Statistical Analysis of Random Distribution in M&M Packages

Yoni

20 04, 2025

Intro

Objective of Simulation

This simulation explores the fairness and randomness in the color distribution of M&M-style chocolate candies. Specifically, it investigates:

1. How often are M&M packs evenly distributed across all colors?
2. What's the probability of a pack missing at least one color?
3. How do the number of candies and available colors impact that probability?

while seeing some chocolate packages, I wondered:

```
**Can I finish a pack of M&M eating two at a time, without ever mixing colors in a bite?**
```

This playful question leads to a deeper statistical exploration of random sampling, distribution fairness, and packaging quality in candy production.

Methodology

Since production data from M&M isn't publicly available, I simulate packages based on common retail sizes and the standard 6-color set. Each simulation randomly draws a sample of candies, assigning each a color. We repeat this process hundreds of times to analyze statistical properties across "virtual" packages.

Through this approach, I estimate the probability of:

1. Getting a perfect pack (i.e., equal counts of each color).
2. Receiving a pack missing at least one color.
3. Seeing how these probabilities change with pack size or number of available colors.

My hypothesis is that perfectly balanced packs are extremely rare, especially when six colors or more are involved. Larger packs may contain all colors more consistently but still tend to be uneven in distribution.

Parameters

Basic parameters:

We define a "pack" as a vector of integers representing the count of each color.
Each simulation uses random sampling with replacement to mimic real-world packaging.
Key variables:

- `n-color`: Number of distinct colors
- `n-unit`: Total candies in the pack

```
#parameters
n= 1000                #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 package
```

```
## [1] "The avarage number of lentils per color is 45.8"
```

Creating the Sample

General Sample

In order to test the theoretical data, I 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.

key parameters for `sample_MnM`

- `n`: Number of packages in the sample
- `x_units`: Total candies in each package
- `n_colors`: Number of distinct colors in each package

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

```
##       1  2  3  4  5  6
## [1,] 15 17 19 14 14 22
```

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

```
##       Red Blue Green Orange Yellow Brown
## Bag_1  15   14    14     23     19    15
## Bag_2  19   11    13     19     20    18
## Bag_3  21   13    16     14     15    22
```

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. `low_col`- sum true if one color's count is lower than $\frac{2}{3}$ of expected value
5. `Variance`- variance of lentils per color
6. `min`- the lowest color in each row
7. `all_even`- are all colors even

here are the first rows:

Table 1: M&M sample random rows

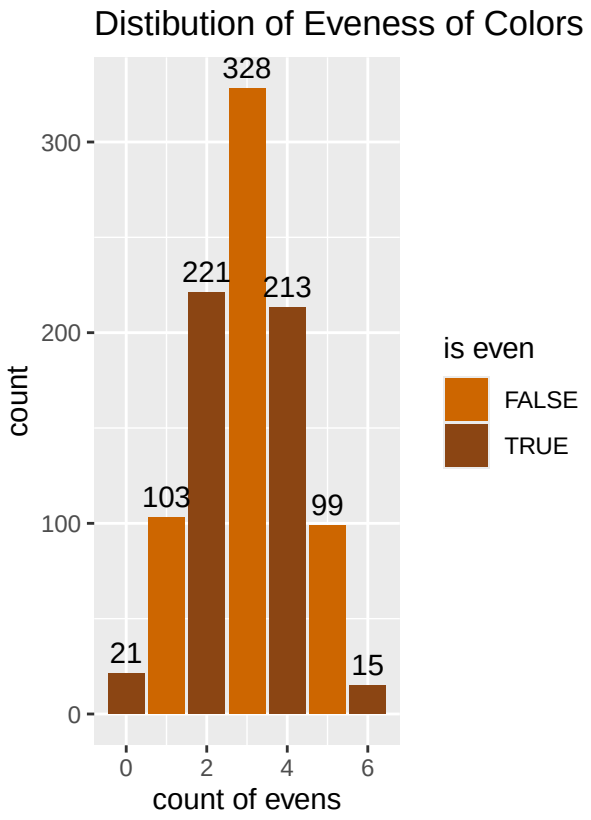
Red	Blue	Green	Orange	Yellow	Brown	even_count	even_evens	low_col	Variance	min	all_even
43	45	54	31	51	50	2	TRUE	0	67.87	31	FALSE
45	46	45	51	55	33	1	FALSE	0	55.37	33	FALSE
45	56	43	43	43	44	2	TRUE	0	26.27	43	FALSE
47	44	40	37	56	51	3	FALSE	0	49.37	37	FALSE

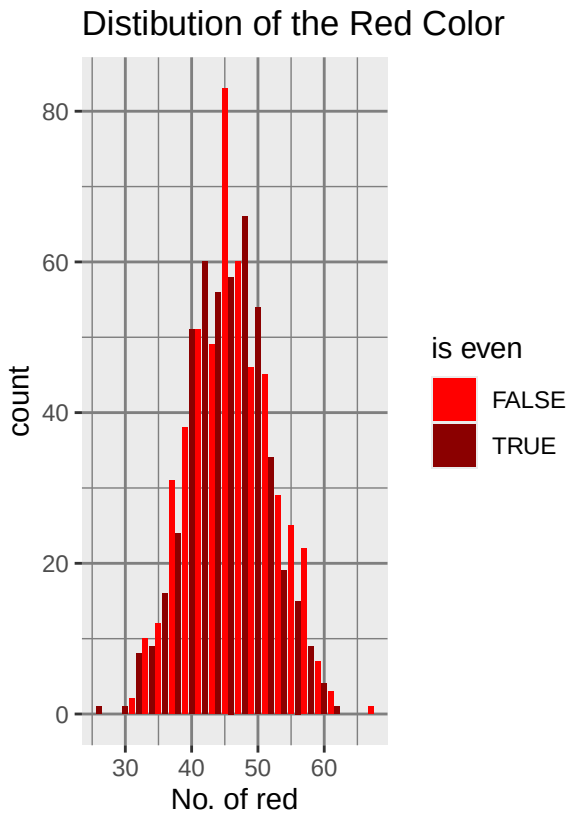
I summarized the sample by color bellow

Table 2: summary of all colors Distibution

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Var
Red	26	41	45	45.696	50.00	67	36.79
Blue	28	41	46	45.649	50.00	64	37.22
Green	19	41	45	45.615	50.00	67	38.79
Orange	29	42	46	45.690	50.00	65	38.03
Yellow	30	42	46	46.133	50.25	65	36.65
Brown	29	41	45	45.747	50.00	69	40.34

