**PART E. ANOVA: Robberies Per Population versus Region**

**a) Code**

See appendix.

**b) Statistical Procedure**

Our goal for this portion of the project is to determine if there is any significant variation in robbery rate (RobberiesPerPopulation) between different regions of the US (Northeast, North Central, South, and West). If there is, we want to know where the variation lies.

This analysis should be conducted using a one-way ANOVA procedure, including both a hypothesis test and a Tukey multiple comparison procedure. The one-way ANOVA procedure is appropriate because we are attempting to analyze the variance of a parameter (mean Robberies Per Population) between multiple different populations (one for each region) that differ by a single factor (Region). The Tukey method for multiple comparisons is appropriate because we wish to compare the regions in a pairwise fashion to get a better idea of where any differences arise.

We will be performing our procedure with a two-sided hypothesis (at least one pair of distinct means is different).

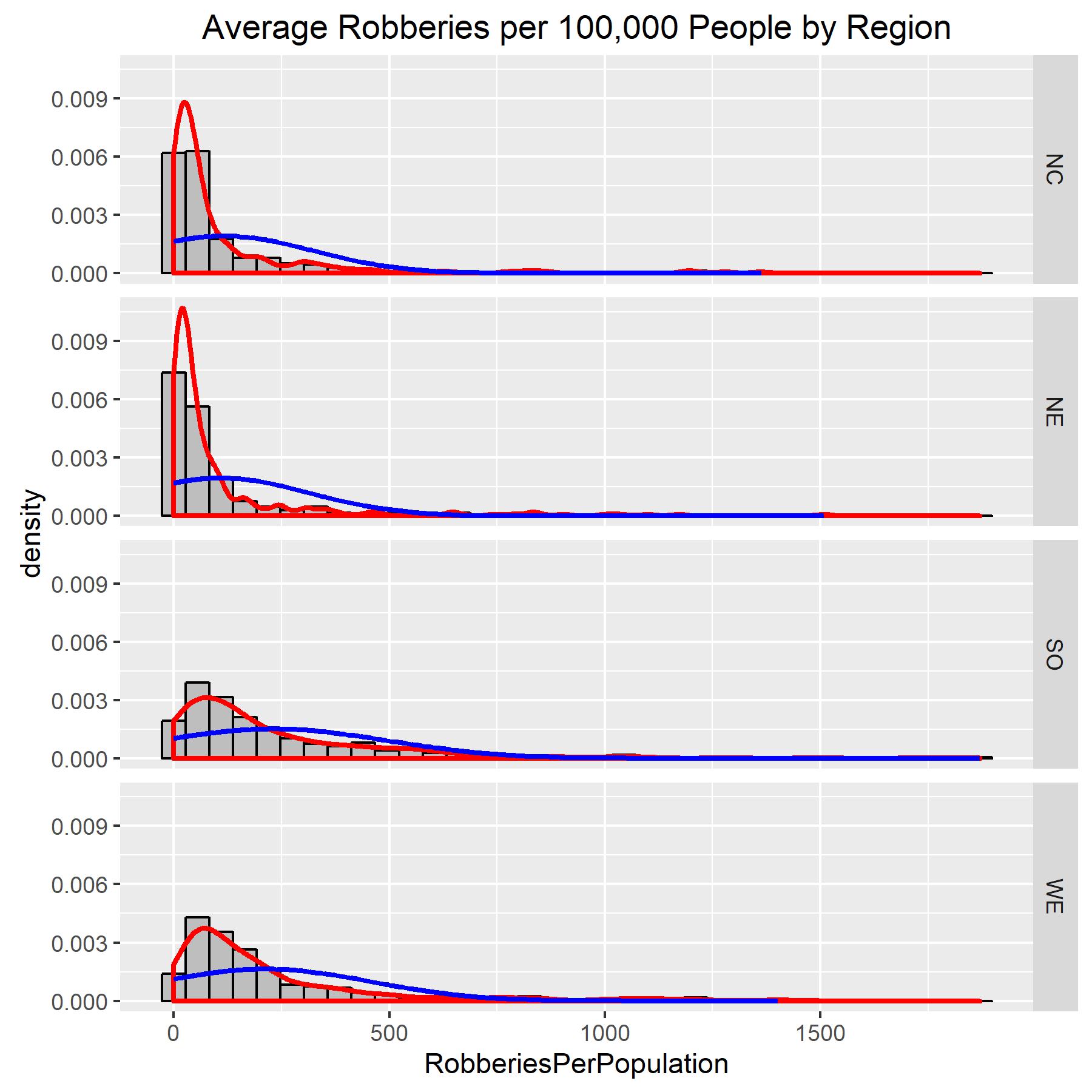
**c) Assumptions**

1) We have k independent SRSs, one from each population.

