Stat 5000Homework #3
Fall 2024 due Fri, Sep 20th @ 11:59 pm Name: Sam 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: Ben, Sabrina, Sarah, Vanessa, **The Hat Man**

Note to Self

Confidence Intervals

Formula:

$$\text{Confidence Interval} = (\bar{Y_1} - \bar{Y_2}) \pm t_{(df-2, 1-\frac{\alpha}{2})} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Where:

$$S_p = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Calculation

```
# Load Data
library(readr)
smsspeed_1 <- read_csv("C:/Users/samue/OneDrive/Desktop/Iowa_State_PS/STAT 5000/Labs/Lab 3/smsspeed-1.cs
```

```
## Rows: 30 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): AgeGroup
## dbl (3): Age, Own Phone, Control
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
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
```

```r
data1 <- smsspeed_1 %>%
  filter(smsspeed_1$AgeGroup == "Over30")

data2 <- smsspeed_1 %>%
  filter(smsspeed_1$AgeGroup == "Teens")

sampleMean1 <- mean(data1$`Own Phone`)
sampleMean2 <- mean(data2$`Own Phone`)
difference <- sampleMean1 - sampleMean2
difference
```

```
## [1] 44.012
```

```r
#Step 2: Finding standard deviation
s1 <- sd(data1$`Own Phone`)
s2 <- sd(data2$`Own Phone`)

#Step 3: Finding sample size
n1 <- length(data1$`Own Phone`)
n2 <- length(data2$`Own Phone`)

numerator <- (n1-1)*(s1^2) + (n2-1)*(s2^2)
denom <- n1 + n2 - 2
pooled <- sqrt( numerator / denom )
sqrtFactor <- sqrt(1/n1 + 1/n2)

tStatDf <-  2.0484

rightSide <- tStatDf*pooled*sqrtFactor

pooled
```

```
## [1] 18.51069
```

```r
rightSide
```

```
## [1] 13.84544
```

```r
lb <- difference - rightSide
ub <- difference + rightSide

lb
```

```
## [1] 30.16656
```

```r
ub
```

```
## [1] 57.85744
```

Assignment

1.

+21

: A major medical center in the Northeastern U.S. conducted a study looking at blood cholesterol levels and incidence of heart attack. Below are summary statistics of blood cholesterol levels from 16 people who had a heart attack and 20 people who did not have a heart attack.

Group	Sample Size (n)	Sample Mean ($\bar{y}$)	Sample Std. Dev. (s)
Heart Attack (1)	$n_1 = 16$	$\bar{y}_1 = 265.4$	$s_1 = 43.645$
No Heart Attack (0)	$n_2 = 20$	$\bar{y}_2 = 193.1$	$s_2 = 21.623$

1. List the assumptions needed to properly use the t-based confidence interval.

- $Y_{1,1}, Y_{1,2}, ..., Y_{1,16}$ are i.i.d. $N(\mu_1, \sigma^2)$
- $Y_{2,1}, Y_{1,2}, ..., Y_{2,20}$ are i.i.d. $N(\mu_2, \sigma^2)$
- Population variances between the two groups are equal (both have σ^2 variance)
- Independence between observations between groups, i.e. $Y_{1,i}$ and $Y_{2,j}$ are independent for all i and j

2. Using the formula from lecture, compute a 95% confidence interval “by hand" for the population difference in mean cholesterol level between the heart attack and the no heart attack groups.

```
# df <- 16 + 20 - 2
# tStatistic <- 34, 0.975

tStatistic <- function(n1, n2, mu1, mu2, sd1, sd2) {
  numerator1 <- (n1-1)*(sd1^2) + (n2-1)*(sd2^2)
  denom1 <- n1 + n2 - 2
  pooled <- sqrt( numerator1 / denom1 )

  numerator2 <- mu1 - mu2
  denom2 <- pooled * (sqrt ( (1/n1) + (1/n2) ) )
  tStatOutput <- numerator2 / denom2
  tStatOutput

  }

ciEstimate <- function(n1, n2, mu1, mu2, sd1, sd2, tStat) {

  numerator <- (n1-1)*(sd1^2) + (n2-1)*(sd2^2)
  denom <- n1 + n2 - 2

  pooled <- sqrt( numerator / denom )
  sqrtFactor <- sqrt(1/n1 + 1/n2)
  rightSide <- tStat*pooled*sqrtFactor

  difference <- mu1 - mu2

  lb <- difference - rightSide
  ub <- difference + rightSide
  output <- c(min(lb, ub), max(lb, ub))
  output

}
```

```
tStatistic(n1 = 16, n2 = 20, mu1 = 265.4, mu2 = 193.1, sd1 = 43.645, sd2 = 21.623)
```

```
## [1] 6.494355
```

```
ciEstimate(n1 = 16, n2 = 20, mu1 = 265.4, mu2 = 193.1, sd1 = 43.645, sd2 = 21.623, tStat = 6.494355)
```

```
## [1] 4.110884e-06 1.446000e+02
```

The Interval of the difference between groups (average for Heart attack - average for non-Heart attack) is (4.110884e-06, 1.446000e+02) or ($\approx$ 0, 144.6) difference in cholesterol levels.

