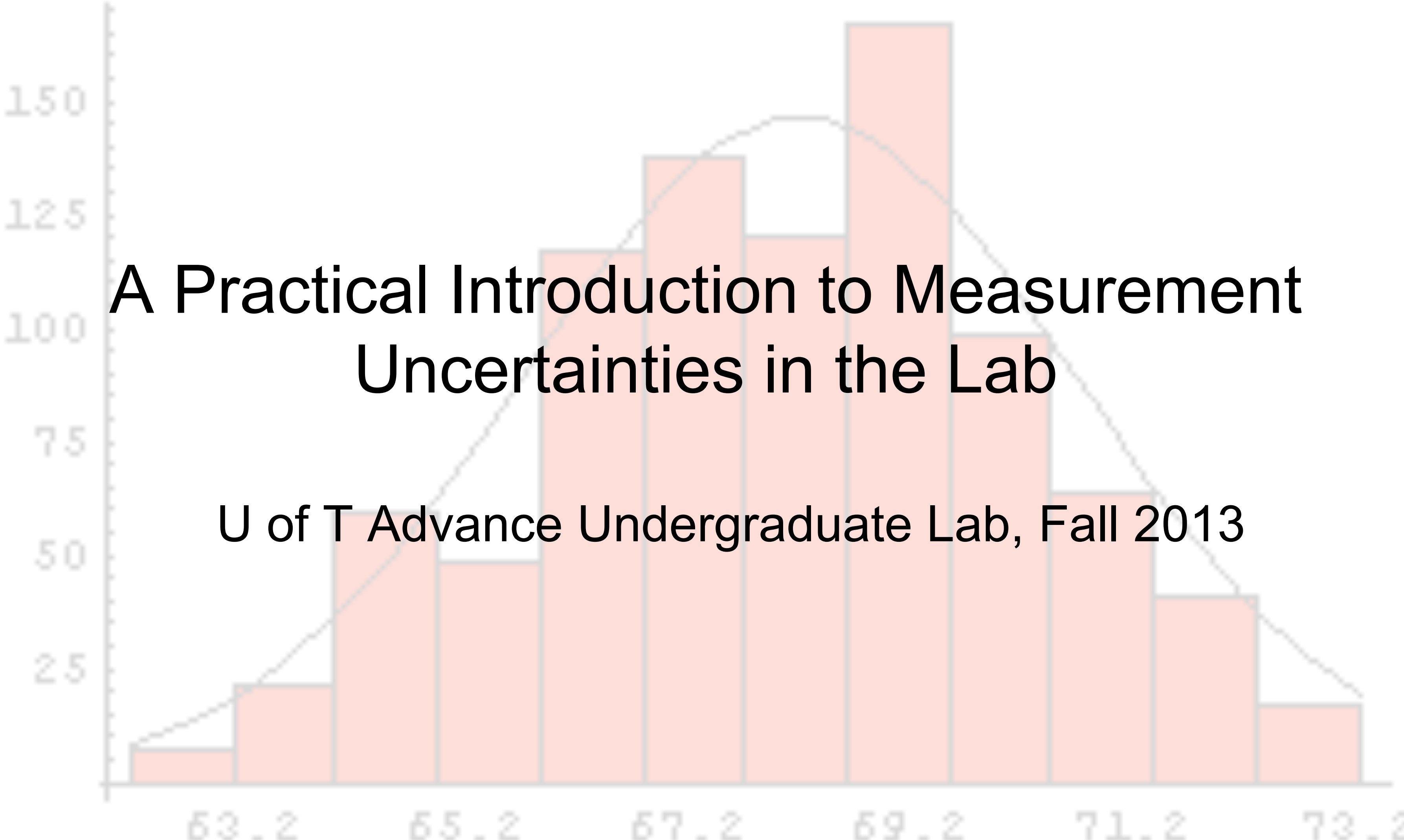
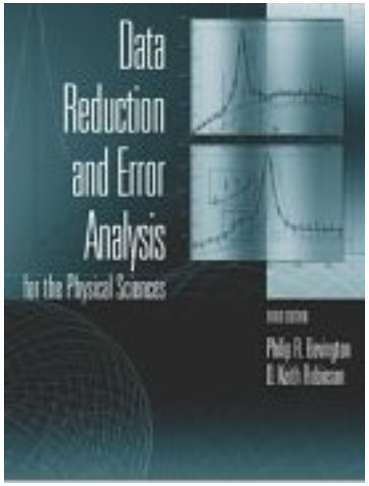
A Practical Introduction to Measurement Uncertainties in the Lab

U of T Advance Undergraduate Lab, Fall 2013



Advanced Lab Data Analysis References



http://www.physics.utoronto.ca/~phy326/links

Data Reduction and Error Analysis for the Physical Sciences. P.R.Bevington & D.K. Robinson (B&R). Available in the U of T Bookstore.

Introductory Lecture from David Bailey (Jan 2013)

Introductory Lecture from Prof. J. Thywissen (Sep 2007)

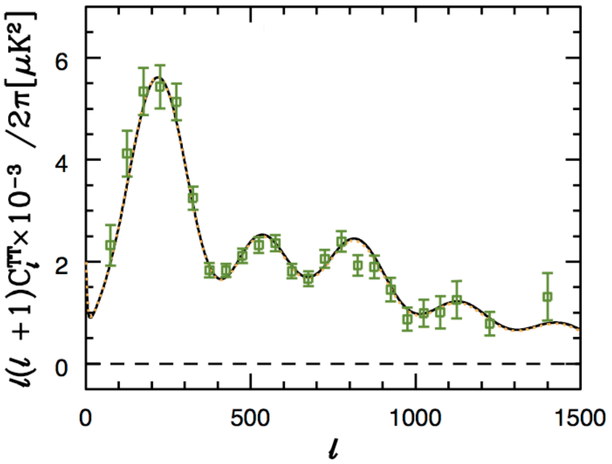
http://www.upscale.utoronto.ca/PVB/Harrison/ErrorAnalysis/

This last reference is a useful brief introduction to a number of the concepts to be discussed today, with nice examples and simulations. I encourage you all to take a couple of hours to look through this.

What's the Point?

Physics students should be able to

- make physical measurements of a system (including error estimates)
- properly calculate derived quantities
- construct mathematical model of a physical system
- solve model analytically or computationally
- compare the measurements with the expectations
- improve model, calculations, or experiment and iterate
- communicate results with others



All of these things involve developing an understanding of the issues surrounding measurement uncertainties, probability and statistics, and fit methods. I will discuss mainly the first of these topics today, with only rather brief discussions of the latter two.

Measurements & Uncertainties

In the physical sciences, a measurement is not just a number. A measured value must also contain an estimate of how close that measured value is expected to be to the “true” value.

If one or more measured values are used in the calculation of some quantity, the errors must be “propagated” to obtain the error on the final quantity (see B&R, chapter 3). For uncertain values of a and b we have (assuming that the errors are uncorrelated):

- For the sum $(a \pm \delta a) + (b \pm \delta b)$ $\quad (a+b) \pm \sqrt{(\delta a)^2 + (\delta b)^2}$
- For the product ab, $\quad \dfrac{\delta(ab)}{ab} = \sqrt{\left(\dfrac{\delta a}{a}\right)^2 + \left(\dfrac{\delta b}{b}\right)^2}$

In general, how errors are propagated depends on the type of error (statistical or systematic) and also on whether any of the errors are correlated (as is often the case for systematic uncertainties).

Start by discussing what we actually mean by the term “error”.

What we ***don’t*** mean is a mistake.

