[redacted]

Intro to Prob. and Stats.

Dr. Jacob

June 8, 2020

Solar Power in the United States.

This data that I found is information about the different solar projects in the U.S and was found in Dr. Jacob's github (https://github.com/nurfnick/Data_Sets_For_Stats) under the Solar_Project_US.xlsx. The thing that really intrigued me about the data set was that it was about a cleaner energy source. Being in Oklahoma where the main source of energy comes from natural gas, seeing that there is some solar energy being used to try and lower the need and usage of natural gas (Table 1.1) is really good news to hear because it is taking a step towards a cleaner environment for the future generations. On Table 1.2 is descriptions of the variable that I chose to use in my analysis. The table that is in the Excel File is bigger, but I chose to not use the longitude and latitude in my analysis. Project Name, City, State, Utility, and Co-located Projects are categorical variables and System Size and Year of Interconnection are quantitative variables, interval and ordinal respectively. The thing that I am hoping to get out of this analysis is understanding of how much solar power is needed to power a community, and if we can eventually lead to a solar powered lifestyle, resulting in a cleaner environment, free from oil and natural gas pollution.

Project Name	City	State	Utility	System Size (kW)	Year of interconnection	Co-located projects
OGE 10 MW Facility	Covington	OK	Oklahoma Gas and Electric Co.	15723.27	2018	NA
TCEC Community Solar	Hooker	OK	Tri-County Electric Cooperative	1536	2016	NA
Mustang OGE Solar Farm South	Mustang	OK	Oklahoma Gas and Electric Co.	2000	2015	NA
Mustang OGE Solar Farm North	Mustang	OK	Oklahoma Gas and Electric Co.	5000	2015	NA
ECOEC Community Solar	Okmulgee	OK	East Central Oklahoma Electric Cooperative	299.25	2017	NA

Table 1.1

Project Name	Project name, if applicable
City	City where system is sited
State	State
Utility	Utility service territory
System Size (kW)	System capacity in kilowatts. Most project capacities should represent rated capacity in kilowatts DC, but errors may exist.
Year of Interconnection	Estimated year that project began serving customers.
Co-located Projects	Some community solar arrays are co-located components of a larger project. If applicable, this field reflects a single project name to link such co-located projects.

Table 1.2

Project Part 2

The frequency table (Table 2.2) is a table that is showing the relative frequencies of states that are using solar power. The two-way table below (Table 2.1) is describing some of the different utilities that have started projects, and the states that they did their projects in. There were many states, as shown in Table 2.2, but I decided to choose only a few states to focus on. I chose some that had a large number of solar projects, like MA and MN, and some that had little, like OK and DC. Looking at the table, it shows that companies don't tend to go out of one area of focus, and there is even competition between companies in some states.

	Xcel Energy	Oklahoma Gas and Electric Co.	National Grid	Eversource	Pepco
CO	46	0	0	0	0
DC	0	0	0	0	12
MA	0	0	72	43	0
MN	63	0	0	0	0
OK	0	3	0	0	0

Table 2.1

State	Frequency	Relative Frequency
AR	4	0.00736648
AZ	6	0.01104972
CO	79	0.14548803
DC	12	0.02209945
DE	1	0.00184162
FL	6	0.01104972
GA	7	0.01289134
IA	10	0.01841621
ID	1	0.00184162
IL	2	0.00368324
IN	3	0.00552486
KS	3	0.00552486
KY	3	0.00552486
MA	121	0.2228361
MD	1	0.00184162
ME	13	0.02394107
MI	7	0.01289134
MN	97	0.1786372
MO	4	0.00736648
MT	5	0.0092081
NC	15	0.02762431
ND	1	0.00184162
NE	6	0.01104972
NJ	1	0.00184162
NM	1	0.00184162
NV	2	0.00368324
NY	19	0.03499079
OH	1	0.00184162
OK	5	0.0092081
OR	9	0.01657459
PA	1	0.00184162
SC	10	0.01841621
TN	4	0.00736648
TX	9	0.01657459
UT	3	0.00552486
VA	1	0.00184162
VT	25	0.04604052
WA	25	0.04604052
WI	20	0.03683241
Total	543	1

Table 2.2

Project Part 3

The two Tables below are statistical charts showing the system sizes of the different solar projects around the US. With the data set having a range of 27,990, the mean was 1325.20402. The five-number summary for the box and whisker chart was, Min=10, Q1= 108.432, Median= 582.08, Q3= 1372.3975, and Max= 28,000. The histogram is skewed to the right because many of the solar projects have a small system size. This can be due to these systems only needing to power a small area, like a rural town or small business. There are less larger projects because they are more expensive, making the histogram skewed right, and have multiple outliers in the box and whisker chart.

