# Fall 2017 STAT 350 Project: Assaults per Population

Jae Yun Park, Dahee Kim

# Appendix

## Codes

### For C, 1-sample t-test

```
** Read in data;
data USDataCleaned;
infile 'W:\usdata_cleaned.txt' delimiter = '09'x firstobs = 2;
input State $ Region $ CountyIndex $ UrbanIndicator $ Population LandArea
PopulationDensity PercentMaleDivorce PercentFemaleDivorce MedianIncome
IncomeCategory $ PercentCollegeGraduates MedianHouseAge RobberiesPerPopulation
AssaultsPerPopulation BurglariesPerPopulation
LarceniesPerPopulation EducationSpending EducationSpendingP2 TestScore ;
run;
** Call t-test for plots, summary stats, intervals and test of
AssaultsPerPopulation;
proc ttest data = USDataCleaned H0 = 4.8 sides = 2 alpha=0.05;
var AssaultsPerPopulation;
run;
** Create Logged Variable;
data logUSDataCleaned;
set USDataCleaned;
logAssaults = log(AssaultsPerPopulation);
run;
** Call t-test for plots, summary stats, intervals and test of logAssaults;
proc ttest data = logUSDataCleaned H0 = 4.8 sides = 2 alpha=0.05;
var logAssaults;
run;
```

### For D, 2-sample t-test

```
** Read in data;
data USDataCleaned;
infile 'W:\usdata_cleaned.txt' delimiter = '09'x firstobs = 2;
input State $ Region $ CountyIndex $ UrbanIndicator $ Population LandArea
PopulationDensity PercentMaleDivorce PercentFemaleDivorce MedianIncome
IncomeCategory $ PercentCollegeGraduates MedianHouseAge RobberiesPerPopulation
AssaultsPerPopulation BurglariesPerPopulation
LarceniesPerPopulation EducationSpending EducationSpendingP2 TestScore ;
run;
** Create Logged Variable;
data logUSDataCleaned;
```

```
set USDataCleaned;
logAssaults = log(AssaultsPerPopulation);
run;
** Run t test on logged data;
proc ttest data = logUSDataCleaned H0 = 4.8 sides = 2 alpha=0.05;
class UrbanIndicator;
var logAssaults;
run;
```

**For E, One way ANOVA**

```
** Read in data;
data USDataCleaned;
infile 'W:\usdata_cleaned.txt' delimiter = '09'x firstobs = 2;
input State $ Region $ CountyIndex $ UrbanIndicator $ Population LandArea
PopulationDensity PercentMaleDivorce PercentFemaleDivorce MedianIncome
IncomeCategory $ PercentCollegeGraduates MedianHouseAge RobberiesPerPopulation
AssaultsPerPopulation BurglariesPerPopulation
LarceniesPerPopulation EducationSpending EducationSpendingP2 TestScore ;
run;
** Create Logged Variable;
data logUSDataCleaned;
set USDataCleaned;
logAssaults = log(AssaultsPerPopulation);
run;
** sort needed for later procs;
proc sort data = logUSDataCleaned;
by Region;
run;
** Make summary statistics, save means for effects plot;
proc means data = logUSDataCleaned; *first calculate the averages;
var logAssaults;
class Region;
output out=means mean=average;
run;
symbol1 v=dot i=join; *to make the effects plot 'pretty';
proc gplot data=means;
plot average*Region;
run;
** QQ PLOT;
```

```
proc sgplot data=logUSDataCleaned;
By Region;
histogram logAssaults;
density logAssaults;
density logAssaults/type=kernel;
run;
** HISTOGRAM;
proc univariate data=logUSDataCleaned noprint;
By Region;
QQplot logAssaults/normal (mu=est sigma=est);
run;
*** ---------;
*** Run ANOVA and make Boxplot;
*** ---------;
proc glm data=logUSDataCleaned alpha=0.05;
class Region;
model logAssaults = Region;
means Region /Tukey cldiff;
run;
```

**A.**

We chose to construct a study about Assaults per population. Our questions are the following:

1) Whether the assaults per population USData is consistent with what FBI suggests, 248.5 per 100,000[1].
2) Whether the assaults per population is significantly different between urban and rural areas.
3) Whether the assaults per population is significantly different among the regions: South (SO), Northeast (NE), North Central (NC), and West (WE).

Security has been the most essential factor that we consider. It is directly related to our life safety, and people would appreciate annual crime rate reports; however, a misleading or wrong information can affect on the decision of such as which areas are safe or dangerous and need security development. The answers to the questions above will suggest whether USData and FBI have a consistent report, whether urban or rural area is safer, and which region is the safest from assaults.

