Lab 7 (100 points): Two-Sample Independent and Two-Sample Paired t Procedures

Objectives: Confidence interval and significance tests for two samples.

A. (10 points) Online Prelab

The statistical procedures introduced in this course depend on the assumption of normality. Sometimes, if the data are not well approximated by a normal distribution, we apply a transformation of the data. In this lab, you will need to apply the log transformation (natural log, or "ln", which is called "log" in both R and SAS), to make the data more normally distributed. When interpreting transformed data, remember to change the data back before looking at any practicality. This does make the problem more difficult to interpret so you should only transform the data if absolutely necessary. You must determine which question it is appropriate for. You can ask your instructor or TA for more information.

Remember that there are two types of two-sample inference used in this Lab, two-sample independent and two-sample paired. Though the code is similar for these two procedures, there are not identical.

B (45 points) Is the median household income Graduates (MedianIncome) significantly different between the Northeast and North Central regions (Region)? (Data Set: Clean US Data)

A researcher was interested in determining the differences in and influencing factors of the median household income for the individuals living in different areas across the United States. As a preliminary analysis, the researcher asked a basic question: "If I compare two regions of the U.S., will there be a statistically significant difference in the median household income on a county level?" The researcher selected the Northeast and North Central regions, and you will be testing whether the means of the median household income in the two regions are significantly different at the 5% level.

1. (5 points) Code. Please be sure to subset the data to only include the Northeast ("NE") and North Central ("NC") regions for this question. Create a new dataset, for example, USDataSubset, to store the selected data.

Solution:

```
###################################
# Lab 7
# Part B
####################################
library(ggplot2)
# Read in data
# using the interface
# Import Dataset --> From CSV --> browse to the file
#    --> set delimiter to tab --> Change name to USData --> Import
# OR
USData <- read.table("W:/STAT350/USData_cleaned_spring.txt",
                     sep = "\t", header = TRUE)
```

```
# Subset
USData.sub <- subset(USData, USData$Region == "NE" |
        USData$Region == "NC")
#
# BOXPLOT FOR EACH GROUP
#   Note: If you get a warning message for stat_summary, try putting
#   mean in double quotes, "mean"
#
windows()
ggplot(USData.sub, aes(x = Region , y = MedianIncome)) +
  geom_boxplot() +
  stat_boxplot(geom = "errorbar") +
  stat_summary(fun.y = mean, col = "black", geom = "point", size = 3) +
  ggtitle("Boxplots of Median Income Between North Central and
          Northeast")
#
# HISTOGRAM FOR EACH GROUP
#
xbar  <- tapply(USData.sub$MedianIncome, USData.sub$Region, mean)
s     <- tapply(USData.sub$MedianIncome, USData.sub$Region, sd)
USData.sub$normal.density <- ifelse(USData.sub$Region == "NE",
     dnorm(USData.sub$MedianIncome, xbar["NE"], s["NE"]),
     dnorm(USData.sub$MedianIncome, xbar["NC"], s["NC"]))
#
windows()
ggplot(USData.sub, aes(x = MedianIncome)) +
   geom_histogram(aes(y = ..density..), bins = 20,
                  fill = "grey", col = "black") +
  facet_grid(Region ~ .) +
  geom_density(col = "red", lwd = 1) +
  geom_line(aes(y = normal.density), col = "blue", lwd = 1) +
  ggtitle("Histograms of Median Income Between North Central and
           Northeast")
#
# QQPLOT FOR EACH GROUP
#
USData.sub$intercept <- ifelse(USData.sub$Region == "NE",
                               xbar["NE"], xbar["NC"])
USData.sub$slope <- ifelse(USData.sub$Region == "NE",
                           s["NE"], s["NC"])
windows()
ggplot(USData.sub, aes(sample=MedianIncome)) +
  stat_qq() +
  facet_grid(Region ~ .) +
  geom_abline(data = USData.sub, aes(intercept = intercept,
              slope = slope))+
  ggtitle("QQ Plots of Median Income by Region")
#-------------------------------------------------------
# Repeat with logged data
#-------------------------------------------------------
USData.sub$logMedianIncome <- log(USData.sub$MedianIncome)
```

```
#
# BOXPLOT FOR EACH GROUP
#   Note: If you get a warning message for stat_summary, try putting
#   mean in double quotes, "mean"
#
windows()
ggplot(USData.sub, aes(x = Region , y = logMedianIncome)) +
  geom_boxplot() +
  stat_boxplot(geom = "errorbar") +
  stat_summary(fun.y = mean, col = "black", geom = "point", size = 3) +
  ggtitle("Boxplots of LOG Median Income Between North Central and
          Northeast ")
#
# HISTOGRAM FOR EACH GROUP
#
xbar.log <- tapply(USData.sub$logMedianIncome, USData.sub$Region, mean)
s.log    <- tapply(USData.sub$logMedianIncome, USData.sub$Region, sd)
USData.sub$normal.density.log <- ifelse(USData.sub$Region == "NE",
     dnorm(USData.sub$logMedianIncome, xbar.log["NE"], s.log["NE"]),
     dnorm(USData.sub$logMedianIncome, xbar.log["NC"], s.log["NC"]))
#
windows()
ggplot(USData.sub, aes(x = logMedianIncome)) +
   geom_histogram(aes(y = ..density..), bins = 20,
                  fill = "grey", col = "black") +
  facet_grid(Region ~ .) +
  geom_density(col = "red", lwd = 1) +
  geom_line(aes(y=normal.density.log), col = "blue", lwd = 1) +
  ggtitle("Histograms of LOG Median Income Between North Central and
          Northeast")
#
# QQPLOT FOR EACH GROUP
#
USData.sub$intercept.log <- ifelse(USData.sub$Region == "NE",
                                   xbar.log["NE"], xbar.log["NC"])
USData.sub$slope.log <- ifelse(USData.sub$Region == "NE",
                               s.log["NE"], s.log["NC"])
windows()
ggplot(USData.sub, aes(sample=logMedianIncome)) +
  stat_qq() +
  facet_grid(Region ~ .) +
  geom_abline(data=USData.sub, aes(intercept = intercept.log,
              slope = slope.log))+
  ggtitle("QQ Plots of LOG Median Income by Region")
#----------------------------------
# Inference
#----------------------------------
t.test(USData.sub$logMedianIncome ~ USData.sub$Region, mu = 0,
       conf.level = 0.95, alternative = "two.sided",
       paired = FALSE, var.equal = FALSE)
```

