Shreya Roy
Jacob Rulli
Rthvik Raviprakash
(Chakraborty 1:30)

A) Introduction and Questions:

Inference 1: One-Sample T-Test on GDP
A student claims that a state has a high GDP if their GDP is greater than 59,927 (GDP per capita for the US, according to the World Bank). Would the average GDP of the states in the top universities list be considered a high GDP?

Inference 2: Two-Sample T-Test on Tuition and University Type
What is the relationship between tuition and the type of school? Would the average tuition at a private university be $10,000 higher (Wignall, College Raptor) than at a public university?

Inference 3: ANOVA Test on Region and Enrollment
Does the average number of enrolled students across different regions differ?

Inference 4 (COVID-19): Linear Regression Analysis on Population Size and Number of Cases
Does the population size of a state significantly affect the number of COVID-19 cases?

In this report, we looked at a dataset that contains information about the top 311 universities in the United States. We used this data to answer three different questions that looked at each university's state GDP, tuition, type, region, and enrollment. The first question, which looks at whether the average GDP would be considered a high GDP, is an important question because it identifies if a state's GDP is related to the fact that one or more of the nation's top universities is in that state. The second, which compares the average tuition for private and public universities, would help students and their families make a decision about which type of university they are willing to pay for. Finally, the last question looks at the difference in the average number of enrolled students across regions, which could help students determine the region in which they want to study based on how populated the schools/areas in those regions are.

The final question we answered related to the current COVID-19 virus pandemic, and the dataset was based on data acquired on March 30th, 2020. We focused on the population size of each state in the US and whether it affects the number of COVID-19 cases per state. Answering this question would prove that the virus does spread faster and wider in more populated areas, and may even justify the need for a quarantine in those areas.

"GDP per Capita (Current US$) - United States." *The World Bank*, 2019, data.worldbank.org/indicator/NY.GDP.PCAP.CD?end=2018&locations=US&start=1960.

Wignall, Allison. "The Average Cost of College in the US for 2017-2018." *College Raptor*, 20 Dec. 2019, www.collegeraptor.com/find-colleges/articles/affordability-college-cost/average-college-costs-2017-2018/.

B) Data:

Variable Name	Type of Variable	Description
GDP	Numerical	GDP of the state the university is located in evaluated in 2017
Tuition	Numerical	Total tuition for one academic year in U.S. dollars
Type	Categorical	Whether the University is Private or Public
Enrollment	Numerical	Undergraduate enrollment
Region	Categorical	Region of the location of each university
Population	Numerical	The population of each state according to the 2000 census.
Cases	Numerical	The total number of reported cases of COVID-19 from the state that the University is in from data from March 31, 2020.

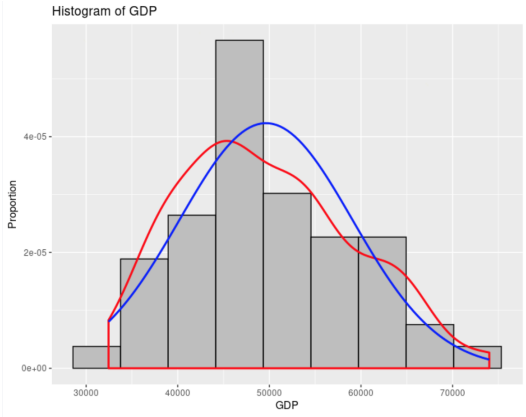
Cleaning the Data

```
> univ_clean <- univsp20[complete.cases(univsp20),]
> View(univ_clean)
> write.table(univ_clean, file="univ_clean.txt", row.names=FALSE, sep="\t")
```

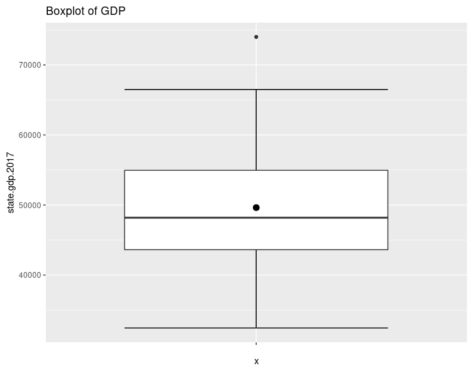
C) Inference 1:

A student claims that a state has a high GDP if their GDP is greater than 59,927 (GDP per capita for the US, according to the World Bank). He wants to know if the average GDP of the states in the top universities list would be considered a high GDP.

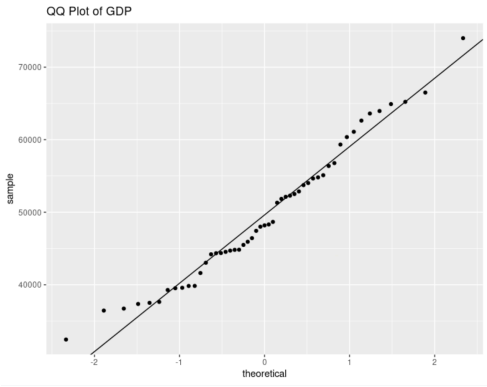
The statistical procedure that will be used here is a one-sample t procedure hypothesis test and a confidence interval. The hypothesis test will be one sided due to the fact the alternative hypothesis is looking at whether the average is larger than a certain value. The confidence interval will also only look at the lower bound of the interval due to the hypothesis test being upper tailed.



