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

DEADLINE:

- Wednesday 6th 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]

   The figure is given in Figure 1.

   ii. Describe the distribution of total seconds.

   [4 marks]

   - Shape: The distribution appears to be right-skewed.
   - Location: The mean is 354.1 seconds, and the median is 267.5 seconds.
   - Spread: The standard deviation is 307.948 seconds and the IQR is 199 seconds.
   - Outliers: One at just less than 1500 seconds, and one at just less than 2000 seconds.

   (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:

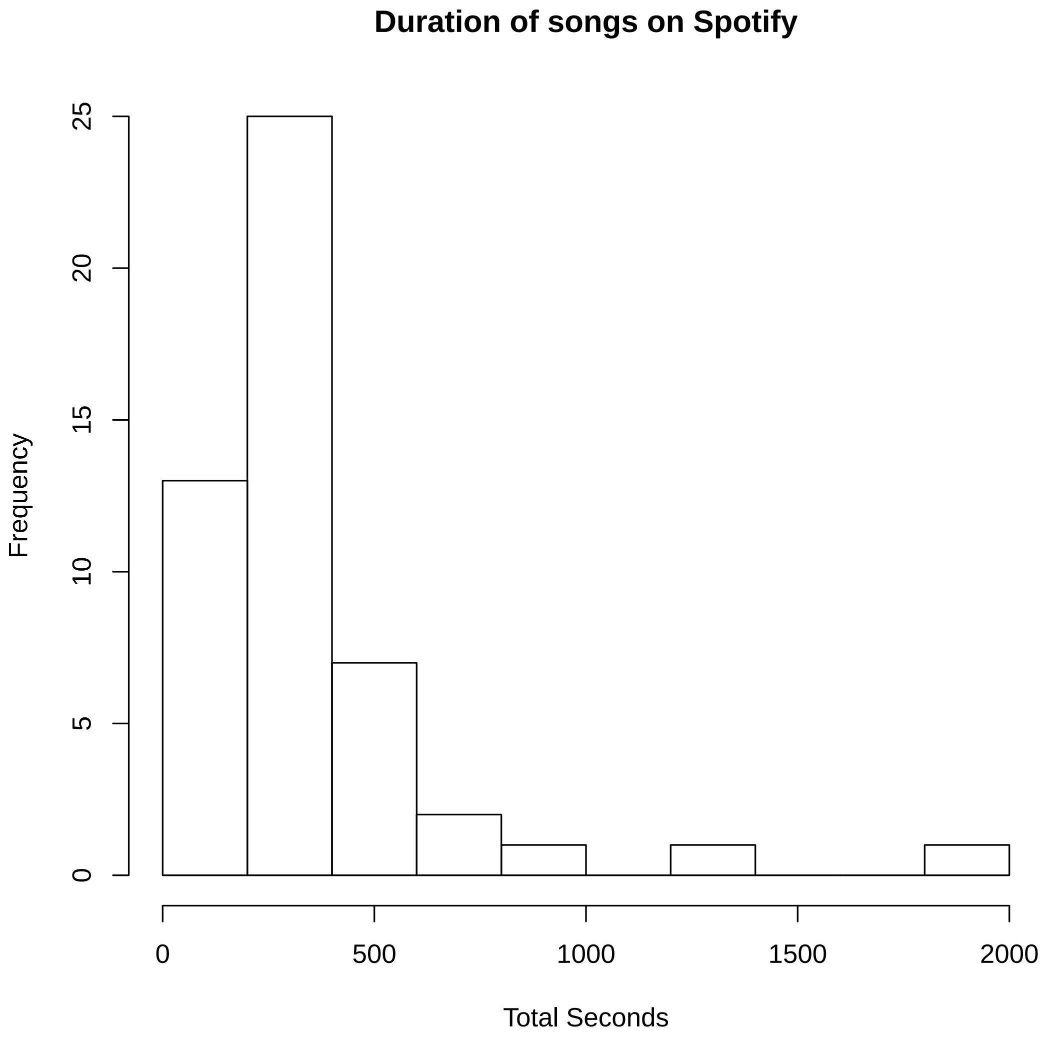

Figure 1: Histogram of the total seconds for the Spotify dataset. The distribution is clearly right skewed.

```
##
##   One Sample t-test
##
## data:  spotify$total_secs
## t = 2.62, df = 49, p-value = 0.01168
## alternative hypothesis: true mean is not equal to 240
## 95 percent confidence interval:
##  266.5823 441.6177
## sample estimates:
## mean of x
##     354.1
```

Table 1: R output for the one-sample t-test of the total seconds in the Spotify dataset showing a significant difference between the mean duration of songs on spotify and the hypothesised duration of 240 seconds.

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

[1 mark]

The R output for the one-sample t-test is given in Table 1.

ii. State the value of the test statistic.

