

MATH1041 Statistics for Life and Social Science

Term 2, 2021

MATH1041 Assignment

Data: Together with this document, you should have received your unique dataset (in the text file format) in an e-mail sent to your official university email address. The data (that is, your dataset) are available in a text file with a name similar to “z1234567.txt”, where z1234567 in the text file name is replaced by your unique student zID. If you have not received your dataset (double check your UNSW email inbox as well as the spam folder), please contact Dr Nicole Mealing (n.mealing@unsw.edu.au).

Submission due date: Friday 30th July (Week 9) before 11:59pm (Sydney time, AEST). Note that a late penalty of 10% (i.e. 5 marks) per day will apply.

Your submission must contain your full name and student zID at the top of your assignment. Submit your assignment through Turnitin via Moodle. See the “Assessments” section on Moodle for further information regarding online submission.

Please submit a neatly typed assignment as a Microsoft Word document (.doc or .docx), see the information and help about the assignment in the assessment section on Moodle. If you use Google Docs, LaTeX, or similar then you need to submit your assignment as a PDF document (.pdf).

Verify that your assignment has been submitted correctly by downloading the submission receipt and clicking on the link to your file to check that it displays correctly in the Turnitin viewer. If not, it is your responsibility to make the necessary amendment.

Typesetting (*)	/2
Q1	/15
Q2	/33
Q3	/10
Total	/60

(*) See the next pages and the assessment section on Moodle for details, help and explanations about the assignment and typesetting.

Reading the data into RStudio

The data are in a text file with a name similar to: “z1234567.txt”, where z1234567 in the text file name is replaced by your student zID. This file was sent to you by e-mail (see page 1).

The first step is to read the data into RStudio. The data format is similar to what you have already done in the Introduction labs. Follow the instructions given in section R1.4 “How to import a text file into RStudio” of the RStudio “How-To-Manual” available on Moodle. Once you have uploaded the data then you are ready to start your analysis!

Note that your assignment and dataset is unique. **You cannot show your dataset or your assignment to anyone.** It is your responsibility to keep your dataset and your assignment secret. **Also, your assignment must be your own work. You cannot get any outside help in any form.** If you have a question about the assignment, the only places where you can ask it is on the MATH1041 Assignment forum, provided you do not reveal your data, or at a staff consultation or Drop-in Centre consultation.

Computing assignment format

Keep in mind that this assignment is not only about assessing your Statistical skills, it is also about giving you feedback on your Mathematical writing skills. The assignment must be typeset correctly and provide complete explanations in complete English sentences and paragraphs. Think of this as practice for a document you might produce in your future career that includes mathematical explanations.

Here are some more details that may assist you:

- Regarding the overall assignment structure, please answer all questions in the given order (that is, 1.a., 1.b., ... etc). You don't need to re-write the assignment questions again, only their label (write “3.e.” for instance when you start question 3.e.). Keep your answers brief, clear and concise.
- You are required to type up your entire assignment (in Microsoft Word, Google Docs, LaTeX, or similar), including any equations. The only exception are the plots produced by RStudio, for which you can save the figures (use “export” in the bottom right window in RStudio) or take screenshots which you then paste in your assignment. Nothing can be handwritten then scanned. As a UNSW student, you can download Microsoft Word for free, see: <https://student.unsw.edu.au/office-365>.
- As in any properly typeset document containing mathematic symbols, you must use an **equation editor** for all maths symbols. For instance, you should write “ X is normal”, rather than “X is normal” (Notice how the ‘ X ’ looks different?) and you should write “ $t_{obs} = 1.23$.”, rather than “tobs = 1.23”.

The marking scheme for this criteria is the following: Are mathematical symbols typeset using the equation editor? 2 marks for ‘almost always’, 1 mark for ‘sometimes’, 0 marks for ‘rarely’. Help about Microsoft equation editor can be found in a document called [Microsoft Word Equation editor help for MATH1041](#) located on Moodle in the assessment tab, at the bottom of the assignment section.

- You should add some working out for the questions involving calculations; do not just give the final answer. Note that you will get marks for clear explanations and a correct method even if you get the wrong answer. However, try to keep your solutions brief and concise. Depending on what the question is asking, your working could consist of **RStudio** commands or perhaps the main steps explaining how you arrived at your answer.
- Keeping your results to 3 or 4 significant figures should be fine. If there are multiple steps in a calculation, it is best to round in the final step only.
- There is no requirement for font size and line spacing but obviously do not make things too small.

