[redacted]

Intro to Probability & Statistics

MATH-1223-01

Dr. Jacob

The Truth is Heavy:

The Epidemic of Obesity in the United States

It is no secret that United States is a very obese country. About 40% of adults alone suffer from obesity. The data set I have chosen comes from https://wallethub.com/edu/fattest-cities-in-america/10532 and it is a sample that provides 2021's top 100 fattest cities in the US, as well as their obese percentages and some other interesting information. I will be discussing the top 25 of these 100 cities in part one. The categorical variables include the year and the city with state and the quantitative variables are the states by rank and the percentage by population. The reason why I have chosen this data is because it pertains to my degree, kinesiology which I would later wish to pursue a physical therapy career field. As a person who prefers a healthier lifestyle involving adequate nutrition and daily exercise, I find that the issue of obesity needs to be addressed in order to improve American's lifestyle choices as it is clear that obesity is severely on the rise, which effects life expectancy and overall quality of life. With the data set, I hope to use this as a way of convincing Americans to consider better and healthier lifestyle choices and to further my understanding of the epidemic known as obesity.

Year	Rank	City	State	% by Population
2021	1	McAllen	Texas	84.73
2021	2	Memphis	Tennessee	84.18
2021	3	Baton Rouge	Louisiana	83.65
2021	4	Little Rock	Arkansas	83.22
2021	5	Shreveport	Louisiana	83.18
2021	6	Birmingham	Alabama	82.51
2021	7	Jackson	Mississippi	82.41
2021	8	Mobile	Alabama	81.69
2021	9	Lafayette	Louisiana	81.40
2021	10	Knoxville	Tennessee	81.23
2021	11	Chattanooga	Tennessee	80.88
2021	12	Tulsa	Oklahoma	80.39
2021	13	Augusta	Georgia	79.94
2021	14	Greenville	South Carolina	79.68
2021	15	Fayetteville	Arkansas	79.60
2021	16	Myrtle Beach	South Carolina	79.24
2021	17	San Antonio	Texas	78.77
2021	18	Wichita	Kansas	78.68
2021	19	New Orleans	Louisiana	78.43
2021	20	Nashville	Tennessee	78.24
2021	21	Oklahoma City	Oklahoma	78.11
2021	22	Toledo	Ohio	77.90
2021	23	Huntsville	Alabama	77.60
2021	24	Louisville	Kentucky	77.35
2021	25	Charleston	South Carolina	77.32

For part two of this project, I will be discussing certain categorical variables with frequencies and relative frequencies. One thing that is fairly noticeable with the table I have shown, and that is with the exception of Ohio, they are all southern states. That might play into the whole southern comfort eating and dining but that is a different study. Below is a data table showing the frequency and relative frequency of the number of cities the individual state that has high obesity rates.

State	Frequency	Relative Frequency
Alabama	3	0.12
Arkansas	2	0.08
Georgia	1	0.04
Kansas	1	0.04
Kentucky	1	0.04
Louisiana	4	0.16
Mississippi	1	0.04
Ohio	1	0.04
Oklahoma	2	0.08
South Carolina	3	0.12
Tennessee	4	0.16
Texas	2	0.08
Total:	**25**	**1**

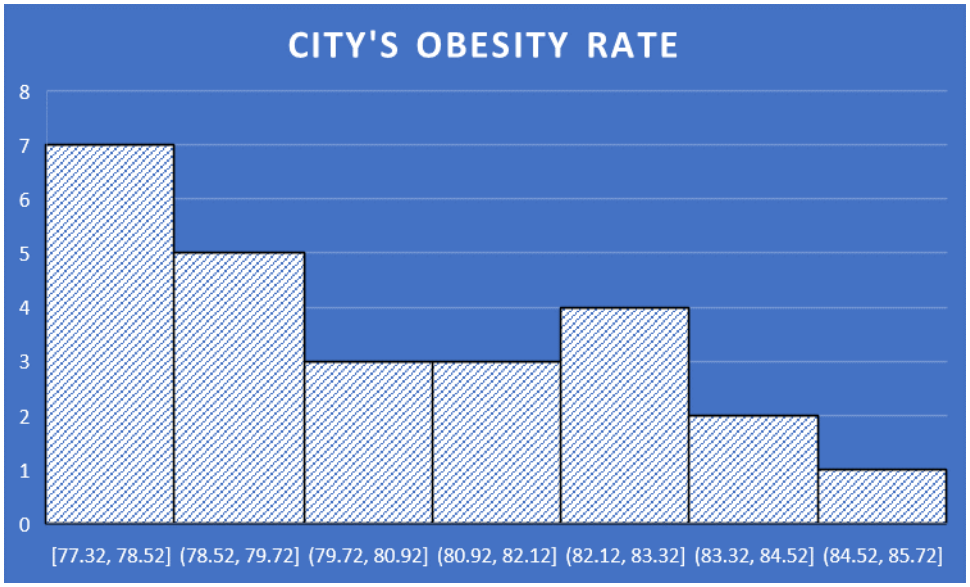
As shown, there is a tie between Louisiana and Tennessee for having four cities each making that 16%. Combined, that makes up almost a third of the top 25 cities with high obesity rates. On the flipside, 8% of the 25 had at least four cities with high obesity rates; whereas, 20% only had one city. This is good because for the states with fewer cities with high rates serve as a passive aggressive way to combat obesity.

