STATS121 Lab 4

Due 8th April at 11:30 PM NZST

Question 1 [2 marks]

1(a)

What is a sampling frame?

1(b)

What is the key difference between a sample survey and a census?

Question 2 [5 marks]

The famous iris dataset collected by Anderson (1935) gives the measurements in centimetres (cms) of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of the three species of iris. Sepals are the outer whorl of a flower (usually coloured green), while petals are the colourful inner whorl of a flower. The species are *Iris setosa*, *versicolor*, and *virginica*.

The `irisLengths.csv` dataset contains information about the 150 iris flowers and three out of the five variables listed here:

- `Sepal.Length` – The sepal length of a flower (in cm).
- `Petal.Length` – The petal length of a flower (in cm).
- `Species` – The species of a flower, either “`setosa`”, “`versicolor`”, or “`virginica`”.

Anderson, Edgar (1935). The irises of the Gaspe Peninsula, *Bulletin of the American Iris Society*, **59**, 2–5.

2(a)

Anderson was interested in estimating the average sepal and petal lengths of iris flowers for each species. Which data collection method, out of randomised experiment, observational study, or sample survey, best describes how the iris dataset was collected?

2(b)

Do you think the relationship between the sepal lengths and the petal lengths depends on the species of the iris?

2(c)

Write R code to produce a scatter plot of the sepal lengths by petal lengths. So the sepal lengths are along the y-axis, and the petal lengths are along the *x*-axis.

> *You can plot the species of each flower by adding* `col = Species.f` *argument to the* `plot()` *function. If you do, delete the comment, #, in front of the* `legend()` *function.*

2(d)

Briefly describe any features of the scatter plot produced in 2(c).

Question 3 [6 marks]

Airlines often have a policy of routinely overbooking flights. Let X be the number of passengers who cannot be boarded because there are more passengers than seats. The following table lists the incomplete probability distribution of X.

x	0	1	2	3	4	5
$\mathbb{P}(X = x)$	0.051	0.141	0.276	0.329	0.157	—

3(a)

What is the probability that exactly five passengers cannot board a flight? That is, $\mathbb{P}(X = 5)$.

3(b)

Update the object named `X.5` in the code chunk underneath the 3(b): Update X.5 heading to be equal to your answer for 3(a).

3(c)

What is the probability that at most two passengers cannot board a flight?

3(d)

What is the probability that more than two passengers cannot board a flight?

3(e)

What is the expected number of passengers who cannot board a flight? That is, $E(X)$.

3(f)

The function named `X.plotr(n)` in the code chunk underneath the 3(f): Simulation heading simulates n realisations of X and plots the observed proportions for $X = 0, \dots, X = 5$ against the provided probability distribution of X. For example, `X.plotr(10)` simulates ten realisations of X and plots the observed proportions of $X = 0, \dots, X = 5$.

Use the `X.plotr(n)` function to simulate X for n = 5, n = 50, and n = 500. Then, briefly comment on whether the observed proportions for each simulation is similar to the probability distribution of X.

3(g)

Comment on how your answer to 3(f) is related to the frequentist view of probability. *(**Hint:** Review Chapter 5.)*