Intro

In my Math 1040 Group project we conducted a sample of 46 2.17 oz bags. In this sample we determined what kind of sample it was. Also, came up with a 5 number summary for the data. Along with that, there is visuals. Finally, confidence intervals. Overall, the project was to help us conduct a study and determine if it is reliable or not.

Group Project

After viewing all the data and graphs from the class as a total, as well as, my individual data we can see that comparing my bag from the proportion of the total doesn't represent the data. For example, I had a total number of 17 purple skittles making it 27.87% of my bag, whereas, comparing it to all 46 bags purple makes about 20.37% of all 46 bags. Judging from my 2.17 ounce bag of Skittles I logically thought that one color would be significantly less, but that isn't the case. For the most part, every color and count of skittles is proportionate. If we were able to do perhaps a sample of 100 bags of skittles I'm sure the proportions would be almost exact. To add on, in the 46 bag of skittles there isn't an outlier, on the whole all bags are between 53-66 skittles per bag. Overall, it's clear to say the total of skittles in the 46 bags does not match the distribution of my individual bag.

Data Collection: My 2.17 ounce Original Skittle bag

	Count Red	Count Orange	Count Yellow	Count Green	Count Purple
My Bag	11	18	8	7	17
Class Counts out of 46 bags	566	522	577	528	561

Total number of candies in my bag: 61

Total candies of all bags: 2754

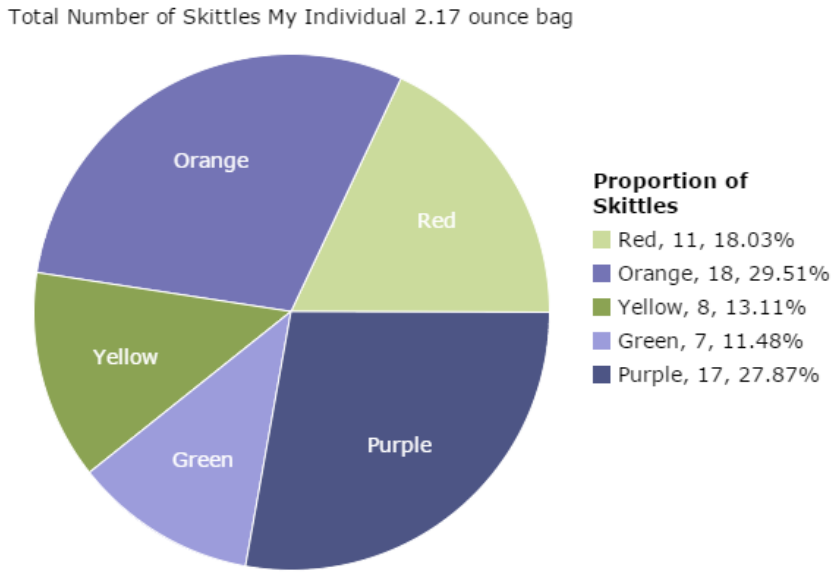
Class Data Collection: Forty-six (46) 2.17 ounce Original Skittle Bags