2. (5 points) Should you use a two-sample independent or two-sample paired procedure to analyze the data? Please explain your answer by discussing the statistical issues related to the analysis, instead of using the format of the dataset. In real studies, you will need to know what method you will use to analyze the data – paired or independent – to know what data to gather. If this is a paired situation, please state the common characteristic that makes these data paired. Do this part before you do any coding.

Solution:

The two-sample independent procedure should be used. The two different regions of the United States can be considered independent with respect to the question of interest. Of course, they are correlated by similarities that exist throughout the culture and politics of the United States, but we are not interested in isolating those similarities, which should exist among all regions and will not influence our test of the difference of mean median income. We only are concerned with the differences in medium income, and there are no obvious confounding factors that impact the difference.

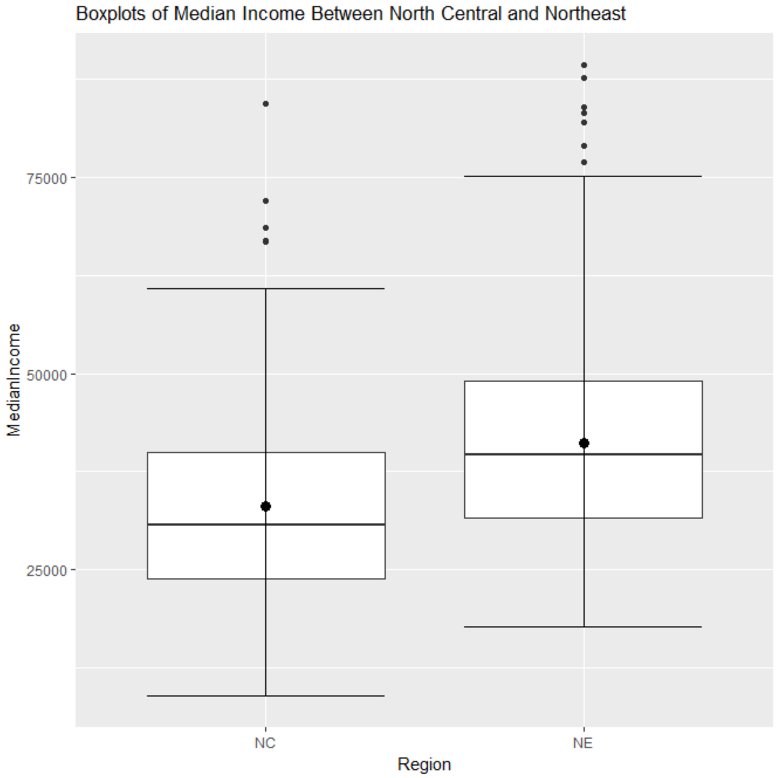
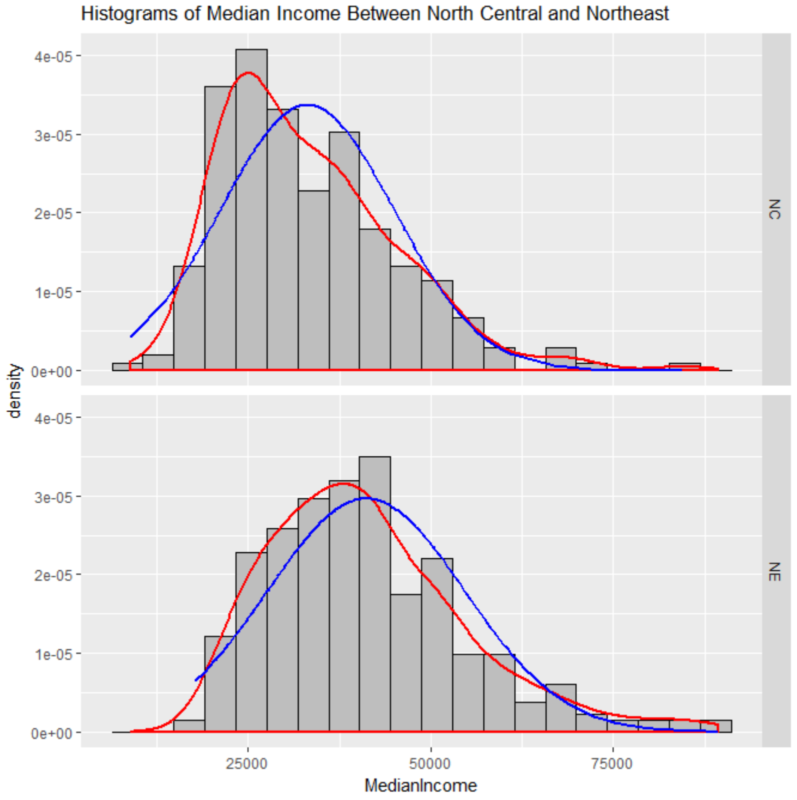
3. (5 points) Should you use a one-sided or two-sided alternative for this analysis? Explain your decision. Do this part before you do any coding.

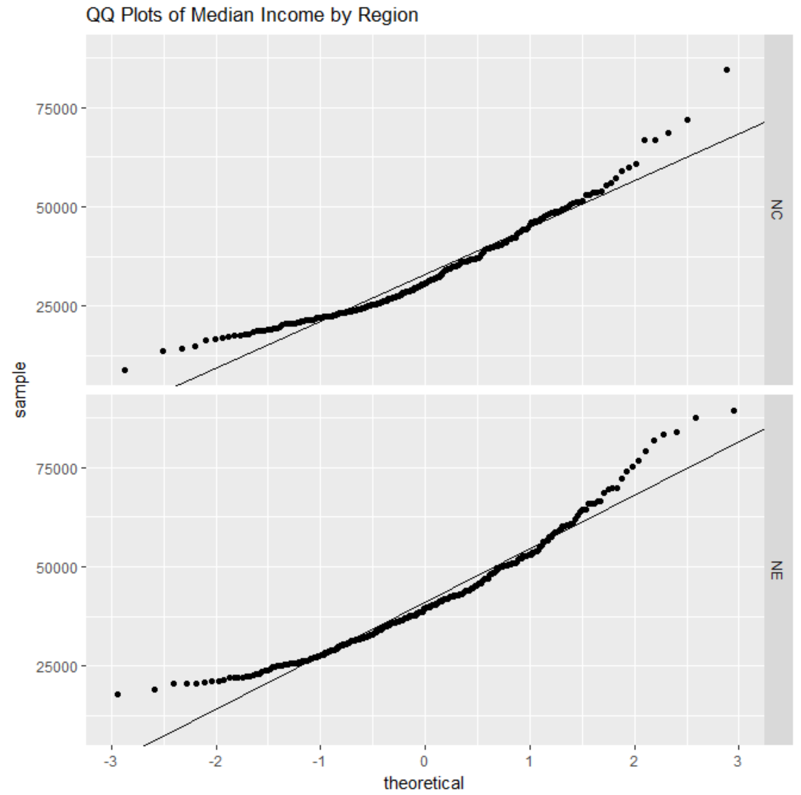
Solution:

Two-sided. The researcher only wanted to know whether a difference exists, and did not have a prior conviction or curiosity about whether one region has a higher mean than the other.

