STATS 1000 / STATS 1004 / STATS 1504
Statistical Practice 1
Assignment 4
2020

DEADLINE:

- Wednesday 6^{th} May 2020 (Week 8) 5:00pm

CHECKLIST

☐: Have you shown all of your working, including probability notation where necessary?

☐: Have you given all numbers to **3 decimal** places.

☐: Have you included all R output and plots to support your answers where necessary.

☐: Have you made sure that all plots and tables each have a caption.

☐: If before the deadline, have you submitted your assignment via the online submission on MyUni?

☐: Is your submission a single word document or pdf file - correctly orientated, easy to read? If not, penalties apply.

☐: Penalties for more than one document - 10% of final mark for each extra document. Note that you may resubmit and your final version is marked, but the final document should be a single file.

☐: If after the deadline, but within 24 hours, have you contacted us via the enquiry page on MyUni and then submitted your assignment online via the online submission on MyUni?

☐: Penalties for late submission - within 24 hours 40% of final mark. After 24 hours, assignment is not marked and you get zero.

☐: Assignments emailed instead of submitted by the online submission on MyUni will not be marked and will recieve zero.

☐: Have you checked that the assignment submitted if the correct one, as we cannot accept other submissions after the due date.

☐: Do not write directly on the question sheet.

1. **One-sample t-test in R**

   This question must be typed in Word

   Fifty songs were randomly selected from Spotify and their duration measured. The data is given in songs.xlsx on MyUni.

   This data has one variable:

   - seconds: the duration calculated in total number of seconds.

   A music critic reckons that all good songs are 4 minutes long. In this question, we will test if the data from Spotify is consistent with this statement.

   (a) First examine the distribution of the total seconds by doing the following:

      i. Produce a histogram of the variable `seconds` in R and include in your assignment.

      [1 mark]

      ii. Describe the distribution of total seconds.

      [4 marks]

   (b) Perform a one-sample t-test of the null and alternative hypotheses

   $$H_0 : \mu = 240,$$
   $$H_a : \mu \neq 240,$$

   where μ is the true population mean duration of songs in seconds on Spotify.

   To do this, complete the following steps:

      i. Perform a one-sample t-test in R and include the output in your assignment.

      [1 mark]

      ii. State the value of the test statistic.

      [1 mark]

      iii. State the P-value.

      [1 mark]

iv. State the distribution of the test statistic if the null hypothesis is true.

[2 marks]

v. State whether you reject or retain the null hypothesis at the 5% significance level? Justify your decision.

[2 marks]

(c) Using your R output, calculate a 95% confidence interval for the mean song duration in seconds. Interpret this interval in context.

[3 marks]

(d) Check the assumption of normality of the sample mean with the following steps:

i. Produce a normal QQ-plot of the total seconds and include in your assignment.

[1 mark]

ii. Using the normal QQ-plot, decide if the assumption of normality is reasonable for total seconds. If not, which theorem, that states that the sample mean is approximately normally distributed for large sample sizes, could be used to justify the use of the t-test?

[2 marks]

(e) Presentation marks (1 for word, 1 for informative table captions, 1 for informative figure captions.)

[3 marks]

[Total: 21]

2. **Two-sample T-test in R**

For full marks, this question must be typed in Word, all required R output should be included and captioned.

The Bechdel test[1] is used to classify movies according to the following criteria:

- The movie has to have at least two women in it,
- who talk to each other,

[1] https://en.wikipedia.org/wiki/Bechdel_test

- about something besides a man.

Does passing the Bechdel test have an effect on a movie's profits?

The dataset `bechdel.xlsx` was obtained from http://fivethirtyeight.com/. The data has been filtered such that only movies that were released from 2010 onwards are included. The column `bechdel` indicates whether the movie passes or fails the Bechdel test, and the column `profit` is calculated as

$$\frac{gross - budget}{budget},$$

where all amounts are converted to 2013 US dollars to account for inflation. The units of profit are proportion of budget.

(a) Import the dataset into R and produce a histograms of `profit` for each level of `bechdel` (include your captioned plot in your final submission). What is the shape of the distribution of profit for each type of movie? Does profit look to be normally distributed?

[3 marks]

(b) Perform a two-sample t-test in R (include your output in your submission). Note you should define Group 1 as PASS and Group 2 as FAIL.

i. Write down appropriate null and alternative hypotheses for the two-sample t-test. Remember to define any parameters used.

[3 marks]

ii. What is the observed value of the test-statistic?

[1 mark]

iii. What is the distribution of the test statistic if the null hypothesis is true? *Note: use the R output for the degrees of freedom.*

[2 marks]

iv. What is the P-value?

[1 mark]

v. Do you reject or retain the null hypothesis at the 5% significance level? Why?

[2 marks]

(c) Use R to calculate the 95% confidence interval for the difference in the population mean profit for movies that pass the Bechdel test to those that fail the Bechdel test. Interpret this confidence interval in context.

[3 marks]

(d) Produce in R, and include in your submission, appropriate plots to test the assumption that the observations in each group are from a normal distribution. Is this assumption reasonable for this dataset? If not, why is the two-sample t-test still reasonable in this case?

[4 marks]

(e) Presentation marks (1 for Word, 1 for informative figure captions, 1 for informative table captions).

[3 marks]

[Total: 22]

[[Assignment total: 43]]