Here we can see the distribution of all colors to be even and of one example color (red)





Statisics Checking of the Simullation

Test Expected Value

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

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

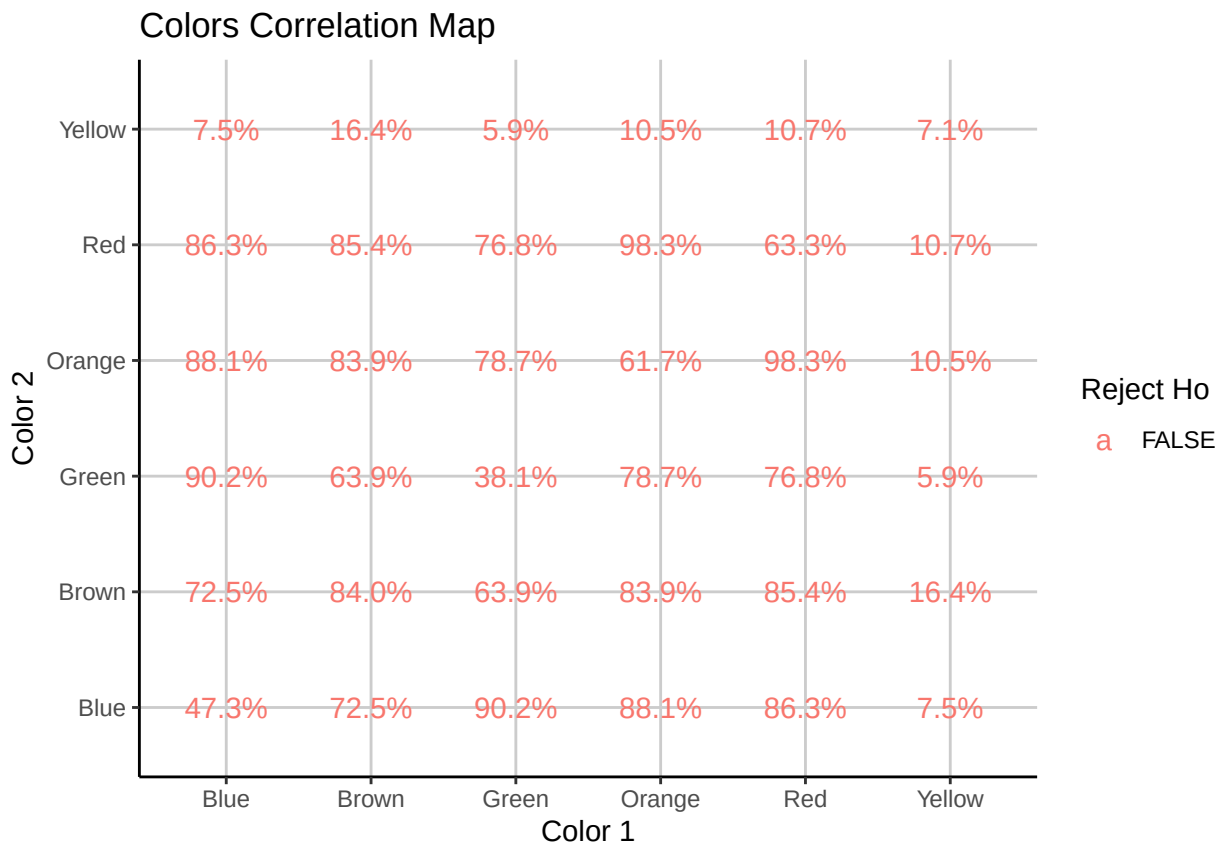
p.value of $H_0 : \mu = \frac{n-unit}{n-color}$

```
##     Red    Blue   Green  Orange  Yellow   Brown
## "63.3%" "47.3%" "38.1%" "61.7%"  "7.1%" "84.0%"
```

