Homework 8 : Inference for the Difference Between Two Population Means

NAME: Michael Darfler

NETID: mbd25

DUE DATE: November 18, 2019 by 11:59pm

For this homework, it will be helpful to have a copy of the knitted version of this document to answer the questions as much of it is written using mathematical notation that may be difficult to read when the document is not knitted.

Instructions

For this homework:

1. All calculations must be done within your document in code chunks. Provide all intermediate steps.
2. Include any mathematical formulas you are using for a calculation. Surrounding mathematical expresses by dollar signs makes the math look nicer and lets you use a special syntax (called latex) that allows for Greek letters, fractions, etc. Note that this is not R code and therefore should not be put in a code chunk. You can put these immediately before the code chunk where you actually do the calculation.

Problem 1

Are mean pulse rates different when students are taking a quiz versus when they are sitting in lecture? The *QuizPulse20* data contains the pulse rates for 20 randomly selected students from a large psychology class under two different scenarios: 1) when they were sitting in class taking a quiz and 2) when they were sitting in class during lecture.

μ_x = mean pulse rate for students taking a quiz

μ_y = mean pulse rate for students sitting in lecture

a. Read the `QuizPulse20` data into this homework document.

```
QuizPulse20 <- read.csv("https://raw.githubusercontent.com/mdarfler/BTRY_6010/master/Homework/HW%208/QuizPulse20(3).csv", header = T)
```

b. Add a new column to this data frame containing the differences between the pulse rates for each student (Quiz - Lecture). Name this column “Differences.”

```
QuizPulse20$Differences <- QuizPulse20$Quiz - QuizPulse20$Lecture
head(QuizPulse20)
```

	Student <int>	Quiz <int>	Lecture <int>	Score <int>	Differences <int>
1	1	75	73	80	2
2	2	52	53	82	-1
3	3	52	47	85	5
4	4	92	88	80	4
5	5	56	55	76	1
6	6	84	70	90	14

