Lecture 4 – errors and experimental practice

- Focus on "errors":
  - Why do these matter.
  - Random and systematic types.
  - Key results for random errors.
  - Mitigating systematic errors.

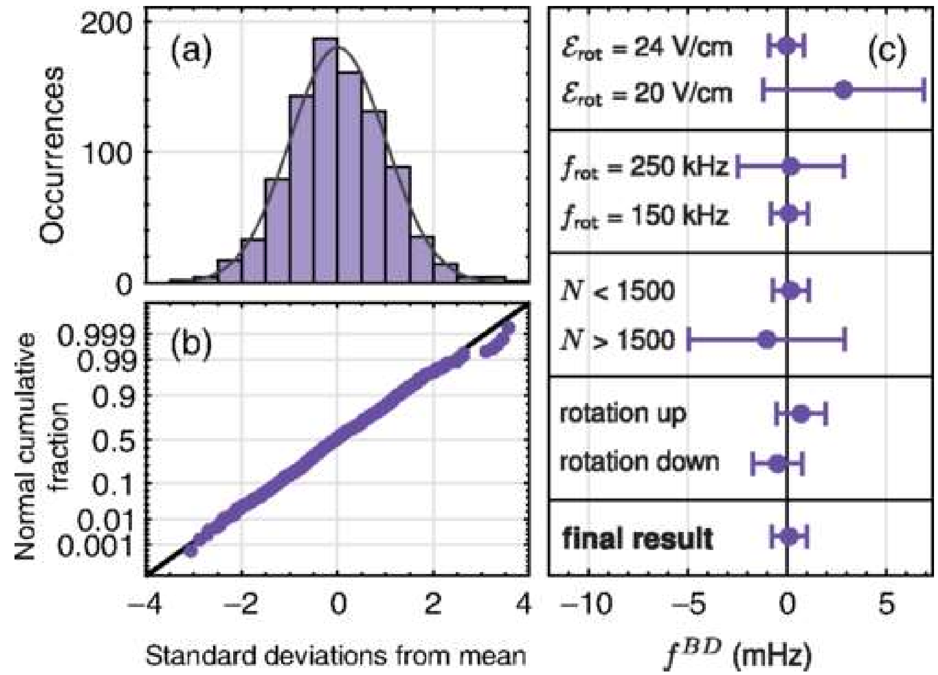


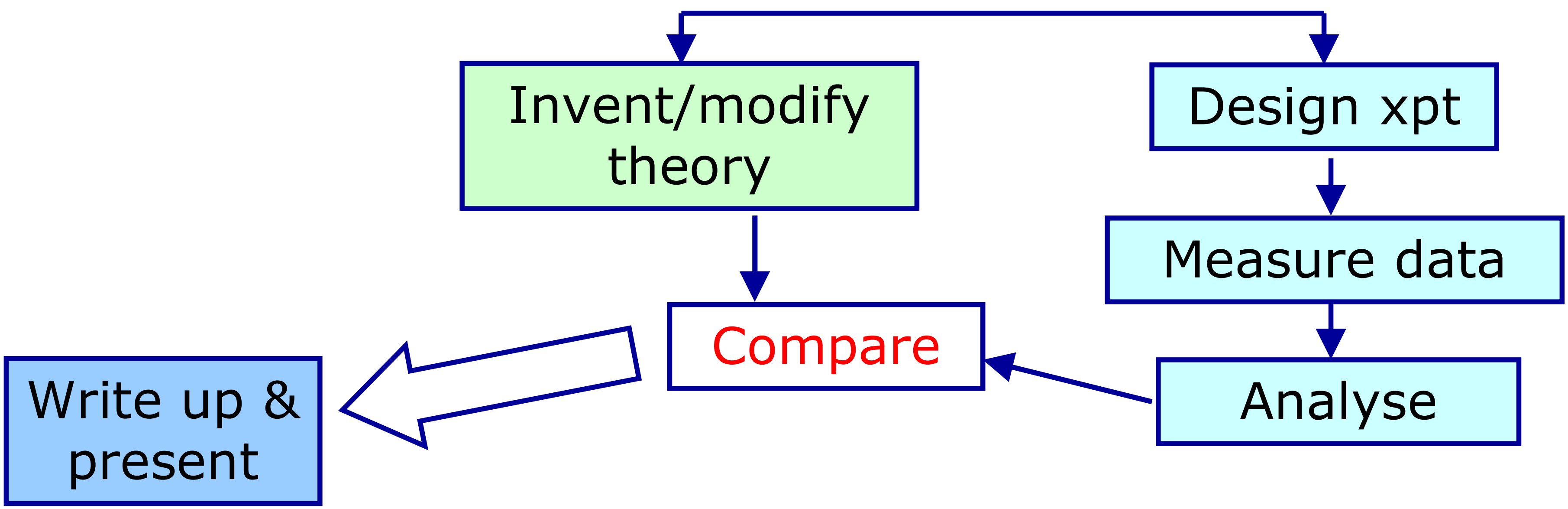
Figure 4. Summary of eEDM data set after cuts and scaling δf by $\sqrt{(\chi^2_r)}$ to account for overscatter. (a) Histogram of normalized, centered eEDM-sensitive frequency measurements $(f^{BD} - \langle f^{BD} \rangle)/\delta f$. (b) Normal probability plot of the same data set, showing a linear trend suggesting that the data are consistent with a normal distribution. (c) Subsets of eEDM data taken under different values of experimental parameters, and the overall average of f^{BD}. Here N is the average number of trapped HfF^+ ions per run.

