STATS 2107
Statistical Modelling and Inference II
Assignment 2

Matt Ryan

Semester 2 2021

ASSIGNMENT CHECKLIST

- Have you shown all of your working, including probability notation where necessary?
- Have you included all R output and plots to support your answers where necessary?
- Have you included all of your R code?
- Have you made sure that all plots and tables each have a caption?
- Is the submission a single pdf file - correctly orientated, easy to read? If not, penalties apply.
- Assignments emailed instead of submitted by the online submission on Canvas will not be marked and will receive zero.
- Have you checked that the assignment submitted is the correct one, as we cannot accept other submissions after the due date?
- Assignments submitted no later than 24 hours after the deadline will still receive 100% credit. This is in acknowledgement of that fact that people can get sick or have other commitments occur or have internet difficulties. However, assignments can not be submitted more than 24 hours after the deadline.
- Other variations to the assessment on medical/compassionate grounds will only be considered if suitable documentation is provided. In this case you should email Matt Ryan as soon as possible. Please do not ask your tutor for an extension or try to submit an assignment directly to them. Tutors are not able to grant extensions or accept any assignment submissions.

Q1

This questions may be typed or hand written and scanned in as a pdf.

The aim of this question is is to build your experience with power and sample size calculations. You are introduced to calculating the power for a two-sided normal Z test, and applying the sample size calculation to this test.

A study was conducted to compare finger dexterity of music students who study piano to those who study singing. A random sample of 137 students from the piano group had mean dexterity score of 37.25 and standard deviation 4.34. An independent sample of 137 singing students had mean dexterity score of 35.91 and standard deviation of 5.19. You may assume the data is normally distributed. Let μ_1 and σ_1^2 denote the true mean and variance dexterity score of piano students and μ_2 and σ_2^2 denote the true mean and variance dexterity score of singing students.

(a) Test to see whether there is sufficient evidence in the sample to indicate that piano students have a different mean dexterity score than singing students. You may use a significance level of 0.05 and make the assumption $\sigma_1 \approx 4.34$ and $\sigma_2 \approx 5.19$.

[4 marks]

(b) Calculate the power of the test used in part (a) when μ_1 - $\mu_2 = 3$.

[4 marks]

(c) Find the sample size needed for the test to achieve a significance level of 0.05 and power of 0.95, when μ_1 - $\mu_2 = 3$. You may assume equal sample size for both group and that $\sigma_1^2 = \sigma_2^2 = \sigma^2$. **Hint.** Estimate σ^2 with

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} .$$

[2 marks]

[Question total: 10]

Q2

This questions may be typed or hand written and scanned in as a pdf.

The aim of this question is to give you experience at performing hypothesis tests and confidence intervals for the population variance. You will be given an opportunity to consolidate your learning of this material from Week 4.

Consider the study from the previous question to compare finger dexterity of music students who study piano to those who study singing. A random sample of 137 students from the piano group had mean dexterity score of 37.25 and standard deviation 4.34. An independent sample of 137 singing students had mean dexterity score of 35.91 and standard deviation of 5.19. You may assume the data is normally distributed. Let σ_1^2 and σ_2^2 denote the true variance of dexterity score of piano students and singing students, respectively.

(a) State the sampling distribution of $\dfrac{(n_1 - 1)S_1^2}{\sigma_1^2}$, where n_1 is the number of students in the piano group and S_1^2 denotes the sample variance of the piano group.

[1 mark]

(b) Calculate a symmetric 95% confidence interval for σ_1^2.

[4 marks]

(c) The researcher has questioned whether they can assume equal variance between the two groups. Perform the hypothesis test

$$H_0 : \sigma_1^2 - \sigma_2^2 = 0 \quad \text{vs} \quad H_a : \sigma_1^2 - \sigma_2^2 \neq 0 ,$$

at the $\alpha = 0.05$ level of significance. Your hypothesis test should include:

- Your test statistic and the distribution of the test statistic under the null hypothesis.
- The observed value of your test statistic.
- The critical region for this hypothesis test.
- Your conclusion in context.