Scenario

Western Australia (WA) is home to a special species of giant trees, not found anywhere else in the world. These are *Eucalyptus diversicolor*, but more commonly known as Karri trees. Mature Karri trees are thought to range in height from about 45-60m high, but some people believe there are Karri trees in WA that are as tall as 80m. When a bushfire moves through the forest it can hollow out the tree trunk, which can be an impressive sight to behold. In the recent past, Karri trees have been logged for hardwood timber and used as fire lookouts (where people climb the tree to spot any fires that have started). Some trees have iron spikes wrapping up and around their trunks that tourists can climb for a view.

An arborist team wishes to understand Karri trees in more detail. From satellite imagery, they are aware of the location of all Karri trees and can see that there are n forests containing Karri trees. The team drives to each of the n regions and measures data for the first tree they see that is safe for them to climb to the top.

The team collects three measurements. It is straightforward for the team to measure the Diameter at Breast Height (DBH) of their selected Karri trees. This is a standard height set in Australia as the diameter of the tree at 1.4m above the ground. New leaves on Karri trees tend to be egg-shaped and older leaves tend to be lance-shaped. The team randomly selects a leaf from each selected tree and records the type (Youth or Adult) and length of the leaf (in centimeters). It is more difficult to measure the height of these giant trees but this is managed (in metres).

Your job is to assist the arborist team by analysing the dataset provided to you. The data is provided on the right of this page and also as a separate text file attachment sent by email. The text file contains your unique dataset of length n in separate rows consisting of four variables: `Tree.Height`, `Tree.Diameter`, `Leaf.Type`, and `Leaf.Length`.

```
"Leaf.Type" "Leaf.Length"
"Tree.Height" "Tree.Diameter"
"Youth" 13.4 51.72 6.52
"Youth" 10.4 52.82 7.86
"Adult" 11.4 55.43 13.34
"Adult" 9.4 50.58 10.51
"Adult" 10.7 55.78 11.83
"Adult" 9.8 51.11 10.29
"Youth" 15.1 56.18 12.96
"Youth" 9.7 56.14 11.32
"Adult" 9.1 54.78 6.95
"Adult" 10.2 52.78 13.22
"Adult" 9.3 53.24 8.35
"Adult" 10.1 52.22 8.54
"Adult" 10.2 52.24 8.64
"Adult" 10.7 55.44 10.12
"Adult" 10.2 52.64 6.4
"Adult" 9.9 55.73 10.32
"Adult" 9.1 55.03 10.96
"Youth" 9.1 54.88 7.77
"Youth" 10.2 54.46 10.38
"Youth" 7.6 51.03 9.15
"Adult" 10 55.09 11.56
"Adult" 9.8 54.29 9.87
"Adult" 10.6 57.21 12.15
"Adult" 9.4 53.16 8.54
"Adult" 10.4 52.49 7.15
"Youth" 10.2 53.06 7.23
"Adult" 10.5 55.14 10.38
"Youth" 10.5 53.26 7.61
"Adult" 9.8 51.58 7.04
"Adult" 10.3 57.25 11.78
"Youth" 12.7 54.71 7.06
"Adult" 9 50.36 7.54
"Adult" 11.3 53.12 8.7
"Adult" 9.2 54.06 10.83
```

The Analysis Tasks

The questions you need to answer in your assignment submission are given below.

Q1. The arborists start with an exploratory analysis on the Diameter at Breast Height (DBH) of the Karri tree data collected.

- 1.a. Calculate the sample mean and sample standard deviation of the DBH (`Tree.Diameter`) of the trees in your sample.
- 1.b. Produce a histogram for the `Tree.Diameter` measurements. Include this histogram in your submitted assignment properly labelled.
- 1.c. Comment on the shape (skewness/symmetry) of your histogram from Part 1b.
- 1.d. A common technique that can be used to remove skewness in data is known as a log-transformation. That is, for each value in your data (denoted by x_i), you can log-transform it as $y_i = \log(x_i)$. The function in `RStudio` that performs a log-transformation on a value is `log()`. Produce a histogram for the `log(Tree.Diameter)` measurements. Include this new histogram in your submitted assignment properly labelled.
- 1.e. Do you think this log-transformation reduced any skewness identified in Part 1c? Explain briefly.