Cairncross et al, PRL 119, 153001 (2017)

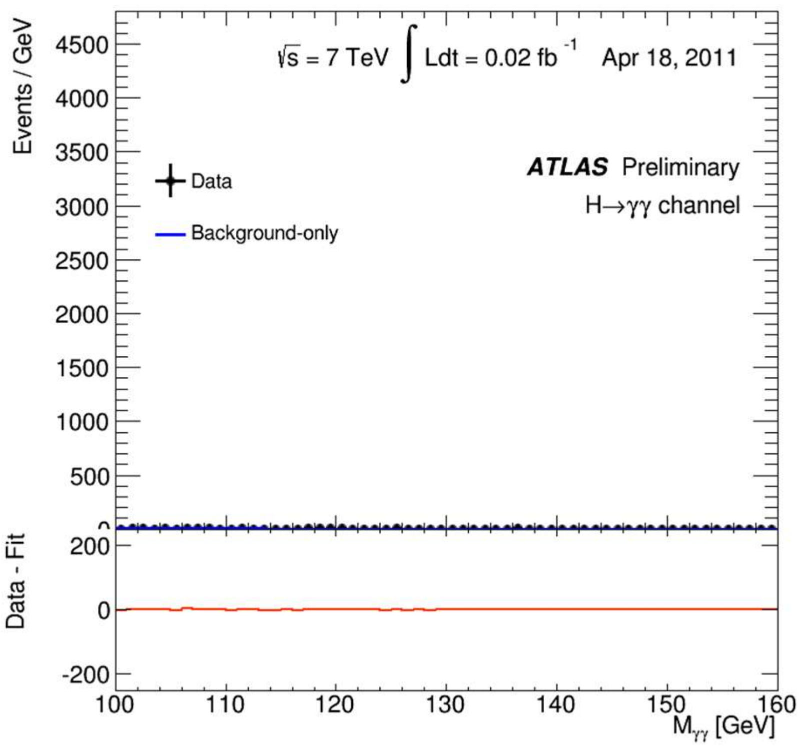
Errors matter in physics

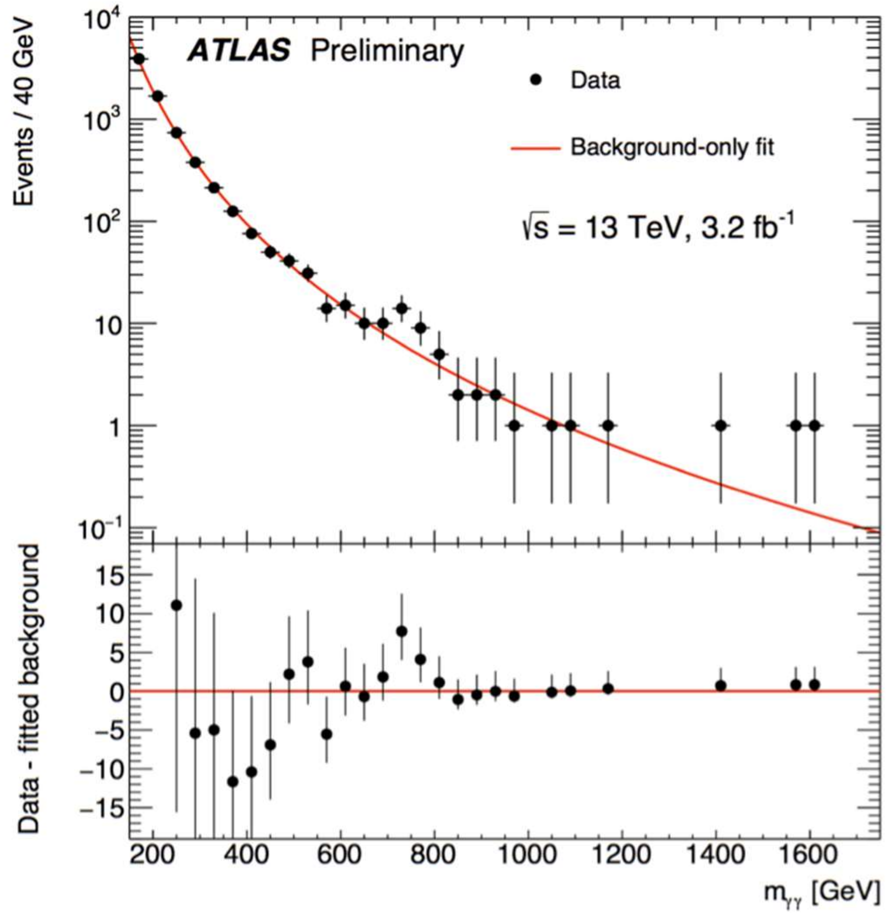
- In confronting a physical model with an experimental result, we need the experimental value and an estimate of its error to know if the experimental value is significantly different from the theoretical prediction.

- Error estimation is a critical part of doing real physics:



Lessons from particle physics



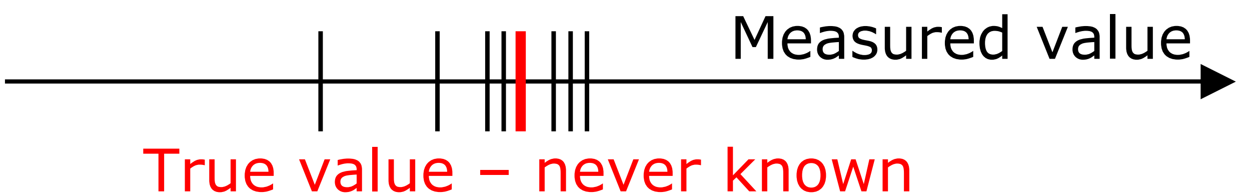


Possible digamma detection (3.5 sigma, Dec 2015)

We can distinguish between two types of errors (i) – random error

- Definition: the error (i.e., the true value minus the measured value) is equally likely to be +ve or –ve. More precisely, the average error is zero.