[1 mark]

The test-statistic is $t = 2.62$.

iii. State the P-value.

[1 mark]

The P-value is 0.012.

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

[2 marks]

The distribution of the test statistic if the null hypothesis is true is a t-distribution with 49 degrees of freedom:

$$t_{49}$$

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

[2 marks]

We reject the null hypothesis at the 5% significance level as the P-value is less than 0.05.

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

[3 marks]

The 95% confidence interval, from Table 1, for the mean song duration on Spotify is

$$(266.582, 441.618).$$

We are 95% confident that the mean song duration of songs on Spotify lies between 266.582 and 441.618 seconds.

(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]

The normal QQ-plot is given in Figure 2.

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]

If the assumption of normality is reasonable, then we expect the points in the normal QQ-plot (Figure 2) to roughly lie along the line.

There is obvious curvature in the points and so we conclude that the assumption of normality is unreasonable for this dataset.

As the sample size is large (50 observations), we can use the central limit theorem, which states that the sample mean will be approxi-

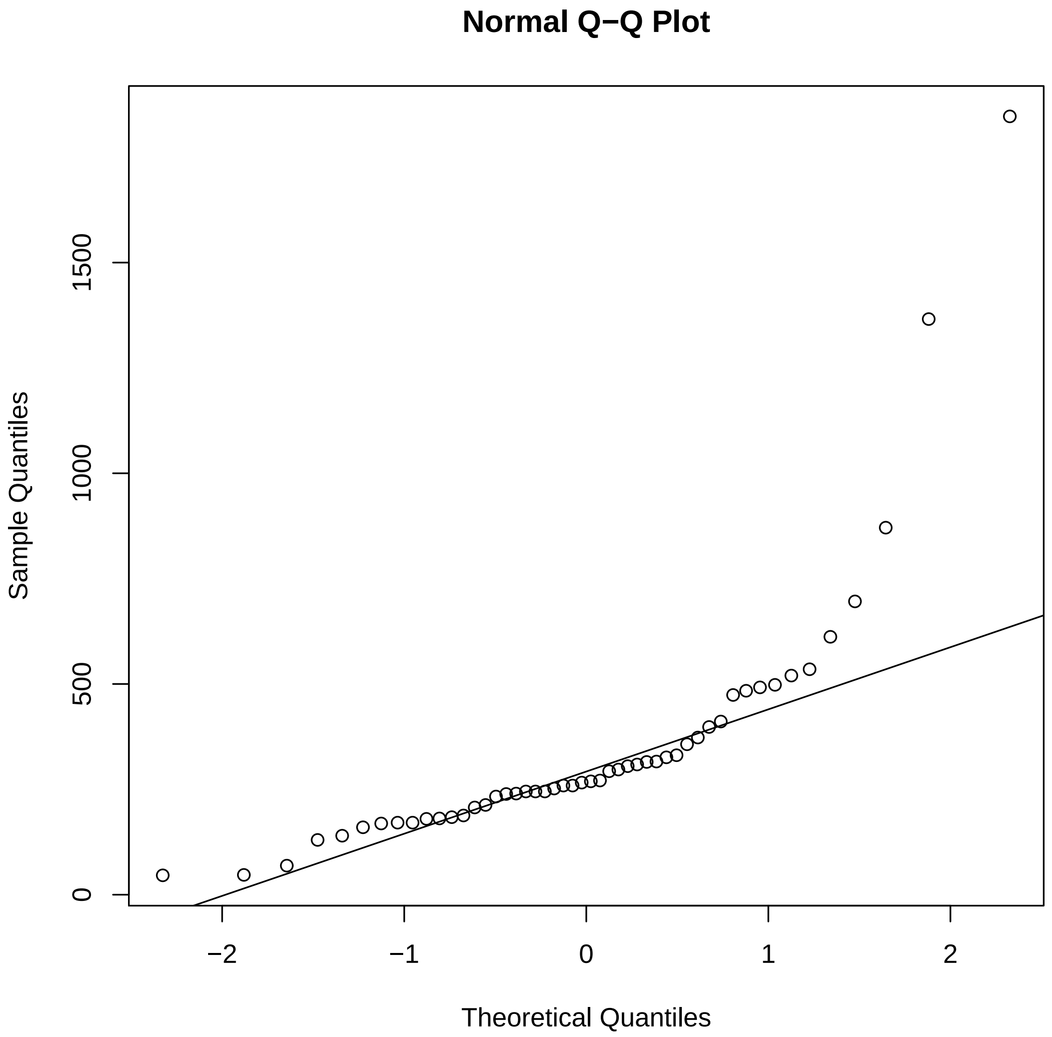

Figure 2: Normal QQ-plot of total seconds for the Spotify dataset. This shows the data is clearly not normally distributed.

mately normally distributed and hence using the one-sample t-test is reasonable.

