In the previous question, we used R's wilcox.test to test the hypothesis that the median of the cars$dist data was equal to 100. In this question, we will work through the test by hand to show how the test works. (Note that only a mathematician would be interested in doing this test by hand in general, but it is useful for you to do it by hand once.)

It is recommended that you have your "Explanation" tab of your "Wilcoxon Tests" page of your Math 325 Notebook and follow the five steps for the "One Sample Example." Answer the following questions to test your understanding of these five steps.

Recreate the dot plot used in the previous question by running the code:

```
ggplot(cars) +
  geom_dotplot(aes(x=dist)) +
  geom_vline(xintercept=100) #Adds the vertical line representing the null hypothesis
```

Before beginning **Step 1** of the Wilcoxon Test, we need to calculate the "differences". As stated in your textbook, "The differences are obtained by subtracting the hypothesized value for the [median] from all observations." This could be done in R with the code:

```
> cars$dist - 100
```

As shown in your dot plot there is only one positive difference, this is the single dot to the right of the vertical line in your dotplot. All the other values are negative as they are to the left of the hypothesized value of 100.

What is the value of the single positive difference for this data?

(In other words, how far above 100 is the maximum dot in the plot?)

[20]

What rank will this single positive difference be given? Hint the first eight negative differences (in order of magnitude) are

-7, -8, -15, -16, -20, -24, -30, and -32?

(Remember: if there is a tie between two ranks, then the ranks of the two values are averaged together.)

The value of the single positive rank will be: [5.5]

Thus, the sum of the positive ranks will thus also be: [5.5]

Note that this answer is exactly the same as the value of the test statistic that you obtained in your test in R from the previous question! To see this, rerun the wilcox.test() from the previous question with the code:

```
> wilcox.test(cars$dist, mu = 100, alternative = "two.sided")
```

What letter does the wilcox.test() output in R use for the Wilcoxon Test statistic?

[V]

This shows that the test statistic of the Wilcoxon Test is the sum of the ranks from the group with the fewest ranks.

There are two things needed to calculate the p-value of a hypthesis test. Which of these two things have we calculated so far?

- [x] A test statistic.
- [ ] The distribution of the test statistic (that is found by considering all the possible values that the test statistic could have been, assuming the null hypothesis is true).

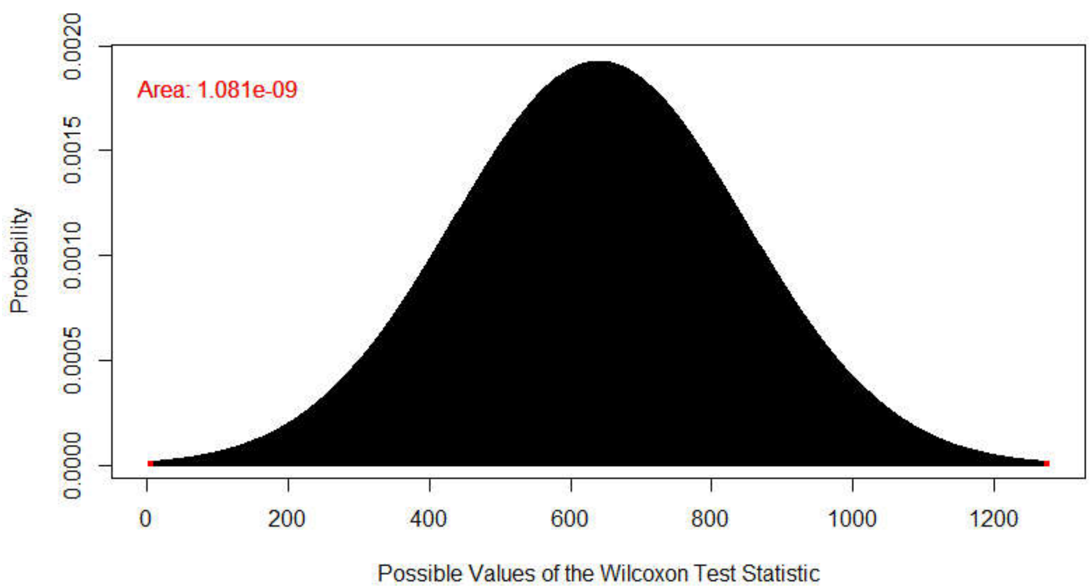
So, to calculate the p-value, we must now come up with all of the possible values that the test statistic could have been, and the probability of each value occurring. This is known as the distribution of the test statistic.

First, note that it could have been possible that none of the values were greater than 100. In this case, the sum of the positive ranks would have been 0

Also, it could have been possible that all of the values were greater than 100. In this case, the sum of the positive ranks would have been 1275 , which is the sum of the ranks 1, 2, 3, ..., 50. You can find this in R with the command:

```
> sum(1:50)
```

This means that the test statistic of the Wilcoxon Test could have been any whole number between 0 and 1275 . That is a lot of possibilities! It is so many possibilities in fact, that drawing the plot of the distribution of all possible test statistics in a way similar to the plots of vertical lines shown in your textbook provides the following image. The lines are so close together it looks like a solid normal distribution. This is an amazing result because the Wilcoxon Test is a nonparametric test, so no assumptions about normality were made. All the same, we shade the region that is as extreme or more extreme than the observed test statistic of 5.5. This results in the following picture where the very tips of the two tails are shaded in red.



Ths, the p-value is calculated by adding up the probabilities of all the possible sums of ranks that are as extreme or more extreme than the observed sum of ranks. This gives a p-value of 1.081e-09