Jeffrey Worley
Mr. Hutchinson
Math 1040 Statistics
1 August 2016

A Statistical Exploration of Skittles

Throughout the math 1040 course, I have learned many methods of data analysis and data representation. The goal of this paper is to examine the statistical distributions and properties of a sample of skittles. Alongside that, this paper is an opportunity and challenge to apply maths from the classroom to the real world, which is often the most difficult aspect of mathematics. The processes by which the statistical distributions and properties of Skittles will be found are several methods of data representation, 5 number summaries, confidence intervals, and Hypothesis testing.

Figure 1

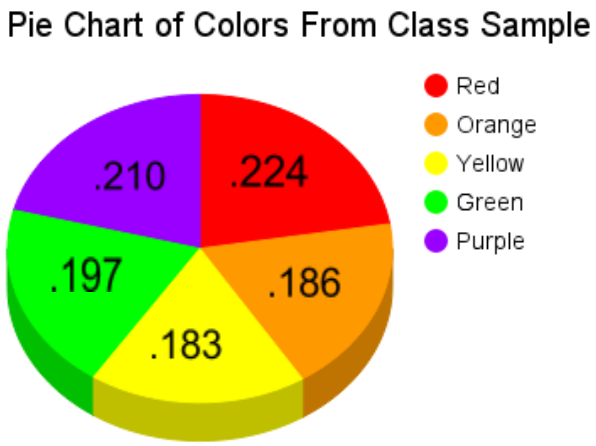
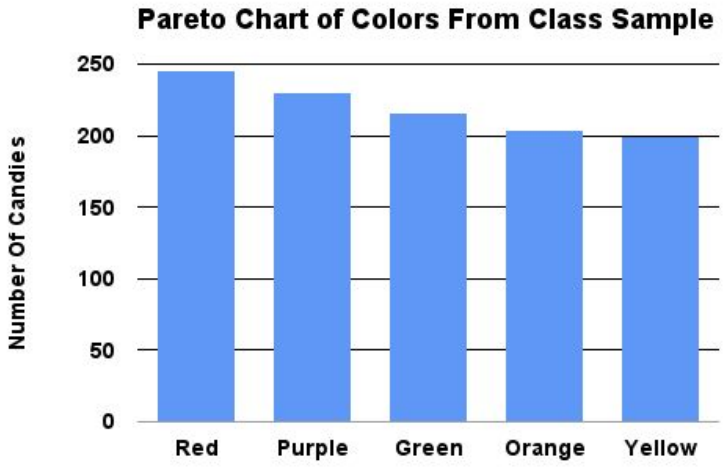



Figure 2



**Class Sample Values**

Table 1

| Color | Number of Candies | Proportion of Total |
|---|---|---|
| Red | 246 | .225 |
| Purple | 230 | .210 |
| Green | 216 | .197 |
| Orange | 204 | .186 |
| Yellow | 200 | .183 |

**Personal Sample Values**

Table 2

| Color | Number of Candies | Proportion of Total |
|---|---|---|
| Red | 16 | .258 |
| Purple | 13 | .210 |
| Green | 13 | .210 |
| Orange | 16 | .258 |
| Yellow | 4 | .064 |

Figure 1 shows the proportion of candies to the class total. In this graphical representation, one can see that the range of proportion is a mere 4%. This is significantly smaller than the roughly 20% range in my personal sample (Table 2). This is what I would expect, because as the sample size increases, I would expect the proportions of colors to approach uniform values. Additionally, the Pareto chart in figure 2 shows that while there is a difference in the amounts of candies per color, it's not as substantial as the difference in my sample. So, the data currently matches what I would expect to see in a sample that is expected to have uniform values for each color.

After inputting the individual amount of candies per bag into $L_1$ on my calculator, the five number summary, mean, and standard deviation were calculated using the 1-Var Stats function and are as follows:

**Minimum = 25**
**Quartile 1 = 57**
**Median = 59**
**Quartile 3 = 61**
**Maximum = 64**
**Mean = 57.7**
**Standard Deviation = 8.29**

Before I continue with a histogram and boxplot of the data, there is something obviously wrong with the data. The minimum of 25 seems to be an outlier compared to the rest of the data. The way to calculate outliers is any number outside of the range:

$$Q1 - (1.5 * IQR) \leq X \leq Q3 + (1.5 * IQR)$$

$$Interquartile\ Range\ (IQR) = Q3 - Q1$$

So, the range for normal values for our distribution is:

$$IQR = 61 - 57 = 4$$

$$57 - (1.5 * 4) \leq X \leq 61 + (1.5 * 4)$$

$$= 51 \leq X \leq 67$$

From this range of values, it can be concluded that the data entry of 25 candies in a bag is an outlier that is probably the result of an error in data collection. So, after throwing out that data point, the new 5 number summary, mean, standard deviation, and proportions are:

