My M&M OCD

Yoni

14 04, 2025

Intro

Objective of Simulation

The goal of this simulation is to test the statistics of M&M and other stacks even Chocolate lentils by color.

I wanted to know what is the chance of my finishing the package of M&M without mixing any color in one bite, eating them 2 by 2

In addition, here are some BI incite that needed to be checked:

1. What is the probability of M&M packages packaged fairly?
2. What is the probability of M&M packages packaged without one color?
3. How does the size of the package or number of colors affect this probability?

Method

I do not have an inner information of how does (M&M)[https://www.mms.com/en-us] make their delicious snacks nor we know how they make sure each package have fair amount of each color.

Therefore, the method I chose is based of simulation of some M&M bags, according to the most common sizes of packages.

Each time we sample x lentils(units of M&M), name them by colors (V1,V2…), and see the results for many packages as a statistic data.

My hypothesis is that the probability of perfect package (aka a package with all colors number been even) is very small, at least for a standard 6 colors pack.

Parameters

Basic parameters:

```
n= 800                #numbers of bags per sample
n_color= 6            #unique colors of M&M
gram= 0.91            #weight of one M&M
bag_g= 250            #common weight of M&M package
n_unit= bag_g/gram    #M&M per packagenm,
```

```
## [1] "The avarage number of lentils per color is 45.79"
```

Creating the Sample

General Sample

In order to test the theoretical data, we need to simulate it using customize functions. here are there:

- Create_bag- function to create one snack package for chosen package size and number of colors.
- sample_MnM- function to create n bags from the Create_bag function.

```
## [1] "One bag of 100:"
```

```
##       1  2  3  4  5  6
## [1,] 17 14 13 17 26 13
```

```
## [1] "3 bags of 100:"
```

```
##       Red Blue Green Orange Yellow Brown
## Bag_1  19   14    16     13     20    18
## Bag_2  17   14    17     15     20    17
## Bag_3  10   25    11     24     13    17
```

Preview Graph

Now will be creating n bugs of M&M

columns:

1. V1:V6- the number of lentils per color
2. even_count- how many evens colors there are
3. even_evens- are the uneven colors even
4. Variance- variance of lentils per color
5. low_col- sum true if one color's count is lower than $\frac{2}{3}$ of expected value
6. min- the lowest color in each row

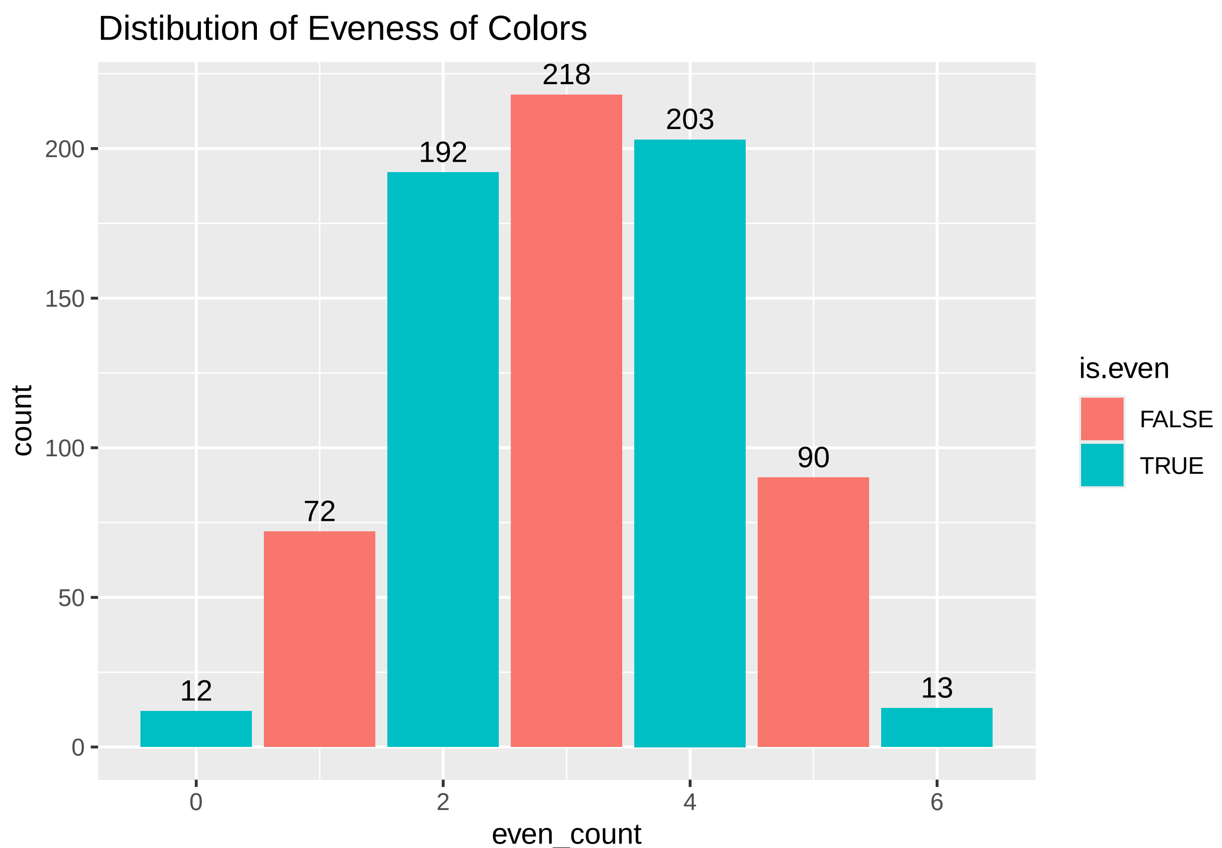
here are the first rows:

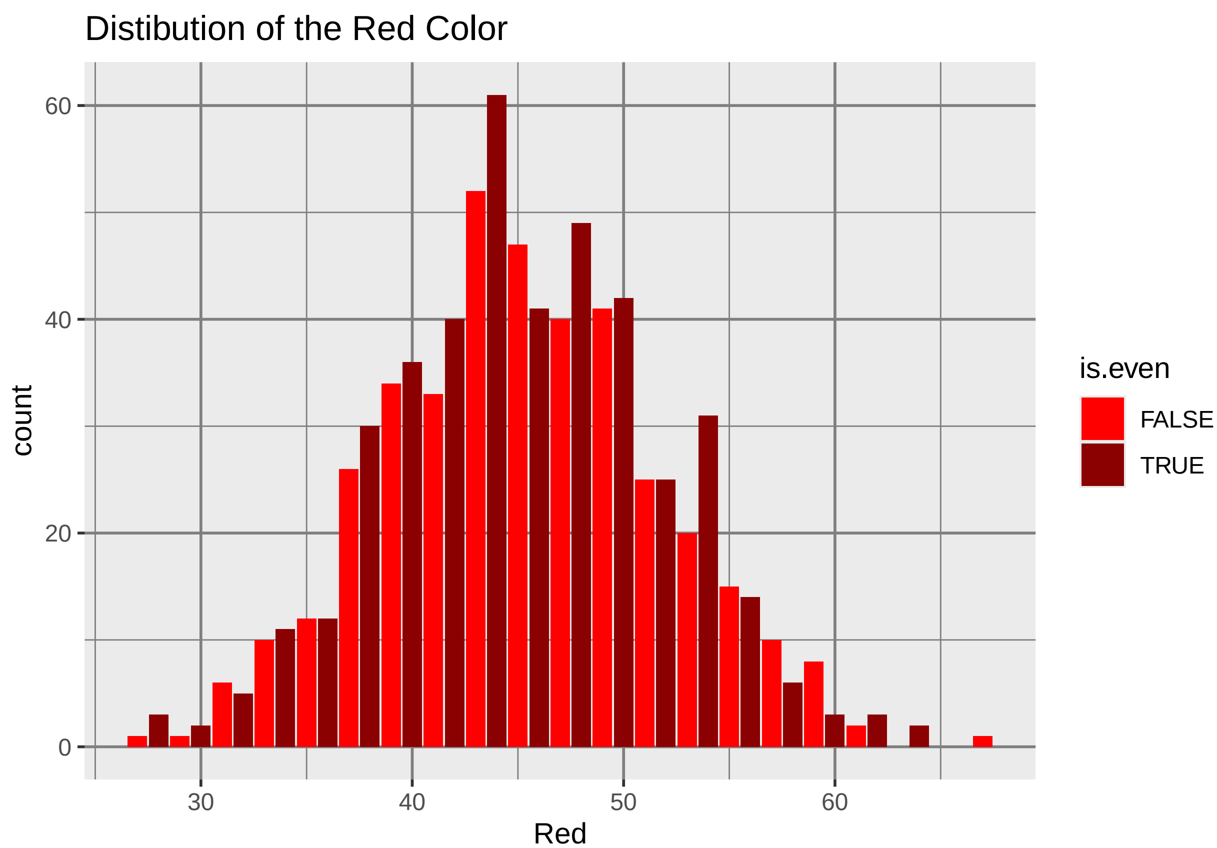
```
## # A tibble: 6 x 12
##     Red  Blue Green Orange Yellow Brown even_count even_evens low_col Variance
##   <int> <int> <int>  <int>  <int> <int>      <dbl> <lgl>        <dbl>    <dbl>
## 1    61    41    45     39     40    48          2 TRUE             0     67.9
## 2    59    44    42     46     45    39          3 FALSE            0     47.8
## 3    40    42    44     38     54    57          5 FALSE            0     61.0
## 4    62    46    43     37     48    38          4 TRUE             0     82.7
## 5    54    42    50     46     35    47          4 TRUE             0     43.5
## 6    55    43    42     42     51    41          2 TRUE             0     34.3
## # i 2 more variables: min <int>, all_even <lgl>
```

plot the M&M sample sample

```
## [1] "summary of all colors Distibution:"
```

```
##         Min. 1st Qu. Median     Mean 3rd Qu. Max.      Var
## Red       27      41     45 45.32000      50   67 42.47069
## Blue      28      43     46 46.22625      50   68 33.82735
## Green     27      41     46 45.70125      50   66 37.96696
## Orange    27      42     45 45.42750      50   65 35.98472
## Yellow    30      41     46 45.88625      50   64 39.19731
## Brown     26      42     46 45.91375      50   64 38.54950
```





Statisics Checking of the Simullation

Test Expected Value

to see is the μ of the lentils per color are fair, we will test it per column with t.test for each color.

