Hugo Escobar

Final Project Report

The dataset I have chosen is New Jersey’s Top 50 High Schools Based on Their SAT scores. The dataset I use ranks these schools by combining both the math scores and their reading scores. The dataset also separates the scores while listing these individual schools. It also includes the county. In this report, I will be applying every technique we have learned in this class. The techniques are mean, median, mode, standard deviation, variance, combination, permutation, unions, intersections, compliment, binomial probability distribution, binomial expected, binomial variance, binomial standard deviation, geometric probability distribution, geometric expected, geometric variance, geometric standard deviation, hypergeometric probability distribution, hypergeometric expected, hypergeometric variance, hypergeometric standard deviation, Poisson probability distribution, Poisson expected, Poisson variance, Poisson standard deviation, Tchebyshev’s, uniform probability distribution, uniform expected, uniform variance, and the uniform standard deviation.

The first thing I will be doing is the mean, median, mode, standard deviation, and variance.



In the image above you can see that I used the stats library implemented from this class to test my answers. (I have also used calculators online to verify these are correct). So the average SAT score for all 50 schools is 1293 if you decide to round up.

Next is the median, the stats library says the median is 1266. 

As seen in the image above, the median is 1266. It is quite literally in the middle of the 50 schools thus making it the median.

Next up is the mode, the mode of this dataset is also 1266. 

This combined score occurred the most, a total of three times, while also being in the middle of the dataset. This would make it the mode as well as the median.

Next is the Standard deviation as well as the variance

As seen in the image above, the standard deviation of all 50 scores is 77.867 and the variance is 6063.239. This was pretty simple to check since the standard deviation is just the square root of the variance.

Next is the combinations, permutations, unions, intersections, and the compliments.

For combinations I decided to find out how many combinations were possible in schools from both Camden County and Somerset County.

As seen in the image above, the answer is 7 combinations. 7 of the schools are from Somerset County and only 1 school in the dataset is from Camden County.

Next was the permutations, for this I asked the question, “How many different ways can I list these schools from Somerset County?”



In the image above, it shows that there are 5040 ways of making an order of these schools, specifically all in the dataset that are in Somerset County.

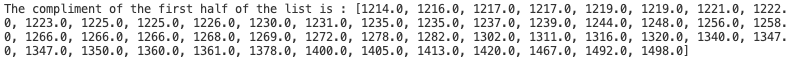
For the next three sections, I split the dataset into two lists, one containing the first 25 schools and the second list containing the rest of the schools.

First came the union, where I simply just combined the two lists.

Here you can clearly see that the union worked as it combined every possible SAT score from the dataset. Unfortunately, it removes duplicates.

Next is the intersection. 

It seems that both of these lists share 1266 as a value. This makes sense because it is the median and also the mode. If you were to split the entire list in two, the middle element would be 1266 and both lists would have 1266 because it shows up three times in the dataset.

Finally, the compliment.

For this example, I used the compliment of the first half of the dataset. To find this, I took the complete dataset and compared it to the first half arraylist. If the arraylist did not match with what it was being compared to in the complete arraylist, then it would be outputted on the screen, thus showing what values where in the complete dataset and not in the first half arraylist.

The next section contains all the distributions, their expected, variance, and standard deviation.

First is the binomial distribution. Here are the results. The question I asked was “What are the odds of getting four schools from Somerset County if you chose randomly from the entire list?”



The percent chance of choosing four schools from Somerset County is 8.58%. I found this by finding the total, the percent of choosing a Somerset County school, the percent of not choosing a Somerset County school, and the amount being selected so it is four in this case. For the expected all that was needed was the Somerset County percent to give you 7.14% which is fairly close to the actual answer. The variance of this being 43.8 and the standard deviation is 6.6.

Next is the geometric distribution. For this I asked, “What are the odds of picking a Somerset County school out of the list on the 5th try?”



For this question, I used the probability of picking a Somerset County school, the probability of not picking a Somerset County school, and the amount of

tries it takes. The answer being 7.65%. The expected being 7.14. The variance is 43.877 and the standard deviation is 6.62.

For the next distribution, it was the hypergeometric one. I asked, “What are the odds of choosing two Monmouth County Schools that have a combined score over 1300?”



For this I used the total number of schools, the amount of Monmouth County schools, the amount of Monmouth County schools with a score over 1300, and the amount being chosen. After using the stats library, the answer is 12.23%. The expected was .7% so it was off by a lot. The variance was .55 and the standard deviation was .74.

Next was the Poisson distribution. For this I had to make a hypothetical rate for lambda. So the question I asked was, “If 50 representatives for the military visit a top school each year in New Jersey, what are the odds of all schools within the dataset get a visit this year?”



For this all I needed was the amount of schools being tested and the rate of representatives visiting. After doing the calculations the percent was 5.63%. The expected was 50 schools being visited. The variance was also 50 and the standard deviation was 7.07.

Next was the application of Tchebyshev’s theorem. I asked, “What are the odds of picking a school within the upper half of the list?”



As seen above the odds of picking a school from the upper half of the list is 65.97%. I made the upper limit the middle of the list and the lower limit the first item of the list.

Finally, I used the uniform probability distribution on the dataset. The questions I asked was “What are the odds of picking the last 5 schools in the list?”



Using the uniform distribution, I got an answer of 8.1%.

This concludes the section of the report that displays my research on the dataset I found. Next is the images and code for the Matlab section. Using matlab I read in the salted data in the csv file from project 2. This csv file has 1000 inputs and the y values should be different numbers since they are salted. Here is what the salted graph looks like in matlab.Chart

Description automatically generated

As you can see it is just a jagged line going in a diagonal direction. This is because it is based off the slope graph. Here is what the code looked like:Text

Description automatically generated

It reads in the csv file as an array object, sets col1 to the first column in the csv file and col2 as the second column. The it plots the columns and gives the appropriate labels and title.

Next is the salter plotter in matlab. It is similar to the salter graph and code but all it does is smooth the data before the graphing. Here is what the graph looks like:

Chart, line chart

Description automatically generated

Compared to the salter graph, you can see that it is a lot more “smoother” so it definitely worked. Here is the code for it:

Text

Description automatically generated

This is extremely similar to the code used in the salter but just has a smoother method to smooth out the data then plot it.

Next is the JFree chart graphs. First is the salted data graph.

Chart, line chart

Description automatically generated

Although it may be a little hard to see, you can see the jagged line and see that the data is salted like in the matlab graphs.

This is the smoother graph that is a little harder to see you can see that the line is smoothed out and going in the correct direction.

Chart, line chart

Description automatically generated