Uncertainty, Accuracy and Precision

Any error (regardless of the source, other than a actual mistake) is actually better referred to as an uncertainty, though the term "error" is in common usage.

Precision refers to the size of the uncertainty on a measurement. The absolute precision on a measurement of quantity x is simply the total uncertainty δx. The relative precision is $\delta x/x$.

Accuracy refers to how close the result is to the "true" value.

Example: ratio of circumference of a circle to its diameter.

• π=3.14159265358979323846264 3… (theory)	
• 3±1	accurate, not precise
• 2.93846±0.00001	precise, not accurate
• 3.14159±0.00001	precise and accurate
• 15±3	neither

Why Make Repeated Measurements?

We saw before that if we add two uncertain values a and b, the uncertainty on the sum (assuming uncorrelated δa and δb) is given by

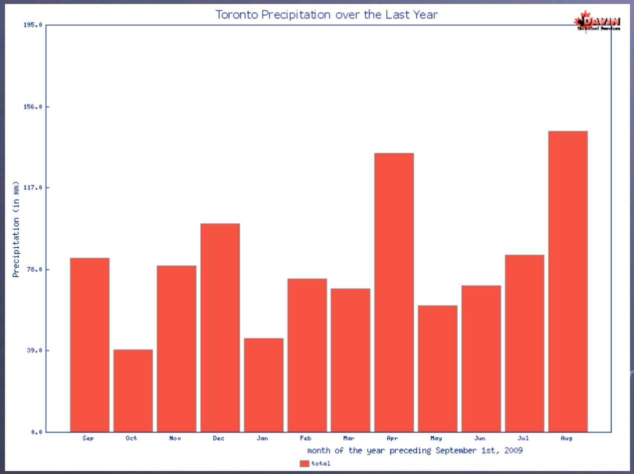
$$\delta(a+b) = \sqrt{(\delta a)^2 + (\delta b)^2}$$

If we want to measure some quantity x, and can make n independent measurements each with (uncorrelated) error δx then the average is given by

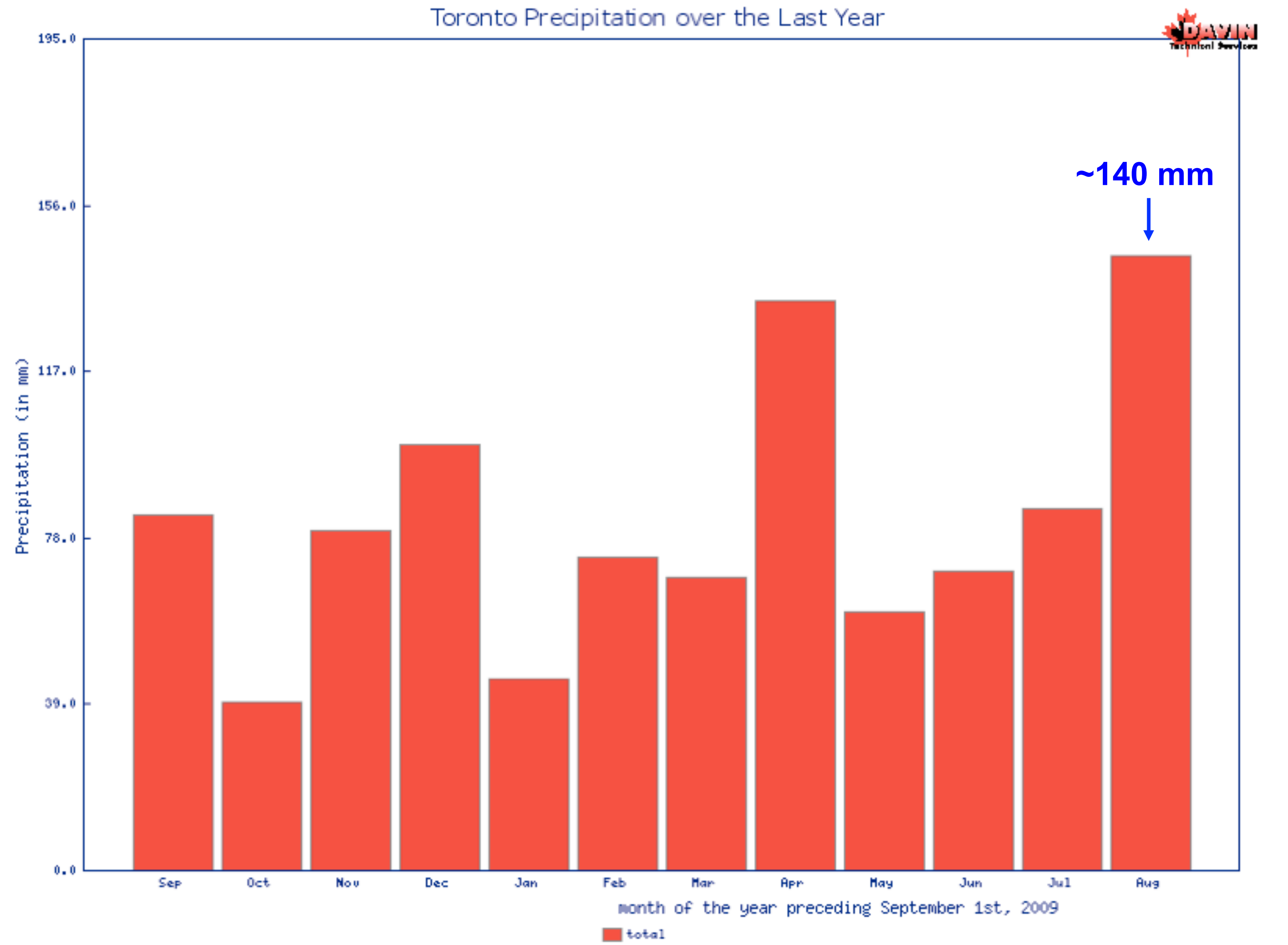
$$\bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$$

and the uncertainty is given by $\delta\bar{x} \equiv \frac{1}{n}\sqrt{\sum_{i=1}^{n}(\delta x_i)^2} = \frac{\delta x}{\sqrt{n}}$.