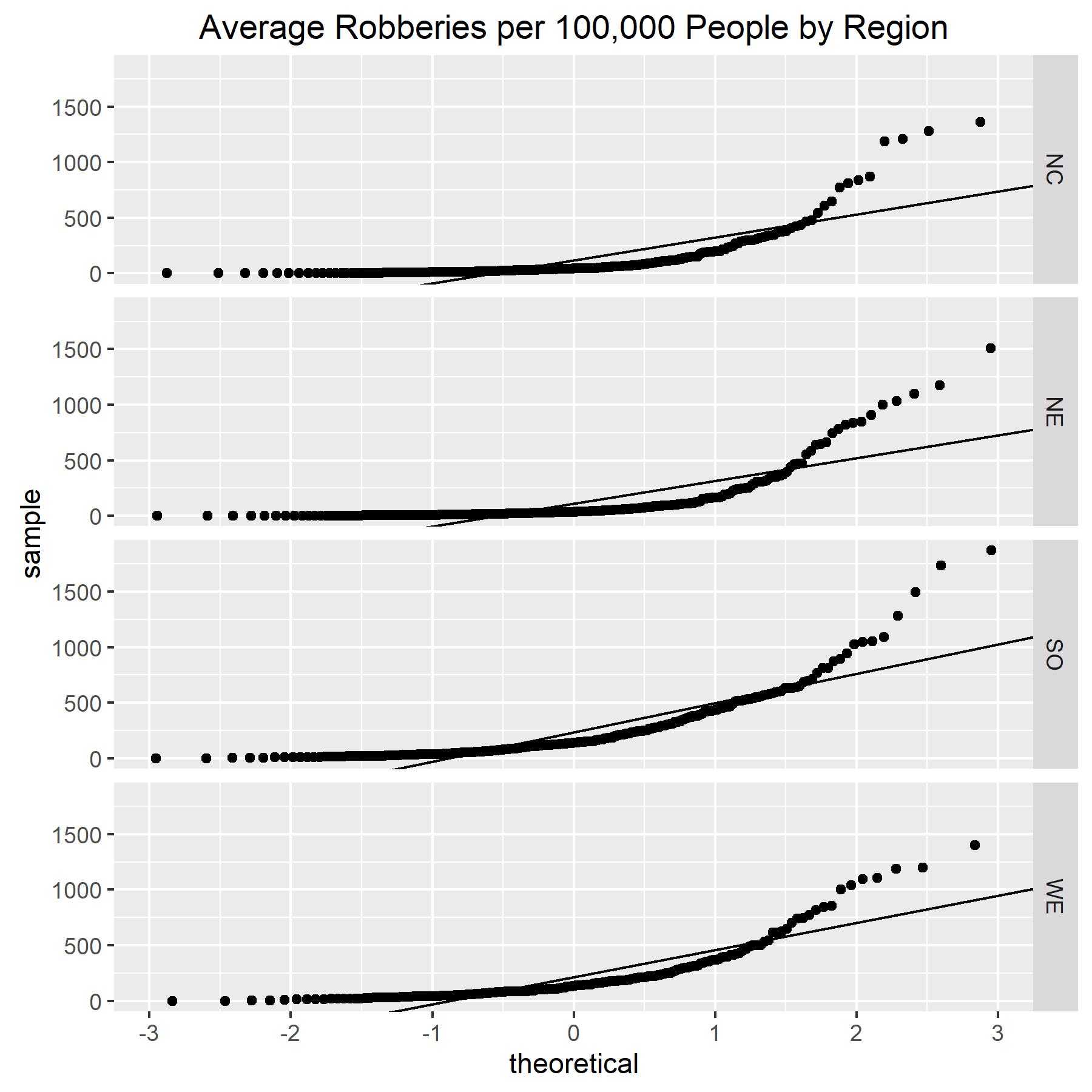
This assumption cannot be tested using the available information; we must simply assume that the US Data was collected through independent Simple Random Samples.