**B.**

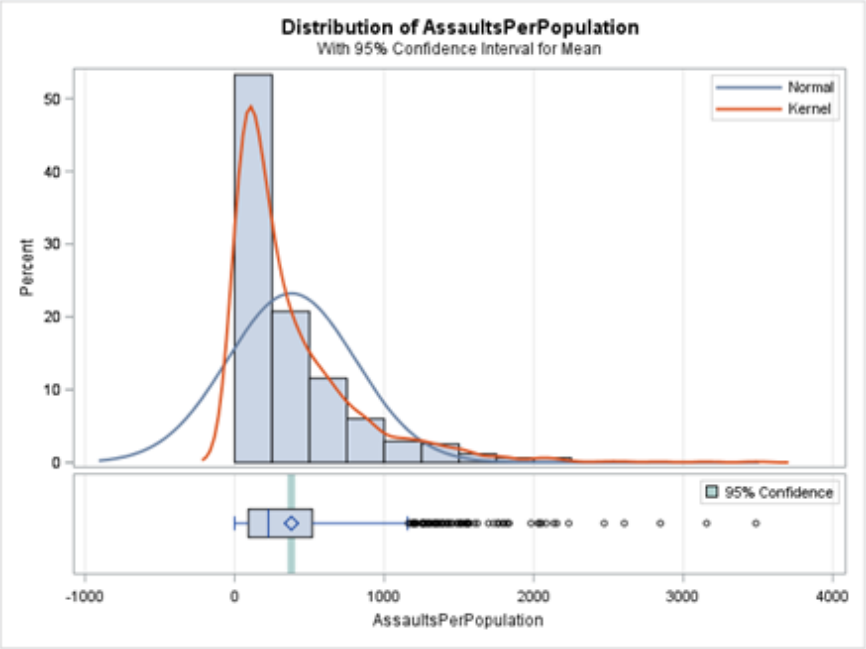
The variable we used are as follows:

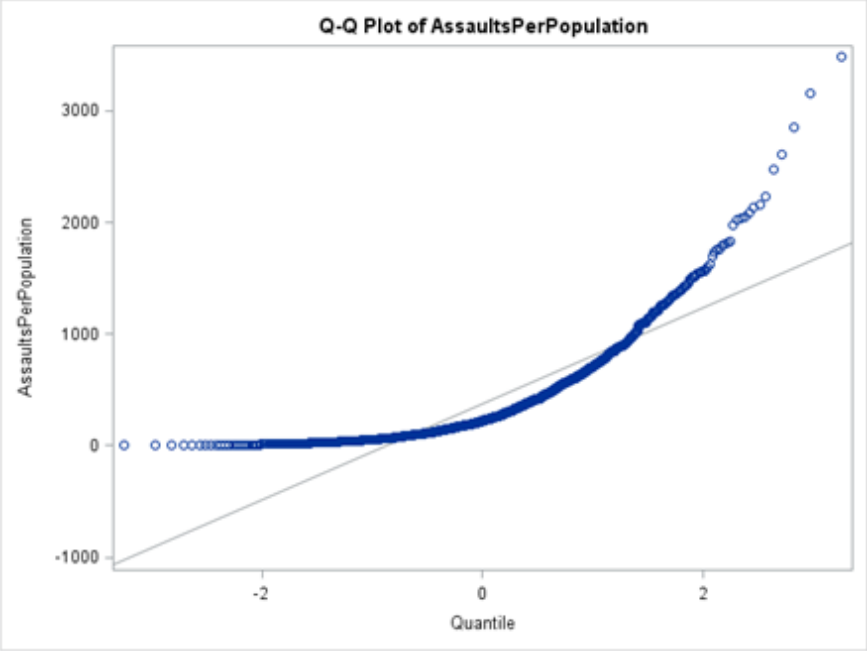
- Region
  - Categorical
  - Region the state belongs to
  - South (SO), Northeast (NE), North Central (NC), and West (WE)
- UrbanIndicator
  - Categorical
  - Indicates whether a country is urban (1) or rural (0)
- AssaultsPerPopulation
  - Numeric
  - Number of assaults out of 100,000 people

[1] https://ucr.fbi.gov/crime-in-the-u.s/2016/crime-in-the-u.s.-2016/topic-pages/aggravated-assault

**C. (1-sample t-test)**

b)

