HW 2 - due 04/21 at 11:59 pm.

Math 181B, Spring 23, Rava

Follow closely the 'Hw guide' under Files in the folder 'Course Contents' on how to write, scan and submit your homework.
On any problem involving R, you should include your code and output as part of your answer. You may take a screenshot of the code/output, or write it by hand.
Be careful with notation, remember to define the parameters and the random variables you intend to use.

1 Exercise 1

In class we studied the F test, that is used to compare the variances of two independent populations. One key assumption of the test is that the two populations have normal distribution. While many inferential procedures based on normal distributions perform well as normality is lost as long as sample sizes are big enough, the F test is quite sensitive to the normality assumption. This means that the F test cannot be used when normality is lost even when sample sizes are big. In this exercise you will perform simulations in R to convince yourself that this is the case.

a) [5 points] In R sample 50 numbers from a normal distribution with mean 1 and standard deviation 2. Sample also 50 numbers from a normal distribution with mean 0 and standard deviation 2. Compute the sample variances of the two samples (s_X^2, s_Y^2) and then compute their ratio $f = \frac{s_Y^2}{s_X^2}$. Repeat the process 1000 times and record in a vector the 1000 obtained $f = \frac{s_Y^2}{s_X^2}$. (Advice: To do that create a function that samples the two samples and return the value of f. Use then function replicate to apply the function created 1000 times. Remember that you can always use ?replicate to ask R for help).
At this point you should have a vector that records 1000 realizations of $F = \frac{S_Y^2}{S_X^2}$, that according to the theory, has a $F_{49,49}$ distribution. (Make sure that you know why the degrees of freedom are 49 and 49).
Plot a histogram of your 1000 realizations of F and superimpose to the histogram the density of the $F_{49,49}$ distribution. (Hint: make sure to not plot an histogram of the frequencies. Use ?hist to understand how).
Do your simulations agree with the theory?

b) [5 points] We are now going to use simulations to convince ourselves that indeed the F test described in class guarantees probability of type I error equal to α, as long as the two populations have normal distribution. In R sample 50 numbers from a normal distribution with mean 1 and standard deviation 2. Sample also 50 numbers from a normal distribution with mean 0 and standard deviation 2. Compute the sample variances of the two samples (s_X^2, s_Y^2) and then compute their ratio $f = \frac{s_Y^2}{s_X^2}$. According to the test described in class you reject the null $H_0 : \sigma_X^2 = \sigma_Y^2$ at significance level $\alpha = 0.01$ if $f \leq F_{0.005,49,49}$ or $f \geq F_{0.995,49,49}$. Check if your f falls into the rejection region and report 1/TRUE if it does, 0/FALSE if it doesn't.
Repeat the process 1000 times and record in a vector the results of the 1000 tests. At this point you should

have a vector that records the results of 1000 tests. TRUE/1 means that you rejected the null, FALSE/0 means that you have failed to reject the null. According to the theory, the proportion of TRUE/1 should be around $\alpha = 0.01$. (Remember that in this case we are simulating from two populations with equal variance so we are simulating under the null).
Verify that this is the case and that indeed the procedure described in class guarantees probability of type I error equals to α.

c) [4 points] Repeat exercise a) and b) but now simulate both samples from an exponential distribution with $\lambda = 10$. What do you notice?

2 Exercise 2

Your friend ripped his wetsuit and you are trying to advice him on a special glue to fix it. A company markets two brands of glue - regular and a more expensive one that claims to dry faster. Your friend decide to test this claim. On a big piece of neoprene, they perform 20 equal tears. They then apply the regular glue on 10 of them, chosen at random, and the fast glue on the others. They record the drying times. The results can be found in the files 'Regular.csv' and 'Fast.csv' available on Canvas. You want to use the data to conduct an hypothesis test with $\alpha = 0.03$.

a) [6 points] First you need to understand if the variances of the two populations can be assumed equal. Conduct an HT with significance level 0.05. You can use R to find the quantities needed but you cannot just use a built-in function that performs the test for you. Make sure to report all the lines of code needed, including the ones used to import the two files. Make sure to define parameters, hypotheses, comment on assumptions and write a meaningful conclusion.

b) [2 points] Based on your results of part a), do you think a 95% CI for the ratio between the two variances would contain 1? Explain your reasoning without actually constructing the CI.

c) [6 points] According to your result of part a) conduct the appropriate HT to test if indeed on average you tend to have longer surfer sessions in the morning. Use $\alpha = 0.03$. You can use R to find the quantities needed but you cannot just use a built-in function that performs the test for you. Make sure to report all the lines of code needed, including the ones used to import the two files. Make sure to define parameters, hypotheses, comment on assumptions and write a meaningful conclusion.

d) [2 points] Verify your results of part a), b) and c) using the R built-in functions that perform the tests for you.

3 Exercise 3

You have two roommates, Marla and Kate, that are fighting. Marla thinks that she does a better job cleaning. Kate of course thinks that it is not true. They are asking what you think. You decide to collect some data and use your statistical knowledge to decide whether to support Marla or Kate. You start to record whether Marla and Kate clean the stove after they cook. You watch Marla cooking 50 random times. She cleans the stove right after being done 23 times. You watch Kate cooking 50 random times.

a) [4 points] You report [-0.275,0.115] as 95% CI for the true difference between the proportion of time Marla cleans the stove after cooking and the proportion of time Kate cleans the stove after cooking. Find the number of time Kate cleans the stove right after being done in your sample.