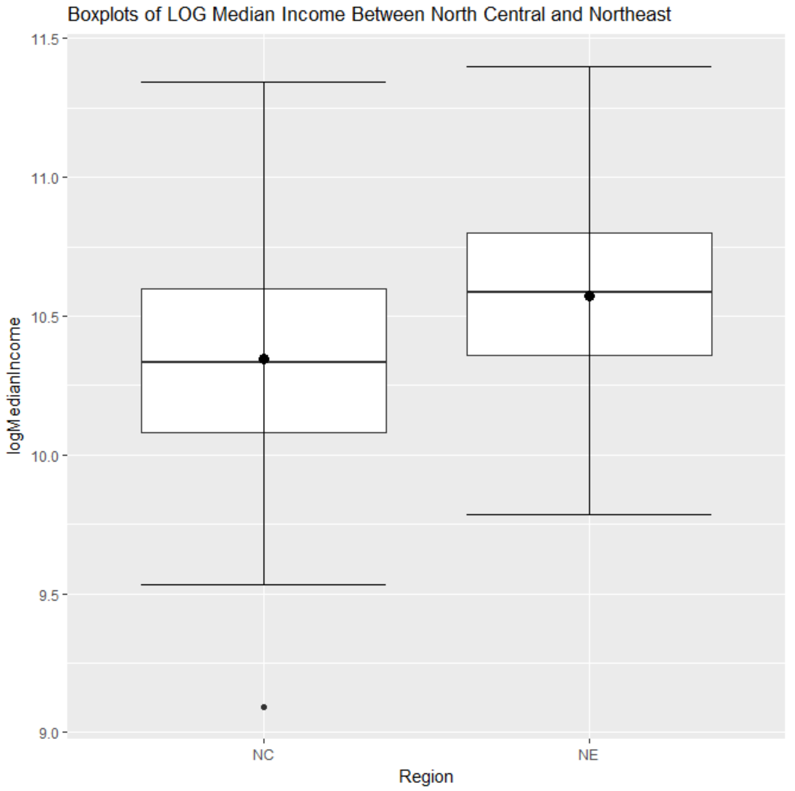
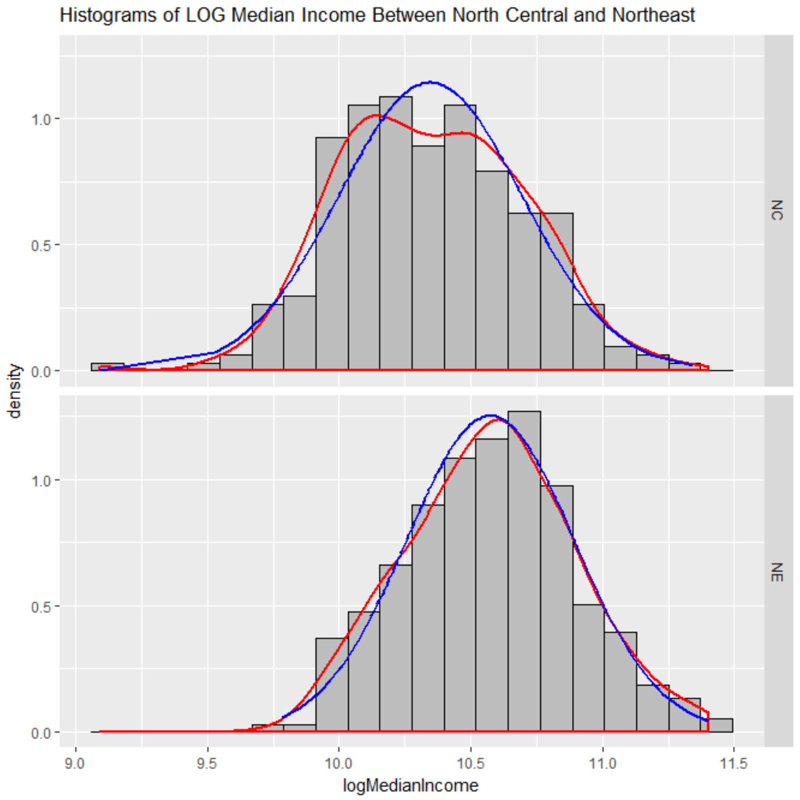
4. (10 points) Create the three diagnostic plots (boxplot, histogram with the estimated kernel density and the estimated normal density curves, and QQ plot) for the appropriate variable(s). Do you think these data are normally distributed? If not, please apply a log transformation, show the diagnostic plots of the transformed data, and comment on the normality of the transformed data. **If you apply a transformation, use the transformation for the remainder of Part B**. Write a short summary of your findings being sure to comment on each graph and providing the answer on whether the appropriate variable is normal or not.