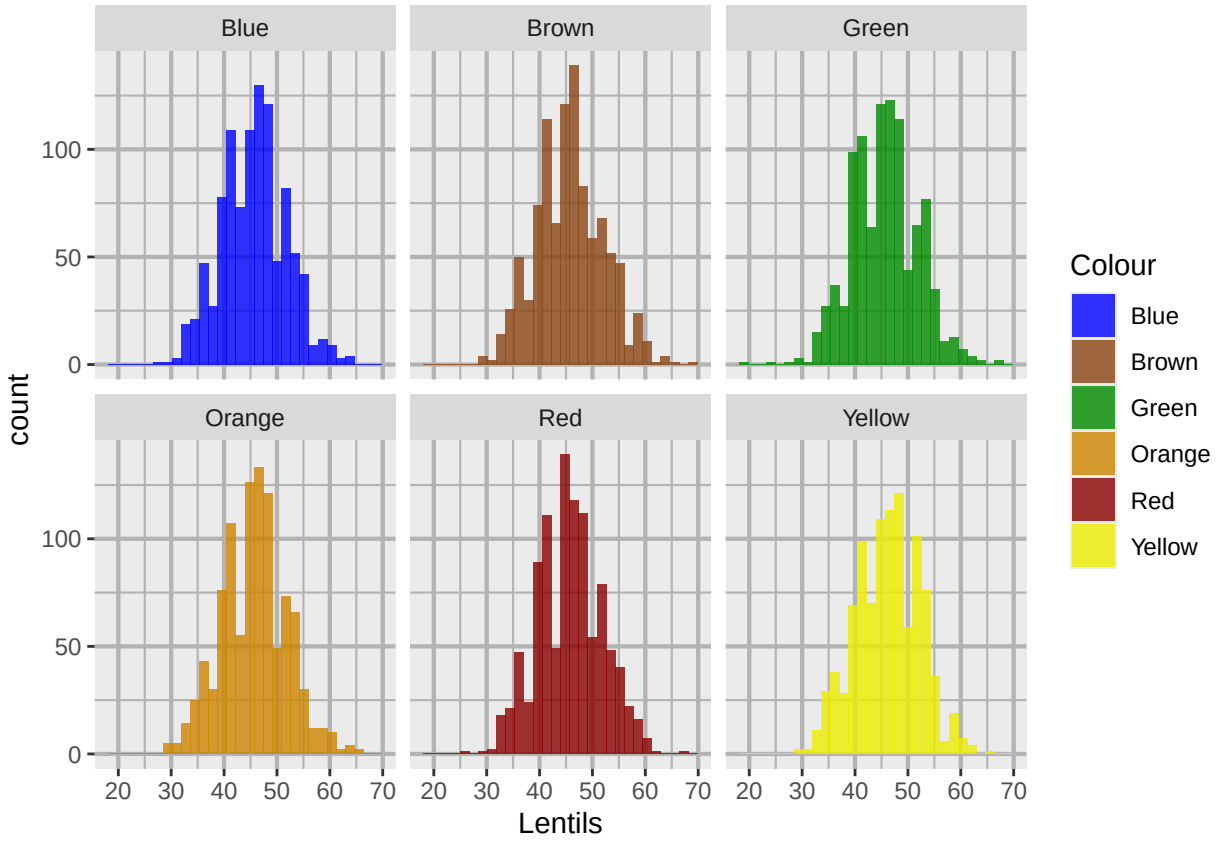
Now I will do the same checking for 2 samples, to see whether there is correlation between each 2 colors distribution.