The histogram indicates approximate normality due to the symmetry and by comparing the blue and red curves. There are no obvious outliers.



The boxplot indicates that the distribution is approximately symmetric. This is due to the fact the whiskers are approximately the same length. There appears to be one outlier of a high value.



The points on the probability plot somewhat closely follow a straight line with no pattern. The points are randomly above and below the line. This indicates that the distribution is approximately normal.

Assumptions:

- The data is a random sample
- The sample has a normal distribution

The data is assumed to be an SRS so that assumption is confirmed. The sample is proven to have a normal distribution through the use of the histogram, box plot, and especially the normal probability plot as seen above. This means all assumptions are met and the t test can be used.

T-Test

```
        One Sample t-test

data:  states_new$state.gdp.2017
t = -7.8023, df = 50, p-value = 1
alternative hypothesis: true mean is greater than 59927
95 percent confidence interval:
 47417.4     Inf
sample estimates:
mean of x
 49629.29
```

Hypothesis Test:

Step 1 - Terms:

μ is the population mean average GDP of the states in the top universities list

Step 2 - Hypothesis:

H_0: $\mu = 59{,}927$ H_a: $\mu > 59{,}927$

Step 3 - Test statistic, df, p-value:

Test statistic = -7.8023

Df = 50

P-value = 1

Step 4 - Conclusion:

Since $1 > 0.05$, we should not reject H_0

The data does not provide evidence ($p = 1$) to the claim that the population mean average GDP of the states in the top universities is higher than the GDP per capita for the US (59,927).

Confidence Interval:

Lower interval bound = 47417.4

This means we are 95% certain that the true mean of the average GDP of the states in the top universities is above 47417.4. Due to the fact there is a large amount of possible true mean values between this value (47417.4) and the GDP per capita for the US (59,927), this is not evidence that the average GDP of the states in the top universities list is a high GDP.

Conclusion:

If you assume that a state has a high GDP if their GDP is greater than 59,927, which is the GDP per capita for the US, you can use this to determine whether US states are above or below this number. This assumption was used to determine whether or not the average GDP of the states in the top universities list would be considered a high GDP. This was done using a one-sample t procedure hypothesis test and a confidence interval. The hypothesis test gave an output of p-value 1, which is much higher than the significance level used, 0.05, meaning that the alternative hypothesis of the average GDP of the states in the top universities list being considered a high GDP was not proven accurate. The confidence interval also showed a lower bound of 47417.4, which is much lower than the GDP per capita for the US (59,927), meaning this is not evidence that the average GDP of the states in the top universities list is a high GDP. With the results of the hypothesis test and confidence interval, we can safely say one cannot assume the average GDP of the states in the top universities list is necessarily a high GDP.

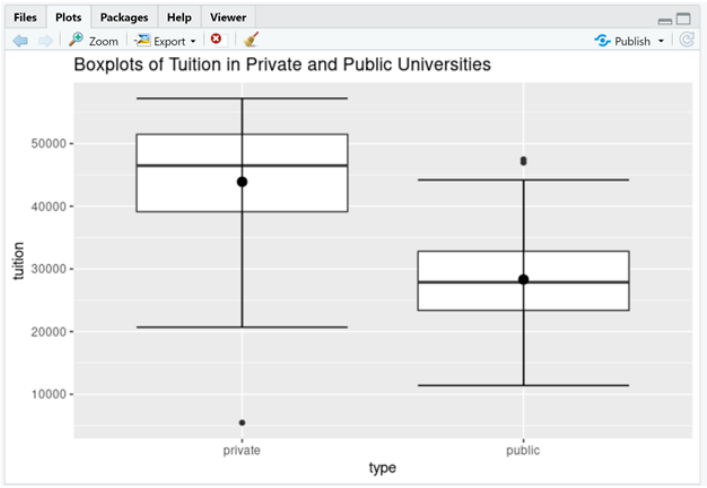
D) Inference 2:

According to the College Raptor Blog the average cost for Public four-year is $36420 and the average cost for a private four year is $46950. But here we wanted to check if the average tuition for the private universities in the dataset is greater than the average tuition for the public universities by $10000.

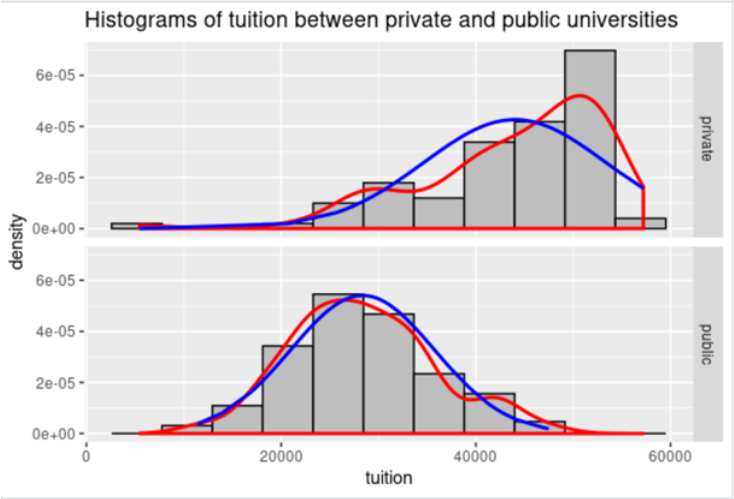
We had to satisfy the following assumptions before using the two-sample t test,

SRS: This assumption is made based on the collection of the data. Here, the data were independently collected at random so this assumption is true.