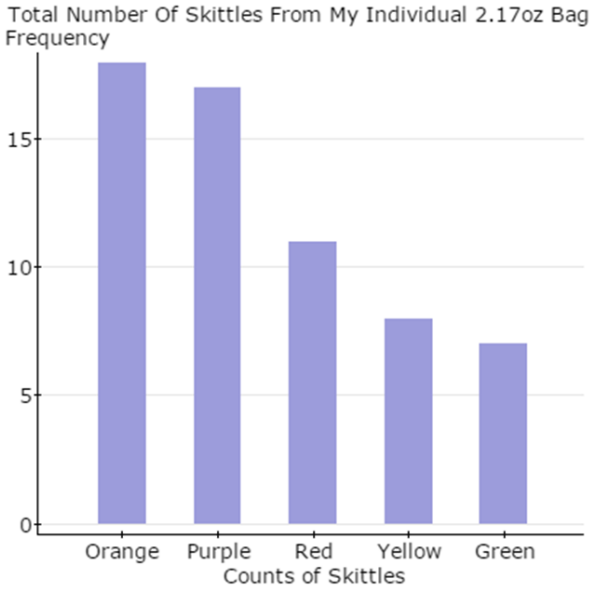
Hbhhb=My bag

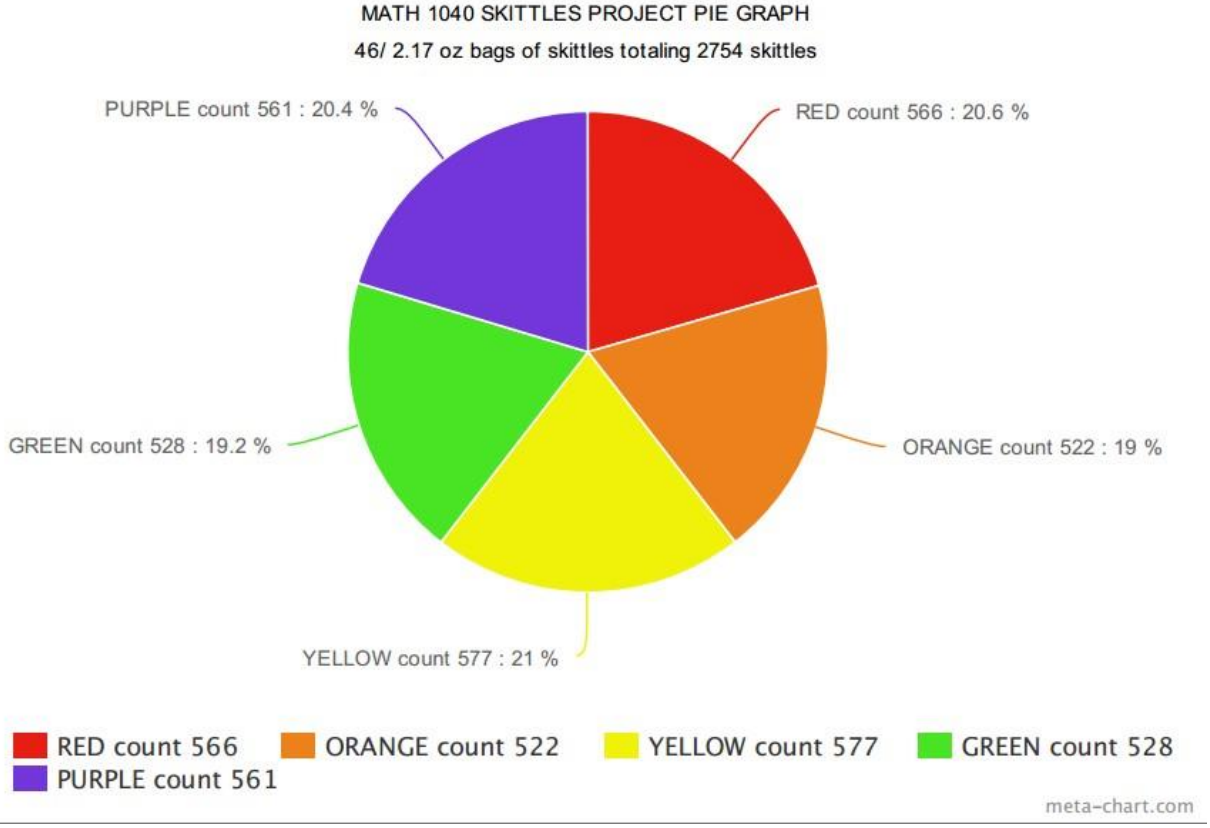
Bag		Count Red	Count Orange	Count Yellow	Count Green	Count Purple	Total Candies in each bag
1		19	12	17	8	10	**66**
2		19	12	17	8	10	**66**
3		14	13	15	9	11	**62**
4		7	13	13	13	15	**61**
5		14	14	7	15	10	**60**
6		12	10	18	7	16	**63**
7		10	11	10	16	16	**63**
8		11	18	8	7	17	**61**
9		4	6	13	18	19	**60**
10		18	12	9	9	11	**59**
11		16	6	14	12	12	**60**
12		11	13	15	14	7	**60**
13		11	13	15	14	7	**60**
14		6	16	12	14	12	**60**
15		12	11	10	13	13	**59**
16		13	10	16	11	8	**58**
17		7	19	9	16	9	**60**
18		12	8	11	18	11	**60**

