Taste the rainbow, this slogan probably brings to mind the popular fruity Skittles candy. Have you ever wondered how many colors of each candy there is in a bag? Maybe you thought each bag had the same amount of candies and colors of those candies. Each candy bag varies and if you thought all Skittles bags were equal, you just might be wrong. This term project is about organizing and analyzing the data found from bags of Skittles. These bite-sized chewy candies with a colorful candy shell will help us apply the statistic skills we have learned this semester and use it in a real life situation.

First, each member of the class purchased a 2.17 ounce bag of Original Skittles. I searched far and wide, and after going to several grocery stores, found my bag of Skittles in a 7-Eleven. Each student then counted the number of red, orange, yellow, green, and purple candies from the bag. My bag of skittles had 10 red, 11 orange, 13 yellow, 16 green, and 8 purple skittles. We then submitted our data online to Professor Santistevan. She combined the data and put it in an easy to read chart so we could use the information and know how many Skittles were in everyone else’s bags and how many of each color there were. The Number of Candies of Each Color in the Sample
As you can see above, I have created a Pie Chart and Pareto Chart to show the number of candies of each color in the Skittles bags for the entire class sample. These graphs do reflect what I expected to see, I hypothesized that the number of colors of Skittles I had in my bag would most likely represent what the entire class had as well. Although I did assume that the numbers for each color would vary a little bit as well as the number of Skittles in each person's bag. The overall data from the class sample reflects very similarly with my results that I received. Below are tables reflecting the results I got from my own bag of Skittles and the results for the classroom sample.

### Personal Skittles Bag

<table>
<thead>
<tr>
<th>Number of red candies</th>
<th>Number of orange candies</th>
<th>Number of yellow candies</th>
<th>Number of green candies</th>
<th>Number of purple candies</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>11</td>
<td>13</td>
<td>16</td>
<td>8</td>
</tr>
<tr>
<td>.172</td>
<td>.190</td>
<td>.224</td>
<td>.276</td>
<td>.138</td>
</tr>
</tbody>
</table>

### Classroom Sample of Skittles Bags

<table>
<thead>
<tr>
<th>Number of red candies</th>
<th>Number of orange candies</th>
<th>Number of yellow candies</th>
<th>Number of green candies</th>
<th>Number of purple candies</th>
</tr>
</thead>
<tbody>
<tr>
<td>321</td>
<td>292</td>
<td>306</td>
<td>316</td>
<td>276</td>
</tr>
<tr>
<td>.212</td>
<td>.193</td>
<td>.209</td>
<td>.213</td>
<td>.183</td>
</tr>
</tbody>
</table>

### Summary statistics

<table>
<thead>
<tr>
<th>Column</th>
<th>n</th>
<th>Mean</th>
<th>Variance</th>
<th>Std. dev.</th>
<th>Median</th>
<th>Range</th>
<th>Min</th>
<th>Max</th>
<th>Q1</th>
<th>Q3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Candies per Bag</td>
<td>25</td>
<td>60.44</td>
<td>19.006667</td>
<td>4.3596636</td>
<td>60</td>
<td>24</td>
<td>53</td>
<td>77</td>
<td>59</td>
<td>61</td>
</tr>
</tbody>
</table>
In the above histogram and Boxplot it shows the distribution of the candies per bag for the classroom sample. In the histogram and boxplot you can see how the distribution is skewed to the right of the graph. There is also a large section missing from 65-75. One bag of candy had 77 candies in the bag which pulled the mean to the right of the graph. I did not expect to see the candies per bag to jump from 62 to 77 candies. I thought for sure that there would be amounts of candies that fell in-between those two values but the data clearly shows that 77 is an outlier compared to the rest of the classes’ bags. The classroom sample data of Skittles per bag is very close but does not agree with my own Skittles candy bag. I had 58 candies in my bag and that is very close to the first quartile but I am one Skittle off.

<table>
<thead>
<tr>
<th>Total # of Bags in Sample</th>
<th>Skittles Candies in my Bag</th>
</tr>
</thead>
<tbody>
<tr>
<td>25 Bags</td>
<td>58 Candies</td>
</tr>
</tbody>
</table>

There are two basic divisions of data, quantitative and categorical. Quantitative data is values that can be measured or counted, you will sometimes hear it called numerical data. Some examples of quantitative data would be weight in pounds, length in inches, and time in seconds. Categorical data is values or observations like names or labels that can be sorted into groups or categories but cannot be measured. Categorical data can take on numerical values in some cases but those numbers don’t have mathematical meaning. Examples of categorical data would be models of cars, gender, pass or fail, and eye color.