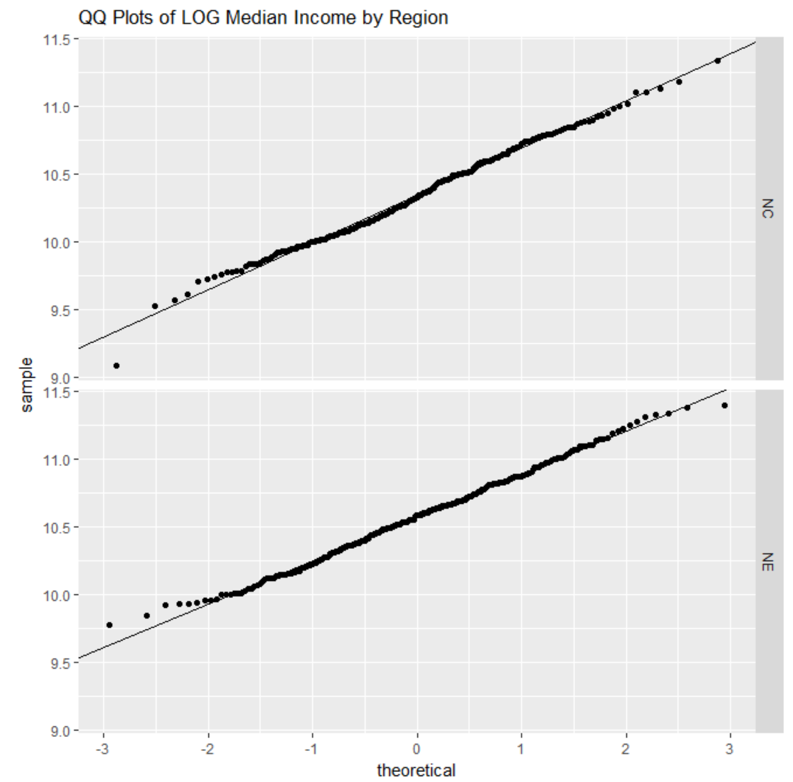
Solution:

Original data	
 Boxplots of Median Income Between North Central and Northeast	The mean is noticeably larger than the median with long tails to the right, suggesting a right skewed distribution.
 Histograms of Median Income Between North Central and Northeast	We see that the distribution is clearly right skewed, and hence is not approximated well by a normal distribution. Also note that the distributions look similar.

	The convex pattern (sometimes called "concave up") indicates that the distribution is right skewed.

The data do not appear normal, so we will consider a log transformation.

Log Data	
 	The distribution appears to have tails of about equal length and the means are close to the medians.
 	The distributions appear to be close to symmetric. The two curves look closer to each other, indicating that the distributions are reasonably approximated by normal distributions.

	In both cases, the data follow the line closely. There are minor deviations in the tails, but not to an extent that causes any concern with our data size.

The original data were skewed and hence did not closely resemble a normal distribution. After performing the log transformation, the data distributions are much closer to normal distributions and can be considered as 'normal' enough with the size of the data set.

5. (5 points) No matter how you answered parts 3/4, determine and interpret the 95% confidence interval of the population parameter of interest (which depends upon whether it is a paired or independent sample situation).

Solution:

This output is for North Central region – Northeast region. Since R alphabetizes the regions, this is the default order.

```
        Welch Two Sample t-test

data:  USData.sub$logMedianIncome by USData.sub$Region
t = -7.9868, df = 508.25, p-value = 9.325e-15
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2841656 -0.1719640
sample estimates:
mean in group NC mean in group NE 
        10.34629         10.57436 
```

The interval is (-0.2841656, -0.1719640).
We are 95% confident that the difference in population mean of the log median income between the North Central region and the Northeast region is covered by the interval from -0.2841656 to -0.1719640.