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



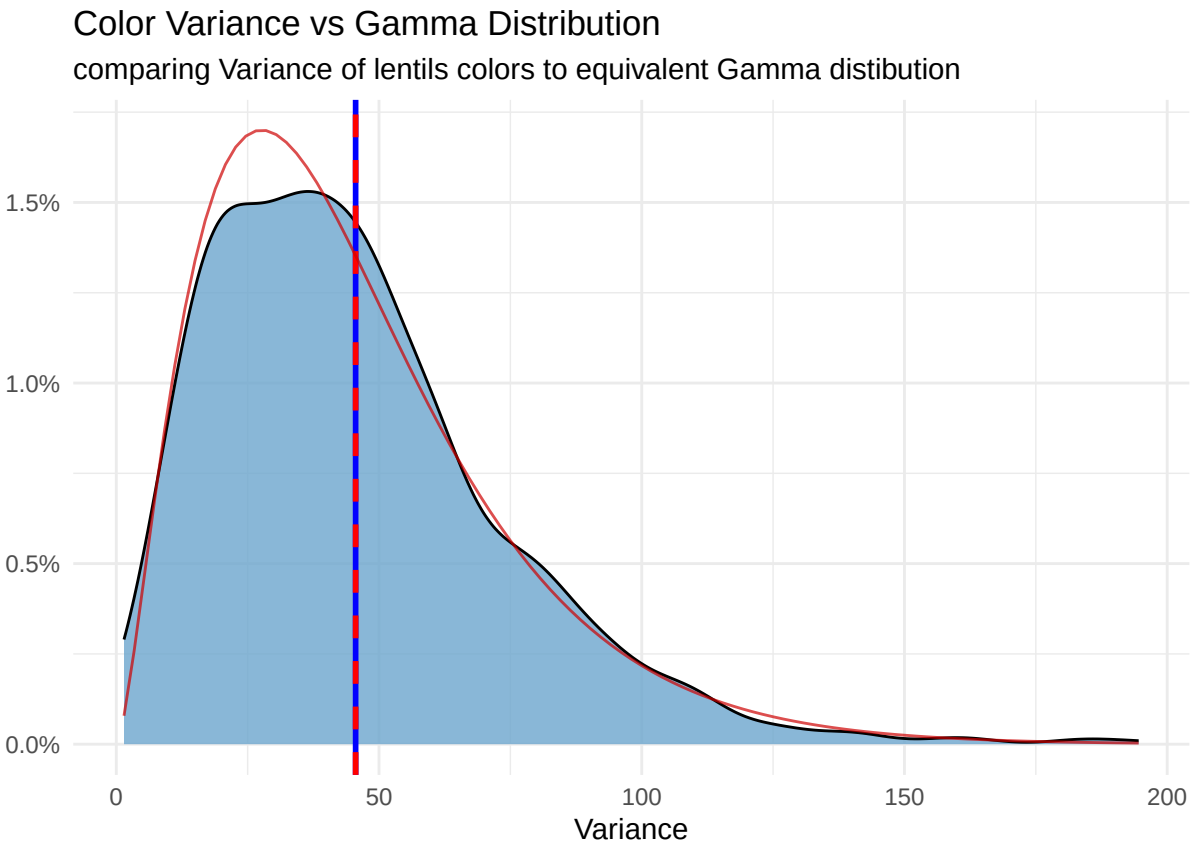
Variance Distribution Checking

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

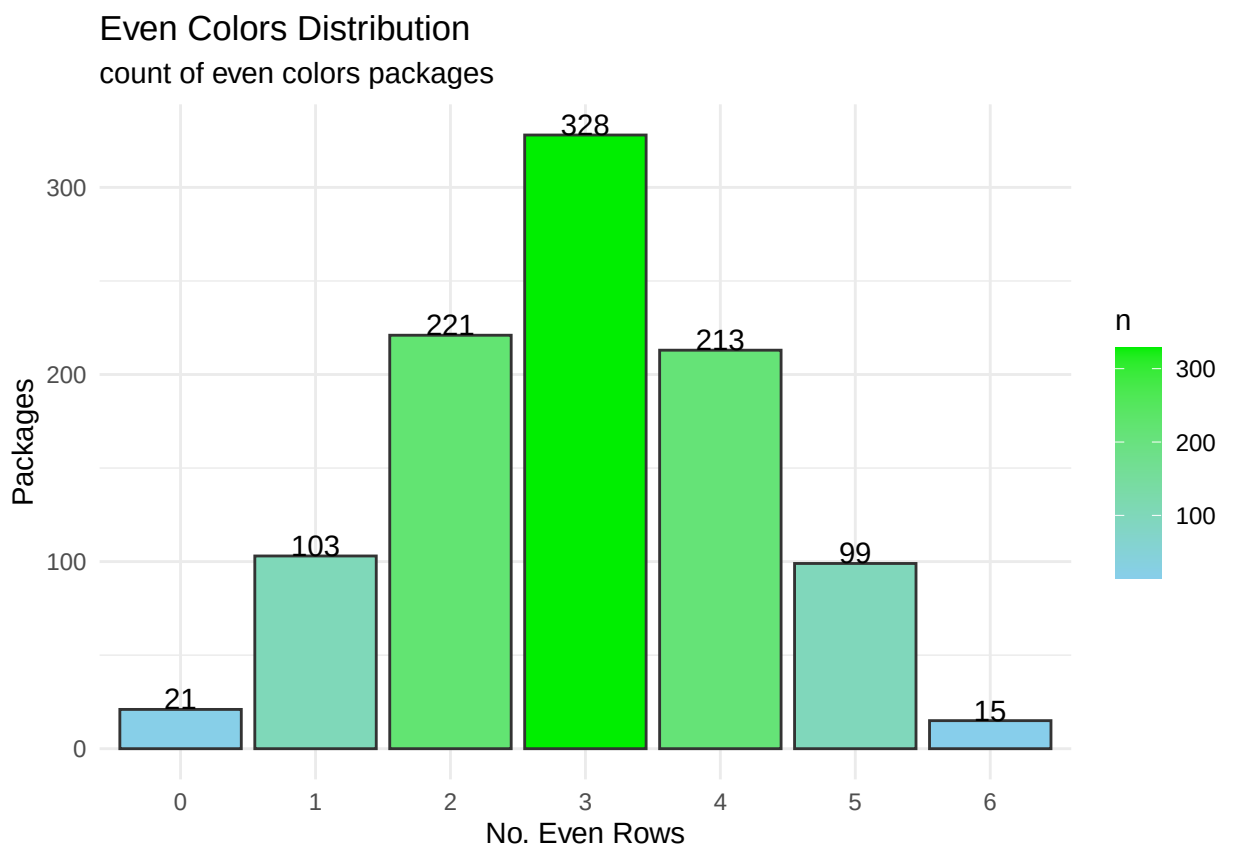
$$f(x) = \frac{1}{\Gamma(\alpha)\theta^{\alpha}} x^{\alpha-1} e^{-x/\theta}$$

I 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 2.545 and scale 17.896"
```


Color Variance vs Gamma Distribution
comparing Variance of lentils colors to equivalent Gamma distibution
1.5%
1.0%
0.5%
0.0%
0
50
100
150
200
Variance

Are All Even in the Sample?



Multiple Types of Samples

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

I will make the multiple sample. parameters:

```
n_color<- 2:8 #Number of distinct colors in each package option
grams_op<- c(25,45,150,250,330,500,750,1000) #Weight of each package option
n_unit_op<- grams_op/gram #Total candies in each package option
nn<- 1200 #Number of packages in the sample
```

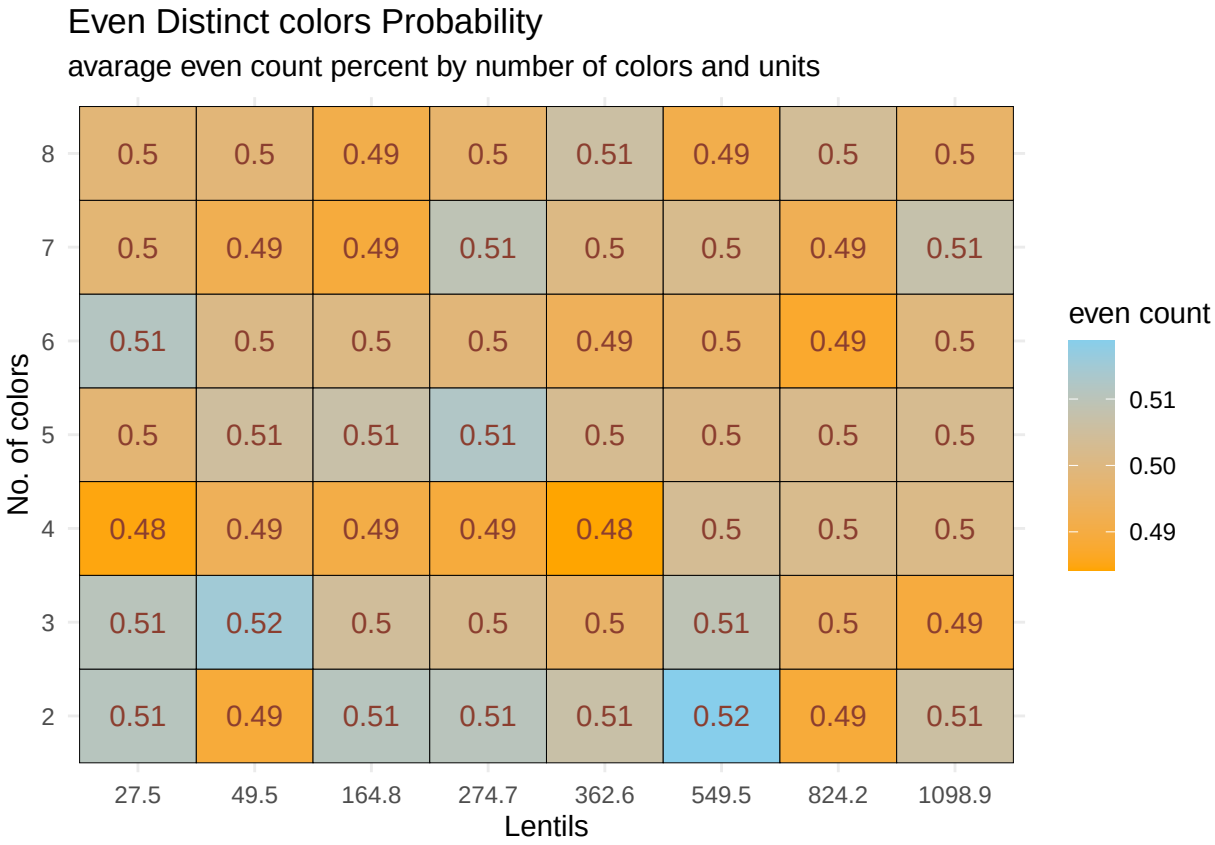
Here is some random rows:

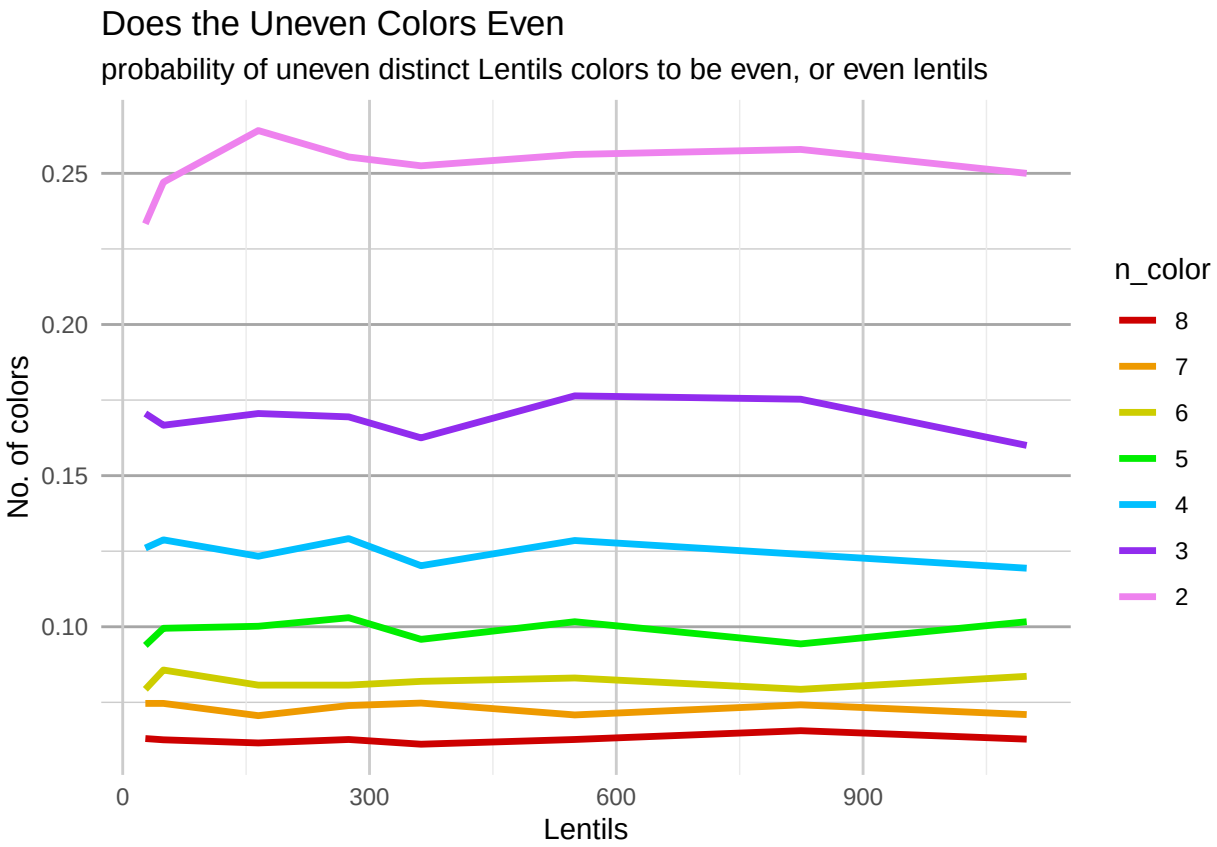
Table 3: Multiple sample example rows

n_unit	n_color	even_count	even_evens	var_col	all_even	low_color	smallest_col
27.5	3	0.510	0.171	9.446	0.144	0.290	2
27.5	2	0.511	0.233	14.308	0.244	0.115	6
164.8	4	0.491	0.123	39.028	0.053	0.013	24
1098.9	3	0.489	0.160	367.420	0.109	0.000	305
49.5	4	0.493	0.129	11.963	0.066	0.192	3

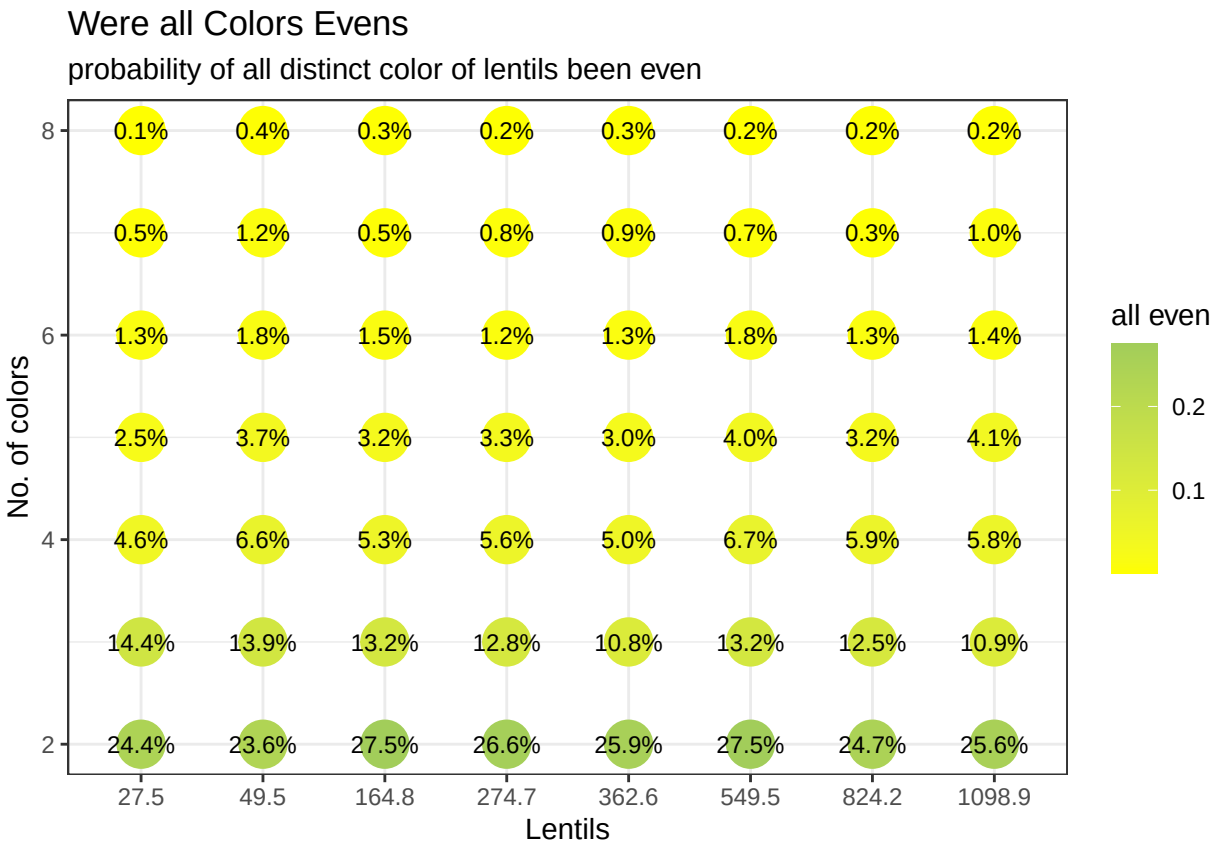
Deep Insight on the Data

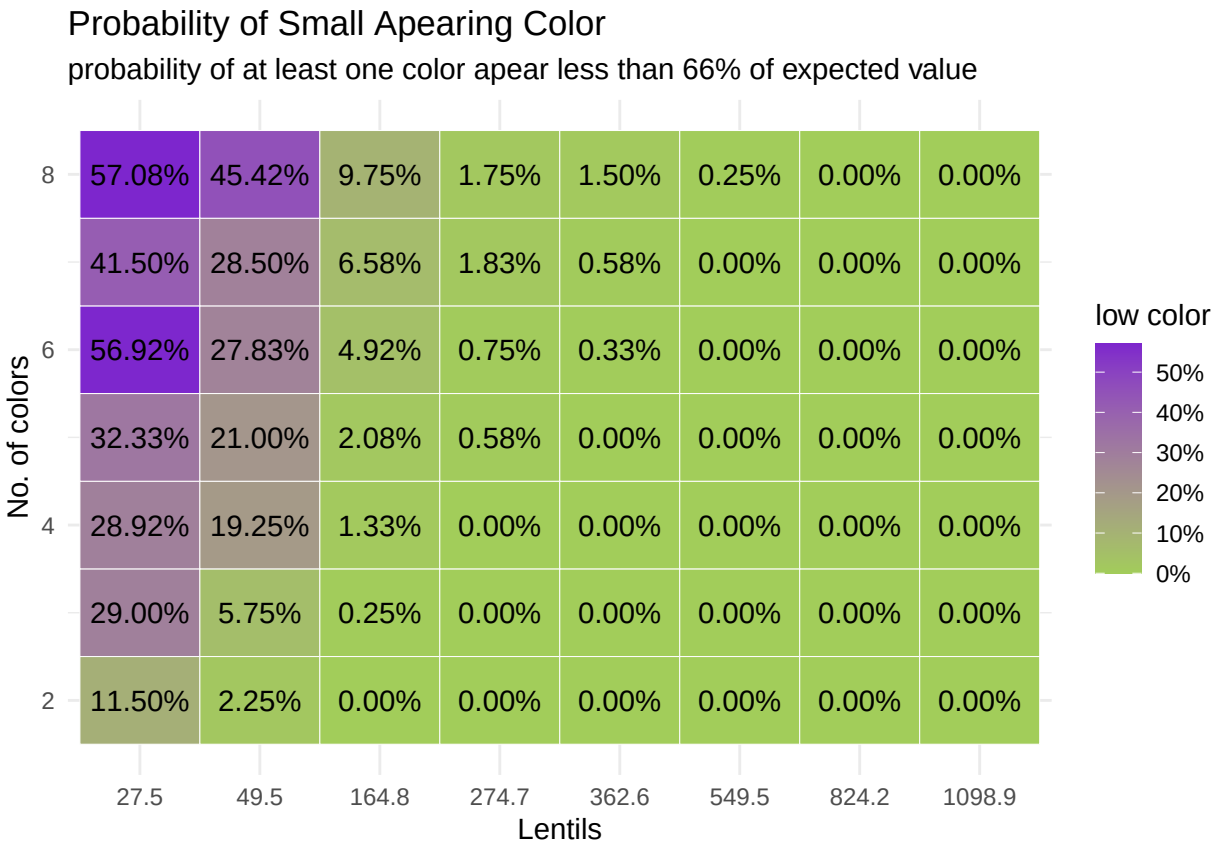
