**Math 1040 Skittles Term Project**

Report Introduction

This project is a way to put in practice what we have learned throughout the semester. It will allow us to pull together many of the concepts we learned in class, including organizing and analyzing data, drawing conclusions using confidence intervals and hypothesis tests. Also, presenting work in a well-organized paper. The overall report will be a narrative that will explain the processes and the conclusions, including mathematical calculations (in some cases It may be handwritten and scanned). But in general, summaries and written conclusions are going to be presented in this report. After finishing it, our instructor asked us to put in in our e-portfolio as a single document.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Number of  red candies | Number of orange candies | Number of yellow candies | Number of green candies | Number of purple candies |
| 13 | 12 | 21 | 8 | 9 |

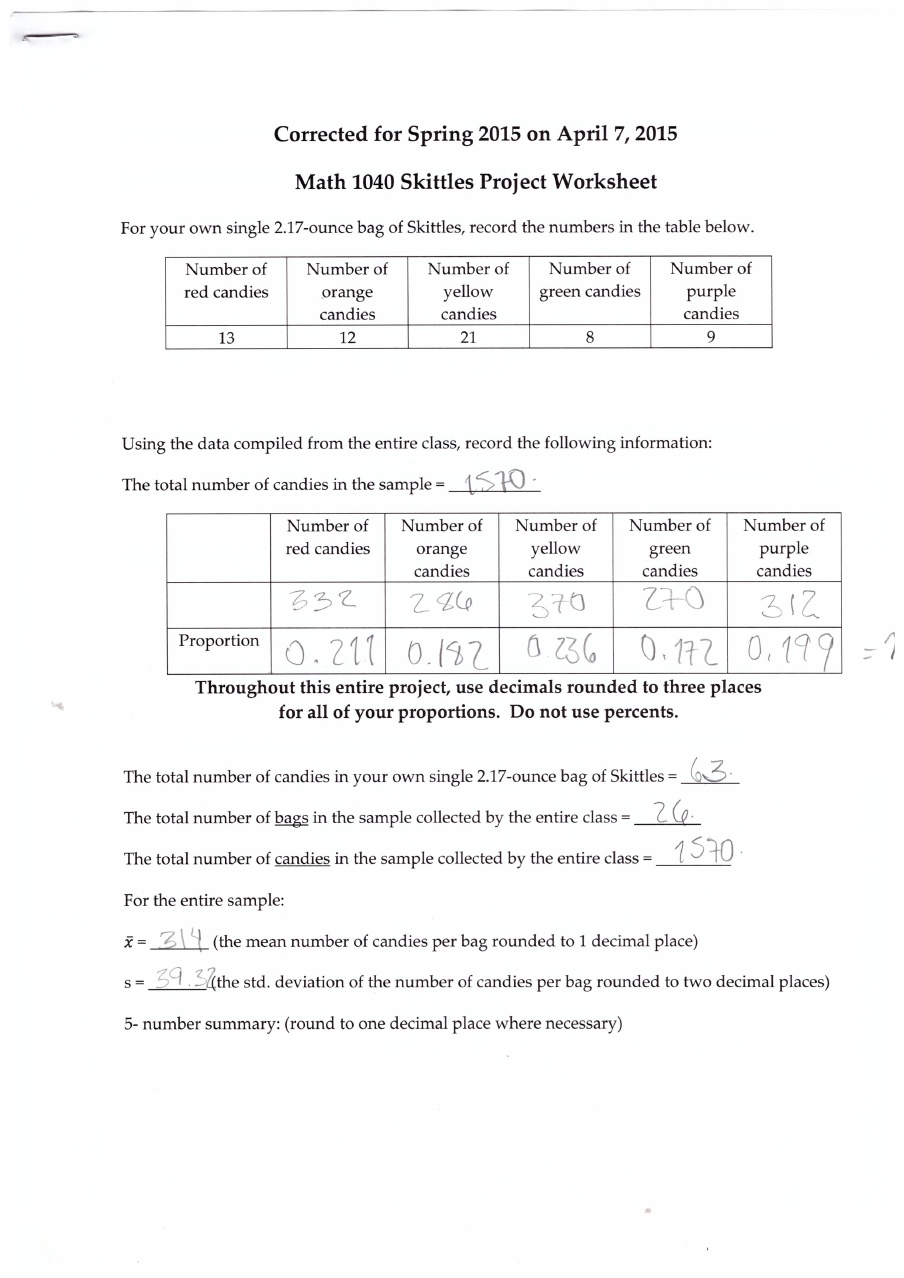
The processes of the project will be clearly detailed through the entire paper, but to give and overview, first we will be working with a 2.17-ounce bag of Original Skittles and we (individually) recorded the following data:

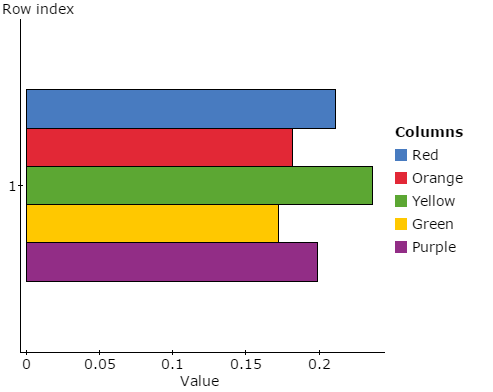
Secondly, our instructor compiled all the data from the whole class, along with the total number of candies in each bag, the total number of bags, and the total number of candies in a master chart. An Excel file containing this information was send out to us so we can work with the entire classroom data. This is the data set that I would be using to complete the rest of the term project.

Organizing and Displaying Categorical Data: Colors

First, to organize and display categorical data of color we will have to determine the proportion of each color within the overall sample gathered by the class. To do this we will have to create a Pie Charta and a Pareto chart. It is important to mention that as we work with the variables, we have to recognize the sample size of the data (1570). The proportions will be decimals rounded to the third place.

On the following graphs, we can see a variation in colors depending on the variable. For example, in the pie chart I was able to match the color with the variable but in the Pareto chart I was not able to do so because It was made in Start Crunch. According to the results and de data that we obtained the graphs do reflect what I expected to see. A higher value representations of the yellow group in both graphs. In comparison, the overall data collected by the whole class agreed with my own data from a single bag of candies, because in contrast with the other sets of data, the yellow skittles values are much higher than the other colors.

****

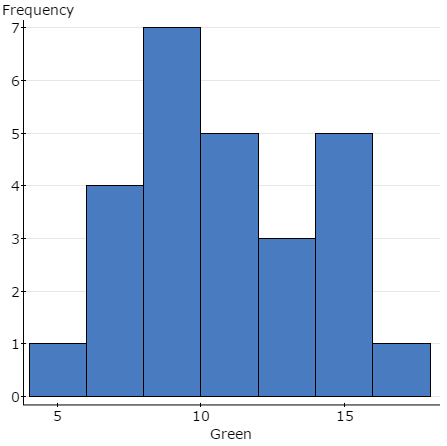


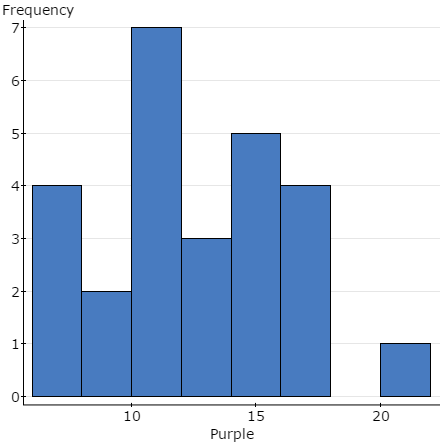
Organizing and Displaying Quantitative Data: the Number of Candies per Bag