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

[3 marks]

1 for word, 1 for table captions, 1 for figure captions.

[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,
- 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]

The histograms are given in Figure 3. Both distributions are right skewed. The variable profit does notappear to be normally distributed.

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

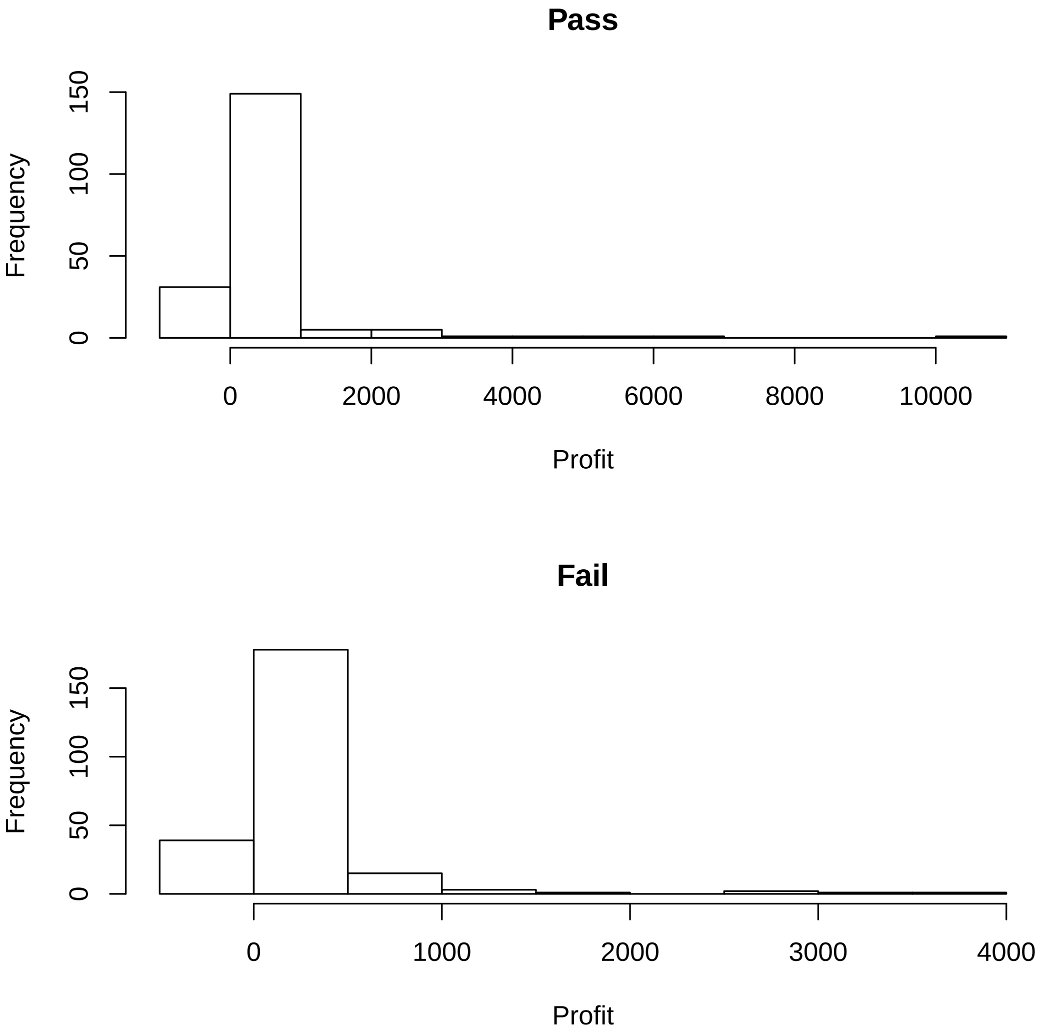

Figure 3: Panel histograms of profit for movies that pass and fail the Bechdel test showing that each group is right skewed.

```
##
##  Welch Two Sample t-test
##
## data:  profit by bechdal
## t = 2.6915, df = 248.22, p-value = 0.007596
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   59.82417 386.23760
## sample estimates:
## mean in group PASS mean in group FAIL
##           464.3615           241.3306
```

Table 2: R output for two-sample T-test for Bechdel dataset showing a significant difference in the amount of profit made between movies that passed and those that did not pass the Bechdel test.

(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]

The output is given in Table 2. The appropriate null and alternative hypotheses are

$$H_0 : \mu_1 - \mu_2 = 0,$$
$$H_a : \mu_1 - \mu_2 \neq 0,$$

where μ_1 is the population mean profit for movies that pass the Bechdel test, and μ_2 is the population mean profit for movies that fail the Bechdel test.

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

[1 mark]

The observed value of the test statistic is 2.692.

