MTH 324, Lab #10, 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.

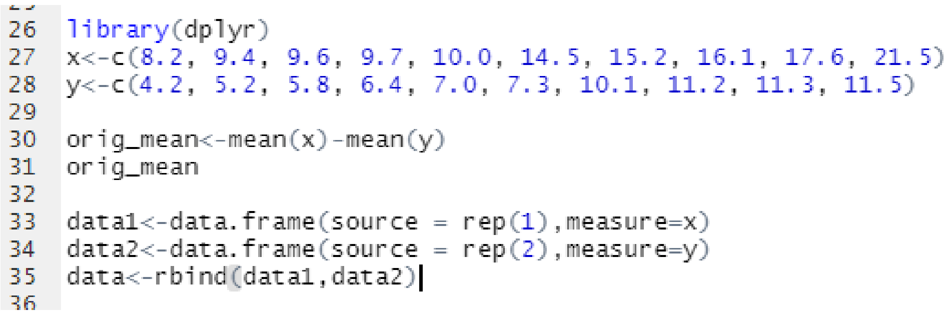
Computational models: Permutations and Bootstrapping

Let's look at some examples of bootstrapping and permutation tests. While we have looked at simple cases earlier in the course, we are going to look at some slightly more complex cases. The first will be to return to some data we used in an example in a previous lecture for the two-sample Wilcoxon test. (These are the sample examples from the lecture.)

Example.
A study of guinea pigs tests the effects of orange juice vs. synthetic ascorbic acid on odontoblasts. The data is below.

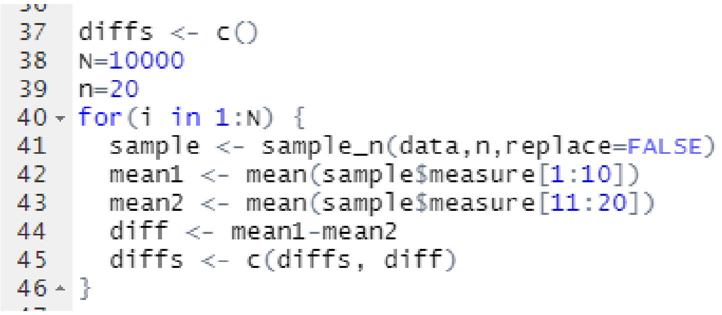
Orange Juice	8.2	9.4	9.6	9.7	10.0	14.5	15.2	16.1	17.6	21.5
Ascorbic Acid	4.2	5.2	5.8	6.4	7.0	7.3	10.1	11.2	11.3	11.5

We are going to go through the example again, but now is your chance to repeat it in R. First, we'll enter the data and load the required package.



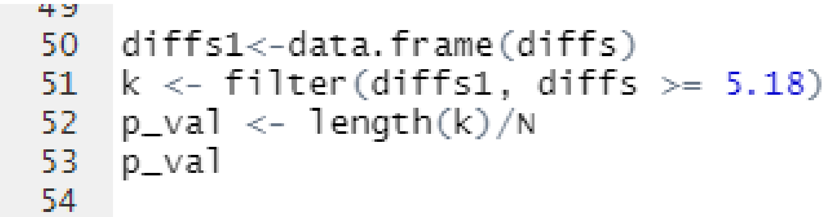
```r
library(dplyr)
x<-c(8.2, 9.4, 9.6, 9.7, 10.0, 14.5, 15.2, 16.1, 17.6, 21.5)
y<-c(4.2, 5.2, 5.8, 6.4, 7.0, 7.3, 10.1, 11.2, 11.3, 11.5)

orig_mean<-mean(x)-mean(y)
orig_mean

data1<-data.frame(source = rep(1),measure=x)
data2<-data.frame(source = rep(2),measure=y)
data<-rbind(data1,data2)
```

Then we set up our sampling procedures and compute our differnces of means. Starting with the permutation test.



