Analyzing the Gaussian Distribution Function Through Stochastic Processes

Sandy Spicer, Maddy Hagen, and Kristin Ludwicki
PHYS 310
(Dated: 2017 Sep 13)

The goal of this experiment is to create and quantify a stochastic process in order to become better familiar with the Gaussian distribution function. In doing this, we strengthen our experimental, computational, and critical thinking skills. Our experiment involves calculating the sum of two randomly chosen numerically-valued colored post-it notes out of a cup to see the distribution of sums after 60 trials. We learn how to use trial and error to obtain a reasonable fit to our data. In our particular experiment, we learn more about the truncated Gaussian distribution function and how our experiment limits certain sums but enhances others. However, we find that our results don't match with our expectations. Our results show that 9 is the most common sum, although it is more likely for it to be 7 since it has the most sum combinations. We can conclude that running more than 60 trials will likely result in a peak at 7. Regardless of how many trials we run, this experiment will always be modeled by a truncated Gaussian function because of our limits in the sum values we can produce.

I. INTRODUCTION

This introductory lab focuses on the design and measurement of a stochastic process in order to model the distribution of our data. The random process we chose involves picking two pieces of colored paper out of a cup. With each color representing a different number, we take note of the sum of our two picks. We then model our distribution with a Gaussian function that we obtain through trial and error. The purpose of this experiment is to develop an understanding of the quantities that describe the one-dimensional Gaussian function and apply it to our own experiment (amplitude, mean/center of distribution, and standard deviation). Before the experiment, we did not know that a truncated Gaussian function exists and that it applies to our experiment. It is necessary to conduct 60 trials to show this because our sum values are restricted and no outliers are possible. In order to obtain the best fit possible, we use trial and error to estimate the parameters of the Gaussian function. The main question we are trying to answer about our experiment is: How can we model the distribution of our data using a Gaussian function, and what does this tell us about our experiment?

II. PROCEDURE

To prepare our experiment, we gather six different colored post-it notes and a plastic cup. We crumple the post-it notes into balls and assign them different numerical values. Pink = 1, Yellow = 2, Blue = 3, Green = 4, White = 5, and Black = 6. We put all of the crumpled post-it notes into the cup, shake it, and pick a color without looking. We write down what color it is and put it back into the cup for a second drawing. We do this 60 times and record our results in Python. We create a large array in Python that calculates the sum of each drawing for us. We then plot the data in a histogram. Finally, we use trial and error to obtain the parameters of our Gaussian function in order to analyze our data distribution.

III. DATA

In this experiment, we conduct 60 trials, all of which are shown in Table 1 on page 2. The letters correspond to the following: b=blue, y=yellow, k=black, w=white, g=green, and p=pink

IV. ANALYSIS

The analytical goal of this lab is to use a Gaussian function to model our data distribution. The Gaussian distribution is a continuous function which approximates the number of times an event occurs in a given amount of trials. The Gaussian distribution is also referred to as the "normal distribution" and represents a bell-shaped curve. The formula is as follows:

Trial	Colors	Sum	Trial	Colors	Sum	Trial	Colors	Sum
1	b+k	9	21	g+y	10	41	g+w	9
2	y+b	5	22	g+w	6	42	y+b	5
3	y+g	6	23	w+p	9	43	b+p	4
4	w+w	10	24	p+p	6	44	b+w	8
5	k+g	10	25	w+b	2	45	p+p	2
6	w+w	10	26	k+g	8	46	w+k	11
7	y+w	10	27	y+w	10	47	k+k	12
8	b+w	7	28	w+p	7	48	w+g	9
9	w+p	8	29	w+k	6	49	y+w	7
10	w+k	6	30	p+b	11	50	g+y	6
11	k+g	11	31	k+w	4	51	g+w	9
12	k+y	10	32	k+y	11	52	g+k	10
13	g+p	8	33	b+k	8	53	p+p	2
14	b+g	5	34	g+k	9	54	b+p	4
15	w+g	7	35	g+k	10	55	y+y	4
16	p+b	9	36	y+k	10	56	w+y	7
17	g+w	4	37	k+k	8	57	w+g	9
18	b+b	9	38	y+w	12	58	w+g	9
19	k+y	6	39	w+y	7	59	k+k	12
20	g+k	8	40	w+p	7	60	k+p	7

TABLE I: Above is a table of our experimental data which includes trial, color, and sum. All 60 trials of our two-color picks and the sum of those color picks are shown.

$$G(x) = Ae^{-(x-\bar{x})^2/2\sigma^2}$$

where A is the amplitude, $\bar{x}$, is the mean or center of distribution, and σ is the standard deviation. Through trial and error, we found A to be 10, $\bar{x}$ to be 8.5 and σ to be 3.

