STAT101-25S2

Assignment

Due 11.59pm Friday 19 September 2025

You may work on your own for this assignment or in a group with no more than three members. All group members must list their full names and ID numbers in the space below. If you only have two members or you are doing the assignment by yourself, leave the unnecessary space as is:

This Word.doc includes the questions for the assignment. Under each question is the designated editable space for your answer. The format of the editable spaces is fixed and cannot be changed. If you cannot fit your answer into the space provided, you are writing more than is required/necessary and should edit your answer to make it more concise. For the questions that require a screenshot of your plot, click on the editable space to upload an image.

This Word file needs to be edited in an installed Microsoft Word app (not the web app). As a UC student you have access to Office 365 which includes Microsoft Word.

If you need support installing Microsoft Word on your personal device, see the Course Information Section on Learn for instructions on accessing Microsoft 365. For further support, please see Te Pātaka | Student Hub - IT Digital Services on Level 2, Puaka-James Hight.

Microsoft Word is accessible on UC computers including the Jack Erskine 035 Lab where our Tutorials are. If you do not want to install Microsoft Word on your personal device, you can copy and paste your answers into this Word file using Microsoft Word on those Lab computers.

Once you have finished answering the questions and are ready to submit the assignment, save the Word file as a PDF. The Assignment Drop Box - Individual/Group Submission will only accept PDF uploads.

Any individual member of a group can upload/edit/remove the group's submission and only one file per group is required. If another member uploads a submission, it will override any previous submission from anyone in the group. All group members will receive the same mark.

Group Member 1:

Dax Babaria 65232935

Group Member 2:

Group Member 3:

DECLARATION:
In submitting an assignment to the drop-box on Learn, you confirm that you have read the Academic Integrity section in the General Information for Students (see link below) and verify that this assignment submitted for assessment is entirely the work of the student/s submitting the assignment.

https://www.math.canterbury.ac.nz/static/policy/student_info.pdf

Question 1 (Refer to Sections 1 & 2)

This question uses the data file 'Student Data.csv' linked in the Assignment drop-box. This dataset contains a sample from a previous occurrence of STAT101.

