STATS 2107
Statistical Modelling and Inference II
Assignment 3

Matt Ryan

Semester 2 2022

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

The aim of this question is to introduce you to asymptotic pivotal quantities and the Fisher-Z transformation. Fisher is the godfather of modern statistics, and did fundamental work developing many of the methods we still use today. The Fisher-Z transformation is a commonly applied transformation used to study the correlation between two normal random samples.

Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n) \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be bivariate normal observations with covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{bmatrix}.$$

This question is concerned with inference for the unknown correlation coefficient ρ. Let s_X^2 and s_Y^2 denote the sample variances of the X_i and Y_i respectively, and let s_{XY} denote the sample covariance between the X_i and Y_i. An obvious estimator for ρ is the Pearson sample correlation coefficient

$$r = \frac{s_{XY}}{s_X s_Y}.$$

We consider the Fisher-Z transformation given by

$$z = \operatorname{arctanh}(r) = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right),$$

and you can assume that, asymptotically,

$$z \sim N\left(\operatorname{arctanh}(\rho), \frac{1}{n-3}\right).$$

(a) Show that $H = \sqrt{n-3}\,(z - \operatorname{arctanh}(\rho))$ is a pivotal quantity (asymptotically) for ρ.

[2 marks]

(b) Using the pivotal quantity from Part (a), construct a symmetric $(1-\alpha)100\%$ confidence interval for ρ.

[4 marks]

(c) On MyUni, you will find a `.csv` file called `bivariate_normal_data.csv`. This data contains 100 observations from a bivariate normal distribution. Using this data, construct a 95% confidence interval for the true correlation ρ between X and Y. Round your confidence interval to 3 decimal places where appropriate, and provide any R code used in your calculations.

[6 marks]

(d) Using your confidence interval from Part (c), test the hypothesis

$$H_0 : \rho = \frac{1}{2}$$

at the $\alpha = 0.05$ level of significance. Be sure to give a precise conclusion of your findings.

[3 marks]

[Question total: 15]

Q2

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

The aim of this question is to explore the simple linear regression estimates through a different light. When you move to multiple regression, it is better to evaluate your estimates in "matrix form", as we like to call it. This question is to demonstrate that using the matrix form for these estimates will give you the simple linear regression estimates as you know and love them.

Consider

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \text{and} \quad X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}.$$

(a) Show that:

i. $X^T X = \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & \sum x_i^2 \end{pmatrix}$.

[2 marks]

ii. $\det(X^T X) = nS_{xx}$.

[2 marks]

(b) State a necessary and sufficient condition on $x_1,\ x_2,\ \ldots, x_n$ for $X^T X$ to be invertible.

[2 marks]

(c) Show that if $X^T X$ is invertible, then

$$\left(X^T X\right)^{-1} X^T \boldsymbol{y} = \begin{pmatrix} \bar{y} - \frac{S_{XY}}{S_{XX}} \bar{x} \\ \frac{S_{XY}}{S_{XX}} \end{pmatrix}$$

[9 marks]

Hint. Use the facts that

$$S_{XX} = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2\,,$$
$$S_{XY} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\,.$$

[Question total: 15]

Q3

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

The aim of this question is to explore a hypothesis test on a linear combination of normal random variables. You will extend the notion of the pooled variance estimator, and derive the distribution of a test statistic you create.

The three most popular menu items of a local burger restaurant are beef burger, chicken burger, and fish burger. Let W, X, and Y denotes the weekly sales of beef, chicken, and fish burgers, respectively. Suppose the random variables W, X, and Y are independent and normally distributed with means μ_1, μ_2, and μ_3 and variances σ^2, $a\sigma^2$, and $b\sigma^2$, respectively, where a and b are known constants. The profit for a beef burger is \$2, a chicken burger is \$1, and a fish burger is \$0.40. Hence, the expected weekly profit from these three menu items is $\theta = 2W + X + 0.4Y$.