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]

$$T \sim t_{248.215}$$

iv. What is the P-value?

[1 mark]

The P-value is 0.008.

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

[2 marks]

We rejectthe null hypothesis at the 5% significance level as the P-value is < 0.05.

(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]

The 95% confidence interval is

$$(59.82417, 386.2376).$$

We are 95% confident that the population mean profit for movies that pass the Bechdel test is 59.824 to 386.238 (proportion of the budget) higher than the population mean profit for movies that fail the Bechdel test.

(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]

What: Normality.

Where: Look at normal QQ-plots of the observations in each group (Figure 4).

Expect: If the observations are normally distributed, we expect the points to be roughly linear.

See: There is definite curvature in both plots.

Conclude: So the assumption of normality is unreasonable.

Each group has a large number of observations so by the central limit theorem, the sample means are approximately normally distributed and so we can use the two-sample t-test.

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

[3 marks]

1 for Word, 1 for informative figure captions, 1 for informative table captions.

[Total: 22]

[[Assignment total: 43]]

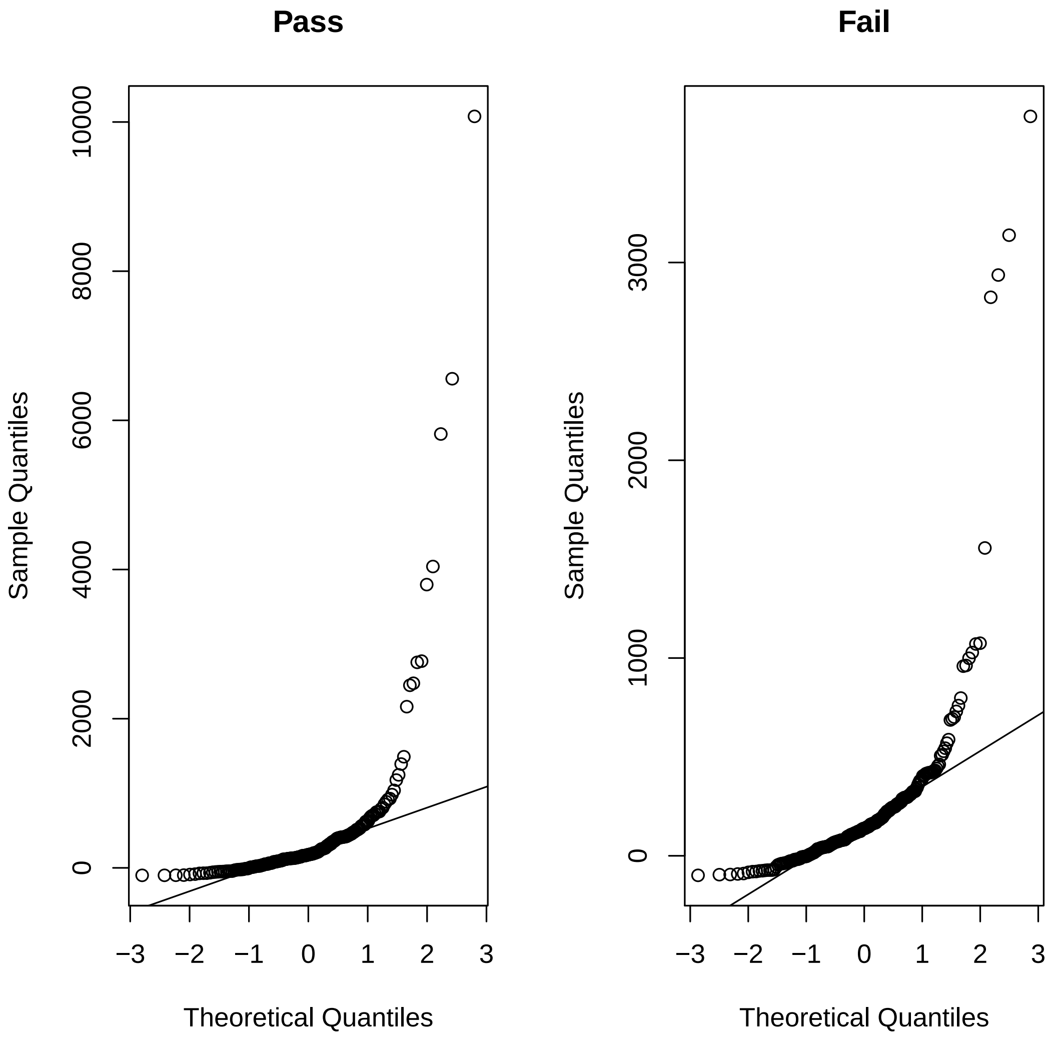

Figure 4: Normal QQ-plots for the observation in each group for the Bechdel dataset displaying clearly non-normal data.