- Repeated measurements show this type of distribution:



- Combining the measured values in some way (TBD) is beneficial.
- These, random errors, are the ones treated in statistics books.
- The source of random scatter in experiment can come from

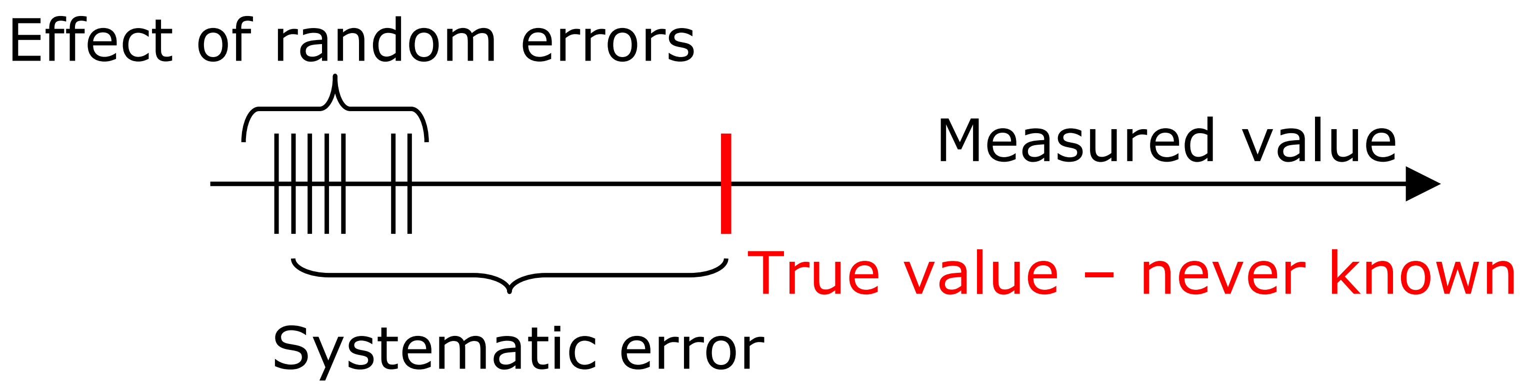
  (1) technical or (2) fundamental noise.

  e.g. diffraction limit, shot noise, …

And a more insidious type
(ii) – systematic error

- These are ≈ constant (or drift) in time.
- Best definition: all the errors that are <u>not</u> random.

  E.g. a miscalibrated oscilloscope timebase (reading, say, 5% low); a counter with counts lost due to a finite dead-time etc; a background signal mimicking the true quantity.



- No combination of measurements will reduce systematic errors.
- Only thoughtful experimental design can reduce these.
- Off-topic in statistics books and often ignored – at peril – in real research.

What strategies can be useful?

Example of resistance measurement with a DVM

- Is there a random error present?
  - Keep repeating the measurement.

- Is there a systematic error present? Approach is generally different in every case – so chasing these is hard.

  - Lead/connector resistance – does the DVM read zero when the leads are connected?
  - Do different DVMs agree?
  - What happens if the measurements are made in a different way, e.g., when using a bridge.
  - What happens with time:
    - Does the measurement disturb the system?

- Don't forget – you only need to reduce errors to the extent needed to achieve the required accuracy.

Skip in lecture

An aside about significant figures: presenting results

- 1.5678 ± 0.1 K is ridiculous.
- But 1.57 ± 0.1 K is necessary if you are recording an intermediate result and want to avoid rounding errors.
- Don't be fooled by your calculator – or by other people's "significant" figures:

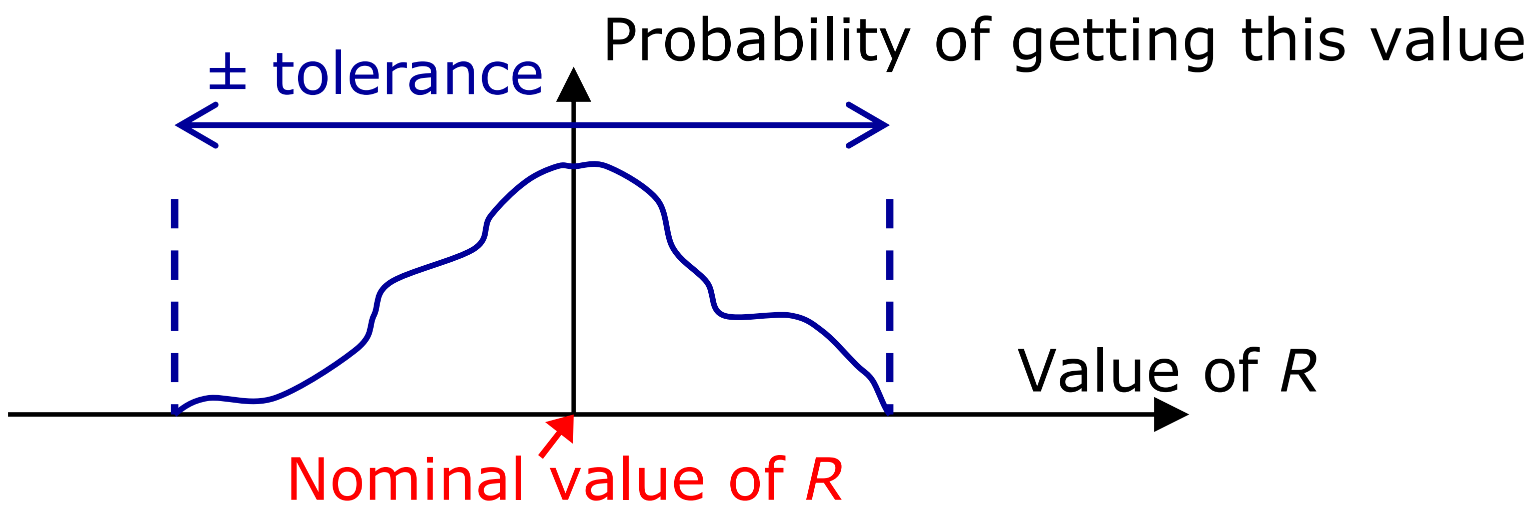
Suppose we are told that in 5 measurements with a ruler, the length of widget Z is 171, 171, 171, 172, 171 mm. What can we conclude?