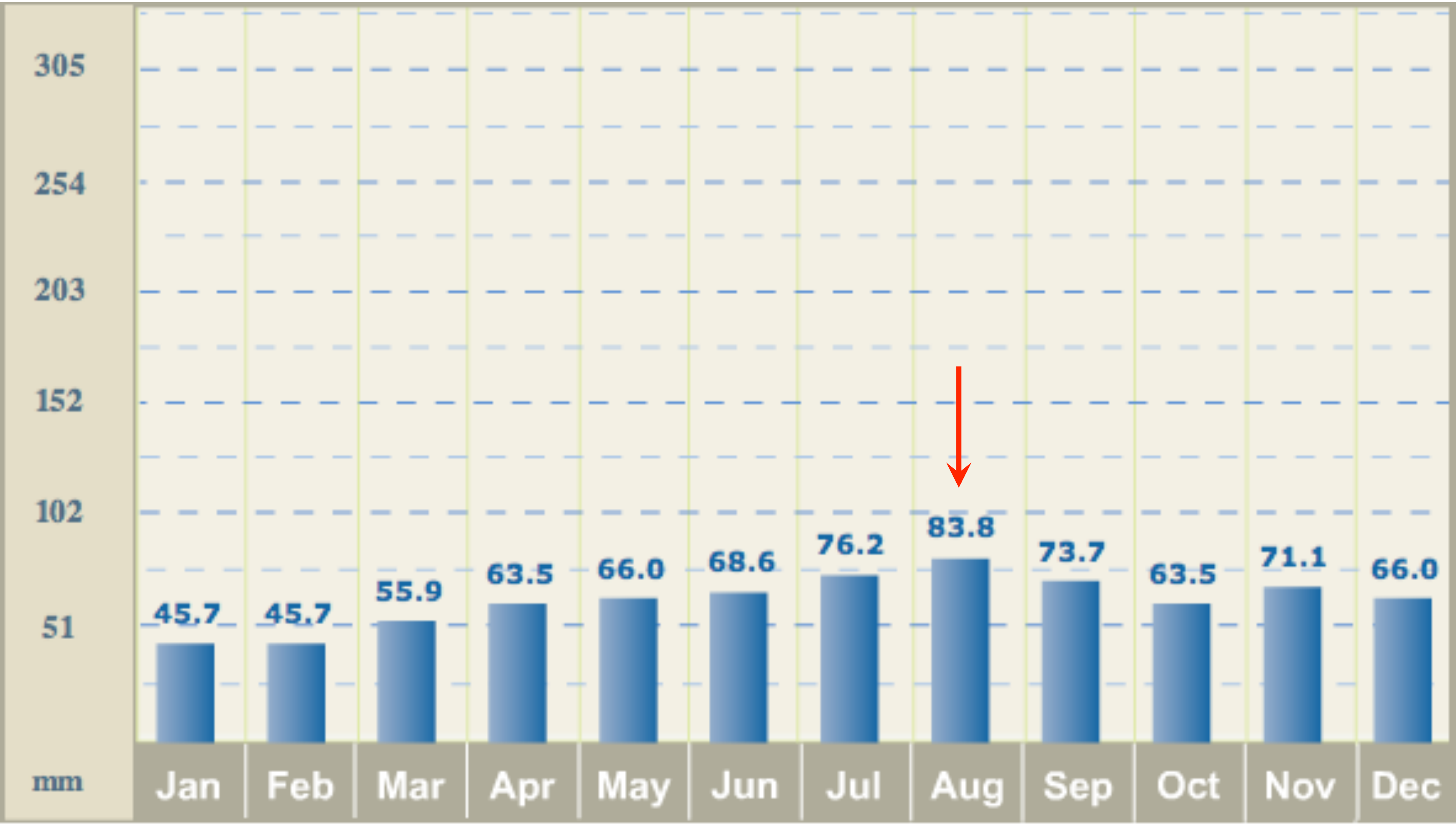
So repeated measurements improve the precision we can achieve on a measurement of x (provided that the errors are random, *e.g.* uncorrelated). We will discuss which type of errors this applies to, and also how many repeated measurements are enough.



Toronto, Summer 2009


Toronto Precipitation over the Last Year
195.0
156.0
117.0
78.0
39.0
0.0
Precipitation (in mm)
~140 mm
Sep
Oct
Nov
Dec
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
month of the year preceding September 1st, 2009
total

Toronto Average Rainfall Statistics



Toronto Average Rainfall Statistics

or, version 2

Precipitation

	J	F	M	A	M	J	J	A	S	O	N	D
Rain (mm)	23	26	40	57	67	68	69	80	76	60	67	45
Snow (cm)	29	25	19	6	0	0	0	0	0	0	5	27
Total (mm)	51	50	60	63	67	68	69	80	76	61	71	73
Snow Depth(cm)	7	4	0	0	0	0	0	0	0	0	1	3

There's a lesson to be learned here. If you are calculating a result that relies on an input that you have not measured yourself you may want to check multiple sources. If they report different values try to understand why. If you can't resolve the discrepancy, you can take the variation as a systematic uncertainty (more on systematics later on).

So, average August rainfall in Toronto is about 80mm.

In August 2009 we had about 140mm. How unusual is this?

Asked another way, if the average is 80mm, how likely is it that the rainfall in any given August will be ≥ 140mm?

The answer is (e): it's impossible to say. Any distribution (such as the distribution of rainfall in Toronto in August) is characterized not only by a mean (or average) but also by a quantity that is a measure of how broad the distribution is. To answer the above question you need to know both.

Distributions: Means and Widths

A given distribution might be characterized by a number of parameters, but we will be mainly concerned with only two of these: the mean (or average) and the variance (which is a measure of the width).

Consider a set of N measurements x_i (for example, the height of thirty-year old men). The average height is given by:

$$\bar{x} \equiv \frac{1}{N}\sum_{i=1}^{N} x_i$$

and a measure of the width, the variance σ^2, is given by the expression

$$\sigma^2 \equiv \frac{1}{N-1}\sum_{i=1}^{N} (x_i - \bar{x})^2$$

The smaller the variance, the more closely the measurements are clustered, and the more likely the probability that additional measurements will lie "close" to the mean.

We often take σ as the uncertainty on a single measurement. If the mean August rainfall in Toronto is 80mm with a σ of 20mm, then we were really unlucky in 2009. If, however, σ = 60mm, then 140mm would not be so unusual.

Probability Distributions: Binomial Distribution

Q: If you flip a coin 20 times, how many times will it come up heads?

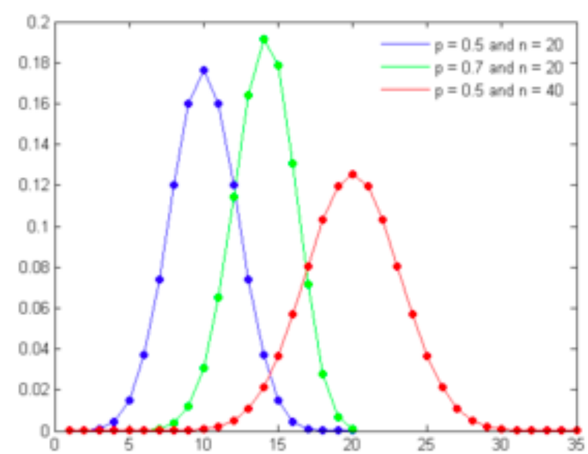
A: You can't say. You can only quote a probability for each possible answer (within the physical bounds of 0-20).

The probability p of a single flip coming up heads is $p=1/2$, so you expect (if you repeat this experiment many times) that the mean value is 10, but what about the width of the distribution (e.g. how often will the answer be 20)?

Can also consider the case where $p \neq 1/2$ for each event, *e.g.* you can roll a dice 20 times and ask how many time does a 6 appear: $p=1/6$.

In each case the answer is governed by the binomial distribution.

$$P_B(x;n,p) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$



The mean and variance of this distribution can be show to be

$$\mu = np \quad \text{and} \quad \sigma^2 = np(1-p)$$

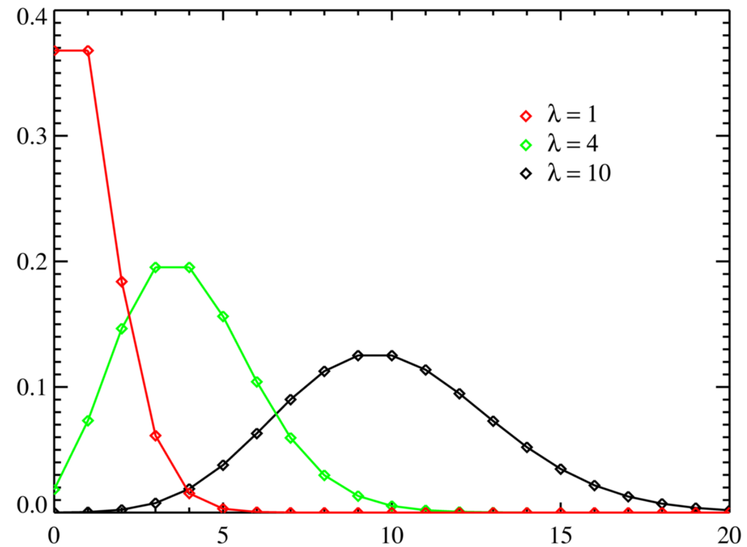
Probability Distributions: Poisson Distribution

