In this test the data we will be using tracks the rate of deaths due to heart disease. Different regions throughout the United States will be considered in order to answer the question, “Is there a significant difference between different regions of the country and their corresponding heart rates?”

The regions include all states and are split in regions using the Census Bureau’s region and division assignments. The District of Columbia is also included with the southern region. The data includes the total number of deaths by heart disease and the rate of death per 100,000 people in each state. First we will take a look at some basic summary statistics and graphics.

|  |  |  |  |
| --- | --- | --- | --- |
| Census Bureau Regions | | | |
| Northeast | **Midwest** | **South** | **West** |
| Connecticut | Indiana | Delaware | Arizona |
| Maine | Illinois | District of Columbia | Colorado |
| Massachusetts | Michigan | Florida | Idaho |
| New Hampshire | Ohio | Georgia | New Mexico |
| Rhode Island | Wisconsin | Maryland | Montana |
| Vermont | Iowa | North Carolina | Utah |
| New Jersey | Kansas | South Carolina | Nevada |
| New York | Minnesota | Virginia | Wyoming |
| Pennsylvania | Missouri | West Virginia |  |
|  | Nebraska | Alabama |  |
|  | North Dakota | Kentucky |  |
|  | South Dakota | Mississippi |  |
|  |  | Tennessee |  |
|  |  | Arkansas |  |
|  |  | Louisiana |  |
|  |  | Oklahoma |  |
|  |  | Texas |  |
| **Table 1: Contains each state and the region to which it belongs** | | | |

Table 1 below includes the averages of rates for each state per region and also the averages of all observations. We can compare each region to the overall statistics in order to try to gather some insight into the data. We can see right away that the western region not only contains the minimum observation, but it is lower than overall in every statistic. The northeast and southern regions seem to be higher on average, and the Midwestern region is very close to the overall averages. The next figure, a histogram, tells the same story. West is visibly lower that the other regions. It is also the box with the smallest quartile range, thought there are some outliers. It is worth noting that this region also has the least number of observations.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Northeast | Midwest | South | West | All |
| Min. | 179.6 | 107.9 | 152.0 | 96.0 | 96.0 |
| 1st Qu. | 194.7 | 184.3 | 189.7 | 151.0 | 165.5 |
| Median | 207.4 | 192.9 | 218.2 | 161.2 | 194.7 |
| Mean | 207.8 | 194.9 | 213.5 | 155.5 | 193.4 |
| 3rd Qu. | 224.1 | 227.8 | 249.3 | 166.0 | 224.4 |
| Max | 247.6 | 244.1 | 258.0 | 197.4 | 258.0 |
| Table 2: Summary statistics, deaths per 100k due to heart disease | | | | | |

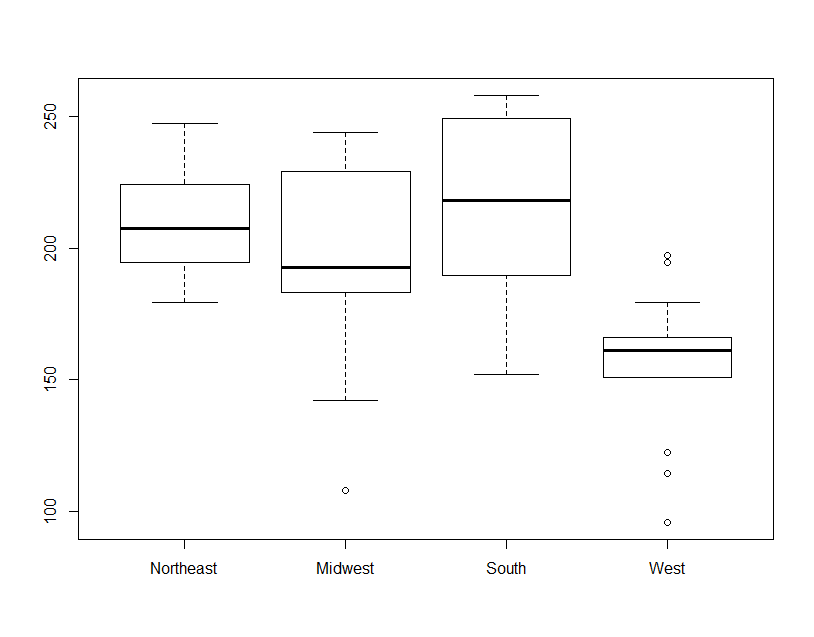


Figure 1: Boxplot comparison of regions

The next figure contains histograms of each region. We can check the normality of our data with histograms. None of the histograms seem to be normally distributed. Three of them are clearly left-skewed. This means that while some regions have similar rates of death amongst states, there are also nearby states with very different rates. The fact that the data is not normal means that nonparametric tests will be our best option.

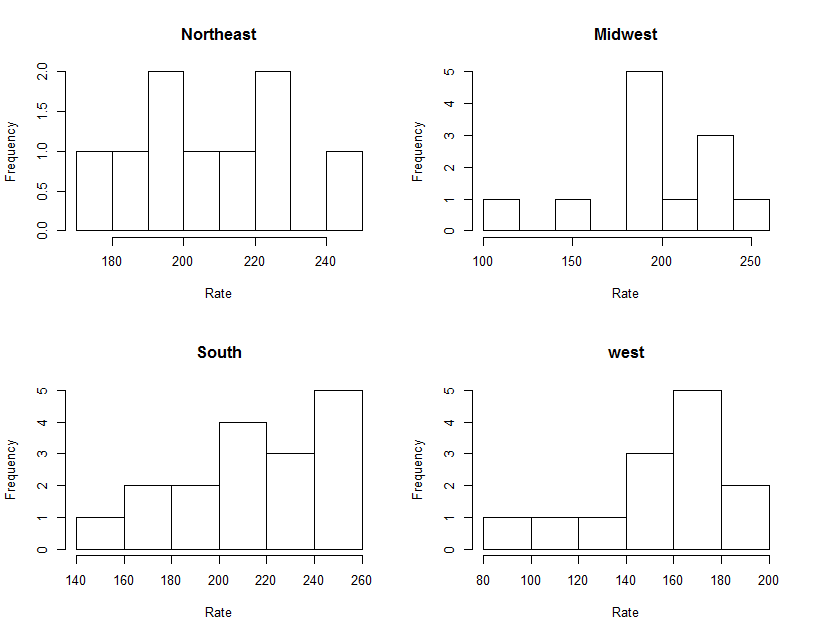


Figure 2: Histogram of each region

The test we will do is a Kruskal-Wallis test. The null hypothesis for the test is that the mean ranks of the different treatments are the same. A low (< 0.05) p-value would mean that at least one group is different from the others. This test does not require normally-distributed data. It does however require that the groups have equal variances. In order to test this assumption we will first use a permutation test on deviances. This test will check each pairwise combination of regions for unequal variance.

Table 3 below contains the results of both tests. Since the minimum p-value of the test on deviances was 0.29, which is not less than 0.05 we can assume that the regions have equal variances. The Kruskal-Wallis test found an extremely low p-value. Since the p-value is less than 0.05 we can reject the null hypothesis. We can conclude that there are differences in the means of the different regions.

|  |  |  |
| --- | --- | --- |
| Test | Statistic | Value |
| Permutation Test on Deviances | min(p-value) | 0.2946154 |
| Kruskal-Wallis Test | Chi-squared | 17.589 |
| p-value | 0.0005346 |
| Table 3: Test results | | |