- Best estimate of length is mean, i.e. sum/5 = 171.2 mm? NO
- Its likely that the measurements were quantised in units of a millimeter.
- Length is 171 ± around ½ mm, assuming no systematic error.

Defining what we mean by an error is critical

- Those of you who build electronics may have met the resistor color code. This indicates the resistance (at room temperature, at DC …) to within the "tolerance".

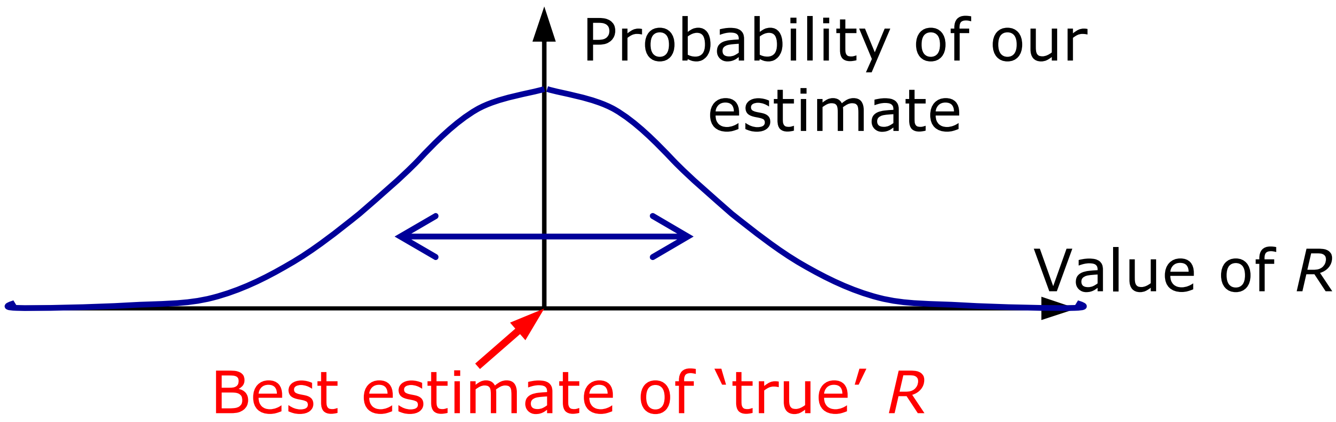
- Tolerance is a definition, usually, associated with the full range within which a value may lie:



One is (supposedly) 100% confident that R actually has a value in the range nominal ± tolerance.

In physics we use a different – but standard – approach when we quote errors

- We interpret the error, again, in terms of the probability distribution of the unknown quantity.



- Unless otherwise stated, the probability distribution is assumed Gaussian, the error then being the associated σ.
- The actual value of the quantity can lie outside the range $\pm\sigma$.
- If we know about the probability distribution itself, we can be more specific and discuss "confidence intervals".

How to treat data subject to random errors

- Assume all we have is a sample of *N* "equivalent" measurements of quantity *x*. Call each measured value x_i. We do not know the true value of *x*.

- The best estimate of the true value, assuming we have reduced systematics as far as possible is:

  [for distributions with finite variance]

$$\text{Mean, } \bar{x} = \frac{1}{N}\sum_i x_i \,.$$

- To characterise the spread of the x_i, we work with the square of the deviation $(x_i - \bar{x})$ either side of the mean value since the sign of the deviation does not matter. There are two important definitions:

$$\boxed{\text{Variance,}}\ \mathrm{Var(x)} = \frac{1}{N}\sum_i (x_i - \bar{x})^2 \equiv \overline{(x - \bar{x})^2} \equiv \langle (x - \bar{x})^2 \rangle \,.$$

$$\text{Its square root, the } \boxed{\text{standard deviation}} = \sqrt{\frac{1}{N}\sum_i (x_i - \bar{x})^2} \,.$$

Textbook confusion: *N* vs (*N*-1)

$$\sqrt{\frac{1}{N}\sum_i (x_i-\bar{x})^2}\ .$$

- In the previous expression we have used a factor of *N* to average the squared deviations.

- If, however, we wish to use the spread of the data to estimate the random error in a ***single datum***, then the expression we use is:

$$\text{best estimate of error is given by } \sqrt{\frac{1}{N-1}\sum_i (x_i-\bar{x})^2}\ . \quad = SD(x)\sqrt{\frac{N}{N-1}}$$

This involves a factor *N-1*, since we have had to use the data themselves to estimate what the mean value is (i.e. we do not know what the true value of the quantity is).

- If all we are interested in is quantifying the spread of our data, then the formula with *N* is correct.

Is averaging the correct thing to do and can we quantify its benefit?

- Yes, if the measurements are all independent and otherwise identical (i.e. associated with the same error process).