The arborists now explore the leaf data they collected.

- 1.f. Produce comparative boxplots for `Leaf.Length` against `Leaf.Type`. Include this plot in your submitted assignment, properly labelled.
- 1.g. Describe any differences or similarities in the distribution of the length of leaves among newer ('Youth') and older ('Adult') leaves using your comparative boxplots from Part 1f. Include in your answer comments on shape, location, spread and outliers. Your answer should be written in plain English so that others not looking at your boxplots can understand the distribution of these leaf lengths.

Q2. The Australian GeoZight Magazine published an article which claims that the average height of Karri trees is 55m. However, the arborists believe that the true average height is actually lower than 55m and ask you to investigate their claim.

Let μ be the true mean height of Karri trees, which are endemic to the southern region of Western Australia.

- 2.a. Produce a normal quantile plot of your sample of `Tree.Height` values (see Section R2.6 "How to produce a normal quantile plot using `RStudio`"). Include this plot in your submitted assignment, properly labelled.
- 2.b. By referring to the normal quantile plot obtained in Part 2a, briefly discuss if the Karri tree heights are approximately normally distributed.

- 2.c.** Carry out an appropriate hypothesis test by completing the following steps.
- State the null (H_0) and alternative hypotheses (H_a) relevant to the research objective. Write the hypotheses out in plain English and mathematically.
 - State the expression for a suitable test statistic.
 - Calculate the observed value of the test statistic using your sample.
 - Write down the sampling distribution for the test statistic when the null hypothesis, H_0 , is true.
 - Calculate the P -value for this test, showing some working.
 - Explain what the P -value you obtained in Part 2(c)v calculates, by describing this probability in words and referring to what you observe in your sample.
 - Write a conclusion to your test in plain English (that is, with no technical mathematical or statistical language, and referring to the scenario given).
- 2.d.** Some assumptions need to be satisfied for the sampling distribution of the test statistic (as given in Part 2c) to be valid. State these assumptions, and briefly discuss whether these assumptions are satisfied.
- 2.e.** Produce a 95% confidence interval for μ , the true mean Karri tree height. For this question you may assume that it is appropriate to use a t -distribution. Write down all the required steps to calculate this interval.
- 2.f.** Interpret your confidence interval (constructed in Part 2e) in plain English (that is, with no technical mathematical or statistical language, and referring to the scenario given).
- 2.g.** Explain whether your confidence interval (constructed in Part 2e) is consistent with the results of your hypothesis test in Part 2c.
- 2.h.** Like all realised confidence intervals, your confidence interval (constructed in Part 2e) is of the form: *estimate* \pm *margin of error*. Does the margin of error you calculated include errors due to measurement error (e.g. if the arborist measuring the tree height sometimes wrote down a slightly inaccurate value)? Briefly justify your answer.

Q3. The arborists acknowledge that it is quite difficult to take height measurements of such tall trees. They ask you to investigate if they can use the diameter at breast height (DBH) of the Karri trees to predict the height of these trees. You will use the log of the DBH measurements in your analyses from Part 1c (i.e. `log(Tree.Diameter)`).

- 3.a.** Construct an appropriate graphical summary to visualize the relationship between `Tree.Height` and `log(Tree.Diameter)`. Include this plot in your assignment, properly labelled.
- 3.b.** Summarise the key features of your plot from Part 3a.
- 3.c.** Suggest an appropriate numerical summary to quantify the strength of the linear relationship between `Tree.Height` and `log(Tree.Diameter)`. Report and briefly comment on this value.

- 3.d.** The Giant Tingle Tree is a Karri tree near Walpole in Western Australia, which is famous due to its huge fire-hollowed trunk. This tree is 30m tall with a diameter of 22.3m. (There are rumours that cars may have parked within its trunk in the 1980s!)

The arborists want to predict the height of the Giant Tingle Tree from its diameter measured at breast height using a least squares regression line calculated from your dataset. Would you recommend they do this? Explain your answer briefly.

END OF ASSIGNMENT

Did you know? Karri trees are endemic to Western Australia and the Giant Tingle Tree is real.