Lab 2 (100 points + 10 points BONUS) - Describing Distributions with Graphs and Numbers
Objectives: Numerical Summaries, Histograms, and Boxplots

Remember to use the cleaned data set that you generated in Lab 1.

A. (10 points) Online Prelab

B (45 pts) Average Test Scores (Data Set: USData cleaned) We are interested in the graphical and numeric summaries for the adjusted average test score for a college admission exam (TestScore).

1. 10 points) Code. The code is the script (commands) either in R or in the Editor in SAS. Remember that you need to include the code or procedure for inputting the data into your software package.

Solution:

You can read in the data using the SAS interface.

```
File --> Import Data --> Tab Delimited File (.txt) --> Next > -->
browse for the file --> Library: Work, Member: USData_cleaned (or an
appropriate name) --> Finish
```

If you use this method, the data statements that use this file must be prefaced by WORK.

```
/* ********************************** */
/* STAT 350 Spring 2018 Lab 2         */
/* ********************************** */
** Read in data;
data USData;
  infile "W:\USData_cleanSAS.txt" delimiter = '09'x firstobs = 2;
  length IncomeCategory $ 11
  input State $ Region $ CountyIndex $ UrbanIndicator $Population
      LandArea PopulationDensity PercentMaleDivorce
      PercentFemaleDivorce MedianIncome IncomeCategory
      PercentCollegeGraduates MedianHouseAge
      RobberiesPerPopulation AssaultsPerPopulation
      BurglariesPerPopulation LarceniesPerPopulation
      EducationSpending EducationSpendingP2 TestScore;
run;
** Print to make sure it read in properly...;
proc print data = USData(obs = 6); run; *This is optional;
** 2), 7) Summary Statistics on TestScore;
proc univariate data = USData;
  var TestScore;
run;
** 3) Print the outliers optional;
proc print data = USData;
  var State CountyIndex TestScore;
  where TestScore > 1763.245 OR TestScore < 1443.525;
run;
```

```
** 4) Boxplot;
data USData1;
  set USData;
  index = 1;
run;
proc boxplot data = USData1;
  plot TestScore * index / boxstyle = schematic idsymbol = circle;
run;
** 5) Histogram;
proc sgplot data = USData;
  histogram TestScore;
  density TestScore;
  density TestScore / type = kernel;
run;
```

2. (2 points) Find the five-number summary for these data.

Solution:

Quantiles (Definition 5)	
Quantile	**Estimate**
100% Max	1754.28
99%	1735.32
95%	1699.64
90%	1679.07
75% Q3	1643.35
50% Median	1604.04
25% Q1	1563.42
10%	1530.67
5%	1509.40
1%	1452.44
0% Min	1377.15

Min: 1377.15 Q_1: 1563.42 Median: 1604.04 Q_3: 1643.35 Max: 1754.28

3. (5 points) Calculate the 1.5 IQR upper and lower limits for the outliers. Are there any outliers according to the 1.5 IQR rule (just answer yes or no, and explain why you know)? This part may be done by hand. If done by hand, all work needs to be provided. If done via computer code, then the code must be listed.

Solution:

$IQR = Q_3 - Q_1 = 1643.35 - 1563.42 = 79.93$

$LowerFence = Q_1 - 1.5(IQR) = 1563.42 - 1.5(79.93) = 1443.525$
$UpperFence = Q_3 + 1.5(IQR) = 1643.35 + 1.5(79.93) = 1763.245$

The minimum 1377.15 is less than the lower limit of 1443.525, we know there is at least one outlier. We can get more specific information with SAS:

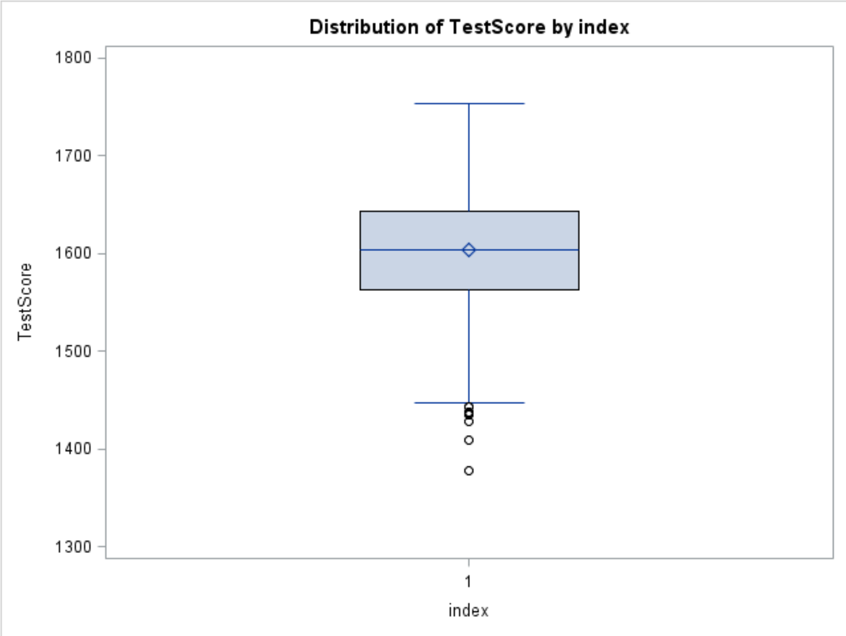
```
** Print the outliers;
proc print data = USData;
  var State CountyIndex TestScore;
  where TestScore > 1763.245 OR TestScore < 1443.525;
run;
```

Obs	State	CountyIndex	TestScore
3	Alabama	3	1436.83037
946	Texas	12	1435.145507
1055	West Virg	2	1437.173229
1057	West Virg	4	1443.277933
1058	West Virg	5	1377.150907
1059	West Virg	6	1443.492787
1060	West Virg	7	1427.777812
1062	West Virg	9	1408.522925

It looks like 6 counties in West Virginia are outliers on the low end. Plus one from Alabama and one from Texas.

4. (5 points) Make a modified boxplot. Describe the distribution by stating whether it is symmetrical, left, or right skewed, and if there are outliers. Please indicate which features of the plot you used to classify the type of skewness and determine whether outliers are present.

Solution:

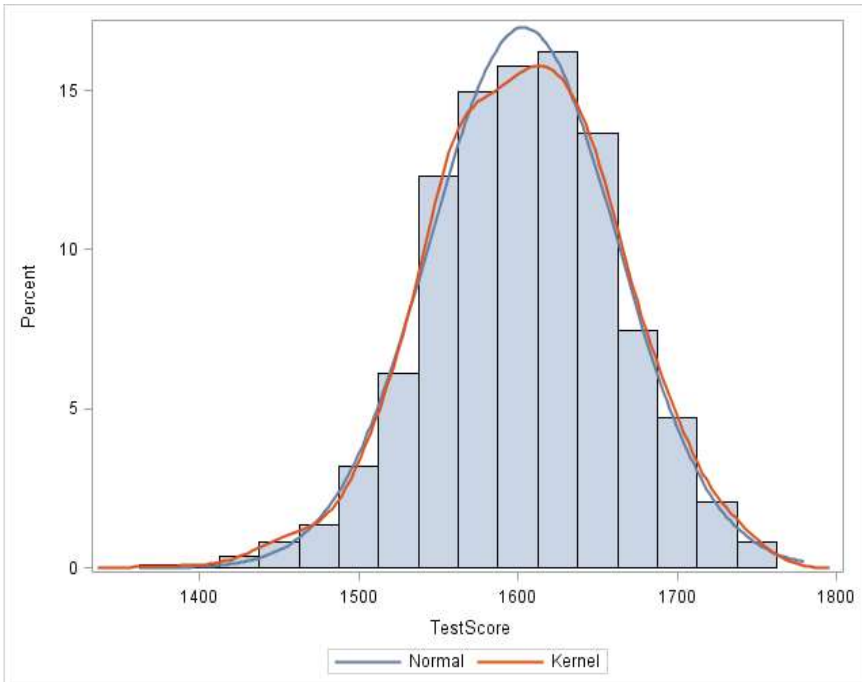


The distribution is approximately symmetric; the boxplot does not stretch out in one direction more than another direction.

It looks like there is no outlier at a higher value and at least six outliers at lower values. It is hard to tell the number because one of the circles looks bold which means that there is more than one point there. This is consistent with the optional work shown in Question 3.

5. (5 points) Make a histogram of the data. Describe the distribution by stating whether it is symmetrical, left, or right skewed, and if there are outliers. Please indicate which features of the plot you used to classify the type of skewness and determine whether outliers are present.

Solution:



The distribution looks symmetric though it has a slight tail at lower values. The distribution exhibits a very similar shape on both sides of its mode, suggesting symmetry. In addition, the two curves are close. I do not see any points that are not close to the rest of the data; therefore, there are no outliers.