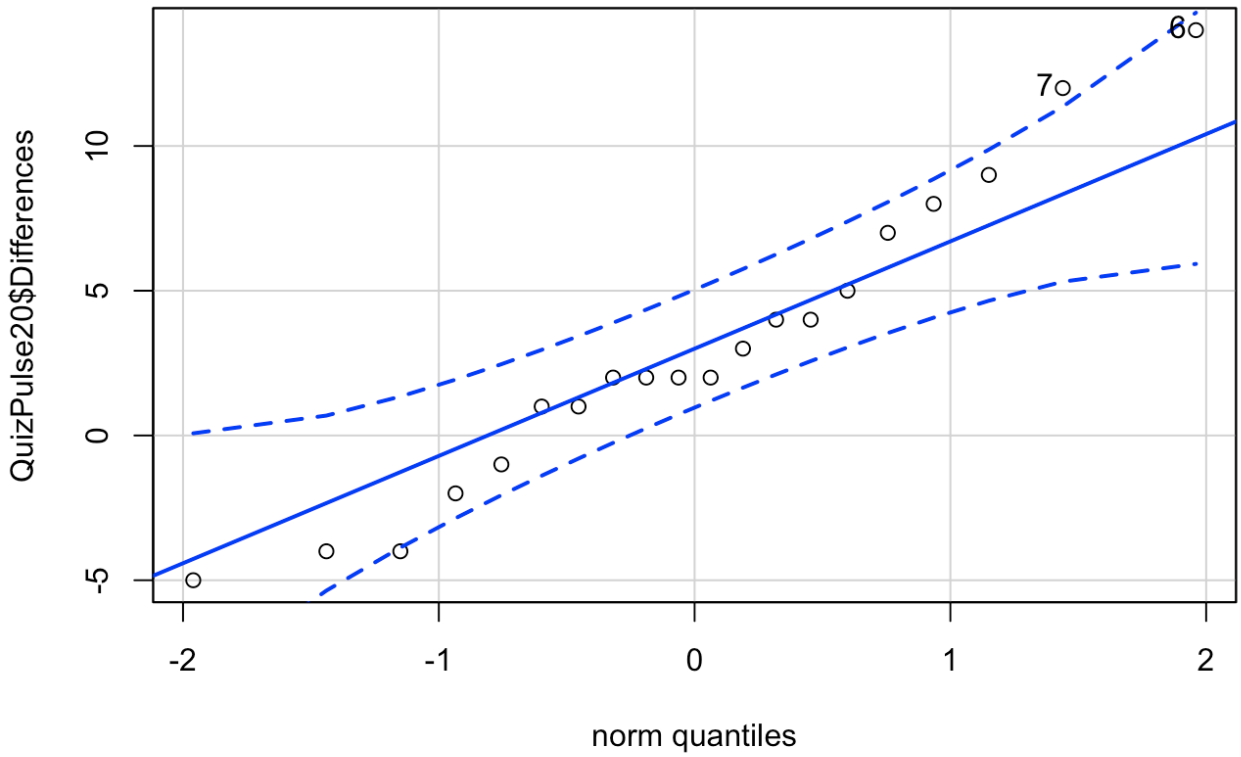
6 rows

c. Since the sample of differences is small, it makes sense to check whether it seems reasonable that the differences are normally distributed. One way to check this is by looking at a Q-Q plot of the differences. The “car” package in R includes the function `qqPlot()`. To use this function, the “car” package will need to be installed. To install this package, from the R Studio menu choose *Tools* > *Install Packages…*. A window will pop up in which you can specify “car” as the package to be installed. Then, just click on “Install”. After this package is installed, run the following code to create the Q-Q plot. Knit your document to look at this plot. Do most of the points lie within the confidence bands? If so, it is reasonable to assume the differences are normally distributed.

```
library(car)
```

```
## Loading required package: carData
```

```
qqPlot(QuizPulse20$Differences)
```




```
## [1] 6 7
```

Yes. it is reasonable to assume that the differences are normally distributed

d. Create a 95% confidence interval for the difference between student mean pulse rate during a quiz and student mean pulse rate during lecture using the `t.test()` function in R. Do this by passing in the differences.

```
results <- t.test(QuizPulse20$Differences)
results
```

```
##
##  One Sample t-test
##
## data:  QuizPulse20$Differences
## t = 2.6153, df = 19, p-value = 0.01702
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.5991368 5.4008632
## sample estimates:
## mean of x
##         3
```

e. Interpret the confidence interval created in (d) in terms of the study.

We are 95% confident that the true population difference μ_d lies between 0.599 and 5.401

f. Complete the following steps to perform the test relevant to this study.

```
i) State the null and alternative hypotheses for this study
```

$H_0 : \mu_d = 0$
$H_A : \mu_d \neq 0$

```
ii) The output from (d) can also be used to perform a paired t-test.  What is the
decision based on the p-value of this test if the significance level is set at 0.
05?  State your answer in the context of this study.
```

Given a significance level of 0.05, i.e. 5% chance of a type-I error, and a p-value of 0.017 we would reject the null hypothesis in favor of the alternative hypothesis $\mu_d \neq 0$.

Problem 2

Is there an association between increased pulse rate and test performance? The variable, `Score`, in the `QuizPulse20` data indicates each student's score on the quiz as a percent. Here we will investigate the relationship between an increase in pulse rate during a quiz and quiz performance. Due to measurement error, it was decided that a student's pulse will only be denoted as having increased during the quiz if it is at least 3 beats per minute higher than his/her pulse rate during lecture.

μ_x = mean score for students whose pulse increased during the quiz

μ_y = mean score for students whose pulse did not increase during the quiz

a. In a code chunk create two vectors. The first vector should contain the scores for students whose pulse increased during the quiz. The second vector should contain scores for students whose pulse did not increase during the quiz (according to the criterion stated in the description of the study).