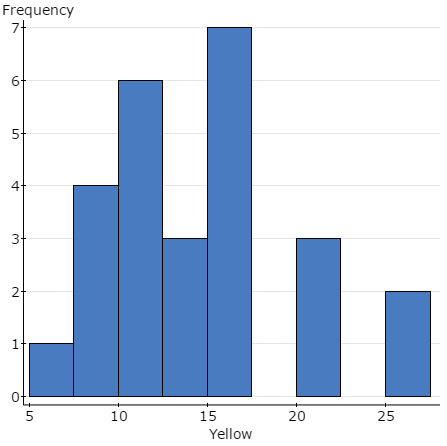
**Summary statistics (Class sample): Summary statistics (sample data):**

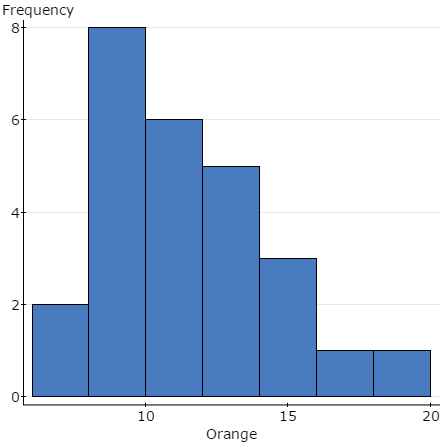
| **Column** | **Mean** | **n** | **Std. dev.** | **Median** | **Min** | **Max** | **Q1** | **Q3** | **IQR** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| red | 13 | 1 | NaN | 13 | 13 | 13 | 13 | 13 | 0 |
| Orange | 12 | 1 | NaN | 12 | 12 | 12 | 12 | 12 | 0 |
| yellow | 21 | 1 | NaN | 21 | 21 | 21 | 21 | 21 | 0 |
| green | 8 | 1 | NaN | 8 | 8 | 8 | 8 | 8 | 0 |
| purple | 9 | 1 | NaN | 9 | 9 | 9 | 9 | 9 | 0 |

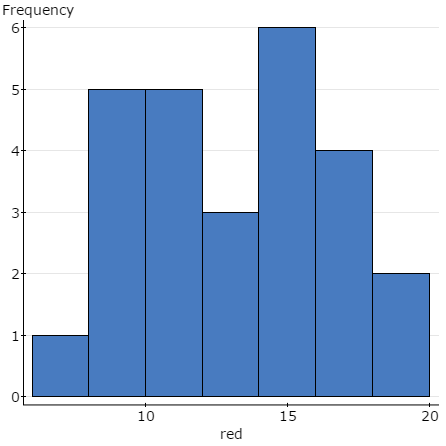
| **Column** | **Mean** | **Std. dev.** | **Median** | **Min** | **Max** | **Q3** | **Q1** | **IQR** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Red | 12.8 | 3.37 | 13 | 7 | 19 | 15 | 11 | 4 |
| Orange | 11 | 3.02 | 10 | 7 | 18 | 13 | 9 | 4 |
| Yellow | 14.2 | 4.89 | 14 | 7 | 25 | 16 | 10 | 6 |
| Green | 10.4 | 3.21 | 10 | 5 | 16 | 13 | 8 | 5 |
| Purple | 12 | 3.73 | 12 | 6 | 20 | 15 | 10 | 5 |

