PHYS 314A: Electric and Magnetic Fields Lab

Statistics and Analysis Tools

Statistics comes in two varieties: the dry, arcane stuff they teach you in statistics classes, and the stuff you actually need to know to analyze data. Similarly, spreadsheets can be used for two purposes: to make pretty tables, and to save you a ton of time doing arithmetic on data. The goal of this handout is to teach you how to employ statistics and spreadsheets to analyze data, without wasting your time and without a lot of useless jargon.

Statistics

Averages

Let's start off with the basics. If I take a large number of measurements of the same phenomenon – say, the mass of a bunch of marbles – I won't always get the same answer. The marbles will all be slightly different in size, and my scale isn't perfectly accurate. Most of the time, the best way to estimate the typical measurement is the **sample mean**:

$$\bar{x} = \frac{\Sigma\, x_i}{N} \qquad (1)$$

where x_i represents each of the actual measurements, and *N* is the total number of measurements I make. Add 'em all up and divide by how many you've got.

When the data has outliers – data points that are far bigger or far smaller than average, it makes more sense to report the **median** – the measurement which is in the middle, with half of the measurements greater than it, and half less. Student test grades, for example, might be between 85% and 95% except for one student who got a 20% – in that situation, the median is better than the mean at describing how a typical student performed.

Uncertainty

Suppose I measure a bunch of marbles, each of which is a little bigger or smaller than average. Suppose I want to estimate how far from average my next marble likely to be. This is the **uncertainty** or **scatter** of the measurement, and can be found by calculating the **sample standard deviation** of the data:

$$\sigma_x = \sqrt{\frac{\Sigma\,(x_i - \bar{x})^2}{N-1}} \qquad (2)$$

The denominator is *N-1* rather than *N* when we use equation (1) to calculate the mean $\bar{x}$, which is usually the case. The standard deviation has the same units as the original measurements, and is set up so that about ⅔ of the measurements will lie within 1 standard deviation of the average, and 95% will lie within 2 standard deviations. 2 standard deviations is usually called the **"95% confidence interval"**. That is, if I weigh marbles and calculate $\bar{x}$ = 4.8 grams and σ_x = 1 gram, the 95% confidence interval is 2, so I can be confident that 95% of the marbles will weigh between 4.8 – 2 = 2.8 and 4.8 + 2 = 6.8 grams.

Very often, we want to know how accurate our estimate of the *average* is. Suppose I measure the average mass of a set of 10 marbles: how similar will my results be if I average a different set of 10? Surprisingly, the answer is not σ_x! The **standard deviation of the mean** is

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{N}} \qquad (3)$$

where *N* is the number of measurements and σ_x is from (2). If I take the average of ten marbles a whole bunch of times, about ⅔ of my averages (not the individual marbles!) will lie within the range $\pm\sigma_{\bar{x}}$.

Notice a *very important* fact: I can get a very good estimate of the average (small $\sigma_{\bar{x}}$) even if the measurements themselves are very uncertain (large σ_x). I just have to take a lot of measurements.

For example, if I'm trying to find the average mass of my marbles, and I know that the masses have a range of uncertainty of $\sigma_x = 1$ gram, if I take an average of 4 of marbles, I know that the mean has an uncertainty of $\sigma_{\bar{x}} = 1/\sqrt{4}$ = ½ gram. If I take an average of 9 marbles, the mean is certain to within ⅓ gram, and so on. Notice that this depends on the square root of the number of measurements, so if I want an uncertainty of 1/1000, I'm going to have to make 1,000,000 measurements! Repeating your experiments as many times as possible is the key to good accuracy.

In general, **when I ask you to measure something, you should always take as many measurements as you can, and report the mean (equation 1) and the 95% confidence interval, which is which is 2 times the standard deviation of the mean (equation 3).** So if you calculated $\bar{x}$= 4.8 grams and $\sigma_{\bar{x}}$= 0.25 grams, I would say that "the average marble weighs 4.8 ± 0.5 grams."

