# Multivariate Probability and Mathematics Statistics Project G5098

203090 202340 200939 200886

**M&M Colour Distribution Analysis** 

# **Contents**

Introduction	3
Section 1 - Colour Distribution of M&Ms	4
Section 2 - Student T-Test	8
Section 3 - Chi-Square Test	10
Section 4 - Kolmogorov-Smirnov Test	12
Bibliography/References	17
Appendix	18

### Introduction

The Mars company is currently worth £70 Billion, with M&M's being one of their most popular confectionery items sold. With their extensive research and product inspection, one would expect that a simple bag of M&M's would be "perfect", i.e having identical cardinality for each colour. Through this project, we will be analysing the 2510 M&M's against our null and alternative hypotheses stated below;

**H**<sub>0</sub>: Each packet will contain an equal distribution of each coloured M&M **H**<sub>1</sub>: Each coloured M&M is unequally distributed with random distribution within each packet.

Within Section 1, we processed atomic data that we collected ourselves and have then analysed it to find the mean and standard deviation for each of the individual colours. These values then allowed us to deduce their distribution. Here, you will also find visual representations of their distributions.

Utilising our observations from section one, the potential of normally distributed data offers the chance to apply another useful test. Here, in Section 2, we conduct a Student T-Test to understand whether our sample has a specific mean to our null hypothesis.

In Section 3, we applied the Chi-Square test to our data set in order to help us to decide whether we can reject or accept our null hypothesis by comparing our observed values for each colour with what we initially expected.

We concluded our research in section 4. We use the Kolmogorov-Smirnov test to identify if each colour followed a similar distribution and thus, allowing us to determine the comparability of data. The method includes finding the maximum distance between blue's CDF and any true value from all other colours. This test can be used at any significance level and allowed for more complicated distributions to be used as opposed to the other methods which have limited use.

### **Section 1 - Colour Distribution of M&Ms**

Analysis - Is each coloured M&M equally distributed in a regular M&M's packet?

 $H_0$ : Each packet will contain an equal distribution of each coloured M&M.  $H_1$ : Each coloured M&M is unequally distributed with random distribution within each packet.

The analysis in question determines whether the distribution of each coloured M&M follows a common distribution within each packet. This is utilised to confirm or refute several hypotheses and query the comparability of the data. Ultimately this section can decide if the distribution of our data collected on each coloured M&M gives us an indication as to which hypothesis to accept.

In order to construct our analysis we collected our own raw data. Taking a sample size of 24 bags, 2510 M&M's, we counted totals, calculated means and standard deviation (See Table 1.1).

Packet	Blue	Brown	Green	Orange	Red	Yellow	Total
1	15	20	17	19	10	20	101
2	7	16	26	22	17	21	109
3	15	18	21	15	21	19	109
4	12	16	19	16	19	20	102
5	17	16	19	21	16	15	104
6	15	13	15	22	20	17	102
7	16	16	16	20	18	18	104
8	13	14	15	22	13	32	109
9	17	20	18	20	12	22	109
10	18	14	19	20	15	20	106
11	16	13	20	17	20	20	106
12	17	16	23	17	18	17	108
13	17	12	16	19	21	18	103
14	16	25	14	14	19	18	106
15	13	20	15	22	17	19	106
16	17	21	18	11	15	19	101
17	19	16	22	16	14	16	103
18	18	17	18	22	18	12	105
19	18	17	26	10	12	19	102
20	20	10	21	20	14	19	104
21	20	17	21	17	16	12	103
22	20	16	17	13	15	20	101
23	16	16	23	13	11	23	102
24	18	17	11	18	18	23	105
Total	390	396	450	426	389	459	2510
Percentage Total	15.538%	15.777%	17.928%	16.972%	15.498%	18.287%	100.000%
Mean	16.250	16.500	18.750	17.750	16.208	19.125	104.583
Standard Deviation	2.908	3.190	3.710	3.639	3.162	3.916	2.733

Table 1.1 - M&M Colour Distribution Data Table

Observing our data sheet in *Table 1.1*, we noticed a combination of similarities and differences among the six coloured M&M's. Notably, we see that the average of each coloured M&M in each packet is roughly between 17-18, with a lowest recorded mean of 16.208 for Red and a highest of 19.125 for Yellow. In addition, we had a mean of 104.583 M&M's per packet.

To assess the varied number of each coloured M&M in each packet, we found the standard deviation of each colour. With the data from *Table 1.1* we constructed a bar graph, with standard deviation error margins (See Fig. 1.2). This enabled us to comparably see how much each coloured M&M deviates from each mean and clearly analyse the spread of each colour. A smaller standard deviation indicates a closer distribution of data and a larger value would suggest greater spread from the mean.

