Submit one .pdf file that includes the answers to both sets of questions. Also submit your .R script that you used to answer the questions in the R part.

R Question

1. (4 points) In lab you investigated the validity of a test to violations of assumptions. In this question you will examine the validity of t-based **confidence intervals** in different situation: small sample sizes and non-Normal distributions. We will use the case of a Gamma(2, 2) distribution, which has a mean of 1, to investigate.

   (a) Make a plot in `ggplot2` of a Gamma(shape = 2, rate = 2) density curve. Look back at the Module 5 homework if you can not quite remember how.

   (b) Write a function that (i) draws a sample of size n from a Gamma(2, 2) distribution, (ii) performs a t-test with `t.test()`, (iii) extracts the 95% confidence interval, and (iv) returns `TRUE` if the interval contains the true mean, and `FALSE` if it does not. (Hint: it's easiest to write code that works for one specific example, that is, pick a sample size, and write steps (i), (ii), (iii) and (iv), then turn that code into a function).

   (c) Use your function, along with `mean()` and `replicate()`, to find the proportion of t-based 95% confidence intervals in 100,000 samples of size 5, that contain the true Gamma(2,2) population mean.

   (d) Now repeat for samples of size 25, 50, and 100.

   (e) In two sentences interpret what is happening to the confidence intervals as the sample size increases, and why.

Conceptual Questions

Answer **all** of the following questions.

Each of the following scenarios describe a study that violates one of the assumptions of the proposed analysis. For **each** scenario:

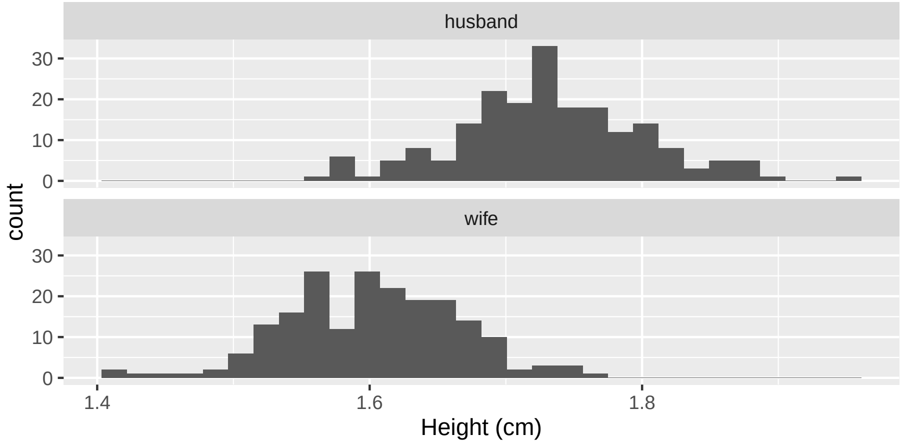
(a) Describe which assumption is most likely violated, and the evidence you have for the violation.

(b) Comment on whether we should expect any robustness from the procedure against the violation.

(c) Regardless of the robustness, make a suggestion for how the study or analysis could be improved to diminish (or remove entirely) the effect of the violation of the assumption.

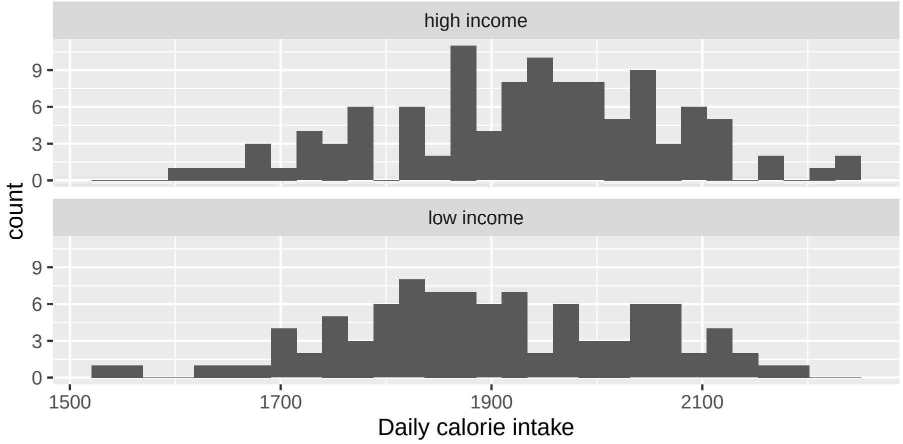
2. (2 points) The Great Britain Office of Population Census and Surveys collected data on a random sample of 170 married, opposite sex, couples in Britain, recording the age (in years) and heights (in cm) of the husbands and wives.

   They conduct a two-sample t-test to compare the mean height of husbands to the mean height of wives. Histograms of the heights are shown below.



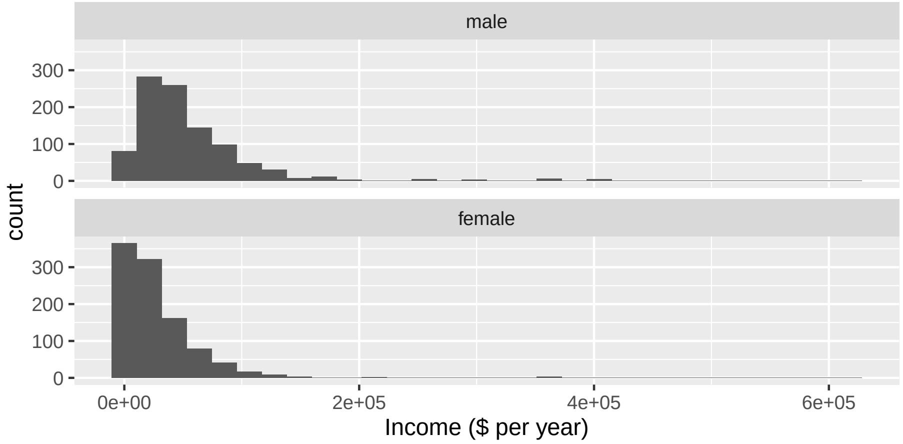
3. (2 points) In a study on the differences in diet between high and low income households, 50 low income and 50 high income households are randomly selected. Every adult in each household records their calorific intake for one week, and this is summarised to a daily average for each person. In total there are 110 adults in the high income households and 96 adults in the low income houses. The mean *average daily caloric intake* is compared between adults living in low and high income households using a two sample t-test.

   Histograms of the average calorie intakes in the study are shown below.



4. (2 points) In an effort to quantify gender inequality in income, the State of Oregon collects a random sample of 2000 residents with comparable qualifications and years of experience (in practice this is really hard to do, but for the purpose of this problem assume it was done well). They compare the mean income of females to the mean income of males using a two-sample t-test.

   Histograms of the incomes are provided below.


male
female
count
300
200
100
0
300
200
100
0
0e+00
2e+05
4e+05
6e+05
Income ($ per year)