UNCERTAINTY ANALYSIS

FOR PHYSICS 201 & 202

E. M. EDLUND

LAST UPDATED: AUGUST 12, 2019

TOPICS COVERED

Section 1: The normal distribution

Section 2: Mean and standard deviation

Section 3: Definition of uncertainty

Section 4: Making histograms from raw data

Section 5: Fitting a normal distribution to a histogram

Section 6: Calculating a linear fit to data

Section 7: Uncertainty for a quantity composed of multiple independent measurements

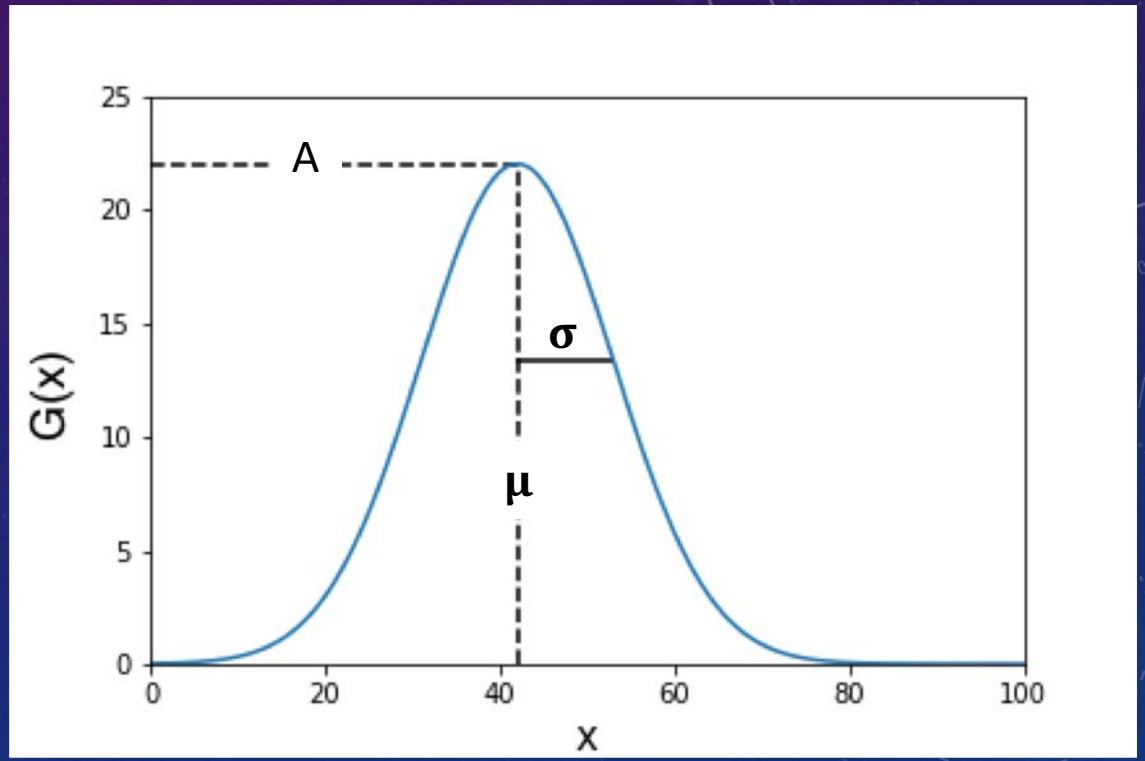
Section 8: When it is acceptable to use an R^2 value

SECTION 1: THE NORMAL DISTRIBUTION

- The normal distribution, also called a Gaussian distribution or sometimes a bell-curve, is an extremely common curve encountered in statistical analysis of random processes.

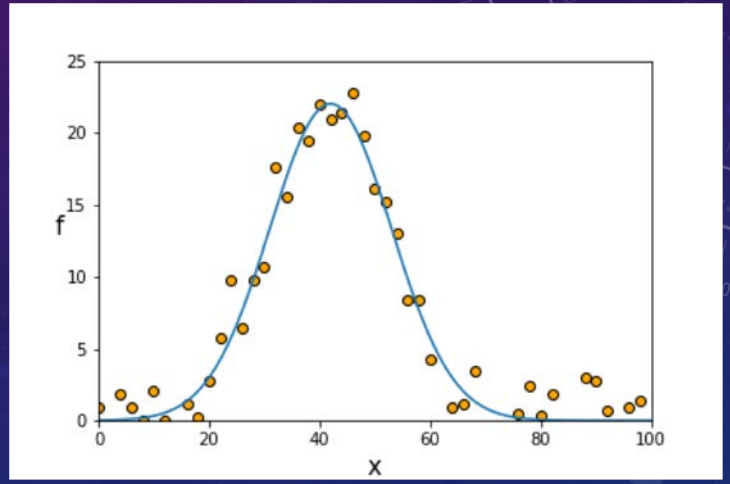
$$\mathrm{G}(x) = A \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

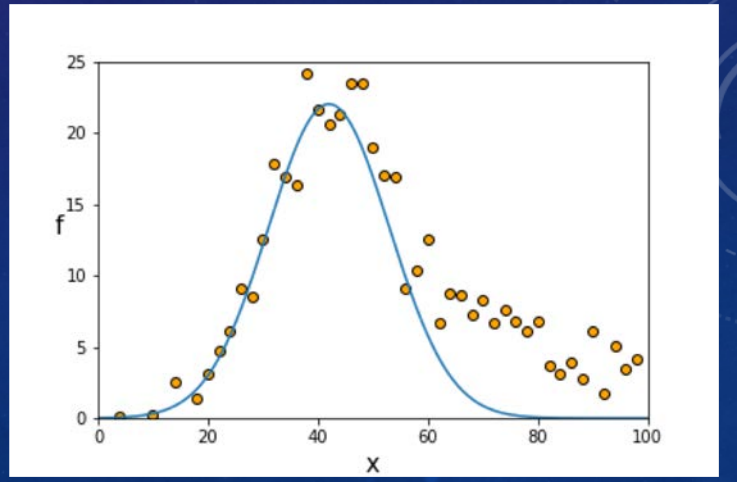
- In this equation the distribution G(x) has amplitude (height) of A, a mean (average value) of $\boldsymbol{\mu}$, and a standard deviation (width) described by $\boldsymbol{\sigma}$.
- What do we mean by random process?
  - An experiment where there is some natural randomness for which we cannot account.
  - No experiment is perfect, so there will always be some ``noise" or randomness, hence all experiments are a random process of a sort.
  - Our task as experimentalists is often to figure out how to separate the underlying trend (the pattern we want) from the random noise that we cannot control.



SECTION 1: THE NORMAL DISTRIBUTION

- It is fairly common to encounter data presented in form of a histogram, a measure of how often a set of measurements occur.
- In the data on the right, the actual values that are recorded in a set of measurements are represented by x (the horizontal axis). The frequency at which those values occurred, which is how many times each value of x appeared in your lab notebook, is represented by f (the vertical axis).
- The blue curves are some kind of fit to the data, something we have to generate, kind of like calculations of $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$.
- In the examples at the right we have two data sets that have peaks.
  - It would seem to be a reasonable conclusion that the behavior of the experiment in the top plot is likely a gaussian process, though we can't say for sure because there is some uncertainty in this analysis.
  - The data in the bottom plot have a region that seems to match with the gaussian distribution, but the long tail to the right of the peak indicates that a gaussian process is not the proper model for this system.





