Stat 5000 Homework #1
Fall 2024 due Fri, Sep 6th @ 11:59 pm
Name: Samuel Olson

Directions: Type or clearly handwrite your solutions to each of the following exercises. Partial credit cannot be given unless all work is shown. You may work in groups provided that each person takes responsibility for understanding and writing out the solutions. Additionally, you must give proper credit to your collaborators by providing their names on the line below (if you worked alone, write "No Collaborators"):

Collaborators: Sarah Kilgard, Ben Moolman

1. [+10]: Some researchers were interested in studying the effects of different fertilizer amount (Low Nitrogen and High Nitrogen) and different genotypes (one energy line and one grain line) of sorghum on biomass. For each genotype, six pots of one-week-old seedlings were available, and each pot held one seedling. For each genotype, the researchers randomly assigned three pots to high nitrogen treatment (H) and the remaining three pots to low nitrogen (L) treatment. After 2 weeks, the fresh weight for each seedling was measured. In total, there are 12 observations.

1. Identify the experimental units.
   A pot containing one one-week-old seedling

2. Identify the observational units.
   [A pot containing] one one-week-old seedling

3. Identify the treatments. The treatments were either a high nitrogen treatment (H) or a low nitrogen treatment (L)

4. Identify the response variable. The biomass of the seedling after 2 weeks of being treated (weight of the sorghum)

5. Does the experiment utilize replication? Answer yes/no and provide a brief justification.
   Yes, because a treatment is applied independently to two or more experimental units; for each genotype we have two or more (three to be exact) who receive either the high or low nitrogen treatment.

6. Does the experiment utilize blocking? Answer yes/no and provide a brief justification.
   Yes, as similar experimental units are grouped together (genotype), followed by treatments being assigned to the same number of experimental units (via balance), as researchers first divided up the seedlings by genotype and then randomly assigned the treatments. A block is a particular genotype, and there are differences in the treatment a particular genotype is given.

7. Does the experiment utilize randomization? Answer yes/no and provide a brief justification.
   Yes, the experiment utilizes randomization First to detail the two types of randomization for our purposes: Randomization to units within a group (genotype), and secondly randomization in the selection of units within a group. At a high level randomization is occuring as there is the random assignment of treatments to experimental units (seedlings), the first of the aforementioned points. However, the second point is not true, as the allocation of experimental units to groups is not random, as the genotypes are predefined.

2. **[+10]:** A statistics teacher wanted to determine if having business students use clickers to respond to questions posed in a business statistics class would improve student learning. The teacher decided to have students use clickers in one class of introductory business statistics 226. She did not have students use clickers in a second class of business statistics 226 that she taught during the same semester. She tossed a coin to select the class to use the clickers. She used the same book and the same lectures in both classes and gave the same assignments and same exams to both classes. There were 90 students in each class. At the end of the semester, she compared the final exam scores for the students in the class that used clickers to the final exam scores for the students in the class that did not use clickers.

1. Identify the experimental units.
   The class of business students taking Statistics 226(0)

2. Identify the observational units.
   A student of the class (individual)

3. Identify the treatments.
   Whether a clicker was provided for use in class (or not)

4. Identify the response variable.
   The final exam scores of students in a class (one final exam score per student)

5. Does the experiment utilize replication? Answer yes/no and provide a brief justification.
   No because the treatment is not applied independently to two or more experimental units (the class). Because there are only two classes (experimental units), we only have one treatment group at the experimental unit level, i.e. only one, not multiple, classes receive the clickers.

6. Does the experiment utilize blocking? Answer yes/no and provide a brief justification.
   No, and this is a bit tricky. We only have two experimental units, meaning only one experimental unit is given the treatment, and all observational units within the group receive the treatment (clikcer access). Though the decision of which experimental unit receives the treatment was done independently, one block is a single class, and within a single class all observational units (students) receive the same treatment (clickers or no clickers).

7. Does the experiment utilize randomization? Answer yes/no and provide a brief justification.
   Yes, there is randomization. However, there are two types of randomization to consider: Randomization to units within a group, and secondly randomization in the selection of units. We have randomization in the selection of units for the treatment (experimental, each class). However, within a group or block (a single class), we don't have random assignment of whether they receive treatment, as that decision is determined at the group (experimental unit) level. However, as the choice of which class would receive the clickers was randomly determined through a coin toss this experiment is randomized.

