

# Dartboard: Gaussian Statistics

## Objective

In this experiment, you will be demonstrating that when a measurement is subject to many sources of random error, the distribution of measurements will obey a Gaussian. You will be proving this by throwing a “large” number of darts at a dartboard and additionally analyzing a real-life data set of your choice. The purpose of this experiment is to understand how to treat random error, as well as understand the Gaussian distribution which occurs frequently in scientific analysis and nature. For the theory associated with the experiment, you will need to refer to Chapters 4, 5, and 10 in Taylor.

## Procedure

Create the target shown in Figure 1 by drawing parallel vertical lines on a sheet of newsprint using the special meter stick that is 1.25 cm wide. Shade the center strip (zero) of the target sheet and label the other strips as indicated. Wrap the target sheet around a square section of pressed board and fasten it with masking tape.

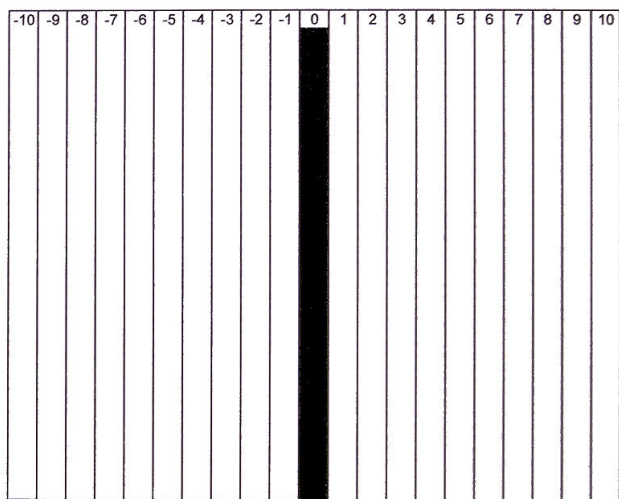


Figure 1. Target layout.

Use the attached data sheet having columns that correspond to the target sheet. Prop the target against the wall and throw darts at it from a distance of approximately 2 meters. Try to throw each dart at the central column without being

influenced by the positions of previous throws. If a dart falls out of the board before it is recorded or if a dart misses the marked columns, throw it again. Each time a set of 5 darts has been thrown, record where they landed on the data sheet.

Throw 30 sets of 5 darts (150 darts altogether). If two or three people are working as partners, they can decide between them how to alternate the throwing. When the 150 darts have been thrown, add the columns vertically to determine how many darts hit in each strip.

## Data Analysis

The total of 150 readings is taken to represent a distribution of single measurements. A distribution is a collection of a very large number of readings. Although 150 is not a very large number, it is adequate for the present purpose.

### Part 1. Mean and Standard Error

After you have thrown all 150 darts, calculate the mean, the standard error, and standard error in the mean for the distribution. As you have read in Chapter 4 of Taylor, we use the mean as our best guess of the true value (the one we would get if we took an infinite amount of measurements). This makes sense to use an estimate, since if we knew the true value, we wouldn't be doing the experiment in the first place and obviously we don't have an infinite amount of time! We also can only give our best guess of the standard error in the measurements:

$$\sigma \approx \sqrt{\left(\frac{1}{N-1}\right) \sum d_i^2} \quad (1)$$

where  $N$  is the number of measurements and  $s$  is known as the standard deviation of the sample, given by Taylor Eq. 4.9.

You also need to calculate the standard error in the mean  $\sigma_m$ . Make sure you know the difference

between  $\sigma$  and  $\sigma_m$ . You will be using a spreadsheet in Excel to help in managing this large data set as well as perform the various calculations.

**Questions** you should address in your lab write-up are the following:

1. If we assume the true value of our measurement is  $X = 0$  (the center of the dartboard where we are aiming), what is the true standard error  $\sigma$ ? How does this compare to our estimate? Note that if your throwing is biased you will need to assume a different “true value” for  $X$ .
2. What does the standard error  $\sigma$  mean? What does it say about your measurements or making future measurements?
3. What is the meaning of the standard error in the mean  $\sigma_m$ ? How many more dart throws would you have to make to gain 100 times more precision?

### *Part 2. Properties of the Gaussian.*

You will be using the program *Origin* to plot a histogram for all 150 readings, showing the frequency that each reading occurs. You will then use your data from *Part 1.* to plot a Gaussian distribution, superimposing it on the histogram.

Next you will use your histogram data to plot an experimental curve of the Gaussian integral function  $\varphi(z)$ . Just as you did with the histogram, you will plot the theoretical integral curve over your data points to compare.

### *Part 3. Plot and analyze a Gaussian distribution of data you find on your own.*

For this part you will be doing a little bit of research. You need to find some raw data that you surmise is Gaussian, histogram it as you did your darts in *Part 1.* and overlay the theoretical Gaussian curve as you did in *Part 2.* Don't worry about the integral curve here. However, be sure to discuss your results in your write up, i.e. what is their significance?