2) The *i*th population has a normal distribution with an unknown mean *μi*.

We can check this assumption by creating histograms and normal probability plots of Robberies Per Population for each Region.

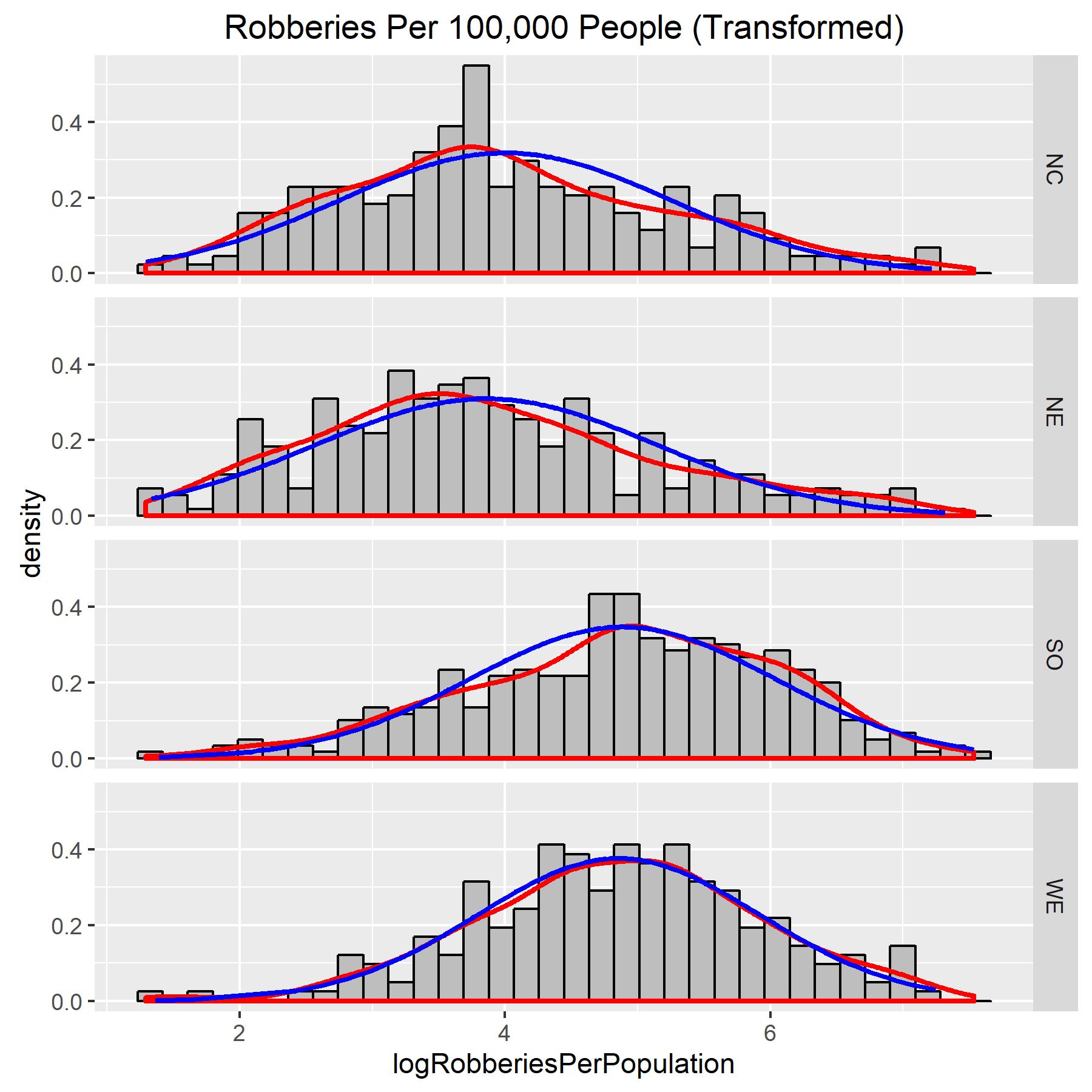


Our histograms do not appear to show normal distributions, but instead strong positive skews. We can check again with normal probability plots.