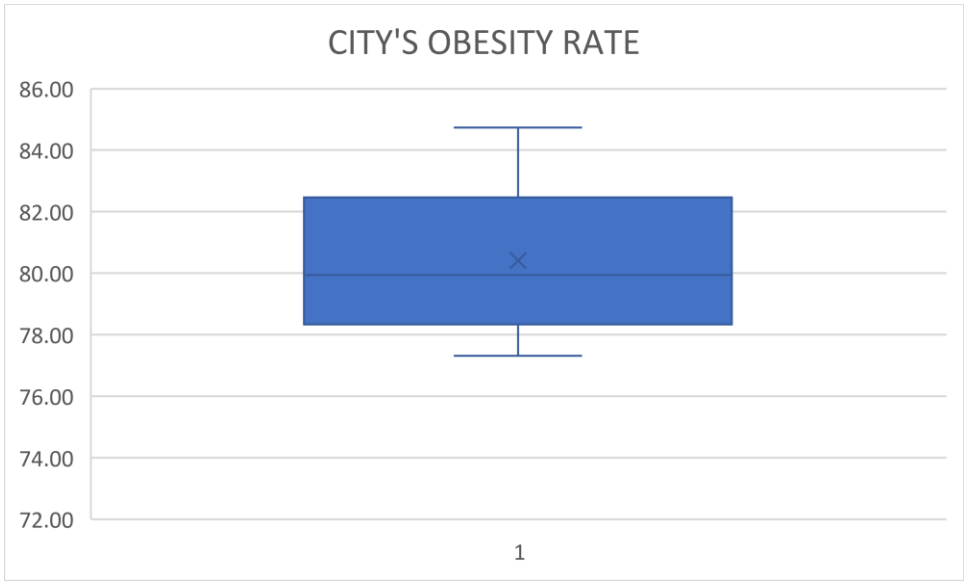
Moving on, the next table shows the city's states percent of the population that are obese. For example, there is one city in Alabama that is between 70%-79% and another is there are three cities in Louisiana that fall between 80%-89%.

States	50%-59%	60%-69%	70%-79%	80%-89%	90%-100%	Total:
Alabama	0	0	1	2	0	3
Arkansas	0	0	1	1	0	2
Georgia	0	0	1	0	0	1
Kansas	0	0	1	0	0	1
Kentucky	0	0	1	0	0	1
Louisiana	0	0	1	3	0	4
Mississippi	0	0	0	1	0	1
Ohio	0	0	1	0	0	1
Oklahoma	0	0	1	1	0	2
South Carolina	0	0	3	0	0	3
Tennessee	0	0	1	3	0	4
Texas	0	0	1	1	0	2
Total:	**0**	**0**	**13**	**12**	**0**	**25**

This is important to show because it allows us to see the percent of the population that live among us. It is eye opening! Being an Okie myself and the fact that I don't live too far from the cities under Oklahoma, it is shocking. Between 70% and 89% of those two Oklahoma cities are obese! Another interesting note of the 25 cities is that is almost split between. 52% are between 70%-79% and 48% are between 80%-89%. Hopefully we never see it go into the 90% and above range.

For part three, I will be further analyzing this sample with all top 25 cities chosen by way of histogram and box plot. I will also provide a mean, standard deviation, and a five number summary. This is the main quantitative variable being analyzed due to the fact this is the entire population of the individual city and the article didn't share other demographics.





Mean: 80.41

Standard Deviation: 2.29

Minimum: 77.32

Q1: 78.34

Median: 79.94

Q3: 82.46

Maximum: 84.73

It is obvious in the histogram it is skewed right representing the heaviest cities. With the interquartile range, I was able to calculate if there were any outliers and to my surprise and somewhat relief, there weren't any. If there were, they would have had to be less than 76.16% and 88.64%.