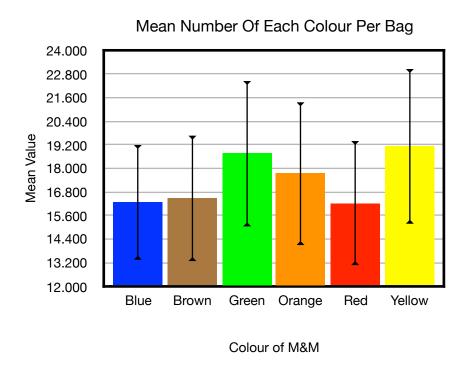


Figure 1.2 - Mean of Each Colour per Packet with Standard Deviation Error Margins

As you can see from *Fig. 1.2*, the standard deviation errors bars barely differ in length. From our data table *(Table 1.1)* there's only a minor range of about 1.008 from 2.908-3.916 within our standard deviations. However, we believe the standard deviation values may be even smaller if it was not for a couple of possible outliers, highlighted by the pink cells in *Table 1.1*. With one packet containing just 7 blue M&M's and another containing 32 yellow M&M's we felt best to construct a histogram for each colour. This way we can correlate any possible trends in the distribution of the coloured M&M's using the frequency density.

After plotting the histograms we spotted a partial trend between each coloured M&M's graph, indicating the possibility the colours of M&M's are normally distributed. Taking our means and standard deviations for each colour we utilised the "NORMDIST" function of Mac Numbers to determine the probability density function for each respective colour of M&M. These probability density values were then used to approximate a normal distribution curve for each histogram (See Fig. 1.3). However, we cannot be certain that the colours are normally distributed, to be more confident we would require a much greater sample size.

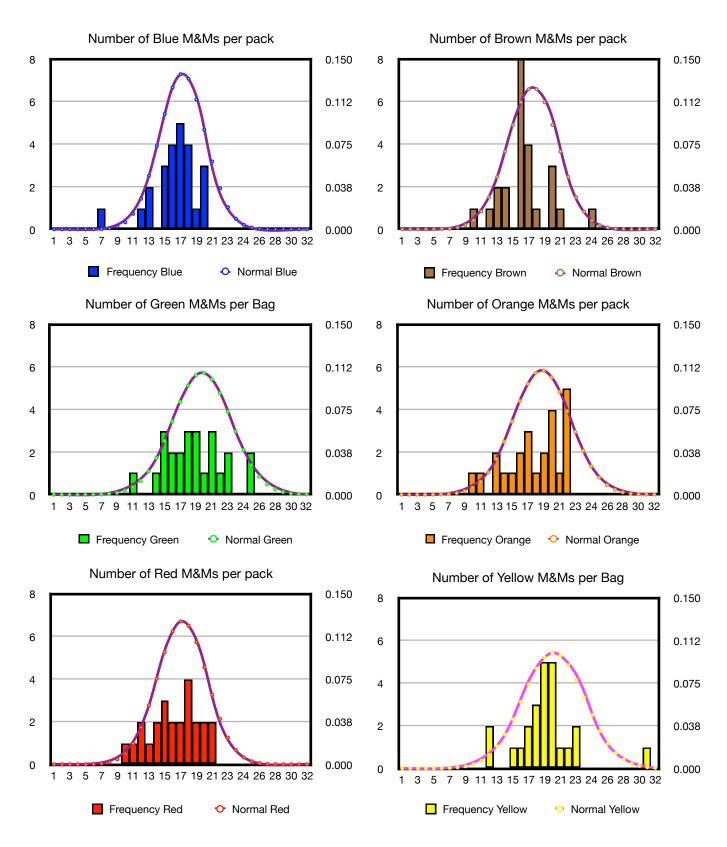


Figure 1.3 - Histogram and Normal Curves for each Coloured M&M

Fig. 1.3 does suggest that each coloured M&M is partially normally distributed but we cannot be sure of our assumption without considering a possible theorem first, as follows:

**Theorem 1.1** - Central Limit Theorem (CLT). Suppose  $X_1$ ,  $X_2$ , ... are independent, with a common distribution having mean  $\mu$  and variance  $\sigma^2$ . Define  $Z_n$  by (1). Then the distribution  $\mathbf{Z}_n$  tends to standard normal as  $\mathbf{n} \to \infty$ , in the sense that for each fixed  $\mathbf{z}$ ,

$$P(Z_n \leq z) \longrightarrow \Phi(z) = P(Z \leq z) \quad (n \longrightarrow \infty),$$

with,

$$Z_n = (S_n - n\mu)/(\sigma \sqrt{n}) \quad \text{where} \quad S_n = X_1 + ... + X_2. \tag{1}$$