[7 marks]

[Question total: 12]

Q3

THIS QUESTION IF FOR POSTGRADUATE STUDENTS ONLY. This questions may be typed or hand written and scanned in as a pdf.

The aim of this question is is to derive a 1-sample hypothesis test for the population variance. You will be guided through constructing a hypothesis test for σ^2 based on normal data.

Let $Y_1, Y_2, \ldots, Y_n$ be i.i.d. normally distributed data with true mean μ and true variance σ^2. Let S^2 denote the sample variance of the data. Consider the null hypothesis

$$H_0 : \sigma^2 = \sigma_0^2 .$$

(a) Propose an appropriate 2-sided alternative hypothesis.

[1 mark]

(b) Propose an appropriate test statistic under the null hypothesis, including a statement of the appropriate reference distribution.

[2 marks]

(c) Working at the α level of significance, what is the critical region for this hypothesis test?

[2 marks]

(d) Suggest an appropriate p-value based on your test.

[1 mark]

(e) A researcher is interested in the variability of canine toe nail length. They have measured the toe nail length of 164 randomly selected dogs and found a sample standard deviation of 1.5cm. Assuming the data is normally distributed, test the null hypothesis that the true variance of toe nail length in dogs is 2 at the $\alpha = 0.05$ level.

[6 marks]

[Question total: 12]

Q4

Please submit your answer to this question online using MyUni. You will be asked to answer quiz questions and to upload an R script file.

The aim of this question is to help develop your R skills. You will be asked to construct a function to calculate the sample size for a given significance level, proportion, and margin of error. You will then use this function to explore the sample size calculations.

A researcher is planning a pilot study to estimate the proportion of people who are regularly taking over-the-counter paracetamol in South Australia.

(a) Write a function to calculate the sample size required to estimate a population proportion (`p`), with a desired significance level (`alpha`) and margin of error (`delta`) for a two-sided hypothesis test. Your function should follow the format below, and prints the required sample size (as numeric):

```r
proportion_ss <- function(alpha, p, delta){
  ## your code here
}
```

(b) Suppose 20% of the population is regularly taking paracetamol, use your function to calculate the required sample size for 5% significance level with ±10% margin of error. **Hint**: remember to always round up to the nearest integer.

(c) With COVID-19, the researcher is concerned that the proportion of South Australians regularly taking paracetamol would increase. They wanted to know how this would affect the required sample size.

  i. Use your function to produce a plot of the required sample size (Y-axis) against proportions (X-axis) ranging from 10% to 90%, assuming 5% significance level with ±10% margin of error.

  ii. What did you observe? (Answer this in the quiz on MyUni.)

  iii. What is the required sample size if 60% of the population is regularly taking paracetamol, assuming 5% significance level with ±10% margin of error?

(d) The researcher wanted to aim for a lower margin of error.

  i. Use your function to produce a plot of the required sample size (Y-axis) against proportions (X-axis) ranging from 10% to 90%, with separate lines showing the required sample sizes for each of ±5%, ±8% and ±10% margin of error. **Hint**: remember to clearly label each line with a figure legend.

  ii. What did you observe? (Answer this in the quiz on MyUni.)

  iii. What is the required sample size if 50% of the population is regularly taking paracetamol, assuming 5% significance level with ±7% margin of error?

(e) The researcher is granted ethics approval to approach 100 participants only.

  i. What margin of error can be achieved with this sample size, assuming a population proportion of 50% and 5% significance level?

ii. A margin of error of 8% is desired. Assuming 5% significance level, what is the population proportion/s that can be estimated with this level of margin of error?

Your R script file should contain the function `proportion_ss`, the code to produce the required plots, and the code used to answer the questions in the quiz on MyUni.

For full marks you must include commented code to explain the steps in your function.

[20 marks]

[Question total: 20]

[[Assignment total: 54]]