6. (5 points) Are the data points that you considered outliers in the histogram and the boxplot the same or different? If they are different, please provide a possible explanation for the difference.

Solution:

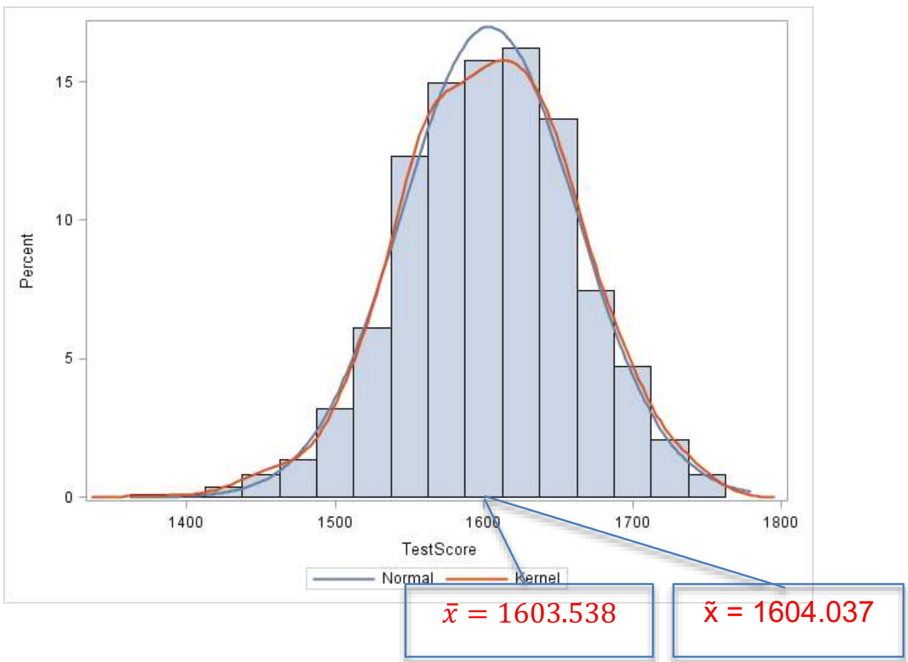
They are different. I identified more than six outliers using the boxplot and no outlier(s) using the histogram. One possible reason for the difference is that the boxplot employs a “rule of thumb” based on the IQR and its criterion for identifying outliers is not guaranteed to align with one’s impression obtained from visually inspecting a histogram. In addition, the outliers are not far above or below the fence. It is not surprising that different methods will suggest different categorizations of these “boundary” points.

7. (5 points) Obtain the sample mean, $\bar{x}$, and the sample standard deviation, s. Indicate the location of the mean and median on the histogram in Question 5. This may be done by hand or computer software. $\bar{x}$ is the average of the average values and the standard deviation of the average values for the counties in the sample.

Solution:

Basic Statistical Measures			
Location		Variability	
Mean	1603.538	Std Deviation	58.74789
Median	1604.037	Variance	3451
Mode	.	Range	377.12514
		Interquartile Range	79.92726

$\bar{x} = 1603.538$, $s = 58.74789$. Also, the median is 1604.037 (the same as in the other table).



8. (3 points) Do you think the median is close to the mean? Please explain your rationale.

Solution:

Yes, the median is very close to the mean both my just looking at the numbers, where they are placed on the histogram and looking at the boxplot.

$$\frac{mean - median}{maximum - minimum} = \frac{1603.538 - 1604.037}{1754.28 - 1377.15} = -0.001323$$

$$\frac{mean - median}{standard\ deviation} = \frac{1603.538 - 1604.037}{58.74789} = -0.00849$$

9. (5 points) If you only had one measure to describe the central location of the distribution of TestScore, which would you choose? Please explain your answer.

Solution:

I would choose the mean. Both the mean and median are appropriate when the distribution is approximately symmetric. However, the mean is easier to calculate so is usually used in symmetric situations.

C (45 pts). Larcenies (Data Set: USData Cleaned). We are interested in the graphical and numeric summaries of number of larcenies (thefts of personal property) out of 100,000 people (LarceniesPerPopulation).