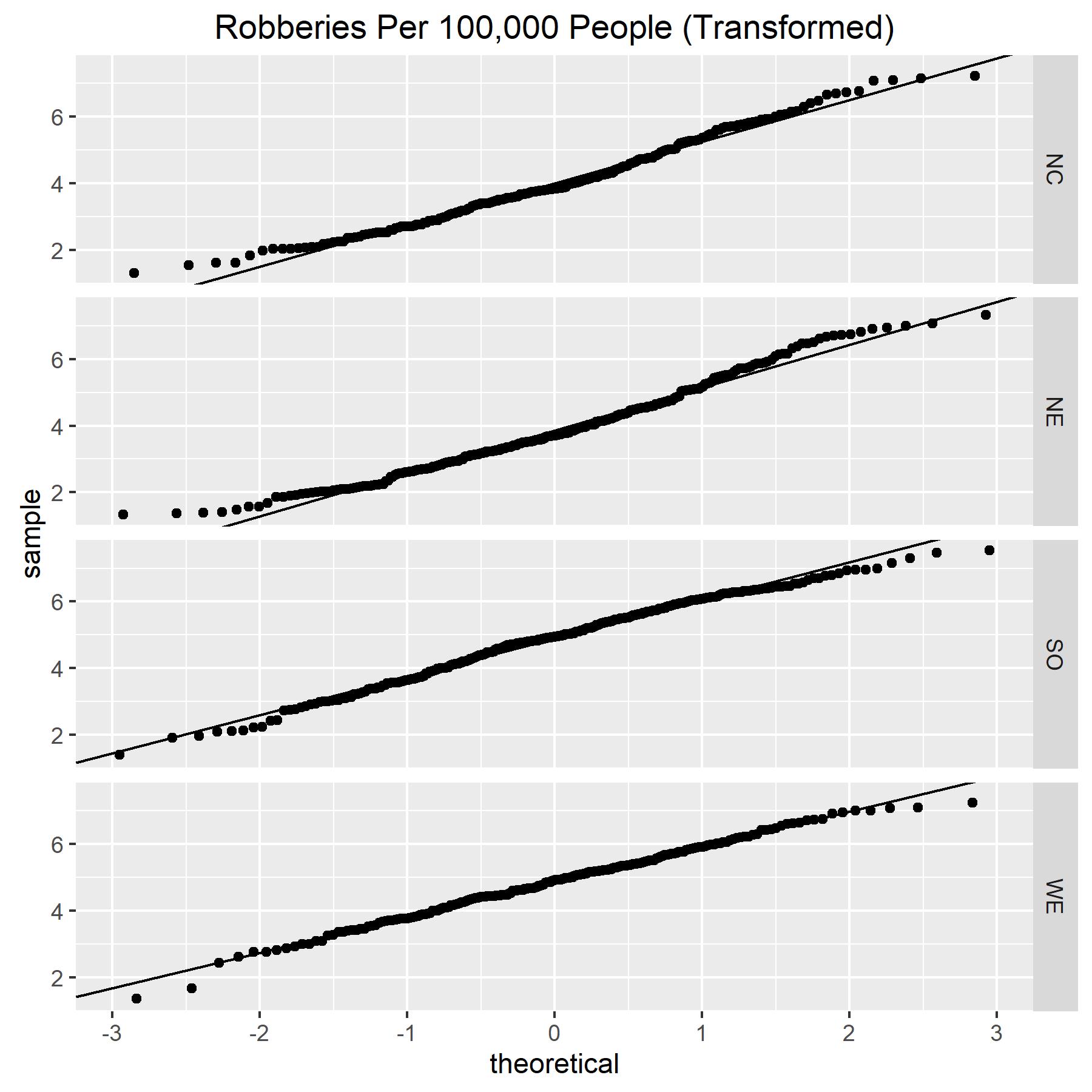


Indeed, these normal probability plots show a significant positive skew.

Clearly, the distributions of Robberies Per Population in each region are not normal. We will need to transform the data. The positive skew suggests a logarithmic transformation may be helpful. We can check again using histograms and normal probability plots of the transformed data.



These histograms of the transformed data appear to show normal distributions, with the density lines being similar to the normal lines. We can also check this with normal probability plots.



The normal probability plots also show a nearly normal distribution for the transformed data: most of the data points lie very close to their respective lines of normality.

Thus, if we perform our analysis using the transformed data, our second assumption is valid. Let us now return to our assumptions – we still need to check the third assumption.

3) All populations have the same variance, σ2, whose value is unknown.

