University of Alabama
Department of Physics and Astronomy

PH115 / LeClair August 23, 2019

Laboratory 1: Uncertainty Analysis

Hypothesis: A statistical analysis including both mean and standard deviation can reveal whether a set of cards is "stacked" or not, even without seeing all the cards.

Goals: Understanding of uncertainty and basic statistical quantities, comparisons of data sets, visualization of data.

Symbols:

x	The quantity being measured.
x_i	A particular measurement of that quantity.
σ or σ_x	The standard deviation of the quantity being measured (the "spread").
n	The number of measurements taken.

1 Introduction

In this exercise we are interested in learning how to correctly analyze data. The main point here is that any time one measures a quantity, one must be able to tell the accuracy of this quantity. Otherwise, there is no way to tell whether the number agrees with the predictions of a theory, and there is no way for another scientist to check the experiment.

For example, suppose I measure the circumference of a circle, then its diameter, and divide the circumference by the diameter. The result ought to be π which we know is approximately 3.14159 .... If my result is 3.15, have I proved that π is not approximately 3.14159...? In this case, of course, we know the accepted result. If the uncertainty of my measurement is 0.01 or more, then my result is consistent with the value that we are familiar with.

No matter how many measurements of a quantity we make, we will never know its true value. If we make a large number of measurements under nominally identical conditions, then the average of this collection of measurements gives us an *estimate* of the true value. We might logically expect that the more measurements we make, the better our estimate of the true value. In some cases, the underlying statistics of the randomness in the measurements allows us to determine how far our estimate is from the true value. Repeated measurements of independent, random events occurs often in physics, and the goal of this laboratory is to learn how to analyze such experiments using processes more familiar to everyday life.

What we will first explore in this laboratory are data analysis techniques that will allow you to determine, from a series of measurements, what the uncertainty in the measured quantity is. In a follow-up experiment, we will learn how to analyze data that appear to follow a linear relationship.

1.1 Standard Deviation

Suppose a series of measurements is made of the value of some unknown quantity. Usually these measured values will not all be the same. A statistical analysis of the measured values estimates the quantity and its variability. Analysis of the uncertainty determines the probability that the "true" value lies within a certain range. Of course, the percent difference between two measured values gives some idea of the range of measured values to be expected, but this is not a very reliable indicator. **Whenever a measurement is reported, an estimation of the accuracy of the measurement procedure and a determination of the reliability of a measurement are equally important to report.** The mathematical methods used to determine reliability (statistical uncertainty) are commonly referred to as uncertainty analysis.

A mathematically complete treatment of error analysis is beyond the scope of this course, but is necessary to understand the basic methods of uncertainty analysis to properly report the results of our the experiments. Some starting assumptions are useful to estimate the uncertainties encountered in the measurements and analysis of data. **First**, it is assumed that differences in measurements are due to small random fluctuations that are just as likely to make the measurement higher as it is to make it lower. **Second**, in some cases there is a systematic error which always makes a measurement smaller or larger than the "true" value. Examples of systematic errors include parallax in reading a meter stick, friction the bearings in a balance or meter, tightening of an adjustment screw too much, failure to account for air resistance, etc.

In well-designed experiments, systematic errors are accounted for, noted and measured. Under these conditions, a very large number of measurements of the same quantity should distribute themselves symmetrically about the average, which is the "best" value of the quantity. The expected variations of the measurements can be described by a quantity called the "standard deviation"

The standard deviation is computed in a straightforward manner. Suppose the quantity x is measured n times. The measured values are labelled x_1, x_2, $\ldots x_n$. First, we calculate the *mean*, or average of all the values, denoted $\overline{x}$. This is just as you would expect: add up all the measurements

$(x_i = x_1, x_2, x_3, \ldots)$ and divide by how many there were (n).

$$\text{average } x = \overline{x} = \frac{x_1 + x_2 + x_3 + \ldots x_n}{n} = \frac{\text{sum of measurements}}{\text{number of measurements}} \tag{1}$$

In more compact notation (don't worry if this is not familiar):

$$\overline{x} = \sum_{i=1}^{n} \frac{x_i}{n} \tag{2}$$

Now we want an idea of how much the individual measurements collectively scatter about the average. For this we want the difference between each measurement and the average, $x_i - \overline{x}$. We can't just add these all together, however: if the uncertainty is random, there will be about as many measurements above the average as below, and the differences will all cancel out. To make sure we only have positive quantities to characterize the deviations from the mean - we just care about deviation, not whether it is too high or too low - we square the deviations: $(x_i - \overline{x})^2$. We add all the squared deviations together, divide by the number of measurements n minus one (since we need at least two measurements to have a collection), and take the square root of the result. [i]

$$\text{standard deviation of } x = \sigma_x = \sqrt{\frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + (x_3 - \overline{x})^2 + \ldots (x_n - \overline{x})^2}{\text{number of measurements - 1}}} \tag{3}$$