**Minimum = 56**

**Quartile 1 = 58**

**Median = 59**

**Quartile 3 = 61**

**Maximum = 64**

**Mean = 59.5**

**Standard Deviation = 2.53**

**Red = .224**

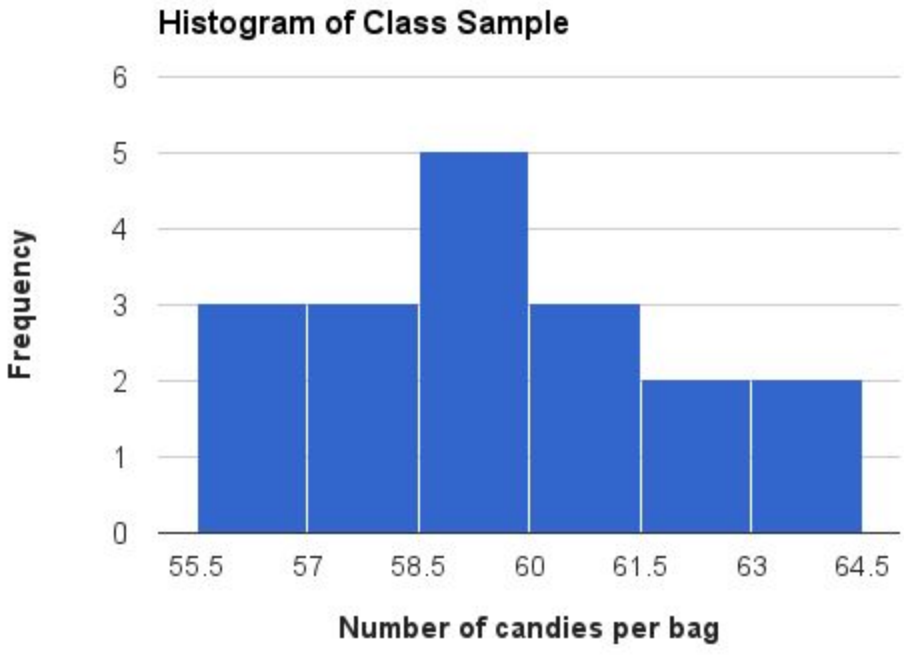
**Purple = .213**

**Green = .197**

**Orange = .184**

**Yellow = .182**

Figure 3



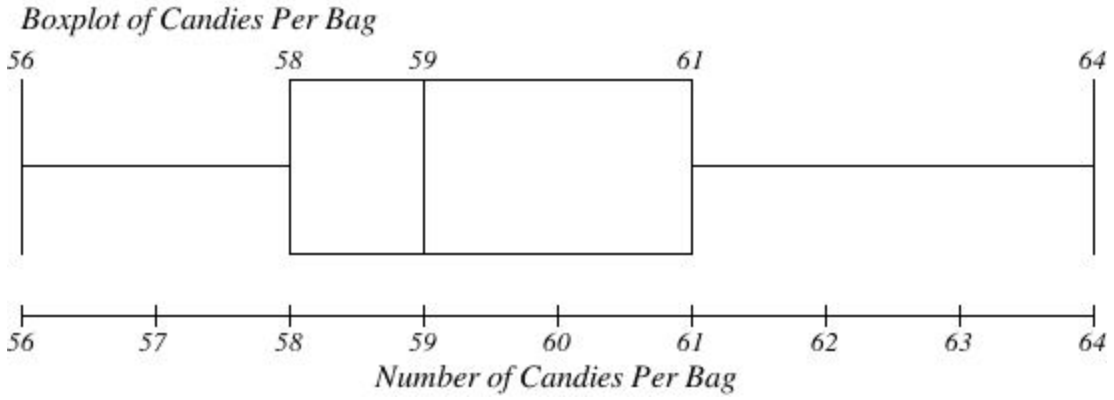



Figure 4

Figures 3 and 4 were made with a sample size of 18. The original sample size was 19, but the outlier was removed for accuracy. The Histogram in figure 3 and the Boxplot in Figure 4 both appear to be very slightly skewed right. The mode is slightly to the left of the center and there is a slightly longer tail to the right of the mode than to the left of the mode. This is somewhat surprising because I would assume the distribution to be approximately normal. However, this graph is barely skewed right, so I am very certain that if the sample size were to increase, the histogram and boxplot would normalize. My personal data of 62 candies in a bag does agree with the rest of class because it has a Z-score of approximately .988, which puts it within the range of normal z-scores.