3. **[+30]:** The file `guinea_pigs.csv` (available on Canvas) contains data on survival times (in days) of guinea pigs that were randomly assigned either to a control group or to a treatment group that received a dose of tubercle bacilli (Doksum, K. (1974), *Annals of Statistics*, pp 267-77).

1. Use R to compute the following summary statistics for each treatment group:

Figure 1: Yeah, let's pump em' full of Bacilla

```
# knitr::opts_chunk$set(echo = F)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
guinea_pigs <- read.csv("C:/Users/samue/Downloads/guinea_pigs.csv")
# summary(guinea_pigs)
controlData <- guinea_pigs %>%
      filter(Treatment == "Control") %>%
      na.omit()
treatmentData <- guinea_pigs %>%
      filter(Treatment == "Bacilli") %>%
      na.omit()

summary(controlData$Time)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    18.0   141.8   316.5   345.2   570.8   735.0
```

```
IQR(controlData$Time)
```

```
## [1] 429
```

```r
sd(controlData$Time)
```

```
## [1] 222.2139
```

```r
summary(treatmentData$Time)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    76.0   161.0   214.5   242.5   306.0   598.0
```

```r
IQR(treatmentData$Time)
```

```
## [1] 145
```

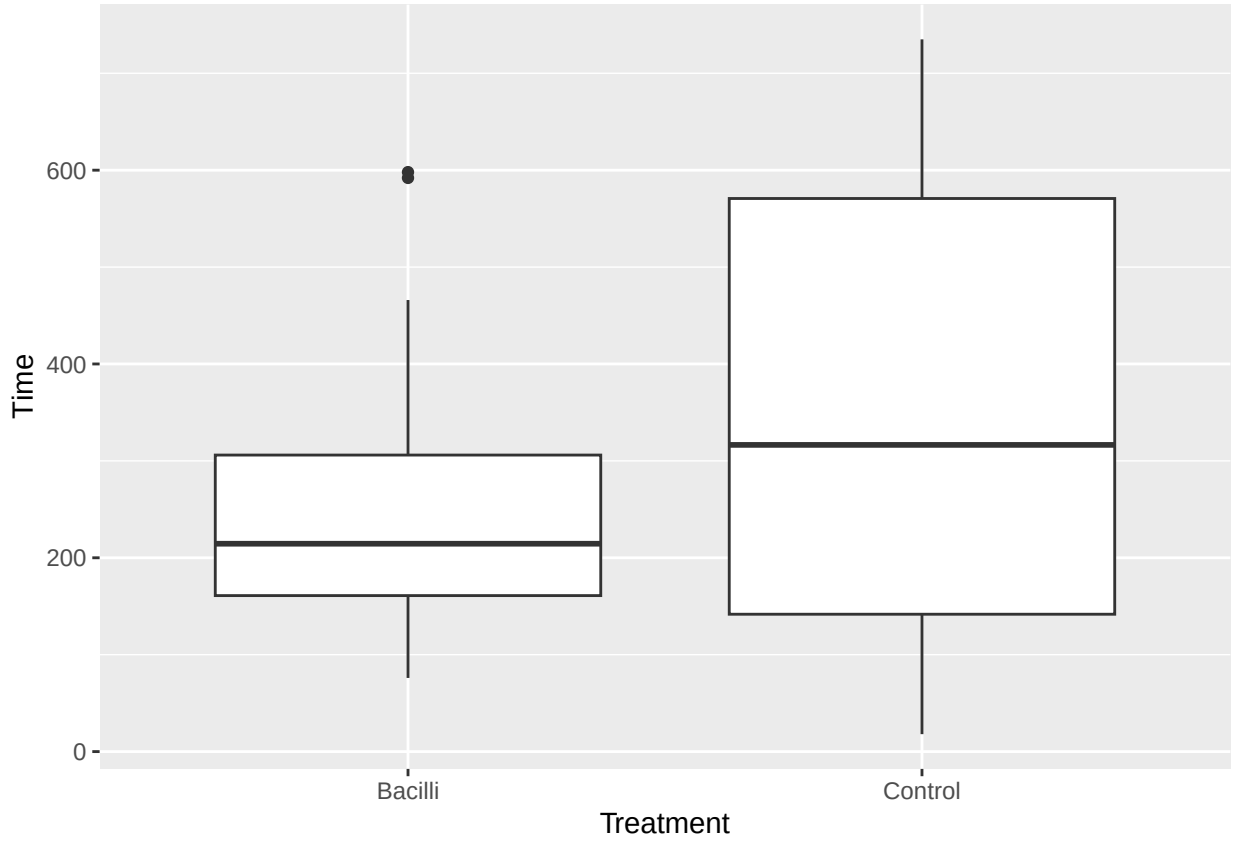
```r
sd(treatmentData$Time)
```

```
## [1] 117.9309
```