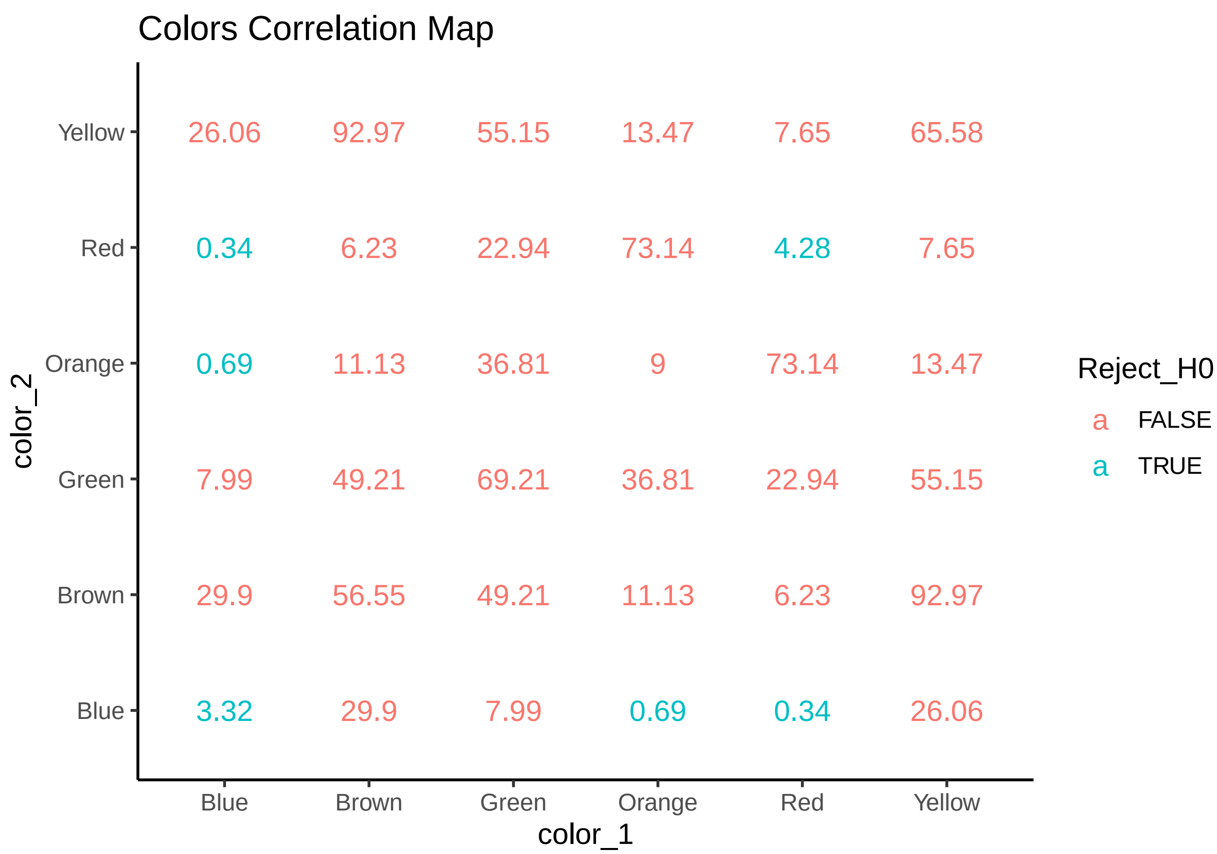
Here is the result, none of them bellow 5% P. value

p.value of $H_0 : \mu = \frac{n_u nit}{n_c olor}$

```
##      Red     Blue    Green   Orange   Yellow    Brown
##  "4.28%"  "3.32%" "69.21%"  "9.00%" "65.58%" "56.55%"
```

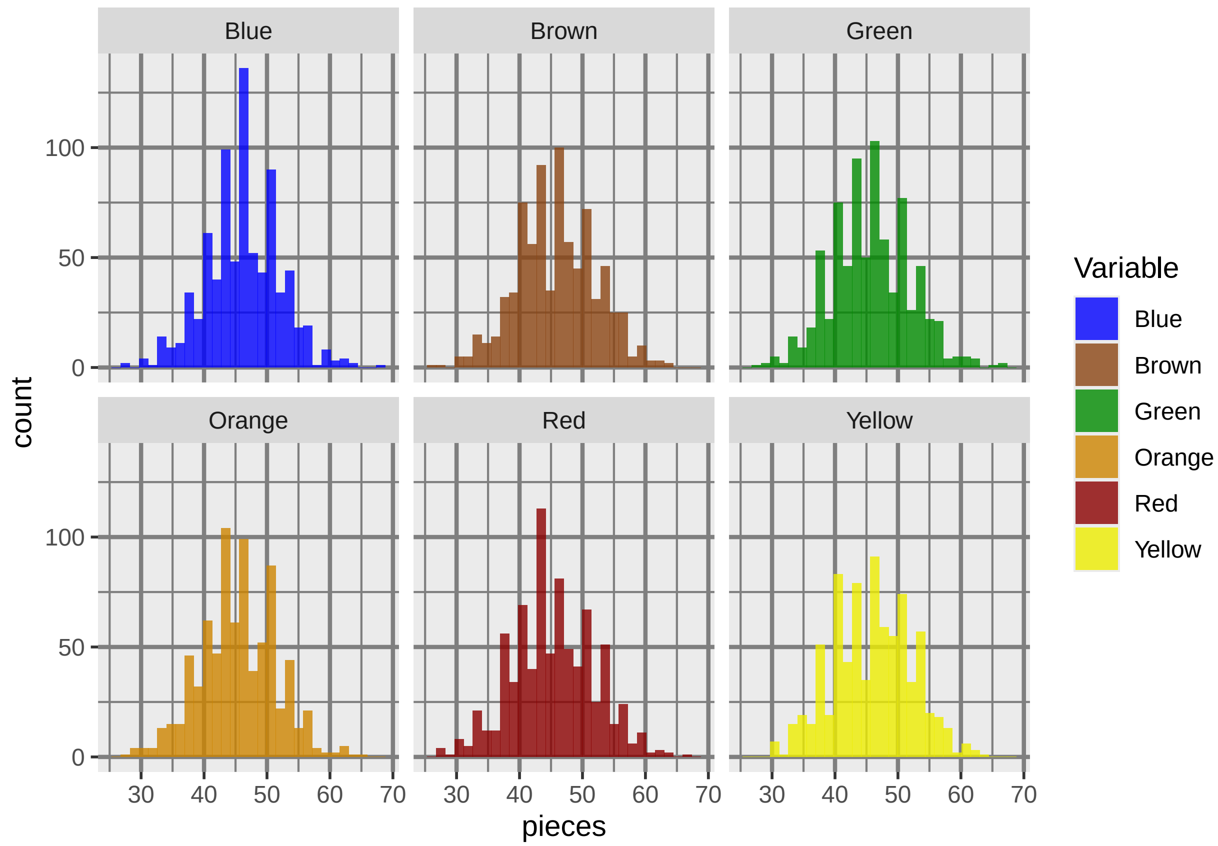
Now we will do the same checking for 2 samples, to see whether there is correlation between each 2 colors distribution.

for each row i and column j, 1) if i==j, this it the check from before of the expected value to n_unit/n_color 2) if i!=j, this is two samples test of same expected value hypothesis