Graphs that make sense with categorical data would be pareto chart, bar graph, and pie chart. Scatterplot, stem-and-leaf plot, time-series graph, and dot-plot are example of graphs that make sense using quantitative data. In quantitative data you could add, multiply, subtract, and divide the data. Calculations with quantitative data still makes sense mathematically when manipulated but you could not make calculations with the categorical data. Such as the colors of Skittles in this project, does it make sense to add or multiply the categories yellow + green? The answer is no, this would not provide mathematical meaning.

A confidence interval is a range of values used to estimate the value of a population parameter. Statisticians use a confidence interval to describe the amount of uncertainty associated with a sample estimate of a population parameter. The range of the confidence interval is defined by the sample statistic plus or minus the margin of error. (Confidence interval = sample statistic ± Margin of error) What a confidence interval does is allow us to estimate the range in which our true population parameter falls, given what we know about the population from the sample we’ve observed. Because no estimate can be 100 percent reliable, we must be able to know how confident the estimates are. For example if repeated samples were taken and the 95% confidence interval was determined for each sample, 95% of the intervals would contain the population mean. Confidence intervals can be computed for various parameters, not just the mean.

That may be difficult to understand so to better explain confidence intervals, I am going to start at the beginning and I am going to give an example before we jump in and start working
with the Skittles to find confidence intervals. In order to understand confidence intervals we have to understand sampling. A sample is a set of observations/data taken from a population. A sample may be a set of oranges in a grove. We wish to know how big the oranges are at a point in the season. We can’t weigh all the oranges one at a time (that would take forever) so we take a sample and we are going to infer information about all the oranges in the grove based on this sample. Because it is just a sample it will never be a perfect representation of the population. Different samples of the same population will give different results and so there will always be sampling error no matter how hard we try.

Now back to confidence intervals, when we express an estimate of the data we can give it as a confidence interval this will determine how accurate our estimate is likely to be. The question is what is the mean weight of all the oranges in the grove? We are going to take the sample data of the oranges we weighed and calculate the sample mean. We are then going to use a confidence interval to express the range in which we think the population parameter lies. For example the apple mean weight is 147g ± 4g.

There are two things that can affect the confidence interval. The first is variation of population. If the oranges that we got were all very similar is size then the sample is going to have little variation and we will have a small confidence interval. Our estimate is going to be closer to the estimated mean. On the other end of the spectrum if the population has lots of variation then the bigger confidence interval will be. We will be further away from the estimated mean we found. Sample size also affects the confidence level. If we take a small sample, the data it will vary more; but we take a larger sample, we have more information and unusual values will not affect the data as much.

The first confidence interval estimate that I calculated was for a 99% confidence level of the true proportion of yellow candies. The answer was .176 < p < .230. or written as .203 ± .027. To interpret this information it means that I am 99% sure that the interval, 203 plus or minus the error of .027 contains the true proportions of all yellow candies.
The second confidence interval estimate performed was a 95% confidence interval for the true mean number of candies in per bag. The result was 58.6 < µ < 62.2 or written as 60.4 ± 1.800. This means that I am 95% sure that all Skittles bags will fall between the interval, 60.4 plus or minus the error of 1.800 for the mean number of candies in per bag.

The last was a 98% confidence interval estimate for the standard deviation of the number of candies per bag. The result for this was 3.258 < σ < 6.483. So I believe that I am 98% sure that for all bags of Skittles ever made that the standard deviation would fall between 3.258 < σ < 6.483.

A statistical hypothesis is an assumption about a population parameter. This assumption may or may not be true. Hypothesis testing refers to the procedures used by statisticians to accept or reject a statistical hypothesis. Statisticians typically examine a random sample from
The first hypothesis test is to use a 0.05 significance level to test the claim that 20% of all Skittles Candies are red and the alternative hypothesis is that it is not true that 20% off Skittles Candies are red. This is a two tailed test and I found that the test statistic 1.209 is less than the critical value 1.96 so we fail to reject the null hypothesis. There is sufficient evidence to support the claim that 20% of all Skittles candies are red.
The Second hypothesis test we used a 0.01 significance level to test the claim that the mean number of candies in a bag of Skittles is 55. The claim is the null hypothesis and the alternative hypothesis is that the mean number of candies in a bag is not 55. This is a two tailed test and I found that the test statistic 6.193 is greater than the critical value 2.797 so we fail to reject the null hypothesis. There is not sufficient evidence to support the claim that the mean number of candies is a bag of Skittles is 55.