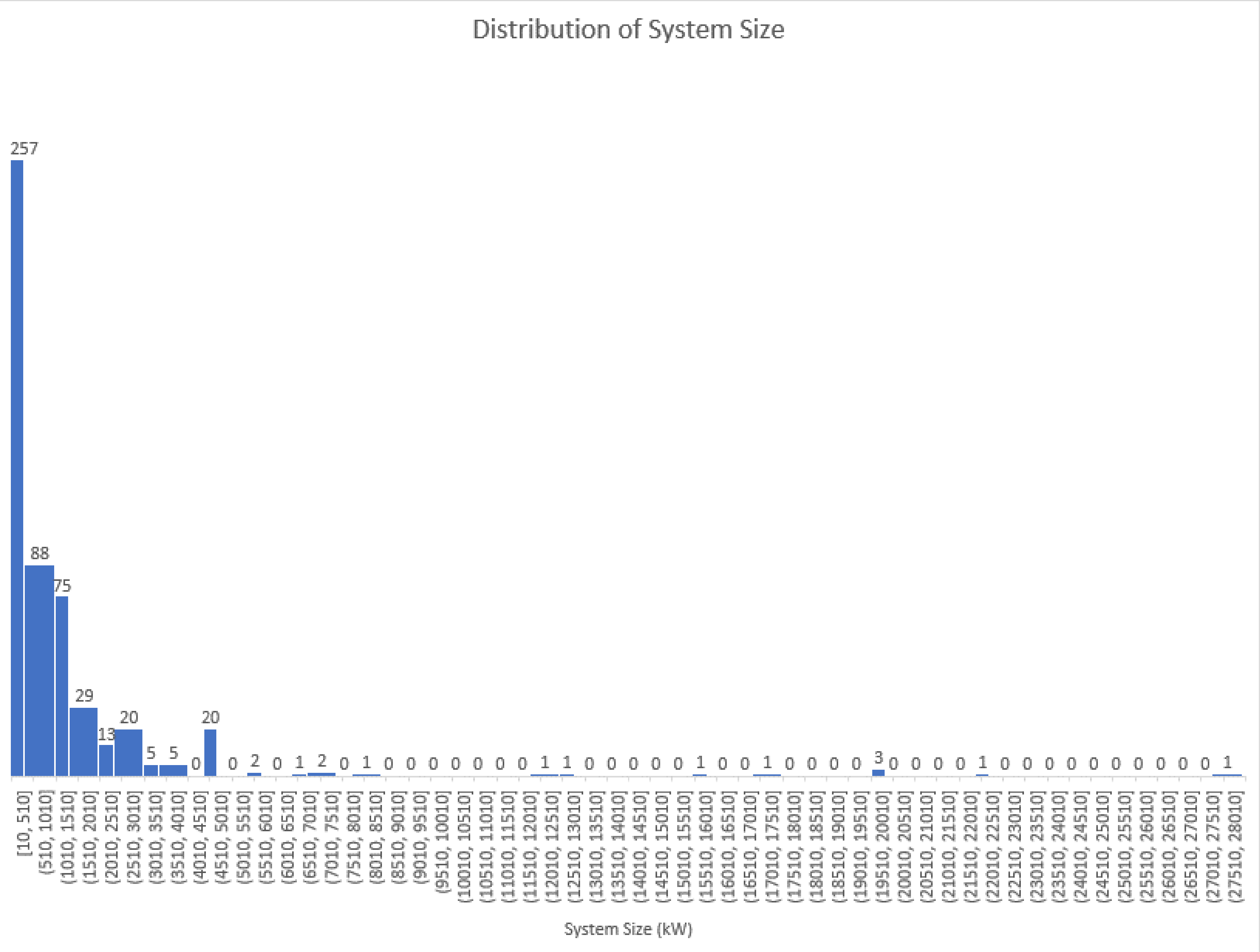

Table 3.1

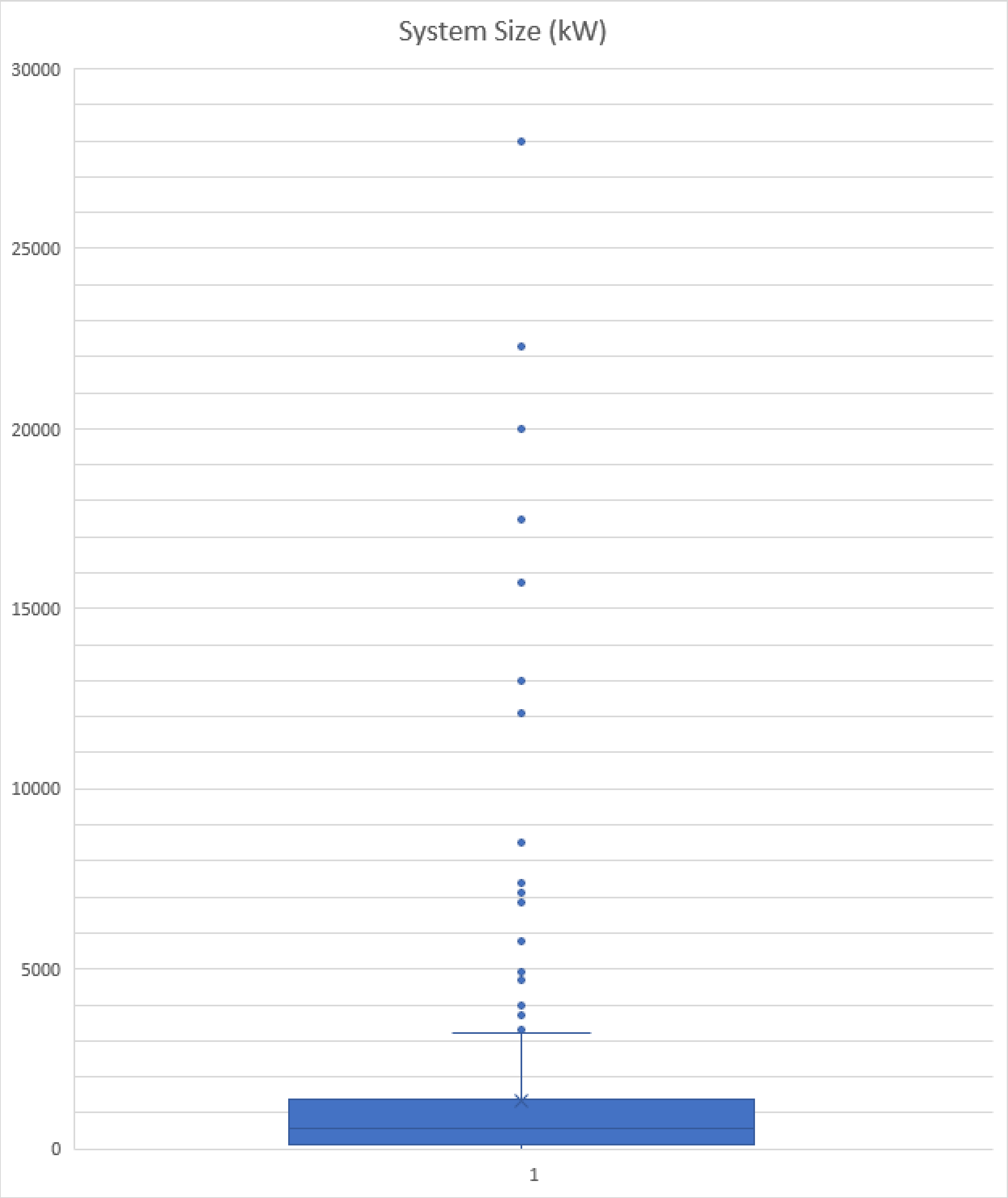

Table 3.2

Project Part 4

For this week, I decided to look at the system size for my quantitative hypothesis.

H_o: μ=600

H_a: μ≠600

I wanted to look at the mean because a lot of the projects are really small, so I wanted to see if the mean computed earlier is actually near the true mean.

For the categorical variable, I decided to look into which state has the highest number of solar projects.

H_o: The proportion of solar projects in MA is p=.25.

H_a: The proportion of solar projects in MA is p<.25.