This is the *standard deviation*, a characterization of the overall spread of all measurements about the average. If the measurements are very scattered, it is high; if the measurements are all close to the average, it is low. Again, in more compact notation:

$$\sigma_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2} \tag{4}$$

To recap: a large standard deviation indicates that the data points are on the whole spread far from the mean, while a small standard deviation indicates that they are on the whole clustered closely around the mean. *The standard deviation of a group of measurements gives an indication of the reliability and uncertainty of those measurements, as well as the expected range within which subsequent measurements are expected fall.* Put another way, the standard deviation tells you "how much can I trust this measurement" and "in what range to I expect most of subsequent measurements to fall?"

[i] We are ignoring the distinction between *population* and *sample* standard deviation here. While it is an important conceptual point, for the amount of data we are going to take the operational difference is nil.

When deciding whether measurements agree with a theoretical prediction, the standard deviation of those measurements is of crucial importance: if the mean of the measurements too many standard deviations away from the prediction, then the theory being tested probably needs to be revised.

1.1.1 Standard deviation and distribution of data about the mean

When performing a series of measurements, any given observation is rarely more than a few times the standard deviations from the mean. A mathematical result known as *Chebyshev's inequality* tells us, for all distributions of measurements in which standard deviation can be meaningfully defined, the number of measurements we expect within a certain number of standard deviations of the mean, summarized in the table below.

minimum population in range	**distance from mean**	**expected frequency outside range**
50%	$\pm$ 1.4σ	1 in 2
75%	$\pm 2\sigma$	1 in 4
89%	$\pm 3\sigma$	1 in 10
94%	$\pm 4\sigma$	3 in 50
96%	$\pm 5\sigma$	1 in 25
97%	$\pm 6\sigma$	3 in 100

Table 1: *Minimum expected fraction of the data lying within a certain number of standard deviations for an* arbitrary *distribution.*

According to this result, there is a 75% probability that any additional measurement made of the quantity x will lie within $\pm 2\sigma$ of the mean and a 94% probability that it will lie within $\pm 4\sigma$ of the mean. In most of the experiments of this course, measurements are repeated about five or ten times. Using the above analysis for less than five independent measurements of a quantity is generally not considered to be very reliable.

This result is quite general, and in fact rather conservative. If we know that our data should follow a *particular* distribution, such as the normal distribution (Gaussian or "bell curve" distribution), the constraints can be even more stringent. The table below shows the expected fraction of data within a certain number of standard deviations for data following a normal (bell curve) distribution.

The central limit theorem of statistics says that the distribution of an average of many independent, identically distributed random measurements tends toward the famous bell-shaped normal distribution, and it is encountered very often. This is most often the situation in our laboratory experiments, and we will typically assume it to be the case. The normal distribution is strongly peaked about the mean, making it very unlikely to see measurements more than a 2 or 3 standard deviations from the mean. For example, if one observes an event which occurs once per day, a 4σ

population in range	distance from mean	expected frequency outside range
50%	$\pm 0.67\sigma$	1 in 2
68%	$\pm 1\sigma$	1 in 3
90%	$\pm 1.65\sigma$	1 in 10
95.4%	$\pm 2\sigma$	1 in 22
99.7%	$\pm 3\sigma$	1 in 370
99.994%	$\pm 4\sigma$	1 in 16000
99.99994%	$\pm 5\sigma$	1 in 1,700,000

Table 2: *Minimum expected fraction of the data lying within a certain number of standard deviations for a* normal *(bell curve) distribution.*

event occurs every 43 years, a 5σ event occurs only once every 5000 years, and a 6σ event only once every 1.5 million years!

1.2 Relationship of mean and standard deviation

The standard deviation is a measure of the random statistical uncertainty in a set of measurements, and it becomes part of the experimental error associated with a measurement. If you have measured an average to be 695 with $\sigma = 26$, and another experimenter has measured an average of 680, then their count agrees with your count, within the statistical uncertainty. Nothing is made of the difference between the values 695 and 680 because, in all probability, the two results are the same since they differ by less than σ. That is, because the difference between the two is less than σ, there is a better than even chance the difference is meaningless.

