Skittles Color Vs. Mass Statistics

Jaymie Woolsey

March 01, 2024

Introduction

The purpose of this exploratory data analysis project is to use R to learn about possible relationships of statistics of a particular interest. This analysis includes exploring data from mini bags of Skittles brand candy that were collected by 133 students at Chico State University from Spring 2021- Spring 2023. There are 5705 pieces of Skittles candy total (observations), and we will be exploring the relationship between Color and Mass (variables). The research question to be explored will be related to the composition of color dye and how it affects candy mass. Here, I hypothesized that the Yellow Skittles Candy will have significantly consistent less mass than the other colors in Skittles Candy. This hypothesis came about due to Yellow being a primary color that can be mixed in varying amounts across the other candy colors.

First, we need to load the packages and tools we need in the first code chunk.

```
knitr::opts_chunk$set(warning=FALSE, message=FALSE, fig.height=4, fig.width=5, fig.align='center')
library(tidyverse)
library(readxl)
library(modeest)
library(ggplot2)
```

Second, we must load our data.

```
historical_candy <- read_excel("historical_candy_data.xlsx")
```

Univariate Variable Analysis

Variable 1: Color

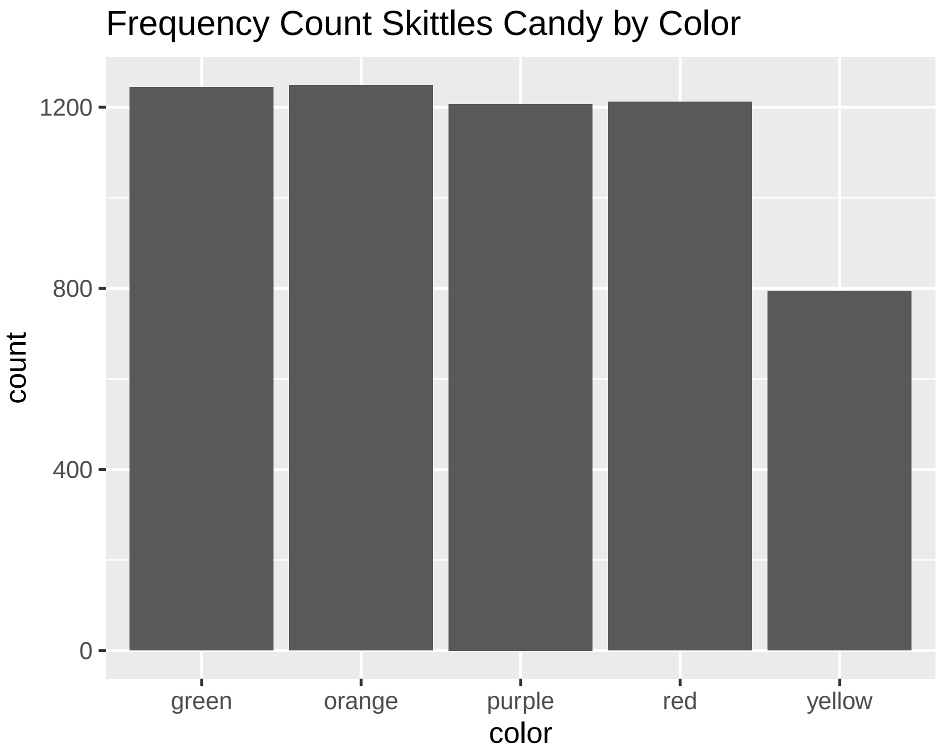
There are 5 colors that are currently produced by the Skittles brand company to be distributed to the public. Knowing how many of each Color is produced might give us some helpful information. We can explore the frequency of each Color in our data set as follows:

```
table(historical_candy$color)
```

```
##
##  green orange purple    red yellow
##   1244   1248   1207   1212    794
```

The table shows that the Yellow candy is the least frequent Color produced. We can further visualize this data using a bar chart.

```r
ggplot(historical_candy, aes(x=color)) + geom_bar() +
  labs(title="Frequency Count Skittles Candy by Color")
```



First, a table summarizing the individual counts of each Color of Skittles was made to visualize any significant difference in data. And as we can see here, using the bar plot helps visualize that the Yellow candy seems to have a significantly lower count historically than the other colors.