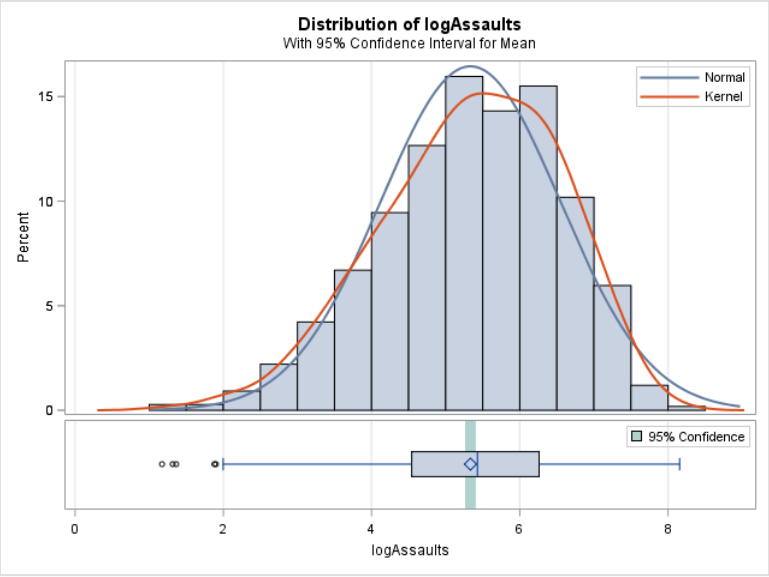


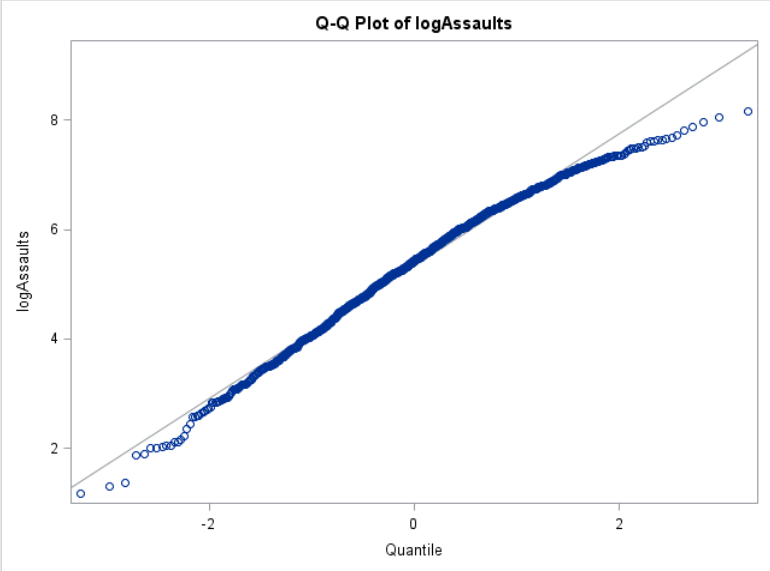




We proceeded one sample t-test with AssaultsPerPopulation to see the consistency between the assaults per population of USData and what FBI suggested, 248.5 per 100,000. Since we are checking whether both values from USData and FBI are different or not, we performed two-sided inferences.

Also, since we used transformed log Assaults per Population, we also logged the value from FBI, which is 5.515, in order to proceed hypothesis tests.

c) Before we created logged variable of AssaultsPerPopulation, the histogram clearly shows right skewedness. Also, the boxplot shows that several outliers exist. Moreover, since the QQ plot shows concave up, this distribution does not show normality.




After making AssaultsPerPopulation logged, the histogram shows normality even if it is slightly left skewed. Also, the boxplot shows that there are only five outliers. Moreover, since the plots of QQ plot follow along the line, this distribution shows normality.

Finally, this is appropriate to analyze these data using the t procedures with assumptions of Simple Random Sample (SRS), and the data come from a normal distribution to take advantage of CLT. In our case, the sample size is large and the distribution is normal.

d)

| N | Mean | Std Dev | Std Err | Minimum | Maximum |
|---|---|---|---|---|---|
| 1090 | 5.3341 | 1.2127 | 0.0367 | 1.1725 | 8.1566 |

The sample mean is 5.3341, the sample standard deviation is s = 1.2127, and the standard error of the mean is 0.0367.
This means that the data is centered at 5.3341 with a relatively small spread. The standard deviation of the test statistics is much smaller than the sample standard deviation so the sampling distribution has a much narrower spread than the population does.

e)

| Mean | 95% CL Mean | | Std Dev | 95% CL Std Dev | |
|---|---|---|---|---|---|
| 5.3341 | 5.2620 | 5.4062 | 1.2127 | 1.1638 | 1.2658 |

The 95% confidence interval is (5.2620, 5.4062).
We are 95% confident that the population mean of log Assaults per Population is covered by 5.2620 and 5.4062.

Based on the 95% confidence interval, we would reject the claim that the mean log Assaults per Population is 5.515 at 5% level, since the confidence interval does not contain the value 5.515.

To check these data does provide evidence that the average log Assaults per Population is different from 5.515, we used the four-step procedure.

| DF | t Value | Pr > \|t\| |
|---|---|---|
| 1089 | 14.54 | <.0001 |

Step 1: Define Parameters
Let $\mu$ represents the population mean of log Assaults per Population

Step 2: State Hypothesis
$H_0$: $\mu = 5.515$
$H_a$: $\mu \neq 5.515$

Step 3: Test Statistic, Degrees of Freedom, P-value
$t_{ts} = 14.54$
df = 1098

$p < 0.0001$

Step 4: State Conclusions

Since the p-value is much smaller than 0.05, we should reject the null hypothesis. The data shows strong support ($p < 0.0001$) to the claim that the population mean of log Assaults per Population is not equal to 5.515.