Although the goal of this lab is to use a "normal" Gaussian function to analyze our stochastic process, we fit our data using a truncated Gaussian function, as depicted in green in Figure 1 on page 3. This means that there are no outliers in our experiment where the probability approaches zero. Since we assign values to each color and take the sum, we limit ourselves to a small amount of possible sums and restrict our domain. For example, it is impossible to attain a sum that is less than two or greater than twelve, therefore, our Gaussian function is truncated and limited by these min and max values. Additionally, different color combinations may result in the same sum (blue + black, green + white). This shows up as the same value on our histogram rather than two separate sums, which contributes to a higher probability than other sum values. Although our peak is at nine, it should be centered on seven since that has the most sum combinations. If we conduct more trials, the peak will likely shift toward seven rather than nine.

V. DISCUSSION

Our results indicate that this experiment cannot be modeled by a normal Gaussian function. Certain sum values are impossible while others are much more likely due to multiple combinations. Our results show that higher sums like nine and ten are more likely to be achieved when that should not be the case. There are only four combinations that result in a sum of nine while only two combinations result in a sum of ten. This results in a truncated Gaussian function that is shifted toward higher sum values which is incorrect. Our truncated Gaussian model does not describe our dataset well since the curve peaks at a sum of nine rather than around seven. A sum of seven is most likely to be achieved since there are six different combinations of the numbers 1-6 that result in seven. The physical significance of the amplitude in our experiment is the "most- likely" sum value that is achieved, which is nine. The center of distribution represents the average sum out of our 60 trials, which is approximately 7.68. The standard deviation represents the width of all possible sums in our experiment. In this case, it spans from two to twelve. The histogram bin size depends on the amount of sum combinations in our experiment. In this case, there are eleven sum combinations, therefore we have eleven bins.

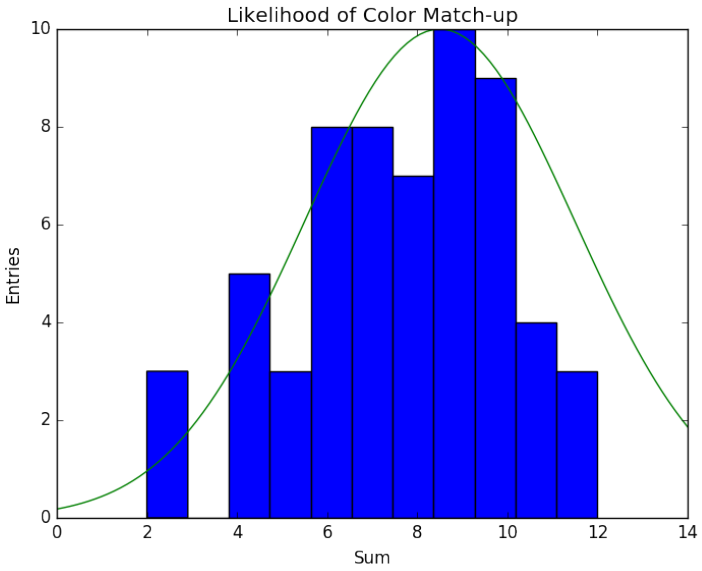

FIG. 1: Depicted above is a histogram of the likelihood of color combination sums from our experiment. Our data is in blue and a one-dimensional Gaussian function is in green. The x-axis represents the sum of the two color values that we picked from the cup while the y-axis represents how many times that sum is achieved. We use eleven bins since there are eleven different sum combinations that can occur in our experiment.

This experiment could be improved by running more than 60 trials. Our results will still be consistent with a truncated Gaussian distribution, however we would expect our amplitude to be closer to seven. Rather than fitting our dataset through trial and error, we could use the scipy package "norm.fit"

VI. CONCLUSIONS

Overall, this experiment aided in my understanding and implementation of the Gaussian function on stochastic processes. By creating our own random experiment, we learned that not all Gaussian functions are normal. Our color match-up experiment led us to the conclusion that truncated Gaussian functions are possible when limitations are introduced into our experiment. Since sum values less than two or greater than twelve were impossible, we were faced with no outliers and a restricted domain. In other words, the ends of the bell curve appear to be cut off since the sums in our experiment are limited. Although our experiment does not represent a normal Gaussian distribution, we still learned about the Gaussian function itself and learned what causes the function to truncate.