(a) Find $E[\theta]$ and $\text{var}(\theta)$. State the distribution of θ.

[4 marks]

(b) The owner has made independent observations of W, X, Y for the past n weeks. Let $\bar{W}$, $\bar{X}$, and $\bar{Y}$ denote the sample means, and S_W^2, S_X^2, and S_Y^2 their respective sample variance. Let

$$\hat{\theta} = 2\bar{W} + \bar{X} + 0.4\bar{Y} \quad \text{and} \quad S_p^2 = \frac{1}{3}\left(S_W^2 + \frac{1}{a}S_X^2 + \frac{1}{b}S_Y^2\right).$$

Show that $T = \dfrac{\hat{\theta} - \mathrm{E}[\hat{\theta}]}{S_p\sqrt{\frac{4+a+0.16b}{n}}} \sim t_{3(n-1)}$.

[8 marks]

(c) Develop a hypothesis test for $H_0 : \theta = \theta_0$ versus $H_a : \theta \neq \theta_0$, where θ_0 is a constant, at the α significance level.

[3 marks]

[Question total: 15]

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 develop your R skills to answer questions about t-tests. In this question, you will be required to properly implement the `t.test` function in R, as well as demonstrate your understanding of the t-test by writing code to do it yourself. You will also have to develop a plot to submit to MyUni along with your Rscript.

Cores drilled from the sea floor provide information to better understand climatic changes, glaciation, and the evolution of deep-water and shallow-water patterns. The manager of an ocean drill company OzDrill, who operates drills in the Antarctic region, is interested in investing in drills in the Pacific region. A keen collaborator who operates drills in the Pacific region, Core Drill, has provided data on the performance of their drills. OzDrill would like to know if the performance of drills, measured as drilling time in hours, in the Pacific region is **shorter than** those in the Antarctic region.

The data file `Drill.csv` is available from MyUni. Details of the dataset:

Table 1: Variables in Drill.csv.

Variable	Details
Ship	Unique Drill ID
Time_hr	Driling time (in hours)
Region	Region of drill

(a) Let μ_1 be the true drill time (in hours) of drills in the Antarctic region, and let μ_2 be the true drill time (in hours) of drills in the Pacific region.

(i) What is the sample size for each region?
(ii) Calculate the mean and standard deviation of drilling time (in hours) for drills in each region.
(iii) Which test is appropriate?
(iv) Perform a Welch t-test using R, assuming 5% significance level, to test the hypothesis

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{vs} \quad H_a : \mu_1 - \mu_2 > 0\,.$$

For this hypothesis, state the test statistic, the P-value, and the corresponding 95% confidence interval.

(b) Another company, AllDrill, who operates drills in the Atlantic region, have also expressed their interest in collaborating with OzDrill. They have provided a summary performance of their drills:

$$\begin{aligned} \text{Number of drills} &= 38 \\ \text{Mean drilling time (in hours)} &= 25 \\ \text{Standard deviation of drilling time (in hours)} &= 2.14 \end{aligned}$$

(i) OzDrill would like to know if drills in the Atlantic region has a shorter drilling time compared to those in the Antarctic region. Write an R function to calculate Welch t-test, given the mean, standard deviation and sample size of each group. A 5% significance level is assumed.

Your function should follow the template provided below:

```
welch_t <- function(mu1, mu2, sd1, sd2, n1, n2){
  t_stat <- #code to calculate t-statistic here
  welch_df <- #code to calculate DF here
  p_val <- #code to calculate p-value here
  return(
    tibble(t_statistic = t_stat,
           DF = welch_df,
           P = p_val)
  )
}
```