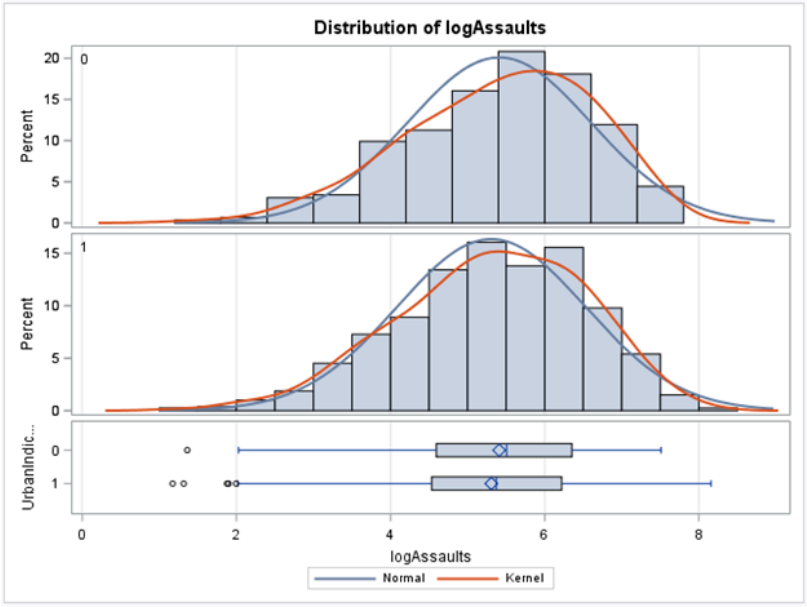
f)

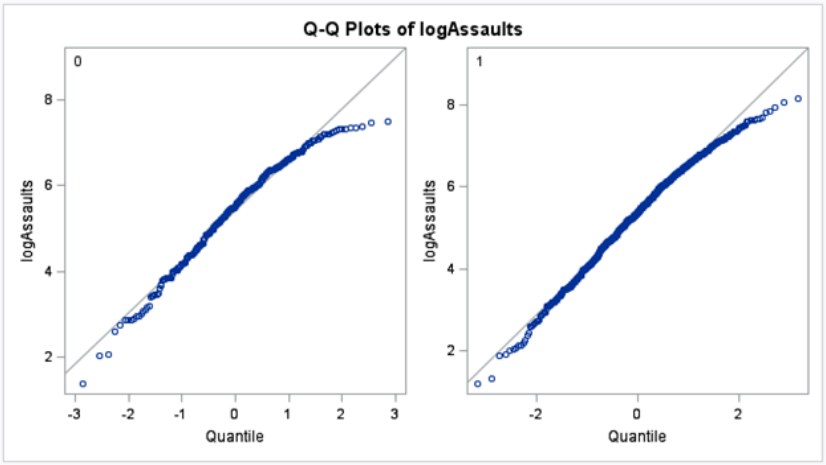
The assumption of a mean of 5.515 is very close to the confidence range of 5.2620 to 5.4062. We would say that this small difference from confidence interval and the assumption value can be ignored when we look for the consistency of values of log assaults per population from USData and FBI. Therefore, we would say that the mean of the log Assaults per Population could be 5.515.

**D. (2-sample t-test)**

b)
We proceeded two sample independent t-test with log AssaultsPerPopulation and UrbanIndicator to see whether the assaults per population is significantly different between urban and rural areas. Since we are checking whether they are significantly different or not, we performed two-sided inferences.

c)




As we mentioned at part C, we logged AssautlsPerPopulation in order to make the data normally distributed. Therefore, the histograms show unimodal and symmetric even if they look slightly left skewed. Even though some outliers can be seen, it seems they are not extreme.
Also, the plots of QQ plot follow the line closely. Even though there is a slight curve at the right top of QQ plot of rural county, it can be regarded as normally distributed.

This is appropriate to analyze these data using the t procedures with assumptions of Simple Random Sample (SRS), and the data come from a normal distribution to take advantage of CLT. In our case, the sample size is large and the distribution is normal.

d)

| UrbanIndicator | Method | Mean | 95% CL Mean | | Std Dev | 95% CL Std Dev | |
|---|---|---|---|---|---|---|---|
| 0 | | 5.4097 | 5.2727 | 5.5466 | 1.1912 | 1.1020 | 1.2964 |
| 1 | | 5.3063 | 5.2215 | 5.3912 | 1.2200 | 1.1629 | 1.2831 |
| Diff (1-2) | Pooled | 0.1033 | -0.0592 | 0.2659 | 1.2124 | 1.1635 | 1.2655 |
| Diff (1-2) | Satterthwaite | 0.1033 | -0.0576 | 0.2643 | | | |

This shows that 95% confidence interval is (-0.0576, 0.2643).
We are 95% confident that the difference in population mean of the log Assault per Population between rural and urban region is covered by the interval -0.0576 and 0.2643.

e)

| Method | Variances | DF | t Value | Pr > \|t\| |
|---|---|---|---|---|
| Pooled | Equal | 1088 | -56.70 | <.0001 |
| Satterthwaite | Unequal | 531.62 | -57.33 | <.0001 |

Step 1: Define Parameters

Let $\mu_{rural}$ represents the population mean of log Assaults per Population for rural county, and $\mu_{urban}$ represents the population mean of log Assaults per Population for urban county.