You need to pick something of interest and something that is easy to find. To make a reasonable plot and analysis, choose a data set that has more than 20 measurements (but less than 200 to save time). Note that the more data points you have the analysis is easier in a sense- your continuous curve will match much better.

The following websites provide data sets for all sorts of studies: pollution, census data, baseball statistics, etc.

<http://lib.stat.cmu.edu/datasets/>  
<http://www.seattlecentral.edu/qelp/Links.html>  
[http://factfinder.census.gov/home/saff/main.html?\\_lang=en](http://factfinder.census.gov/home/saff/main.html?_lang=en)

You are welcome to, and encouraged to find anything you like. It may be a distribution of ACT scores at Augie, the height of American females, or the percentage of African American population in a sample of Illinois counties.

I have attached a sample analysis in the appendix to show you what I am looking for.

### **Using the Spreadsheet**

Using one of the PC's in lab, login as *phlab*, password *electron*. Click on the **Start** button, then **Lab Experiments, PH 350, Dartboard07**. Insert your data for all 150 darts in column A. I have calculated the mean  $\langle x \rangle$  for you at the bottom of the column in the “averages” section. You will need to write the appropriate equations in the various cells to compute the rest: residuals for each throw  $d_i$ , and the square of the residuals  $d_i^2$ . You will then average these values to find the average of the residuals, and the average of the square of the residuals  $s^2$ . Make sure to check that these averages make sense physically. For example do you get what you expect for the average of the  $d_i$ 's? The averages will help you find  $\sigma$  and  $\sigma_m$ .

Note that there are built-in functions in Excel to find the mean and standard deviation. They are AVERAGE(num1:numN) and

STDEV(num1:numN) respectively. How do your calculations compare to the built-in functions?

To assist you in graphing, there are two additional columns on the spreadsheet. The  $x$  column contains the position numbers and the  $n$  column the number of darts stuck in each position. Enter the data from the bottom of your raw data sheet in the  $n$  column. The next column “Frequency  $n$ ” will compute the fraction of throws occurring for a particular value  $x$ . In this way, the sum of the column will be normalized to 1 rather than 150. Note that the  $x$  column has some extra values you did not make on your target. These are to be used if your data is biased to one side, i.e. if the mean is +2 instead of 0. You can globally shift the data in this case. This only should be done if your mean is outside the range of -0.5 and +0.5.

In addition to including a hard copy of the spreadsheet in your lab report, **save a copy of this file and send it to your instructor.**

## Graphing

You will be using the program *Origin* to produce high quality graphs of your results. **Graphs must be properly labeled and titled.**

### *Histogram and Gaussian.*

Use your spread sheet data from Excel to plot a histogram of the distribution in Origin. A histogram is a plot of the frequency of measurements. You bin the data in the  $x$ -coordinate and display the number of times (or frequency) a measurement occurred within these bins. Typically one normalizes the data so that the fraction of measurements is displayed as bin height rather than the raw number of counts. In your data set, you will have 21 bins (-10 through +10) with a width of 1.0 each. Ideally, the center bin, which ranges from -0.5 to +0.5, will contain the largest number of hits and describe the mean of a Gaussian distribution. For more information on histograms, see Taylor Chapter 5. An example of a histogram is also shown in the Appendix of this lab.

To plot a histogram in Origin, use the **Column Graph** button/function.

For this experiment you also need to overlay the theoretical continuous Gaussian curve. To plot a function in Origin, choose **File, New, Function, OK**. *Carefully* type in the equation you wish to plot. In your case you will use the Gaussian function whose parameters are determined by the calculations in your spreadsheet.

To overlay the function on your histogram, use the “layer icon.” This is the button is labeled “1” in the left-hand corner of your particular graphing window. Double-clicking this icon will allow you to combine elements of different plot windows on the same graph. Be sure to uncheck “rescale on ok” before plotting. It is not “ok” and will ruin all the work you did getting that theoretical curve to match the histogram appropriately.

**Don’t forget** you will need to repeat this procedure for the set of real-life data you find through the suggested websites or through your own research.

### *Integral Function & Fraction of Throws.*

This graph will contain an experimental curve of the Gaussian integral function from your histogram data and the theoretical curve overlaid.

For the experimental plot, graph the total fraction of darts that lie within a given  $\pm x$  (range of bins) against  $x$ , i.e. the total fraction in  $\pm x$  will be the  $y$ -data point, and just  $+x$  will be the  $x$ -data point. To continue the theme of normalization, you should actually plot the total fraction of darts within  $\pm x/\sigma$  vs.  $x/\sigma$ . Be careful how you choose your values for  $x$ , since the value  $x$  for the integral function really corresponds to  $2x$  bins worth of darts. Use the typical scatter plot method for graphing this data by choosing **Plot, Scatter** from the menu and picking the appropriate columns for  $x$  and  $y$  data.