We can check this assumption by calculating the sample variances of each of our transformed populations and checking if the maximum is twice the minimum (in which case this assumption is invalid).



The maximum standard deviation is in the Northeast region (smax = 1.289248) and the minimum standard deviation is in the West region (smin = 1.060813).

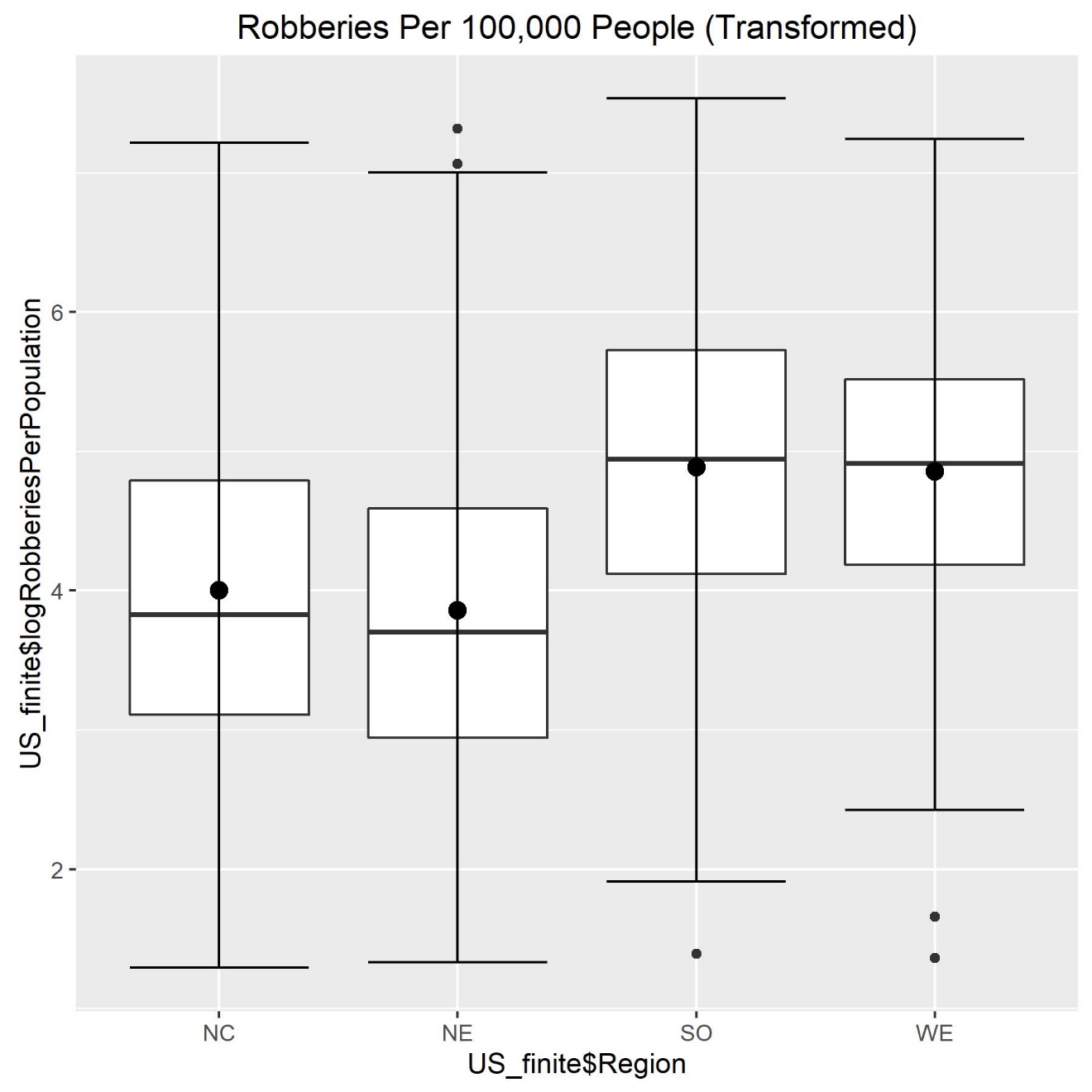
Thus smax / smin = 1.289248 / 1.060813 = 1.215 < 2.

This assumption is valid.