For part four of my project, I will be creating a null and alternative hypothesis for a quantitative variable and a categorical variable. For my quantitative variable, I will set my null parameter to be that on average, the top 25 cities are at least 78% obese. Unfortunately, it's a little worse. On average, 80.41% of the top 25 cities are obese give or take 4%. The sad thing is, the city closest to this average is Tulsa, Oklahoma at 80.39%!

H_0: $\mu = .78$

H_a: $\mu \neq .78$

For my categorical variable null and alternative hypothesis, I will say that 95% of these cities in 2021 come from southern states. That reasoning coming from southern comfort eating, fried foods, or maybe even ethnic demographics, but that is a different study. I wasn't far off with my findings. Of the 25, 23 of them or 92% were indeed from southern states. The other 8% or two states were Ohio and Kansas.

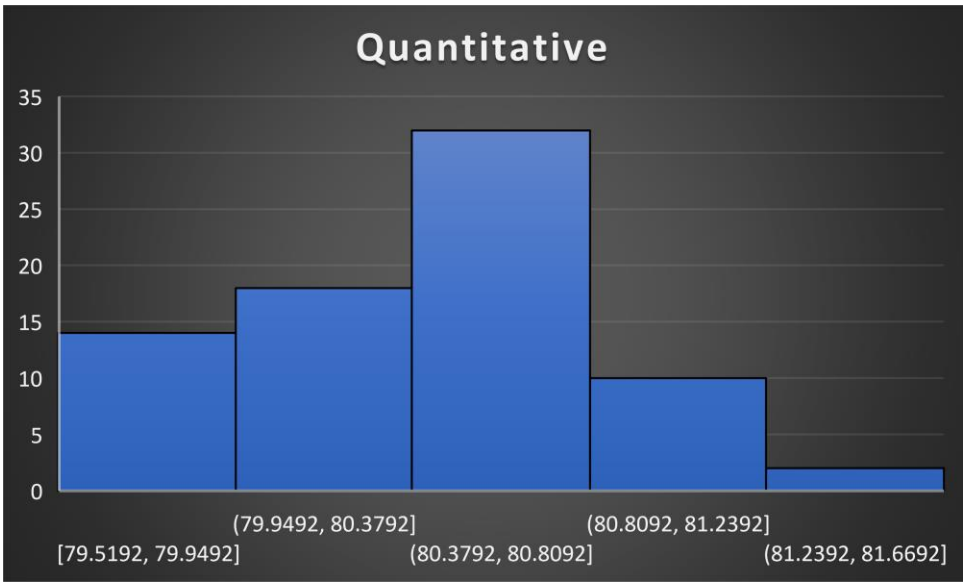
H_0: $p = 0.95$

H_a: $p < 0.95$

In the 5^{th} part of my project, I will be utilizing the bootstrap method to find the standard error for my quantitative variables and categorical variables. Depending on each of the outcomes, I will decide whether to accept or reject the results.

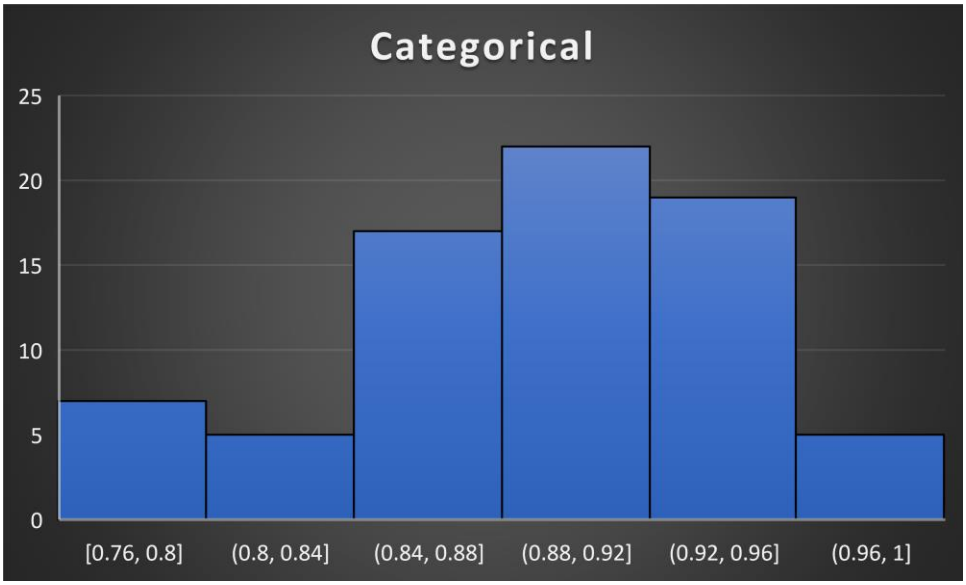
Quantitative Variable

My standard error for my quantitative is 0.435790719. With this information, I am able to calculate for my 95% confidence interval. The distribution is between [79.55, 81.29]. This bootstrap sample result was tested 50+ times. Knowing this allows me to decide that I fail to reject the null hypothesis because the data calculated leaned toward the bootstrap mean of 80.51.