SECTION 2: MEAN AND STANDARD DEVIATION

…. of a single quantity (a list of measurements):

- Imagine that you make a series of measurements of a single quantity, maybe the distance (the range) that an object flies through the air, under the same experimental condtions. Let these values be R_1, R_2, R_3, … R_N.
- We might like to know the mean value and the standard deviation for this collection of measurements.
- If we had an infinite number of measurements we could calculate the true values of the mean and standard deviations for our experimental measurements.
- However, given that we have limited experimental data, the best we can do is to calculate some estimate of the true values of the mean and standard deviation.
- These estimates are given by:

mean value of R:

$$\mu_R = \frac{1}{N}\sum_{i=1}^{N} R_i$$

The mean represents the average value of all measurements.

standard deviation of R:

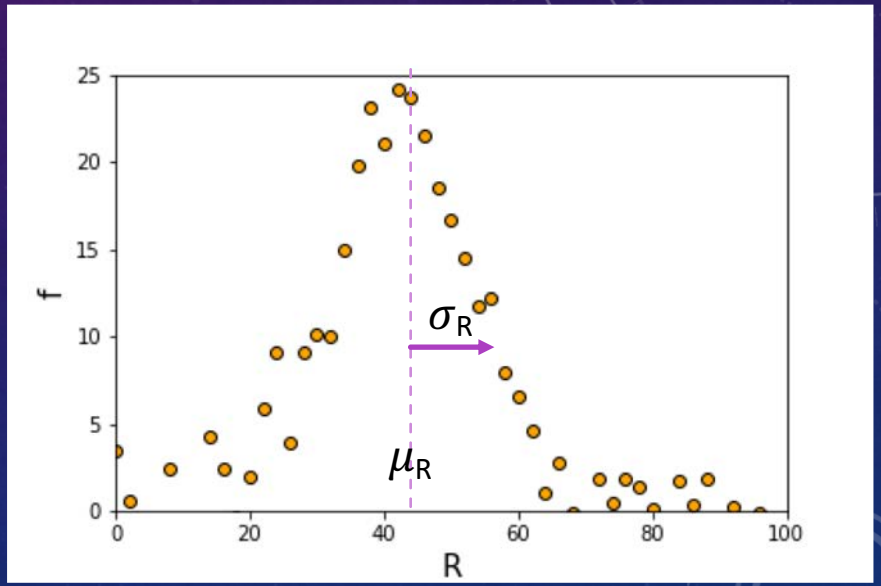
$$\sigma_R = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(R_i - \mu_R)^2}$$

The standard deviation measures the spread in the values around the mean. Small standard deviation (compared to the mean value) indicates that the measurements are tightly clustered around the mean value.

SECTION 2: MEAN AND STANDARD DEVIATION

…. of a distribution function:

- Imagine that you compiled your data for the range measurements (R) of the previous slide into a histogram.
- This is just kind of a condensed version of the prior table, with some values occurring more often than others.
- The prior equations can be transformed to give the following equations:



Total number of measurements: $N = \sum_{i=1}^{k} f_i$

mean value of R: $\mu_R = \frac{1}{N} \sum_{i=1}^{k} f_i \, R_i$

standard deviation of R: $\sigma_R = \sqrt{\frac{1}{N-1} \sum_{i=1}^{k} f_i (R_i - \mu_R)^2}$

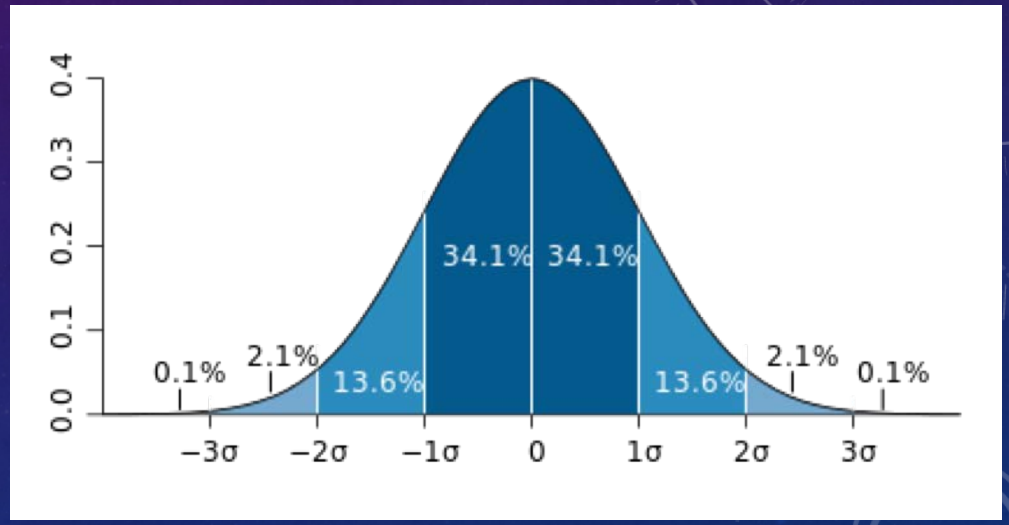
NOTE: If you have created a histogram that faithfully represents your data, the mean and standard deviations from these calculations should be nearly identical to those for the raw data on the previous page.

SECTION 3: DEFINITION OF UNCERTAINTY

- There are many different definitions and standards quantifying uncertainty. Each field of science, or even individual scientists, may have their own preferred methods for these things.
- In the general sense, however, what we want something that captures the notion of the “typical” range of measurements.
- For this course we are going to use the following definition:

  The uncertainty of a quantity R will be designated **Δ**R and will be taken to be equal to

$$\Delta R = 2\sigma_R$$

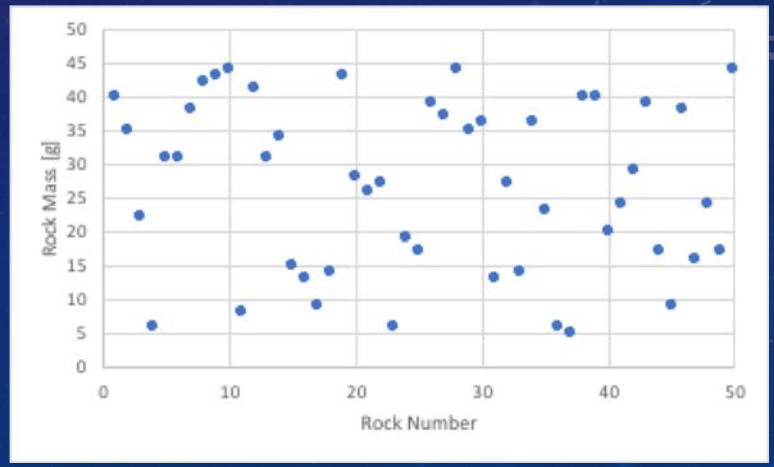


- This definition means that roughly 95% of all of our measurements fall within +/- **Δ**R of the mean value.