Step 2: State Hypotheses

$H_0$: $\mu_{rural} - \mu_{urban} = 0$

$H_a$: $\mu_{rural} - \mu_{urban} \neq 0$

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

$t_{ts} = -57.33$

df = 531.62

$p < 0.0001$

Step 4: State Conclusions

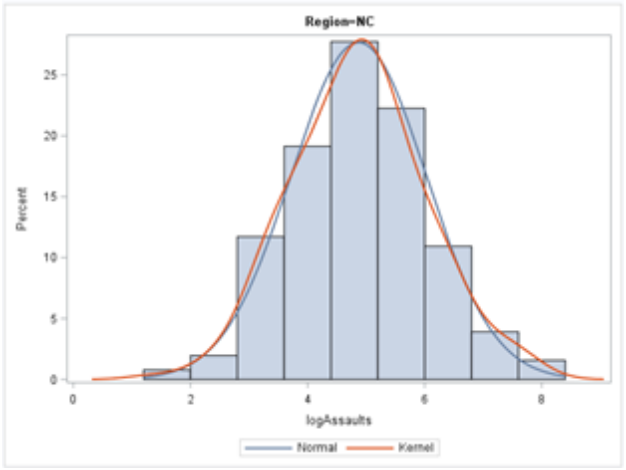
Since the p-value is much smaller than 0.05, we reject the null hypothesis. The data shows strong support ($p < 0.0001$) to the claim that the population mean difference of log Assaults per Population is different between the rural and urban counties.

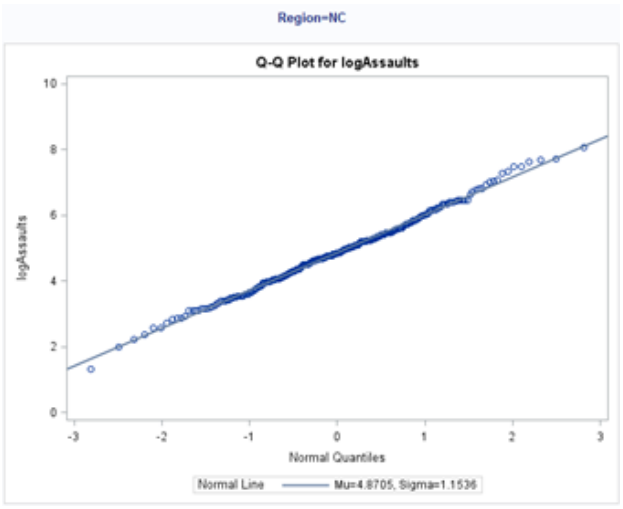
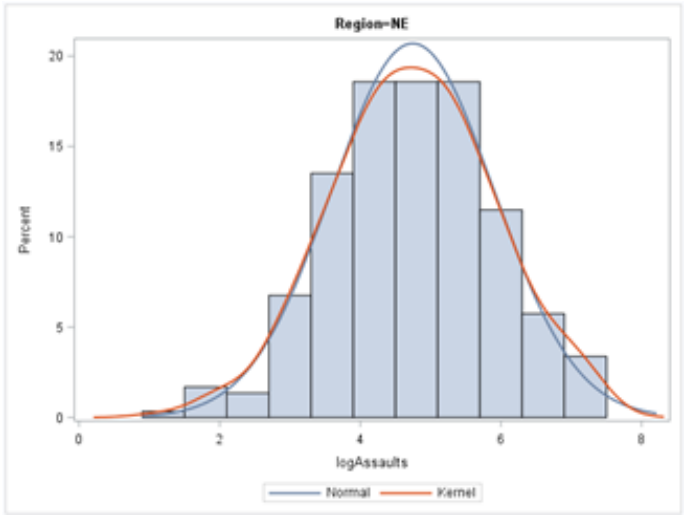
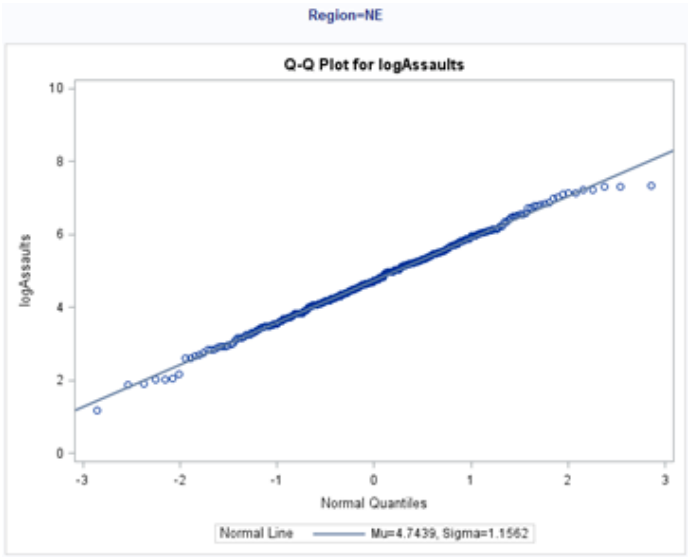
f)
The 95% confidence interval includes 0, and the range of interval is so close to 0. Also, the four-step hypotheses test show that the p-value is smaller than the alpha value, which is 0.05. We can derive a conclusion that the values of assaults per population from rural and urban area are close to each other.