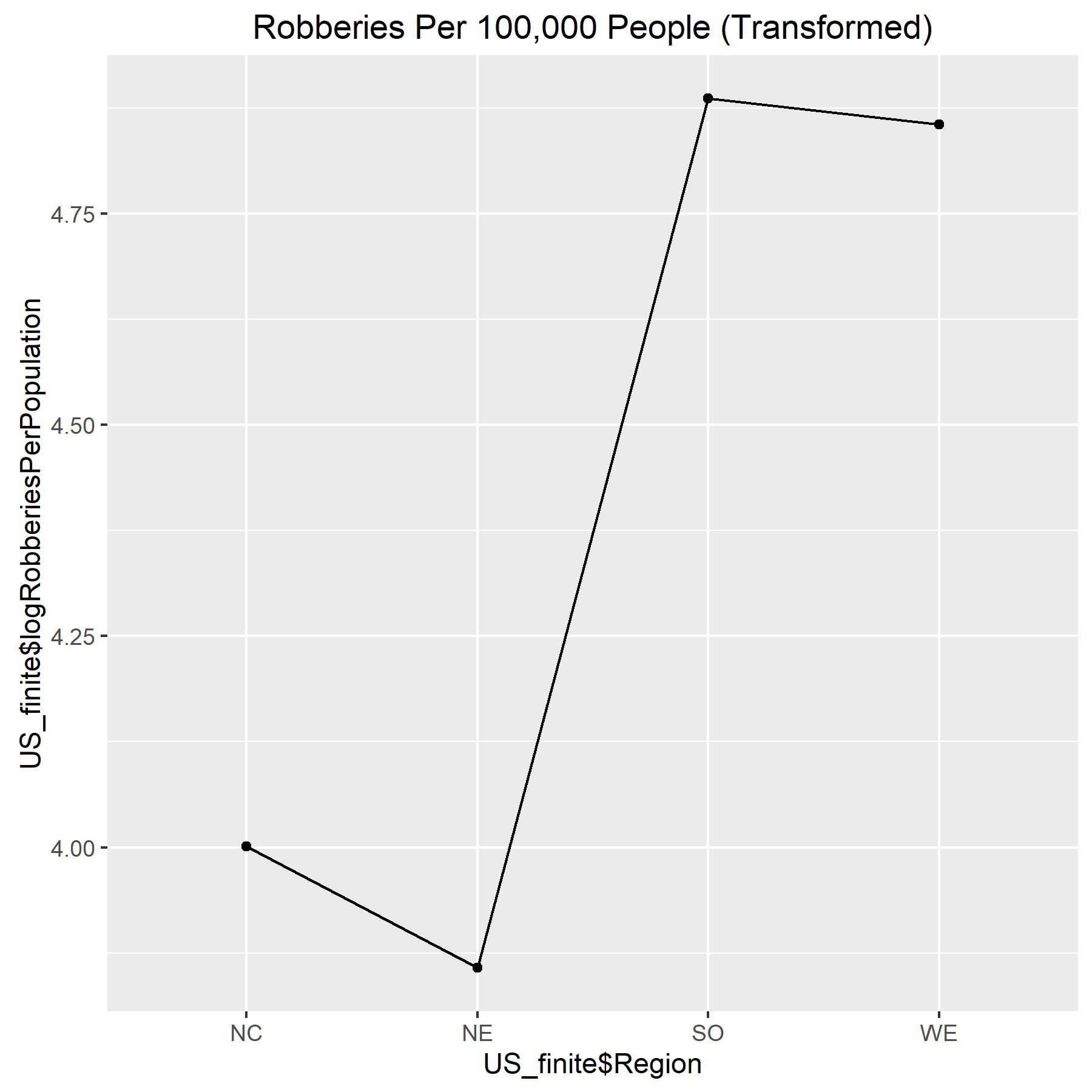
Thus, all necessary assumptions are satisfied.

**d) Graphs**

For a one-way ANOVA procedure with multiple comparison, we should begin by graphing the (transformed) data on side-by-side boxplots and an effects plot.



These side-by-side boxplots appear to show no significant difference in the mean log of robberies per population between different regions, as all the boxes appear to overlap along some horizontal region. However, it is difficult to draw a confident conclusion from this plot alone.



This effects plot appears to show differences in mean log of robberies per population between regions, especially between the Northern regions and the South/West regions. However, it is not clear from this plot alone how significant these differences may be. A one-way ANOVA procedure and Tukey procedure should provide clearer information.