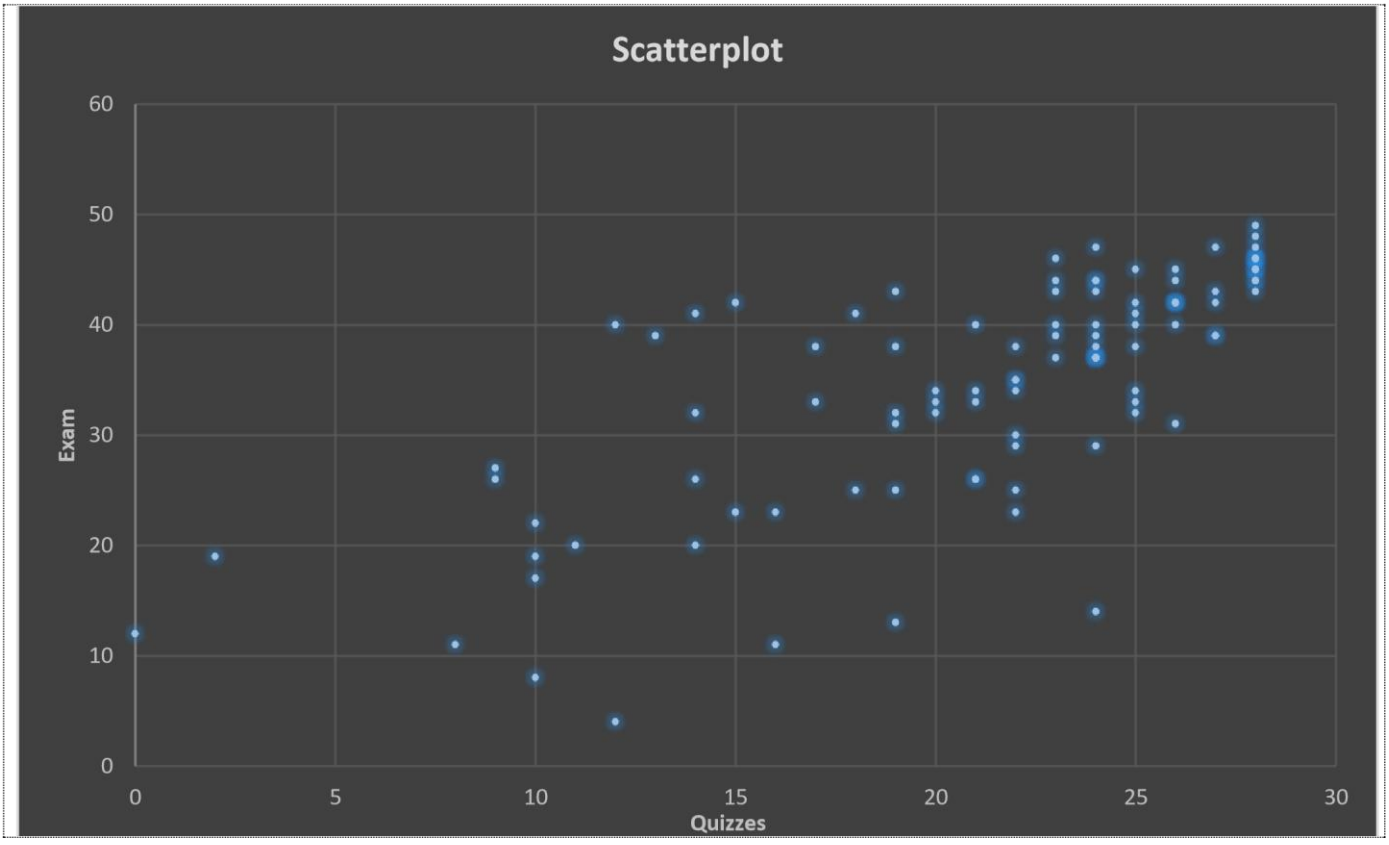
The variables included in the data are:
'Programme' – Degree the student was enrolled in.
'Age' – The age of the student in years
'First Person in Family – Whether the person is the first in their family to attend University.
'Attendance' – The mark the student received for attendance on the course (max 6).
'Quizzes' – The mark the student received for the quizzes on the course (max 28).
'Exam'– The exam mark the student received on the course (max 50).
'Grade' – The final letter grade the student received for the course.

Your task in the question parts below is to create a linear regression model using the variable 'Quizzes' to predict variable 'Exam'. State all values to 3 decimal places when answering the following questions.

a) Explain in one or two sentences which variable is being treated as the explanatory, and which variable is being treated as the response for the linear regression model described above.

The quizzes variable is the explanatory variable as its been used to predict the response variable which is exam.

b) Using StatKey, Excel, or another technology, create a scatterplot for the two variables 'Quizzes' and 'Exam'. Include the scatterplot in the space below with appropriate labelling. Do not include any extraneous plots/statistics etc.



Question 1 continued

c) Report the sample correlation between the variables Quizzes and Exam. Explain in one or two sentences what this correlation statistic indicates about the strength and direction of the association between Quizzes and Exam marks.

The correlation between the variables quizzes and exams is positive as one variable increases other increases as well.
The variables shows strong correlation as the correlation is 0.726

d) State the equation for the linear regression model in context.

exam^=9.094+1.215×quizzes

e) In one or two sentences interpret the slope of the linear regression model in context.

The slope is 1.215 and it suggests that for every 1 mark increased in quizzes, the predicted exam score increases by 1.215 marks on average.

f) In one or two sentences interpret the y-intercept of the linear regression model in context.

The intercept is 9.094 and it predicts the exam score when quizzes is 0. In this context, if a student scores 0 on quizzes, the model predicts they would score about 9.094 on exam.

g) Explain in one or two sentences whether the interpretation for the y-intercept is meaningful in this context?

The interpretation is not really meaningful as its very unlikely that a student will score 0 on quizzes if he/she is attempting all of them.
Statistically it shows relationship but in real world context its not very meaningful.

h) Use your model to predict the Exam mark for a student with a Quizzes mark of 24. Show your working.

exam^=9.094+1.215×quizzes
Exam marks when a student scores 24 marks on Quizzes
= 9.094 + 1.215 x 24
= 38.254

i) The first student (Case ID #1) had a Quizzes mark of 24 and an Exam mark of 44. Calculate the residual for this case using your answer from part h). Show your working.