This is not the only sort of uncertainty we are interested in, however. If we make repeated measurements of a quantity, we would expect that the more measurements we take the more accurate our mean becomes. This makes some sense - we would expect that our average after 100 measurements should be much more accurate than our average after only 10. What we are really asking is how close is the mean value we have measured to the *true* mean, determining which would require an infinite number of measurements. As it turns out, this quantity is nothing more than the standard deviation divided by the *square root* of the number of measurements.

$$\text{uncertainty in average} = \frac{\sigma_x}{\sqrt{n}} \tag{5}$$

The quantity tells us how far our measured mean could be from the true value, based only on random uncertainty. It tells you that if you measure x repeatedly, the average result $\overline{x}$ *itself* has an uncertainty $\sigma_x/\sqrt{n}$ compared to the true mean. Since the uncertainty in the average goes down as n gets larger, that means that the more measurements you perform, the smaller the uncertainty in the mean value of your measurements. This makes sense - of course making more measurements should give you a better result! The uncertainty is reduced as $1/\sqrt{n}$, which will never reach zero, but it can be reduced to an arbitrarily small value simply by taking more and more measurements. Of course, this only works if you have arbitrary amounts of time – while your accuracy increases as $1/\sqrt{n}$, the amount of time your measurement takes grows more quickly (as n), so you are fighting a losing battle! For example, to double your accuracy compared to a single measurement, you need $n = 4$ so $\sqrt{n} = 2$. This takes four times as long - annoying, but not terrible. To improve your accuracy by ten times requires $n = 100$ so $\sqrt{n} = 10$, a hundred times longer than a single measurement! There is then a question of how much time you can afford, and whether the thing you are measuring is stable for that long.

Put another way, if you are primarily interested in the average value of x, then $\sigma_x/\sqrt{n}$ tells you the uncertainty in that average, which you can use to properly report average quantities with statistical uncertainty:

$$\text{(best value of } x) = \overline{x} \pm \frac{\sigma_x}{\sqrt{n}} \qquad \text{(68\% confidence if normally distributed)} \tag{6}$$

This tells the reader not just the average, but how accurately you were able to determine it. Whether one reports $\pm\sigma_{\overline{x}}$, $\pm 3\sigma_x/\sqrt{n}$, or even $\pm 5\sigma_x/\sqrt{n}$ as the margin of uncertainty varies from discipline to discipline. In the present experiment, we will use $\pm\sigma_x/\sqrt{n}$.

2 Example: is the deck stacked?

As an everyday example of how standard deviation can be used, we will consider the following problem: how could we tell whether or not a deck of playing cards is legitimate without seeing all of the cards? Making the problem more concrete, we will imagine that we have a collection of several decks of cards shuffled together (say, 4 decks), used to play a two-person game of poker. During this game, ten cards are dealt and counted, and then returned to the deck which is thoroughly shuffled. Seeing only 10 cards at a time, could we determine if the deck is legitimate? Moreover, is there a technique which would work no matter how many decks are shuffled together?

First, we must find a way to quantify the cards. We will number the cards ace through king with the numbers 1 through 13. The numbered cards simply have their face value, and we assign Ace = 1, Jack = 11, Queen = 12, King = 13. In a normal deck of cards, there are an equal number of each type of card, so you can quickly convince yourself that after enough deals, the *average* value of all cards seen should be $\overline{x}=7$. We really just need to average over the 13 cards in one suit, since each of the four suits in the deck have the same numbers, and all decks shuffled together are the same:[ii]

$$\overline{x}_{\text{suit}} = \overline{x}_{\text{decks}} = \frac{\text{sum of all cards}}{number of cards} = \frac{1+2+3+4+5+6+7+8+9+10+11+12+13}{13} = 7 \qquad \text{legitimate deck}$$

[ii]You can calculate this more quickly by noting that $\sum_{i=1}^{n} i = \frac{1}{2}n(n+1)$.

Let's say someone removed all of the aces. With our numbering scheme (Ace = 1), we now have fewer low cards, so the average will be a bit too high (note also that there are only 12 cards per suit now, 2 through King):

$$\overline{x}_{\text{suit}} = \overline{x}_{\text{decks}} = \frac{2+3+4+5+6+7+8+9+10+11+12+13}{12} = 7.5 \qquad \text{no aces in deck}$$

Right away, by observing the average of many cards we can see something is wrong, *even without seeing all the cards*. This is purely theoretical at the moment. For an actual measurement, we would need make sure we did enough measurements such that the 0.5 difference for the stacked deck was larger than the uncertainty in our measured mean, using the standard deviation of the mean. In practice, this means perhaps 50 or 75 measurements.