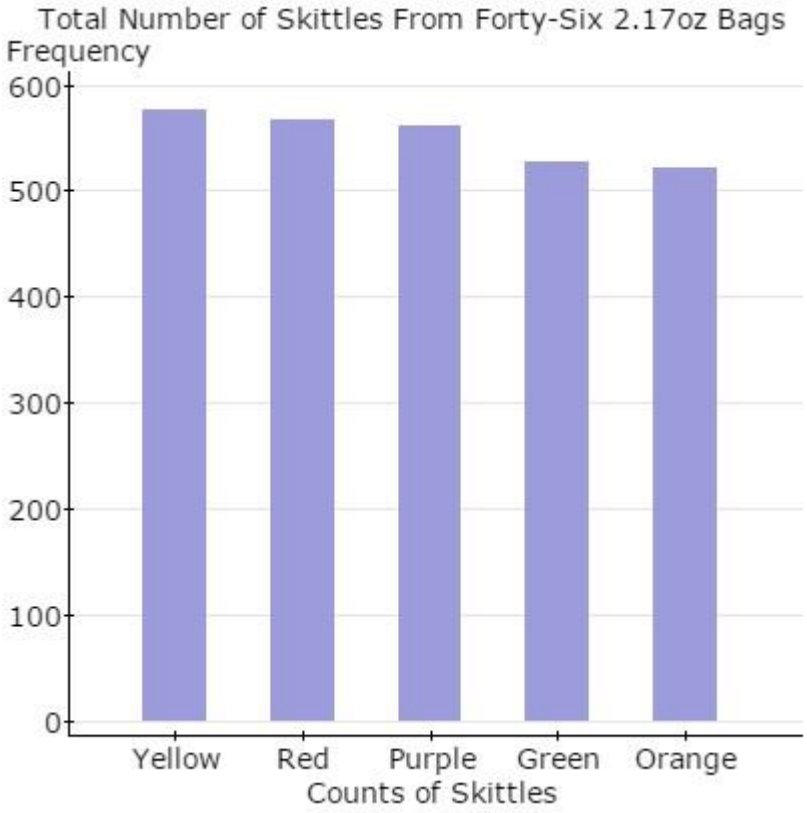
19		12	16	10	11	11	**60**
20		10	12	17	8	12	**59**
21		6	13	16	12	12	**59**
22		5	12	14	15	14	**60**
23		18	7	8	11	15	**59**
24		12	6	14	11	16	**59**
25		10	6	15	15	11	**57**
26		11	14	9	12	8	**54**
27		17	5	7	13	11	**53**
28		11	13	11	12	11	**58**
29		16	8	10	13	14	**61**
30		11	9	13	15	11	**59**
31		10	15	17	9	11	**62**
32		15	12	11	8	11	**57**
33		13	14	10	10	12	**59**
34		8	11	18	8	16	**61**
35		12	11	14	7	18	**62**
36		12	15	11	10	13	**61**
37		19	12	4	10	16	**61**
38		10	10	16	12	9	**57**
39		18	12	12	8	11	**61**
40		10	7	14	14	15	**60**
41		12	13	11	12	13	**61**
42		17	12	14	9	10	**62**
43		12	7	18	11	15	**63**
44		14	11	10	9	13	**57**
45		17	11	11	12	10	**61**
46		12	11	13	9	8	**53**
		Total: 566	**Total: 522**	**Total: 577**	**Total: 528**	**Total: 561**	**2754**

Individual Organized Data: My 2.17oz Bag of Original Skittles




Total Number Of Skittles From My Individual 2.17oz Bag
Frequency
15
10
5
0
Orange
Purple
Red
Yellow
Green
Counts of Skittles


MATH 1040 SKITTLES PROJECT PIE GRAPH
46/ 2.17 oz bags of skittles totaling 2754 skittles
PURPLE count 561 : 20.4 %
RED count 566 : 20.6 %
GREEN count 528 : 19.2 %
ORANGE count 522 : 19 %
YELLOW count 577 : 21 %
RED count 566
ORANGE count 522
YELLOW count 577
GREEN count 528
PURPLE count 561
meta-chart.com



Through the discussions, we determined that some of us expected skewed numbers in the color count of Skittles while some of us expected results that were similar to the total count of Skittles. This was definitely caused by our own personal experience with our own bag; we all expected what was in our bags to reflect the total count. This project is a perfect example of why repetition and proper sampling are so vital in research because numbers and data can be inaccurate or biased when only a small sample is observed. It gives us a good look at what “tunnel vision” can do to the actual set of facts.