Now here Is visualization of the actual data per color

```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```



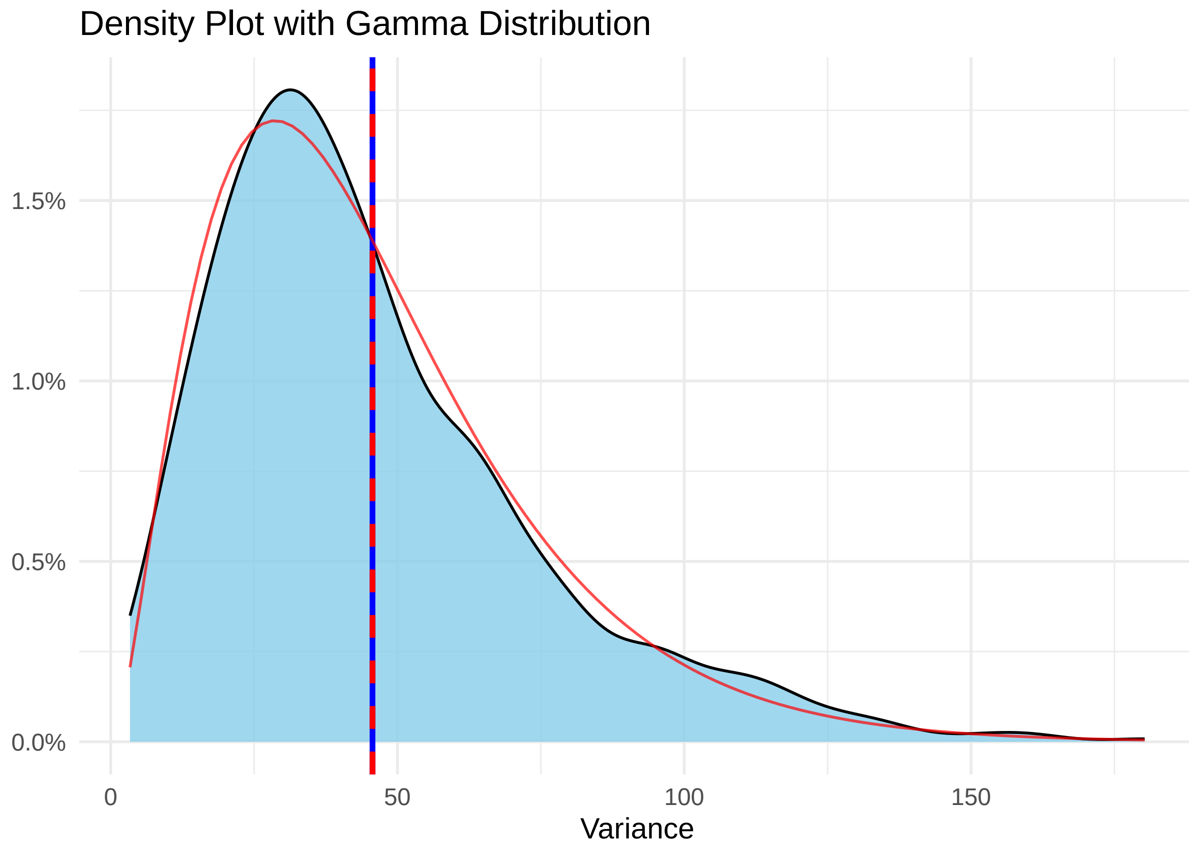
Variance Distribution Checking

We know that the distribution of variance is approximately Gamma distribution:

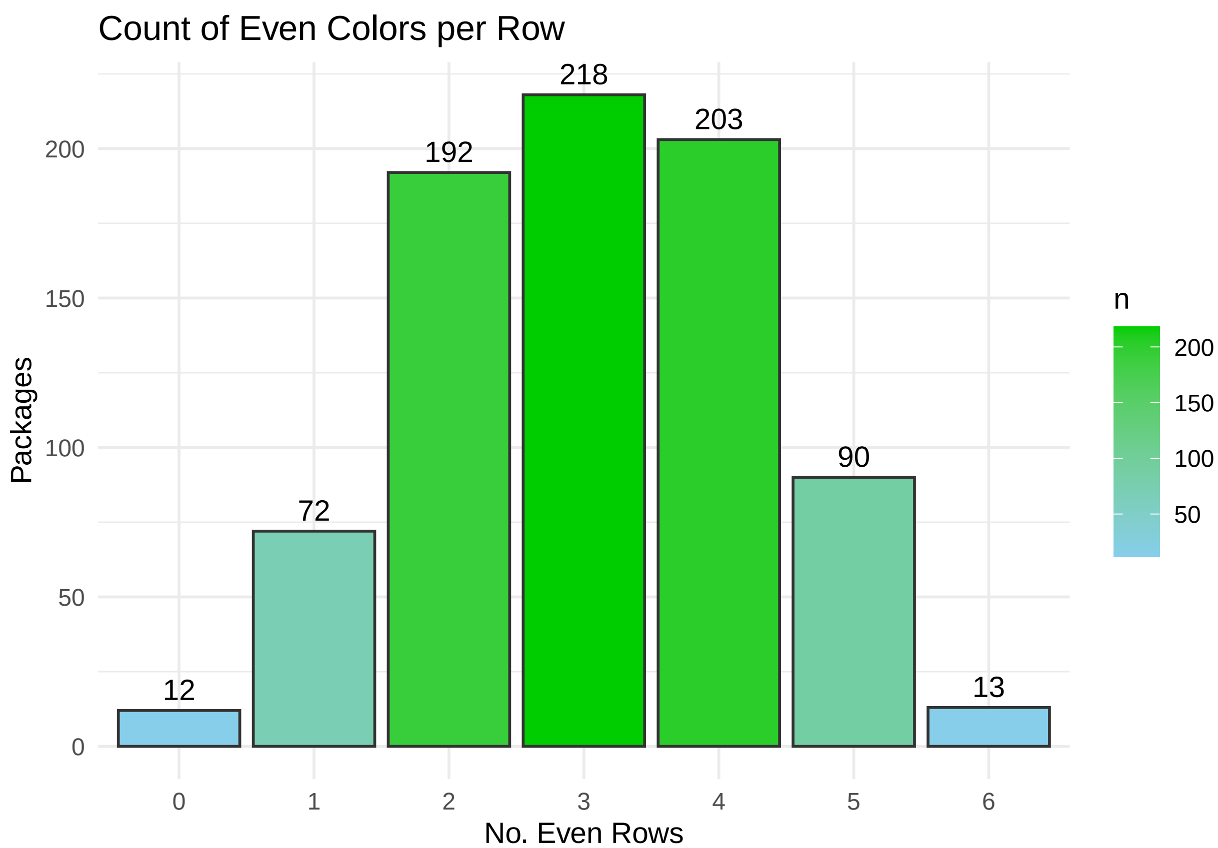
$$f(x) = \frac{1}{\Gamma(\alpha)\theta^{\alpha}} x^{\alpha-1} e^{-x/\theta}$$

We can see that the variance distribution is Gamma like with shape and rate as seen below