**e) Procedures**

**Hypothesis Test**

1. Parameters

μNE is the population mean log of Robberies per Population for the Northeast region.

μNC is the population mean log of Robberies per Population for the North Central region.

μSO is the population mean log of Robberies per Population for the South region.

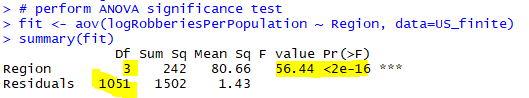
μWE is the population mean log of Robberies per Population for the West region.

2. Hypotheses

H0: μNE = μNC = μSO = μWE

Ha: at least two μi’s are different.

3. Test statistic (Fts), degrees of freedom (DF1, DF2), and p-value (p)



Fts = 56.44

DF1 = 3, DF2 = 1051

p < 2\*10-16

4. Conclusion

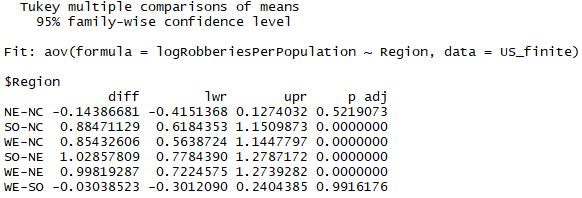
α = 0.05

p < 2\*10-16 🡪 p << α 🡪 reject H0

The data provides strong evidence (p < 2\*10-16) to the claim that the population mean log of Robberies per Population for at least one of the four US regions is different from the rest.

**Multiple Comparison**

α = 0.05 🡪 95% confidence intervals



To determine which pairs are significant, we can check whether the interval (lwr, upr) contains 0. This would mean we are 95% confident that an interval containing 0 captures the true difference in population means between these two regions (in which case there is no significant difference). Alternatively, we can check whether p adj is smaller than α (in which case there is a significant difference). Whichever method we use, this procedure yields statistical evidence that the population mean log of Robberies per Population is different between the South and North Central regions; between the West and North Central regions; between the South and Northeast regions; and between the West and Northeast regions. However, we do not have evidence that the population mean is different between the Northeast and North Central regions or between the West and South regions.

We can represent this information graphically as well, using sample means we have calculated in our code.



|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| 3.857433 | 4.001299 | 4.855625 | 4.886011 |

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Note that the highest upper bound of any confidence interval that does not include zero is 1.2787172. This means the highest difference we have evidence for is between the South and Northeast regions, at which the natural logarithm of the difference is no greater than 1.2787172. Since e1.2787172 = 3.59202, we can say we can say the maximum difference we found in population mean robberies per population between regions ins 3.59202 robberies per 100,000 people.

**f) Conclusion**

Our statistical analysis gives reason to believe that robbery rates (robberies per 100,000 people) are, in general, higher in the West and South regions of the United States than it is in the Northeast and North Central regions; however, we have no evidence of a difference in average robbery rate between the West and South regions, or between the North and North Central regions. It is important to note that this only tells us this difference exists; we have no evidence to say what causes the difference. It is also important to note these are only differences in *average* robbery rate per region, based on per-county data. So, it may still be the case that certain counties or cities in the Northern regions may have higher robbery rates than certain counties or cities in the West and South regions. Finally, one should note that although we can be very confident that these differences exist, the available information suggests these differences are no greater than 4 robberies per 100,000 people, or 0.004%. In practical terms, this is of little significance.

**Appendix E. Code for Part E**

#####

# Jordan Mayer

# STAT 350

# Project

# April 19, 2018

#####

### setup ###

setwd("C:/Users/jordan/Google Drive/Courses Spring 2018/STAT 350/STAT 350 Labs/Project")

# set working directory

library(ggplot2) # set up ggplot2 for plotting

graphics.off() # close any open figures

USData <- read.table("US\_Data.txt", header=TRUE, sep="\t") # get US Data

US\_clean <- USData[complete.cases(USData),] # clean US Data

# data of interest: robbery rate (RobberiesPerPopulation) across

# regions of US (Region)

### Part b ###

# check normality via histograms and normal probability plots

title <- "Robberies Per 100,000 People"

# calculate theoretical density curves

xbar <- tapply(US\_clean$RobberiesPerPopulation, US\_clean$Region, mean)

sd <- tapply(US\_clean$RobberiesPerPopulation, US\_clean$Region, sd)