3. Interpret the confidence interval in the context of the study.

We are 95% confidence the true difference between the two groups (those who had heart attacks compared to those who didn't) is between 4.110884e-06 (close to zero!) to 144.6, meaning we believe those who had a heart attack on average have between 116.31 higher to **near zero difference** blood pressure compared to those who didn't have a heart attack..

4. Suppose that the researchers want to replicate the study, varying the targeted sample sizes in each group to obtain the *best* (or least variable) estimate of the difference in group means, given the constraint that they can only afford to collect information from 50 total participants. How many subjects should they recruit for each group? Fill in the table below to help you answer the question (as an example, one of the solutions is already provided).

$$Var(\bar{Y}_1 - \bar{Y}_2) = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

n_1	n_2	$Var(\bar{Y}_1 - \bar{Y}_2) = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \ \backslash$
1	49	$1.020\sigma^2$
5	45	$0.222\sigma^2$
10	40	$0.125\sigma^2$
20	30	$0.083\sigma^2$
25	25	$0.08\sigma^2$
30	20	$0.083\sigma^2$
40	10	$0.125\sigma^2$
45	5	$0.222\sigma^2$
49	1	$1.020\sigma^2$

They should evenly divide the 50 total participants into two groups, making two groups of 25.

2.

+10

: Refer to the data set `cholesterol.csv` (posted in Canvas). This file contains data on 18 randomly sampled individuals diagnosed with high cholesterol who replaced butter in their diets with a brand (A or B) of margarine. The brand of margarine was randomized. Their doctors recorded their blood cholesterol levels at the beginning of the experiment, after four weeks of their diets, and again after eight weeks. The researchers are interested in exploring the question:

How big is the difference in mean cholesterol reduction after 8 weeks between brand A and B for subjects who replaced butter in their diet with margarine?

1. Compute the 99% confidence interval in R and provide a screenshot of the output.

```
library(readr)
library(dplyr)
cholesterol <- read_csv("cholesterol.csv")
```

```
## Rows: 18 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): Margarine
## dbl (4): ID, Before, After4weeks, After8weeks
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
differenceDf <- cholesterol |>
  mutate(difference = Before - After8weeks)
```

```
t.test(differenceDf$difference ~ differenceDf$Margarine, conf.level = 0.99)
```

```
##
##  Welch Two Sample t-test
##
## data:  differenceDf$difference by differenceDf$Margarine
## t = -2.1529, df = 15.981, p-value = 0.04695
## alternative hypothesis: true difference in means between group A and group B is not equal to 0
## 99 percent confidence interval:
##  -0.38758151  0.05869262
## sample estimates:
## mean in group A mean in group B
##       0.5466667       0.7111111
```

The Confidence Interval is (-0.3876, 0.0587) for the average difference between Group A and Group B where the difference in a particular group is given by “Cholesterol Before” - “Cholesterol After 8 Weeks”.

2. Interpret the confidence interval in the context of the study.

We are 99% confident that the true difference in the change in blood cholesterol levels between those who consumed Brand A vs. Brand B during an 8 week period is between (-0.3876, 0.0587) with units being the difference in change in cholesterol levels; we are 99% confident the true difference between the change in cholesterol for consuming Brand A compared to consuming Brand B is between (-0.3876, 0.0587).

3.

+9

: Suppose that the researchers want to replicate the cholesterol study. Help them achieve their study design goals by performing the following sample size determinations in R (provide screenshot of output) or by hand-calculation (show work):

1. Given an approximate pooled sample standard deviation of $S_p = 0.16$, what sample size is needed in each of two equally-sized treatment groups in order for the standard error of the difference in average cholesterol reduction to be no more than 0.02?

Standard Error

Formula

$n = \frac{2S_P^2}{(s.e.)^2}$

Where s.e. is taken as an input

Calculation

$n = \frac{2S_p^2}{(s.e.)^2}$

Where $S_p = 0.16$ and $s.e. = 0.02$

```
n <- 2*(0.16^2)/(0.02^2)
n
```

```
## [1] 128
```

128 sample size for each treatment group.

2. Given an approximate pooled sample standard deviation of $S_p = 0.16$, what sample size is needed in each of two equally-sized treatment groups in order for the width of a 95% confidence interval for the true difference in mean cholesterol reduction to be no more than 0.04?