```
## [1] "The parameters of the gamma shaped variance is shape 2.685 and rate 0.059"
```



Are All Even in the Sample?



n*m types of snacks

I will create a function that create sample for each number of colors and package size we want, and then calculate some interesting parameters

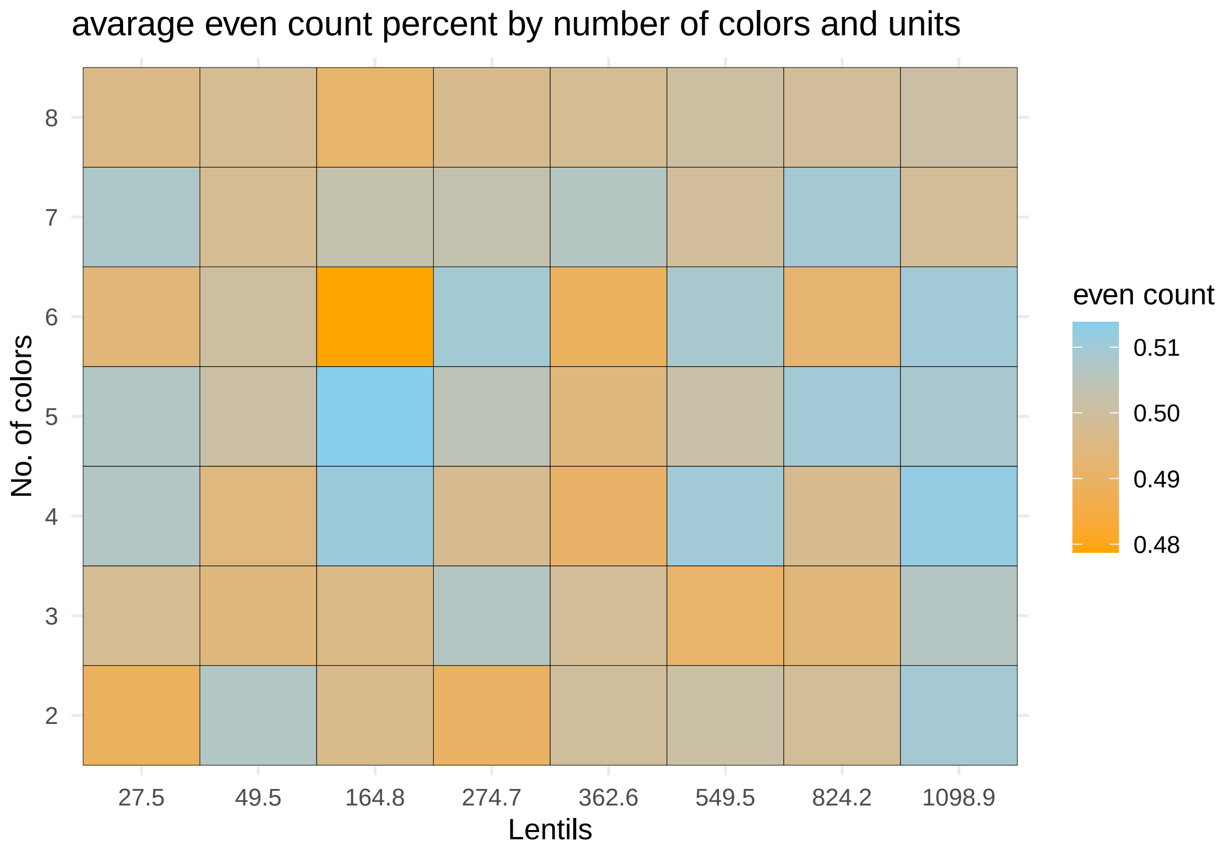
```
color_op<- 2:8
grams_op<- c(25,45,150,250,330,500,750,1000)
n_unit_op<- grams_op/gram
nn=800
```

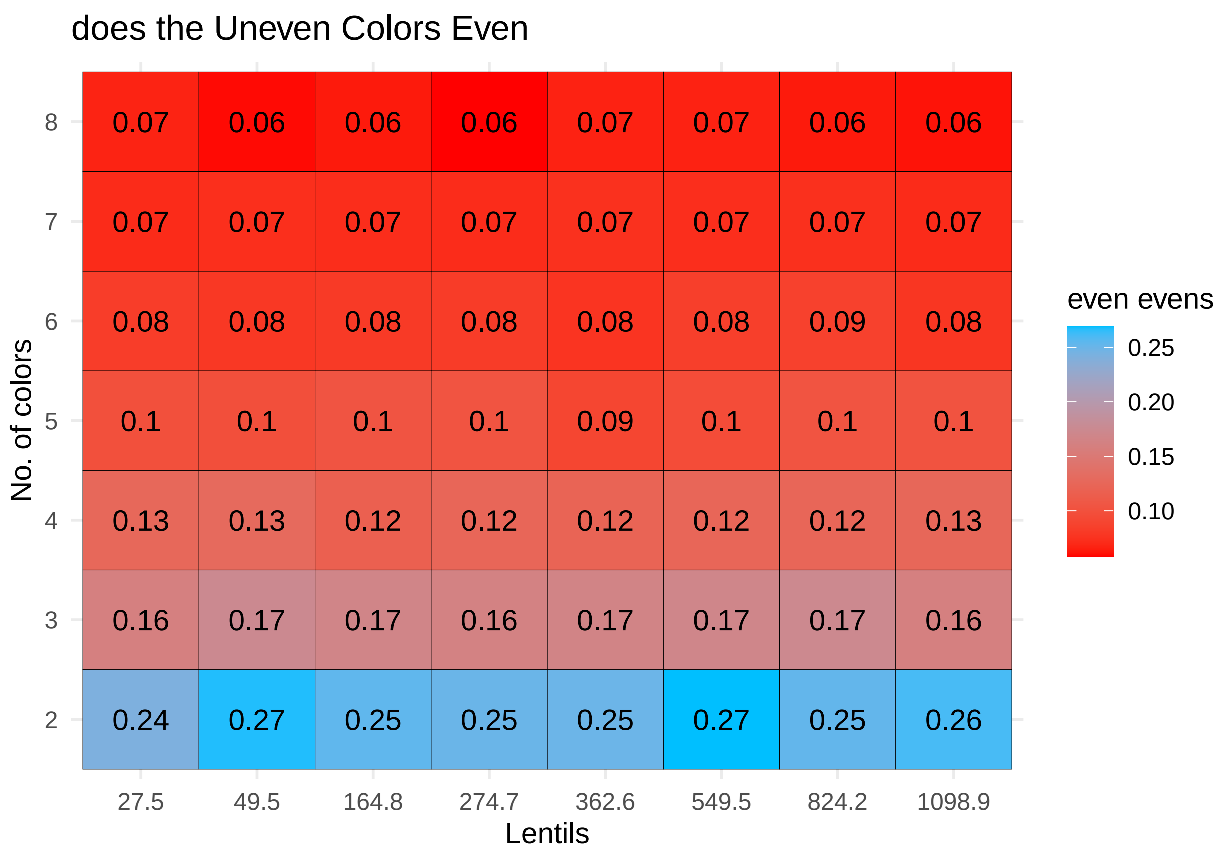
We will make the multiple sample. Here is some random rows:

```
##   n_unit n_color even_count even_evens   var_col all_even low_color
## 1 1098.9       7   0.498750 0.07125000 146.60063  0.00750   0.00000
## 2  362.6       6   0.488750 0.07666667  61.05500  0.01125   0.00375
## 3   49.5       2   0.506875 0.26687500  25.82062  0.27375   0.01750
## 4  824.2       2   0.498750 0.25250000 405.37500  0.25125   0.00000
## 5  549.5       4   0.510000 0.12406250 127.94688  0.06750   0.00000
##   smallest_col
## 1          119
## 2           40
## 3           11
## 4          371
## 5          107
```

Deep Insight on the Data

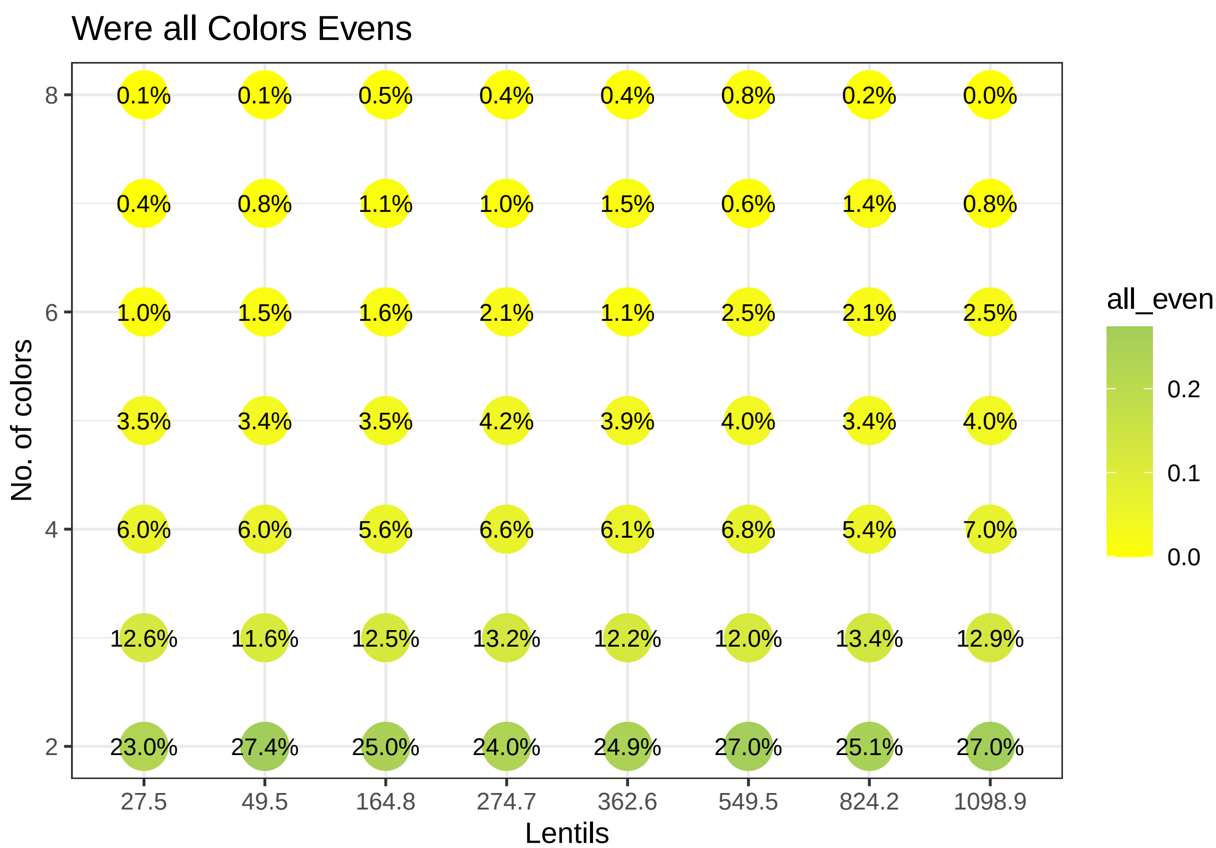
here are some insights:

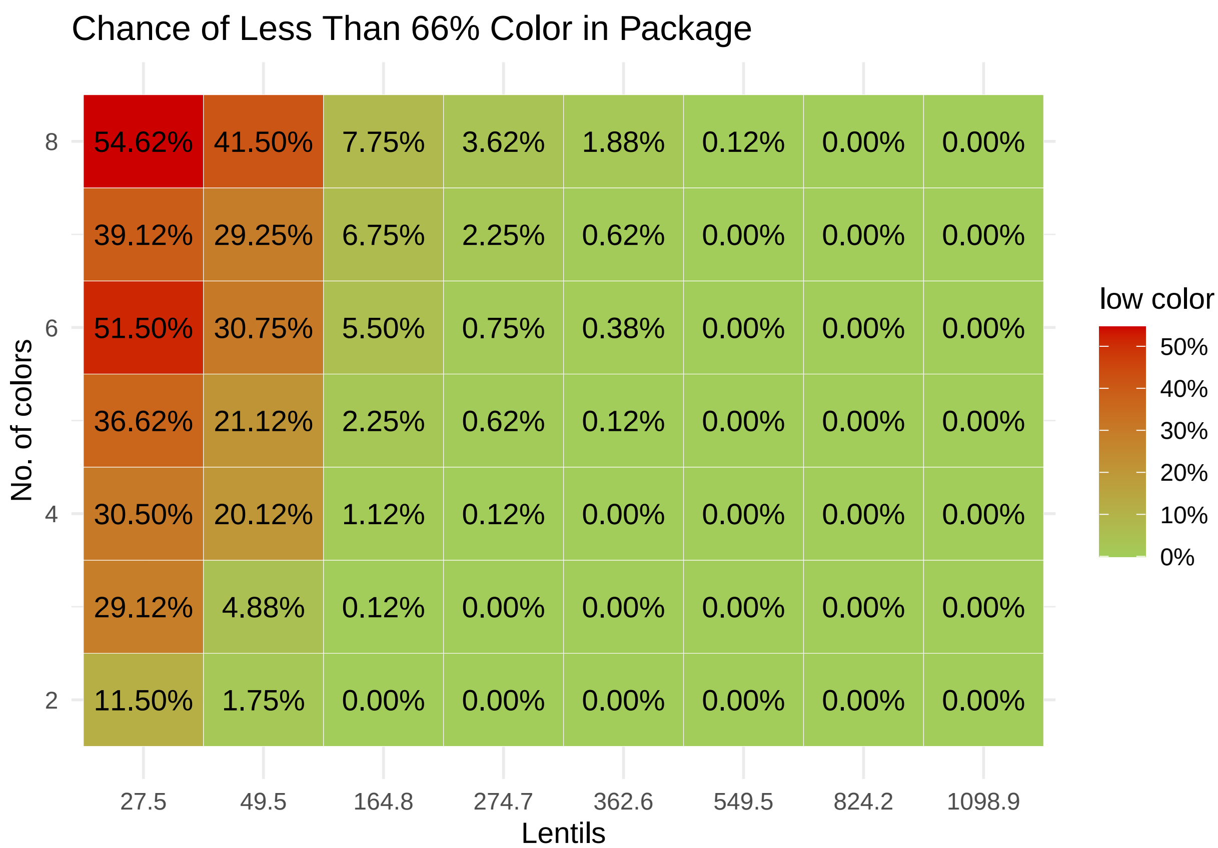


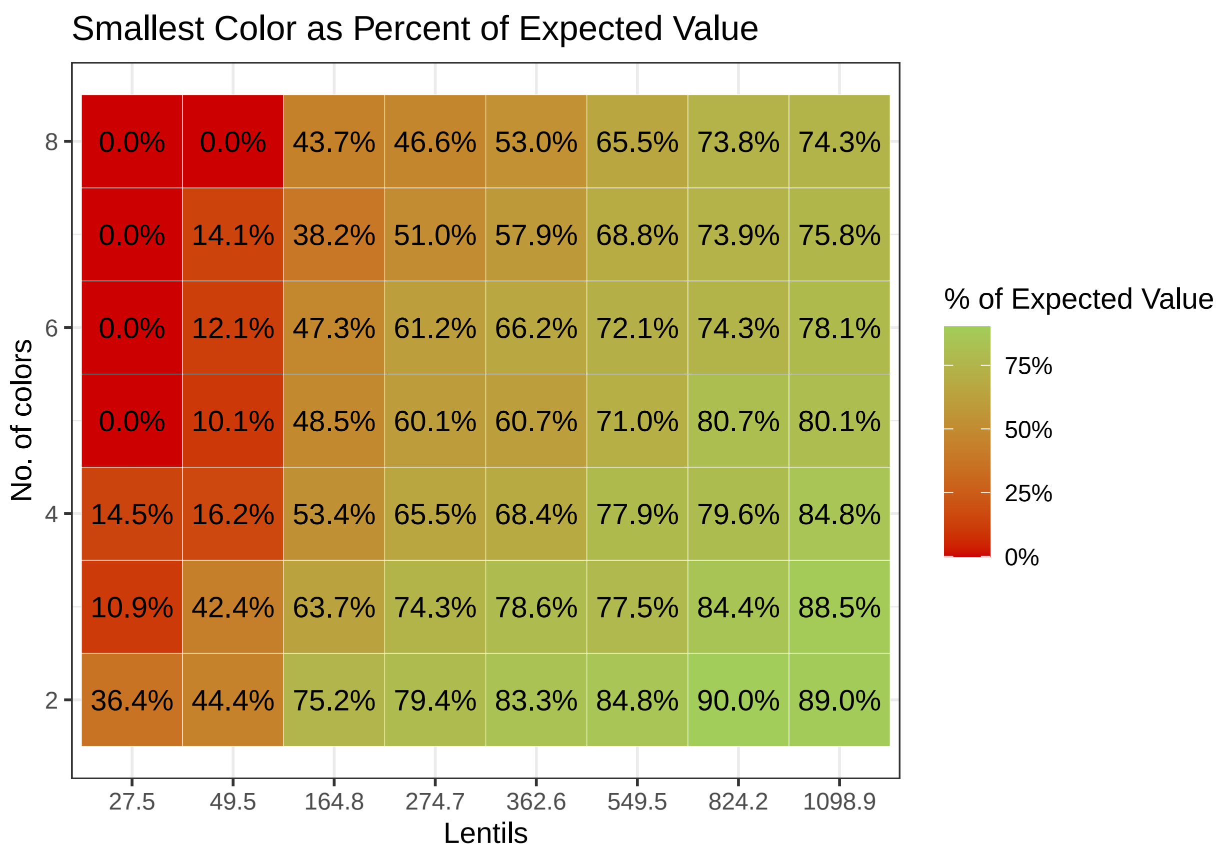


Here is probability of all even, and whether there is pattern.