Figures 1-2 and tables 1-2 represented categorical data. Categorical data is data that fits into a countable number of categories or groups. Some examples would be eye color of people, gender of individuals, and of course the color of Skittles. Some good methods of representing categorical data are bar graphs, pie charts, pareto charts, frequency polygons, and sometimes ogives depending on how one wants to present the data. These charts are good graphical representations of categorical data, because they represent the data in terms of frequencies and proportions. Frequencies and proportions are the only calculations that work with categorical data. This is because the only quantitative value of categorical data is the frequency and proportions.

Figures 3-4 represented quantitative data. Quantitative data is data that can be ordered and measured. Some examples of quantitative data are prices of coffee, gas mileage of cars,

and the amount of Skittles in a package of Skittles. Some good methods of representing quantitative data are histograms, frequency tables, dot plots, and stem and leaf plots. These charts are good graphical representations of quantitative data, because they show all of the data points in a way that the distribution of the data can be observed. The calculations that are useful for quantitative data are 5 number summaries and measures of center. This is because the data points for quantitative data do have a numerical value that can be used to find the overall distribution of the data.

Now that the data has been shown graphically and the sample statistics have been calculated, there are some calculations that can be done to get an idea of the population statistics. One way that this is achieved is confidence intervals. Confidence intervals are intervals that have a confidence level (represented as a percentage). These intervals are the range of values that one can be confident to a certain confidence level that the population statistic sits in. For example, after finding a sample mean, a confidence interval can be created to show a range of values that the population mean sits in with a given confidence level. Confidence levels are often 95% or 99%.

A 99% confidence interval estimate for the true population proportion of yellow candies can be calculated as such:

$$\hat{p} - E < p < \hat{p} + E \text{ where } E = Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$\hat{p} = sample\ p(yellow) = .184$$

$$n = sample\ size = 18 \text{ (n = 18 because the outlier was thrown out)}$$

$$\alpha = 1 - .99 = .01$$

$$Z_{\alpha/2} = Z_{.005} = 2.575$$

$$\therefore E = 2.575\sqrt{\frac{.184(1-.184)}{18}} = .235177068$$

$$\therefore .184 - .235177068 < p < .184 + .235177068$$

$$= -.051177068 < p < .419177068$$

$$\approx -.051 < p < .419$$

A 95% confidence interval estimate for the true population mean number of candies per bag can be calculated as such:

$$\bar{x} - E < \mu < \bar{x} + E \text{ where } E = t_{\alpha/2}\frac{s}{\sqrt{n}}$$

$$\bar{x} = 59.5$$

$$\alpha = 1 - .95 = .05$$

$$t_{\alpha/2} = t_{.025}$$

$$degrees\ of\ freedom = n - 1 = 18 - 1 = 17$$

$$\therefore t_{.025} = 2.110 \text{ (found with student t critical value table)}$$

$$s = sample\ standard\ deviation = 2.53$$

$$\therefore E = 2.110\frac{2.53}{\sqrt{18}} = 1.258249377$$

$$\therefore 59.5 - 1.258249377 < \mu < 59.5 + 1.258249377$$

$$= 58.24175062 < \mu < 60.75824938$$

$$\approx 58.242 < \mu < 60.758$$

A 98% confidence interval estimate for the true population standard deviation of candies per bag can be calculated as such:

$$\sqrt{\frac{(n-1)s^2}{\chi_R^2}} < \sigma < \sqrt{\frac{(n-1)s^2}{\chi_L^2}}$$

$$n = 18$$

$$degrees\ of\ freedom = n - 1 = 18 - 1 = 17$$

$$\chi_L^2 = 6.408 \quad \chi_R^2 = 33.409 \text{ (found using chi-square table)}$$

$$s = sample\ standard\ deviation = 2.53$$

$$\therefore \sqrt{\frac{(17)2.53^2}{33.409}} < \sigma < \sqrt{\frac{(17)2.53^2}{6.408}}$$

$$= 1.80473418 < \sigma < 4.120820813$$

$$\approx 1.805 < \sigma < 4.121$$

The following is the takeaway from the result of the confidence intervals. One can be 99% confident that the true population proportion of yellow candies is on the interval $-.052 < p < .418.$ This would make sense because the expected proportion would be .2, which is in that range. While proportions cannot be negative, confidence intervals for proportions can have negative values. Another conclusion is that on can be 95% confident that the true population mean of candies per bag is in the interval $58.242 < \mu < 60.758.$ This makes sense given our sample mean. The final conclusion is that one can be 98% confident that the true population standard deviation is on the interval $1.805 < \sigma < 4.121.$ This makes sense given our sample data.