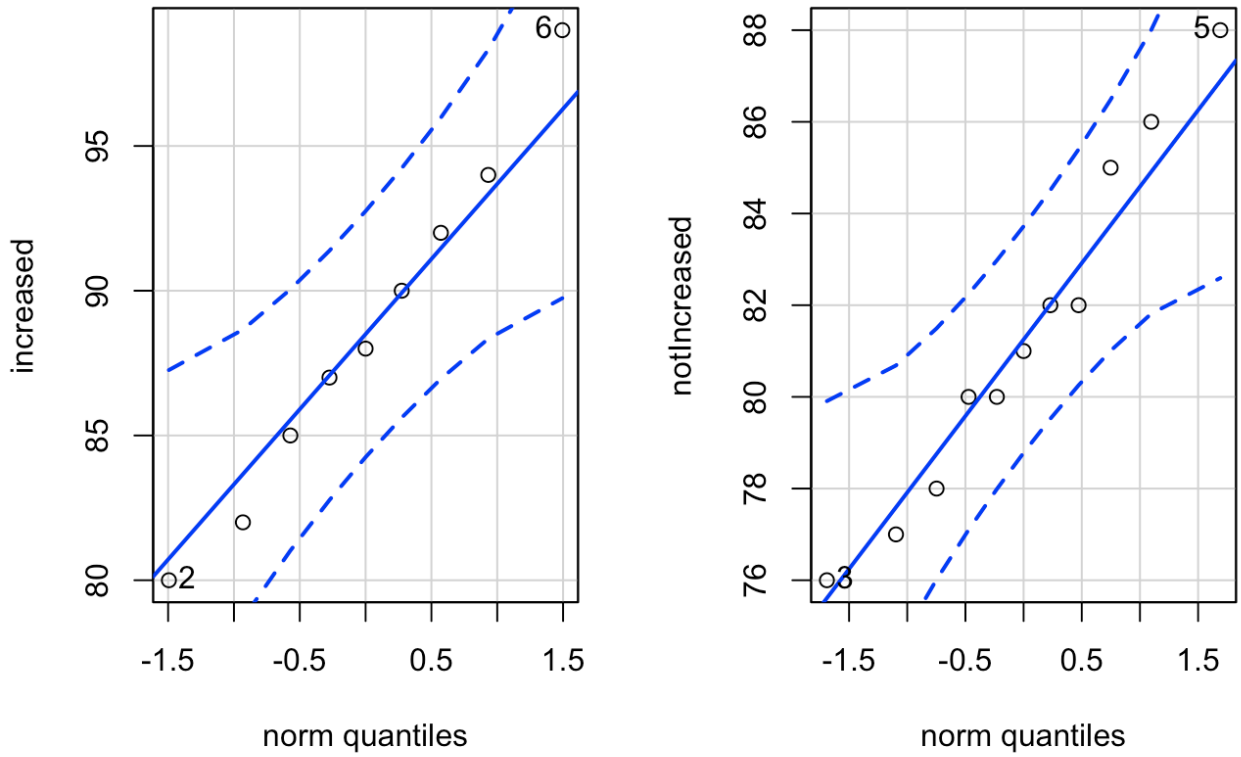
```
increased <- subset(QuizPulse20$Score,QuizPulse20$Differences >= 3)
notIncreased <- subset(QuizPulse20$Score,QuizPulse20$Differences < 3)
```

b. Since the sample size is small for both sets of students, it would be a good idea to look at the Q-Q plots of `Score` for the students whose pulse rate increased during the quiz and also for the sample of students whose pulse did not increase during the quiz. Use the `qqPlot()` function to create the Q-Q plots. Then, give an assessment of whether it seems reasonable to assume the populations these samples are drawn from are normally distributed.

```
par(mfrow=c(1,2))
qqPlot(increased)
```

```
## [1] 6 2
```

```
qqPlot(notIncreased)
```




```
## [1] 5 3
```

The qqPlots show that the data is sufficiently normally distributed.

c. Create a 95% confidence interval for the difference in mean scores for the two groups of students using the `t.test()` function. Pass in the two groups' data as the first two arguments to `t.test`.

```
results <- t.test(increased, notIncreased, paired = F)
results
```

```
##
##  Welch Two Sample t-test
##
## data:  increased and notIncreased
## t = 3.1408, df = 12.994, p-value = 0.007813
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   2.244783 12.139056
## sample estimates:
## mean of x mean of y
##  88.55556  81.36364
```

d. Interpret the confidence interval from (c) in the context of the study.