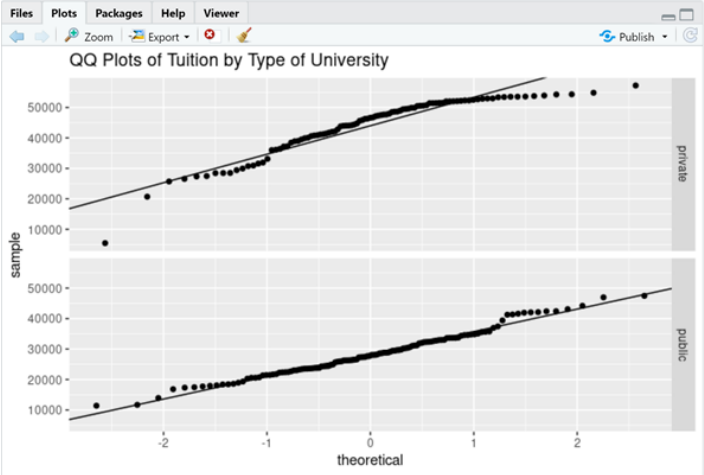
Normality: Satisfy the assumptions of normality for tuition with private and public universities.



-> Here the box plots look normal and there are only 2 possible outliers.



-> The Histograms are bimodal and they are approximately symmetric.



-> The points on the QQ-plot follow the line except the ends are just slightly deviated.

All three of these plots seem to show that the tuition for private and public universities are normal. So, the assumption of normality is satisfied.

As we can see there are no outliers here so whatever outliers that were in the dataset was removed during cleaning.

Output for the 2-sample t-test

```
        Welch Two Sample t-test

data:  univsp20_cleaned$tuition by univsp20_cleaned$type
t = 4.8769, df = 179.17, p-value = 1.184e-06
alternative hypothesis: true difference in means is greater than 10000
95 percent confidence interval:
 13733.26      Inf
sample estimates:
mean in group private  mean in group public
             44009.97              28361.85
```

The Hypothesis Test

Step 1: Define the Parameters

Let μ_{pr} be the true mean of the tuition in private universities

Let μ_{pu} be the true mean of the tuition in public universities

Step 2: State the Hypothesis

H_0: $\mu_{pr} - \mu_{pu} = 10000$

H_a: $\mu_{pr} - \mu_{pu} > 10000$

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

Tts = 4.8769, df = 179.17, p-value = 1.184e-06

Step 4: State the conclusion

Since, 1.184e-06 < 0.05 we reject the null hypothesis. The data shows strong support to the claim that the population mean difference is greater than 10000.

Conclusion:

We can assert with a 95% confidence level that the true mean difference for the tuition of private universities to public universities is higher than the $10000 figure as per the information provided in collegeraptor.com.Private universities tend to have higher tuition costs compared to public universities. Private universities have a cost which is as much as $10000 greater when compared to the average cost of going to a public university as posted on collegeraptor.com. So, it means that households should become more skeptical of their source when figuring out how much they would need to afford college tuition rates.

E) Inference 3:

Does the number of enrolled students for each university differ between regions?

We are trying to determine if there is a significant difference between the true mean values of 4 regions with a single factor, so we can use the ANOVA test.

Assumptions:

1. The four regions are independent, random samples (SSR).
   The data is independently collected at random, so we can assume SRS.

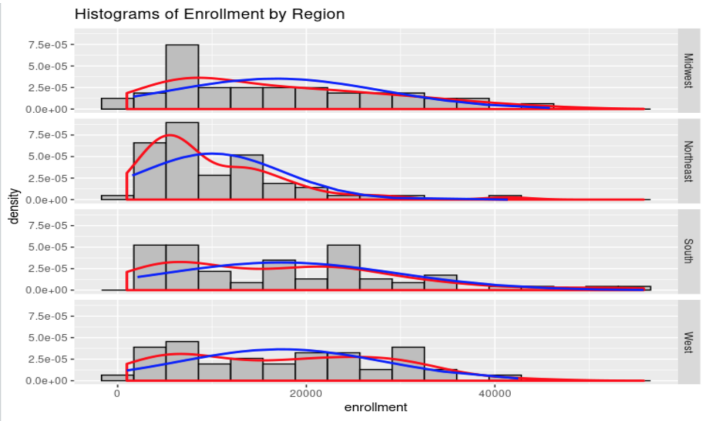
2. Each region has a normal distribution.
   The graphs are bimodal and approximately symmetric.

3. The populations have a constant variance.
   sd_{max}/sd_{min} = 12413.004/7464.519 = 1.663 < 2. The variances are all close to each other.