With the above theorem we have the ability to record how much the means of our coloured M&M's will vary, without the need to take any other sample means to compare it with. Thus; taking our sample size  $\mathbf{n}$  (the number of M&M's counted), finite number of means ( $\mathbf{\mu}$ ) and the variance ( $\mathbf{\sigma}^2$ ) we can approximately normally distribute the sample distribution as the sample size increases. Here using our means and standard deviations from *Fig. 1.1*.

From the normal curve we are able to observe that our findings will fall below the curve with the Empirical Rule (or 68-95-99.7 Rule) percentages for each standard deviation respectively. Hence, we see from *Fig. 1.3*; blue, brown and red each have a greater frequency density aligned with the peak of their respective normal curves, suggesting our observations are close or equal to the mean. We subjectively note that their peaks are also higher than the remaining three peaks, whereas for green, orange and yellow, there are certain similarities between the three. Clearly the histograms of green, orange and yellow suggest a less dense spread of data and are not as similar to their normal distribution curves as evidenced by the wider, shorter curves of the three.

To provide us with a more accurate representation of each colour's normal distribution, we constructed a graph with all the normal curves together (See Fig. 1.4).

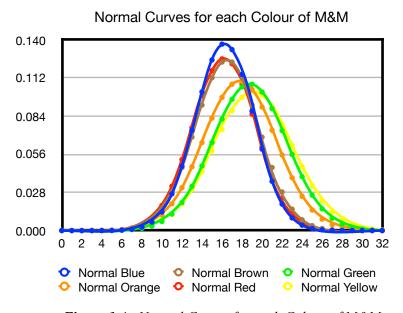


Figure 1.4 - Normal Curves for each Colour of M&M

Firstly, it is evident that blue is a great example of a normal curve. Achieving the highest peak, and narrowest curve, which is consistent with blue having the lowest standard deviation of the six colours of M&Ms. Followed closely by brown and red, both with respectively small standard deviations, their curves peak rather high while remaining narrow. On the other hand; green, orange, yellow each have respectively lower curves with greater width and yellow with the lowest peak of them all. This is consistent with their standard deviations being greater than the previous three.

Relating back to our hypotheses, we conclude the distribution of each colour of M&M can be normally distributed with a certain leniency but it is hard to determine if the distribution is equal. Indeed, we see potential to accept our null hypothesis but cannot disregard our alternate hypothesis without consistency. However, to be more certain of our findings we conducted further analysis but feel a greater sample would have to be collected to determine the correct hypothesis.

### **Section 2 - Student T-Test**

Analysis - Can we determine if our sample has a mean specific to the null hypothesis?

One sample t-test is a statistical procedure used to determine whether a sample of observations could have been generated by a process with a specific mean

 $H_0$ : Each packet will contain an equal distribution of each coloured M&M  $H_1$ : Each coloured M&M is unequally distributed with random distribution within each packet.

We are looking if each colour's actual cardinality could be predicted from the expected value. Since we are testing whether they are perfectly distributed, the mean and expected value would be the same across all colours.

Thus, for any colour;

 $H_0$ : = 418,  $H_1$ :  $\neq$  418,

at 5% significance =0.05, with  $\mathbf{n}$ -1 degrees of freedom, where  $\mathbf{n}$  = number of possible outcomes, we have 6 colours and hence our degree of freedom is 5.

Degrees	Significance level Of.								
of	20%	10%	5%	2%	1%	0.1%			
freedom	(0.20)	(0.10)	(0.05)	(0.02)	(0.01)	(0.001)			
1	3.078	6.314	12.706	31.821	63.657	636.619			
2	1.886	2.920	4.303	6.965	9.925	31.598			
3	1.638	2.353	3.182	4.541	5.841	12.941			
4	1.533	2.132	2.776	3.747	4.604	8.610			
5	1.476	2.015	2.571	3.365	4.032	6.859			



