MTH 324, Lab #7, Fall 2022   Name ____

Instructions: Follow along with the tutorial portion of the lab. Replicate the code examples in R on your own, along with the demonstration. Then use those examples as a model to answer the questions/perform the tasks that follow. Copy and paste the results of your code to answer questions where directed. Submit your response file and the code used (both for the tutorial and part two). Your code file and your lab response file should each include your name inside.

Two-Sample Hypothesis Tests

As we move into two-sample tests, we want less and less to conduct tests using formulas. We can, but we'll focus in this lab on using R's built-in formulas.

Let's start with testing for two proportions.

We used the prop.test() function in the last lab for one-sample proportions. We can use the same function by replacing x and n with vectors.

Consider the example:

A car manufacturer aims to improve the quality of the products by reducing the defects and also increase the customer satisfaction. Therefore, he monitors the efficiency of two assembly lines in the shop floor. In line A there are 18 defects reported out of 200 samples. While the line B shows 25 defects out of 600 cars. At α 5%, is the differences between two assembly procedures are significant?

```
prop.test(x = c(18, 25), n = c(200, 600), alternative = "two.sided")
```

Let's look at the output of the test.

```
        2-sample test for equality of proportions with continuity correction

data:  c(18, 25) out of c(200, 600)
X-squared = 5.9722, df = 1, p-value = 0.01453
alternative hypothesis: two.sided
95 percent confidence interval:
 0.002236347 0.094430320
sample estimates:
    prop 1     prop 2
0.09000000 0.04166667
```

The p-value here is less than the standard significance level, so we can reject the null hypothesis. This is good evidence that the rate of defects is not the same on the two lines.

Behind the scenes, this test is conducted under the assumptions of a χ^2 test, which we haven't covered yet. However, in the two-sample case, the results should be substantially similar.

We will not do a z-test example since we rarely use them, but the test, like the t-test function, does allow for an x and a y input variables, so if you have the special case that allows for a two-sample z-test of means, then the syntax should be similar to our t-test examples. Or check the documentation linked in the last lab.

Our two-sample t-tests come in several flavors, so we are going to start with the dependent (paired) case. Let's look at an example.

From a population of 10 students, the students were given the same test at the beginning and end of the school year. Use the Paired t-Test to determine if the average score of the 2nd test has improved over the average score of the 1st test.

Beginning	End
76	89
52	81
78	89
80	92
67	82
65	86
72	91
57	88
82	95
88	100

We can enter this data into R as vectors.

```
x1<-c(76,52,78,80,67,65,72,57,82,88)
x2<-c(89,81,89,92,82,86,91,88,95,100)
```

Recall that for a paired test, we are essentially doing a one-sample test on the differences. Here, we want $\mu_2 > \mu_1$ as the alternative.

```
t.test(x2-x1,mu=0,alternative="greater")
```

Then we can look at the output.

```
        One Sample t-test

data:  x2 - x1
t = 7.6338, df = 9, p-value = 1.606e-05
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
 13.37367      Inf
sample estimates:
mean of x
     17.6
```

Our P-value is quite low, so we can reject the null hypothesis of equality and conclude that the training did improve test scores.

Alternatively, we can use the paired option set to TRUE to conduct a paired test.

```
t.test(x2, x1 ,mu=0,alternative="greater", paired=TRUE)
```

The output produces the same results.

For the independent test, we had the option to conduct the test with a pooled standard deviation (under the assumption that the standard deviations are the same), or the unpooled case where we don't make any assumptions about whether the standard deviations (variances) are the same or not.

We load the data into R from the lab 7 data file and then separate the data by gender.

```
library(readxl)
data <- read_excel("daemen/324lab7data.xlsx")


library(dplyr)
data1<-filter(data, Gender == 1)
data0<-filter(data, Gender == 0)
```

We can put this in our t.test() function and use the var.equal option to specify TRUE for a pooled test, and FALSE for an unpooled test.

```
t.test(data1$`Annual Salary`,data0$`Annual Salary`,var.equal=FALSE, mu=0, alternative='two.sided')
```

Let's look at the results.

```
        Welch Two Sample t-test

data:  data1$`Annual Salary` and data0$`Annual Salary`
t = -3.0139, df = 171.28, p-value = 0.002971
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -21306.640  -4442.603
sample estimates:
mean of x mean of y 
 65910.08  78784.71 
```

This test says we can reject the null hypothesis of equality at the standard significance level. There is good reason to think that the two genders are paid differently.

Don't forget to test your assumptions for each of these tests. Are the proportion distributions sufficiently normal? Check the rule of thumb. Are the data for the t-tests sufficiently normal? Check a boxplot, histogram, or normal probability plot. If not, is the sample size big enough to overcome that fact? If you are using a pooled test, are the standard deviations about the same? Is the data paired or independent?

Tasks

Conduct the following hypothesis tests. Be sure to clearly state your hypotheses and check all assumptions.

1. The data in the table below shows data about body image separated by gender.

		Body Image			
		About Right	Overweight	Underweight	**Total**
Gender	Female	560	163	37	760
	Male	295	72	73	440
	Total	855	235	110	1200

   Conduct a test to see if the proportions of men and women who think they are underweight the same or different. Paste your results here along with any tests of assumptions.

2. Two operators are checking the same dimension on the same sample of 10 parts. Below are the results. Is there a significant operator measurement error? Test at the 5% significance level. Paste your results here along with any tests of assumptions.

S.No	Operator 1	Operator2
1	63	65
2	56	57
3	62	60
4	59	58
5	62	59
6	50	57
7	63	63
8	61	61
9	56	58
10	63	64

3. A researcher collected data from 11 students to see if the number of study hours was affected by gender. Paste your results here along with any tests of assumptions. Did you use a pooled or unpooled test? Explain your reasoning.

	A	B	C
1	Female	Male	
2	26	23	
3	25	30	
4	43	18	
5	34	25	
6	18	28	
7	52		
8			

References:

1. Discovering Statistics Using R. Andy Field, Jeremy Miles, Zoe Field. (2012)
2. https://book.stat420.org/applied_statistics.pdf
3. https://scholarworks.montana.edu/xmlui/handle/1/2999
4. https://www.rstudio.com/resources/cheatsheets/
5. http://www.sthda.com/english/wiki/two-proportions-z-test-in-r
6. https://sixsigmastudyguide.com/two-sample-test-of-proportions/
7. https://www.solver.com/t-test-paired-two-sample-means
8. https://data-flair.training/blogs/hypothesis-testing-in-r/
9. https://bolt.mph.ufl.edu/6050-6052/unit-1/case-c-c/
10. https://sixsigmastudyguide.com/paired-t-distribution-paired-t-test/
11. https://www.excel-easy.com/examples/t-test.html
12.