The Poisson distribution is an approximation to the binomial distribution in the case where $N \to \infty, p \to 0$ with $np = \lambda$ held constant. In practice this means counting experiments with low statistics (for example the number of decays of some radioactive isotope in some time interval).

$$\lim_{\substack{p \to 0 \\ n \to \infty \\ np = \lambda,\, fixed}} P_B(x; n, p) = P_P(x; \lambda) \equiv \frac{\lambda^x}{x!} e^{-\lambda}$$

$$\langle x \rangle = \sum_{x=0}^{\infty} x \frac{\lambda^x}{x!} e^{-\lambda} = \lambda e^{-\lambda} \sum_{x=1}^{\infty} x \frac{\lambda^{x-1}}{(x-1)!} = \lambda e^{-\lambda} \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} = \lambda$$

$$\sigma^2 = \langle (x-\lambda)^2 \rangle = \sum_{x=1}^{\infty} \left[(x-\lambda)^2 \frac{\lambda^x}{x!} e^{-\lambda} \right] = \lambda$$



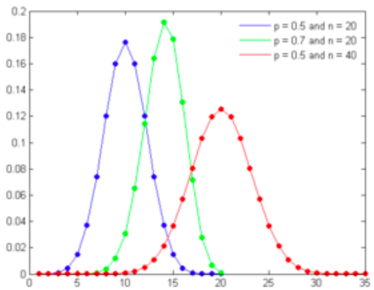
Mean and variance are both = λ

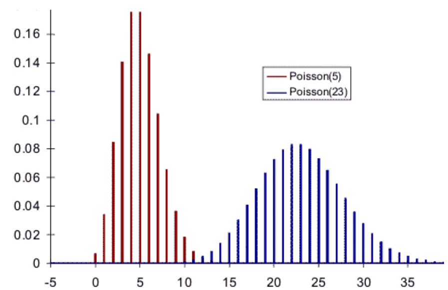
This is the basis of the well known statistical uncertainty associated with counting experiments: $\delta N = \sqrt{N}$.

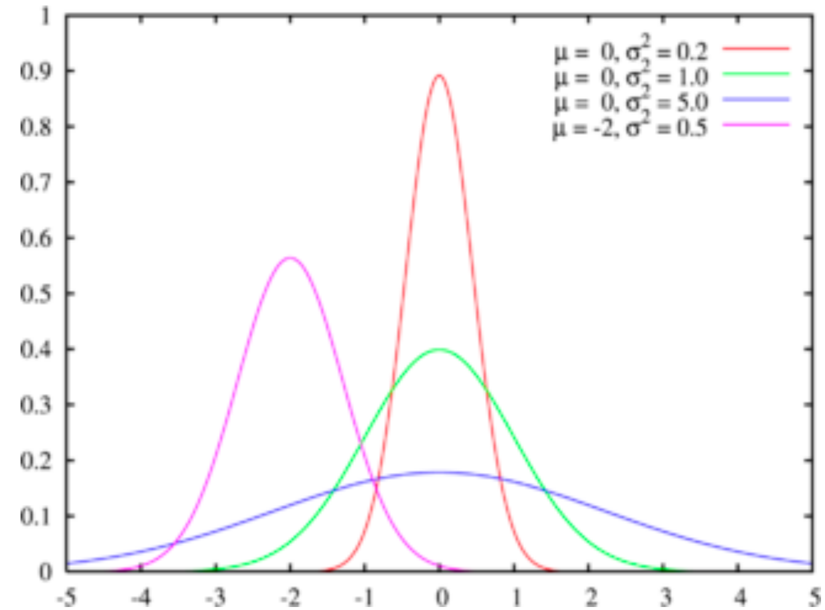
Normal (Gaussian) Distribution

Limit of Binomial/Poisson distributions for large *N* with mean not near zero.

$$P_G(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{(x-\mu)^2}{2\sigma^2}\right)$$

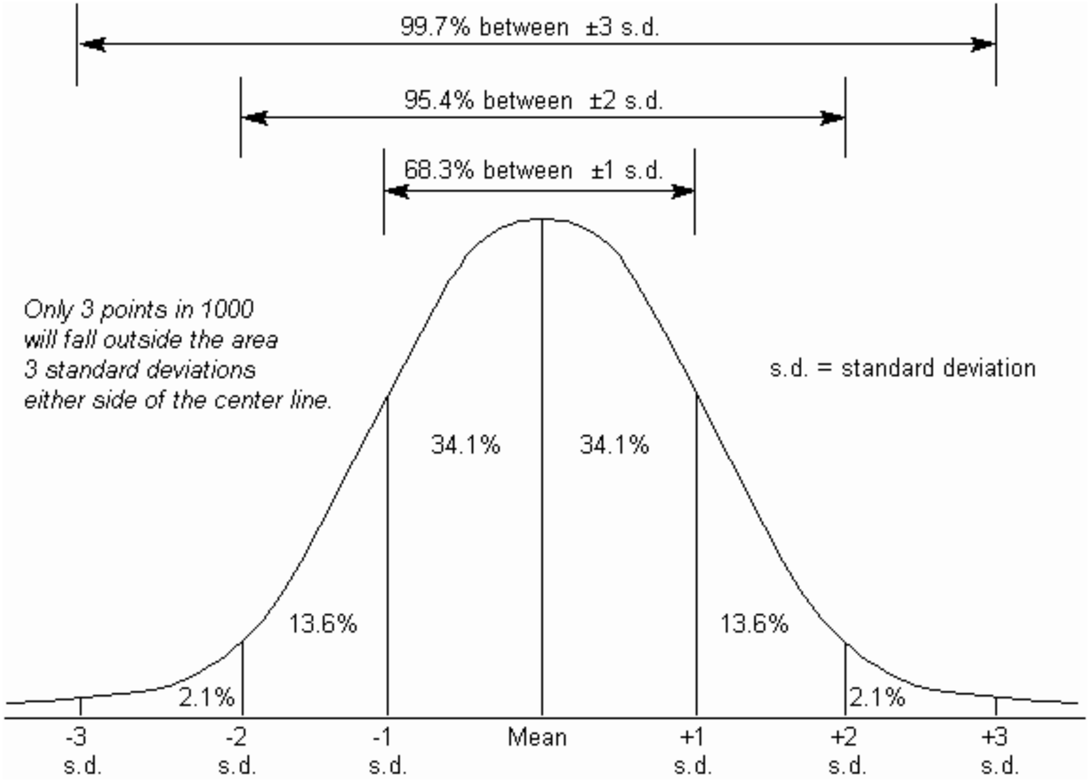






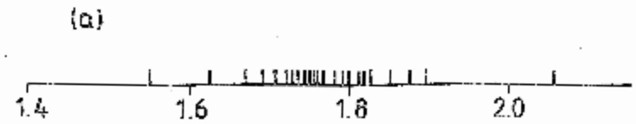
- The Central Limit Theorem says (almost) everything averages to a Gaussian.
- Many resolution functions are at least approximately Gaussian - a blob with a mean and a width (variance). Uncertainties typically treated as Gaussian. Sometimes need to revert to Poisson or Binomial (for low statistics counting experiments).
- It is the only distribution many physicists know anything about.