Finally, we often want to know **whether two quantities are equal**, when one or both of them are uncertain. For instance, suppose you weigh a bunch of marbles, and find that the average weight is 4.8 ± 0.5 grams. But the package says that the marbles should weigh 5 grams. Is the manufacturer cheating you, or is this just an unlucky coincidence? For the purposes of this class, you should simply see whether the one value (5 grams) lies within the range of uncertainty of the other (4.8 ± 0.5 grams). In this case it does, so your marbles are not **significantly different** from the manufacturer's claims.

(By the way, this is a simplified technique. There is a correct procedure for comparing uncertain quantities, called *Student's T-test*. It's more accurate, and should be used for serious publications, but is not necessary for your lab reports.)

Propagation of Uncertainty

Often, we will need to do a mathematical calculation on several uncertain quantities, and want to know how uncertain the answer is. For example, if I have a bag containing 100 ± 5 marbles, each of which weighs 5 ± 1 grams, how much does the bag weigh? Here are the rules:

Addition/Subtraction: if the uncertainty of quantity *x* is Δx and the uncertainty of *y* is Δy, then their sum

$$z = x + y$$

is uncertain by an amount

$$\Delta z = \sqrt{\Delta x^2 + \Delta y^2} \qquad (4)$$

The uncertainties add like independent vectors, using a sort of Pythagorean theorem. Subtraction is exactly the same: if

$$z = x - y$$

then

$$\Delta z = \sqrt{\Delta x^2 + \Delta y^2}$$

Notice that subtracting two uncertain quantities makes the uncertainty *bigger*, even though *z* gets smaller!

For **multiplication by an exact number**, multiply the uncertainty by the same exact number.

$$z = A\,x$$
$$\Delta z = A\,\Delta x$$

Relative uncertainty:
For what follows, it's useful to define the "relative uncertainty" or "relative error" of a quantity. For example, the relative uncertainty of *x* is

$$\frac{\Delta x}{x}$$

This is typically expressed as a percentage: if Δx = 2 and *x* = 20, we would say that *x* has an uncertainty of 10%.

Multiplication/division: If

$$z = x \bullet y$$

then the *relative* uncertainty of *z* is

$$\frac{\Delta z}{z} = \sqrt{\left(\frac{\Delta x}{x}\right)^2 + \left(\frac{\Delta y}{y}\right)^2} \qquad (5)$$

For multiplication, the *relative uncertainties* add using a Pythagorean theorem-like equation. Division uses exactly the same equation.

Note that if *y* is known perfectly (for example, if it's just a fixed number), then the second term in the square root drops out, and the relative error in *z* is equal to the relative error in *x.*

Raising to a Power: If

$$z = x^n$$

then the relative error is

$$\frac{\Delta z}{z} = |n| \frac{\Delta x}{x} \qquad (6)$$

Arbitrary functions: If

$$z = f(x)$$

then

$$\Delta z = f'(x)\, \Delta x$$

Example: Suppose I wanted to measure the energy stored in a spring. I know the spring constant *k* = 50±4 N/m, I know the spring was stretched a distance *x* = .04±0.001 mm, and I know the equation for spring energy:

$$U = k\, x^2$$

The relative error in *k* is 4/50 = 8%, and the relative error in *x* is .001/.04 = 2.5%. From equation (6), I know the relative error in x^2 is thus 5%. This means the relative error in U is

$$\frac{\Delta U}{U} = \sqrt{\left(\frac{\Delta k}{k}\right)^2 + \left(\frac{\Delta x^2}{x^2}\right)^2} = \sqrt{.08^2 + .05^2} = .094 = 9.4\%$$