After we all presented our assumptions about the data, we discussed which type of sample this was. In the beginning, most of us thought that it was a simple random sample. After a bit of debate, we ultimately decided that this didn't really fit the definition of a random sample. The definitions are as follows:

Simple Random - 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.

In the beginning, we thought that this sample was close to being a simple random, but the problem is that there is not a random or equal probability factor. The fact is, we all got our bags at convenient locations, all of which were in or around the state of Utah. This did not account for any sampling on the west coast, the mid west, the east coast, nor south or northern country boarder states. We didn't have any information on whether all the bags came from the same factory or if they came from different factories. This sample also did not account for the time period in which the Skittles were made. For all

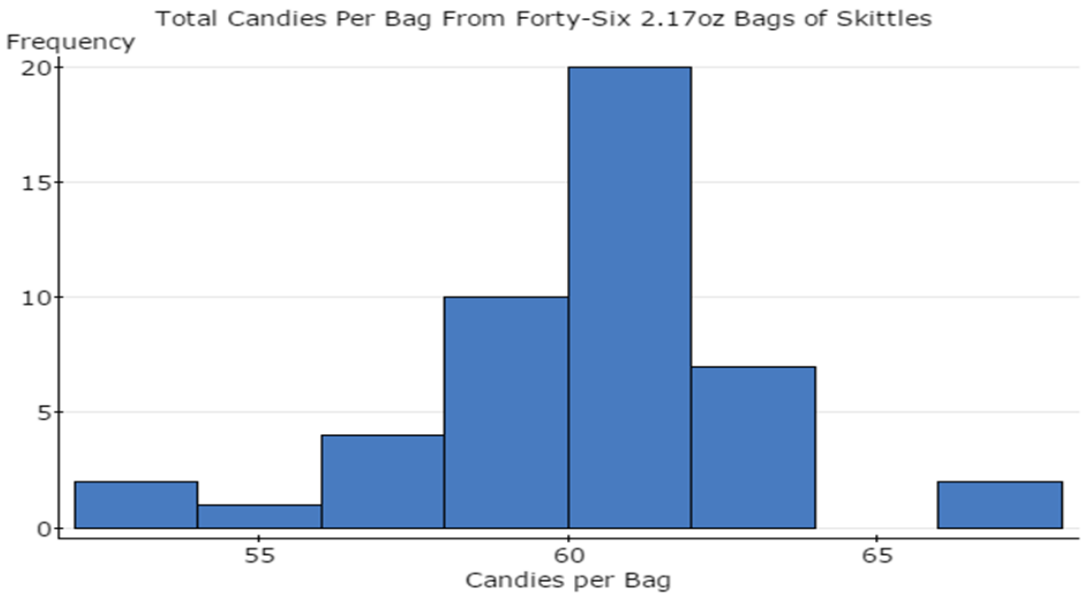
we know, we all bought from the same batch of Skittles. Another thing that we took into account was that the neither students nor stores were chosen at random. Students weren't randomly selected, we all were selected. The stores weren't randomly selected either. We didn't account for every single store in Utah that sells Skittles and randomly select which store out of all of those to go to, we selected the store that was the closest to us. After taking all of this into account, we decided that this is a convenience sample, which fits the definition below much better that the simple random sample.