Gaussian Errors



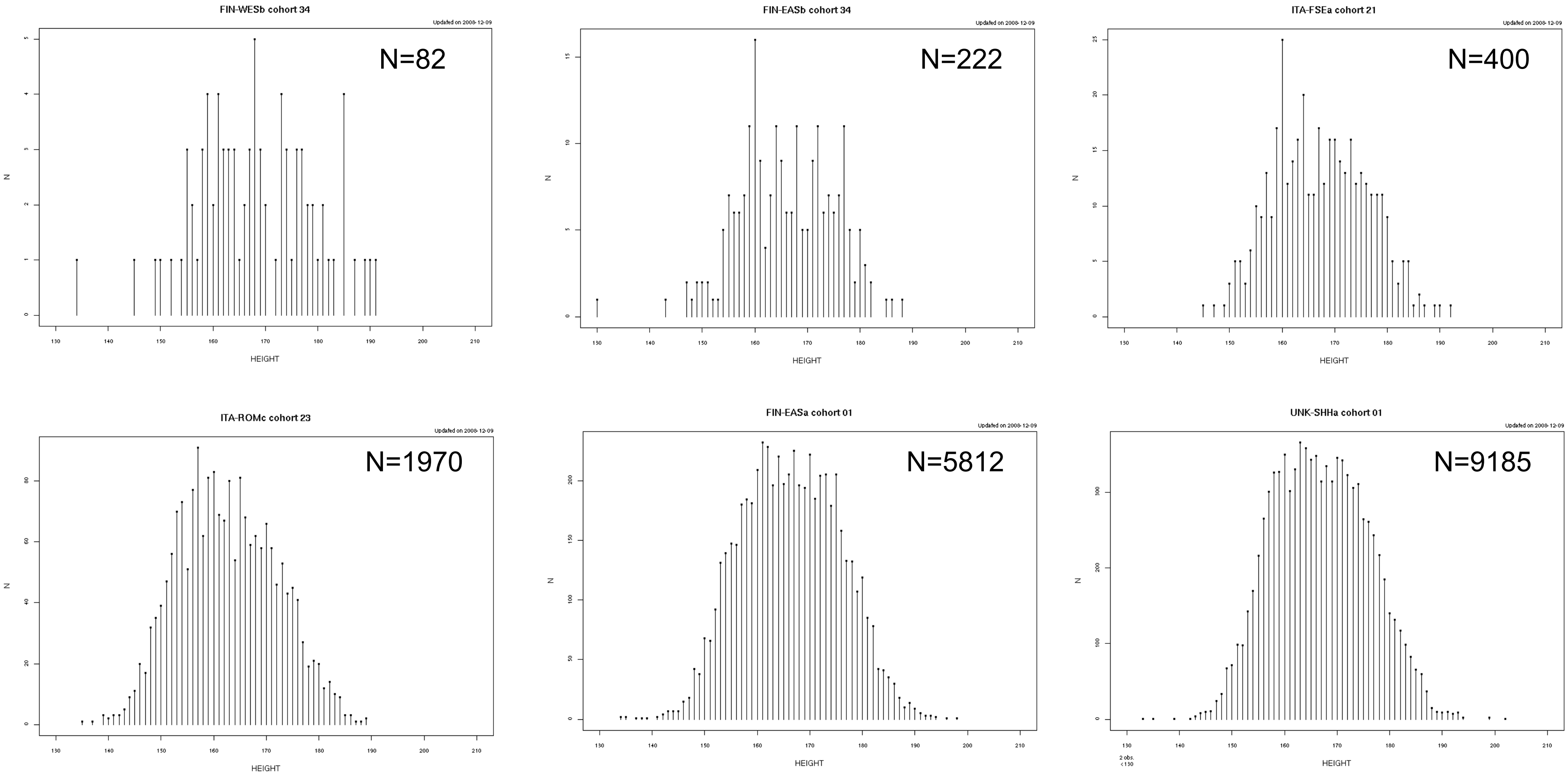
For Gaussian distributed uncertainties, expect that a given measurement should fall within ±1σ of the "true" value 68.3% of the time, within ±2σ 95% of the time, within ±3σ 99.7% of the time…….usually quote 1σ errors.

This is the "nineteen times out of twenty" that one hears quoted when poll results are reported in the news.



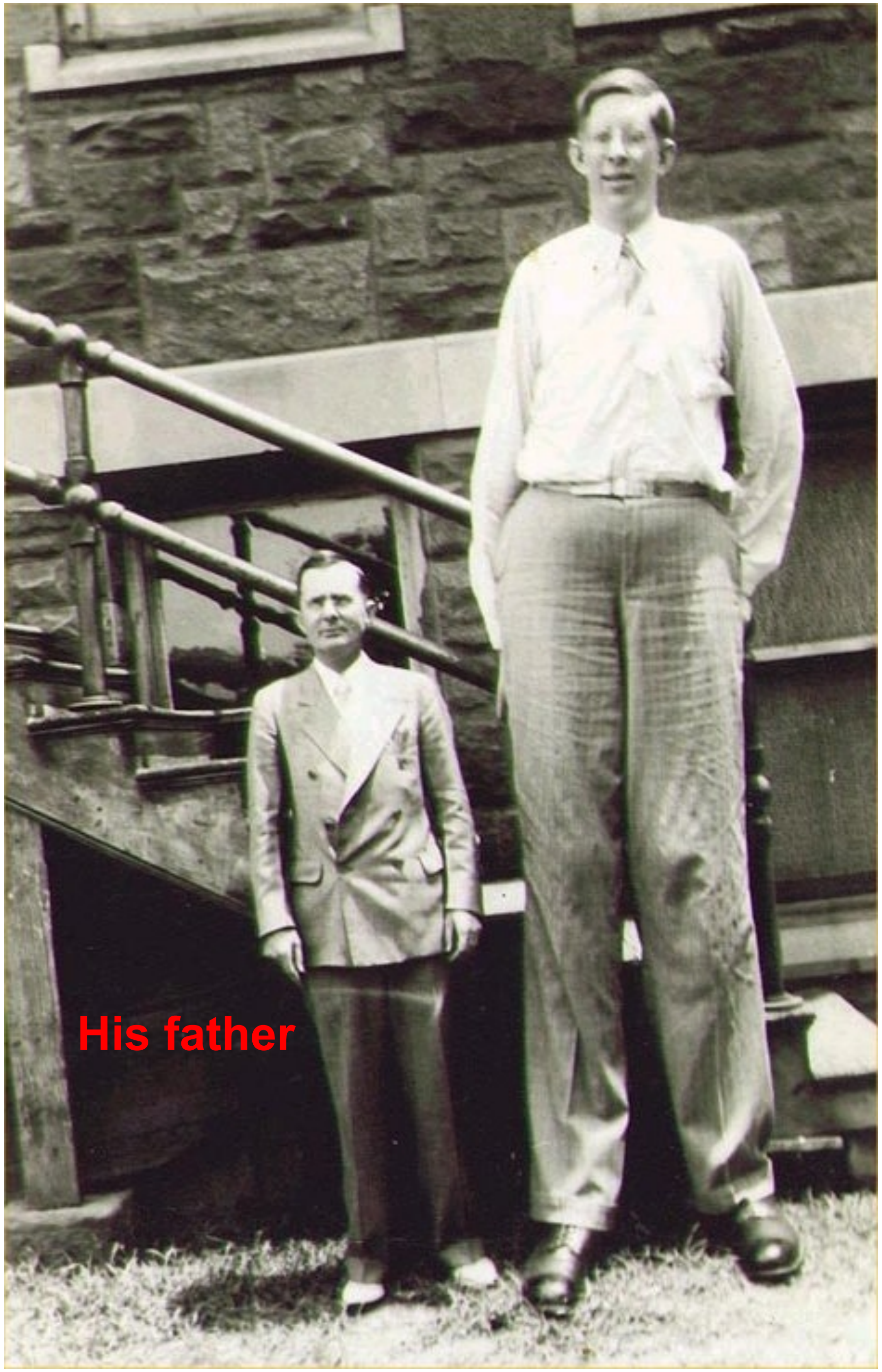
Bar chart of height of 30 year old men (23 measurements)

Real Population Height Distributions



These distributions can never be truly Gaussian, since there are upper and lower limits on adult human heights, while Gaussian distributions have tails out to very large values. From N=9185 distributions we have mean = 167cm, σ = 9cm.

