STAT2203/7203 (S2-2023): Assignment 03

Due: 27-October-2023 @16:59

1. **[5 marks each]**

We have seen how to simulate from a distribution using the inverse-transform method; see §5.8 of the course notes as well as slide 8/14 of Lecture4-3. Another method to simulate random variables from a given distribution is using rejection sampling. This question concerns a particular application of rejection sampling.

Benford's law is a distribution on the integers $\{1, 2, \ldots, 9\}$ with probability mass function

$$f_D(d) = \log_{10}\left(\frac{d+1}{d}\right), \quad d \in \{1, 2, \ldots, 9\}.$$

We would like to be able to simulate the random variable D from this distribution. Suppose X has probability mass function

$$f_X(x) = \frac{1}{9}, \quad x \in \{1, 2, \ldots, 9\}$$

Conditional on $X = x$, the random variable Y has a Bernoulli$(f_D(x)/\log_{10}(2))$ distribution.

(a) Verify that $f_D(x)/\log_{10}(2) \leq 1$ for all $x \in \{1, 2, \ldots, 9\}$.

(b) What is the joint probability mass function of (X, Y)?

(c) Determine $\mathbb{P}(Y = 1)$.

(d) Determine the conditional probability mass function of X given $Y = 1$.

(e) This suggests we can simulate a random variable with probability mass function f_D using the following algorithm

```
Y = 0
While (Y = 0) {
Simulate X from a uniform distribution on {1,2,...,9}
Simulate Y from a Bernoulli distribution with success
probability f(X)/log10(2)
}
Return X
```

In each loop a new pair of random variables (X, Y) is simulated, independent of all previously simulated random variables. Implement this algorithm in `R` (or any programming language of your choice). You will need to use a `while` loop. In `R`, the general form is

```
while (cond) {
expressions
}
```

where `cond` is a length one logical vector.

(f) What is the distribution of the number of pairs of random variables (X, Y) that need to be simulated to simulate a single random variable from Benford's law?

2. **[5 marks each]**

For The following questions, work out your answers 'by hand'. You may still use `R` (or any other programming language) to obtain probabilities and quantiles from the appropriate distributions and calculate your final answers.

A study investigated if psychotherapy combined with limited administration of Methylenedioxymethamphetamine (MDMA) can reduce symptoms of post-traumatic stress disorder. Severity of symptoms was measured via the CAPS-IV score with higher scores indicating more severe symptoms. Forty-eight patients recruited to the study with twenty-four patients being randomly allocated each of the two dosage levels (Low – 40 mg, High – 125 mg). The primary outcome was the reduction in CAPS-IV score one month after the end of treatment.

The forty-eight patients at the commencement of the study had an average CAPS-IV score of 81.35 with a sample standard deviation of 17.54. At the end of treatment, the High dose group experienced an average drop in CAPS-IV score of 24.2 with a sample standard deviation of 23.1. The Low dose group experienced an average drop in CAPS-IV score of 12.7 with a sample standard deviation of 19.4.

(a) Determine a 95% confidence interval of the population mean CAPS-IV score of patients at commencement of the study.

(b) Does the data provide evidence that the high dose MDMA treatment is associated with a decrease in mean CAPS-IV score? State the null and alternative hypotheses, and determine the appropriate test statistic and p-value. What do you conclude?

(c) Researchers would like to determine if patients experience a greater decrease in CAPS-IV score with the high dose MDMA treatment than low dose MDMA treatment. State the null and alternative hypotheses, and determine the appropriate test statistic and p-value. What do you conclude?

(d) A secondary outcome was whether the patient experienced a drop of 20% or more in CAPS-IV score. In the high dose treatment group 11 patients experienced such

a drop in CAPS-IV score. Construct a 95% confidence interval for the population proportion of patients that would experience a 20% or more drop in CAPS-IV score with the high dose treatment.

(e) In addition to the 11 patients in the high dose treatment group who experienced a 20% or more drop in CAPS-IV score, 6 patients in the low dose treatment group also experienced a 20% or more drop in CAPS-IV score. The researchers would like to test if the population proportion of patients that would experience a 20% or more drop in CAPS-IV score is greater with a high dose treatment than the low dose treatment. State the null and alternative hypotheses, and determine the appropriate test statistic and p-value. What do you conclude?

(f) Are the assumptions/approximations you used for the analysis in part (e) valid? Justify your answer.

3. **[4 marks each]**

Exposure to ground level ozone (O_3) is believed to impair airway function in healthy individuals. To investigate this, researchers recruited 60 individuals (34 males and 26 females) and had them exercise for one hour on a cycle ergometer while breathing 0.30 parts per million of ozone. The Forced Expiratory Volume (FEV) and Forced Vital Capacity (FVC) of each subject was measured before and after the test and the change recorded as a percentage.

The file `ozone.csv` contains the following variables:

- `FVC` – Percentage change in Forced Vital Capacity
- `FEV` – Percentage change in Forced Expiratory Volume

(a) Run linear regression for Change in FEV% against Change in FVC% using `R` (or any programming language of your choice) and give the summary output. Produce diagnostic plots, namely scatterplot of residuals against fitted values and the normal quantile plot of residuals, for the linear regression fit. Give these captions and figure numbers and refer to them as needed in later questions.

(b) List the assumptions of the linear regression model. For each, explain whether or not there is evidence that this assumption is violated, based on the diagnostic plots.

(c) A researcher suggest that the linear regression model is not appropriate because the Change in FVC % does not have a normal distribution. Are they correct? Justify your answer.

For the following parts you may assume that the model assumptions hold.

(d) Report a 99% confidence interval for the slope of the linear regression model.

(e) Provide both a 90% prediction interval for the change in FEV% for a healthy

individual with a change in FVC of 10% and a 90% confidence interval for the mean change in FEV% for a healthy individual with a change in FVC of 10%. Briefly explain the difference of between the two intervals.

(f) Researchers believe that the changes in FEV% and FVC% are both due to a change in inspiratory capacity so that the intercept of the regression line should be zero. Is the result of the regression analysis consistent with this belief? State the null and alternative hypotheses, and report the appropriate test statistic and p-value. What do you conclude?

(g) Explain the meaning of the R-squared value in the regression output.

4. **[6 marks each]**

Consider the simple linear regression model as discussed in class where the observations are modeled $Y_i \overset{\text{iid}}{\sim} \mathcal{N}(\beta_0^\star + \beta_1^\star x_i, \sigma^2)$, $i = 1, \ldots, n$. Consider the least squares estimators, $\hat{\beta}_0$ and $\hat{\beta}_1$, for the respective coefficients $\beta_0^\star$ and $\beta_1^\star$.

(a) Show that

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \, .$$

(b) Show that

$$\text{Var}(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \right) .$$

100 marks in total

Note:

- This assignment counts for 10% of the total mark for the course.
- Although not mandatory, if you could type up your work, e.g., `LaTex`, it would be greatly appreciated.
- Show all your work and attach your code and all the plots (if there is a programming question).
- Combine your solutions, all the additional files such as your code and numerical results, **all in one single PDF file**.
- Please submit your single PDF file on Blackboard.