STAT 5000 LAB #1
FALL 2024 OPTIONAL NAME: _____

Randomization Tests in R

The following data set comes from a small experiment on the effect of sodium intake on systolic blood pressure for men with Stage 1 hypertension. Blood pressure is composed of two measurements: systolic and diastolic pressure. Systolic blood pressure measures the amount of pressure in the arteries when the heart beats and diastolic blood pressure is the pressure when the heart is at rest. A normal blood pressure is written as 120/80 mmHg, where 120 is the systolic number and 80 is the diastolic number and the units are millimeters of mercury. Hypertension is a medical condition characterized by higher than normal blood pressure readings. In the first stage of hypertension, systolic pressures range from 140 to 159 mmHg or diastolic pressures range from 90 to 99 mmHg. The research question is this study is:

> Does sodium intake affect systolic blood pressure in men with Stage 1 hypertension?

Twenty men with Stage 1 hypertension are recruited through their primary care physicians to participate in the study. Men are randomly assigned to one of two treatments: either a low sodium diet (50 mmol/day) or a high sodium diet (200 mmol/day). After three months, the systolic blood pressure of the 20 men is measured. The data are provided in the `hypertension.csv` file in Canvas.

The bullet-points below walk through the R commands you will need to complete the lab assignment. Add these commands to a new R program in RStudio as you read through the explanations.

- Download the `hypertension.csv` file from Canvas and save it to your computer. The file has two columns: `sodiumdiet` is a factor variable with two levels - `low` or `high` - to indicate the treatment groups, `bloodpressure` is a numeric variable representing the systolic blood pressure response. Read the data into RStudio using the "Import Dataset" tool. Make sure to change the treatment variable to a `factor`. The associated R code will look something like:

```
library(readr)
hypertension <- read_csv("hypertension.csv",
                 col_types=cols(sodiumdiet=col_factor(levels=c("low", "high"))))
View(hypertension)
```

- Include the data set in your output by using the `print()` function:

```
print(hypertension)
```

- Summary statistics for the response variable within each treatment group can be obtained using the `summary()` and `sd()` functions in combination with the `$` dollar-sign to select columns of the dataframe and `[]` square brackets to subset.

```
summary(hypertension$bloodpressure[hypertension$sodiumdiet=="low"])
sd(hypertension$bloodpressure[hypertension$sodiumdiet=="low"])

summary(hypertension$bloodpressure[hypertension$sodiumdiet=="high"])
sd(hypertension$bloodpressure[hypertension$sodiumdiet=="high"])
```

  Or, if you prefer the *tidy* version explained in the *R for Data Science* book (as compared to *base* R functions), see Sections 4.5.1 and 4.5.2 for details on the following code chunk:

```
library(tidyverse)
hypertension |>
     group_by(sodiumdiet) |>
     summarize(
          Y_min = min(bloodpressure),
          Y_Q1 = quantile(bloodpressure, 0.25),
          Y_med = quantile(bloodpressure, 0.5),
          Y_Q3 = quantile(bloodpressure, 0.75),
          Y_max = max(bloodpressure),
          Y_IQR = Y_Q3 - Y_Q1,
          Y_mean = mean(bloodpressure),
          Y_sd = sd(bloodpressure)
     )
```

- While summary statistics are nice to use to compare values of the response variable between the two groups, a visual representation of these values would also be helpful. There are several options to choose from, but side-by-side box-plots are one of the best in determining differences in values between groups. The `boxplot()` function specifies the response variable before the ~ symbol followed by the treatment variable. The command below also includes formatting commands to label the axes and add a title.

```
boxplot(hypertension$bloodpressure ~ hypertension$sodiumdiet, xlab="Sodium Diet",
        ylab="Blood Pressure (mmHg)", main="Hypertension Experiment")
```

  Or, if you prefer the *tidy* version explained in the *R for Data Science* book (as compared to *base* R functions), see Section 11.5.1 for details on the following code chunk:

```
ggplot(hypertension, aes(x = sodiumdiet, y = bloodpressure)) +
     geom_boxplot() +
     ggtitle("Hypertension Experiment") +
     xlab("Sodium Diet") +
     ylab("Blood Pressure (mmHg)")
```

- The summary statistics and visual displays of the data can provide some indications of how you might answer the research question, but they can't be used directly to perform inference. For inferential purposes, we use a randomization test. There is no nice, easy built-in function in R to perform randomization tests, so we need to write one ourselves:

```
randomization.test <- function(response, treatment, Nsamps=10000, the.seed=500){
     ts <- rep(NA, Nsamps)
     n <- length(response)
     set.seed(the.seed)
     for(s in 1:Nsamps){
          permute <- sample.int(n)
          new.group <- treatment[permute]
          m1 <- mean(response[new.group==levels(treatment)[1]])
          m2 <- mean(response[new.group==levels(treatment)[2]])
          ts[s] <- (m1 - m2)
     }
     hist(ts, main="Randomization Distribution",
          xlab="Test Statistics (Difference in Mean Response)",
          ylab="Count")
```

```
    xbar1 <- mean(response[treatment==levels(treatment)[1]])
    xbar2 <- mean(response[treatment==levels(treatment)[2]])
    obsT <- (xbar1 - xbar2)
    p <- mean(abs(ts)>=abs(obsT))
    return(list(test_stats = ts, observed=obsT, p_value=p))
}
RT = randomization.test(hypertension$bloodpressure, hypertension$sodiumdiet)
```

You can print the results returned by this function, and find and print the values that go into the randomization p-value using the following code:

```
RT
RT$test_stats[which(abs(RT$test_stats)>=abs(RT$observed))]
```

Assignment

1. Calculate the sample mean score for each treatment group. What is the difference in the two sample means?
2. Use R to create a comparative box-plot for the sample mean score for each treatment group. Describe what you see.
3. What are the null and alternative hypotheses for the randomization test necessary to explore the research question?
4. Conduct a randomization test for these data in R (be sure to keep the random seed set at 500 so everyone gets the same answer) and study the reference distribution for the difference in the sample means for the 10,000 random assignments of treatments to subjects. Describe the shape, center and variability of this distribution.
5. Locate the observed difference in the sample means from part (a) on the reference distribution. Given the observed difference in the sample means from part (a), what is the p-value for this randomization test?
6. Interpret the results of the test in the context of the research question.
7. What aspects of the data collection in this experiment would need special attention by the researcher?