Histograms

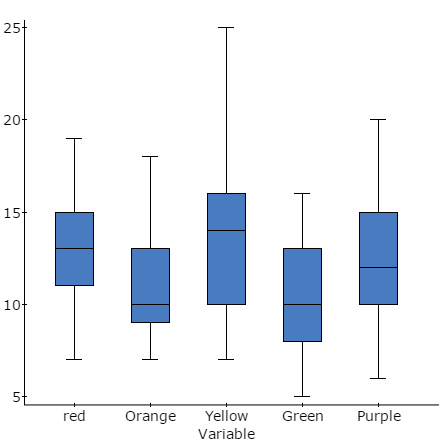






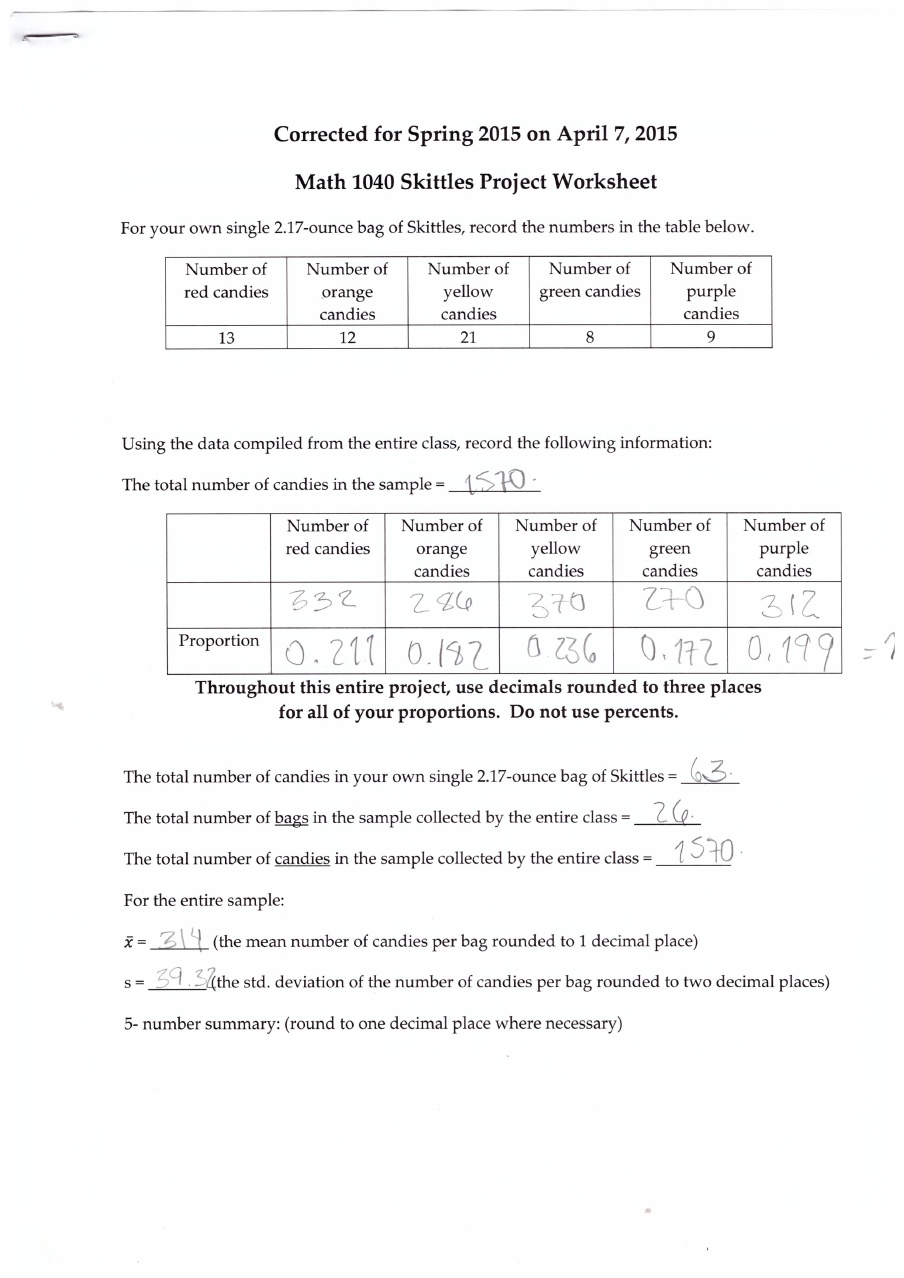


Boxplot



In most of the graphs we can see a not normal distribution. therefore, the graph does not reflect what I expected to see. Because of the fluctuation of data, the distribution is not consistent. Also, the overall data collected by the whole class does agree with my own data from a single bag of candies, because in my sample data the distribution is not normal, even though the highest value is in the middle the distribution has an up-down shape not consistent.