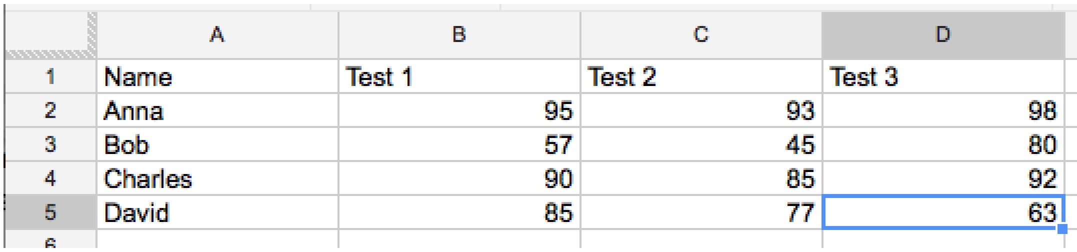
Notice that combining uncertain quantities always makes the uncertainty bigger.

A shortcut: If you have a complicated math equation, and you know that some quantities are known very very accurately, you can ignore them (treat their uncertainty as zero) in your error calculation.

Spreadsheets for Data Analysis

Some of you may have used Microsoft Excel or Google Spreadsheets before, but many students don't know their true power. Spreadsheets are not just a way to lay out tables of data: they allow you to do math on that data. **If you ever find yourself plugging a whole bunch of numbers into your calculator, use a spreadsheet instead!**

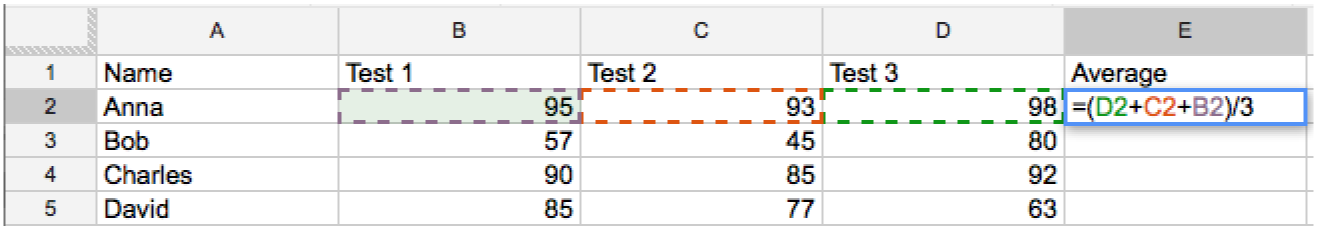
If you haven't used a spreadsheet before, here's a quick example showing how it works: calculating students' average grades for a class. Let's suppose I give three tests, and want to calculate each student's average test grade. I open up my favorite spreadsheet (I like Google Docs, but Microsoft Excel and Apple Pages works the same way) and type in the students' names and raw scores:



	A	B	C	D
1	Name	Test 1	Test 2	Test 3
2	Anna	95	93	98
3	Bob	57	45	80
4	Charles	90	85	92
5	David	85	77	63

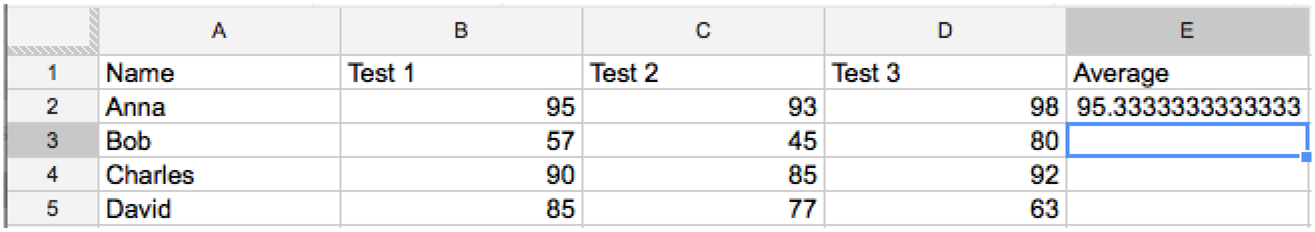
Doing math with spreadsheets

Now, I want to calculate the average test score for each student. In space E2 (column E, row 2), I type "=" to tell the spreadsheet I want an equation, then enter the following:



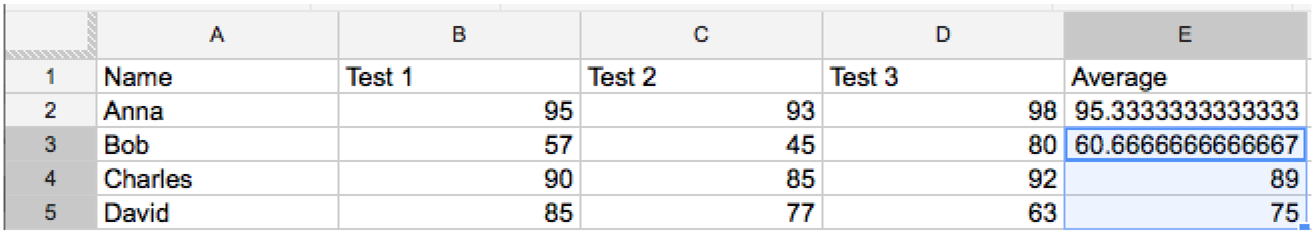
	A	B	C	D	E
1	Name	Test 1	Test 2	Test 3	Average
2	Anna	95	93	98	=(D2+C2+B2)/3
3	Bob	57	45	80	
4	Charles	90	85	92	
5	David	85	77	63	

This means "add up what's in spaces B2, C2, and D2, and divide by 3". You can either type in "B2", etc., or click with the mouse on each square to add it to the formula. When I hit return, the spreadsheet calculates the result:



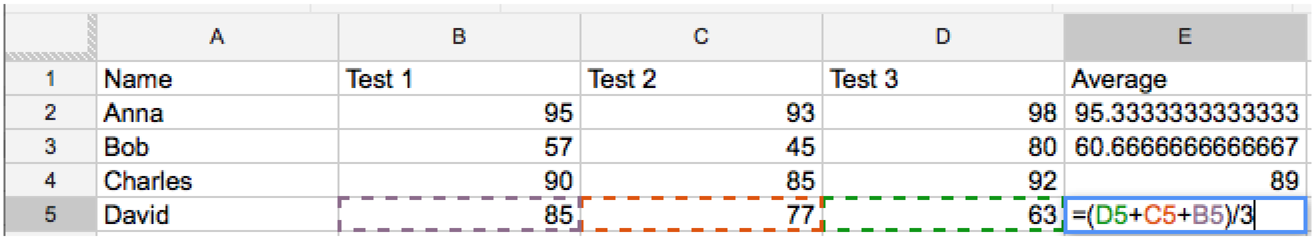
	A	B	C	D	E
1	Name	Test 1	Test 2	Test 3	Average
2	Anna	95	93	98	95.3333333333333
3	Bob	57	45	80	
4	Charles	90	85	92	
5	David	85	77	63	

But here's where it gets cool. To calculate Bob, Charles, and David's grade, I just copy and paste space E2 into spaces E3 and E4: this copies the *formula*, not the number!



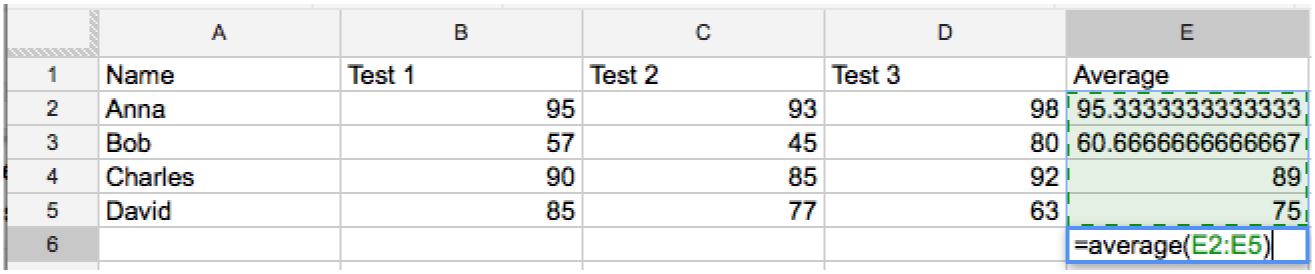
	A	B	C	D	E
1	Name	Test 1	Test 2	Test 3	Average
2	Anna	95	93	98	95.3333333333333
3	Bob	57	45	80	60.6666666666667
4	Charles	90	85	92	89
5	David	85	77	63	75

