STATS 2107
Statistical Modelling and Inference II
Assignment 1

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 purpose of this question is to show that the sample variance is an unbiased estimator of the population variance when you have independent observations. The key point of this question is understanding where the $n-1$ comes from in the formula of the sample variance.

Let $X_1, X_2, \ldots, X_n$ denote a sample of i.i.d. random variables with finite mean μ and finite variance σ^2. Consider the sample sum of squares given by:

$$S_{XX} = \sum_{i=1}^{n} (X_i - \bar{X})^2 \,.$$

a. Show that

$$\mathrm{E}[X_i X_j] = \begin{cases} \mu^2 & \text{if } i \neq j\,, \\ \mu^2 + \sigma^2 & \text{otherwise}\,. \end{cases}$$

[3 marks]

b. Hence show

$$\mathrm{E}[X_i \bar{X}] = \mu^2 + \frac{\sigma^2}{n}\,.$$

[2 marks]

c. Show that

$$\mathrm{E}[S_{XX}] = (n-1)\sigma^2$$

[4 marks]

d. Hence show that the sample variance S^2 is unbiased for σ^2.

[1 mark]

[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 with different types of estimators of parameters. You will be able to practice finding the bias and MSE of a couple of different types of estimators, and use the MSE to determine which is the best.

In this question you may assume the following two facts of independent random variables X and Y:

- $\mathrm{E}[XY] = \mathrm{E}[X]\mathrm{E}[Y]$; and

- $f(X)$ and $g(Y)$ are independent for continuous functions f and g.

We wish to estimate the area $A = bh\frac{1}{2}$ of a triangular display, with base b and height h. We therefore measure both height and base twice. Suppose these measurements are outcomes of independent random variables $X_1, X_2 \sim N(b, \sigma^2)$ and $Y_1, Y_2 \sim N(h, \sigma^2)$, where σ describes the accuracy of our measuring instrument. There are two natural ways to estimate the unknown area A. We can multiply $\frac{1}{2}$ with the average of base measurements and average of the height measurements, or we can take the average of the two estimated areas, that is,

$$T_1 = \frac{(X_1 + X_2)}{2} \times \frac{(Y_1 + Y_2)}{2} \times \frac{1}{2} \quad \text{and} \quad T_2 = \frac{(X_1 Y_1 \frac{1}{2}) + (X_2 Y_2 \frac{1}{2})}{2}.$$

(a) Show that T_1 and T_2 are both unbiased estimators of A.

[5 marks]

(b) Show that $\mathrm{var}(X_1 Y_1) = \sigma^2(\sigma^2 + b^2 + h^2)$. (Hint. Use the variance formula $\mathrm{var}(X) = \mathrm{E}[X^2] - \mathrm{E}[X]^2$)

[4 marks]

(c) Show that $\mathrm{cov}(X_1 Y_1, X_1 Y_2) = \sigma^2 h^2$.

[3 marks]

(d) Find the MSE of T_1 and T_2. Based on this, which estimator is preferred?

[8 marks]

[Question total: 20]

Q3

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

The purpose of this question is to investigate the properties of different types of estimators. This is to familiarise you with different types of estimators.

Let Y_1, Y_2, Y_3 be independent $\mathrm{Exp}\left(\frac{1}{\theta}\right)$ random variables with density

$$f(y) = \frac{1}{\theta} e^{-\frac{y}{\theta}} \quad \text{for } y > 0.$$

Consider the following estimators of θ:

$$\hat{\theta}_1 = Y_1, \quad \hat{\theta}_2 = \frac{Y_1 + 2Y_2 + 3Y_3}{6}, \quad \hat{\theta}_3 = \bar{Y}, \quad \hat{\theta}_4 = \min(Y_1, Y_2, Y_3).$$

(a) Find the bias of each of these estimators.

[5 marks]

(b) Find the variance of each of these estimators.

[2 marks]

(c) Which estimator has the smallest MSE?

[3 marks]

[Question total: 10]

Q4

Please submit your answer to this question online using MyUni. You will be asked to upload an R script file with the commands you used to complete the tasks in this question. Further information is found on MyUni.

The purpose of this question is for you to practise your data cleaning skills that you learnt in Practical 1. You are presented a new dataset and asked to clean it in R using the methods covered in Practical 1.

A survey in 2003 was conducted to study the TV viewing habits of Australians. The data is available on MyUni in an Excel spreadsheet called `survey2003_dirty.csv`. A description of the variables recorded are listed below:

- `Participant ID`: ID for participant survey
- `favourite genre`: Participant's favourite TV show genre (Action, Comedy, or Thriller)
- `sleep hour`: Average hours of sleep per day
- `TV hour`: Average hours spent per day watching TV
- `height`: Height (in cm) of participant
- `weight`: Weight (in kg) of participant
- `gender`: Participant's gender

For each of the following variables:

- `favourite_genre`
- `sleep_hr`
- `TV_hr`
- `height`
- `weight`
- `gender`

Perform the following:

1. Clean each of the variables using the methods described in Practical 1. This includes:

a. Ensure each variable is the right class (*i.e.* numeric).
b. Make sure NA values are correctly entered.
c. Identify any values that may be incorrectly entered.
d. Where possible, recode factors to the right values.

2. For each of the variables, produce an appropriate plot to look at the data. That is:

a. Look at histograms for numeric variables.
b. Look at bar charts for categorical variables.

3. For each quantitative variable, identify whether it is unimodel or bimodel; also whether it is symmetric, left-skewed or right-skewed. For the categorical variables identify the most common level. (Hint. Look at the distributions without the incorrectly entered data.)

4. Generate five-number summaries for quantitative variables. For categorical variables produce a frequency table. Identify any missing values.
5. Export the cleaned data into a CSV (comma separated values) file. (Hint. Type `?write_csv` into R console.)
6. Answer the questions in Assignment 1 (practical) on MyUni.
7. Upload the R script that you used to complete Tasks 1 to 6 above.

For full marks you must include commented code to explain why and how you cleaned each variable.

Please note the output csv file must follow the naming: `A1_aXXXXXXX.csv`. That is, when you save your clean data, you will call it this file name where `aXXXXXXX` is your Student ID number.

[20 marks]

[Question total: 20]

[[Assignment total: 60]]