Justin Peck

Math 1040

**Introduction**

Everyone in the class was asked to count and color coordinate skittles from a single bag of skittles. Our individual numbers were then combined into a single spreadsheet and were used for statistical analysis. From this data we were able to use information learned during the semester to construct graphs and draw conclusions using confidence intervals and hypothesis testing.

|  |
| --- |
| Math 1040 Class Skittles Proportions |
| Color | Count | Proportion of Total |
| Red Skittles | 564 | 0.199 |
| Orange Skittles | 564 | 0.199 |
| Green Skittles | 566 | 0.199 |
| Purple Skittles | 559 | 0.197 |
| Yellow Skittles | 586 | 0.206 |
| Total Number of Skittles in the class | 2839 | 1 |

**Organizing and Displaying Categorical Data: By Color**



Does the Class data represent a random sample?

Yes, the class data does represent a random sample. Although each student was asked to buy their own bag of skittles and not every bag of skittles in the region had an equal chance of being selected, the distribution of skittles from the central plant/warehouse was most likely random. The skittles company most likely does not count colors as they load the bags and simply loads by weight, and assuming students did not make any biased decisions about which bag to grab off the shelf every bag produced had an equal chance of being shipped to any location in the country and being selected at random by a student in the class.

What would the population be?

In this study, the sample is the class data. Since not everyone in the class is currently living in the same state, the population would be all 2.17 ounce skittles bags in the United States. There are currently different manufacturing plants operating overseas, therefore the population can only reasonably be expanded to include the United States distribution circuit.

Individual Part 2

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|   | Red  | Orange  | Yellow | Green | Purple | Total |
| Individual | 16 | 12 | 16 | 6 | 9 | **59** |
| Relative Frequency | 0.27 | 0.20 | 0.27 | 0.10 | 0.15 |   |
|   |   |   |   |   |   |   |
| Class Totals | 564.00 | 564.00 | 586.00 | 566.00 | 559.00 | 2839 |
| Relative Frequency | 0.20 | 0.20 | 0.21 | 0.20 | 0.20 |   |

When I did my initial sample of the one pack of skittles I was surprised to see the large gap between green(6) and red/yellow(16). This was further seen when I calculated the relative frequency. When I received the class data I could see small differences in the numbers, but the relative frequencies were all within 1% of each other. This was more of what I expected the initial outcome to be. I believe this shows how sample size can greatly impact the results. The smaller sample was probably not the best indicator of the whole population. When more individuals were added to the sample, it can be assumed that the population is better characterized. This is also taking into account simple random sampling.

**Organizing and Displaying Quantitative Data: The Number of Candies per Bag**

 

The next step in our class was to analyze this data and formulate means, standard deviations and 5 number summaries. The results to these questions are listed below.

(a) The mean number of candies per bag is 59.1 candies.

(b) The standard deviation per bag is 6.4 candies.

(c) The 5-number summary is 34-58-60-62-71.

**Findings about the variable “Total candies in each bag”**

Thus far this assignment has allowed me to apply what we are learning in the assignments to something concrete. The frequency histogram indicates a generally symmetric shape with the peak being the range of 60-62. The histogram also indicates several outliers in the 34-36 range. When comparing these results to my bag of candy, I was able to see my bag of 58 skittles is the median of the lower 25% of the total class values. I found this by using the data from the box plot. The box plot shows that Q1 to be at 58 skittles. From this data I would be able to say this is generally what I expected to find. Had my original results been one of the outliers, I would have to say that the data did not correspond with my original findings. This is why it is important to draw conclusions from a larger sample.

**Difference between categorical and quantitative data**

The easiest way to define categorical data and quantitative data is to explain them both separately. Categorical data, or in other words, Qualitative data is mainly used to put variables into categories based on properties. These properties can be anything from male/female, salty/sweet or even what kind of car an individual drives. Quantitative data on the other hand, are data, which are measurable or quantifiable. This is data such as weight, height, or number of skittles per bag. The type of graph used to represent the data is very important. For categorical data, it is best to use graphs that separate the data into different categories. Graphs that can do this are bar graphs and pie charts. The bar graphs will break out the different categories and the height will usually give the frequency of the occurrence. The pie chart is a quick effective way of showing the difference between categories frequency. There are many graphs that can be used for quantitative data. Graphs such has histograms, stem and leaf plots, or box plots are just a few. These graphs separate that data into numerical categories. They will give you an idea of the distribution of data and also give you values such as the Median. As with the graphs, Categorical data can use calculations such as frequency and relative frequency. Quantitative data uses a wide variety of calculations that include mean, median, range, standard deviation, and many others. These calculations help decipher and analyze the data to help draw conclusions. With each experiment it is important to understand the types of data and also what calculations and graphs can be used to draw conclusions.