To plot the theoretical trace on top of your data points, use the data provided in Appendix A of Taylor. You don’t have to input all of them into your work sheet, but enough so that the curve looks as it should (use more points in the beginning where the curve changes the quickest). Start by

making a scatter plot as you did with your experimental data. Use the “B-spline” option for **Line Connect** in the **Plot Details** window to connect this data in a nice continuous curve. How does it match to theory? What can you conclude about your data and the distribution given this good or bad match?

## Appendix

This is an example of plotting some real-life data. I used the exact same method as *Part 2.* above, just with a different data set.

What I wanted to find out was the percent chance that Barry Bonds hits a home run when at bat. I used 22 data points corresponding to 22 baseball seasons. I used Excel to divide the number of homeruns Barry hit each season by the number of at bats for that season (the column labeled %; note to get percent here I really should have multiplied by 100). I calculated the mean  $\langle x \rangle$  to be 0.256 and the standard error  $\sigma$ , to be 0.096. I then made a histogram of this data (Fig. 2) and overlaid the theoretical Gaussian using my values of  $\langle x \rangle$  and  $\sigma$ .

**Table I. Data**

Season	HR	At bats	%
1986	16	413	0.038741
1987	25	150	0.166667
1988	24	144	0.166667
1989	19	159	0.119497
1990	33	151	0.218543
1991	25	153	0.163399
1992	34	140	0.242857
1993	46	159	0.289308
1994	37	112	0.330357
1995	33	144	0.229167
1996	42	158	0.265823
1997	40	159	0.251572
1998	37	156	0.237179
1999	34	102	0.333333
2000	49	143	0.342657
2001	73	153	0.477124
2002	46	143	0.321678
2003	45	130	0.346154
2004	45	147	0.306122
2005	5	14	0.357143
2006	26	130	0.2
2007	27	116	0.232759

$\langle \% \text{ HR} \rangle =$	0.256216
$\sigma =$	0.095374

The only tricky thing here was to make a judicious choice of bins. For my plot in Fig. 2, I chose six bins with spacing of 0.05 in order to observe a more Gaussian-like distribution. Note that one has to be careful with normalization when plotting the

theoretical curve atop the histogram. In *Part 2* of the lab, we “chose” a bin width of 1.0. This ensured that the area under the histogram was also 1.0 so that the normalized continuous Gaussian curve of area 1.0 matched nicely. When you arbitrarily choose a bin width, this will not be true, you either need to renormalize the bins, or renormalize the Gaussian. I chose the latter and multiplied the Gaussian by 0.05, i.e. my bin width. The moral of the story is to just make sure you compare curves of the same area. To compare apples to apples, the continuous Gaussian curve and the discrete bins need to have the same area.

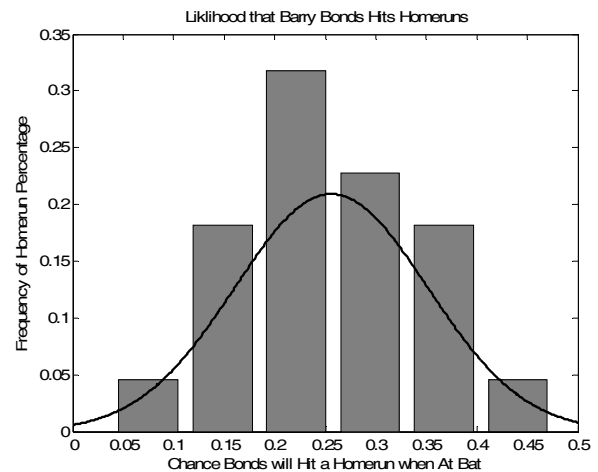


Fig. 2 Likelihood of Barry Bonds hitting a homer

From the data of 22 baseball seasons, we see that Barry Bonds’s average homerun percentage for a season is  $25.6\% \pm 1.9\%$ . This means that roughly one out of any four times he is at bat, he is likely to hit a home run. The standard error in the distribution is 9.5% which means that if we randomly pick a season, we can say with 68% confidence that Barry will perform within  $\pm 9.5\%$  of the 25.6% chance for hitting homeruns. So 68% of the time you will have witnessed a fairly good season, 16.1% chance of hitting a homer for a given at bat on the low end and 35.1% on the high end.

## Data for Dartboard Experiment

	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10
1																					
2																					
3																					
4																					
5																					
6																					
7																					
8																					
9																					
10																					
11																					
12																					
13																					
14																					
15																					
16																					
17																					
18																					
19																					
20																					
21																					
22																					
23																					
24																					
25																					
26																					
27																					
28																					
29																					
30																					
Total																					