MAST20031 Analysis of Biological Data – Assignment 1

Instructions

- The assignment contains 3 problems worth a total of 25 marks.
- Your assignment must be handed in by Friday 3rd April at 11.59pm.
- Your assignment should clearly show your student ID number, your tutor's name, and the time and day of your assigned tutorial class.
- Assignments submitted late will incur a penalty of 1% per hour. If you have exceptional circumstances that prevent you from meeting the deadline, please email Paul and Luke well in advance (i.e. not at the weekend), and we may be able to grant an extension. Note that there have been reports of internet connection trouble due to the COVID-19 measures, so we recommend that you don't leave submission until the last moment.
- Tutors may not help you directly with assignment questions. They may, however, provide some appropriate guidance.
- It is recommended that you use R for Problems 2 and 3 whenever appropriate.
- You may report graphs, results of analyses and main working of the problems in your answers. Lengthy, or not strictly necessary, R code should be attached at the end of your assignment as an appendix.
- Comments should be brief and concise: two sentences will suffice in most cases. *There is a **page limit** of 10 single-sided pages (including graphs, but excluding the appendix).* Pages exceeding this limit will not be assessed.

Leaf data

Problems 2 and 3 use the data collected by the class of 2019 in Data Collection Exercise 2, which is available on Canvas (go to Modules, then click Data Collection Exercise 2). The data consist of 2731 rows, with each row containing measurements on a particular leaf. Observed characteristics for each leaf are: side of the tree (north or south), length and width. Each student provided measurements of 5 leaves – 5 from the north, and 5 from the south (see the "student" and "rep" columns).

The rationale for collecting the data was to test the idea that leaves grow differently, depending on whether they are on the sunny northern side of the tree, or on the shaded southern side of the tree. On the north side, where sun is abundant, water loss might be a bigger issue for the leaf than access to sun. On the southern side of the tree the reverse is likely true. Because of this, we might expect the mean leaf size on the south to be different to the mean leaf size on the north.

Problem 1 [6 marks]

(a) **[3 marks]** Your original dataset (DCE2) contains four variables of interest (ignore `rep`). Name them and for each variable identify the type of variable, and describe its sample space.

(b) **[3 marks]** What is your statistical population in this study? Describe, in less than three sentences, how you achieved a random sample from that population.

Problem 2 [10 marks]

In this problem you will work with the raw data on leaf widths. W denotes the width of individual leaves.

(a) **[2 marks]** Compute and report numerical summaries on W (each summary containing min, max, 1st, 2nd and 3rd quartiles + mean) for the north and south sides of the tree.

(b) **[2 marks]** Produce and report a paired box plot (two boxplots within a single graph) for W on the south and north sides. Label the y-axis appropriately.

(c) **[2 marks]** Based on (a) and (b), briefly comment on differences or similarities in terms of centre and spread of the distributions of widths in the north side and compared to the south sides. Provide a biological justification to your findings.

(d) **[2 marks]** An outlier is an observation that falls outside some pre-determined "fences". An observation smaller than $\hat{q}_{0.25} - k \times IQR$ or larger than $\hat{q}_{0.75} + k \times IQR$, with $\hat{q}_p$ denoting the p-sample quantile and $IQR = \hat{q}_{0.75} - \hat{q}_{0.25}$. Mild ($k = 1.5$) and strong ($k = 3$) outliers fall outside these two "fences". Set $k = 1.5$ and count the number of outliers for W in both directions beyond these fences. Count the number of mild and strong outliers, for leaves with above- or below-average width, then fill in the 4 cells of a table similar to the one shown below:

	No. mild outliers ($1.5 < k < 3$)	No. strong outliers ($k > 3$)
Below-average width		
Above-average width		

(e) **[2 marks]** Briefly comment on the presence (and potential origin) of outliers in the leaf data with respect to W. Based on your findings, which of the following summaries of {centre, spread} seems to be the most appropriate to summarise width W: $\{\hat{q}_{0.5}, IQR\}$ or $\{\bar{x}, s\}$? Why?

Problem 3 [9 marks]

Assuming you have loaded the data and assigned it to an object called DCE2, the following command produces a new data set containing 528 within-student average measurements:

```
> DCE2.agg <- aggregate( . ~ studentID + side, DCE2, mean)
```

This command aggregates all the variables in the raw data DCE2 for all the combinations of the factors studentID and side. As a result, each row in the aggregated dataset DCE2.agg reports the average leaf width and length for each side of the tree for each student.

In part (a) you will work with raw and aggregated data. In parts (b)–(e) you will work with the raw leaf width data DCE2.

(a) **[3 marks]** Compute the sample mean, $\bar{x}$, and sample standard deviation, s, for the width of leaves using the aggregated data (528 rows) and original data (2731 rows). Plot histograms for the raw leaf widths and for mean leaf widths in the aggregated data. Use the arguments `xlim=c(0, 250), breaks=20` for both plots. Briefly compare and contrast the findings from the original and averaged data. Particularly, explain why s in the original data is bigger than the same quantity computed from the averaged data.

(b) **[1 mark]** Let X be the width of a randomly selected leaf from the South side of the tree. Using the data provide an estimate for $p = P(X > 200)$, i.e. the probability that a randomly selected leaf from the south side of the tree exceeds width of 200 millimetres.

(c) **[1 mark]** Using the result in (b) compute the (approximate) probability that at least one out of five leaves randomly picked by a student from the south side exceed 200 millimetres width.

(d) **[2 marks]** Let Y be the number of leaves wider than 200 millimetres out of 5 leaves sampled from the south side of the tree. Give the distribution of Y and specify $E(Y)$ and $Var(Y)$.

(e) **[2 marks]** Your lecturer collected a random sample of 5 leaves from the south side of a tree finding that 3 out of 5 leaves are wider then 200 millimetres. Is there enough evidence suggesting that your lecturer collected a sample from a tree different from the one in Data Collection Exercise 2? Justify your answer using an appropriate probability calculation.