- To answer the 2nd question we use a result from statistics:
  - "The variance of a sum of independent random variables is = the sum of the variances of each of these".

Suppose we have two independent quantities y and z. For convenience let $\bar{y} = \bar{z} = 0$.

These terms are both zero-mean

Now consider the variance of their sum:

$$Var(y+z) = \frac{1}{N}\sum_i \big(y_i + z_i - (\bar{y}+\bar{z})\big)^2 = \overline{(y+z)^2}\,.$$

Now expand $\overline{(y+z)^2} \rightarrow \overline{y^2 + z^2 + 2yz} = \overline{y^2} + \overline{z^2} + 2\overline{yz}\,.$

The variance of a sum of 2 independent quantities – contd.

$\overline{yz} - \bar{y} \cdot \bar{z} = CoVar(y,z)$

But y and z are independent of each other, so $\overline{yz} = \bar{y} \cdot \bar{z} = 0$.
So, $Var(y+z) = \overline{y^2} + \overline{z^2}$.

So, variance of sum = sum of variances for independent quantities.

- We can apply this to our set of *N* measurements of *x* (each of which had some error) to establish the error in the mean of the set:
  - This is because the mean is just a scaled sum of these independent quantities.

Why averaging helps (when errors are random)

For one particular measurement we can say that:
$\text{error}^2 = (x_i - x_{true})^2 \approx (x_i - \bar{x})^2 = \text{var(one measurement)}.$

For the complete set of measurements our statistical result says:

Var(sum of measurements) = ∑ (variance for each measurement)

Var(N · mean) = ∑ (variance for each measurement)

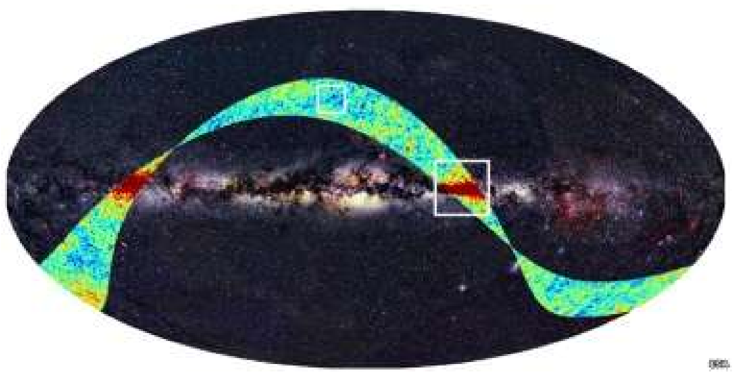
N^2 Var(mean) = N · (average variance for each)

Var(mean) = (average variance for each)/N

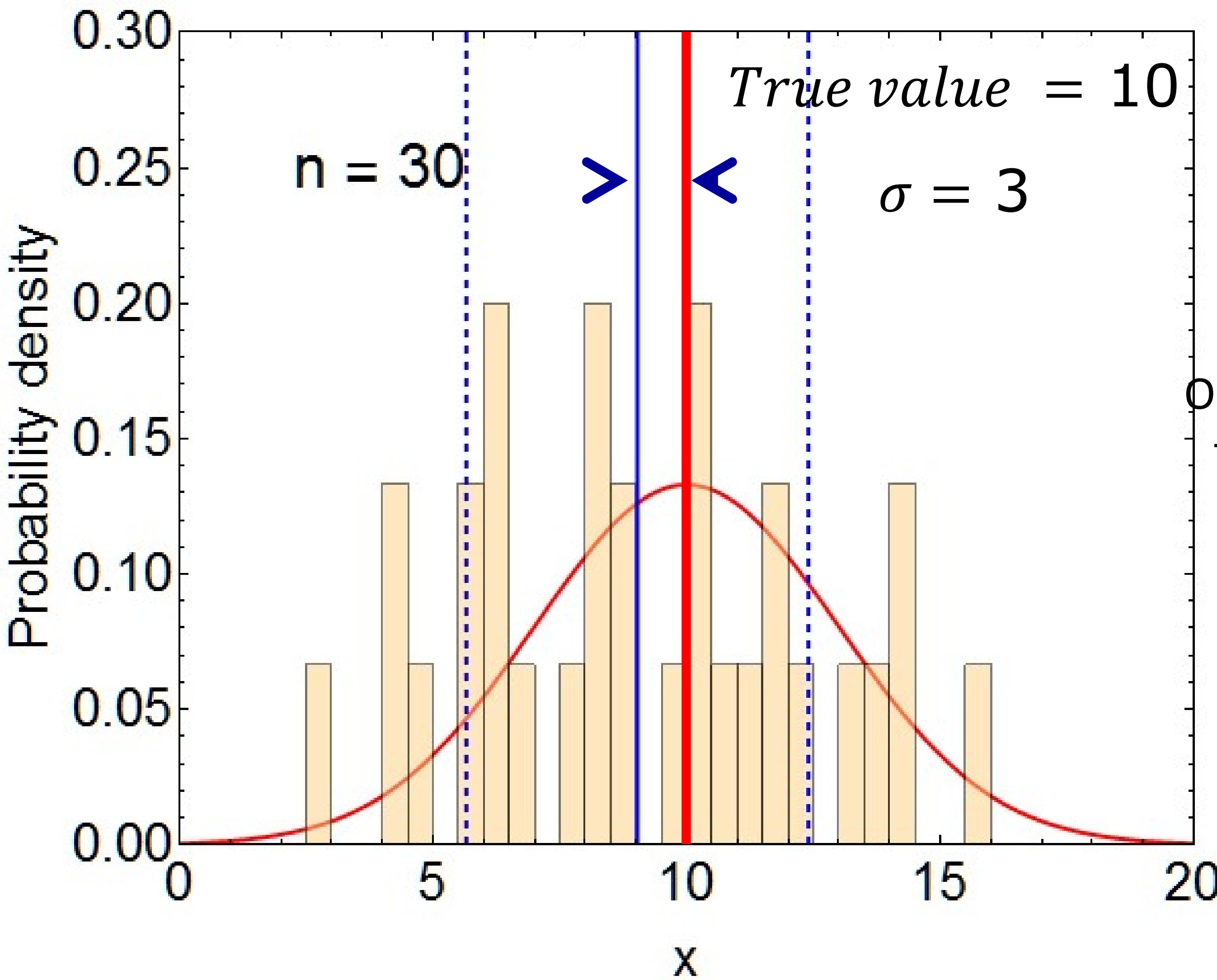
$$\text{Uncertainty in mean} = \frac{\text{uncertainty in each value of } x}{\sqrt{\text{N}}}.$$