Residual = Observed – Predicted
Residual = 44 – 38.254
Residual = 5.746

Question 1 continued

j) Explain in one or two sentences whether the data collected in the sample is likely to have been observational or experimental.

The data is likely to have been observational as the teacher just collected the quiz and exam scores and not intervention was done during examination, making it observational.

k) Explain in one or two sentences whether it is appropriate to imply a causal relationship between 'Quizzes' and 'Exam'.

No, it doesn't imply causal relationship , as there are some other confounding variables such as study habits, prior knowledge etc, can influence both the quiz and exam scores

Question 2 (Refer to Section 3)

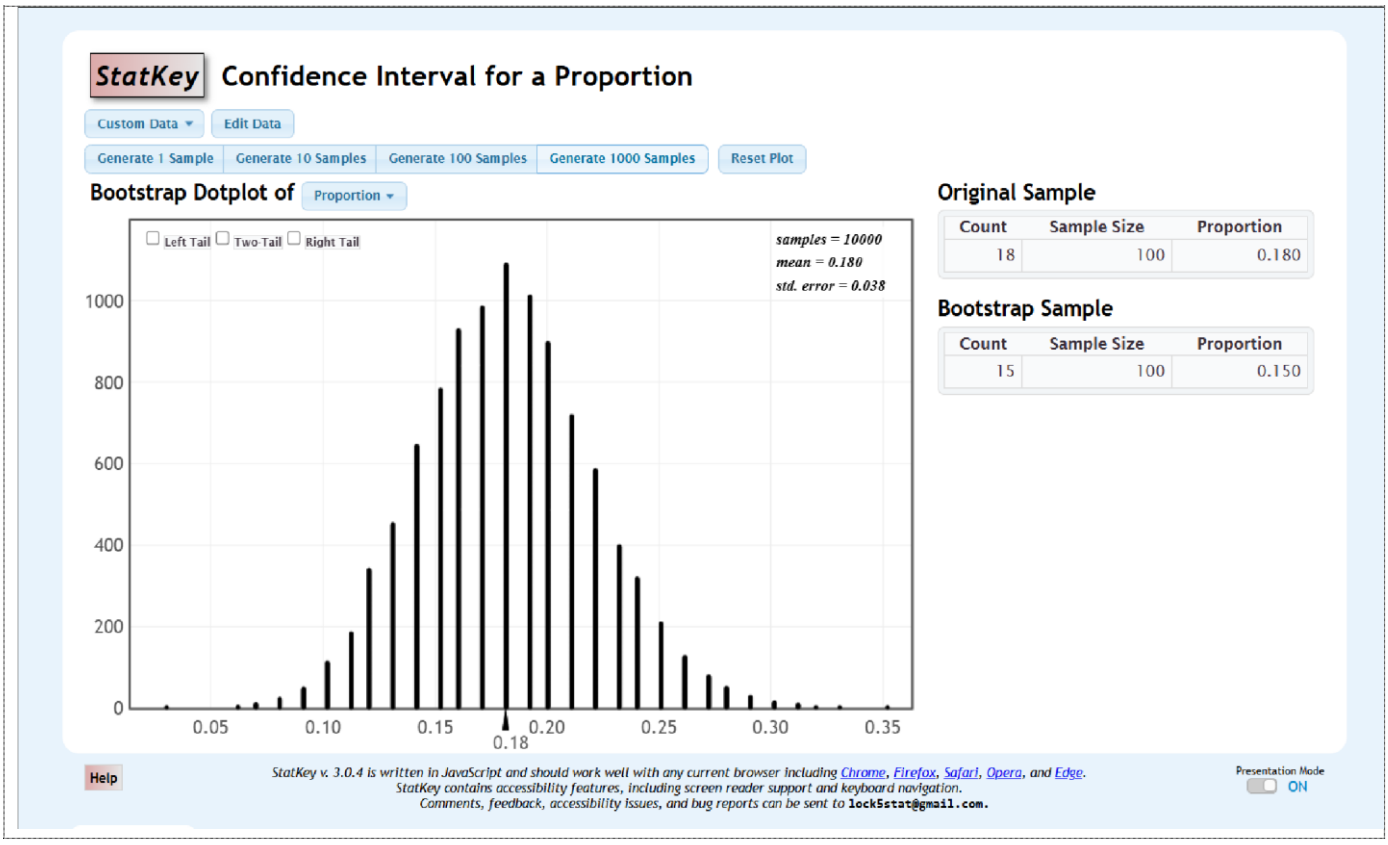
This question uses the same data file as used in Question 1, 'Student Data.csv' linked in the Assignment drop-box. This dataset contains a sample from a previous occurrence of STAT101.