Another method of testing statistical properties of a population is hypothesis testing. Hypothesis testing is the method by which a claim about a population can be tested by calculating if there is enough evidence to support or fail to support a claim. In a hypothesis test, there is the null hypothesis and the alternate hypothesis. To see if the null hypothesis should be rejected or fail to be rejected, one finds the significance of the alternate hypothesis. Once the hypothesis test is complete, either the null or the alternate is supported and the other is rejected.

For example, a hypothesis test for the claim that 20% of skittles are red given a .05 significance level can be calculated as such:

$$H_0 : p(red) = .20$$

$$H_1 : p(red) \neq .20 \quad \text{(two tailed test)}$$

$$test\ statistic\ (Z) = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{.224 - .2}{\sqrt{\frac{.2(.8)}{18}}} = \pm .2545584412$$

$$p - value = 2 * normalcdf(.2545584412,\ 1E99,\ 0,\ 1)$$

$$= 2 * .3995321337 = .7990642674$$

$$.7990642674 > .05$$

$\therefore$ fail to reject null because the p-value is greater than the significance level

There is sufficient evidence to support the claim that 20% of skittles are red

Another example is a hypothesis test for the claim that the mean number of candies per bag is 55 given a .01 significance level can be calculated as such:

$$H_0 : \mu = 55$$

$$H_1 : \mu \neq 55 \text{ (two tailed test)}$$

$$test\ statistic\ (t) = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{59.5 - 55}{\frac{2.53}{\sqrt{18}}} = \pm 7.546198851$$

$$p - value = 2 * tcdf(7.54619885,\ 1E99,\ 17)$$

$$= 2 * 4.004184553E - 7 = 8.008369107E - 7 = .0000008008369107$$

$$.01 > .0000008008369107$$

$\therefore$ reject the null because the p-value is less than the significance level

There is sufficient evidence to reject the claim that the mean number of candies per bag is 55

The first hypothesis test found that there is evidence to support the claim that 20% of all skittles are red. This would make sense as I would expect the total population of skittles to be 20% red. The second hypothesis test found that there is enough evidence to reject the claim that the mean number of candies per bag is 55. This would make sense given that no one in the class sample had a bag with 55 candies in it. This is shown in figures 3 and 4.

Throughout the above confidence intervals and hypothesis tests, there were a few conditions that needed to be known for the calculations to take place. Confidence intervals require either knowing your margin of error or your population and knowing whether you are calculating the intervals for the mean, standard deviation, or proportion. A hypothesis test requires knowing whether you are dealing with a population or a sample, the size of your population or sample, and whether you are testing for the mean, standard deviation, or proportion. A possible error that could exist with the data is a falsified or incorrect data entries. However, the one known outlier was thrown out for all the calculations for the confidence intervals and hypothesis tests.

The conclusions from all of these calculations is that the expected uniform distribution of colors of skittles is a definite possibility given the result of the confidence intervals and the hypothesis tests. On top of that, One can say with 95% confidence that the true population mean of the amount of skittles per bag is in the range of $58.242 < \mu < 60.758$. Another important conclusion is the fact that an increased sample size created more realistic results. Had the confidence intervals or the hypothesis tests been calculated with my personal sample data, the findings would not reflect the real world population statistics.

# Reflection

I think the most important thing that I will take away from this statistical analysis and the class as a whole is the education on critical thinking when it comes to data. Often times in the news you will hear about studies that seem preposterous. The skills developed in this statistics class and this statistical analysis have helped me look more in depth to see what studies are actually finding. I often find myself reading the actual publication of a paper now and seeing if the writer actually had enough of an understanding of statistics to fully understand what publications are finding. However, this was not the case before this class. Usually I would just accept an article at face value and not question whether or not the writer understood the mathematical analysis necessary to accurately communicate the findings of a publication to a reader. In the past week alone I have found a few studies that I was reading about where the author of the article completely misinterpreted the findings of the publications.

The mathematical skills that I have learned in this class have already come in handy for my next mathematics course. The final section in our summer homework was all about 5 number summaries, histograms, etc. As I was working with my friends, I had to help teach those that are taking statistics alongside this further math class next year. The assignment that took me about an hour to complete took them 3-5 hours to complete when they were trying to teach themselves the topic.

Statistics will especially be helpful next year, because my IB HL mathematics test will have a substantially large section on statistics. Statistics is so fundamental to the test that passing is not feasible without a solid understanding of statistics.