Figure 2.1 - Significant Values Table with Degree of Freedom

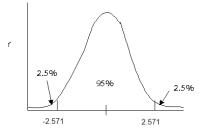


Figure 2.2 - Critical Value Region for our 95% Significance

Therefore if **t** is less than -2.571 or greater than 2.571 then we will reject our null hypothesis. Where

$$t = \frac{\overline{X} - \mu}{\frac{s}{n}}$$

Colour	Blue	Brown	Green	Orange	Red	Yellow
$ar{\mathbf{X}}$	17.43	17.43	17.43	17.43	17.43	17.43
μ	16.25	16.5	18.75	17.75	16.21	19.13
σ	2.908	3.19	3.71	3.63	3.16	3.92
n	6	6	6	6	6	6
t	2.44	1.75	-2.133	-0.53	2.32	-2.60

Figure 2.3 - Table to determine t

As you can from *Fig. 2.3*, since the t-values for blue, brown, green, orange and red lie within -2.571 and +2.571, we accept the null hypothesis for these colours. However as yellow's t-value exists outside this boundary, we reject the null hypothesis. Cleary, all our colours, except yellow (from student t-test), suggests that our observations could have been generated by a process with a specific mean. Nevertheless, since one of the colours would have been 'incorrect' we can confirm that they are not all distributed evenly/perfectly between the bags so reject the null hypothesis.

## **Section 3 - Chi-Square Test**

**Analysis** - Can we determine the difference between our observed and expected findings to conclude on a hypothesis?

Commonly, the Chi-Square Test is used to test for independence between two variables, i.e. no relationship exists between them. This test of independence is used to create a comparison between expected and observed data.

The definition of a random variable with a Chi Square distribution is stated below:

We say a random variable W has **chi** square(**d**) distribution with n degrees of freedom), denoted as  $\chi^2(n)$ , if W can be written as a sum of squares of independent standard normal random variables i.e.

$$W = Z_1^2 + \ldots + Z_n^2$$

where  $Z_i \sim \mathcal{N}(0, 1)$ .

**Definition 3.1 -** (Chi-Square Distribution). Multivariate Probability and Mathematical Statistics, An Introduction, Minmin Wang, Spring 2020.

In our case, we used this test to deduce whether bags of M&M's are evenly distributed over all 6 colours or not. This was completed by comparing the number of each colour we would expect to find with how many we observed there to be.

Our hypotheses are stated below:

H<sub>0</sub>: Each packet will contain an equal distribution of each coloured M&M H<sub>1</sub>: Each coloured M&M is unequally distributed with random distribution within each packet.

We have chosen a significance level of 5% or 0.05.

We counted a total of 2510 M&M's across 24 bags, and with 6 possible colours, we would therefore expect a total of 418 (418.33) for each colour, with the assumption that they are equally distributed. Below is a table showing our expected values against the observed data we collected, taken from *Table 1.1*.

Colour	Blue	Brown	Green	Orange	Red	Yellow
Expected	418	418	418	418	418	418
Observed	390	396	450	426	389	459

Figure 3.2 - Observed/Expected table

The Chi-Square Test requires a contribution table. This will show us which colours most fit an even distribution, i.e. closest to the expected value, and which ones 'stray' the most. The contribution value was calculated by

(observed-expected)<sup>2</sup>/expected.

In our case, we could replace the expected value by 418 for each colour.

Colour	Blue	Brown	Green	Orange	Red	Yellow
Contribution	1.876	1.158	2.450	0.1531	2.012	4.022

Figure 3.3 - Contribution Table

The smaller the contribution value, the closer the expected value is to our observed. Therefore, from *Fig. 3.3* it's clear that the orange M&M's were closest to what we expected with a contribution value of just 0.1531, whereas yellow was furthest with 4.022 as its contribution value. This was already anticipated, since there were just 6 more orange M&M's than expected but a big difference of 41 for the yellow ones.

We then calculated the Chi-Square statistic using the following definition:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

where  $\mathbf{f_0}$  represents the observed value and  $\mathbf{f_e}$  the expected.

Definition 3.4 - https://www.statisticssolutions.com/using-chi-square-statistic-in-research/

To do this, we could simply add up the contribution values we calculated in *Fig. 3.3*, giving us a value of 11.6711 for our Chi-Square statistic.

As we're completing a one-way Chi Square test, the degrees of freedom, n, is the number of cells minus one. For our test, the number of cells is equivalent to the number of possible outcomes; the 6 colours. Hence, our value for n is 5.

With our chosen significance level of 5% and our n degrees of freedom, we used the Chi-Square Distribution table to find the critical value (See Fig. 3.5).