Your task in the question parts below is to create a bootstrap distribution and use it to estimate the proportion of students that get an A+ grade. State all values to 3 decimal places and use appropriate notation when answering the following questions

a) Calculate the sample proportion (3dp) of students that received an A+ grade. You may want to explore the 'Format as Table' or 'Sort' tools in Excel to help with this.

18/100 = 0.180

b) Use StatKey (or another technology) to generate a bootstrap distribution of 10,000 sample proportions of students that received an A+ grade. Include the bootstrap distribution in the space below with appropriate labelling. Do not include any extraneous plots/statistics etc.

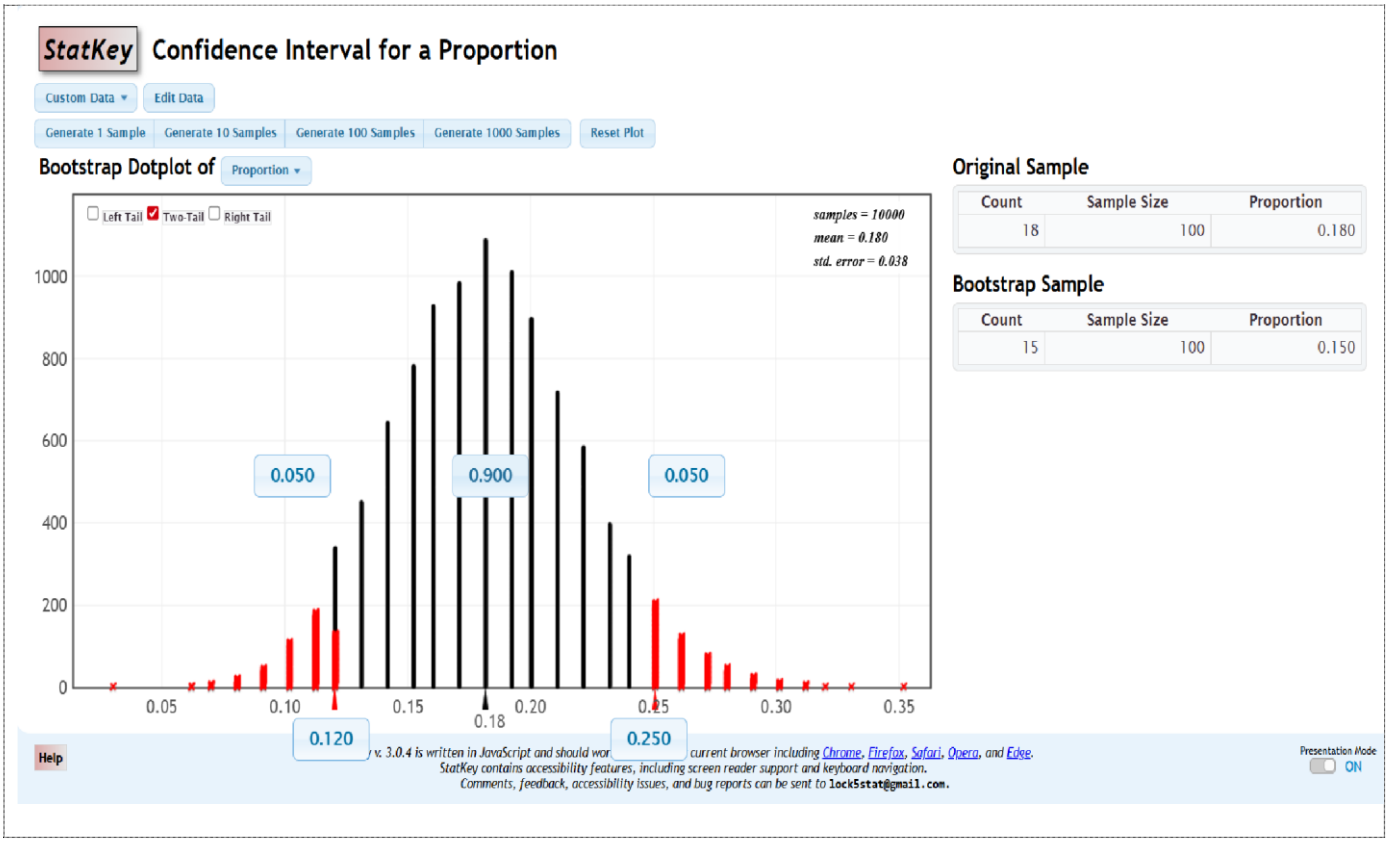


c) Describe in one or two sentences the shape and centre of your bootstrap distribution.

The bootstrap distribution is symmetric and bell shaped, centered around the mean value of 0.180.

Question 2 continued

d) Using percentiles (not the standard error) from your bootstrap distribution, find a 90% confidence interval for the proportion of students that receive an A+ grade. Include the bootstrap distribution showing the 90% confidence interval in the space below. Do not include any extraneous plots/statistics etc.



e) In one or two sentences interpret the 90% confidence interval in context.

The 90% confidence interval for the population proportion is (0.120, 0.250). This means we are 90% confident that the true proportion of the population lies between 12% and 24%.

f) Explain in one or two sentences whether it is plausible that the proportion of all STAT101 students that get an A+ grade is 0.20 (20%).