272 cm
(+11.6σ)



236.2cm
(+7.5σ)

73 cm
(-10.4σ)

Measurement Errors

Where are these errors (uncertainties) from?

There are different sources of uncertainty and different types, but at the heart of it, is the fact that an experiment always measures something, and nothing can be measured with infinite precision. There will always be some uncertainty on any measured quantity, though it may be very small.

There are two types of error to deal with, those which are *statistical* in nature (and thus benefit from repeated measurement) and those which are "*systematic*", e.g. mis-calibration of a measuring device, which cannot benefit do not improve with repeated measurements.

Measurement Errors

If I asked each student in the class to measure the width of one of the lab tables using a metre-stick I would inevitably get a variety of different results, if people make the measurement to the best of their abilities.

You might all agree to the same number of *mm*, but your estimate of the next significant digit would likely vary, depending both on how you lay down the metre-stick and how you “eyeball” the reading.You should always record a measurement to the best precision you feel you can achieve.

Such reading errors are assumed to be random, if they are symmetrically distributed around the “true” value, then a given reading is equally likely to be higher or lower than the “true” value. This is why repeated measurements allow you to improve the precision of the result, since the average difference between the measured value and the true value must tend to 0.

Measurement Errors Cont'd

If the metre-stick is perfectly calibrated (*e.g.* it is exactly one metre long and is divided *perfectly* into 1000 $1mm$ divisions) and the table has exactly the same width at all points, then one expects that the average of the repeated measurement will provide a good estimate of the "true" value, since some people will measure a little too high and some a little too low.

We expect that the width of the distribution of measurements will tell us something about how well we believe the mean (which is our estimate of the true values) to be determined. Furthermore, the measurement uncertainty due to such random (Gaussian distributed) errors can be decreased by repeating the measurement: the error on the mean of a set of (independent) measurements decreases as $N^{-1/2}$ with N the number of measurements.

Systematic Errors

If the metre-stick from the last example is mis-calibrated (for instance if it is actually only 0.996 metres in length) then one will get measurements that are always larger than the measurement one would get with an unbiased (*e.g.* properly calibrated) metre-stick. This is an example of a systematic uncertainty.

This type of uncertainty does NOT improve with repeated measurements, since each measurement is off by the same amount (rather than by a random amount that can be either positive or negative).

Of course, the same applies to any measurement apparatus: voltmeters, ohmmeters, pressure gauges,……

This is simply an example of a correlated (rather than random) error. That is, the error is the same on each measurement.

A measurement typically has both a statistical error and a systematic error.

Systematic Errors

Other sources of systematic uncertainty (for example):

- Uncertain inputs: for example, if I asked you to determine the volume of a set of irregularly shaped pieces of copper, one good method might be to measure their weights and then divide by the density of copper. Since the density is only know to some accuracy (say for example 0.1%) then you could not provide an estimate of the volume that was better than 0.1%. (In this example the purity of the copper might also need to be accounted for).

- Model dependence: the parameters you extract from your measurements often depend on some model. Imagine dropping an object and timing how long it takes to fall to the floor, as a way to measure the acceleration due to gravity using $d = at^2/2$.
  - This is probably a pretty good model for a ball bearing, where drag is not an issue.
  - It's maybe not such a good model for a ping-pong ball.
  - It's unquestionably a terrible model for a feather

- For some measurements one needs to consider things like detection efficiencies and physics or instrumental backgrounds.

Estimation of Systematic Errors

Identifying sources of systematic uncertainty is mostly straightforward, but sometime not.

Estimation of a systematic uncertainty is a kind of art form. After all, you are often estimating the 68% confidence level on some quantity "by eye".

There is usually no one "right way of doing this". However…..

Saying that is ***not*** the same as saying there is no wrong way. Some judgement is required. There might be more than one sensible way to arrive at an estimate, but they should not disagree too much……..if they do you should think about things some more.

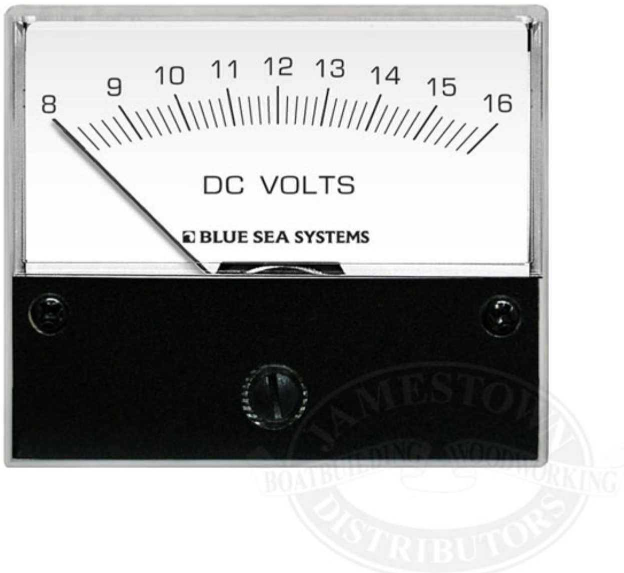
Caveat: Correlated Uncertainties

Systematic uncertainties are an example of errors which are correlated from one measurement to the next.

In the trivial examples we have used, the correlation has been 100%.

Correlations can of course exist with correlation coefficients that are less than 1. Treatment (propagation) of such errors is a little more complicated.

Reading Errors from Apparatus



For analog equipment, you are in much the same boat as with a meter-stick. You have to assign a reading uncertainty based on your judgment of how well you can read the device.

For this meter, can easily read to 0.1V (half a division, and perhaps even to 0.05V)

e.g. if you hook it up to 10.50V, does it actually read 10.50V ?

You also still need to know how well calibrated it is (or to test this with a more precise voltmeter, in this example). Usually a device will have a statement about it's accuracy somewhere upon it, or at least in the manual.

Digital Readout Apparatus

For reading error, generally assume ± half of the last digit (meaning ±0.005 in this case), *e.g.*

1.85 must be closer to 1.85 than to either 1.84 or 1.86, so in the range 1.845-1.855.

But you still need to know the calibration….

12061 Model Digital MultiMeter

- 6½ digits resolution.
- 11 types of measurement characteristics
  - DC voltage/current (1000V/3A max)
  - AC voltage/current (750V/3A max)
  - Resistance 2 or 4-wire ohms measurement
  - Period & frequency - Diode & continuity
  - Temperature (Thermocouple & RTD)
- DC voltage accuracy : 0.0015%
- AC voltage accuracy : 0.04%

so ±0.3V at 750V AC

- Up to 2000 readings memory storage

"There are known knowns. There are things we know that we know. There are known unknowns. That is to say, there are things that we now know we don't know. But there are also unknown unknowns. There are things we do not know we don't know."