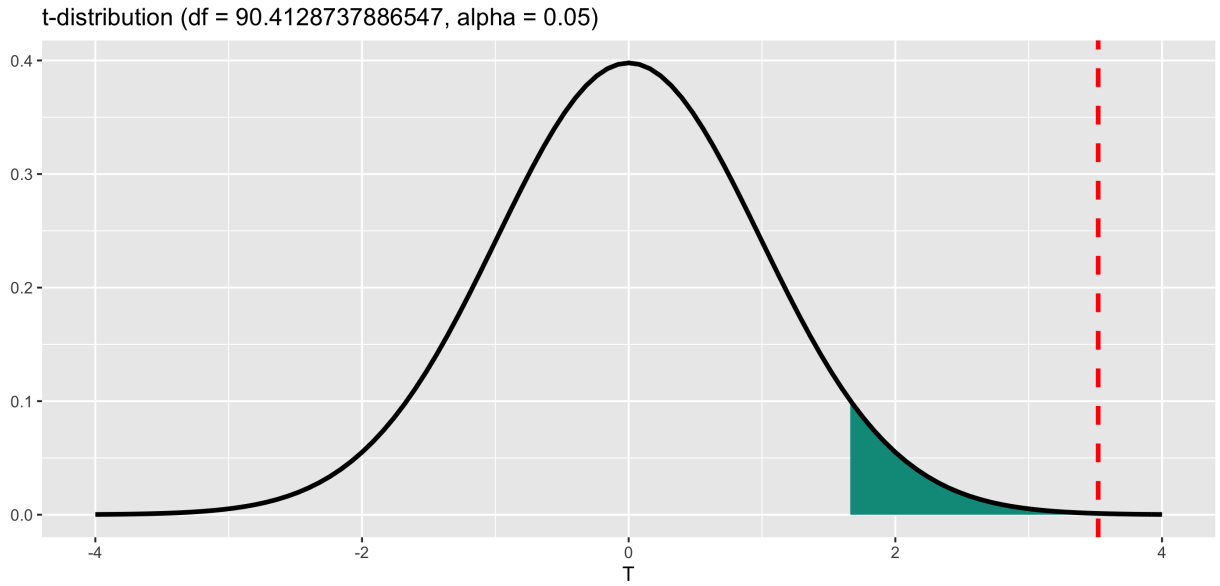
where:

- `mu1` and `mu2` is the mean of group 1 and 2, respectively,
- `sd1` and `sd2` is the standard deviation of group 1 and 2, respectively, and
- `n1` and `n2` is the sample size of group 1 and 2, respectively.

(ii) Perform a t-test with Welch approximation using your function to test whether the drilling time of drills in the Atlantic region is **shorter** than those in the Antarctic region, assuming 5% significance level. State the value of the test statistic, the degrees of freedom using Welch's approximation, and the P-value.

(iii) What can you conclude from the t-tests (from part a and b above) on the drilling time of drills in the Pacific and Atlantic region compared to the Antarctic region?

(iv) Produce a density plot for the t-distribution (similar to the example below), with the degrees of freedom from Welch's approximation as obtained in part (ii). The plot should also show the value of t-statistic as obtained in part (ii) (indicated by a red dashed line), with 5% significance level critical region shaded.



HINT: The example code below will produce a plot of normal density curve with $\mu = 3$ and $\sigma = 1$, with the region where $X \leq 3$ shaded:

```
ggplot(tibble(c(0,0)), aes(c(-3,9))) + # Get the plot set up, plot from -3 to 9
  geom_area(xlim=c(-3,3), # Define the area you want to shade
            stat="function", # What stat do you want
            fun=dnorm, # What function, in this case a normal density
            args=list(mean=3,sd=1), # parameters for dnorm
            fill="#00998a") + # choose a colour
  stat_function(fun=dnorm, # Time to add the density plot
                args=list(mean=3), # Give the parameters needed
                size=1.2) + # Make a little bigger
  labs(x = "X-AXIS", # Make it pretty
       y = "Y-AXIS",
       title = "TITLE")
## This next command will save your plot in your current working directory.
ggsave("A3_a1234567.png")
```

You will be asked to submit the required plot as a `.png` file as well as your R script. Your R script file should contain the function `welch_t`, the code to produce the required plot, 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: 65]]