```
## <Guides[1] ggproto object>
##
## colour : "none"
```


Were all Colors Evens
No. of colors
Lentils
all_even
0.0
0.1
0.2
8
6
4
2
27.5
49.5
164.8
274.7
362.6
549.5
824.2
1098.9
0.1% 0.1% 0.5% 0.4% 0.4% 0.8% 0.2% 0.0%
0.4% 0.8% 1.1% 1.0% 1.5% 0.6% 1.4% 0.8%
1.0% 1.5% 1.6% 2.1% 1.1% 2.5% 2.1% 2.5%
3.5% 3.4% 3.5% 4.2% 3.9% 4.0% 3.4% 4.0%
6.0% 6.0% 5.6% 6.6% 6.1% 6.8% 5.4% 7.0%
12.6% 11.6% 12.5% 13.2% 12.2% 12.0% 13.4% 12.9%
23.0% 27.4% 25.0% 24.0% 24.9% 27.0% 25.1% 27.0%


Chance of Less Than 66% Color in Package
No. of colors
Lentils
low color
50%
40%
30%
20%
10%
0%
8
6
4
2
27.5
49.5
164.8
274.7
362.6
549.5
824.2
1098.9
54.62% 41.50% 7.75% 3.62% 1.88% 0.12% 0.00% 0.00%
39.12% 29.25% 6.75% 2.25% 0.62% 0.00% 0.00% 0.00%
51.50% 30.75% 5.50% 0.75% 0.38% 0.00% 0.00% 0.00%
36.62% 21.12% 2.25% 0.62% 0.12% 0.00% 0.00% 0.00%
30.50% 20.12% 1.12% 0.12% 0.00% 0.00% 0.00% 0.00%
29.12% 4.88% 0.12% 0.00% 0.00% 0.00% 0.00% 0.00%
11.50% 1.75% 0.00% 0.00% 0.00% 0.00% 0.00% 0.00%



As we can see, only the small package (less than 50 lentils) have high probability of at least one color to appear severely lower.

Therefore, splitting package by color on the big ones should be relatively even.

using regression for correlation check