Region	Midwest	Northeast	South	West
n	47	62	67	45
xbar	16774.74	10056.61	17280.96	17356.69
sd	11299.498	7464.519	12413.004	10920.100

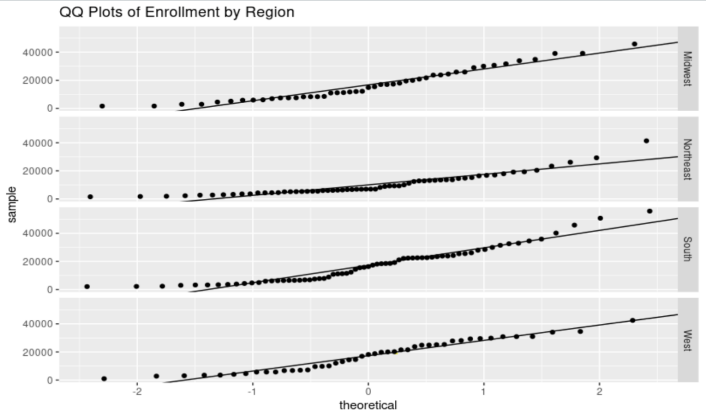
Normality of the data can be proved by looking at the histograms.

The graphs are bimodal and approximately normal.

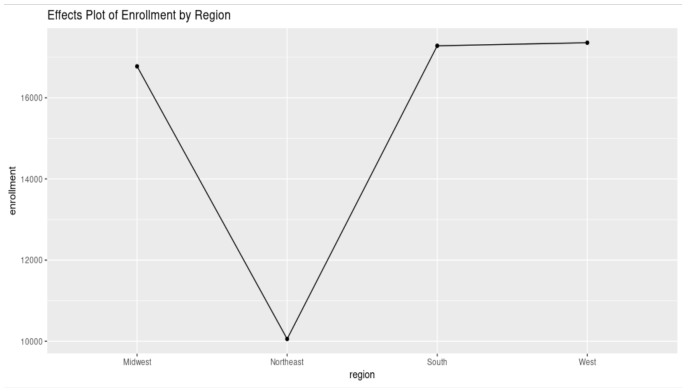


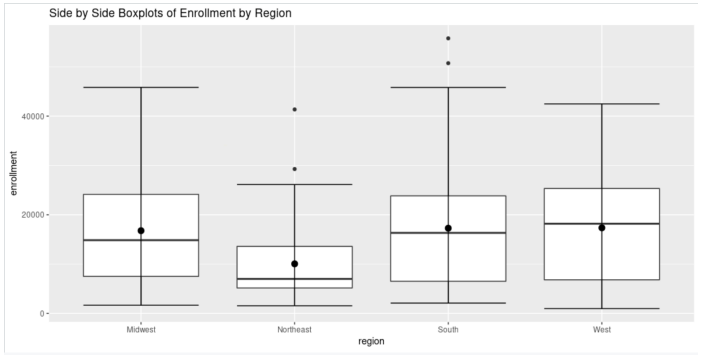
Normality of the data can also be proved by looking at the normal probability plots.

In this case, it looks like the data is pretty normal with some deviation at the ends.



The effect plot and box plots show that the enrollment rate in Northeastern schools is significantly less than that of the other three regions, with the West being the highest.





ANOVA Test Output

```
             Df    Sum Sq   Mean Sq F value   Pr(>F)
region        3 2.256e+09 751950259   6.609 0.000271 ***
Residuals   217 2.469e+10 113771771
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Conclusion 1: The ANOVA test suggests that the null hypothesis can be rejected because the P-value (0.000271) is less than 0.05. This provides support for the claim that at least two true means of enrollment are statistically different. Therefore, there will be at least one region whose enrollment rates are significantly higher or lower than at least one of the others.

Tukey Test Output

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = enrollment ~ region, data = univ_clean)

$region
                           diff        lwr       upr     p adj
Northeast-Midwest -6718.13178 -12059.174 -1377.089 0.0071130
South-Midwest       506.21054  -4748.196  5760.617 0.9945374
West-Midwest        581.94421  -5177.708  6341.597 0.9937135
South-Northeast    7224.34232   2357.815 12090.870 0.0009114
West-Northeast     7300.07599   1891.943 12708.209 0.0032040
West-South           75.73367  -5246.855  5398.322 0.9999820
```

Interval:

Northeast Midwest South West

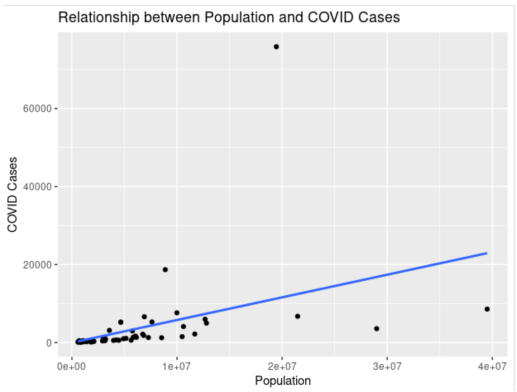
Conclusion 2: As determined by the Tukey test, the Northeast region seems to have a significantly lower enrollment rate on average than the other three regions, within a 95% confidence interval. This means that the Midwest, South, and West regions all have the highest enrollment rates because they're statistically the same.