Here are some insights:



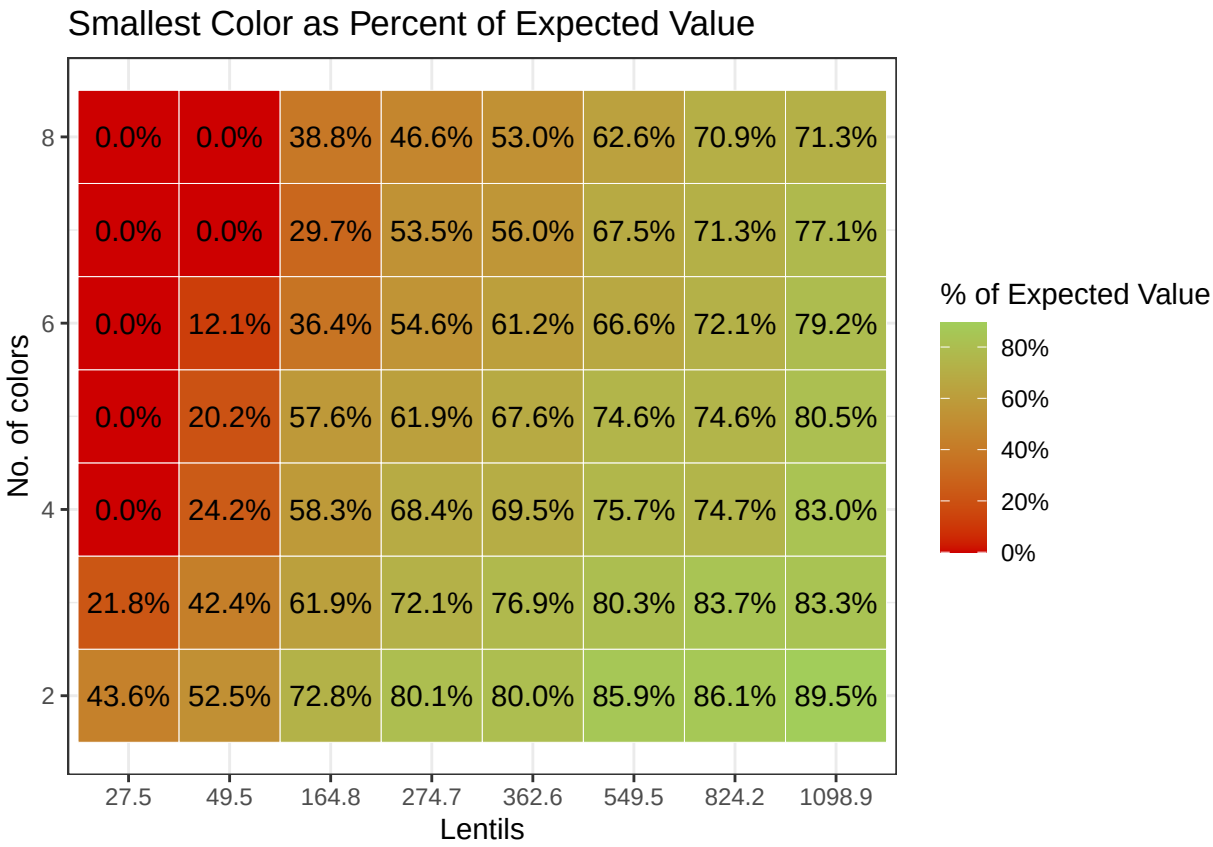


Here is probability of all even, and whether there is pattern.


Were all Colors Evens
probability of all distinct color of lentils been even
No. of colors
8
6
4
2
0.1% 0.4% 0.3% 0.2% 0.3% 0.2% 0.2% 0.2%
0.5% 1.2% 0.5% 0.8% 0.9% 0.7% 0.3% 1.0%
1.3% 1.8% 1.5% 1.2% 1.3% 1.8% 1.3% 1.4%
2.5% 3.7% 3.2% 3.3% 3.0% 4.0% 3.2% 4.1%
4.6% 6.6% 5.3% 5.6% 5.0% 6.7% 5.9% 5.8%
14.4% 13.9% 13.2% 12.8% 10.8% 13.2% 12.5% 10.9%
24.4% 23.6% 27.5% 26.6% 25.9% 27.5% 24.7% 25.6%
27.5 49.5 164.8 274.7 362.6 549.5 824.2 1098.9
Lentils
all even
0.2
0.1



Here we can see the smallest % of Lentils in one color as seen in my sample:



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