SECTION 4: MAKING HISTOGRAMS FROM RAW DATA

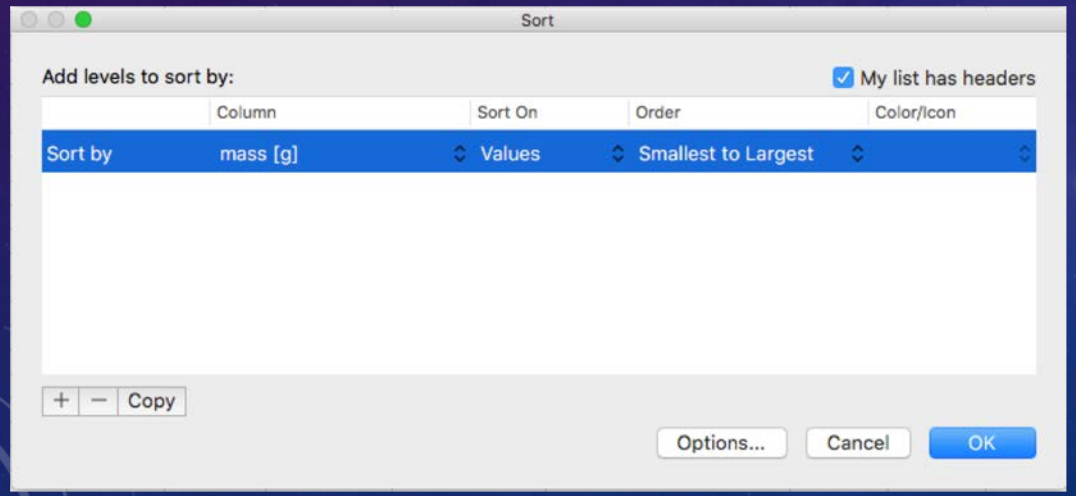
- How would we go about transforming a list of measurements, perhaps the mass of a set of rocks, into something that can be more easily studied by way of a graph?
- Imagine we have 50 different rocks, and we label each rock with a number in the range of 1-50. The order in which we grabbed the rocks and labeled them was random, this this number has no meaning, other than to make sure we don't measure the same rock twice (effectively it is a name).
- Plotting the rock mass versus rock label would give us something that is utterly not helpful. Worse than that, actually, it would give us something that is misleading, because when you present a plot to someone the implication is that there is something interesting to look at there. If there isn't (as there isn't in this case) then DON'T MAKE SUCH A PLOT.

Rock #	mass [g]
1	34
2	5
3	17
4	2
5	21
6	25
7	46
8	13
9	11



SECTION 4: MAKING HISTOGRAMS FROM RAW DATA

- Creating a histogram in Excel:
- First, sort your data from lowest to highest value. You can do this using the sort command after highlighting the data range you want to sort. And if you want to sort the measurement number (the Rock number in our example) with the mass, then choose to sort by the measurement (mass): Sort -> custom sort and select the column you want to sort on.



- Your data should now look something like this:

Rock #	mass [g]
4	2
17	4
2	5
28	5
50	6
20	7
45	8
24	9
31	9
12	9
42	10
39	11
14	11
9	11
43	12
8	13
11	14

SECTION 4: MAKING HISTOGRAMS FROM RAW DATA

- What we want to do next is to organize the data into something that makes sense. We want to group the measurements into reasonable sized “bins”, and then count the number of rocks that have masses in these ranges.
- Maybe we define the bins as follows:

  bin 1: 1 to 2 grams

  bin 2: 3 to 4 grams

  and so on…
- But we could also have done something like:

  bin 1: 1 to 5 grams

  bin 2: 6 to 10 grams

  and so on…
- But how do we choose the best bin size?

SECTION 4: MAKING HISTOGRAMS FROM RAW DATA

- There is no simple answer here, and it really is a matter of trial and error. If you choose a bin size that is too small then your data will look ragged and bumpy and you won't see any significant pattern. If you choose a bin size that is too large then you are going to start averaging out the features of the distribution that you want to see.
- There are many ways that one could estimate a good bin size, which we will call h. Here are a few of those:

method 1:

$$h = \frac{\max(x) - \min(x)}{N^{1/2}}$$

method 2:

$$h = \frac{\max(x) - \min(x)}{\log_2(N) + 1}$$

method 3:

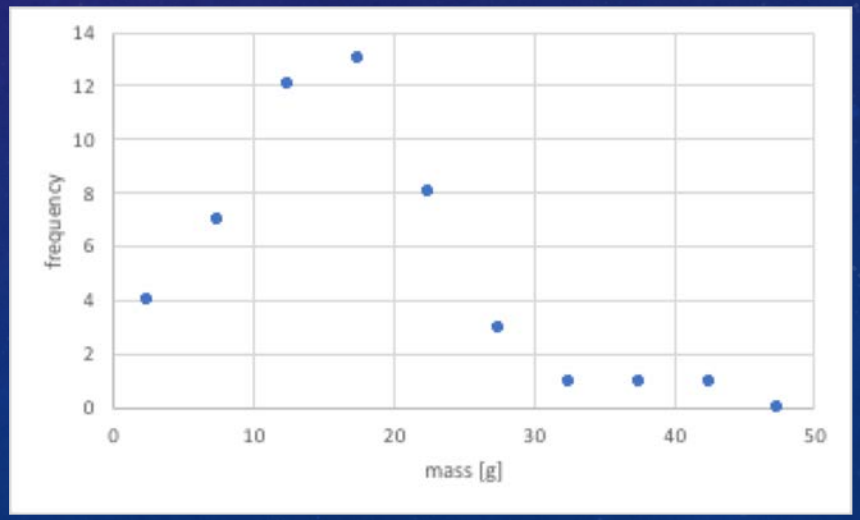
$$h = \frac{2\sigma_x}{N^{1/3}}$$

- Once you make a guess using one of these methods, you first probably want to round your number off to something easy to work with (like 1-5, 6-10, etc.), and the secondly you definitely have to make sure that it makes sense! If it doesn't, then adjust and try again.
- I like method 3. However, the thing that you need to be careful about with method 3 is that it can easily over-estimate the ideal bin size. Because it uses the standard deviation, and the standard deviation is very sensitive to outliers, it can report bin sizes that are too large. If you think you have outliers, remove them from your estimate of $\boldsymbol{\sigma}$ and try calculating h again.

