Final R project

Group size: 2 students
Deadlines:

- have one member of the group e-mail me with your group names **by April 1st**. If I do not hear from you by 03/30 at 11:59PM, you will be assigned a random partner and will be notified via email.
- submit your final report by 9:00AM on **April 29th** by email to ehector@ncsu.edu. I will confirm when I have received it.

Guidelines:

- You are asked to prepare two reports, one for each part. In both parts, you should submit your R code. All code should be commented and annotated so that it is immediately obvious to me what your answer is. Make liberal use of white space to improve readability of your code.
- The quality of the writing of your reports will be taken into account in the grading. The reports should be written like papers, not like exams in which you are answering bulleted questions. There should be introductory and concluding paragraphs, and a clear flow between paragraphs. Use complete sentences, not sentence fragments. The report must be typewritten and contain no typographical errors (spelling, duplicate words, poor grammar, punctuation mistakes).
- Do not expect the reader to fill in explanatory text between displayed equations, if any, to explain their meaning. Every notation, symbol, and acronym should be defined prior to its first use. Every sentence should begin with a proper English word and not a symbol or number. Text or formulas should not extend past the right margin. The order of mathematical parentheses is $[(\cdot)]$.
- You should submit one document that contains both reports and the R code for both parts.
- Each report should include the authors' names and a bibliography at the end if any resources are cited. You are encouraged to use LaTeX but it is not required.
- **At the end of the two reports, you should detail the contribution of each member of the team, and explain who completed which parts**. It is not sufficient to state that all members contributed equally, you must provide details.

Part I (40 points)

A dataset is available on wolfware. This dataset represents two independent samples of systolic blood pressure (`sysBP`) for smokers (`currentSmoker=1`) and nonsmokers (`currentSmoker=0`) in the Framingham heart study. If you are curious, the full documentation on the Framingham heart study is available here [1]. We will make the assumption that the data are both random samples from normal distributions with parameters (μ_1, σ_1^2) and (μ_2, σ_2^2) respectively, where (μ_1, σ_1^2) are the mean and variance for nonsmokers and (μ_2, σ_2^2) are the mean and variance for smokers. We wish to make inference on the difference of means.

There are two tests we can use in this situation: the two-sample t-test when equal variance is assumed (pooled) and when unequal variances are assumed (with Satterthwaite approximation for the degrees of freedom). The test statistics and forms for the rejection regions are given in the notes for Chapter 9.

You must conduct a hypothesis test for this dataset.

1. You should download the data from Wolfware and read it into your R session.

2. Perform both two-sample t-tests on the data at significance level $\alpha = 0.05$. You should test your hypothesis two ways: using the p-value and confidence interval methods. For both, you should provide the values of all component parts and report the decision for each test. You must write code to do the tests yourself; you may not use a function/package in R.

3. Using plots, comment on how well the normality assumption is met by the data.

4. Discuss in detail which test you prefer for this dataset and why. Make sure you justify your answer with theoretical arguments, additional tests (if any are required) and your observed results.

5. Write a short report (no more than 1 page) with your findings from both tests, comment on Normality assumption, and discussion on which test to use. Include your commented and annotated R code with plots after your report. Be sure to set up the hypothesis test using all required details so as to make your report self-contained. The reader should not have to refer back to the prompt to understand what you are doing.

[1] `https://biolincc.nhlbi.nih.gov/media/teachingstudies/FHS_Teaching_Longitudinal_Data_Documentation.pdf?link_time=2021-03-10_15:22:56.639690`

Part II (60 points)

Conduct a simulation study to compare the two hypothesis testing procedures from Part I. To perform a simulation study for a hypothesis testing procedure, you should proceed through the following steps:

- Generate many data sets, say 100, from a null or alternative situation based on the assumptions of the test.
- Determine if you reject or fail to reject for each generated dataset.
- Look at the sample proportion of times you reject as a measure of the Type-I error rate α or the power.

The level of detail should be sufficient that a reader can replicate your simulation study based solely on the information in your report. Remember that a simulation study always includes replication for each situation, here 100.

We want to look at the performance of the two tests in terms of controlling α and having larger power as we vary the following items (look at all combinations of these!)

1. True variance: consider $\sigma_1^2 = 1, 4, 9$ and $\sigma_2^2 = 1$.
2. Sample size: consider $n_1 = 10, 30, 70$ and $n_2 = 10, 30, 70$.
3. True means equal vs true mean difference $(\mu_1 - \mu_2)$ of -5, -1, 1, 5.

Write a short report for your simulation study (no more than 2 pages). This should include the purpose of the simulation, the design (including all situations considered), the results of your simulations using figures, and your conclusions. You are expected to discuss why a numerical simulation study can be used to evaluate the two testing procedures. Include your commented and annotated R code after your report.

There are a lot of situations to compare above. You should give your results in terms of plots for ease of visualization. You can overlay multiple results in one plot using appropriate legends. The quality of your graphical visualizations will be taken into account in the grading of your reports.