1. (10 points) Code

Solution:

```
** Summary Statistics on LarceniesPerPopulation;
proc univariate data = USdata;
  var LarceniesPerPopulation;
run;
** Determine the number of outliers;
data count;
  set USData;
  if LarceniesPerPopulation > 7925.225;
run;
proc means data = count;
  var LarceniesPerPopulation;
run;
** Boxplot;
* You do not need to remake the new data set if you have already run
   the code in part B;
data USData1;
  set USData;
  index = 1;
run;
proc boxplot data = USData1;
  plot LarceniesPerPopulation * index / boxstyle = schematic idsymbol =
circle;
run;
** Histogram;
proc sgplot data = USData;
  histogram LarceniesPerPopulation;
  density LarceniesPerPopulation;
  density LarceniesPerPopulation / type = kernel;
run;
```

2. (2 points) Find the five-number summary.

Solution:

Quantiles (Definition 5)	
Quantile	**Estimate**
100% Max	12274.59
99%	8445.63
95%	6977.86
90%	5911.06
75% Q3	4370.17
50% Median	3071.26
25% Q1	2028.42
10%	1371.66
5%	1063.13
1%	685.96
0% Min	170.16

Min: 170.16 Q_1: 2028.42 Median: 3071.26 Q_3: 4370.17 Max: 12274.59

3. (5 points) Calculate the 1.5 IQR upper and lower limits for the outliers. Are there any outliers according to the 1.5 IQR rule (just answer yes or no, and explain why you know)? This part may be done by hand. If done by hand, all work needs to be provided. If done via computer code, then the code must be listed.

Solution:

$$IQR = Q_3 - Q_1 = 4370.17 - 2028.42 = 2341.75$$

$$LowerFence = Q_1 - 1.5(IQR) = 2028.42 - 1.5(2341.75) = -1484.205$$
$$UpperFence = Q_3 + 1.5(IQR) = 4370.17 + 1.5(2341.75) = 7882.795$$

Since the Lower Fence is negative and the data has to be positive, there cannot be lower outliers according to this rule. However, since the max of 12274.59 is greater than the 7882.795, there is at least one outlier with a higher value.

You can use SAS to determine the number of outliers.

```
** Determine the number of outliers;
data count;
  set USData;
  if  LarceniesPerPopulation > 7882.795;
run;
proc means data = count;
  var LarceniesPerPopulation;
run;
```

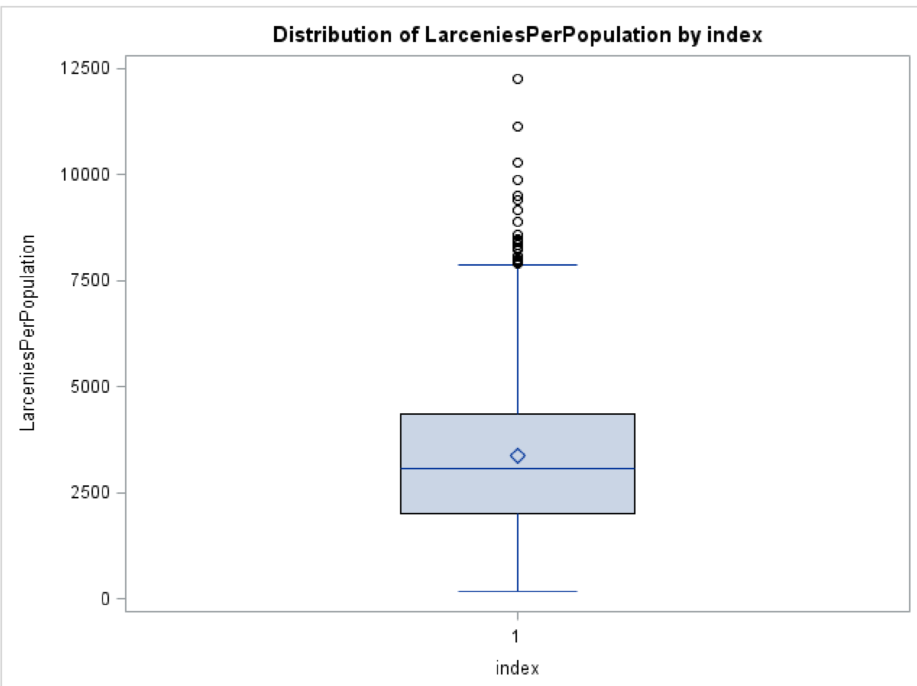
We will be discussing these commands later in the semester.

Analysis Variable : LarceniesPerPopulation				
N	Mean	Std Dev	Minimum	Maximum
25	8794.46	1097.22	7898.68	12274.59

There are 25 outliers for this variable.