Categorical Variable

For the categorical variable null and alternative hypothesis, I said 95% of the 25 states were southern and on the contrary, it was 92%. To test this with boot strapping, here were the results. The bootstrap mean comes back to .9168 with a standard deviation of .0534. My 95% confidence interval came between [0.810098274, 1.023585937]. I can conclude that I fail to reject my null hypothesis.



` In the next part of my project, I will be retesting my hypothesis from my categorical variable data with various formulas to further see if the numbers come back different or similar. Previously, I had guessed that 95% of these states were southern and it turned out that 92% were in fact southern. My standard deviation came to be .0534 and my 95% confidence interval was [0.810098274, 1.023585937]. Using some formulas, let's see how true these numbers are.

N	**Sample size**	**25**
X	**Southern States**	**23**
Proportion	**From Hypothesis**	**95%**
P Value	**NORM.DIST(Z STAT, .92, 1, TRUE)**	**0.053819105**
P Hat	**X/N**	**92%**
SE	**SQRT(p(1-p)/n**	**0.0436**
Z Statistic	**(P-Hat-p)/SE**	**-0.6882**
Z*	**NORM.S.INV (0.975)**	**1.959964**
CI Low	**.92-(2*SE)**	**0.8328**
CI High	**.92+(2*SE)**	**1.0072**

As you can see, in comparison to the previous data collected, the numbers are fairly similar. These calculations will have some rounding error as a result of myself doing a majority of the calculations. Overall, through boot strapping and formulas, I fail to reject the null hypothesis.

Just like above, I will now use the same formulas to find if my original data is accurate with my quantitative variables. I originally set my null hypothesis to 78% and ended up getting 80.41%.

MU	**From Hypothesis**	**78%**
X-bar	**Average**	**80.41%**
Sigma	**Standard Deviation**	**2.294**
N	**Sample Size**	**25**
SE	**Sigma/SQRT(n)**	**.46**
T Statistic	**Xbar-MU/SE**	**5.23**
Alpha	**For 95% CI**	**.05**
T*	**T.INV(alpha,n-1)**	**-1.71**
CI Low	**Xbar-T-Value*SE**	**79.62%**
CI High	**Xbar+T-Value*SE**	**81.20%**

As you can see from the formulas above, the confidence intervals are almost exact with my original numbers being [79.55, 81.29] and these numbers come [79.62, 81.20]. That's due more than likely to rounding error. When you take those and get the average, you get exactly 80.41! With the help of these formulas, I fail to reject the null hypothesis.

For the final part of my project, I will be using a formula to determine conditional probabilities from the table of part 2. The formula I will be using is

$$P(A|B) = P(A \cap B)/P(B)$$

City(x)State(y)	McAllen	Memphis	Baton Rouge	Little Rock	Shreveport	Birmingham	Jackson	Mobile	Layfayette	Knoxville	Chattanooga	Tulsa	Augusta	Greenville	Fayetteville	Myrtle Beach	San Antonio	Wichita	New Orleans	Nashville	Oklahoma City	Toledo	Huntsville	Louisville	Charleston	Total
Alabama	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	3
Arkansas	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	2
Georgia	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
Kansas	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
Kentucky	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
Louisiana	0	0	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	4
Mississippi	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Ohio	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
Oklahoma	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	2
South Carolina	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	1	3
Tennessee	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	4
Texas	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	2
Total	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	25

Because I have mentioned I am an Okie, I will make this probability dear to me. My first probability will be to determine the cities are from Oklahoma coming out to be 2/25. This probability is simple because I would take the two given cities in Oklahoma and divide that by the total amount of cities. The two cities are Oklahoma City and Tulsa out of the other 25 cities which is .08%. In the grand scheme of things in respect to the data, this statistic is pretty good but not for the fact that these two cities are still in the top 25 cities in the nation.

My second probability I will create is what if the probability that the city is McAllen if the given state is Texas? Using the formula, we will set A as McAllen and B as Texas. Following the order of operations with the formula $P(A|B) = P(A \cap B)/P(B)$. The intersection of McAllen and Texas is 1 but to figure if the probability is Texas it is 2 making the probability ½ or 50% that the city is McAllen.