SECTION 4: MAKING HISTOGRAMS FROM RAW DATA

- Calculate an estimate of the bin size (h) using your estimate of the sample standard deviation.
- Define your bin ranges, and then tally the number of measurements that fall within each range.
- The value of x (the rock mass) that is recorded for each bin is the mean value of each bin interval. That means, if we have a bin that runs from 1 to 5, we take (1+5)/2 = 2.5 to get the value for that bin.
- Note that the sum of the frequency values should equal the original number of measurements (a good thing to check).
- Say in the rock examples we use a bin size of grams, this would look like:

group	x	f
1 to 5	2.5	4
6 to 10	7.5	7
11 to 15	12.5	12
16 to 20	17.5	13
21 to 25	22.5	8
26 to 30	27.5	3
31 to 35	32.5	1
36 to 40	37.5	1
41 to 45	42.5	1
46 to 50	47.5	0
	total:	50



SECTION 5: CALCULATING A BEST FIT GAUSSIAN

- Building on the prior example, we will now look at how we find a “best fit normal distribution” for our data that is now represented by a histogram. Recall the definition of the Gaussian distribution:

$$\mathrm{G}(x) = A\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- Microsoft Excel does not offer the option to fit a normal distribution to our data.
- We get around this issue by recognizing that the natural logarithm of the normal distribution gives us a quadratic function, which is something we can fit in Excel (being a second order polynomial):

$$\ln(G) = \ln(A) - \frac{(x-\mu)^2}{2\sigma^2}$$

$$= -\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x + \left[\ln(A) - \frac{\mu^2}{2\sigma^2}\right]$$

$$= c_1x^2 + c_2x + c_3$$

$$c_1 = -\frac{1}{2\sigma^2}$$

$$c_2 = \frac{\mu}{\sigma^2}$$

$$c_3 = \ln(A) - \frac{\mu^2}{2\sigma^2}$$

SECTION 5: CALCULATING A BEST FIT GAUSSIAN

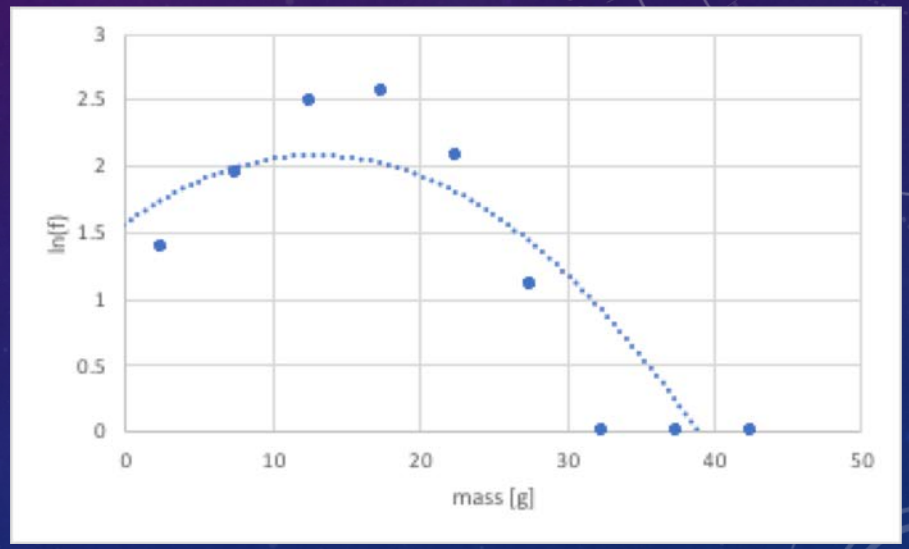
- Our hypothesis for the moment is that our data may be represented by a Gaussian. Therefore, we need to put it in a form that allows us to compare against the Gaussian distribution.
- Because, as described on the last page, we need to plot the natural logarithm of the Gaussian distribution (so that we can do this in Excel), we also need to take the natural logarithm of our frequency data.
- Note that we might get some errors if there are bins for which we have zero counts. This is a problem because ln(0) = -∞, and an infinity is never a good thing to have in your experimental data.
- We need to review our ln(f) table and simply remove any values that fail in this way. We must remove them because Excel interprets any error like this as having a value of 0, which throws-off the fit.

group	x	f	ln(f)
1 to 5	2.5	4	1.38629436
6 to 10	7.5	7	1.94591015
11 to 15	12.5	12	2.48490665
16 to 20	17.5	13	2.56494936
21 to 25	22.5	8	2.07944154
26 to 30	27.5	3	1.09861229
31 to 35	32.5	1	0
36 to 40	37.5	1	0
41 to 45	42.5	1	0
46 to 50	47.5	0	#NUM!
	total:	50	

group	x	f	ln(f)
1 to 5	2.5	4	1.38629436
6 to 10	7.5	7	1.94591015
11 to 15	12.5	12	2.48490665
16 to 20	17.5	13	2.56494936
21 to 25	22.5	8	2.07944154
26 to 30	27.5	3	1.09861229
31 to 35	32.5	1	0
36 to 40	37.5	1	0
41 to 45	42.5	1	0
46 to 50	47.5	0	
	total:	50	

SECTION 5: CALCULATING A BEST FIT GAUSSIAN

- Throw your data onto a plot and add a trendline by right-clicking on the plot and selecting “Add trendline.” Since the function we want to fit is a quadratic, we select “polynomial” and set the value of the order to 2 (which defines a parabola).
- The ln(f) vs. x plot looks similar to the f vs. x plot in overall shape (it should), though the vertical values have changed.
- Note however, that the fit we get here doesn’t look great. Is that the best we can do?
- The problem is that those three points that have low amplitude at the bottom-right of the plot are throwing off the fit. The fitting function wants to do the best job possible to fit all of the data, and doesn’t care that these points are possibly bad. We want to focus on the peak where we have greater faith that the measurements are good and not just a result of noise in the system. Always be wary of small values!



- We need throw out the points that have very low values and apply to fit to places where we believe our data are good and reliable.

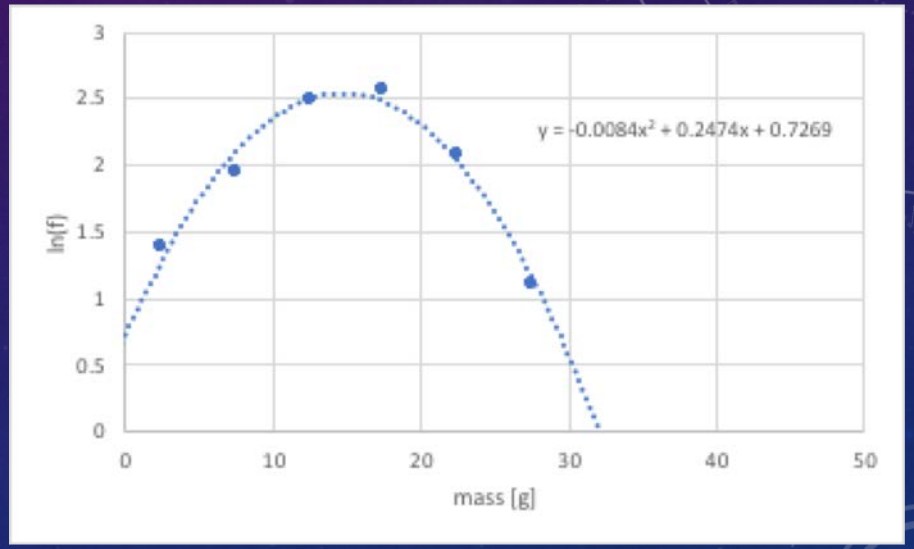
SECTION 5: CALCULATING A BEST FIT GAUSSIAN