***include the number of candies from your own bag and the total number of bags in the sample.***



Reflection

In statistics, it is important to recognize that there are different types of data: Quantitative (discrete and continuous), Categorical and ordinal. We can also define data as pieces of information that you collect through a study. For example: If you ask each of your friend how many candies to they have, the might give you this data set: 2,4,0,5,9. But, not all data are numbers, let’s say that instead of asking them how many candies to they have, we can ask them their gender, you will get the following data: male, female, female, male, male, female. Both types of data are used to measure information, in order to made an inference about it.

Most of the data falls into two groups: Categorical and Quantitative:

Categorical data is also called qualitative variables. It can be put into a countable number of categories or different groups. It can also represent characteristics such as a person’s gender, marital status, hometown, or the types of movies they like. Also, it may or may not have some logical order. For example: payment method, skittle colors, etc. For this type of data pie charts and bar graphs are the best types of graphs to use because they divide information in categories. This type of data also has a subdivision:

*Ordinal* data mixes numerical and categorical data. The data falls into categories, but the number placed on the categories have meaning. For example: rating a hotel on a scale from 0 (lowest) to 5 (highest), it starts given an ordinal data value. Ordinal data are usually treated as categorical, where the groups are arranged when graphs and charts are made. However, unlike categorical data, the information does have mathematical value. For example, if you survey 600 people and ask them to rate a bar on a scale from 0 to 4, taking the average of the 600 responses will have meaning. This would not be the case with categorical data.

Quantitative data have meaning and measurement. For example: weight of a cereal box, blood pressure, IQ, how many pages the newspaper has, etc. It can also be called Numerical Data. For this type of data, histograms, steam and leaf, line graphs, dot plots etc. are useful types of graphs because they organize information in their numerical values. It also has a subdivision of two types: Discrete and continuous.

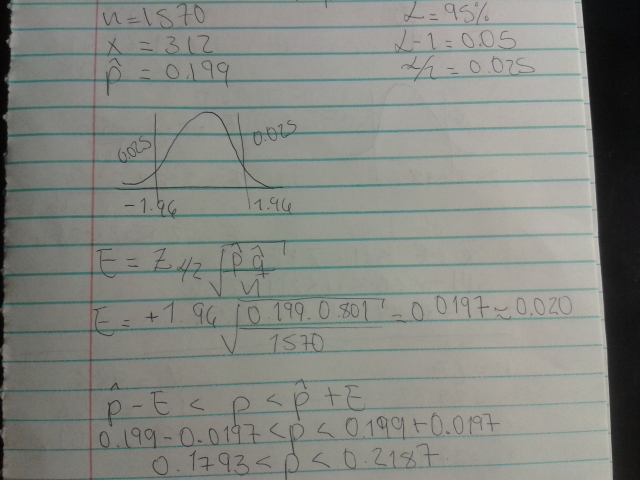
* + *Discrete data* represent items that can be counted. It takes possible values that can be listed out. For example, the number of heads in 160 coin flips takes on values from 0 through 160, but the number of flips needed to get 160 heads takes on values from 160 (the fastest scenario) on up to infinity (if you never get to that 160th heads).
  + *Continuous data*represent measurements. These values cannot be counted and can only be described using intervals on the real number line. For example, the exact amount of gas purchased for cars. With 20-gallon tanks would be continuous data from 0 gallons to 20 gallons, represented by the interval [0, 20].

Confidence Interval Estimates

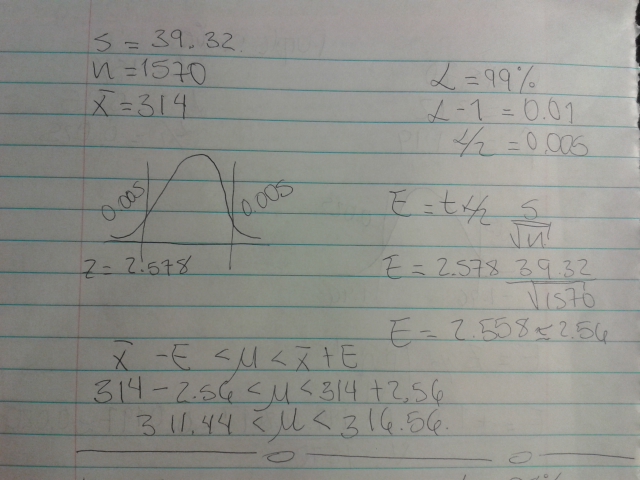
***Explain in general the purpose and meaning of a confidence interval***.