How do we exploit this knowledge in practice?

- Take twice as many data => resultant error in mean should fall by a factor of √2 as compared to a single datum.
- Staring at a flickering ’scope screen four times should halve the reading error – your eye/brain combination will do the averaging for you.
- Measuring a small patch of the 2.7 K cosmic microwave background radiation for 100 times longer should decrease the measurement error by a factor of 10 (as long as systematics don’t creep in…).



Sampling a Distribution



$For\ 30\ samples,$

$\bar{x} = 9.1$

$SD(x) = 3.3$

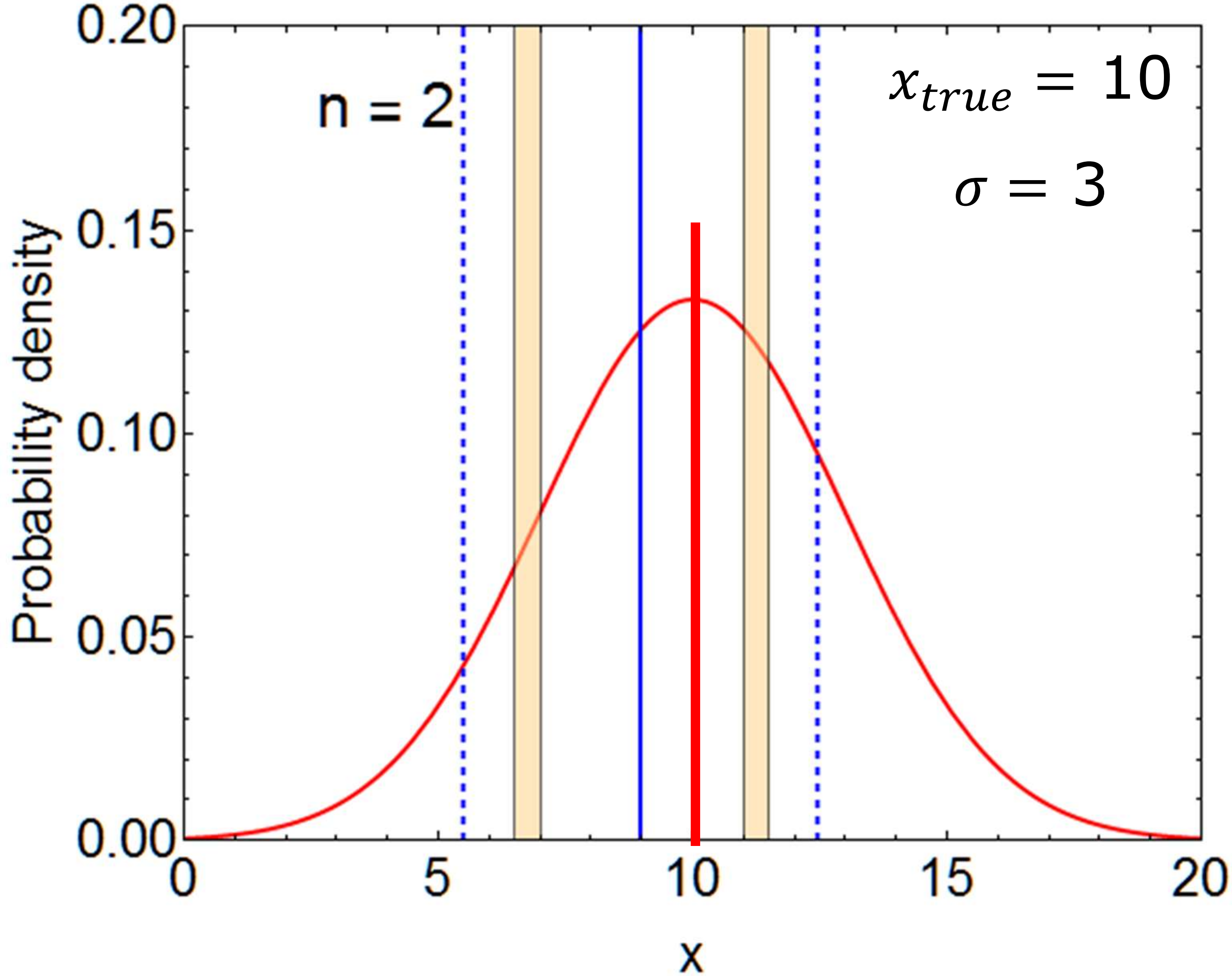
$\mu - \bar{x} = 10 - 9.1 = 0.9$

Our best estimate of the error in the mean (if we didn't known the 'true' value μ):