Double-clicking on the formula for David's grade shows how copying a formula changes the spaces it refers to:



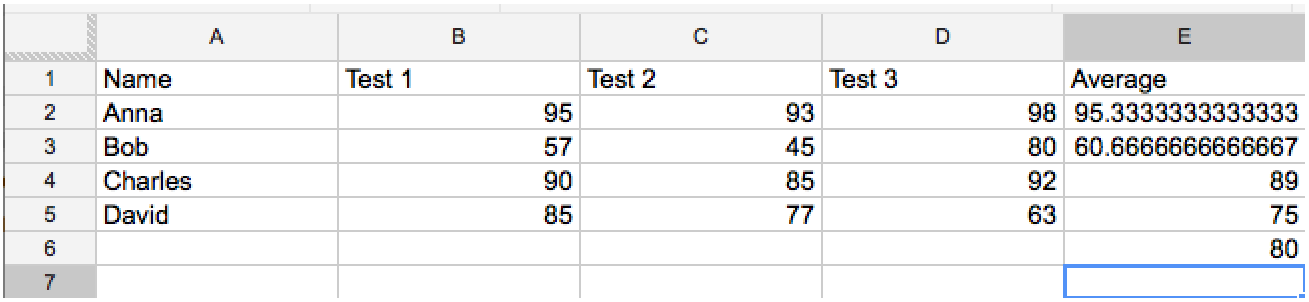
	A	B	C	D	E
1	Name	Test 1	Test 2	Test 3	Average
2	Anna	95	93	98	95.3333333333333
3	Bob	57	45	80	60.6666666666667
4	Charles	90	85	92	89
5	David	85	77	63	=(D5+C5+B5)/3

Now, let's suppose I want to calculate the average grade of everyone in the class. I could use the same kind of formula, but there's an easier way: spreadsheets include a bunch of built-in functions for common things like calculating averages, standard deviations, cosines, square roots, you name it. I take an average of each student's average score like this:



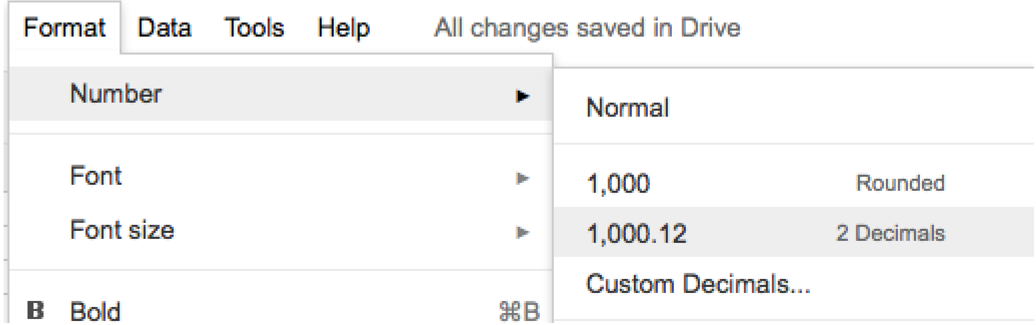
	A	B	C	D	E
1	Name	Test 1	Test 2	Test 3	Average
2	Anna	95	93	98	95.3333333333333
3	Bob	57	45	80	60.6666666666667
4	Charles	90	85	92	89
5	David	85	77	63	75
6					=average(E2:E5)