Donald Rumsfeld, US Secretary of Defense, 12 February 2002

This is both amusing and instructive. For us, the unknown unknowns are worth thinking about. There will ALWAYS be systematics that you have not identified. However, as long as you have identified the **dominant** systematics this doesn't matter.

Note that uncorrelated systematic uncertainties will add in quadrature. So an addition systematic that is significantly smaller than the dominant one, will not add much to the total uncertainty.

How Many Repeated Measurements?

The error on the mean goes down like $N^{-1/2}$, (assuming that the N measurements are really independent) while systematic uncertainties are normally unaffected by repeated measurements.

In general, one likes to have enough measurements to ensure that the statistical error is smaller than the systematic error. However, there is not much value in making it a lot smaller:

$$x = 1.25 \pm 0.15(stat) \pm 0.25(syst)$$

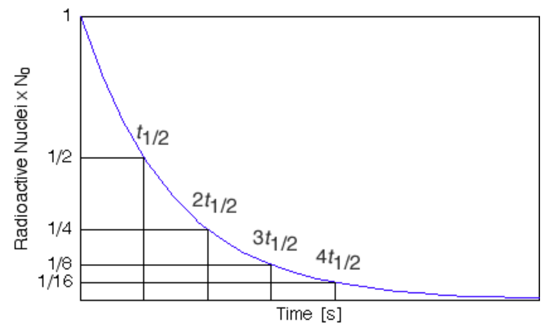
$$x = 1.25 \pm \sqrt{(0.15)^2 + (0.25)^2} \Rightarrow x = 1.25 \pm 0.29$$

No number of repeated measurements can reduce the total uncertainty below the systematic uncertainty. It is often more difficult to reduce systematic errors than statistical, but not always.

Example Measurement

Consider an experiment designed to measure the lifetime of some radioactive nucleus that decays via beta decay. We detect the occurrence of a decay via detection of the emitted electron.

Radioactive decay law: $N(t) = N_0 e^{-t/\tau}$



We measure the time of each decay, and then make a histogram of the number of decays in each time bin. One can fit the resulting distribution to the above exponential form to get an estimate of the lifetime τ (or do a linear fit to a log plot….).

If the lifetime of the material is very long, you might detect only a small number of events, in which case the statistical uncertainty may dominate.

What are the systematic uncertainties likely to be?

Lifetime measurement systematics

N_0 doesn't matter, except for the statistics.

However, the purity of the sample is relevant if there are impurities that also decay via beta decay (presumably with a different lifetime), since electrons from the decay of these impurities would affect your measurement. For instance, even a low contamination with a state having a much shorter lifetime would be a problem. Other background counts may be constant per time interval…..

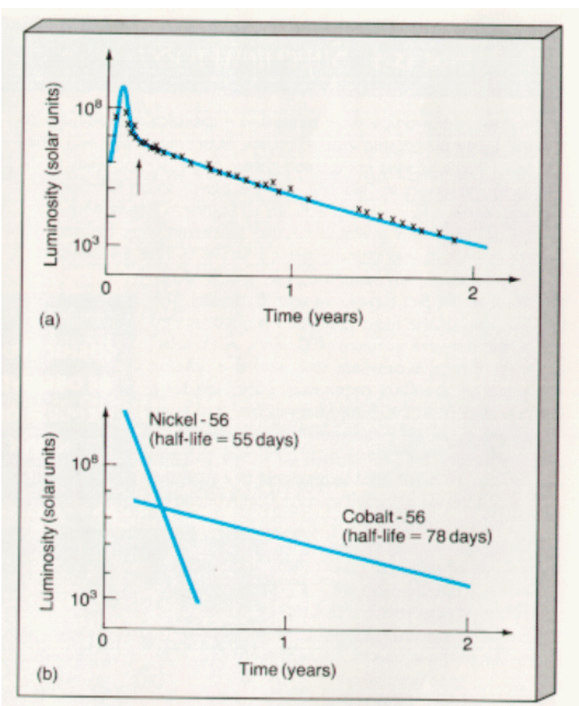
You have to detect the emitted electron, which you can do only with some efficiency ε_e which has some uncertainty, from detector efficiency and/or geometrical acceptance. This also might depend on the electron energy, which varies from 0 up to some kinematic limit. However, all of these effects are presumably independent of the time of the decay, so this also does not contribute.

You have to measure the time of each decay, which you do with some (presumably relatively small) uncertainty.

So statistical uncertainties may dominate in this case (as one example) in which case you need to take data for as long as possible.

Fitting Your Data

- "Fitting" data means adjusting the variable parameters in the physics (mathematical) model so that the model best agrees with the data.
- By convention, the errors on the parameters are those corresponding to the ±34.1% Confidence Interval around the mean value (e.g. the range that contains 68.2% of the distribution).
- A commonly minimized quantity is χ^2 which is a measure of the consistency of the measured data points and the fit function.



Just as a measurement always has an associated error, a fit also has errors associated with the fit parameters. These are important when comparing a result to theoretical predictions to (for example) validate or falsify some theory or model.

Chi-squared (χ^2)

Consider a set of n independent random variables x_i, distributed as Gaussian densities with a theoretical means μ_i and standard deviations σ_i, respectively. The chi-square is the sum

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2$$

In our case we typically have n measurements x_i compared to the fit prediction for the best set of fit parameters. The mean value of the χ^2 should be approximately the number of degrees of freedom, *e.g.* the number of bins (data points) minus the number of (free) fit parameters.

There is a probability distribution to get a certain χ^2 for a given number of degrees of freedom (dof). That probability is often quoted as a measure of the "goodness of fit".

- A χ^2/dof which is very small might indicate that the errors have been overestimated.
- A χ^2/dof which is very large indicates either that the model (fit function) assumed does not describe the data well, or perhaps that errors have been underestimated (some unthought-of systematic?)

Fitting Histograms

When fitting to binned data, it is usually wise to bin your data so that there are adequate statistics in each bin. Many fitting packages *assume* that the errors you have passed to it are Gaussian, so you want to ensure that your fit results are not affected by bins with so few entries that this is not a valid approximation.

Sometimes it may be appropriate to re-bin your data.

Sometimes it may be appropriate to restrict the fit range.

In any case, it is always wise to investigate how sensitive your fit results are to these issues.