US\_clean$normal.density <- apply(US\_clean, 1, function(x) {

dnorm(as.numeric(x["RobberiesPerPopulation"]),

xbar[x["Region"]], sd[x["Region"]])

})

# create histograms

hist <- ggplot(US\_clean,aes(x=RobberiesPerPopulation))+

geom\_histogram(aes(y=..density..),bins=sqrt(nrow(US\_clean))+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(title)

ggsave(hist,filename="hist.jpg",width=6,height=6)

# calculate slopes and intercepts for lines of normality

US\_clean$intercept <- apply(US\_clean, 1, function(x){xbar[x["Region"]]})

US\_clean$slope <- apply(US\_clean, 1, function(x){sd[x["Region"]]})

# create normal probability plots

qq <- ggplot(US\_clean,aes(sample=RobberiesPerPopulation))+

stat\_qq()+

facet\_grid(Region ~ .)+

geom\_abline(data=US\_clean,aes(intercept=intercept,slope=slope))+

ggtitle(title)

ggsave(qq,filename="qq.jpg",width=6,height=6)

# transform data

US\_clean$logRobberiesPerPopulation <- log(US\_clean$RobberiesPerPopulation)

# remove non-finite data

US\_finite <- subset(US\_clean, US\_clean$logRobberiesPerPopulation != -Inf)

title = "Robberies Per 100,000 People (Transformed)"

# calculate theoretical density curves

xbar\_log <- tapply(US\_finite$logRobberiesPerPopulation, US\_finite$Region, mean)

sd\_log <- tapply(US\_finite$logRobberiesPerPopulation, US\_finite$Region, sd)

US\_finite$normal.density.log <- apply(US\_finite, 1, function(x) {

dnorm(as.numeric(x["logRobberiesPerPopulation"]),

xbar\_log[x["Region"]], sd\_log[x["Region"]])

})

# create histograms

hist\_log <- ggplot(US\_finite,aes(x=logRobberiesPerPopulation))+

geom\_histogram(aes(y=..density..),bins=sqrt(nrow(US\_finite))+2,

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(title)

ggsave(hist\_log,filename="hist\_log.jpg",width=6,height=6)

# calculate slopes and intercepts for lines of normality

US\_finite$intercept <- apply(US\_finite, 1, function(x){xbar\_log[x["Region"]]})

US\_finite$slope <- apply(US\_finite, 1, function(x){sd\_log[x["Region"]]})

# create normal probability plots

qq\_log <- ggplot(US\_finite,aes(sample=logRobberiesPerPopulation))+

stat\_qq()+

facet\_grid(Region ~ .)+

geom\_abline(data=US\_finite,aes(intercept=intercept,slope=slope))+

ggtitle(title)

ggsave(qq\_log,filename="qq\_log.jpg",width=6,height=6)

### Part c ###

# display sample statistics

# sample sizes

tapply(US\_finite$logRobberiesPerPopulation, US\_finite$Region, length)

# sample means

tapply(US\_finite$logRobberiesPerPopulation, US\_finite$Region, mean)

# sample standard deviations

tapply(US\_finite$logRobberiesPerPopulation, US\_finite$Region, sd)

### Part d ###

# create side-by-side boxplots

box\_log <- ggplot(US\_finite, aes(x=US\_finite$Region,

y=US\_finite$logRobberiesPerPopulation))+

geom\_boxplot()+

stat\_boxplot(geom="errorbar")+

stat\_summary(fun.y=mean,col="black",geom="point",size=3)+

ggtitle(title)

ggsave(box\_log,filename="box\_log.jpg",width=6,height=6)

# create effects plot

effects\_log <- ggplot(data=US\_finite,aes(x=US\_finite$Region,

y=US\_finite$logRobberiesPerPopulation))+

stat\_summary(fun.y=mean,geom="point")+

stat\_summary(fun.y=mean,geom="line",aes(group=1))+

ggtitle(title)

ggsave(effects\_log,filename="effects\_log.jpg",width=6,height=6)

### Part e ###

# ANOVA significance test

# perform ANOVA significance test

fit <- aov(logRobberiesPerPopulation ~ Region, data=US\_finite)

summary(fit)

# perform multiple-comparison via Tukey procedure

test.Tukey <- TukeyHSD(fit, conf.level=0.95)

test.Tukey