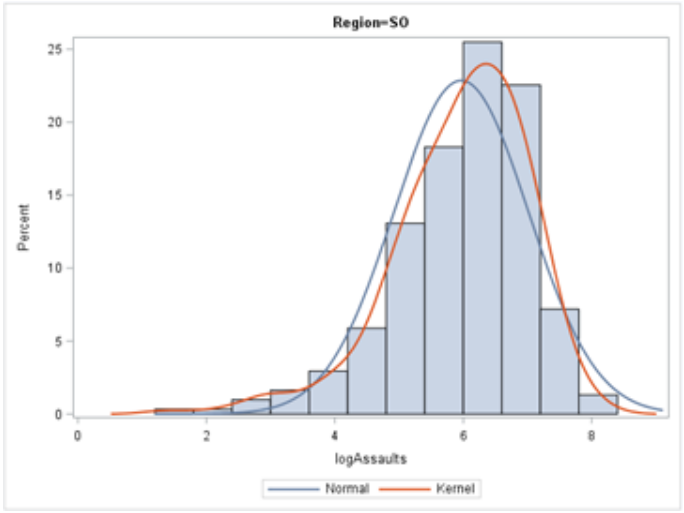
## E. (One way ANOVA)

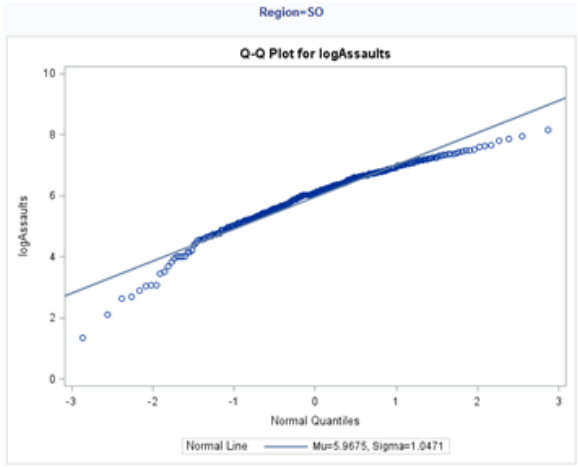
b)
We used ANOVA test by using two variables: log AssaultsPerPopulation and Region. We used this statistical procedure because we wanted to see whether the assaults per population is significantly different among the regions: South (SO), Northeast (NE), North Central (NC), and West (WE). Since we are checking whether they are significantly different from each other or not, we performed two-sided inferences.

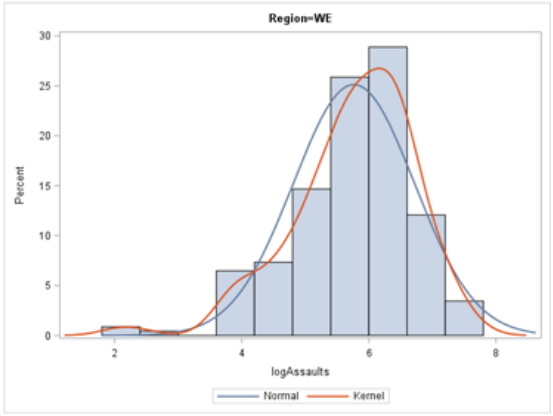
c)

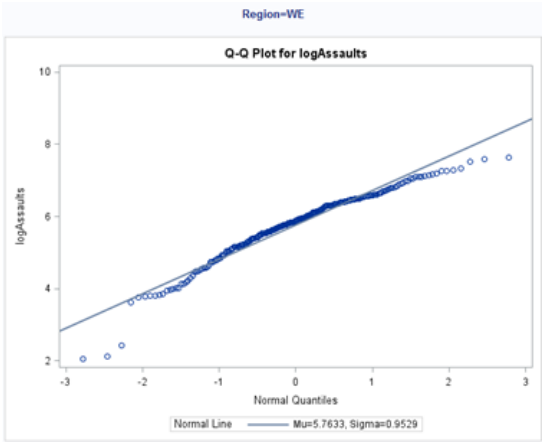
- Assumption 1
  - As we have assumed throughout this semester, all data in USData are SRSs.
- Assumption 2
  - As mentioned in part C and D (one-sample and two-sample), log transformation is used in order to make the data set at least approximately normal distribution.
  - Histograms and qq-plots for the variable logAssaultsPerPopulation for each region indicate that each data set has at least approximately normal distribution.
  - Region NC
  - 