I decided to choose this because the data shows that MA has a lot of projects in that state, so I wanted to see if the majority of solar projects are in MA.

Project Part 5

Going back and testing the quantitative hypothesis, I created a bootstrap sample. I found that the standard error for the sample is 108.91418. I then computed the 95% confidence interval for the mean is between 1071.27496 and 1513.09892. Below is the histogram for the bootstrap distribution of the means (Table 5.1). With this we can reject the null hypothesis because the estimated mean is outside of the confidence interval, therefore it cannot be true.

I also tested the categorical hypothesis using bootstrapping. I found the standard error to be .017833. The 95% confidence interval that I was able to find for the proportion of projects being in MA is between 0.186068 and 0.2574. Below is the histogram for the bootstrap distribution (Table 5.2). With the null hypothesis of p=.25 being within the 95% confidence interval, we cannot reject the null hypothesis.

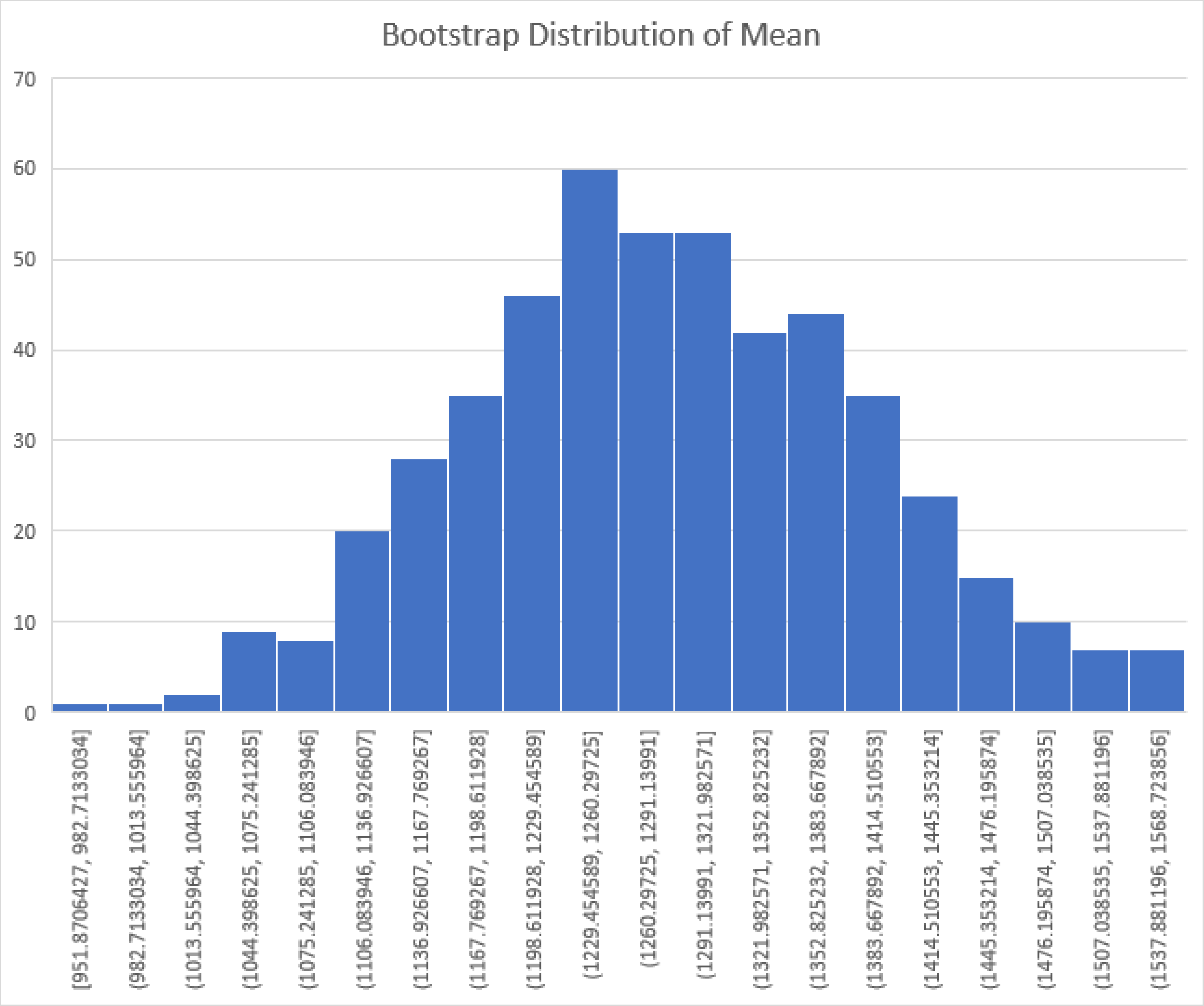
Bootstrap Distribution of Mean
70
60
50
40
30
20
10
0
[951.8706427, 982.7133034]
(982.7133034, 1013.555964]
(1013.555964, 1044.398625]
(1044.398625, 1075.241285]
(1075.241285, 1106.083946]
(1106.083946, 1136.926607]
(1136.926607, 1167.769267]
(1167.769267, 1198.611928]
(1198.611928, 1229.454589]
(1229.454589, 1260.29725]
(1260.29725, 1291.13991]
(1291.13991, 1321.982571]
(1321.982571, 1352.825232]
(1352.825232, 1383.667892]
(1383.667892, 1414.510553]
(1414.510553, 1445.353214]
(1445.353214, 1476.195874]
(1476.195874, 1507.038535]
(1507.038535, 1537.881196]
(1537.881196, 1568.723856]