Degree of	Probability of Exceeding the Critical Value										
Freedom	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.01		
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84	6.63		
2	0.020	0.103	0.211	0.575	1.386	2.77	4.61	5.99	9.21		
3	0.115	0.352	0.584	1.212	2.366	4.11	6.25	7.81	11.34		
4	0.297	0.711	1.064	1.923	3.357	5.39	7.78	9.49	13.28		
5	0.554	1.145	1.610	2.675	4.351	6.63	9.24	11.07	15.09		
6	0.872	1.635	2.204	3.455	5.348	7.84	10.64	12.59	16.81		

**Figure 3.5 -** <u>https://ib.bioninja.com.au/higher-level/topic-10-genetics-and-evolu/102-inheritance/chi-squared-table.html</u>

We find our critical value to be 11.07. Since  $11.07 < X^2 = 11.6711$ , we can reject our null hypothesis. In conclusion, this tells us that our bags are not evenly distributed over all 6 colours.

## Section 4 - Kolmogorov-Smirnov Test

**Analysis** - Is the data collected on each coloured M&M comparable to each respective colour?

The Kolmogorov Smirnov test (KS-test) is a complex method of detecting patterns of data sets. This method is much stronger than other tests such as the Student T-test. This is because the Student T-test requires our sample to be normally distributed. However, we know that there are multiple distributions, for instance Poisson and hence we would use the KS-test. An advantage of using the KS-test is that it makes no assumptions of the distribution (Non-parametric).

Suppose we have the observations  $X_1, \ldots, X_n$ , which we believe comes from a distribution P. The KS-test is used to test

H<sub>0</sub>: Sample comes from P H<sub>1</sub>: Sample does not come from P

We have already tested the given hypothesis using the chi-squared test in Section 3 and the student t-test in Section 2. Using the KS-test we will test if brown, green, orange, red and yellow follow the same distribution as blue.

To apply the KS-test we must use the observed distribution function:

$$F_{obs}(x) = \frac{observations\ below\ x}{observations}$$

With this we calculate the maximum absolute distance between  $F_{obs}$  and  $F_{exp}$  (expected CDF) which is represented by a step graph of the CDF that is associated with our null hypothesis. Taking  $F_{obs}$  for each colour then compare that against that distribution function of blue which is denoted  $F_{exp}$  this the hypothesised distribution  $P_0$ . The Kolmogorov-Smirnov statistic is

$$D_n = \max_{x} \left| F_{exp}(x) - F_{obs}(x) \right|$$

To do this however we must first work out the CDF of blue.

The CDF of a continuous random variable X is,

$$F_{\mathcal{X}}(x) = \int_{-\infty}^{x} f_{\mathcal{X}}(t) dt$$

which we we have use to construct the CDF below. We have also given normal curve for blue (See Fig. 4.1).

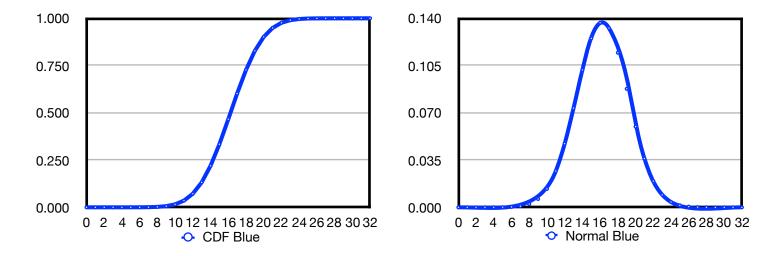


Figure 4.1 - CDF & Normal Curve for Blue

Using the below equations for the construction of the CDF,

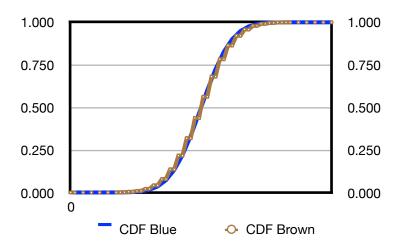
$$F_{exp}(x) = \int_{-\infty}^{x} \frac{1}{2.908\sqrt{2\pi}} e^{\frac{-(t-16.250)^2}{2(2.908)}} dt$$

$$F_{exp}(x) = 0.2932 erf(0.4147x - 6.7382) + 0.2932$$

and the final equation below to create the normal curve of blue

$$f(x) = \frac{1}{2.908\sqrt{2\pi}} e^{\frac{-(x-16.250)^2}{2(2.908)}}$$

This enabled us to use blue as an observation and hence compute step-graphs of each respective colour against blue's CDF (See Fig. 4.2).