- Try again by selecting a narrower range of your histogram data, eliminating the problematic points at the extremes.
- That looks like a decent fit, at least to the region where we think we have good data.
- From here, we can solve for the values that define our normal distribution.
- The equation displayed on the chart gives us the coefficients for the quadratic (parabola) fit to the data.
- Recall:

$$c_1 = -\frac{1}{2\sigma^2}$$

$$c_2 = \frac{\mu}{\sigma^2}$$

$$c_3 = \ln(A) - \frac{\mu^2}{2\sigma^2}$$



- Solving first for **σ** using c_1 we have:

$$\sigma = \sqrt{-\frac{1}{2c_1}} = \sqrt{-\frac{1}{2(-0.0084)}} = 7.7$$

SECTION 5: CALCULATING A BEST FIT GAUSSIAN

- Continuing to c_2 we can solve for **μ**:

$$\mu = c_2\sigma^2 = (0.247)(7.7^2) = 14.6$$

- Lastly, we can use c_3 to solve for the amplitude A:

$$A = \exp\left(c_3 + \frac{\mu^2}{2\sigma^2}\right) = \exp\left(0.727 + \frac{14.6^2}{2\times 7.7^2}\right)$$

$$A = 12.5$$

- So that our best-fit gaussian function is defined by:

$$\mathrm{G}(x) = 12.5\,\exp\left(-\frac{(x-14.6)^2}{118.6}\right)$$

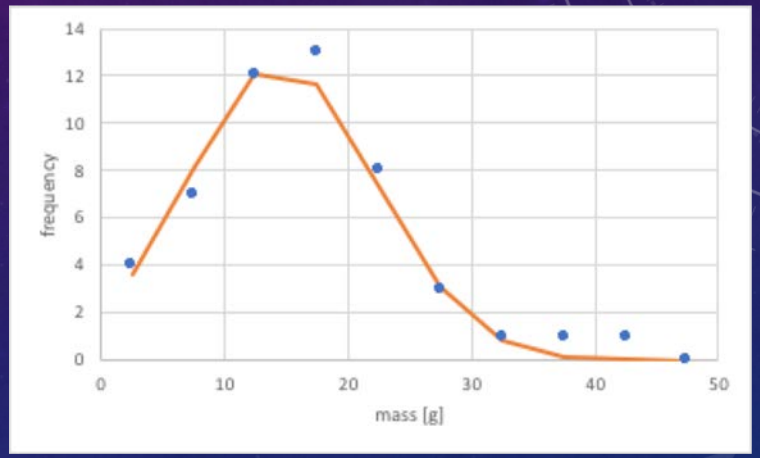
Note: you might need to include a -1 in your formula because Microsoft is run by clowns.

	A. =	12.5		
	mu =	14.6		
	sig =	7.7		
group	x	f	ln(f)	G(x)
1 to 5	2.5	4	1.38629436	3.63654759
6 to 10	7.5	7	1.94591015	8.17119199
11 to 15	12.5	12	2.48490665	12.0436622
16 to 20	17.5	13	2.56494936	11.6441751
21 to 25	22.5	8	2.07944154	7.3847517
26 to 30	27.5	3	1.09861229	3.07213347
31 to 35	32.5	1	0	0.83834209
36 to 40	37.5	1	0	0.15006502
41 to 45	42.5	1	0	0.01762036
46 to 50	47.5	0		0.00135715

- The values for the normal distribution function is entered as an equation, where we use the $ signs to reference the values in the definition of A, **μ** and **σ** that are static.

SECTION 5: CALCULATING A BEST FIT GAUSSIAN

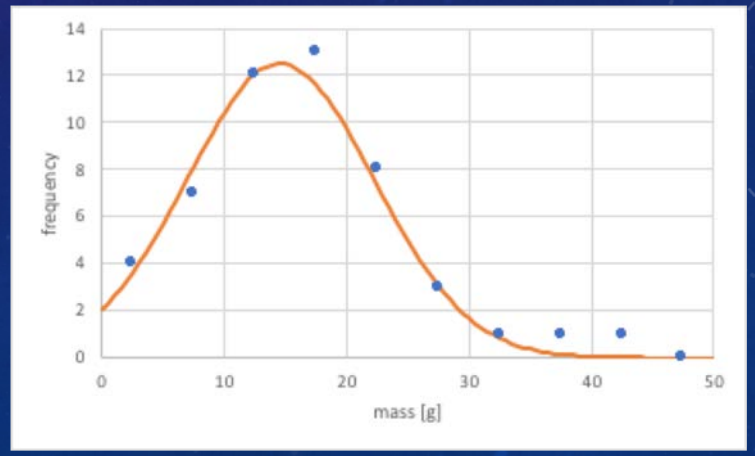
- Finally, we can plot the original histogram and the best-fit normal distribution together.
- That looks ok. Not great. Our normal distribution looks a little chunky and it would be nice if it were smoother.
- Since the N(x) we are plotting is something of a theoretical function, we can define a new table with a much smaller step in x and a define N(x) for each of those values:



x	G(x)
0.0	2.07119068
0.5	2.33763328
1.0	2.6272503
1.5	2.94032467
2.0	3.27686007
2.5	3.63654759
3.0	4.0187355
3.5	4.42240327
4.0	4.84614083

And that looks nice!

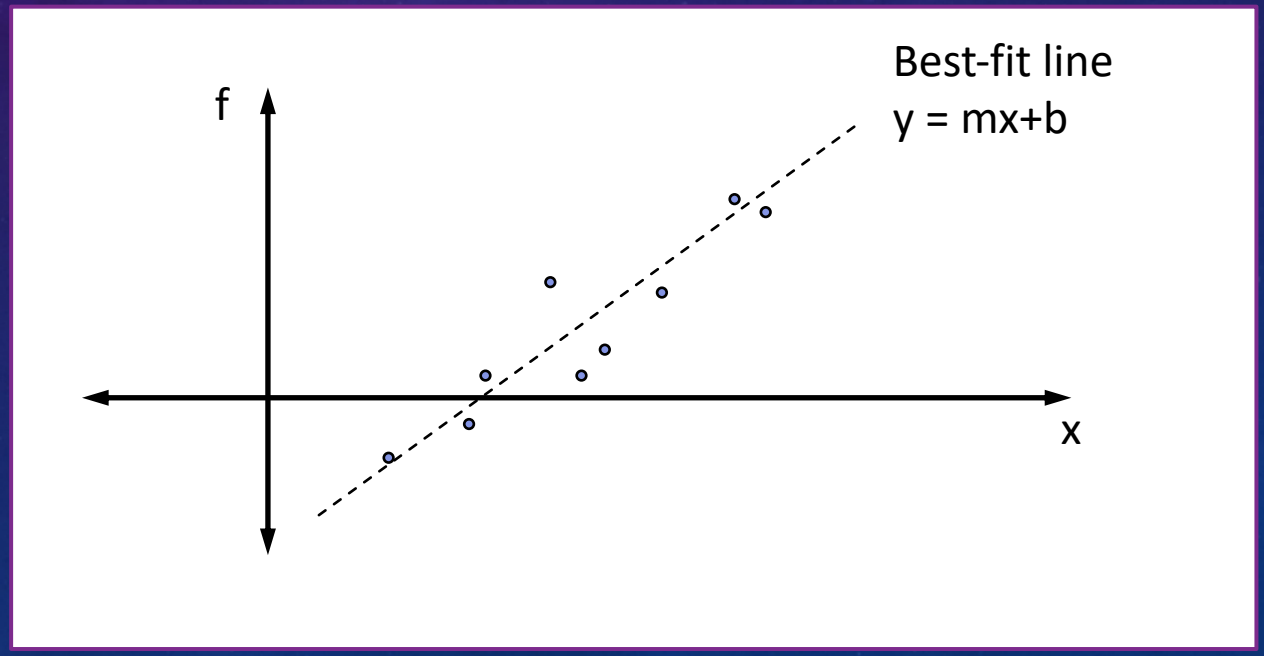
At least as nice as we can expect from Excel. There are, of course, better tools to use, but Excel is a fine place to start.