6. (10 points) No matter how you answered parts 3/4, test the hypothesis that the average median household income is different between the two regions. Assume a 0.05 significance level for the test. Remember to use the full four-step process.

Solution:

```
        Welch Two Sample t-test

data:  USData.sub$logMedianIncome by USData.sub$Region
t = -7.9868, df = 508.25, p-value = 9.325e-15
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2841656 -0.1719640
sample estimates:
mean in group NC mean in group NE 
        10.34629         10.57436 
```

Step 1: Define Parameters

Let μ_{NC} be the true mean of the log medium income for the North Central region and μ_{NE} be the true mean of the log medium income for the Northeast region

Step 2: State Hypotheses

$H_0: \mu_{NC} - \mu_{NE} = 0$
$H_A: \mu_{NC} - \mu_{NE} \neq 0$

Step 3: Report Test Statistic, Degrees of Freedom, and P-value

$t_{ts} = -7.9868$
Degrees of freedom: 508.25
$p = 9.325 \times 10^{-15}$

Step 4: State Conclusions

Since $9.325 \times 10^{-15} \leq 0.05$, we reject the null hypothesis.
The data show strong support (p = 9.325×10^{-15}) to the claim that the population mean difference in log medium income is different between the Northeast and North Central regions.

7. (5 points) Are the conclusions from parts 5 and 6 consistent? Please explain your answer.

Solution:

Yes, they are consistent. In 5, the confidence interval did not contain 0. In 6, we rejected the null hypothesis that the difference in means is 0.

C (45 points) **Is the average amount of money spent on each pupil in period 1 (EducationSpending) different from that in period 2 (EducationSpendingP2)? (Data Set: Clean US Data)** A psychology professor studying relationships wants to know whether the average amount spent on each pupil in period 2 is at least 70 dollars greater than that of period 1 (that is, larger by $70 or more) on average in the United States.

1. (5 points) Code.

Solution:

```
###################################
# Part C
####################################
USData$diff <- USData$EducationSpendingP2 - USData$EducationSpending
# Boxplot
#   Note: If you get a warning message for stat_summary, try putting
#   mean in double quotes, "mean"
windows()
ggplot(USData, aes(x = "", y = diff)) +
  stat_boxplot(geom = "errorbar") +
  geom_boxplot() +
  ggtitle("Boxplot of Difference in Education Spending: P2 - P1") +
  stat_summary(fun.y = mean, col = "black", geom = "point", size = 3)
# Histogram
xbar <- mean(USData$diff)
s <- sd(USData$diff)
windows()
ggplot(USData, aes(diff)) +
  geom_histogram(aes(y = ..density..), bins = 20,
                 fill = "grey", col = "black") +
  geom_density(col = "red", lwd = 1) +
  stat_function(fun = dnorm,  args = list(mean = xbar, sd = s),
                col = "blue", lwd = 1) +
  ggtitle("Histogram of Difference in Education Spending: P2 - P1") +
  xlab("Difference P2 - P1") +
  ylab("Frequency")
# QQ Plot
windows()
ggplot(USData, aes(sample = diff)) +
  stat_qq() +
  geom_abline(slope = s, intercept = xbar) +
  ggtitle("QQ Plot of Difference in Education Spending: P2 - P1")
# Inference
t.test(USData$EducationSpendingP2, USData$EducationSpending, mu = 0,
      paired = TRUE, alternative = "two.sided", conf.level = 0.95)
```

2. (5 points) Should you use a two-sample independent or two-sample paired procedure to analyze the data? Please explain your answer by discussing the statistical issues related to the analysis, instead of using the format of the dataset. In real studies, you will need to know what method you will use to analyze the data – paired or independent – to know what data to gather. If this is a paired situation, please state the common characteristic that makes these data paired. Do this part before you do any coding.

Solution:

We should use a paired procedure. We want to isolate the differences between period 2 and period 1. However, education spending may differ for many different reasons, influenced by lurking variables such as “economics.” So, to remove the confounding lurking variables, we must take differences within each county, leading to the paired procedure.