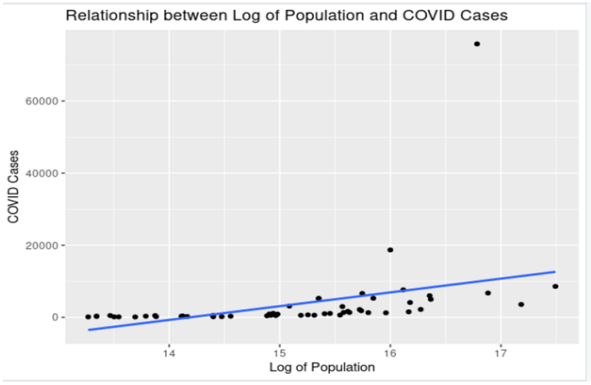
F) Inference 4:

Does the population size of a state significantly affect the number of COVID-19 cases?

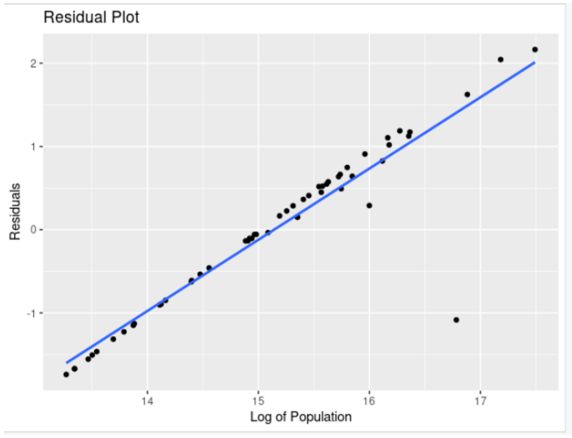
The statistical procedure that we will be using here is a linear regression hypothesis test. We are using this procedure because it is best for determining whether there is a significant linear relationship between an independent variable (state population) and a dependent variable (state's total COVID-19 cases). Also, the correlation coefficient will help show how strong the correlation between the two variables, if there is a relationship to begin with.

The residuals histogram was right skewed so we had to transform the data, specifically the population variable, using the log function. This transformed data is what we will use for the remainder of this inference.

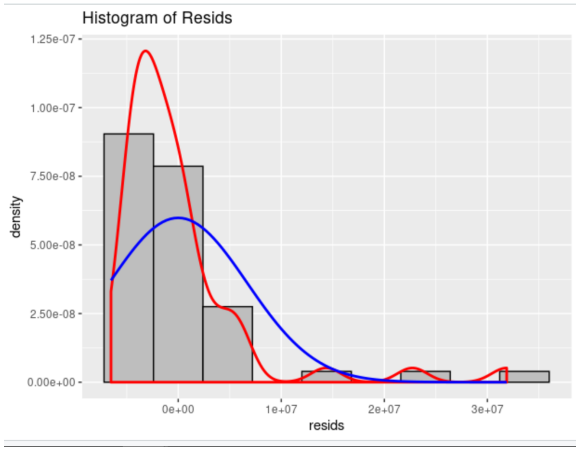


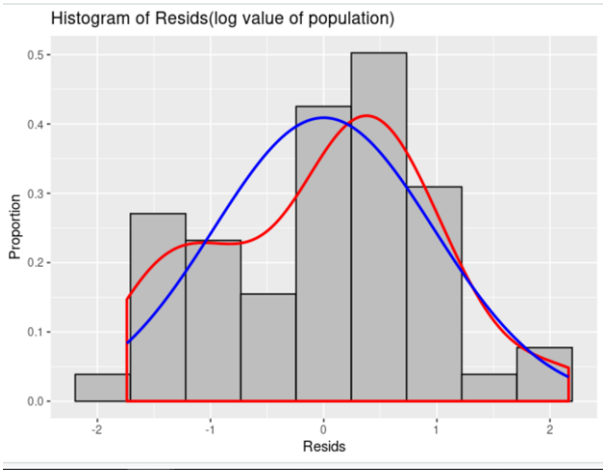


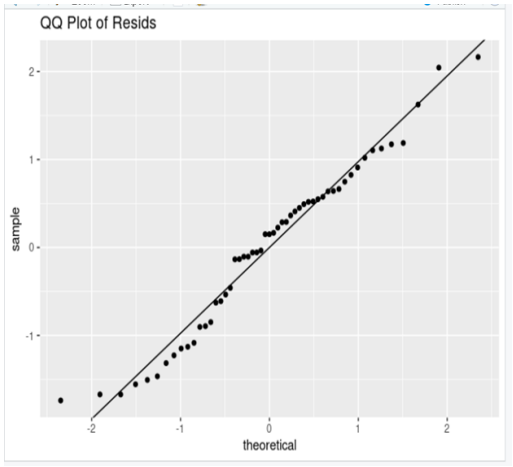
The scatter plot looks linear with a positive direction. There is an outlier with a high Covid-19 cases value.

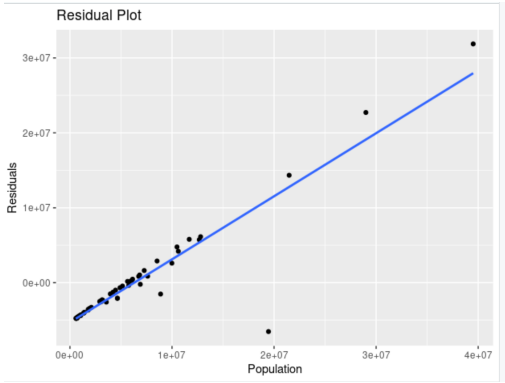


I see no pattern here so the association between Population and COVID-19 Cases seems to be linear. There is nothing strange that might infer the possibility that the standard deviation is not constant. There is again an outlier.