Statistic	Control	Bacilli
Median	316.5	214.5
Q1	141.8	161.0
Q3	570.8	306.0
IQR (Q3 - Q1)	429 (570.8 - 141.8)	145 (306 - 161)
Sample Mean	345.2	242.5
Standard Deviation	222.2139	117.9309

2. Use R to construct side-by-side box plots of survival times for the two treatment groups and include it with this assignment.

```r
library(ggplot2)

# Horizontal
# plot <- ggplot(guinea_pigs, aes(Time, Treatment))
# Vertical
plot <- ggplot(guinea_pigs, aes(Treatment, Time))
plot + geom_boxplot()
```



3. Use the box plots from part (b) and the summary statistics from part (a) to describe and compare features of the distributions of survival times for the two treatment groups.
   The mean survival time for guinea pigs treated with Bacilli is lower than the mean surival time of the Control Group. However, there is greater variance (and standard deviation) in the surival times of the Control Group guineas.

4. There is no function built into R to easily perform the randomization test to determine whether the average survival times of the guinea pigs in the two treatment groups is the same or different. Write your own function to perform the test (there is an example in the optional R lab). Interpret the results by providing

   1. the null and alternative hypotheses;
      There are two hypotheses that come to mind, the first being descriptive and the second being specified quantitatively :
      1: H_0 : Treatment of Bacilli on guinea pigs has no effect on survival time, with H_a : Treatment of Bacilli on guinea pigs has an effect on survival times.
      2: $H_0 : \bar{Y_1} = \bar{Y_2}$ with alternative hypothesis $H_a : \bar{Y_1} \neq \bar{Y_2}$ where $\bar{Y_1}$ denotes the mean survival time of pigs who received the treatment (Bacilli) and $\bar{Y_2}$ denotes the mean survival time of pigs who received no treatment (Control).

   2. observed test statistic;

```r
# function assumes only two unique category values
# of interest

tidyDiff <- function(df, categoryCol, meanCol) {
  tidyData <- df %>%
    as.data.frame %>%
    group_by({{categoryCol}}) %>%
    summarize(mean({{meanCol}}))

  mean1 <- tidyData[[1,2]]
  mean2 <- tidyData[[2,2]]

  diffInMeans <- mean2 - mean1
  diffInMeans
}
```

```r
tidyDiff(guinea_pigs, Treatment, Time)
```

```
## [1] 102.6843
```

The average difference between the Treatment and Bacilli group is 102.68 days, meaning guinea pigs who received the Bacilli treatment on average survived for 102.68 days less than their Control (non-treated) counterparts.

3. randomization histogram;

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.1
## v readr     2.1.5
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
guinea_pigs %>%
    group_by(Treatment) %>%
    summarize(meanTime = mean(Time)) %>%
    pivot_longer(Treatment) %>%
    pivot_wider(names_from = value, values_from = meanTime) %>%
    mutate(meanDiff = Control - Bacilli) %>%
    select(meanDiff)
```

```
## # A tibble: 1 x 1
##   meanDiff
##      <dbl>
## 1     103.
```

```
permFunc <- function(df, meanCol) {
  # one <- as.character(expression({{one}}))
  sampleDf <- sample_n(df, nrow(df))