Variable 2: Mass

The Mean of the Masses of the total Historical Candy Data in grams might give us some helpful information:

```r
# get the mean
historical_mean <- mean(historical_candy$mass)

# print the results
print(paste("Historical mean:", historical_mean))
```

```
## [1] "Historical mean: 1.0576691849255"
```

The overall Historical Mean looks to be around 1.057 grams.

But let's also print the Median and Mode just for fun:

```r
# get the median
historical_median <- median(historical_candy$mass)

getmode <- function(v) {       # create a getmode function to get the get the statistical mode
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}

# get the mode (using our getmode function)
historical_mode <- getmode(historical_candy$mass)


# print the results

print(paste("Historical median:", historical_median))
```

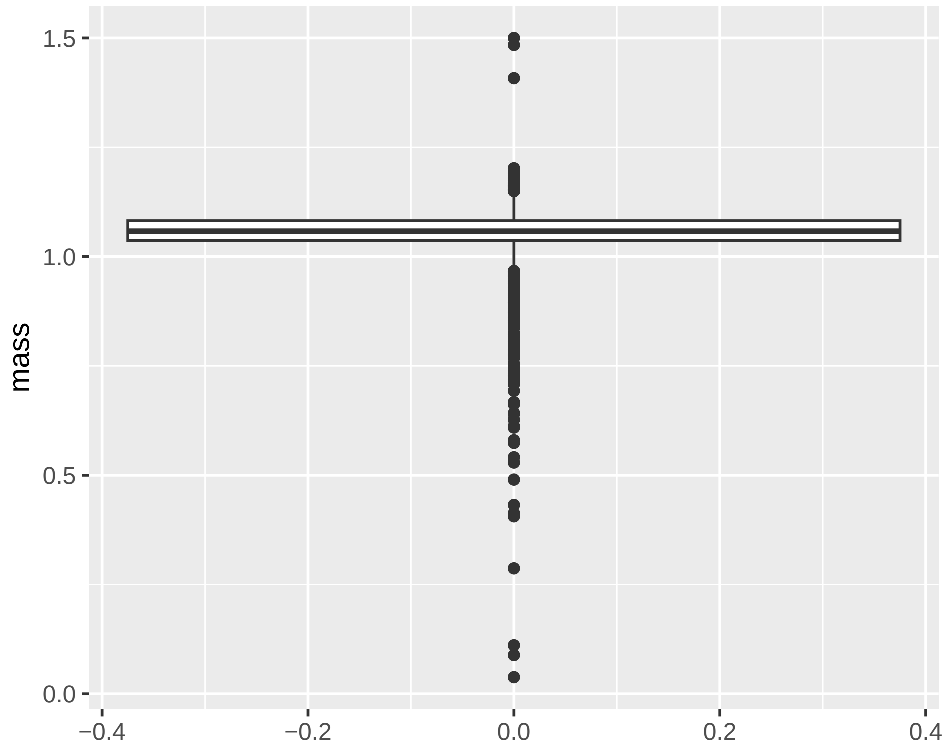
```
## [1] "Historical median: 1.058"
```

```
print(paste("Historical mode:", historical_mode))
```

```
## [1] "Historical mode: 1.042"
```

Visualizing the Mean:

```
ggplot(historical_candy, aes(y=mass)) + geom_boxplot()
```



Most of the candy pieces of the total Historical Candy Data set fall into an overall Mean of about 1.057 grams, with a few outliers distributed above and below this number. Now, we can take Variable 1 (Color) and Variable 2 (Mass) and combine the two variables and explore whether or not our Hypothesis seems to be true.

Bivariate Exploration

The 2 variables under consideration here are Color (Yellow, Red, Purple, Orange, Green) and Mass (in Grams).