Left residual plot graph: default values Right residual plot graph: transformed values (log)

The residuals appear normal because on the normal probability plot the points are close to the line without systematic deviation. The histogram reveals a symmetric, unimodal pattern, and the curves appear close enough to one another, as they both share a similar bell shape. There are once again outliers in the histogram on both ends of the curve.

Assumptions:

- SRS with the observations independent of each other
- The relationship between Population and COVID-19 Cases is linear in the population
- The residuals have a normal distribution
- The standard deviation of the residuals is constant

The SRS can be assumed to be an SRS so that assumption is confirmed. The relationship between Population and COVID-19 Cases is linear in the population as shown in the scatterplot above. The residuals have a normal distribution as proven in both the residuals scatterplot and the QQ plot after the

log transformation was used. Finally, the standard deviation of the residuals appears to be constant from the residuals scatterplot and the QQ plot, again as shown above.

Linear Regression Hypothesis Test:

```
Call:
lm(formula = logpop ~ state.cases, data = CovidNew)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7394 -0.8488  0.1508  0.6373  2.1644

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.500e+01  1.432e-01 104.802  < 2e-16 ***
state.cases 3.777e-05  1.288e-05   2.932  0.00503 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9847 on 51 degrees of freedom
Multiple R-squared:  0.1443,    Adjusted R-squared:  0.1275
F-statistic: 8.597 on 1 and 51 DF,  p-value: 0.00503
```

Step 1 - Terms: (unnecessary)
Step 2 - Hypothesis:

H_0: there is no association between Population and COVID-19 Cases
H_a: there is an association between Population and COVID-19 Cases

Step 3 - Test statistic, df1, df2, p-value:

Fts = 8.597
df1 = 1
df2 = 51
p-value = 0.00503

Step 4 - Conclusion:

Since 0.00503 ≤ 0.05, we should reject H_0. The data provide evidence (p-value = 0.00503) to the claim that there is a positive association between Population and COVID-19 Cases.

Correlation Coefficient: 0.3798101

```
> cor(CovidNew$logpop, CovidNew$state.cases)
[1] 0.3798101
```

(Have to change this)This correlation coefficient proves there is correlation between Population and COVID-19 Cases because 0.3798101 has a fair distance from 0. Although the relationship is not necessarily strong, as 0.3798101 is also not very close to 1.

Conclusion:

We have successfully proven that the population size of a state affects the number of COVID-19 cases. We can see this is true from the output of the linear regression hypothesis test, where we found a p-value of 0.00503, which is less than the confidence level of 0.05, proving there is a correlation present. The correlation could then be measured with the correlation coefficient, which came out to be 0.3798101. This number is far enough from 0 that we can determine there is a correlation present. Both of these procedures prove that there is in fact a strong positive association between Population and COVID-19 Cases.

G) Conclusions:

Inference 1 asked the question that if you claim that a state has a high GDP if their GDP is greater than 59,927 (GDP per capita for the US, according to the World Bank), would the average GDP of the states in the top universities list be considered a high GDP? To determine whether this is true, we used a one-sample t procedure hypothesis test and a confidence interval. The hypothesis test gave an output of p-value 1, a much higher value than the significance level of 0.05, meaning that the alternative hypothesis of the average GDP of the states in the top universities list being considered a high GDP was not proven accurate. The confidence interval showed similar results. The confidence interval also showed a lower bound of 47417.4, which is much lower than the GDP per capita for the US (59,927), meaning this is not evidence that the average GDP of the states in the top universities list is a high GDP. Both of these procedures prove that the average GDP of the states in the top universities list is necessarily a high GDP. In **Inference 2**, we tested if the mean difference of tuition in private and public universities in the US was greater than $10000. We compared it with the information that we retrieved from an article on collegeraptor.com which compared average tuition between private and public universities in the US during 2017. Based on the data we gathered and performed tests on we are confident at a 95% confidence interval that private universities on an average cost about $10000 or more compared to public universities. The tuition amount that we have considered in the data is for 1 year and typically a college student graduates in 4 years, so going to a private university will cost as much as $40000 more than going to a public university which is a lot of money. This means that people should become more skeptical of their source when figuring out how much they would need to afford college tuition rates. In **Inference 3**, we wanted to determine if there is a significant difference between the true mean values of 4 regions with a single factor using the ANOVA test. There were three assumptions that we needed to check to do this test: SRS, normality, and constant variance. To answer the question, we performed an ANOVA test, which gave us a p-value of 0.000271. This value is lower than alpha (0.05), so we can reject the null hypothesis. This means that there is evidence that supports the claim that at least two true means of enrollment are statistically different. Then, we performed the Tukey test to confirm these results. This test suggested that within a 95% confidence interval, the Northeast region seems to have a significantly lower enrollment rate on average than the other three regions. The regions with the highest enrollment rates would be the Midwest, South, and West regions because they're statistically the same. This means that students are looking to study at a less populated school/area, they will most likely feel the most comfortable at a university in the Northeast.