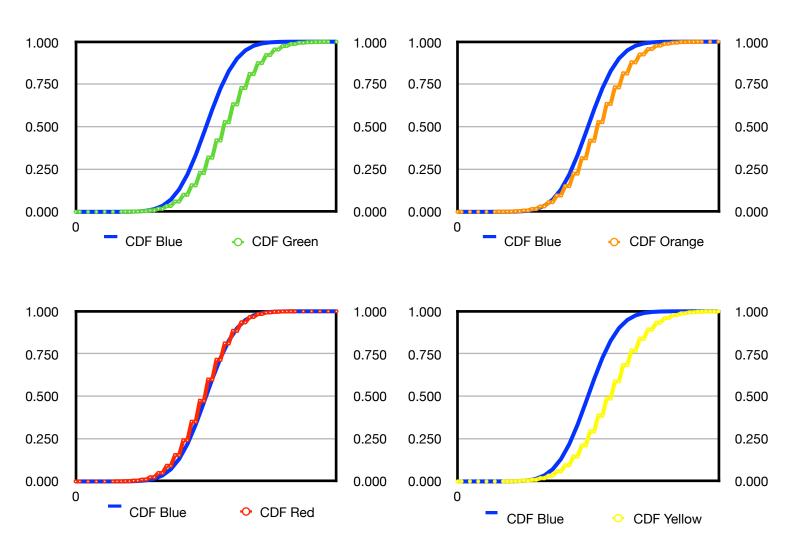


Figure 4.2 - Step-graph for each respective colour against blue's CDF

Finally, we computed a table of values for  $D_n$  from the step-graphs. This will give us an indication whether the distributions of each colour are comparable and how closely they may follow the same distribution (See Table 4.3).

Blue		Brown	Green	Orange	Red	Yellow	D_n Brown	D_n Green	D_n Orange	D_n Red	D_n Yellow
х	F_exp(x)	F_obs(x)	F_obs(x)	F_obs(x)	F_obs(x)	F_obs(x)					
0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0
7	0.001	0	0	0	0	0	0.001	0.001	0.001	0.001	0.001
8	0.002	0	0	0	0	0	0.002	0.002	0.002	0.002	0.002
9	0.006	0	0	0	0	0	0.006	0.006	0.006	0.006	0.006
10	0.016	0.042	0	0.042	0.042	0	-0.026	0.016	-0.026	-0.026	0.016
11	0.036	0.042	0.042	0.083	0.083	0	-0.006	-0.006	-0.047	-0.047	0.036
12	0.072	0.083	0.042	0.083	0.167	0.0833	-0.011	0.03	-0.011	-0.095	-0.0113
13	0.132	0.167	0.083	0.167	0.208	0.0833	-0.035	0.049	-0.035	-0.076	0.0487
14	0.22	0.25	0.167	0.208	0.292	0.0833	-0.03	0.053	0.012	-0.072	0.1367
15	0.334	0.25	0.25	0.25	0.417	0.125	0.084	0.084	0.084	-0.083	0.209
16	0.466	0.583	0.333	0.333	0.5	0.167	-0.117	0.133	0.133	-0.034	0.299
17	0.602	0.75	0.458	0.458	0.583	0.25	-0.148	0.144	0.144	0.019	0.352
18	0.726	0.792	0.583	0.5	0.75	0.375	-0.066	0.143	0.226	-0.024	0.351
19	0.828	0.792	0.625	0.583	0.833	0.583	0.036	0.203	0.245	-0.005	0.245
20	0.901	0.917	0.75	0.75	0.917	0.792	-0.016	0.151	0.151	-0.016	0.109
21	0.949	0.958	0.792	0.792	1	0.833	-0.009	0.157	0.157	-0.051	0.116
22	0.976	0.958	0.875	1	1	0.875	0.018	0.101	-0.024	-0.024	0.101
23	0.99	0.958	0.917	1	1	0.958	0.032	0.073	-0.01	-0.01	0.032
24	0.996	1	0.917	1	1	0.958	-0.004	0.079	-0.004	-0.004	0.038
25	0.999	1	1	1	1	0.958	-0.001	-0.001	-0.001	-0.001	0.041
26	1	1	1	1	1	0.958	0	0	0	0	0.042
27	1	1	1	1	1	0.958	0	0	0	0	0.042
28	1	1	1	1	1	0.958	0	0	0	0	0.042
29	1	1	1	1	1	0.958	0	0	0	0	0.042
30	1	1	1	1	1	0.958	0	0	0	0	0.042
31	1	1	1	1	1	0.958	0	0	0	0	0.042
32		1	1	1	1	0.958	0	0		0	0.042

**Table 4.3** - Results for  $D_n$ 

As you can see from the above table for brown the largest absolute value for  $D_n$  is 0.148. Since 0.148<0.270 we can accept the null hypothesis at 95% level with our critical value c=0.270.