SECTION 6: CALCULATING A LINEAR "BEST-FIT" LINE

- If we have a set of measurements of a single independent variable x and a dependent variable f, we might like to plot this.
- If we expect that the relationship between f and x should be linear, then we might like to know what equation of a line is the best representation of this relationship. But because all data has noise and uncertainty in it, our data will never fall on a perfect line, so we need a statistical method for determining the best fit line.
- Consider a series of measurements of some quantity x (the independent variable) and another quantity f (the dependent variable). Imagine you now have N measurements of these quantities. Arranged in a table and graph the data look like:

x	f
x_1	f_1
x_2	f_2
⋮	⋮
x_N	f_N



SECTION 6: EQUATIONS FOR THE SLOPE AND OFFSET

$$m = \frac{\sum f_i x_i - \frac{1}{N}\left(\sum f_i\right)\left(\sum x_i\right)}{\sum x_i^2 - \frac{1}{N}\left(\sum x_i\right)^2}$$

$$b = \frac{1}{N}\,\frac{\left(\sum f_i\right)\left(\sum x_i^2\right) - \left(\sum f_i x_i\right)\left(\sum x_i\right)}{\sum x_i^2 - \frac{1}{N}\left(\sum x_i\right)^2}$$

SECTION 6: IMPLEMENTING THE EQUATIONS IN EXCEL

- If we had 21 measurements of x and f, our spreadsheet might look like the following: (including the relevant formulas to do these calculations)

	A	B	C	D	E	F
1		Exp. #	x	f	x^2	f * x
2		1	x_1	f_1	=C2*C2	=C2*D2
3		2	x_2	f_2	=C3*C3	=C3*D3
⋮		⋮	⋮	⋮	⋮	⋮
22		21	x_{21}	f_{21}	=C22*C22	=C22*D22
23	sum	=COUNT(B2:B22)	=SUM(C2:C22)	=SUM(D2:D22)	=SUM(E2:E22)	=SUM(F2:F22)
24						
25					m:	= (F23-D23*C23/B23)/(E23 – C23^2/B23)
26					b:	=(1/B23) (D23*E23-F23*C23)/(E23 – C23^2/B23)

SECTION 7: UNCERTAINTY OF A QUANTITY COMPOSED OF MULTIPLE INDEPENDENT MEASUREMENTS

- Consider now that we want to know the uncertainty in a quantity that is composed of several different measurements. Let us consider a quantity E (an energy) composed of some different measured quantities, say h (the height), m (mass), g (gravitational acceleration) and θ (launch angle). This means that E can be expressed as some function of these quantities.

$$E = E(h, m, g, \theta)$$

- We will assume that we know the theoretical form of this function. If we don't have a theoretical model to work from then we would need to use other methods that we will not go into here.
- Let us consider a totally fabricated example. Suppose we have done some theoretical analysis and have reason to suspect that the energy we are looking for should be given by the following expression:

$$E = mgh\sin(\theta)$$

- The general idea is that this new quantity, E, inherits uncertainty from all of the separate values from which it is composed and we need some way of representing this composite uncertainty.

SECTION 7: CONTINUED

- Let us assume that we know the uncertainties in h, m, g and θ, either because we previously calculated them or this number was provided to us (perhaps the manufacturer of the apparatus told us what a typical uncertainty in θ is).

- Let us also take as granted that these quantities and their uncertainties are independent. That is, if we change the angle, for example, this does not change the value of h, m or g.

- Given these uncertainties ($\boldsymbol{\Delta}$h, $\boldsymbol{\Delta}$m, $\boldsymbol{\Delta}$g and $\boldsymbol{\Delta}\theta$) we calculate $\boldsymbol{\Delta}$E as follows:

$$\Delta E = \sqrt{\left(\frac{dE}{dh}\Delta h\right)^2 + \left(\frac{dE}{dm}\Delta m\right)^2 + \left(\frac{dE}{dg}\Delta g\right)^2 + \left(\frac{dE}{d\theta}\Delta\theta\right)^2}$$

SECTION 7: CONTINUED

- Clearly, to continue we need the derivatives of E with respect to the independent variables. But, because we have a theoretical model to work with, we can calculate these.

$$\frac{dE}{dh} = mg\sin(\theta) = \frac{E}{h} \qquad \frac{dE}{dg} = mh\sin(\theta) = \frac{E}{g}$$

$$\frac{dE}{dm} = gh\sin(\theta) = \frac{E}{m} \qquad \frac{dE}{d\theta} = mgh\cos(\theta) = \frac{E}{\tan(\theta)}$$

- Let us look at the first term under the square root:

$$\left(\frac{dE}{dh}\Delta h\right)^2 = \left(\frac{E}{h}\Delta h\right)^2 = E^2\left(\frac{\Delta h}{h}\right)^2$$

SECTION 7: CONTINUED

- A similar result applies to the other terms, and we notice that in each of these has a factor of E^2, which can be removed from the square root to get:

$$\Delta E = E\sqrt{\left(\frac{\Delta h}{h}\right)^2 + \left(\frac{\Delta m}{m}\right)^2 + \left(\frac{\Delta g}{g}\right)^2 + \left(\frac{\Delta \theta}{\tan(\theta)}\right)^2}$$

- Where we see the quantities E, h, m, g, and θ on the RHS of the prior equation we should think that we want the mean values of those quantities. That is:

$$\Delta E = \bar{E}\sqrt{\left(\frac{\Delta h}{\bar{h}}\right)^2 + \left(\frac{\Delta m}{\bar{m}}\right)^2 + \left(\frac{\Delta g}{\bar{g}}\right)^2 + \left(\frac{\Delta \theta}{\overline{\tan(\theta)}}\right)^2}$$

- Notice that the value in the bottom of the last term is the average value of $\tan(\theta)$, which is **NOT** the same thing as the tangent of the average value of θ.

SECTION 7: CONTINUED

- To illustrate this last point consider the following set of measurements of the angle θ:

Measurement #	θ(radians)	tan(θ)
1	0.70	0.84
2	0.90	1.26
3	1.00	1.56
4	1.10	1.96
5	1.30	3.60
average value:	1.00	1.85
	tan(1.0)=.707	

- This is something of a subtle point and something to be aware of. The best thing to remember is that when calculating the uncertainties of a composite quantity, we always want to consider using the mean values of the functions, not the mean values of the independent variables.