There are three conditions for confidence Interval for estimating a population proportion p

1) The sample is a simple random sample.
2) Either or both of these conditions are satisfied: the population is normally distributed or n>30.
3) There are at least 5 successes and at least 5 failures.

Conditions for Confidence Interval for Estimating a Population Mean with \(\sigma\) not known

1) The sample is a simple random sample.
2) Either or both of these conditions is satisfied: The population is normally distributed or n > 30.

Conditions for Confidence Interval for estimating a population Standard Deviation or Variance

1) The sample is a simple random sample.
2) The population must have normally distributed values (even if the sample is large). The requirement of a normal distribution is much stricter here than in earlier sections, so departures from normal distributions can result in large errors.

The sample we took met all the conditions. The requirements are the same between interval estimates and hypothesis tests. The sample was a simple random sample. A simple random sample is a subset of a statistical population in which each member of the subset has an equal probability of being chosen. A simple random sample is meant to be an unbiased representation of a group. This qualification was not met because everyone in this class was assigned the project and purchased a random bag of Skittles. The sample also met the condition that the population is normally distributed and that the population must have normally distributed values (even if the sample is large). The sample was less than 30 but it was normally distributed and it just had to meet one requirement between the two. I can speculate that all bags of Skittles have a normal distribution. Another requirement for population proportions is there are at least 5 successes and at least 5 failures. This requirement is also met because the proportion of units in the population that belong to the success group is the population success proportion. This population success proportion is denoted by the lower case letter \(p\). The
population success proportion $p$ is a parameter. The population failure proportion $1 - p$ is also a parameter.

Mistakes could also be made gathering this data. One type of error could be recording incorrect data. This could happen if the person counted incorrectly or wrote the wrong quantity down for that color. Non-response error is also something that could affect the data. Each person in the class was assigned to buy a Skittles bag but if someone never bought a bag to record the data then we are missing part of our intended sample.

After calculating confidence intervals and hypothesis tests I have learned that it is important to get a simple random sample because it can affect the accuracy of the results you get when gathering data. I have also been able to learn that by examining a sample we can estimate what the population would be for the mean, standard deviation and proportion. I have been able to learn how to properly set up follow the procedures of a hypothesis test and how to determine if I need to fail to reject or reject the null hypothesis.

Reflection on Term Skittles Project

I still remember looking at the instructions for this project and thinking that I was reading something in another language. I had a hard time working on it because of how difficult it seemed at first. I was amazed at how much work this was actually going to require. We needed to create a random sample of data, have that data organized, create graphs, charts and interpret what the information means. Throughout the semester we were thought a great variety of statistical concepts that gave us the necessary tools to be able to complete this project. Little by little I realized that I was not only capable of gradually understanding the instructions, but I was also able to perform the correct sequence of steps for each one of the exercises that were part of this project. As in any other discipline and class, after learning the theory, practice makes the whole difference.

This project allowed us to put into practice key principles studied throughout the term, from using a sampling method to performing hypothesis testing. The most challenging aspects of the project were really understanding each concept, and how it applied to the population of Skittles not just our sample.
This specific project has helped me to better understand statistics and how statistics works in the real world. I have learned a lot about statistics and how they work. It turns out, statistics are actually used in the real world and they play an important part in our everyday lives. They provide information that is beneficial to society. Since I have decided to follow a professional career in the medical field, I have found myself reading many scientific medical articles based on studies that have been done using statistics.

This project and the class in general have gave me the tools to differentiate between valid professional papers from those that end up being questionable sources of information when analyzing things like graphs, and confidence levels and intervals that can make the whole difference when trying to see if the study is well done. To be able to understand the language behind the statistical analysis of studies with simple, but important terms such as media, range, mean, and mode that are so frequently use in so many instances.

Throughout the entirety of this course I have learned a lot. I have learned how to apply these techniques in a daily life situation as well as learned how to use a great program, StatCrunch to create visual representation of the data. I was able to create graphs and find statistical data by simply imputing the given data.

I have always struggled with math but found that this course allowed me to see how math does apply to real life, and was refreshing to take a course outside of the classroom and math book. I have always gritted my teeth at story problems but his course gave a whole new approach and meaning to the information provided. I never thought I would say this, but math can be fun at times, especially when you can understand how it applies to the real life situations helping us to better understand the world around us.