n	α 0.01	α 0.05	α 0.1	α 0.15	α 0.2	
18	0.371	0.309	0.278	0.259	0.244	
19	0.363	0.301	0.272	0.252	0.237	
20	0.356	0.294	0.264	0.246	0.231	
25	0.320	0.270	0.240	0.220	0.210	
_	18 19 20	n 0.01 18 0.371 19 0.363 20 0.356	n 0.01 0.05 18 0.371 0.309 19 0.363 0.301 20 0.356 0.294	n 0.01 0.05 0.1 18 0.371 0.309 0.278 19 0.363 0.301 0.272 20 0.356 0.294 0.264	n 0.01 0.05 0.1 0.15 18 0.371 0.309 0.278 0.259 19 0.363 0.301 0.272 0.252 20 0.356 0.294 0.264 0.246	n     0.01     0.05     0.1     0.15     0.2       18     0.371     0.309     0.278     0.259     0.244       19     0.363     0.301     0.272     0.252     0.237       20     0.356     0.294     0.264     0.246     0.231

Figure 4.4 - Critical Values Table

Similarly with orange having a largest absolute value of  $D_n$  being 0.245. Since 0.245<0.270 we can accept the null hypothesis at 95% level with our critical value c=0.270.

Additionally, red has the largest absolute value of  $D_n$  being 0.095. Since 0.095<0.270 we can accept the null hypothesis at 95% level with our critical value c=0.270. Similarly with green having the largest absolute value of  $D_n$  being 0.203. Since 0.203<0.270 we can accept the null hypothesis at 95% level with our critical value c=0.270.

However, yellow has the largest absolute value of D<sub>n</sub> being 0.352. Since 0.352>0..270 we can reject the null hypothesis at a 95% level.

Therefore throughout the KS-test we can conclude that brown, green, orange and red follow the same distribution as blue. However yellow does not follow the same distribution (according to the KS-test at 95% significance).

Throughout our analysis we collected data, conducted multiple tests and compared findings. Within each section we noticed many similarities and of course, many differences. What was constant throughout was the closeness of the data collected on blue, brown and red, in comparison to the variation from green, orange and yellow. With blue, brown and red leading with the smallest values of standard deviation and the most viable normal curves, it was down to the many tests to help determine greater confidence in our initial findings.

What is apparent from Section 2, is the acceptation of the null hypothesis for all the colours but yellow. Visually, yellow had the worst distribution and greatest standard deviation, followed closely by green and orange. Perhaps we find yellow to reject the null hypothesis in Section 2 due to it following a weaker normal distribution. We required further analysis to achieve an accurate answer to our hypotheses.

With the final two sections potentially being much more accurate, we felt best to decide upon a conclusion taking in to account what tests are best suited for our data. Concluding in Section 3, we rejected our null hypothesis in favour of the alternate hypothesis and this was backed up by the findings in Section 4 with Section 4 remaining the most reliable source of data. The Kolmogorov-Smirnov test requires no need for normally distributed data, and since half our data is not amazingly normally distributed it seems wise to assume it is our most accurate finding. Clear to us, we noted yellow was not equally distributed throughout our analysis and to conclude with the same findings in Section 4 we finalised our conclusion.

We see no reason to accept our null hypothesis. Yes, five of our six colours may be equally distributed but for yellow this is not the case. Without yellow being equally distributed we cannot accept  $H_0$ , hence we much reject our null in favour of our alternate hypothesis  $H_1$ .

# Bibliography/References

http://www.maths.qmul.ac.uk/~bb/CTS\_Chapter3\_Students.pdf

 $\underline{http://www.stats.ox.ac.uk/\!\!\sim\!\!massa/Lecture\%2013.pdf}$ 

https://www.statisticssolutions.com/using-chi-square-statistic-in-research/

https://ib.bioninja.com.au/higher-level/topic-10-genetics-and-evolu/102-inheritance/chi-squared-table.html

Introduction to Probability Vladislav Vysotsky Autumn 2019

Multivariate Probability and Mathematical Statistics, An Introduction, Minmin Wang, Spring 2020

# **Appendix**

### **Probability Mass Function and Cumulative Distribution Function value tables**

The data collected in the below tables was calculated using the NORMDIST function with (x, mean, standard deviation, cumulative?). Each with the respective values coming from the analysis in the first section. The question, "cumulative?" requiring the answer of True or False, determining either the cumulative distribution function or probability mass function values, as given below.