- 
Region=NC
Q-Q Plot for logAssaults
logAssaults
Normal Quantiles
Normal Line
Mu=4.8705, Sigma=1.1536
- Region NE
- 
Region=NE
Percent
logAssaults
Normal
Kernel
- 
Region=NE
Q-Q Plot for logAssaults
logAssaults
Normal Quantiles
Normal Line
Mu=4.7439, Sigma=1.1562

- Region SO
  - 

  - 

- Region WE
  - 

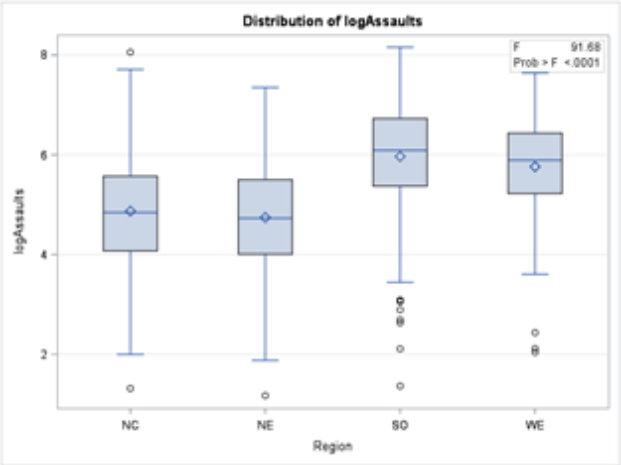
- 


- Assumption 3
  - Empirical rule of thumb
  - The MEANS Procedure

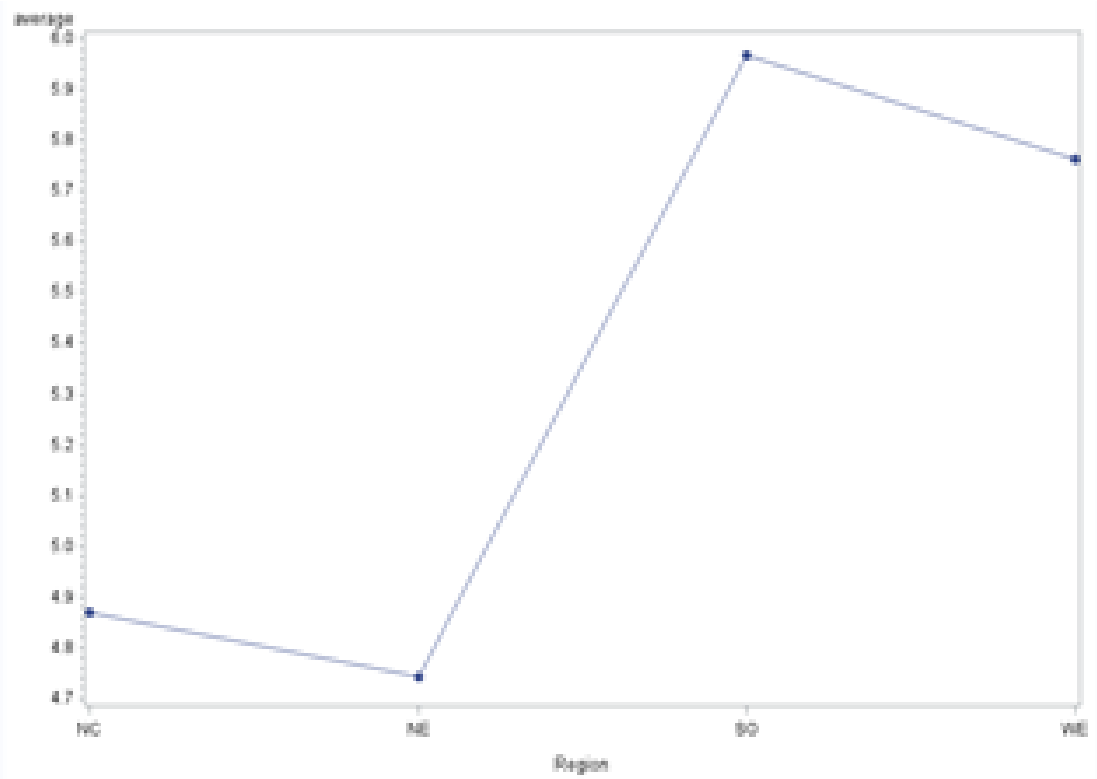
| Analysis Variable : logAssaults | | | | | | |
|---|---|---|---|---|---|---|
| Region | N Obs | N | Mean | Std Dev | Minimum | Maximum |
| NC | 260 | 256 | 4.8705234 | 1.1535700 | 1.3164082 | 8.0570005 |
| NE | 299 | 296 | 4.7438841 | 1.1561632 | 1.1724821 | 7.3515818 |
| SO | 307 | 306 | 5.9674587 | 1.0471244 | 1.3635374 | 8.1565504 |
| WE | 232 | 232 | 5.7633490 | 0.9528754 | 2.0425182 | 7.6417661 |