I wanted to see if there is statistic correlation of the number of distinct colors of package size to the probability of all colors have equal counts of each color. as a result, I chose to check this claim with regression. Furthermore, I did the same regression adding another potential correlated parameter: the evenness of the number of distinct colors.

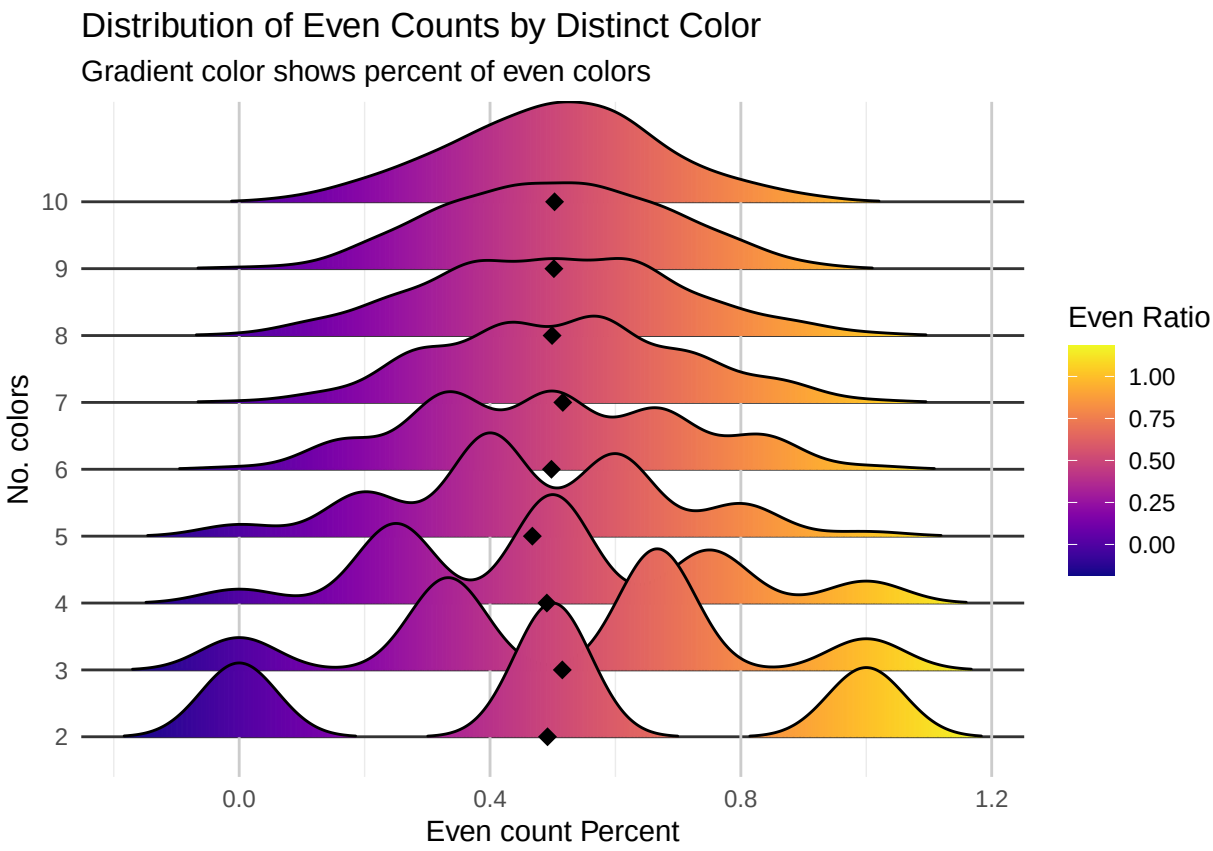
```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                               all_even
##                           (1)            (2)
## ------------------------------------------------
## n_color                -18.924        -9.516
##                       (869.022)      (458.865)
##
## n_unit                  -0.001         -0.001
##                        (0.0005)       (0.0005)
##
## color_No2                              11.184
```