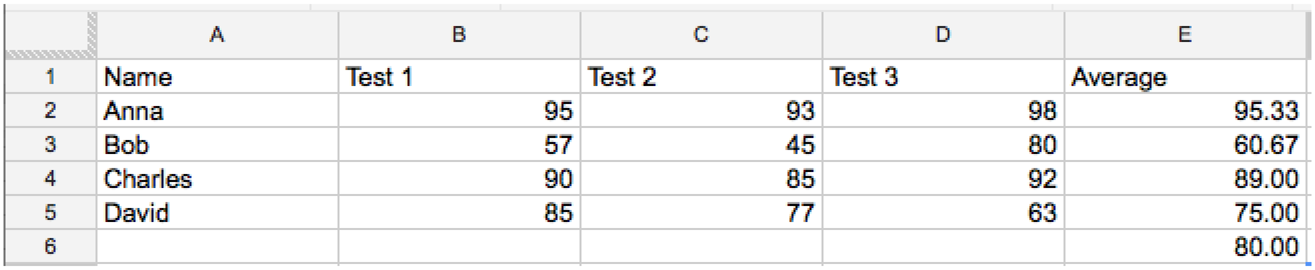
And boom! I'm done.



	A	B	C	D	E
1	Name	Test 1	Test 2	Test 3	Average
2	Anna	95	93	98	95.3333333333333
3	Bob	57	45	80	60.6666666666667
4	Charles	90	85	92	89
5	David	85	77	63	75
6					80
7					

But those repeating decimals are kind of ugly. There's a menu item to change the "number format":

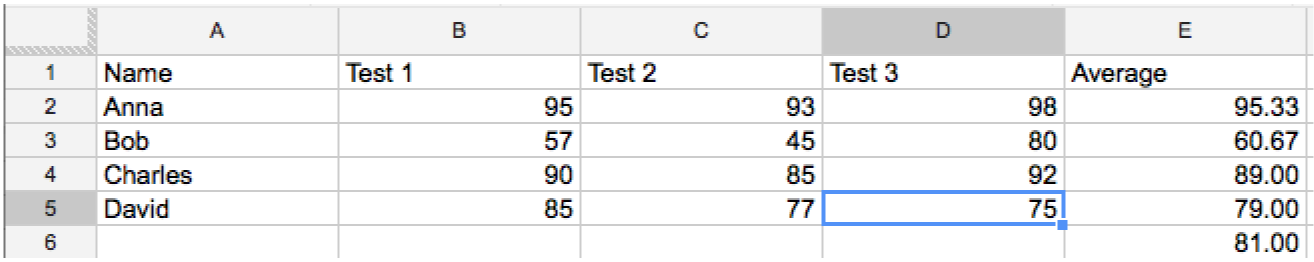




	A	B	C	D	E
1	Name	Test 1	Test 2	Test 3	Average
2	Anna	95	93	98	95.33
3	Bob	57	45	80	60.67
4	Charles	90	85	92	89.00
5	David	85	77	63	75.00
6					80.00

(Notice there's also a scientific notation format, which will come in handy for this class!)

And now, the best part. Suppose David comes into my office to complain about a mistake with my grading, and convinces me he should get a 75 for Test 3. All I have to do is change that one number, and all my formulas automatically get recalculated!



	A	B	C	D	E
1	Name	Test 1	Test 2	Test 3	Average
2	Anna	95	93	98	95.33
3	Bob	57	45	80	60.67
4	Charles	90	85	92	89.00
5	David	85	77	75	79.00
6					81.00

I encourage you to use a spreadsheet for most of the data analysis you do in the E&M lab.

Useful functions

Here are some useful mathematical functions supported by all popular spreadsheet programs. They work the same way as the average() function described above.

average()	Mean of the data
median()	Median of the data
stdev()	Sample standard deviation of the data
count()	Number of elements of data
sum()	Sum of data
sumsq()	Sum of squares of data
sin(), cos(), sqrt()...	Standard mathematical functions