  testDf <- df

  testDf[{{meanCol}}] <- sampleDf[{{meanCol}}]
  return(testDf)
}

head(permFunc(guinea_pigs, "Time"))
```

```
##   Pig Time Treatment
## 1   1  185   Control
## 2   2  107   Control
## 3   3  253   Control
## 4   4  455   Control
## 5   5  270   Control
## 6   6  459   Control
```

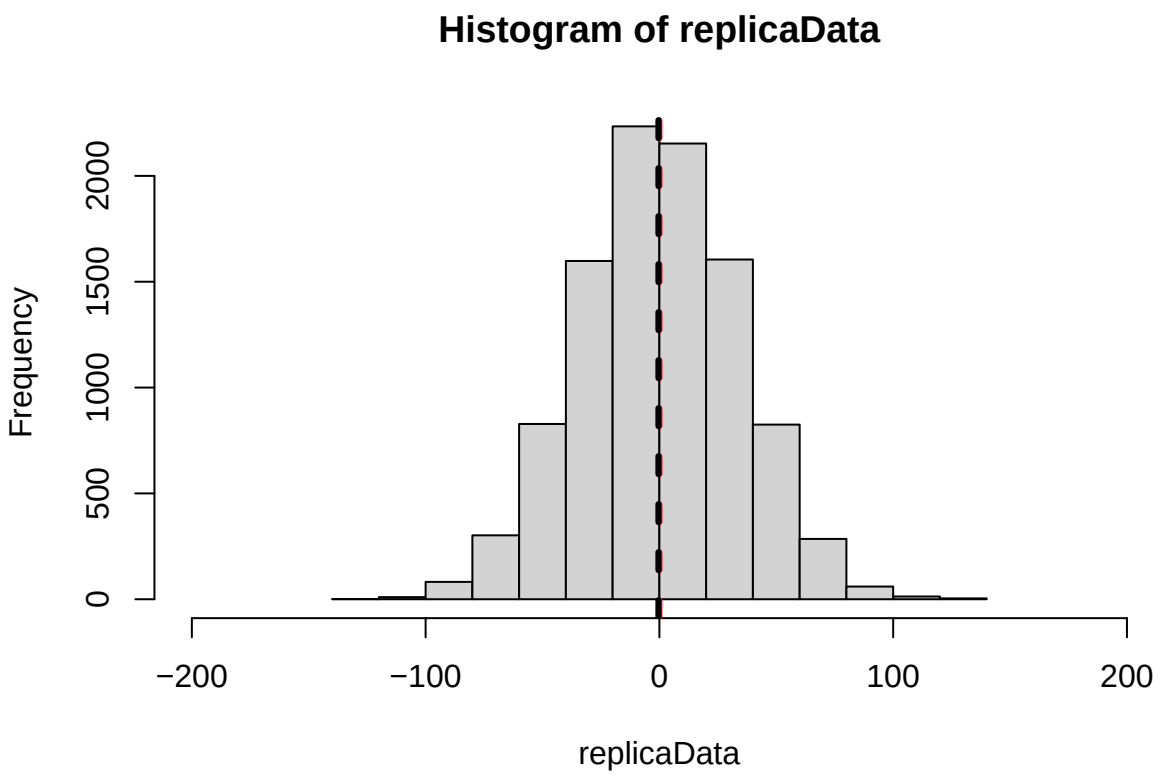
```
head(guinea_pigs)
```

```
##   Pig Time Treatment
## 1   1   18   Control
## 2   2   36   Control
## 3   3   50   Control
## 4   4   52   Control
## 5   5   86   Control
## 6   6   87   Control
```

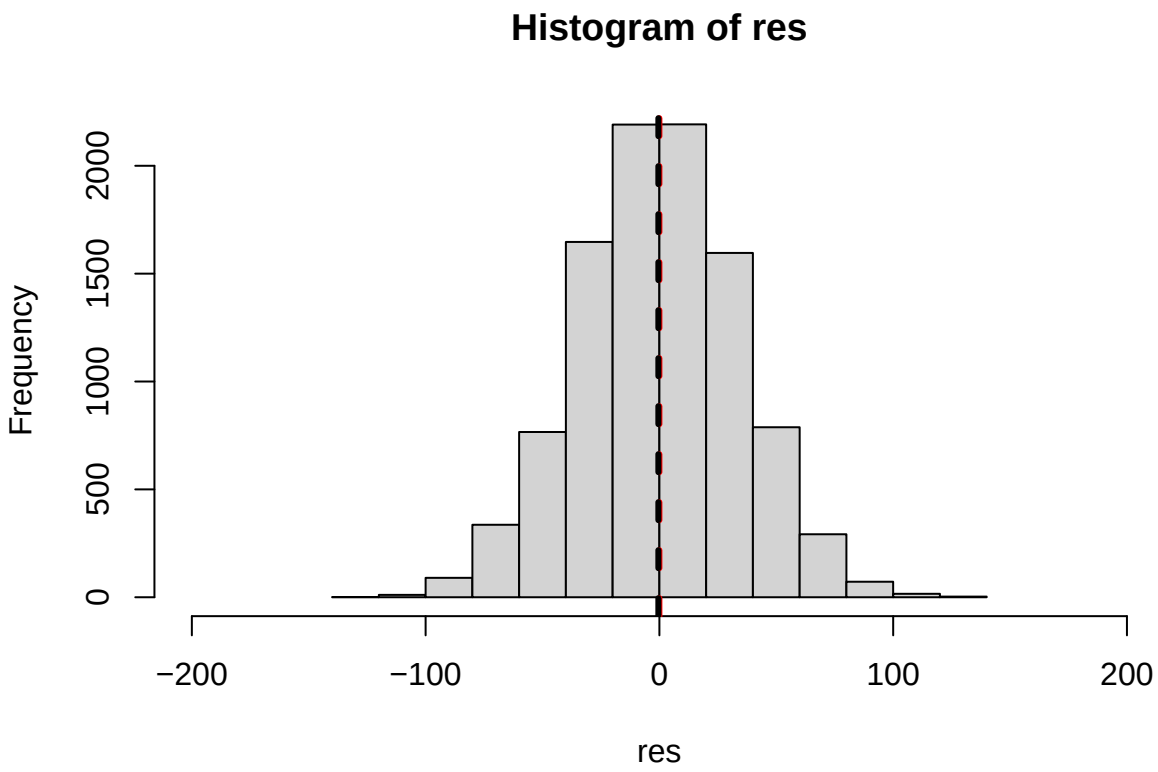
```
compositeFunction <- function(df, meanChar, meanCol, categoryCol) {
  newDf <- permFunc(df, meanChar)