Yes, it is plausible that the proportion of all STAT 101 students who get an grade A+ is 0.20, because 0.20 lies within the 90% confidence interval range.

Question 3 (Refer to Section 4)

This question uses the same data file as used in Question 1 & 2, 'Student Data.csv' linked in the Assignment dropbox. This dataset contains a sample from a previous occurrence of STAT101.

Your task is to determine if there is sufficient evidence to answer the question:

"Do students who are not the first in their family to attend university have a different average mark for quizzes than those who are the first in their family to attend university?".

State all values to 3 decimal places and use appropriate notation when answering the following questions. For the question parts below, use the methods described in Section 4 (not Section 6.3).

a) Calculate the sample means for quiz marks of students that are first in their family to attend university and for students that are not the first in their family to attend university.

Sample mean for the students that are first in their family to attend university = 16.313 Sample mean for the students that are not in their family to attend university = 22.083

b) Calculate the difference in the sample means. This is the observed statistic for the hypothesis test.

Not First - First = 22.083 – 16.313 = 5.77

c) State the null and alternative hypotheses for the question "Do students who are not the first in their family to attend university have a different average mark for quizzes than those who are the first in their family to attend university?".

Null Hypothisis = Marks of not first = Marks of first
Alternative Hypothesis = Marks of not first != Marks of first

d) Explain in one or two sentences whether this is a one-tail or a two-tail hypothesis test.