4. (5 points) Make a modified boxplot. Describe the distribution by stating whether it is symmetrical, left, or right skewed, and if there are outliers. Please indicate which features of the plot you used to classify the type of skewness and determine whether outliers are present.

Solution:

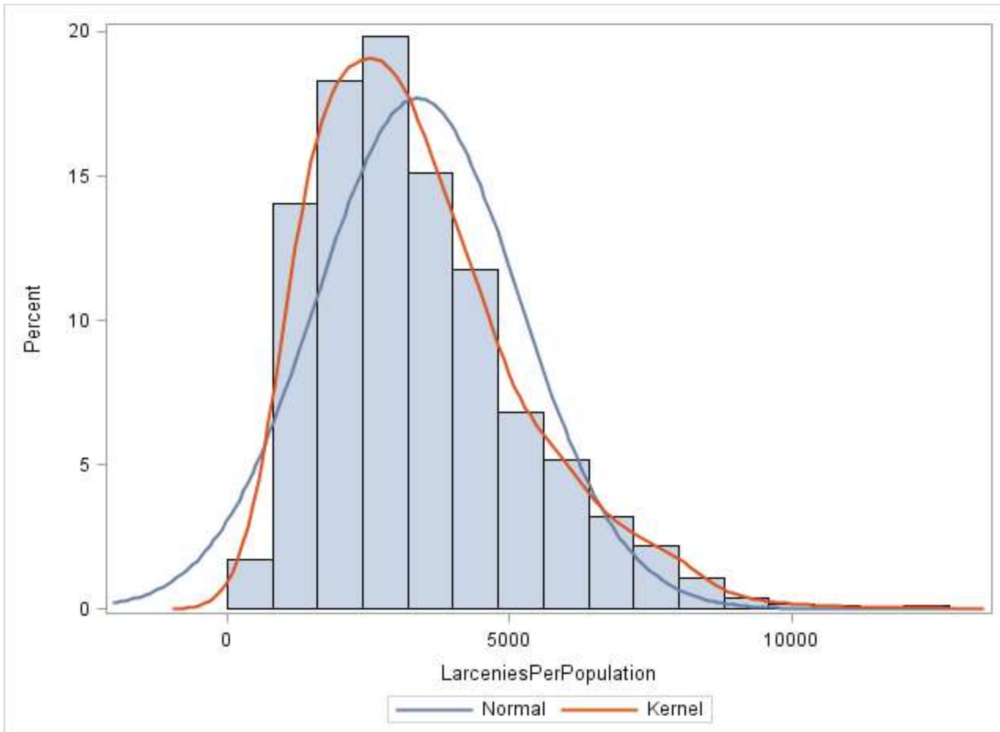


The distribution is right skewed. The boxplot seems more "stretched out" above the median than below. Moreover, the mean is slightly larger than the median.

There appear to be many outliers, as evidenced by the many dots above the upper whisker.

5. (5 points) Make a histogram of the data. Describe the distribution by stating whether it is symmetrical, left, or right skewed, and if there are outliers. Please indicate which features of the plot you used to classify the type of skewness and determine whether outliers are present.

Solution:



The distribution is right skewed, as can be seen by the long right tail. Remember that SAS automatically determines the scale; therefore, if the range is larger than you see points, there have to be data there.

I would suspect that there are points far from the rest of the data even though they can't be seen.

6. (5 points) Are the data points that you considered outliers in the histogram and the boxplot the same or different? If they are different, please provide a possible explanation for the difference.

Solution:

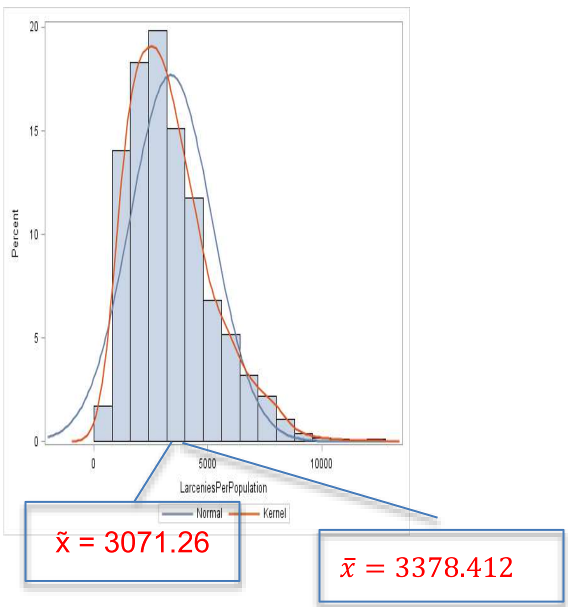
They are different. One possible reason for the difference is that the boxplot employs a "rule of thumb" based on the IQR and its criterion for identifying outliers is not guaranteed to align with one's impression obtained from visually inspecting a histogram. In addition, some outliers above the upper inner fence are not far above the fence. It is not surprising that different methods will suggest different categorizations of these "boundary" points.