  - $\frac{Smax}{Smin}=\frac{1.1561632}{0.9528754}=1.21334143 \leq 2$

d)



Box plot for log(AssaultsPerPopulation) of each region, in the order of NC, NE, SO, and WE from the left.




Effects plot, in the order of NC, NE, SO, and WE from the left.

The MEANS Procedure

| Analysis Variable : logAssaults | | | | | | |
|---|---|---|---|---|---|---|
| Region | N Obs | N | Mean | Std Dev | Minimum | Maximum |
| NC | 260 | 256 | 4.8705234 | 1.1535700 | 1.3164082 | 8.0570005 |
| NE | 299 | 296 | 4.7438841 | 1.1561632 | 1.1724821 | 7.3515818 |
| SO | 307 | 306 | 5.9674587 | 1.0471244 | 1.3635374 | 8.1565504 |
| WE | 232 | 232 | 5.7633490 | 0.9528754 | 2.0425182 | 7.6417661 |

Standard deviations of each region are very similar to each other, less than 0.3 of difference. According to the box plot, regions SO and WE have relatively higher mean value than the rest. Effects plot also indicates the same.

e)

One-way ANOVA test

**Step 1**

$\mu_{NC}$= population mean log(Assaults per population) in NC

$\mu_{NE}$= population mean log(Assaults per population) in NE

$\mu_{SO}$= population mean log(Assaults per population) in SO

$\mu_{WE}$= population mean log(Assaults per population) in WE

**Step 2**

$H_0$: $\mu_{NC}=\mu_{NE}=\mu_{SO}=\mu_{WE}$

$H_a$: at least one $\mu_i$ is different from the rest

**Step 3**

$F_{ts}$=91.68

df1=3

df2=1986

p-value<0.0001

**Step 4**

p-value<α

Therefore, $H_0$ is rejected, suggesting at least one $\mu_i$ is different from the rest.

Tukey method for comparison

| Comparisons significant at the 0.05 level are indicated by ***. | | | | |
|---|---|---|---|---|
| Region Comparison | Difference Between Means | Simultaneous 95% Confidence Limits | | |
| SO - WE | 0.20411 | -0.03886 | 0.44707 | |
| SO - NC | 1.09694 | 0.86054 | 1.33333 | *** |
| SO - NE | 1.22357 | 0.99604 | 1.45111 | *** |
| WE - SO | -0.20411 | -0.44707 | 0.03886 | |
| WE - NC | 0.89283 | 0.63984 | 1.14582 | *** |
| WE - NE | 1.01946 | 0.77474 | 1.26419 | *** |
| NC - SO | -1.09694 | -1.33333 | -0.86054 | *** |
| NC - WE | -0.89283 | -1.14582 | -0.63984 | *** |
| NC - NE | 0.12664 | -0.11157 | 0.36485 | |
| NE - SO | -1.22357 | -1.45111 | -0.99604 | *** |
| NE - WE | -1.01946 | -1.26419 | -0.77474 | *** |
| NE - NC | -0.12664 | -0.36485 | 0.11157 | |

$\bar{x}_{NE}$ $\bar{x}_{NC}$ $\bar{x}_{WE}$ $\bar{x}_{SO}$

f)

This study was constructed to compare the average assaults per population (per 100,000) of each region, SO, NE, NC, and WE. The results indicate that the average assaults per population is different in each regions, with 95% confidence level. More specifically, average assaults per population in NE and NC were similar to each other, and significantly smaller than that in WE and SO, which were also similar to each other.

## F. Final conclusion

According to our results, FBI and USData have at least approximately consistent information to each other. There were a small statistically significant difference between the mean values provided each data, but this is believed to be due to an enormous sample size. Additionally, the hypothesis test concluded that their mean values are different, but the 95% confidence interval for ($\mu_{rural}$-$\mu_{urban}$) included 0, therefore the difference is not significant. Finally, regions WE and SO seemed to have a higher values than regions NE and NC. Practically speaking, region WE had the highest minimum value which were greatly above the rest, and region SO had the highest maximum value. To sum up, US data, which is consistent with FBI and is a reliable source, suggests that urban and rural areas have approximately similar mean assaults per population values, but regions WE and SO have significantly higher mean assaults per population value and need further security development.