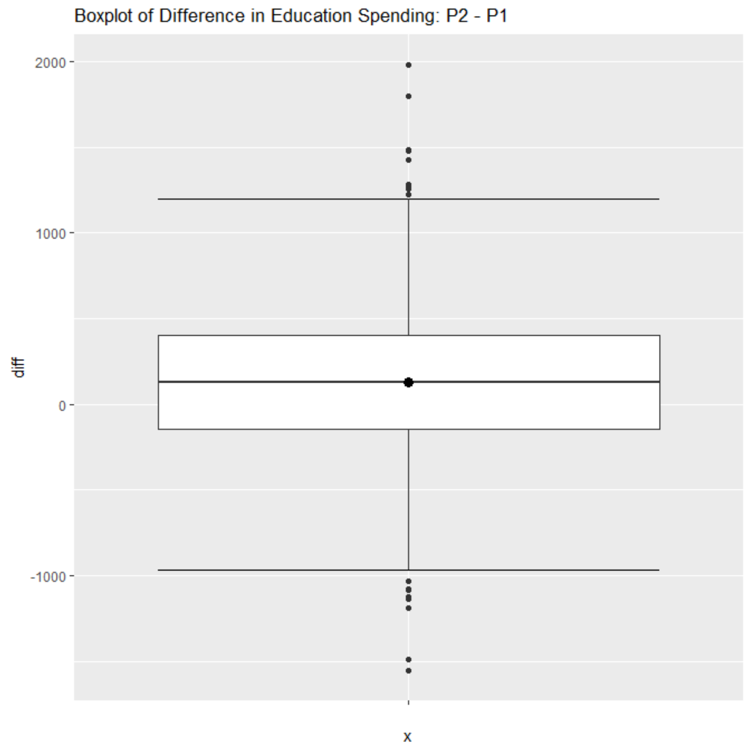
3. (5 points) Should you use a one-sided or two-sided alternative for this analysis? Explain your decision. Do this part before you do any coding.

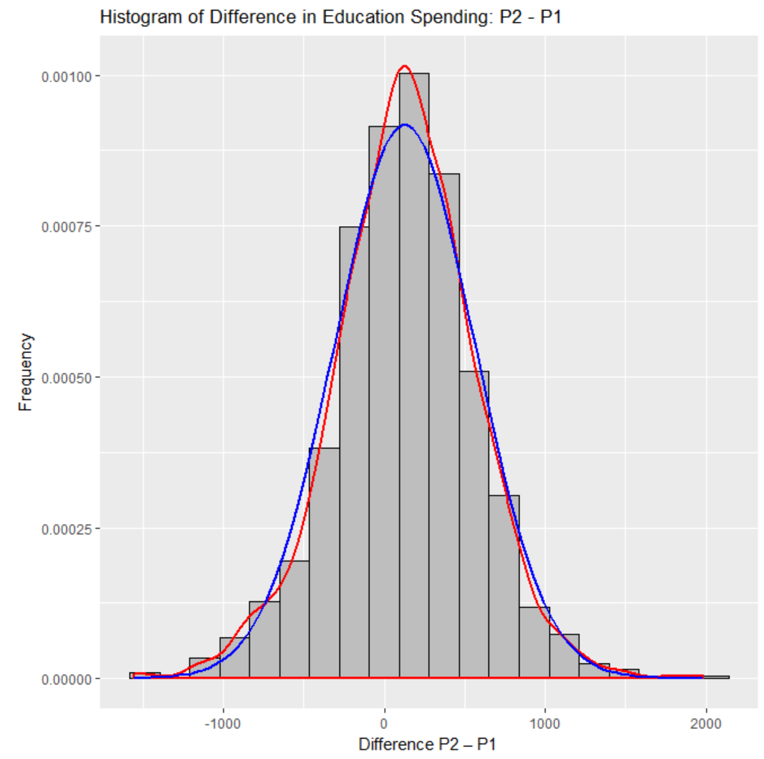
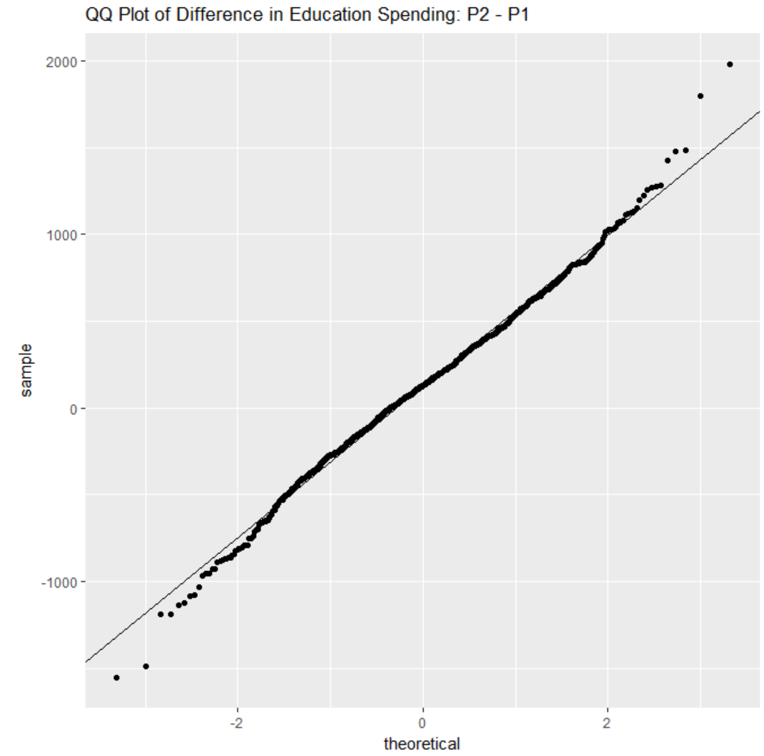
Solution:

A one-sided alternative should be used. The researcher has a specific goal of determining if the education spending in period 2 *exceeds* that of period 1 by *at least* $70.

4. (10 points) Create the three diagnostic plots (boxplot, histogram with the estimated kernel density and the estimated normal density curves, and QQ plot) for the appropriate variable(s). Do you think these data are normally distributed? If not, please apply a log transformation, show the diagnostic plots of the transformed data, and comment on the normality of the transformed data. **If you apply a transformation, use the transformation for the remainder of Part C**. Write a short summary of your findings being sure to comment on each graph and providing the answer on whether the appropriate variable is normal or not.

Solution:

	The distribution is symmetric because the tails look about the same length and the mean and median are approximately in the same location. Even though there are outliers explicitly shown on the boxplot, they do not look like outliers because they are not that far from the rest of the data.

	The distribution is symmetric, but the estimated red curve peaks higher and dies off sooner than the blue line. This suggests deviations in the tails.
	The points closely follow the line except for at both tails.

The distribution is symmetric and unimodal with no significant outliers. While there are deviations in the tails, it will pose no problem for an assumption of normality at a sample size of 1098.

5. (5 points) No matter how you answered parts 3/4, determine and interpret the 95% confidence interval of the population parameter of interest (which depends upon whether it is a paired or independent sample situation).

Solution:

```
	Paired t-test

data:  USData$EducationSpendingP2 and USData$EducationSpending
t = 9.5986, df = 1097, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 100.2113 151.7082
sample estimates:
mean of the differences 
               125.9598
```

The interval is (100.2113, 151.7082).
We are 95% confident that the difference between the population average Education Spending in period 2 and that of period 1 is covered by the interval from 100.2113 to 151.7082.

6. (10 points) No matter how you answered parts 3/4, test the hypothesis that the average amount spent on each pupil is different between the two periods. Assume a 0.05 significance level for the test. Remember to use the full four-step process.

Solution:

```
	Paired t-test

data:  USData$EducationSpendingP2 and USData$EducationSpending
t = 9.5986, df = 1097, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 100.2113 151.7082
sample estimates:
mean of the differences 
               125.9598
```

Step 1: Define Terms
Let μ_D be the difference of the population means of EducationSpendingP2 and Education Spending.

Step 2: State Hypotheses
$H_0: \mu_D = 0$
$H_A: \mu_D \neq 0$

Step 3: Test Statistic, Degrees of Freedom, P-value
$t_{ts} = 9.5986$
Degrees of Freedom: 1097
$p < 2.2\text{e} - 16$

Step 4: State Conclusion

Since $2.2 \times 10^{-16} \leq 0.05$, we reject the null.

The data provide strong support (p < 2.2× 10^{-16}) to the claim that the population mean of education spending in period 2 is different from the population mean in period.

7. (5 points) In one English sentence, explain whether there is evidence that the average amount spent on each pupil in Period 2 is at least 70 dollars greater than that of Period 1. If so, explain whether the amount is greater than $70 by a magnitude of practical significance. If you needed to transform the data, keep this in mind when interpreting the result.

Solution:

Note that because of a problem in the assignment, you can not completely answer this question with the data that is provided. Therefore, we will be lenient on the interpretation.

The confidence interval does not include $70 in it. In fact the lowest part of the interval is more than $100. I would think that an extra $30 per student is practically significant.