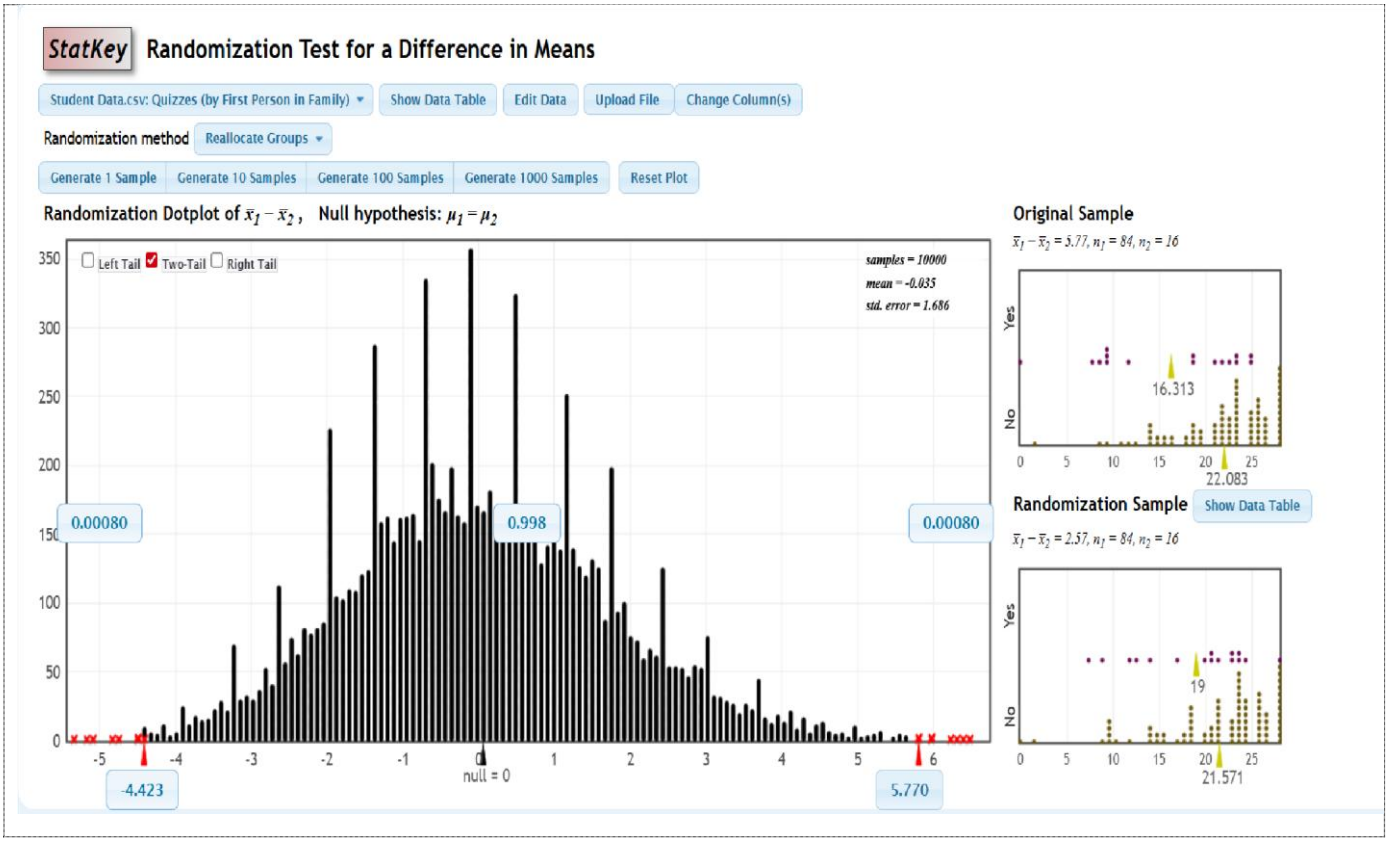
It's a two tailed test as we have to find if there's a difference which means difference can be positive as well as negative so we have consider both the tails.

e) Explain in one or two sentences what the main assumption made when creating a randomisation distribution is?

The main assumption when creating a randomization distribution is that we have to consider that the null hypothesis is true.

Question 3 continued

f) Use StatKey (or another technology) to generate a randomisation distribution for your statistic using 10,000 randomisation samples. Include the observed statistic and p-value in the diagram. Include the randomisation distribution plot in the space below with appropriate labelling. Do not include any extraneous plots/statistics etc.



g) Describe (one or two sentences) the shape and centre of your randomisation distribution.

The distribution is bell shaped and symmetrical and is centered around the null.

h) State the p-value displayed in part f) and use a significance level of 5% (α=0.05) to make a decision and a conclusion for the hypothesis test in context.

As this is a two tailed test we have to sum both the tail's P-value ie, 0.00080 + 0.00080 = 0.0016

Significance level > P – value, that means we reject the null as we have enough evidence for the alternative hypothesis. Therefore, there is a difference between the average marks of students who are not first in their family to attend university in compare of students who are first to attend university.