Confidence intervals are the ranges of values (intervals) that act as good estimates of the unknown [population parameter](https://en.wikipedia.org/wiki/Population_parameter); however, the interval from a particular sample does not necessarily include the real value of the population parameter. For example, when we say, "we are 99% confident that the true value of the parameter is in our confidence interval", we express that 99% of the hypothetically observed confidence intervals will hold the true value of the parameter. If a corresponding hypothesis test is performed, the confidence level is the complement of respective level of [significance](https://en.wikipedia.org/wiki/Statistical_significance). For example, a 95% confidence interval reflects a significance level of 0.05.

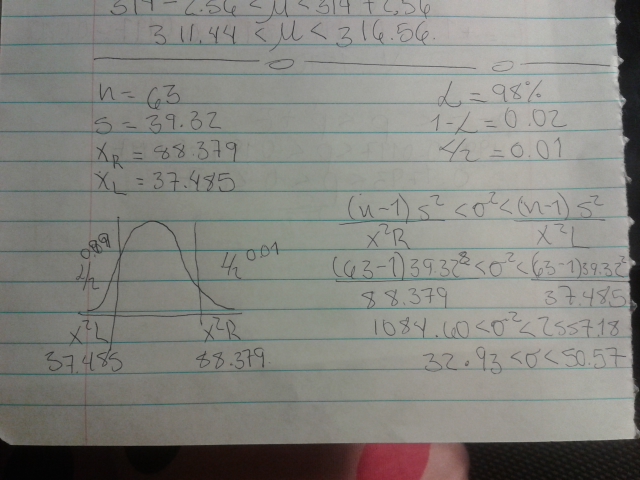
***Construct a 95% confidence interval estimate for the true proportion of purple candies.***



***Construct a 99% confidence interval estimate for the true mean number of candies per bag.***



***Construct a 98% confidence interval estimate for the standard deviation of the number of candies per bag.***

******

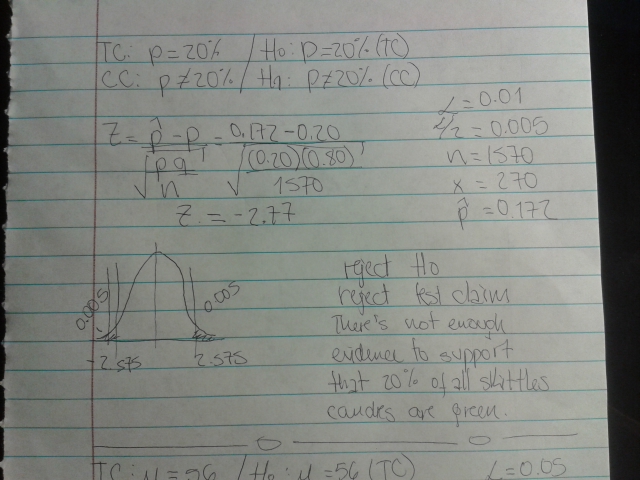
According to the results that we obtained, we can see in some cases how the interval estimates are low comparing it to the confidence level (95%,99% and 98%) this means that if a candy bag is opened, there is a 99% confidence level that the mean will be in between 311.44 and 316.06 (class data). This procedure of calculating confidence intervals helps us to estimate the chances that a set of data will fall in this interval. As a result, I can infer that the data that we obtained is similar to the confidence levels that we got in this problem (for example, sample mean obtained and mean number of candies per bag are similar 314)