  tidyDiff(newDf, Treatment, Time)
}

compositeFunction(df = guinea_pigs, "Time", Time, Treatment)
```

```
## [1] 19.43373
```

```
replicaData <- do.call(rbind, replicate(10000, compositeFunction(df = guinea_pigs, "Time", Time, Treatm

replicaMean <- mean(replicaData)

hist(replicaData, xlim = c(-200,200))
abline(v = 0, col="red", lwd=3, lty=2)
abline(v = mean(replicaMean), col="black", lwd=3, lty=2)
```




```
set.seed(101) ## for reproducibility
nsim <- 10000
res <- numeric(nsim) ## set aside space for results

for (i in 1:nsim) {
    ## standard approach: scramble response value
    perm <- sample(nrow(guinea_pigs))
    bdat <- transform(guinea_pigs,Time=Time[perm])
    ## compute & store difference in means; store the value
    res[i] <- mean(bdat$Time[bdat$Treatment=="Control"])-
        mean(bdat$Time[bdat$Treatment=="Bacilli"])
}

obs <- mean(guinea_pigs$Time[guinea_pigs$Treatment=="Control"])-
        mean(guinea_pigs$Time[guinea_pigs$Treatment=="Bacilli"])
## append the observed value to the list of results
res <- c(res,obs)

hist(res, xlim = c(-200,200))
abline(v = 0, col="red", lwd=3, lty=2)
abline(v = mean(res), col="black", lwd=3, lty=2)
```



Simulating sampling between the two groups of guinea pigs and averaging the difference in their mean survival times yielded a distribution of differences centered at mean 0.

4. p-value;
   Bane of my existence.

"p-value": Probability, given the above distribution, that we observed a difference of 102.6843?

```
length(replicaData)
```

```
## [1] 10000
```

```
length(which(replicaData >= tidyDiff(guinea_pigs, Treatment, Time)))
```

```
## [1] 15
```

```
length(which(replicaData > tidyDiff(guinea_pigs, Treatment, Time))) / length(replicaData)
```

```
## [1] 0.0015
```

```
length(res)
```

```
## [1] 10001
```

```
length(which(res >= tidyDiff(guinea_pigs, Treatment, Time)))
```

```
## [1] 15
```

```
length(which(res > tidyDiff(guinea_pigs, Treatment, Time))) / length(res)
```

```
## [1] 0.00139986
```

p-value: 0.0014 (for a one-sided p-value)

5. interpretation of the test results (stated in the context ofthe problem).
   The data provided would lend support to the theory that guinea pigs treated with Bacilla would on average have lower survival rates than untreated guinea pigs. Furthermore, with a p-value of 0.0014 for a one-sided test we have further support of this conclusion. We say 0.0014 is the probability that randomization alone leads to a test statistic as extreme as or more extreme than the one observed. We say it is unlikely the null hypothesis (there is no treatment effect) is correct and more likely that the alternative hypothesis, namely that treatment of Bacilla on the guinea pigs, has some effect on their survival times.

Total: 50 points **# correct: %:**