SECTION 7: CONTINUED

- Notice as well that for each of the terms under the square root is a dimensionless number. Thus, if one of those terms is much smaller than the others it can be ignored. Consider as an example the following data:

average h [meters]	average m [kg]	average g [m/s^2]	average tan(θ) [dimensionless]		average E
2.0	20.0	9.8	1.8		323
Δh [meters]	**Δ**m [kg]	**Δ**g [m/s^2]	**Δ**θ [radians]		**Δ**E [J]
0.1	0.5	0.1	0.2		40.3

- In equation form we have the following:

$$\Delta E = 323 \text{ Joules} \sqrt{\left(\frac{0.1}{2.0}\right)^2 + \left(\frac{0.5}{20.0}\right)^2 + \left(\frac{0.1}{9.8}\right)^2 + \left(\frac{0.2}{1.8}\right)^2}$$

$$\approx 323 \text{ Joules} \sqrt{(0.05)^2 + (0.025)^2 + (0.01)^2 + (0.11)^2} = 40.3 \text{ Joules}$$

SECTION 7: CONTINUED

- The reason we are going to such lengths to investigate each of these terms is that they tell us something very important about where most of our uncertainty comes from.

- Notice that the third term under the square root is very small, 0.001^2 to be exact, which is considerably smaller than the other terms. What happens if we ignore that term? We would get the following:

$$\Delta E = 323 \text{ Joules} \sqrt{\left(\frac{0.1}{2.0}\right)^2 + \left(\frac{0.5}{20.0}\right)^2 + \left(\frac{0.2}{1.8}\right)^2}$$

$$\approx 323 \text{ Joules} \sqrt{(0.05)^2 + (0.025)^2 + (0.11)^2} = 40.2 \text{ Joules}$$

- We get essentially the same answer! This means that our uncertainty in the gravitational acceleration is so small that it really doesn't affect the answer, or at least not in a significant way. We can effectively take this number to be as good as perfect, as if it doesn't have any uncertainty.

SECTION 7: CONTINUED

- What about the other terms?
- Let's next eliminate the second term under the square root (the next smallest one), associated with the uncertainty in m. Doing so, we get:

$$\Delta E = 323 \text{ Joules} \sqrt{\left(\frac{0.1}{2.0}\right)^2 + \left(\frac{0.2}{1.8}\right)^2}$$

$$\approx 323 \text{ Joules} \sqrt{(0.05)^2 + (0.11)^2} = 39.4 \text{ Joules}$$

- Still not much change. And if we eliminate the first, the one associated with the uncertainty in h we get:

$$\Delta E = 323 \text{ Joules} \sqrt{\left(\frac{0.2}{1.8}\right)^2}$$

$$\approx 323 \text{ Joules} \sqrt{(0.11)^2} = 35.9 \text{ Joules}$$

SECTION 7: CONTINUED

- This tells us that the far majority of the uncertainty comes from the uncertainty in the angle.
- This might be important to know because if we were an engineer at a company and we needed to reduce the uncertainty in our measurements of the performance of some mechanical engine, and had a limited budget to do so, we might like to know where we should put our dollars I buying better equipment.
- In the case at hand we should put our money toward a launcher with less uncertainty in the angle.
- The case we have been looking at was a fairly complicated analysis because of the nasty $\sin(\theta)$ term that made things difficult. In the following we will investigate a simplifying case.

SECTION 7: CONTINUED

- In the prior example we started with the statement:

$$E = E(h, m, g, \theta)$$

- To move forward, we will first generalize our case in a more formal mathematical language. To do so we will consider the stuff on the LHS of the equation to be the independent variable (the f), and all the stuff on the RHS to be the function of the independent variables (the x's).
- That is, if we let f=E and x_1=h, x_2=m, x_3=g, and x_4=θ then we have the statement:

$$f = f(x_1, x_2, x_3, x_4)$$

- And our calculation of the uncertainty would look like:

$$\Delta f = \sqrt{\left(\frac{df}{dx_1}\Delta x_1\right)^2 + \left(\frac{df}{dx_2}\Delta x_2\right)^2 + \left(\frac{df}{dx_3}\Delta x_3\right)^2 + \left(\frac{df}{dx_4}\Delta x_4\right)^2}$$

SECTION 7: CONTINUED

- Generalizing even further, to the case where instead of only 4 independent variables we might have N independent variables, this would look like:

$$f = f(x_1, x_2, \dots, x_N)$$

- With the associated uncertainty calculation:

$$\Delta f = \sqrt{\left(\frac{df}{dx_1}\Delta x_1\right)^2 + \left(\frac{df}{dx_2}\Delta x_2\right)^2 + \cdots + \left(\frac{df}{dx_N}\Delta x_N\right)^2}$$

- Now, it is often, though certainly not always, the case that our function y may be a simpler function of the independent variables, without any nasty sines or cosines or other functions and when this is true we can use a simpler form for the uncertainty analysis.

SECTION 7: CONTINUED

- If our function y is a simple product of the independent variables, each raised to some power

$$f = c\, x_1^{p_1}\, x_2^{p_2} \ldots\, x_N^{p_N}$$

- then, it follows that the derivatives of y with respect to each of its variables is quite simple. For example the derivative with respect to x_1 would be

$$\frac{df}{dx_1} = c\, p_1 x_1^{p_1-1}\, x_2^{p_2} \ldots\, x_N^{p_N}$$

$$= \frac{c\, p_1 x_1^{p_1}\, x_2^{p_2} \ldots\, x_N^{p_N}}{x_1}$$

$$= \frac{p_1 f}{x_1}$$

SECTION 7: CONTINUED

- A similar result follows for all of the other x's since they have the same basic form in the expression.
- In this way, the expression for the uncertainty becomes:

$$\Delta f = \sqrt{\left(\frac{p_1 \bar{f}}{\bar{x}_1}\Delta x_1\right)^2 + \left(\frac{p_2 f}{\bar{x}_2}\Delta x_2\right)^2 + \ldots + \left(\frac{p_N \bar{f}}{\bar{x}_N}\Delta x_N\right)^2}$$

$$= \bar{f}\sqrt{\left(p_1\frac{\Delta x_1}{\bar{x}_1}\right)^2 + \left(p_2\frac{\Delta x_2}{\bar{x}_2}\right)^2 + \ldots + \left(p_N\frac{\Delta x_N}{\bar{x}_N}\right)^2}$$

- Or, as a fractional uncertainty we have:

$$\frac{\Delta f}{\bar{f}} = \sqrt{\left(p_1\frac{\Delta x_1}{\bar{x}_1}\right)^2 + \left(p_2\frac{\Delta x_2}{\bar{x}_2}\right)^2 + \ldots + \left(p_N\frac{\Delta x_N}{\bar{x}_N}\right)^2}$$

SECTION 7: CONTINUED

- If for example, we had some expression we wanted to consider:

$$E = 2.8\frac{mh^2}{t^2}$$

- We could rewrite this using y=E, x_1=m, x_2=h and x_3=t with p_1=1, p_2=2 and p_3=-2 t get:

$$f = 2.8\ x_1^1\ x_2^2\ x_3^{-2}$$

- From which we would get:

$$\frac{\Delta f}{\bar{f}} = \sqrt{\left(1\frac{\Delta x_1}{\bar{x}_1}\right)^2 + \left(2\frac{\Delta x_2}{\bar{x}_2}\right)^2 + \left(-2\frac{\Delta x_3}{\bar{x}_3}\right)^2}$$

$$= \sqrt{\left(\frac{\Delta x_1}{\bar{x}_1}\right)^2 + 4\left(\frac{\Delta x_2}{\bar{x}_2}\right)^2 + 4\left(\frac{\Delta x_3}{\bar{x}_3}\right)^2}$$