Table 5.1

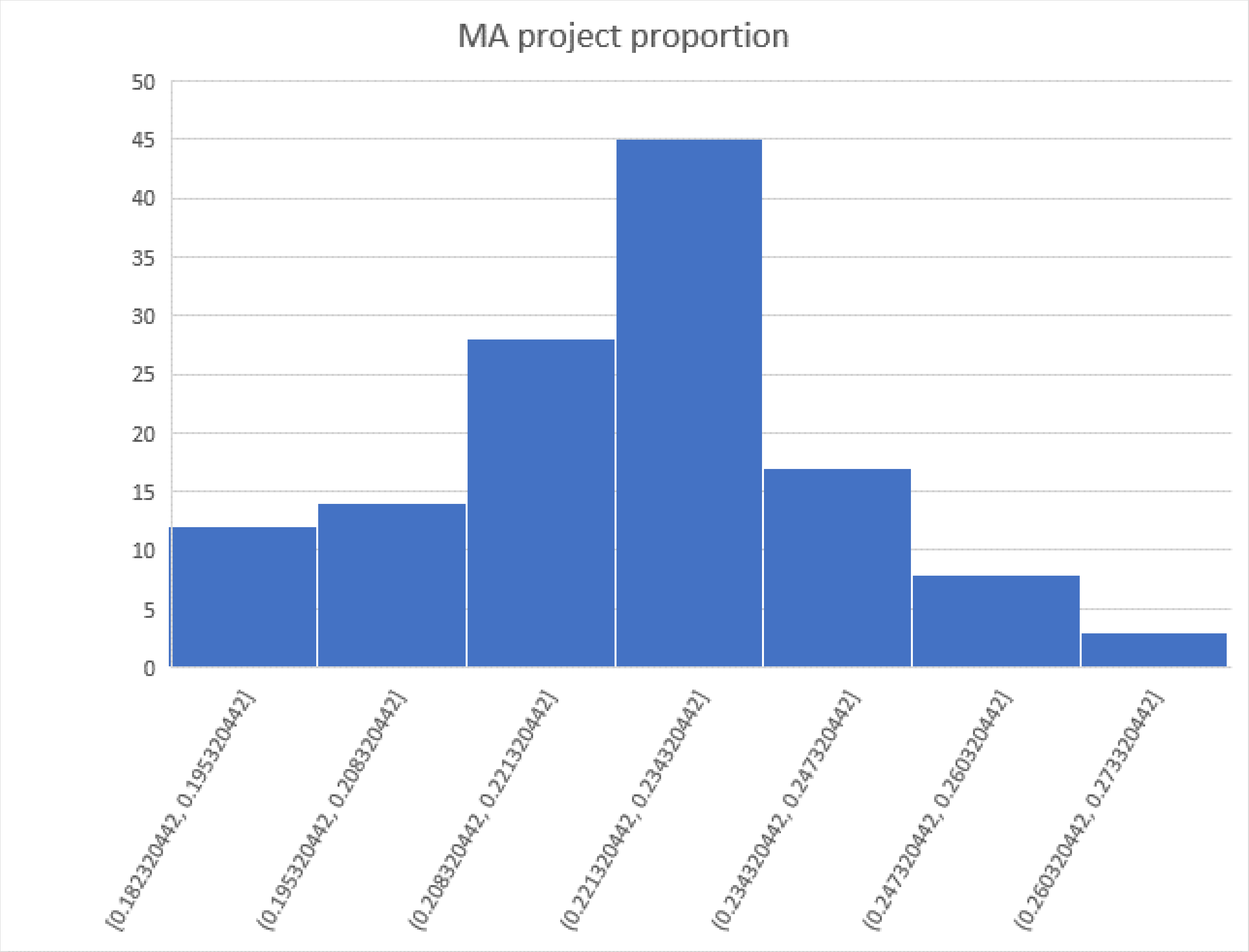
MA project proportion
50
45
40
35
30
25
20
15
10
5
0
[0.182320442, 0.195320442]
(0.195320442, 0.208320442]
(0.208320442, 0.221320442]
(0.221320442, 0.234320442]
(0.234320442, 0.247320442]
(0.247320442, 0.260320442]
(0.260320442, 0.273320442]

