Hw1

1- Which are the top two counties that occur most frequently in your dataset?

2- In the state of Maryland, which two counties occur most frequently in the dataset?

3 - Which are the top two states that occur most frequently in your dataset?

4 - For each state, we can pool all its accidents and then calculate what fraction of them have severity=4. If we do this for every state, which are the top three states in terms of the proportion of severity=4 accidents?

5 - Which are the bottom four states in terms of average temperature of accidents? ("bottom four states" means the four states with lowest average temperature)

Hw 2

1) What is the size of the dataset? (how many datapoints do we have?)

2) What is the mean of this column?

3) What is the standard deviation?

4) What is the maximum value?

5) What is the minimum value?

Calculate the standard scores (also called z-scores) of this column and find:

6) What is the z-score of the largest value?

7) What is the z-score of the smallest value?

Now we investigate the distribution:

8) How many data points are within one standard deviation away from the mean? (that is, how many datapoints are less than mean + sd and greater than mean - sd?

9) What proportion of the datapoints are within one standard deviation away from the mean?

10) What proportion of the datapoints are within two standard deviations away from the mean?

Hw 3

Consider the standard normal distribution:

1) what is the threshold for the bottom 9%? That is, what is the z-value, to the left of which, the area under the curve is 9%?

2) what is the threshold for the top 7%? That is, what is the z-value, to the right of which, the area under the curve is 7%?

3) what is the threshold for the middle 94% around the mean? That is, what is the z-value such that the area under the curve between -z and +z equals 94%?

4) what is the threshold for 8% on the two tails? That is, what is the z-value such that the area outside the region between -z and +z equals 8%?

5) what is the area under the curve to the left of z=1.7?

6) what is the area under the curve to the right of z=2.1?

7) what is the area to the left of z= -1.7?

8) what is the area outside the region between z= -1.3 and z= +1.3?

Now consider the t distribution with df=15:

9) what is the threshold for the bottom 9%? That is, what is the t-value, to the left of which, the area under the curve is 9%?

10) what is the threshold for the top 7%? That is, what is the t-value, to the right of which, the area under the curve is 7%?

11) what is the threshold for the middle 94% around the mean? That is, what is the t-value such that the area under the curve between -t and +t equals 94%?

12) what is the threshold for 8% on the two tails? That is, what is the t-value such that the area outside the region between -t and +t equals 8%?

13) what is the area under the curve to the left of t=1.7?

14) what is the area under the curve to the right of t=2.1?

15) what is the area to the left of t= -1.7?

16) what is the area outside the region between t=-1.3 and t=+1.3?

Now consider the t distribution with df=1500:

17) repeat question 9: what is the threshold for the bottom 9%? That is, what is the t-value, to the left of which, the area under the curve is 9%?

18) repeat question 14: what is the area under the curve to the right of t=2.1?

Consider the dataset of car accidents that you worked with in the previous homeworks.

19) what is the mean of the Temperature column?

Now let's take the first 20 rows to be our sample.

20) what is the sample mean?

21) what is the 95% confidence interval for the mean?

22) what is the 99% confidence interval?

23) what is the 90% confidence interval?

Now let's take the first 200 rows to be our sample.

24) what is the sample mean?

25) what is the 95% confidence interval?