In **Inference 4,** we tested if the population size of the state has a significant impact on the number of COVID-19 cases in that state i.e if they were correlated. By performing linear regression, we found out that there is a correlation between the population and the number of cases as the correlation coefficient was 0.3798101 which is not close to 0. So, states with higher populations are more likely to have more cases. Usually if the state has more population it is more likely to be dense in cities of that state, so there would be more cases present as the virus transmitted to people easily and spread to a larger population. We can see for ourselves that New York has recorded the most number of COVID-19 cases compared to any other state or even many countries around the world, the population of New York state is around 19.45 million and New York city itself has around 8.4 million people. So, there is a much higher risk of

the virus spreading if the population is higher and denser, so cities like New York have to undergo serious initiatives in order to contain the virus and stop it from spreading to many other people.

Appendix:

Inference 1:

```
states_old <- subset(univ_clean, select = c("state", "state.gdp.2017"))
states_new = unique(states_old)
xbar <- mean(states_new$state.gdp.2017)
s <- sd(states_new$state.gdp.2017)
ggplot(states_new, aes(state.gdp.2017)) +
        geom_histogram(aes(y = ..density..), bins = sqrt(nrow(states_new))+2, 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 GDP") + xlab("GDP") +
        ylab("Proportion")
ggplot(states_new, aes(x = "", y = state.gdp.2017)) +
        stat_boxplot(geom = "errorbar") + geom_boxplot() +
        ggtitle("Boxplot of GDP") +
        stat_summary(fun.y = mean, col = "black", geom = "point", size = 3)
t.test(states_new$state.gdp.2017, conf.level = 0.95, mu = 59927, alternative = "greater")
```

Inference 2:

Graphs: Boxplot, Histogram and qq-plot

```
> univsp20_cleaned <- subset(univsp20_cleaned, univsp20_cleaned$type == "private"|univsp20_cleaned$type == "public")
> library(ggplot2)
> ggplot(univsp20_cleaned, aes(x = type , y = tuition))+ geom_boxplot()+ stat_boxplot(geom="errorbar")+ stat_summary(fun.y
=mean, col="black", geom="point", size =3)+ ggtitle("Boxplots of tuition in private and public universities")
> xbar<-tapply(univsp20_cleaned$tuition, univsp20_cleaned$type, mean)
>
> s <-tapply(univsp20_cleaned$tuition, univsp20_cleaned$type, sd)
> n <-tapply(univsp20_cleaned$tuition, univsp20_cleaned$type, length)
> univsp20_cleaned$theoretical.density<-ifelse(univsp20_cleaned$type=="private",
+                                     dnorm(univsp20_cleaned$tuition, xbar["private"] , s["private"]),
+                                     dnorm(univsp20_cleaned$tuition, xbar["public"], s["public"]))
> ggplot(univsp20_cleaned, aes(x =tuition))+
+     geom_histogram(aes(y=..density..), bins = sqrt(min(n))+2,
+                    fill ="grey",col="black")+
+     facet_grid(type~ .)+
+     geom_density(col="red", lwd=1)+
+     geom_line(aes(y=theoretical.density), col="blue", lwd=1)+
+     ggtitle("Histograms of tuition between private and public universities")
> univsp20_cleaned$intercept<-ifelse(univsp20_cleaned$type=="private",
+                             xbar["private"],xbar["public"])
> univsp20_cleaned$slope<-ifelse(univsp20_cleaned$type=="private",
+                         s["private"],s["public"])
> ggplot(univsp20_cleaned, aes(sample=tuition))+
+     stat_qq()+
+     facet_grid(type~ .)+
+     geom_abline(data=univsp20_cleaned, aes(intercept = intercept,
+                                    slope = slope))+ ggtitle("QQ Plots of Tuition by Type of University")
```

Code for 2-Sample t-test:

```
> t.test(univsp20_cleaned$tuition~univsp20_cleaned$type, mu = 10000, paired =FALSE,
alternative ="greater", var.equal=FALSE)
```

Output for 2-Sample t-test:

```
        Welch Two Sample t-test

data:  univsp20_cleaned$tuition by univsp20_cleaned$type
t = 4.8769, df = 179.17, p-value = 1.184e-06
alternative hypothesis: true difference in means is greater than 10000
95 percent confidence interval:
 13733.26      Inf
sample estimates:
mean in group private  mean in group public
             44009.97              28361.85
```

Inference 3:

Graphs

```
#
# Boxplot
#
ggplot(data = univ_clean, aes(x = region, y = enrollment)) + geom_boxplot() +
  stat_boxplot(geom = "errorbar") +
  stat_summary(fun.y = mean, color = "black", size = 3, geom ="point")+
  ggtitle("Side by Side Boxplots of Enrollment by Region")
#
# Effects Plot
#
ggplot(data = univ_clean, aes(x = region, y = enrollment)) +
  stat_summary(fun.y = mean, geom = "point") +
  stat_summary(fun.y = mean, geom = "line", aes(group = 1)) +
  ggtitle("Effects Plot of Enrollment by Region")
#
# Table of descriptive statistics
#
tapply(univ_clean$enrollment, univ_clean$region, length)
tapply(univ_clean$enrollment, univ_clean$region, mean)
tapply(univ_clean$enrollment, univ_clean$region, sd)
#
# CHECK ASSUMPTIONS
#
### Histogram
xbar <- tapply(univ_clean$enrollment, univ_clean$region, mean)
s <- tapply(univ_clean$enrollment, univ_clean$region, sd)
n <- tapply(univ_clean$enrollment, univ_clean$region, length)
univ_clean$normal.density <- apply(univ_clean, 1, function(x){
  dnorm(as.numeric(x["enrollment"]), xbar[x["region"]], s[x["region"]])})
ggplot(univ_clean, aes(x = enrollment)) +
  geom_histogram(aes(y = ..density..), bins = sqrt(230) + 2,
                 fill = "grey", col = "black") +
  facet_grid(region ~ .) +
  geom_density(col = "red", lwd = 1) +
  geom_line(aes(y = normal.density), col = "blue", lwd = 1) +
  ggtitle("Histograms of Enrollment by Region")
### QQ Plot
univ_clean$intercept <- apply(univ_clean, 1, function(x){xbar[x["region"]]})
univ_clean$slope <- apply(univ_clean, 1, function(x){s[x["region"]]})
ggplot(univ_clean, aes(sample=enrollment)) + stat_qq() + facet_grid(region ~ .) +
  geom_abline(data= univ_clean, aes(intercept = intercept, slope = slope)) +
  ggtitle("QQ Plots of Enrollment by Region")
```

ANOVA test

```
fit <- aov(enrollment ~ region, data = univ_clean)
summary(fit)
```

```
              Df    Sum Sq   Mean Sq F value   Pr(>F)
region         3 2.256e+09 751950259   6.609 0.000271 ***
Residuals    217 2.469e+10 113771771
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Tukey Test

```
TukeyHSD(fit, conf.level = 0.95)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = enrollment ~ region, data = univ_clean)

$region
                          diff        lwr       upr     p adj
Northeast-Midwest -6718.13178 -12059.174 -1377.089 0.0071130
South-Midwest       506.21054  -4748.196  5760.617 0.9945374
West-Midwest        581.94421  -5177.708  6341.597 0.9937135
South-Northeast    7224.34232   2357.815 12090.870 0.0009114
West-Northeast     7300.07599   1891.943 12708.209 0.0032040
West-South           75.73367  -5246.855  5398.322 0.9999820
```

Inference 4:

Graphs

Code with the log transformations

```
univ_clean $logpop <- log(univ_clean$state.pop)
CovidOld <- subset(univ_clean, select = c("state", "logpop", "state.cases"))
CovidNew <- unique(CovidOld)

ggplot(CovidNew, aes(x=logpop, y=state.cases))+
+       geom_point() +
+       geom_smooth(method = lm, se = FALSE) +
+       ggtitle("Relationship between Log of Population and COVID Cases") +
+       xlab("Population") +
+       ylab("COVID Cases")
covid.lm <- lm(logpop ~ state.cases, data = CovidNew)
CovidNew$resids = covid.lm$residuals
ggplot(data = CovidNew, aes(x=logpop, y=resids))+
+       geom_point() +
+       geom_smooth(method = lm, se = FALSE) +
+       ggtitle("Residual Plot") +
+       xlab("Log of Population") +
+       ylab("Residuals")
```

```
# Histogram
xbar <- mean(CovidNew$resids)
s <- sd(CovidNew$resids)
ggplot(CovidNew, aes(resids)) +
+     geom_histogram(aes(y = ..density..),
+     bins = sqrt(nrow(CovidNew))+2,
+     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 Resids(log value of population)") +
+     xlab("Resids") +
+     ylab("Proportion")

#qqplot
ggplot(CovidNew, aes(sample = resids)) +
+         stat_qq() +
+         geom_abline(slope = s, intercept = xbar) +
+         ggtitle("QQ Plot of Resids")

#correlation coefficient
cor(CovidNew$logpop, CovidNew$state.cases)

#summary
summary(covid.lm)
```

#Code for graphs without log transformation

```
CovidOld <- subset(univ_clean, select = c("state", "state.pop", "state.cases"))
CovidNew <- unique(CovidOld)

ggplot(CovidNew, aes(x=state.pop, y=state.cases))+
+         geom_point() +
+         geom_smooth(method = lm, se = FALSE) +
+         ggtitle("Relationship between Population and COVID Cases") +
+         xlab("Population") +
+         ylab("COVID Cases")
covid.lm <- lm(state.pop ~ state.cases, data = CovidNew)
CovidNew$resids = covid.lm$residuals
ggplot(data = CovidNew, aes(x=state.pop, y=resids))+
+         geom_point() +
+         geom_smooth(method = lm, se = FALSE) +
```

```
+         ggtitle("Residual Plot") +
+         xlab(" Population") +
+         ylab("Residuals")

# Histogram
xbar <- mean(CovidNew$resids)
s <- sd(CovidNew$resids)
ggplot(CovidNew, aes(resids)) +

+    geom_histogram(aes(y = ..density..),
+    bins = sqrt(nrow(CovidNew))+2,
+    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 Resids") +
+    xlab("Resids") +
+    ylab("Proportion")

#qqplot
ggplot(CovidNew, aes(sample = resids)) +
+         stat_qq() +
+         geom_abline(slope = s, intercept = xbar) +
+         ggtitle("QQ Plot of Resids")
```