Table 5.2

Project Part 6

Following the formulas that we used, I was able to come up with this table below for the proportion of projects in MA. Since I am using less than in my alternative hypothesis, and the z* score is symmetrical on both sides of p, I can use -1.644854 as the beginning of my lower rejection zone. That being said, the z value that I created is not below the critical z value, therefore we are not able to reject the null hypothesis. The 95% confidence interval that I found was between .192271 and .253401. This also does not allow us to reject the null hypothesis. Compared to the bootstrap, which was a 95% confidence interval between 0.186068 and 0.2574, the one computed through formulas had a smaller range between its higher and lower values. This makes me think that using formulas makes us have a more accurate confidence on the true proportion of projects in MA.

Sample	Proportion	Statistic	Significance		
543	0.25	0.222836	0.05		
n	p	p hat	alpha		
				CI	
SE	z	z*	p	Lower	Higher
0.018582	-1.461811821	1.644854	0.071896	0.192271	0.253401
			0.928104		

Project Part 7

For this part of the project, I used a t test to try and find if the mean of the system sizes was not equal to 600 kW. I used 600 as the mu because that was the assumed mean and found x bar by taking the mean of all of my samples. I found the standard error, t score, and t* below using the formulas listed below. Using t*, I found the 95% confidence interval to be between 1096.508 and 1553.9. When I did the bootstrap sampling, I found a confidence interval between 1071.27496 and 1513.09892. Both of these interval are really similar, and by using both, we can reject the null hypothesis because 600 does not fall in between the confidence interval.

mu	x bar	n	stand dev	Alpha
600	1325.204	543	2712.932	.05
t	SE	t*	CI	
6.22904	116.4231	1.96435	Lower	Upper
(x bar-mu/SE)	(SD/SQRT(n))	T.INV(alpha,n-1)	1096.508	1553.9

Project Part 8

Using Table 2.1, I created two conditional probabilities. The first on that I tested was, if one project was chosen at random, what is the probability that the utility involved is Eversource, given that the project is in MA? First, I looked at the probability of getting a project in MA (P(B)), and that was a probability of 0.222836096. Next, I looked at the intersection for how many projects out of all 543 projects that were in MA and have Eversource as the utility. There were 43 out of the 543 that met that criteria, so the P(A cap B) is equal to 0.07918969. To find P(A|B) I used to formula P(A|B)=A cap B/P(B). Plugging in numbers, P(A|B) came to be 0.355371901. There is a 35.54% chance of choosing a project with Eversource given the project is in MA.

The second conditional probability that I wanted to find was, if one project is chosen at random, what is the probability that the project will be in CO, given that the utility provide is Xcel Energy? Following the same steps as before to find P(A|B), I found P(B) to be 0.200736648, P(A cap B) to be 0.084714549, and P(A|B) to be 0.422018349. This means that the answer to the second conditional probability is that there is a 42.2% chance of choosing a project in CO, given that it is run by Xcel Energy.

	Xcel Energy	Oklahoma Gas and Electric Co.	National Grid	Eversource	Pepco
CO	46	0	0	0	0
DC	0	0	0	0	12
MA	0	0	72	43	0
MN	63	0	0	0	0
OK	0	3	0	0	0

Table 2.1