$$\frac{SD(x)}{\sqrt{n-1}} = 0.6$$

We would estimate $\mu = 9.1 \pm 0.6$

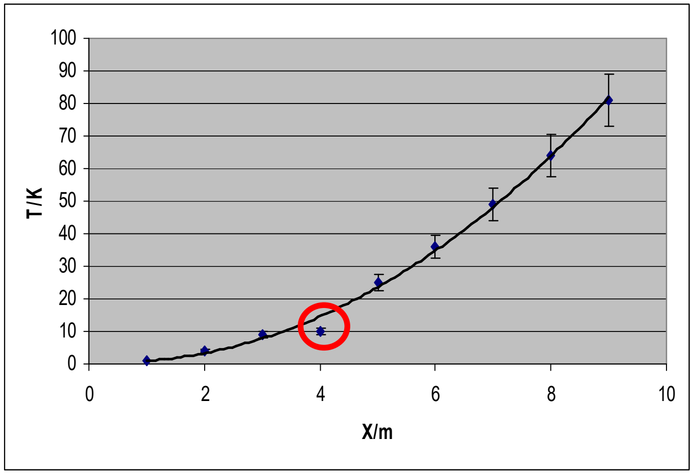
Sampling a Probability Distribution



How many measurements should you make?

- Measure once:
  - Random error ε will be present.
- Measure again:
  - If errors are random, combined error should be $\varepsilon/\sqrt{2}$.
  - But, importantly, you will also have checked for a mistake.

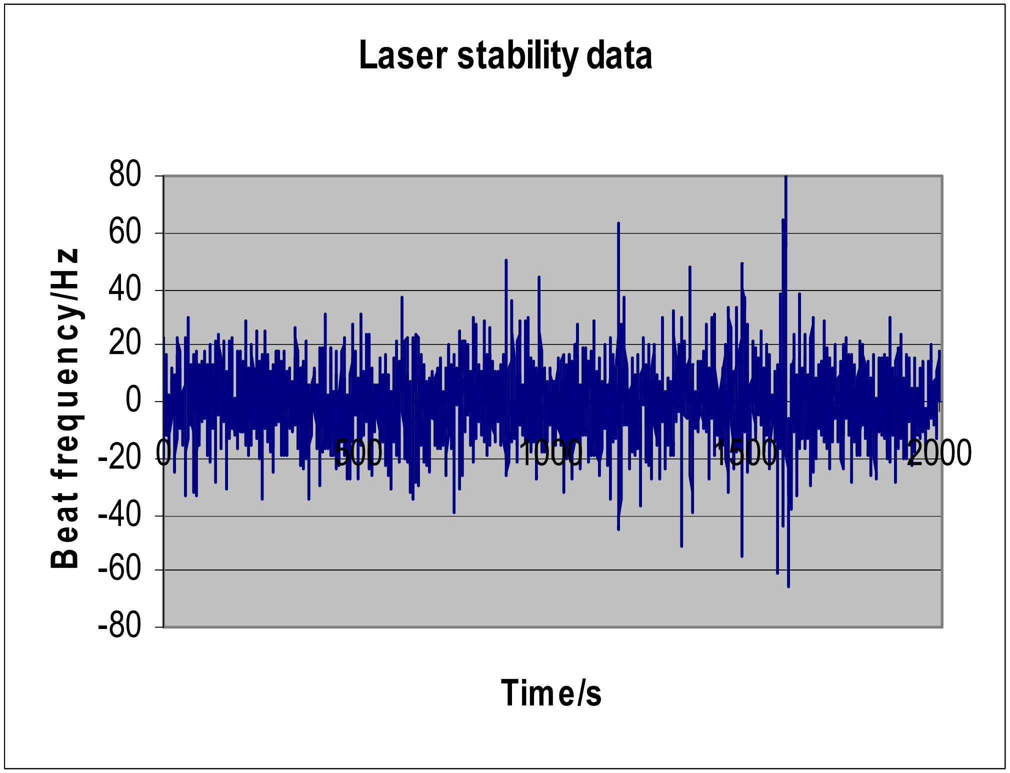
- NB – depending on the context, you may not need to measure with the same nominal setup to reveal a mistake. Neighboring data can alert you.

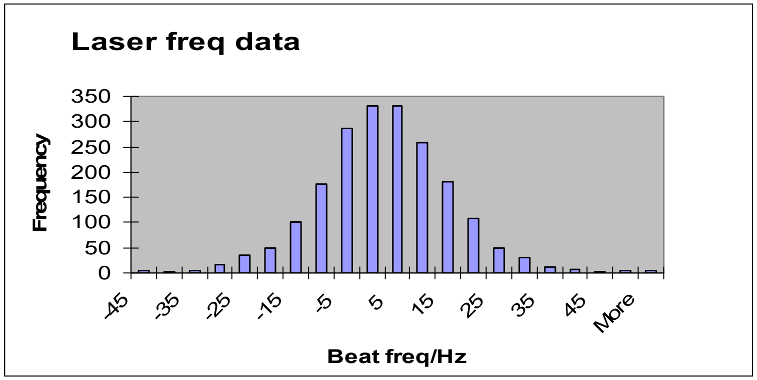


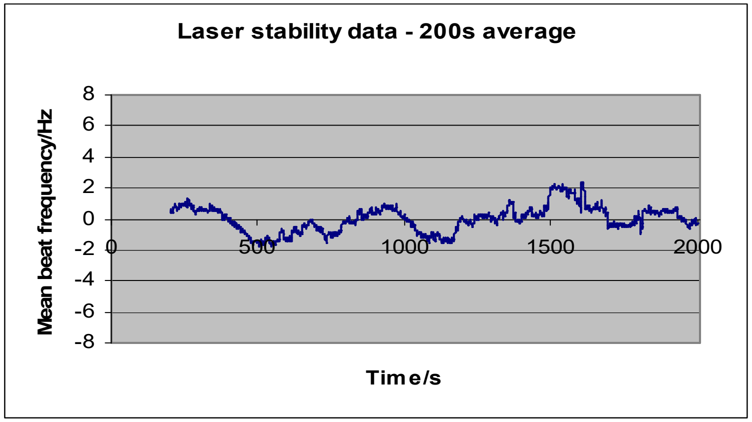
- Measure 100 times:
  - Random error should be reduced to $\varepsilon/10$.
  - May be more efficient to re-design the experiment to lower ε in the first place.

measurements cont[d.]