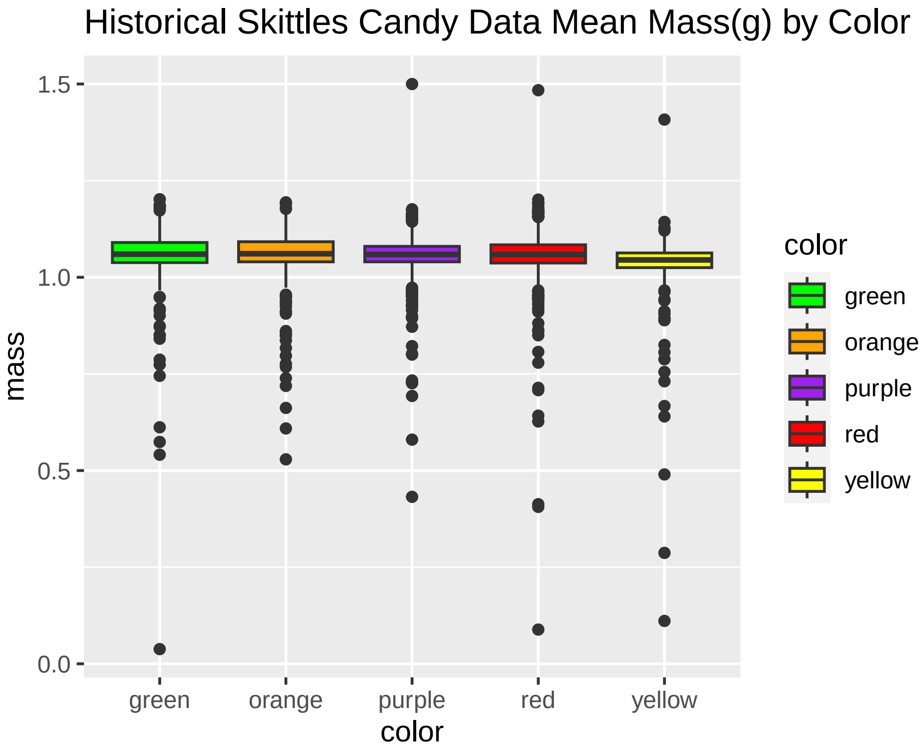
Combining the 2 variables, we can calculate the Mean (and Median and Mode just for fun) of the Masses of each individual Color in grams:

```
getmode <- function(v) {      # create getmode to get the get statistical mode
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}


individual_color_stats <- historical_candy %>% group_by(color) %>%

  reframe(color_mean = mean(mass, na.rm=TRUE),     # get mean
            color_median = median(mass, na.rm=TRUE),  # get median
            color_mode = getmode(mass)        # get mode
            )

individual_color_stats
```

```
## # A tibble: 5 x 4
##   color  color_mean color_median color_mode
##   <chr>       <dbl>        <dbl>      <dbl>
## 1 green        1.06         1.06       1.04
## 2 orange       1.06         1.06       1.04
## 3 purple       1.06         1.06       1.06
## 4 red          1.06         1.06       1.06
## 5 yellow       1.04         1.04       1.04
```

Comparing the Means (average) of each individual Color, we observe that the Yellow candy average Mass in grams is lower than the other colors.

```
ggplot(historical_candy, aes(x=color, y=mass, fill=color)) + geom_boxplot() +
  scale_fill_manual(values=c("green", "orange", "purple", "red", "yellow")) +
  labs(title="Historical Skittles Candy Data Mean Mass(g) by Color")
```



Furthermore, in visualizing this combined data with boxplots side to side, we can see that the Mean Mass in grams for Yellow is visibly lower than the other colors.

Conclusion

My hypothesis is that Yellow Skittles candies have less Mass in grams than the other Skittles colors. This is thought to be due to it being a primary color and therefore does not require mixing with other colors and substances which would contribute to a higher candy mass. There are also less of the Yellow candies overall, which may further support this hypothesis. And in general, the Skittles candy is very consistent in Mass, with there being just a few outliers that are visibly displayed in the boxplots. In addition, there are also more outliers that are smaller than the Mass Mean than larger- and this may be due to the mechanical breaking down of some of the candies in the bags when they are being transported or moved around. When thinking in terms of quality control for a big company, one could see how creating a product consistent in size, flavor, ingredients, nutritional content, and things of that nature would be important to a manufacturer. Being cost efficient would be a consideration as well. Hence, it would make sense if Skittles company intentionally uses Yellow dye across other candy colors to be more efficient in costs and consistent in product mass, flavor, and ingredients. My remaining questions are why are Orange Skittles seemingly the heaviest, and why don't they make Blue Skittles as part of their standard color set since blue dye would hypothetically be needed to create their Green and Purple candies. But to know for sure, one would have to really get ahold of the CEO of Skittles.