard deviation of the distributions for both the group who did not receive any caffeine and the group who received 100ml of caffeine is equal to 2.2, use the command `power.t.test` to compute the power of a t-test with true difference in mean of 1.6 tap per minutes for a significance level $\alpha = 0.05$ when there are 10 students per group. Then, compute the power for different number of students and determine the smallest number of students we would need to have in each group in order to obtain a power at least equal to 0.8. What do you conclude?

4. Perform a similar analysis to test whether there is a significant difference in mean speed of finger tapping between the group who received no caffeine and the group who received 200ml of caffeine.