		Probability M	lass Function	Values Table		
	Normal Blue	Normal Brown	Normal Green	Normal Orange	Normal Red	Normal Yellow
0	0.000	0.000	0.000	0.000	0.000	0.000
1	0.000	0.000	0.000	0.000	0.000	0.000
2	0.000	0.000	0.000	0.000	0.000	0.000
3	0.000	0.000	0.000	0.000	0.000	0.000
4	0.000	0.000	0.000	0.000	0.000	0.000
5	0.000	0.000	0.000	0.000	0.000	0.000
6	0.000	0.001	0.000	0.001	0.001	0.000
7	0.001	0.001	0.001	0.001	0.002	0.001
8	0.002	0.004	0.002	0.003	0.004	0.002
9	0.006	0.008	0.003	0.006	0.009	0.004
10	0.014	0.016	0.007	0.011	0.018	0.007
11	0.027	0.028	0.012	0.020	0.032	0.012
12	0.047	0.046	0.021	0.031	0.052	0.019
13	0.073	0.069	0.032	0.047	0.075	0.030
14	0.102	0.092	0.047	0.064	0.099	0.043
15	0.125	0.112	0.065	0.082	0.117	0.058
16	0.137	0.124	0.082	0.098	0.126	0.074
17	0.133	0.124	0.096	0.107	0.122	0.088
18	0.114	0.112	0.105	0.109	0.107	0.098
19	0.088	0.092	0.107	0.103	0.085	0.102
20	0.060	0.069	0.102	0.091	0.061	0.099
21	0.036	0.046	0.089	0.074	0.040	0.091
22	0.019	0.028	0.073	0.055	0.024	0.078
23	0.009	0.016	0.056	0.039	0.013	0.062
24	0.004	0.008	0.040	0.025	0.006	0.047
25	0.001	0.004	0.026	0.015	0.003	0.033
26	0.000	0.001	0.016	0.008	0.001	0.022
27	0.000	0.001	0.009	0.004	0.000	0.013
28	0.000	0.000	0.005	0.002	0.000	0.008
29	0.000	0.000	0.002	0.001	0.000	0.004
30	0.000	0.000	0.001	0.000	0.000	0.002
31	0.000	0.000	0.000	0.000	0.000	0.001
32	0.000	0.000	0.000	0.000	0.000	0.000

	Cumul	ative Distri	bution Fun	ction Value	s Table	
	Normal Blue	Normal Brown	Normal Green	Normal Orange	Normal Red	Normal Yellow
0	0.000	0.000	0.000	0.000	0.000	0.000
1	0.000	0.000	0.000	0.000	0.000	0.000
2	0.000	0.000	0.000	0.000	0.000	0.000
3	0.000	0.000	0.000	0.000	0.000	0.000
4	0.000	0.000	0.000	0.000	0.000	0.000
5	0.000	0.000	0.000	0.000	0.000	0.000
6	0.000	0.000	0.000	0.001	0.001	0.000
7	0.001	0.001	0.001	0.002	0.002	0.001
8	0.002	0.004	0.002	0.004	0.005	0.002
9	0.006	0.009	0.004	0.008	0.011	0.005
10	0.016	0.021	0.009	0.017	0.025	0.010
11	0.036	0.042	0.018	0.032	0.050	0.019
12	0.072	0.079	0.034	0.057	0.092	0.034
13	0.132	0.136	0.061	0.096	0.155	0.059
14	0.220	0.217	0.100	0.151	0.242	0.095
15	0.334	0.319	0.156	0.225	0.351	0.146
16	0.466	0.438	0.229	0.315	0.474	0.212
17	0.602	0.562	0.319	0.418	0.599	0.294
18	0.726	0.681	0.420	0.527	0.715	0.387
19	0.828	0.783	0.527	0.634	0.811	0.487
20	0.901	0.864	0.632	0.732	0.885	0.588
21	0.949	0.921	0.728	0.814	0.935	0.684
22	0.976	0.958	0.810	0.879	0.966	0.769
23	0.990	0.979	0.874	0.925	0.984	0.839
24	0.996	0.991	0.922	0.957	0.993	0.893
25	0.999	0.996	0.954	0.977	0.997	0.933
26	1.000	0.999	0.975	0.988	0.999	0.960
27	1.000	1.000	0.987	0.994	1.000	0.978
28	1.000	1.000	0.994	0.998	1.000	0.988
29	1.000	1.000	0.997	0.999	1.000	0.994
30	1.000	1.000	0.999	1.000	1.000	0.997
31	1.000	1.000	1.000	1.000	1.000	0.999
32	1.000	1.000	1.000	1.000	1.000	0.999