> We are 95% confident that the true difference between the means, $\mu_{elevated} - \mu_{not}$ is between 2.245 and 12.139.

e. At the 0.05 significance level, is there evidence that the mean score for students whose pulses increase is different from the mean score for students whose pulses do not increase? Base your answer on the confidence interval created in (c).

> At the 5% signifcance level there is evidence that the mean score for students with elevated pulses is different from the mean score for students without elevated pulses. This is because the lower bound of the 95% confidence interval is > 0.

f. Increased pulse rate has been shown to be associated with higher stress levels. Based on the result of this study, would it make sense for the professor of this psychology class to take measures to reduce the stress level of students on days they are taking a quiz in order to increase their scores?

> Based on this data, it would not make sense to reduce stress because our data show that increased heart rate is associated with increased performance.

Power

The power of a test is the probability you will reject the null hypothesis when it is false. In problem 3 we will investigate how the power of a two independent sample t-test is influenced by the following factors:

1. Sample size
2. Effect Size - the absolute difference between the means of the two populations

Problem 3

a. In a code chunk here, define the following variables associated with two independent samples, $X_1, \ldots, X_{16}$ and $Y_1, \ldots, Y_{14}$ where the respective samples are drawn from $X_i \sim N(6, 4)$ and $Y_i \sim N(8, 8)$.

1. `nx` $= n_x$
2. `ny` $= n_y$
3. `mux` $= \mu_x$
4. `muy` $= \mu_y$
5. `sigx` $= \sigma_x$
6. `sigy` $= \sigma_y$

```
nx <-  16
ny <-  14
mux <- 6
muy <- 8
sigx <- 4
sigy <- 8
```

b. Using the variables defined above, simulate two independent samples from the respective normal distributions and perform a t-test for the hypothesis, $H_0 : \mu_x = \mu_y$.

```
x <- rnorm(nx,mux,sigx)
y <- rnorm(ny, muy, sigy)

t.test(x,y)
```

```
##
##  Welch Two Sample t-test
##
## data:  x and y
## t = 0.056144, df = 20.151, p-value = 0.9558
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5.508548  5.813428
## sample estimates:
## mean of x mean of y
##  8.677481  8.525041
```

c. In this part you will perform a simulation to check whether the significance level of the test is in fact 0.05. Start by copying and pasting the code from (a) and (b) into one code chunk here. Steps (i)-(vi) will walk you through writing the simulation.

```r
nx <-  16
ny <-  14
mux <- 6
muy <- 6
sigx <- 4
sigy <- 8

x <- rnorm(nx,mux,sigx)
y <- rnorm(ny, muy, sigy)

t.sampxy <- t.test(x,y)
pvalue <- t.sampxy$p.value

nsim <- 5000
rej <- rep(0,nsim)
signif <- 0.05

for(i in 1:nsim){
  x <- rnorm(nx,mux,sigx)
  y <- rnorm(ny, muy, sigy)
  results <- t.test(x,y)
  pval <- results$p.value

  rej[i] <- pval < signif
}

mean(rej)
```

```
## [1] 0.0552
```

```
i) Remember that the significance level of a test corresponds to the probability
the null hypothesis is rejected when it is true.  So, to estimate this probabilit
y, we should simulate samples under $H_0: \mu_x = \mu_y$.  Let  `muy = 6` so that
the two samples are drawn from populations with the same mean.

ii) Name the output of  `t.test` something like `t.sampxy`.  Save the p-value of
your t-test by including the code `pvalue=t.sampxy$p.value` after the code for th
e t-test.

iii) On the first line in your code chunk create a vector called `rej` that inclu
des 5,000 zeros.

iv) Using a `for` loop, repeat the following process 5,000 times.

1. Simulate two independent samples, one from $N(\mu_x, \sigma_x)$ and the other
from $N(\mu_y, \sigma_y)$ using the parameters already defined in your code chun
k.

2. Perform a t-test for these two samples (using the code in (ii) so that you als
o save the p-value).

3. Set `rej[i] = pvalue <= 0.05` to assign `rej[i]` to 0 in iteration i if the nu
ll hypothesis is not rejected and 1 to `rej[i]` if the null hypothesis is rejecte
d.

v) After the `for` loop created in the last step, estimate the probability that t
he null hypothesis is rejected when it is true by including the following code, `
sum(rej)/5000.`

vi) Does the estimate for the significance level of the test indicate that it is
0.05?  Why or why not.
```