Hypothesis Tests

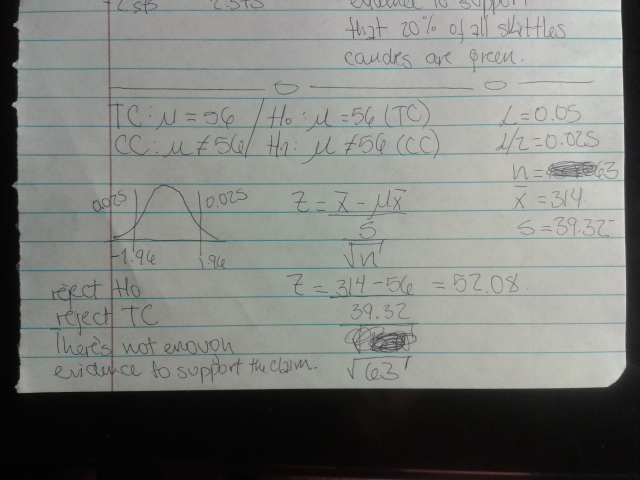
Explain in general the purpose and meaning of a hypothesis test.

A hypothesis test is a statistical test that is used to determine whether there is enough evidence or not in a sample of data to infer that a certain condition is true for the entire population. A hypothesis test examines two opposing hypotheses about a population: the null hypothesis (Ho) and the alternative hypothesis (H1). The null hypothesis is the statement being tested ". The alternative hypothesis is the statement you want to be able to conclude is true. Based on the sample data, the test determines whether to reject the null hypothesis. You use a p-value, to make the determination. If the p-value is less than or equal to the level of significance, which is a cut-off point that you define, then you can reject the null hypothesis.

***Use a 0.01 significance level to test the claim that 20% of all Skittles candies are green.***



***Use a 0.05 significance level to test the claim that the mean number of candies in a bag of Skittles is 56.***



Discuss and interpret the results of each of your two hypothesis tests. Include neatly written and scanned copies of your work.

***According to the results that we obtained during the hypothesis tests, we can see in both cases the test claimed was rejected. This, as a result of the data failing in the critical value area.***

Reflection

Conditions for doing Interval estimates and hypothesis test for population proportions.

Hypothesis tests

* The sampling method is simple random sampling.
* Each sample point can result in just two possible outcomes. We call one of these outcomes a success and the other, a failure.
* The sample includes at least 10 successes and 10 failures.
* The population size is at least 20 times as big as the sample size State the conditions for doing interval estimates and hypothesis tests for population proportions and discuss whether or not your samples met these conditions.

Interval Estimates

* The sampling method is simple random sampling.
* The sample is sufficiently large. As a rule of thumb, a sample is considered "sufficiently large" if it includes at least 10 successes and 10 failures.

My samples did meet this conditions. N was larger than 20 and there were two answers as a result

*Conditions for doing interval estimates and hypothesis tests for population means*

Interval estimates

* The sampling method is [simple random sampling](http://stattrek.com/Help/Glossary.aspx?Target=Simple%20random%20sampling).
* The [sampling distribution](http://stattrek.com/Help/Glossary.aspx?Target=Sampling_distribution) is approximately normally distributed.

Generally, the sampling distribution will be approximately normally distributed when the sample size is greater than or equal to 30.

Hypothesis test

* The sampling method is simple random sampling.
* The sampling distribution is normal or nearly normal.

Generally, the sampling distribution will be approximately normally distributed if any of the following conditions apply.

* The population distribution is normal.
* The population distribution is symmetric, unimodal, without outliers, and the sample size is 15 or less.
* The population distribution is moderately skewed, unimodal, without outliers, and the sample size is between 16 and 40.
* The sample size is greater than 40, without outliers

The samples met this conditions.

*Conditions for doing interval estimates for population standard deviations*

* The sample is a random sample
* The population must have normally distributed values (even if the sample is large).

My samples met this conditions.

By using this data maybe, the variants could have been wrongly counted so, the result might be wrong depending on the individual. In order to improve the sample method, it would have been better to use only a population, and not jump back and forth between sample data and population data.