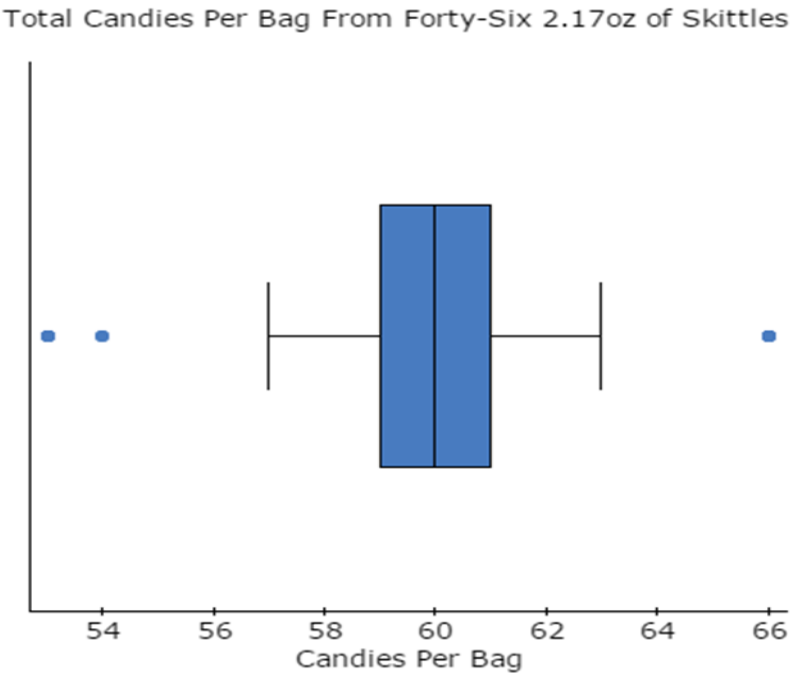
A **convenience sample** is one of the main types of non-probability **sampling** methods. A **convenience sample** is made up of people who are easy to reach. Consider the following example. A pollster interviews shoppers at a local mall.

This is a convenience sample because we all picked up our bags of Skittles at our local grocery store/gas station. We did not travel far or go out of our way to ensure we got accurate results; we did what was the most convenient for us.

Summary Statistics of Candies Per Bag

Mean	59.9
Standard Deviation	2.6
Minimum	53
Q1	59
Q2 (Median)	60
Q3	61
Maximum	66





1. The shape of the distribution judging from both graphs specifically the box plot, it is symmetrical. The boxplot does not have the whiskers more on the left or right meaning symmetrical, however from the box plot we do have three outliers. Also, the histogram, judging from the histogram alone it does appear to skew somewhat, but without a doubt the distribution is symmetrical.

I expected from the start to see a symmetrical distribution since usually name brands have the tech and workers to make sure that each bag of skittles is consistent, which saves money. Compared to, randomly putting skittles in a bag. The Overall data agrees with my own bag of Skittles, since my bag has 61 skittles in it, my bag wouldn't be considered an outlier. Overall, out of the 46 bags in this class sample my bag of 61 skittles is normal.

2. Categorical data or Qualitative data are observations corresponding to a qualitative variable. A qualitative variable allows for classification of individuals based on some attribute or characteristic. Whereas, quantitative data are observations corresponding to a quantitative variable.

Also, a quantitative variable provides numerical measures of individuals. The values of a quantitative variable can be added or subtracted and provided meaningful results. Basically, categorical or qualitative data deals with descriptions, data that can be observed, but not measured, colors, textures, smells, taste, appearance etc. On the other hand Quantitative date deals with numbers, data that can be measured, length, height, area, volume, weight, speed, time, humidity, sound levels, costs, members, age, etc.

Using categorical data or qualitative data, since it involves non numerical data bar graphs and pie charts are best. Using a pie charts and bar graphs can visually show the relationships. For example, for our group project part 2 we classified red to purple skittles in each bag using a bar graph and a pie chart, we saw that the counts were approximately the same amount. Therefore, Pie charts and bar graphs are preferred for using qualitative data.

In contrast, quantitative data which is numerical data it is much better to use a histogram or a stem plot. Histograms can show frequency and relative frequency which is important when organizing numerical data. Histograms allow you to see the distribution of the data, which a pie chart nor bar graph

could show. Stem plot is extremely useful for retrieving raw data and overall great for quantitative data. Although it does lose usefulness when data set is too large.

In conclusion, quantitative data and categorical data or qualitative data are quite different. Qualitative data allows for classification of individuals based on some attribute or characteristic. Whereas, quantitative data involves numerical measurement. Histograms and stem plots are better graphs involving quantitative data. Finally, bar graphs and pie charts are better for qualitative data. Both quantitative data and qualitative data involve organizing data, but they organize different kinds of data.

Confidence Intervals

Confidence Intervals are constructed at a confidence level. Also, Confidence Intervals are a type of interval estimate of a population parameter. To add on, A Confidence Interval is for an unknown parameter that consists of an interval of numbers based on point estimating. Concluding, the Confidence Interval represents the expected proportion of intervals that will contain the parameter if a large number of different samples is obtained.