The final portion of this semester long project was to start making conclusions using hypothesis testing and confidence intervals. This allows us to make predictions about a population using the data from our sample. The information below provides the steps taken to draw the conclusions. To help understand these results, I will first define what a confidence interval is.

Confidence intervals are an estimate range of values calculated from a sample. These range of values are likely to include an unknown population parameter. Confidence intervals allow researchers to use data from an experiment or sample and apply them to larger populations with a certain level of confidence. These can be calculated at varying percentages including 90, 95, 98, and 99. When a 99% confidence interval is used we are saying that we are 99% certain that the true value of a population parameter will fall between the lower and upper bounds of the confidence interval.

**99% Confidence Interval estimate for the population proportion of yellow candies**

X= 586 n= 2839 Z-value for 95% CI = 2.576

p= 586/2839 = 0.206

0.206 +/- 2.576 \* (0.007596) 0.206 +/- 0.01957

99% Confidence Interval Estimate: (0.186, 0.226)

Confidence Intervals estimated from a population proportion are used to determine, with the specified degree of confidence, the proportion of a characteristic found within a population. In relation to the skittles, we are 99% confident that the proportion of yellow skittles in any bag of skittles falls between 0.186 and 0.226.

**95% Confidence Interval estimate for the population mean number of skittles per bag**

n= 49

Sx = 6.38

Sample mean= 59.15

Standard error of the mean = 0.9114

To find the t-value, a t-table was consulted using a degree of freedom of 50. The t-value is 2.009.

59.15 +/– t\*(0.9114)

59.15 + 1.83 = 60.98

59.15- 1.83 = 57.32

95% Confidence Interval Estimate: (57.32, 60.98)

Confidence Interval estimates of the population mean use sample date to extrapolate an interval with the specified degree of confidence that the mean characteristic of a population should fall within. In this case, we are 95% confident that the mean number of skittles in any bag is between 57.32 and 60.98.

  

**98% confidence interval estimate for the population standard deviation of the number of candies per bag**

 

n=49

s=6.378

S2=40.679

χ2 1-a/2 = 0.99

χ2 a/2 = 0.01



On the Chi square distribution chart, 50 degrees of freedom was used. The value for

χ2 1-a/2 was 29.707.

For χ2 a/2 it was 76.154.

√[ s2(df)/Chi value]

|  |
| --- |
| Lower bound: 5.06 |
| Upper bound: 8.11 |

Confidence Interval estimates from the population standard deviation use the sample standard deviation in order to generate an interval that the population standard deviation of the number of candies should fall within, with the specified level of confidence. In this case, we are 98%confident that the population standard deviation is within 5.06 and 8.11 candies. The problem with confidence interval estimates taken from the sample standard deviation is that the sample standard deviation may be quite different from the actual population standard deviation.

REFLECTION

During this semester the group skittles project has allowed me to think outside the box and apply what I am learning to the real world. These include Items such as graphing, probabilities and confidence intervals. Each project had something new to learn and relied on us to figure out how to apply them to our data. This process helped me build on my problem solving abilities to figure out what needed to be done. An example of this is graphing. I thought the first assignment was going to be extremely easy, but quickly found out that manipulating graphs to best represent the data could be quite tricky. I now understand the importance of graphs and how they can be the most effective way to display your results. I also learned this can also be done completely wrong and can help “dishonest add companies” ;) sell a product, by presenting their results in a way that is most beneficial to THEIR objective.

 The thing I love most about statistics is that I can see them being used everywhere I go. Just today I was at a class for work and there was a big poster that read “BETWEEN 85%-99% OF ALAMRS ARE NOT LIFE THREATENING”. I immediately thought of confidence intervals and could guess how they came up with those results. This is just one example of the many that I have come upon this semester. Having these skills is specifically important for me because I want to go into the medical profession. I have to be able to read and analyze the results of various studies. I would then have to decide based on the results if it would be applicable to use with a patient. I have learned so much in this class and I know that I will be able to apply these skills in so many different ways.