SECTION 7: CONTINUED

- This last equation tells us then how to calculate the uncertainty for y (which was just another name for E, the energy, for which we want the uncertainty). Transforming back to the original variable names we have:

$$\frac{\Delta E}{\bar{E}} = \sqrt{\left(\frac{\Delta m}{\bar{m}}\right)^2 + 4\left(\frac{\Delta h}{\bar{h}}\right)^2 + 4\left(\frac{\Delta t}{\bar{t}}\right)^2}$$

- What this means is that if all of the independent variables had the same uncertainty, say 0.010, we have:

$$\frac{\Delta E}{\bar{E}} = \sqrt{(0.010)^2 + 4(0.010)^2 + 4(0.010)^2} = \sqrt{9(0.010)^2} = 0.030$$

- But if we ignore the first term then we have

$$\frac{\Delta E}{\bar{E}} \approx \sqrt{4(0.010)^2 + 4(0.010)^2} = \sqrt{8(0.010)^2} = 0.028$$

- Which reduces the fractional uncertainty in E by only about 7%. So, despite having the same fractional uncertainty, ignoring the uncertainty in the mass has only a small influence on the final answer because of its smaller weighting due to the smaller exponent.

SECTION 8: R^2 VALUES

DISCLAIMER:

- The use of the R^2 value is exceedingly over-used and, usually, incorrectly applied.

- Even when use of R^2 values is justified its interpretation is non-trivial, as explained in the subsequent slides.

- In short: an R^2 value provides us with some measure of how well our data is explained by some model. We must interpret the the meaning of R^2 in this way, and somewhere in there, either explicitly or implicitly, is a comparison of two models or two different experiments – a single R^2 value by itself is meaningless.

SECTION 8: THE MEANING OF R^2

- An R^2 value provides us with a measure of how well a particular model explains the variance in our data.
- This can be interpreted in three ways:
  - Experiment validation: If we have reason to believe that a phenomena must behave in a certain way, or that our measurements are constrained by something extremely robust (like the law of conservation of energy), then the R^2 value derived from comparison of our data to the model can be interpreted, in a sense, as a measure of the quality of our data or the quality of our experimental apparatus. That is, we are taking the theory (or perhaps prior measurements conducted on fancy equipment) as the standard. A low R^2 might indicate that we either have a poor experimental procedure or equipment, or that we overlooked an important energy loss mechanism (e.g. radiation).
  - Predictive capability: If we have data that we believe to be good but include perhaps multiple sources of noise for which we are unable to account, the R^2 value tells us how well our model explains the data, despite the presence of noise. What constitutes a "good" R^2 value is highly dependent on the particular subject of study and the tools being used. For one type of measurement with a lot of inherent complexity and natural variability (e.g. the weather), a model with an R^2 vale of 0.2 might be considered extremely good, in that it offers at least some predictive capability, whereas in modeling growth of crystals in a laboratory where the experimental standards are generally very high (e.g. precision temperature control), an R^2 of 0.90 might be considered fairly bad.
  - Model comparison: If there are multiple theoretical models that could be used to explain our data, then a comparison of the R^2 values, one for each model, is an indicator, though not necessarily the final answer, as to which model may be the better descriptor of the phenomenon we are studying. For example, three different theoretical physicists create three different models describing the loss of ice cream on a hot summer day. Each one included a different effect in their model: eating the ice cream, dripping, and evaporation. These different physical processes lead to different mathematical models of the rate of ice cream loss: a linear function, a hyperbolic tangent function and a decaying exponential function. The R^2 vale would tell us which one of these models, in a way and with caveats, captures the most truth, in the sense of containing the most relevant physical mechanisms. We might have good reason to suspect that the eating model might be the most accurate model, and this then should be reflected in that model having the highest R^2 value of the three.

SECTION 8: RULES FOR WHEN TO USE AN R^2 VALUE

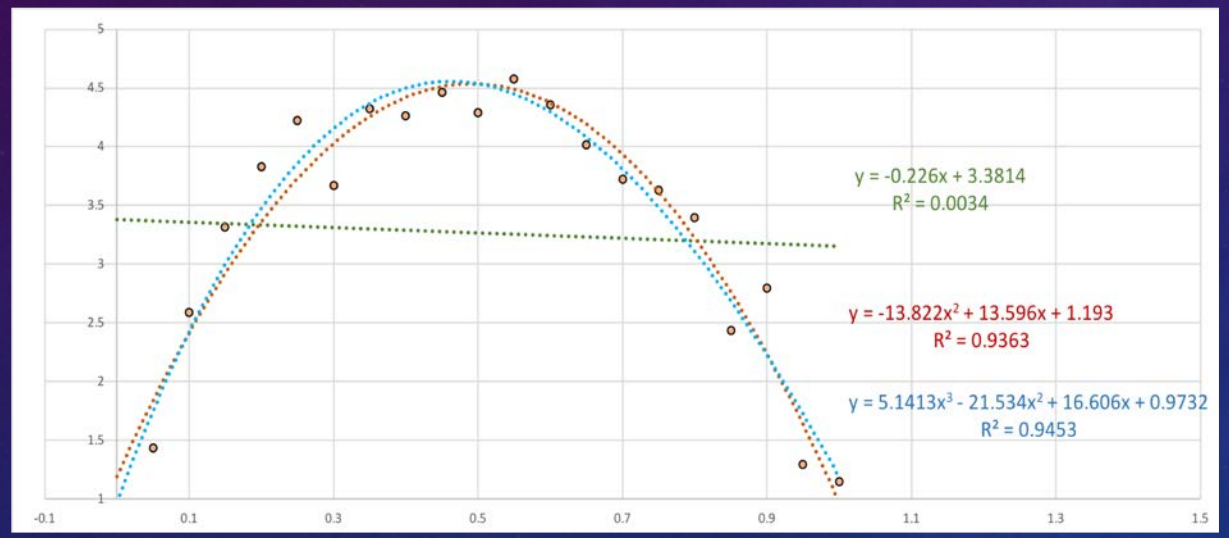
- Rule #1: The meaning of R^2 is somewhat subtle and often requires some additional knowledge about what constitutes a good or bad value of R^2. That additional knowledge could come from prior experiments, or accrued experience about what is typical of your measurements, or comparison to other models within your experiment. If there is a question about whether you should present it, you are most likely better off leaving it out of the analysis.

- Rule #2: An R^2 value should never be presented on a graph unless you can explain why you have it there.

- Rule #3: To assess whether it should be used, see if your intention in using it is congruent with one of the three reasons listed on the prior page.

SECTION 8: R^2 AND MODEL COMPARISON

- As an example, consider the following problem: we have a data set and want to determine whether a linear, parabolic or cubic model represents it best. Comparing the R^2 values for these different models tell us which one we should use.



- The linear model is clearly terrible, and the cubic model has the highest R^2 value. Therefore the cubic model is the best, right?
- Actually no. The parabolic and cubic functions are nearly indistinguishable. The additional term in the cubic is equation is just fitting he noise better, not the underlying trend. We are deceiving ourselves. Why stop at a 3rd order polynomial? Clearly a 4th order, or a 20th order polynomial would be better, right?
- No again. In this case the parabolic fit is the best. It takes more reasoning to show why. The point is to illustrate that this stuff is non-intuitive even when it seems straightforward.