1. TOTAL YELLOW CANDIES IN POPULATION = 577/2754 This is .2095

n=2754 x=577 P hat= .2095 alpha=.01

Critical value $Z^{\alpha}/2 = 2.576$ (calculator function DISTRIBUTION, OPTION 3 INVNORM enter the percentage to the right and close in the parentheses)

Margin of error is +/- 2% E= $Z^{\alpha}/2 * \sqrt{Phat(1-Phat)/n}$

Now using formula P hat - $Z^{\alpha}/2 * \sqrt{Phat(1-Phat)/n} \leq p \leq$ P hat + $Z^{\alpha}/2 * \sqrt{Phat(1-Phat)/n}$ We can calculate the lower and upper bounds of the confidence interval.

$.2095 - 2.576 * \sqrt{.2095(1-.2095)/n} = .1895$

$.2095 + 2.576 * \sqrt{.2095(1-.2095)/n} = .2295$

We can say with 99% confidence that the number of yellow candies in a population proportion of skittles is going to be between 18.95% and 22.95% of the population proportion with a margin of error of +/- 2%.

2. THE POPULATION MEAN NUMBER OF CANDIES PER BAG

n=46 X bar= 59.87 $\alpha = .05$ s= 2.62

Critical value $t^{\alpha}/2 = 2.01$ (calculator function DISTRIBUTION, INVT, area left is .975 and the DF is 1-n or 45)

Margin of error is +/- .78 calculated using formula E= $t^{\alpha}/2 * S/\sqrt{n}$

Now using formula X bar +/- $t^{\alpha}/2 * S/\sqrt{n}$ We can determine the interval of mean number of candies.

$59.87 - 2.01 * 2.62/\sqrt{46} = 59.09$ $59.87 + 2.01 * 2.62/\sqrt{46} = 60.64$

We can say with 95% confidence that the mean number of candies per bag of skittles is going to be between 59.09 and 60.64, with a margin of error of .78 candies.

3. POPULATION STANDARD DEVIATION OF THE NUMBER OF CANDIES PER BAG.

n=46 s=2.62 $\alpha = .02$ $X^2R = 76.154$ $X^2L = 29.707$

Critical values X^2R $\alpha/2$ and X^2L $\alpha/2$ were obtained by using the Formulas and Tables sheet by Mario Triola.

We now use the formula $(n-1)s^2 / X^2R < \sigma < (n-1)s^2 / X^2L$ To calculate the confidence intervals for the standard deviation of the number of candies per bag.

$(46-1)2.62^2 / 76.154 = 2.01$ $(46-1)\ 2.62^2 / 29.707 = 3.22$

We can say with an 98% confidence that the population standard deviation of the number of candies in a bag of Skittles is between 4.06 and 10.40.

Reflection

Statistics is the practice or science of collecting and analyzing numerical data in large quantities, especially for the purpose of inferring proportions in a whole from those in a representative sample. I have learned that Statistics is more important and useful than just organizing data. Statistics is actually really important in predicting outcomes and what we can expect and be able to draw the best conclusions. Also, it aids in collecting trust worthy data. As well as, to analyze data appropriately. It's easy to look at a graph and believe it, because it looks convincing. However, since I've take statistics it's really helped me to double check it. For example, I look at who created the graph as well as the starting point, and draw to a reasonable conclusion. What have found out is I hardly believe any polls going on especially with the election on the way, because I can easily look at it see some error. Knowing statistics and using it in real life is the best tool anyway can have in making a decision.

I'm currently a sociology major and undecided on my second major. I plan to work in a field of work similar to my internship at March of Dimes which is creating templates and finding funding for the program. What I have found out taking statistics is I actually do statistics every day at March of Dimes and it is vital. I have helped in creating presentation and since taking the course I learned some skills in making my charts look favorable, which I'm proud to say have gained a lot of sponsors. Anyone who is familiar with statistics is someone I would trust in any company.

Finally, the team project which had many parts really helped me in conducting and recognizing a reliable study. If I saw team study on the news I would probably ignore it considering it is convenient study meaning it is not a reliable study, since the participants were found on convenience. Statistics is a vital tool for everyday life like deciding if a study is reliable or working in a company where it's your job to gain funding by making graphs that are favorable.