```
##
## Call:
## lm(formula = mega_snack_2$even_count ~ mega_snack_2$n_color +
##     mega_snack_2$n_unit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.53152 -0.17210 -0.00919  0.15337  0.50947
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           5.278e-01  2.139e-02  24.674   <2e-16 ***
## mega_snack_2$n_color -5.594e-03  3.642e-03  -1.536    0.125
## mega_snack_2$n_unit   1.353e-05  2.038e-05   0.664    0.507
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2438 on 1117 degrees of freedom
## Multiple R-squared:  0.002499,   Adjusted R-squared:  0.0007133
```

```
## F-statistic: 1.399 on 2 and 1117 DF,  p-value: 0.2472

##
## Call:
## lm(formula = mega_snack_2$all_even ~ mega_snack_2$n_color + mega_snack_2$n_unit +
##     mega_snack_2$color_No2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.15119 -0.06513 -0.03002  0.01526  0.87161
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 1.428e-01  1.651e-02   8.651  < 2e-16 ***
## mega_snack_2$n_color       -2.679e-02  2.616e-03 -10.238  < 2e-16 ***
## mega_snack_2$n_unit        -2.128e-05  1.464e-05  -1.453    0.146
## mega_snack_2$color_No2TRUE  6.250e-02  1.057e-02   5.911 4.51e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1751 on 1116 degrees of freedom
## Multiple R-squared:  0.1128, Adjusted R-squared:  0.1104
## F-statistic: 47.29 on 3 and 1116 DF,  p-value: < 2.2e-16
```

seeing the 2_{nd} regression we can suggest that more colors is correlated with 3.3% less probability of all colors even, while even numbers of colors is correlated with 7.7% more probability of all colors eve, regardless of any package size.

Conclusions

Data Structure

The simulation created a random samples of snack packs, which was proven to be statistically random. We created with "sample_MnM" one sample with specific size and numbers of colors, and then "mega_snack" that create costume samples and check the relevand indicators fot this project.

We I out that:

- Small packages often lack at least one color, and sometimes contain only one color.
- As the number of colors increases, the chance that all colors have even counts drops significantly.
- For medium to large packages, the probability of any one color being significantly underrepresented (less than $\frac{2}{3}$ of its expected amount) is near zero.

Main Q: Eating M&M by Two

Although there is no clear pattern to the right M&M package for all the colors to have even count, different approach might find a clear reason for more or less couples of M&M. Here is what I did found:

The general probability of all colors to be even in 6 colored pack is 1.5% for small 50g package 2.1% for big 1000g package, and overall 1.5%, which is more than I expected.

For 5 colored pack like Skittles the average is about 2.9%

For 2 colored pack the average is 24.5%, so for 2 colored marshmallow bag this will be the statistics.

Summary

To sum it up, for each medium pack the probability of all even colors is 1.4%, or 1 in a 73 packs of 250g. So I might need to change my snack preference to marshmallow if I want to keep this method.

This project help me implement the method of simulation on a need— as silly as in might be— and check the business question with wider sight in order to understand the hole story.

I applied:

- Simulation logic
- Exploratory analysis
- Hypothesis testing
- Distribution checks
- Outlier detection
- Visualization using R

In addition, I created the infrastructure for similar questions with different parameters to be checked in a reusable, structured way.