Yes, the estimate for the significance level of the test does indicated that it was 0.05. Specifically, the results of our simulation showed that 94.48% of the time the means were the statistically the same.

d. Now let's estimate the power of the test performed in (b). Remember that the power of a test is the probability that we will reject H_0 when it is false. Repeat the simulation from part (c) where now `mux` and `muy` are set to their original values from part (a). What is the approximate power of this test?

```
nx <-  16
ny <-  14
mux <- 6
muy <- 8
sigx <- 4
sigy <- 8

x <- rnorm(nx,mux,sigx)
y <- rnorm(ny, muy, sigy)

t.sampxy <- t.test(x,y)
pvalue <- t.sampxy$p.value

nsim <- 5000
rej <- rep(0,nsim)
signif <- 0.05

for(i in 1:nsim){
  x <- rnorm(nx,mux,sigx)
  y <- rnorm(ny, muy, sigy)
  results <- t.test(x,y)
  pval <- results$p.value

  rej[i] <- pval < signif
}

power = mean(rej)
```

The approximate power for this test is 12.58%

e. Using the code from part (d), include a code chunk for parts (i), (ii), and (iii), below, that changes the power calculation as specified. Then, for all three parts, make a general statement of how changing the parameter as indicated seems to impact the power calculation.

    i. Repeat the power calculation twice. First, increase `nx` to 25, keeping everything else fixed. For the next power calculation change `nx` to 50, keeping everything else fixed.

```
nx <-  c(25,50)

rej <- rep(0,nsim)
result <- rep(0,length(nx))

for(j in 1:length(nx)){
  sample.nx = nx[j]
  for(i in 1:nsim){
x <- rnorm(sample.nx,mux,sigx)
y <- rnorm(ny, muy, sigy)
results <- t.test(x,y)
pval <- results$p.value
rej[i] <- pval < signif
  }
  result[j] <- mean(rej)
}
result
```

```
## [1] 0.1332 0.1254
```

the increase in the sample size of x modestly increased the power

ii. Repeat the power calculation twice. First, increase `nx` and `ny` to 25, keeping everything else fixed. For the next power calculation, change `nx` and `ny` to 50, keeping everything else fixed.

```
nx <-  c(25,50)
ny <-  c(25,50)

rej <- rep(0,nsim)
result <- rep(0,length(nx))

for(j in 1:length(nx)){
  sample.nx <- nx[j]
  sample.ny <- ny[j]
  for(i in 1:nsim){
    x <- rnorm(sample.nx,mux,sigx)
    y <- rnorm(sample.ny, muy, sigy)
    results <- t.test(x,y)
    pval <- results$p.value
    rej[i] <- pval < signif
  }
  result[j] <- mean(rej)
}
result
```

```
## [1] 0.1860 0.3414
```

> increasing the sample size of x and y significantly improved the power. with nx, ny = 50, the power is more than double what it was start

```
iii) Repeat the power calculation twice.  First, increase the effect size, |`mux` - `muy`| to 5, keeping everything else fixed.  For the next power calculation, increase the effect size to 10, keeping everything else fixed.
```

```
nx <-  50
ny <-  50
mux <- 6
muy <- 11

rej <- rep(0,nsim)
result <- rep(0,length(nx))

for(i in 1:nsim){
  x <- rnorm(sample.nx,mux,sigx)
  y <- rnorm(sample.ny, muy, sigy)
  results <- t.test(x,y)
  pval <- results$p.value
  rej[i] <- pval < signif
}
mean(rej)
```

```
## [1] 0.9754
```

```
muy <- 16
for(i in 1:nsim){
  x <- rnorm(sample.nx,mux,sigx)
  y <- rnorm(sample.ny, muy, sigy)
  results <- t.test(x,y)
  pval <- results$p.value
  rej[i] <- pval < signif
}
mean(rej)
```

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

Increasing the effect size has a very significant affect on the power. At a difference of 5 between means, the power is already over 97%. When the effect size is 10, the power is essentially 100%