7. (5 points) Obtain the sample mean, $\bar{x}$, and the sample standard deviation, s. Indicate the location of the mean and median on the histogram generated in Question 5. This may be done by hand or computer software.

Solution:

Basic Statistical Measures			
Location		Variability	
Mean	3378.412	Std Deviation	1804
Median	3071.260	Variance	3255203
Mode	.	Range	12104
		Interquartile Range	2342

$\bar{x} = 3378.412$, $s = 1804$. Also, the median is 3071.26 (the same as in the other table).



8. (3 points) Do you think the median is close to the mean? Please explain your rationale.

Solution:

Yes, the median looks close to the mean but not as close as in part B.

$$\frac{mean - median}{maximum - minimum} = \frac{3378.412 - 3071.26}{12274.59 - 170.16} = 0.0254$$

$$\frac{mean - median}{standard\ deviation} = \frac{3378.412 - 3071.26}{1804} = 0.17$$

Though they look close and the percentage difference with respect to the range is small, the percentage with respect to the standard deviation is fairly large.

This emphasizes the problems of using the comparison of mean and median to determine the skewedness of a distribution. It is much better just to look at the histogram or boxplot to determine skewedness.

9. (5 points) If you only had one measure to describe the central location of the distribution of LarceniesPerPopulation, which would you choose? Please explain your answer.

Solution:

I would use the median since the data is right skewed. The mean is only a good measure if the distribution is close to being symmetric.

D. (10 points) BONUS. We do not discuss how to make graphs of categorical variables in this class; however, this is very important. Make a pie chart for Region. Note: You will not get coding help for bonus questions.

1. (5 points) Code.

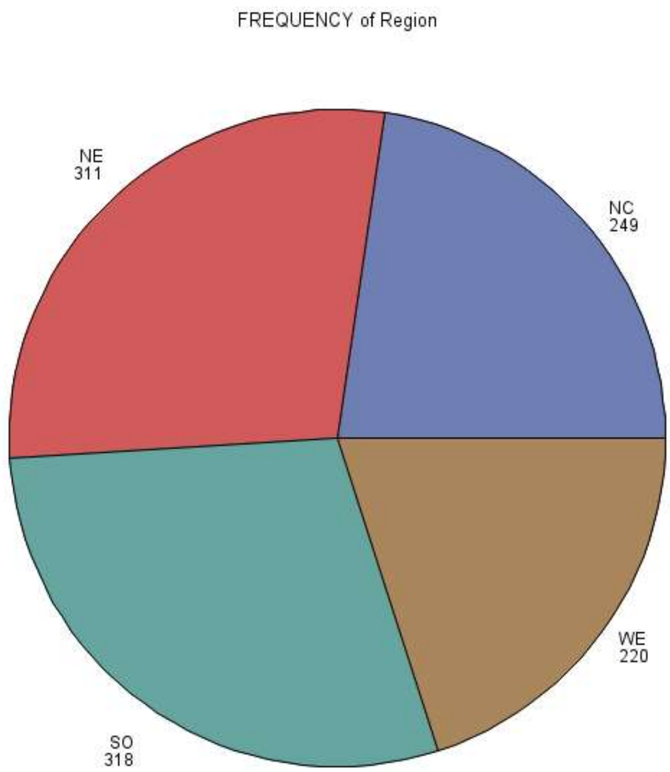
Solution:

```
proc gchart data=USData;
  pie Region;
run;
```

2. (5 points) If the labels are abbreviations which are not clear, state what each label represents. What information did you learn from the pie chart? Please explain your answer.

Solution:

If you provide all of the information required above, you will receive full credit.



Below is the correspondence between the abbreviations and the region names:
NC: North Central
NE: North Eastern
SO: South
WE: Western

Approximately one-fourth of the counties are in each region (which is good ☺). However, the North East and South regions have slightly more counties and the North Central and West have slightly fewer counties. This make sense because we are sampling by state and the states are smaller in the North East and South regions and larger in the North Central and West regions.