Of course, the person stacking the deck may understand this point of mathematics, and can easily devise a method to fool you: remove one high and one low card, such that the average is the same. For instance, if all the Aces and Kings were removed, the average is now (with 11 cards per suit remaining):[iii]

$$\overline{x}_{\text{suit}} = \overline{x}_{\text{decks}} = \frac{2+3+4+5+6+7+8+9+10+11+12}{11} = 7 \qquad \text{no aces or kings in deck} \quad (7)$$

Merely using the average is no help! I actually performed this experiment, drawing 10 cards at a time from 4 decks, reshuffling, and drawing again until I reached 150 cards. Below is a plot of the running average of all cards seen as a function of the number of cards seen. One can see that by about 50 cards the average has stabilized at about 7 as expected, both for a clean deck and a "stacked" one. The insignificance of the difference between the two is more apparent if one calculates the uncertainty of the mean ($\sigma/\sqrt{n}$)at each point and uses it to draw error bars on the plot, also shown below. Also note that as the number of measurements (trials) increases, the error bars get smaller because our uncertainty in the average gets better.

The fact that the error bars for both measurements overlap indicates that, within the statistical accuracy of the measurements, they are not different - at 150 trials the uncertainty in the average is about 0.3, much smaller than the difference observed, meaning the difference is not significant. A simple mean measurement will not tell the decks apart. What to do? By removing the most extreme cards, those farthest from the mean, our opponent has not altered the *mean*, but he or she has altered the *distribution* of cards about that mean. With less cards lying farther from the mean value of 7, we should find a smaller standard deviation, since this is essentially what the standard deviation is designed to measure! Below is a plot of the measured standard deviation for clean and

[iii] Of course this would be obvious after some time too, so you might just remove half the aces and kings, but the logic is the same.

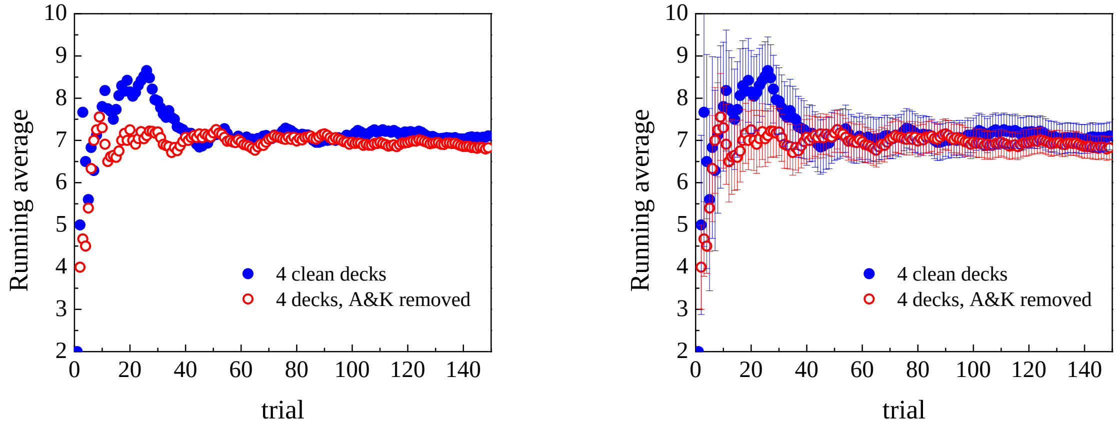

Figure 1: left *Running average as a function of the number of cards drawn. After some initial variability, there is no significant difference between the "stacked" and "clean" decks.* **right** *This is even more apparent when we include error bars representing plus and minus one standard deviation of the mean. When the error bars overlap, there is no statistically significant difference between the two data sets.*

stacked decks as a function of the number of draws.

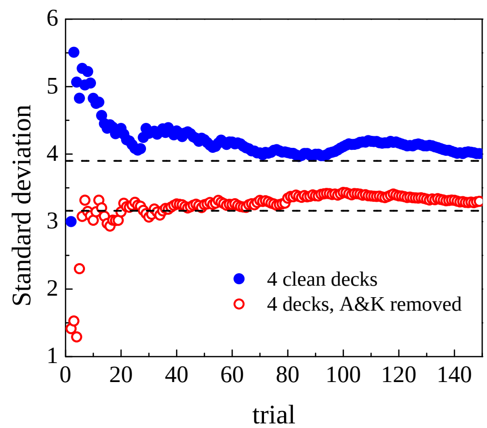

Figure 2: *Running standard deviation as a function of the number of cards drawn. There is a distinct difference between the two decks, reflecting the fact that the "stacked" deck no longer has a uniform distribution of cards. The dashed lines show the theoretically expected standard deviation for each deck.*

It is now apparent that the 'stacked' deck has a much smaller standard deviation, telling us right away that some of the extreme-valued cards must be missing. (The fact that the mean is unchanged tells us that the missing cards must together have an average value of 7.)