Confidence Intervals

Formula

$$n_0 = 8(\frac{z_{1-\frac{\alpha}{2}} S_p}{w})^2$$

$$n = 8(\frac{t_{2(n_0-1),1-\frac{\alpha}{2}} S_p}{w})^2$$

Calculation

$$n_0 = 8(\frac{z_{1-\frac{\alpha}{2}} S_p}{w})^2$$

$$n = 8(\frac{t_{2(n_0-1),1-\frac{\alpha}{2}} S_p}{w})^2$$

Where $w = 0.04$, $S_p = 0.16$

A 95% Confidence Interval gives us:

$\alpha = 0.05 \rightarrow z_{0.975} = 1.96$

$n_0 = 491.7248 \rightarrow t_{492,0.975} \approx 1.984$

```
qt(p = .975, df = 100)
```

```
## [1] 1.983972
```

```
qt(p = .975, df = 492)
```

```
## [1] 1.964797
```

```
n0 <- 8*((0.16*1.96)/(0.04))^2

n <- 8*((0.16*1.965)/(0.04))^2
n0
```

```
## [1] 491.7248
```

```
n
```

```
## [1] 494.2368
```

The sample size needed is 495 participants per group (using $n_0 = 492$).

3. Given an approximate pooled sample standard deviation of $S_p = 0.16$ and an effect size of $\delta = 0.03$, what sample size is needed in each of two equally-sized treatment groups in order for a level $\alpha = 0.05$ two-sided test to have 80% power?

Hypothesis Tests

Formula

$$n_0 = \frac{(z_{1-\frac{\alpha}{2}} + z_{1-\beta})^2(2S_p^2)}{\delta^2}$$

$$n = \frac{(t_{2(n_0-1),1-\frac{\alpha}{2}} + t_{2(n_0-1),1-\beta})^2(2S_p^2)}{\delta^2}$$

Calculation

$$n_0 = \frac{(z_{1-\frac{\alpha}{2}} + z_{1-\beta})^2(2S_p^2)}{\delta^2}$$

$$n = \frac{(t_{2(n_0-1),1-\frac{\alpha}{2}} + t_{2(n_0-1),1-\beta})^2(2S_p^2)}{\delta^2}$$

Where $\delta = 0.03$, $S_p = 0.16$, $\alpha = 0.05$, and $\beta = 0.2 \rightarrow 1 - \beta = 0.8$

```
z1 <- qnorm(p = 0.975)
z2 <- qnorm(p = 0.8)
```

```
numerator <- (z1 + z2)^2 * (2 * 0.16^2)
denom <- 0.03^2
n0 <- numerator/denom
n0
```

```
## [1] 446.514
```

```
t1 <- qt(p = .975, df = 447)
t2 <- qt(p = .80, df = 447)

numerator <- (t1 + t2)^2 * (2 * 0.16^2)
denom <- 0.03^2
n <- numerator/denom
n
```

```
## [1] 448.4689
```

We require a sample size of 449 participants per group.

4.

+10

: Refer to the data set `birthweight.csv` (posted in Canvas) containing the weights of a random sample of babies born at a certain hospital and information about whether the birth mother was a smoker.

1. Using R, find a 90% confidence interval for the difference in average birthweight between babies born to mothers who smoke versus non-smoking mothers, and provide a screenshot of the output.

```
birthweight <- read_csv("birthweight.csv")
```

```
## Rows: 42 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (2): Birthweight, smoker
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
birthweight1 <- birthweight |>
  filter(smoker == 0)
birthweight2 <- birthweight |>
  filter(smoker == 1)
```

90% Confidence $\rightarrow \alpha = .10 \rightarrow 1 - \alpha/2 = 0.95$

Using the following formula, we have:

$$\text{Confidence Interval} = (\bar{Y_1} - \bar{Y_2}) \pm t_{(df-2, 1-\frac{\alpha}{2})} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

We calculate our t statistic as:

```
qt(p = .95, df = 40)
```

```
## [1] 1.683851
```

```
ciEstimate(n1 = nrow(birthweight1),
         n2 = nrow(birthweight2),
         mu1 = mean(birthweight1$Birthweight),
         mu2 = mean(birthweight2$Birthweight),
         sd1 = sd(birthweight1$Birthweight),
         sd2 = sd(birthweight2$Birthweight),
         tStat = 1.66342)
```

```
## [1] 0.07711274 0.67370544
```

Giving us a confidence interval (average non-smoking birthweight minus average smoking birthweight) of (0.077, 0.674).

2. Interpret the confidence interval in the context of the study.

We are 90% confident the true difference between non-smoking birthweights compared to smoking birthweights is between 0.077 to 0.674. This is evidence in support of the observation that smoking mothers will on average have children with lower birthweight compared to non-smoking mothers.

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