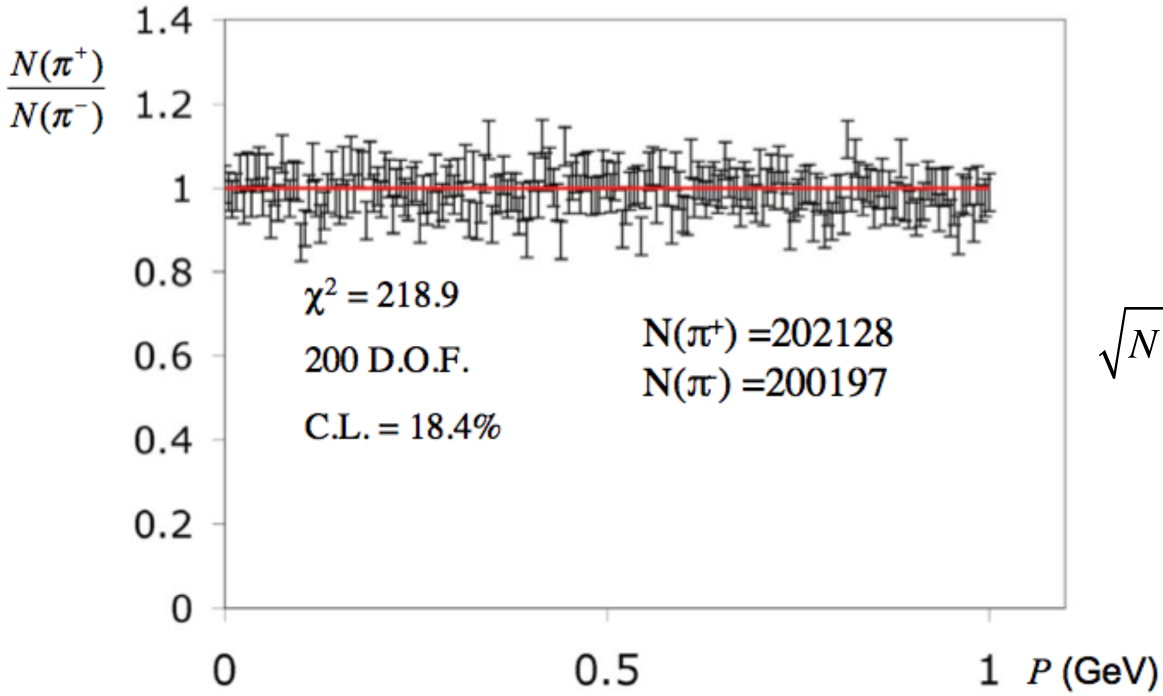
Note that the results of a fit that does not account for the measurement uncertainties is useless, as it cannot provide any uncertainty on the fitted parameters.

Charge Conjugation Invariance Test

In the reaction $p\bar{p} \to \pi^+\pi^-\pi^0$ charge conjugation invariance implies that the π^+ and π^- momentum distributions must be identical.

Test this by taking the ratio of the number of π^+ and π^- in bins of their momentum P.

Fit is to the hypothesis that the distribution is flat and = 1.



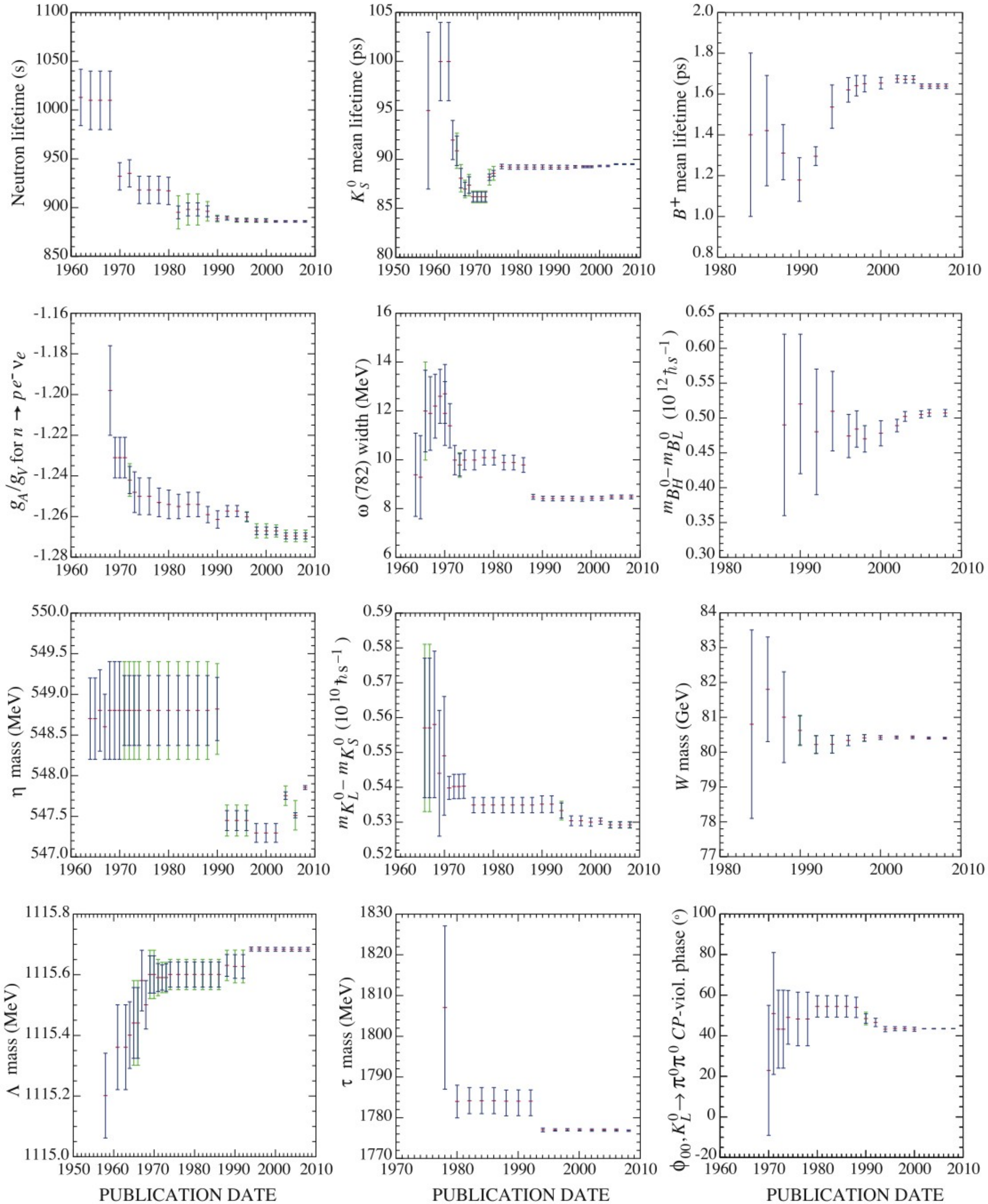
$\sqrt{N(\pi^+)} \approx 500$

Data Analysis / Fitting Software

- Matlab, Octave
- Maple, Mathematica, Reduce, …
- Excel (see comments on web pages)
- Python, C, C++, …
- Faraday, DataStudio, Kaleidagraph, …
- ROOT (C++ like particle physics analysis package)

You should use whatever software you are most comfortable with. We don't care what you use, but we do care that you understand what you do. In particular, when fitting you should know what minimization scheme is being used, and what the metric is for the "goodness of fit" (for example the χ^2/dof if the fit is minimizing the χ^2).

If you are staring from scratch, note that Python is the most supported platform in the APL.



Time-lapse view of some important measurements in experimental high-energy physics.

- Measurements improve as we improve experiments (statistics and systematic uncertainties).
- Physicists make mistakes!
- You will make mistakes too.
  - Try to learn from them.
- But they also learn to do things better
  - Try to do that too !!!

A final comment……

Understanding of the proper treatment of errors (uncertainties) is one of the most important things that you will learn in the lab course. This is also useful elsewhere in life (poll results, clinical trial results……).

By this I mean developing judgment about what uncertainties are relevant to a particular measurement, and also about how meaningful a given result is (for example when you see the results of some poll or medical trial reported in the press, or of course read a scientific paper on some measurement).

Poll results are often stated to be good to within (for example) 3.1% 19 times out of 20 (which is simply a statement that 3.1% = 2σ).

Some food for thought……..

If a newspaper article reports that 100 people were given a pill to test it's efficacy for a given condition and the researchers concluded that a positive outcomes were 20% higher than other results (*e.g.* no improvement or an adverse reaction) do you think that that's a meaningful result? How confident can one be that the pill is actually beneficial? How would you proceed as a researcher?