The plot above shows the measured standard deviation, does it agree with the theoretical value? For a clean deck, we know exactly what cards are present, so we can calculate what the standard deviation would be if we simply looked at all the cards. We start with the mean $\overline{x} = 7$ and our equation for standard deviation.[iv] Since all suits are the same in a deck, and all decks in our stack are the same, a calculation for a single suit ($n = 13$ cards) is sufficient. For a clean deck, the result is

$$\sigma_{\text{suit}} = 3.89 \tag{8}$$

This is in pretty good agreement with my measured result of 4.00.

3 Preparatory Questions

Comment on these questions in your report.

1. Suppose we removed the kings and queens from a deck of cards. Would you expect the mean value to increase or decrease? The standard deviation?
2. How about if we removed the 2's and queens?
3. Calculate the expected mean and standard deviation for a deck of cards in which all of the aces and kings have been removed.

4 Supplies & Equipment

1. Two samples of playing cards
2. Computer with Excel
3. Group of 2-4 students

5 Suggested Procedure

Each group should receive two samples of cards, of roughly 150 cards each. Each sample is taken from a large collection of cards (about 20 decks each), and thus each sample represents only a small fraction of the total number of cards in each collection. One of the two collections is comprised of "clean" decks of cards, the second collection is made up of "stacked" decks. Using statistical analysis, you can tell which set of cards comes from "stacked" decks even though you will not be able to see all the cards.

1. Label your samples of cards A and B and do not mix them. As yet, you do not know which one is from the clean decks, and which is from the stacked decks.
2. Pick one of the samples, and draw out 5-10 cards.
3. Record their numbers (using the table below as an example; using Excel is clever).
4. Return the drawn cards to the sample, and shuffle thoroughly.

[iv] You can also do this very easily in Excel using the stdev() function.

5. Repeat steps 2-4 until you have drawn out a total of about 75 cards.
6. Repeat steps 2-5 for your second sample of cards

draw i	**card**
1	2
2	8
3	13
4	3
...	...

6 Data Analysis

Once you have acquired your data, calculate running mean and standard deviation as a function of the number of points taken for each sample. Given that you have many data points, it is far easier to do the work in Excel, which has a built-in function for calculating standard deviation. The figure below shows an example table, along with the requisite formulas.

draw	card	running average	running standard deviation	
1	2	2.00		
2	8	5.00	4.24	
3	13	7.67	5.51	
4	3	6.50	5.07	
5	2	5.60	4.83	

	A	B	C	D	E
1	draw	card	running average	running standard deviation	
2	1	2	2.00		
3	2	8	=AVERAGE($B$2:B3)	=STDEV($B$2:B3)	
4	3	13	^ drag down	^ drag down	
5	4	3	6.50	5.07	
6	5	2	5.60	4.83	
7					

Figure 3: *Letting Excel do the hard work ... the upper portion of the figure shows a data table and calculated average and standard deviation, the lower portion reveals the formulas required. Type these formulas in the second row, hit enter, and drag them downward to the last row of data.*

Once you have analyzed your data, plot the standard deviation (y axis) as a function of the number of cards drawn (x axis) using Excel. For the entire set of data (i.e., after 75 cards for each sample), calculate the standard deviation of the mean as well.

7 Discussion Topics for Report

- Did you draw enough cards for the mean and standard deviation to stabilize at a roughly constant value?

- Are the average values significantly different for the two samples? How can you *quantitatively* state this?
- Can you tell which deck is "stacked" from your statistical analysis? Why?
- Can you hypothesize about how it was "stacked" from the statistical data alone?
- Would it be possible to devise a stacking of the deck that leaves both the mean and standard deviation unchanged? Why?

8 Format of Report

Your report need not be formal, the format is largely up to you (though we suggest you follow the template). Answer all the questions above, turn in plots of average and standard deviation for each sample of cards, and your overall conclusions. Be sure to note the mean and standard deviation of each sample. Address the discussion topics briefly.