```
##                                  (2,161.438)
##
## Constant              36.994         6.996
##                     (1,738.043)    (2,521.165)
##
## ------------------------------------------------
## Observations          2,160          2,160
## Log Likelihood       -134.303       -134.303
## Akaike Inf. Crit.    274.607        276.607
## ================================================
## Note:              *p<0.1; **p<0.05; ***p<0.01
```

As you can see, there is no statistic clear correlation of the number of colors to probability of all distinct colors to be even.

```
##
## ========================================================
##                            Dependent variable:
##                    ------------------------------------
##                    cbind(n_color, n_color - even_count)
##                           (1)                (2)
## --------------------------------------------------------
## n_color                  0.003              0.003
##                         (0.007)            (0.007)
##
## n_unit                  0.0001*            0.0001*
##                        (0.00005)          (0.00005)
##
## color_No2                -0.006
##                         (0.031)
##
## Constant                0.635***           0.633***
##                         (0.055)            (0.053)
##
## --------------------------------------------------------
## Observations             2,160              2,160
## Log Likelihood         -3,013.535         -3,013.553
## Akaike Inf. Crit.      6,035.069          6,033.105
## ========================================================
## Note:                         *p<0.1; **p<0.05; ***p<0.01
```

Like the previous regression, the second regression that examines the correlation of the number of colors and package size to the percentage of even numbers did not find a statistically significant correlation. In addition, visually there is no clear correlation.



Summary

Data Structure

The simulation created a random samples of snack packs, which was proven to be statistically random with known μ and σ^2. I created one sample with specific size and numbers of colors using "sample_MnM", and costume multiple samples using "mega_snack". Then, I check the relevand indicators fot this project.

I found 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.

suggestion for any random sampler factory (like candies, Lego, toys):

1. Smaller packages need more diversity check
2. Althernatively, I would recomand calculate the amount of each type in small packages

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 25%, so for 2 colored marshmallow bag this will be the statistics.

See all here:

Table 4: Probability of All Colors Even by Pack Colors Number

Colors	All Even Percent
2	25.72
3	12.74
4	5.68
5	3.39
6	1.45
7	0.74
8	0.26

Conclusions

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 allowed me to implement simulation methods in response to a real (albeit silly) question, and evaluate it statistically from end to end..

I applied:

- Simulation by demand
- Exploratory analysis
- Hypothesis testing
- Distribution checks
- Outliers detection
- Visualization using R

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