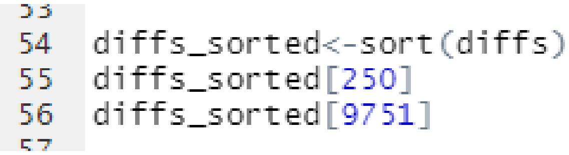
```r
diffs <- c()
N=10000
n=20
for(i in 1:N) {
  sample <- sample_n(data,n,replace=FALSE)
  mean1 <- mean(sample$measure[1:10])
  mean2 <- mean(sample$measure[11:20])
  diff <- mean1-mean2
  diffs <- c(diffs, diff)
}
```

Once we have our differences, we need to find the number that are greater than our original value.



```r
diffs1<-data.frame(diffs)
k <- filter(diffs1, diffs >= 5.18)
p_val <- length(k)/N
p_val
```

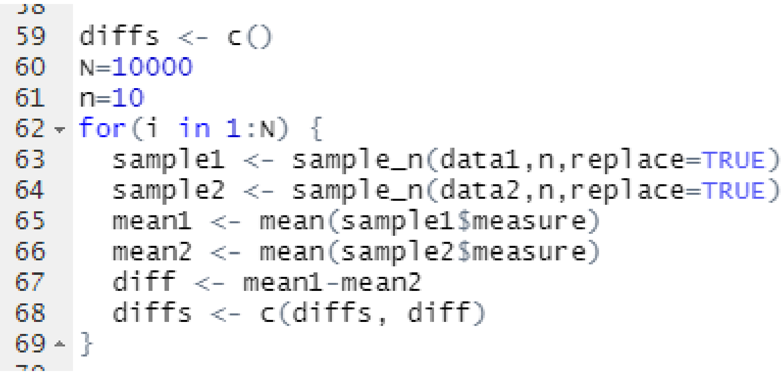
If we want to construct our confidence interval, we need to sort the list to find the relevant margin of error.



```r
diffs_sorted<-sort(diffs)
diffs_sorted[250]
diffs_sorted[9751]
```

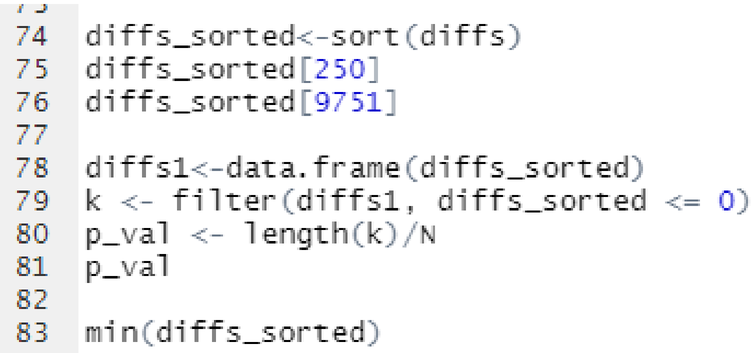
Remember that these values depend on having 10,000 differences. If you increase the number of trials, you'll need to adjust the numbers to correspond to 2.5% in each tail for a 95% confidence interval. Or $\alpha/2$ for any other confidence level.

Now, bootstrapping the same data.



```r
diffs <- c()
N=10000
n=10
for(i in 1:N) {
  sample1 <- sample_n(data1,n,replace=TRUE)
  sample2 <- sample_n(data2,n,replace=TRUE)
  mean1 <- mean(sample1$measure)
  mean2 <- mean(sample2$measure)
  diff <- mean1-mean2
  diffs <- c(diffs, diff)
}
```

Then we can calculate our confidence interval and P-value estimate.



```r
diffs_sorted<-sort(diffs)
diffs_sorted[250]
diffs_sorted[9751]

diffs1<-data.frame(diffs_sorted)
k <- filter(diffs1, diffs_sorted <= 0)
p_val <- length(k)/N
p_val

min(diffs_sorted)
```

See the class notes from the last lecture for additional details. You can also look back at the end of Lab #5 for the one-sample example. After running these examples yourself (and feel free to experiment with the number of trials and the bootstrapping sample size), use these models to complete the tasks.

Tasks

Use the data in the Excel file to complete the tasks. Use at least 10,000 trials.

1. The data on sheet 1 is on husbands and wives purchasing a car at the same dealership. Use bootstrapping to construct a confidence interval for the paired test. How does your result compare with the results of a paired t-test? Check any assumptions with graphs.

2. The data on sheet 2 is on the impact of exercise on cholesterol levels. Conduct a permutation test on this data to determine if there is sufficient evidence to believe that exercise reduces cholesterol. Check any assumptions with graphs.

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. https://statisticsbyjim.com/hypothesis-testing/bootstrapping/
6. https://www.mastersindatascience.org/learning/introduction-to-machine-learning-algorithms/bootstrapping/
7. https://towardsdatascience.com/bootstrapping-statistics-what-it-is-and-why-its-used-e2fa29577307
8. https://online.stat.psu.edu/stat555/node/119/
9. https://data-flair.training/blogs/bootstrapping-in-r/