- Measure 10,000 times:
  - Resultant error should be $\varepsilon/100$ if systematics are unimportant.
  - Feasible iff data are secured electronically – but then is easy.
  - This approach is helpful in spotting systematic drifts.







How do errors in an experiment combine, i.e. consider f(x,y,..) where x, y,.. are measured

Consider a function $f(x, y, \ldots)$. For changes in $x, y, ..$ we know we can write to first order: $df = \frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy + \ldots.$

This is close to what we need, but we have errors, σ_x, σ_y, not changes dx, dy

It turns out that if the errors are Gaussian, as they usually are, the best estimate of σ_f, the resultant error in f is given by:

$$\sigma_f^2 = \left(\frac{\partial f}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_y^2 + \ldots$$

- This is another application of the variance of a sum equalling the sum of the variances
- I.e., we have assumed independent errors in *x, y,* …
- Note how the partial differentials act as "scale factors"

Skip in lecture

$$\sigma_f^2 = \left(\frac{\partial f}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_y^2 + \ldots$$

What does this result mean in practice

If $f = x \pm y$, then $\sigma_f^2 = (1^2)\sigma_x^2 + (1^2)\sigma_y^2 \Rightarrow \sigma_f^2 = \sigma_x^2 + \sigma_y^2$.

Add absolute errors "in quadrature"

So, if $x = 10 \pm 1$, and $y = 20 \pm 1$ then $f = x + y = 30 \pm \sqrt{2}$.

If $f = x^n$, then $\sigma_f^2 = (nx^{(n-1)})^2\sigma_x^2 \Rightarrow \sigma_f = \frac{nf}{x}\sigma_x \Rightarrow \frac{\sigma_f}{f} = n\frac{\sigma_x}{x}$.

Fractional error multiplied by *n*

So, if $\mathrm{f} = x^2$ and $x = 10 \pm 1$ then $f = 100 \pm 20$.

Similarly, if $f = x \cdot y$ or ${}^{x}/_{y}$ then

$$\left(\frac{\sigma_f}{f}\right)^2 = \left(\frac{\sigma_x}{x}\right)^2 + \left(\frac{\sigma_y}{y}\right)^2.$$

Add fractional errors "in quadrature"

What about more complicated cases?

- We still use the "quadrature" formula for any form of *f(x,y,...)*.

$$\sigma_f^2 = \left(\frac{\partial f}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_y^2 + \ldots$$

- Usually its easiest to evaluate the $\partial f / \partial x$ etc by partial differentiation. But if *f* is complicated, it can be quicker to evaluate $\partial f / \partial x$ empirically as:

$$\frac{f(\bar{x} + \sigma_x, \bar{y}, \ldots) - f(\bar{x}, \bar{y}, \ldots)}{\sigma_x}.$$

- Warning: the Taylor expansion doesn't work for large deviations from a best estimate, so you need to look at the full distributions of derived quantities in these cases:
  - Play with the NIST uncertainty machine to explore these: https://uncertainty.nist.gov/

Other applications of the quadrature results

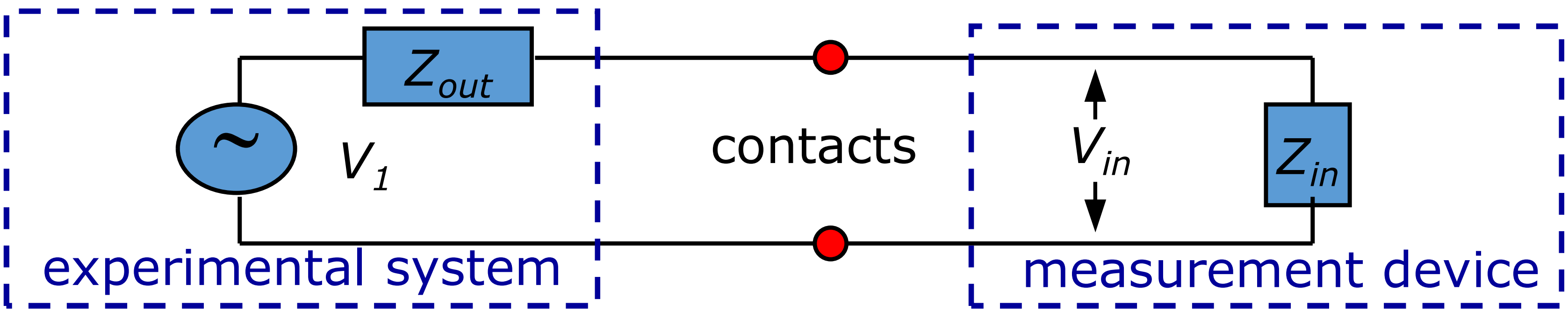
- If you are investigating some relationship of the form, $p=aq^2$ for some $(q_1 \ldots q_n)$, its usually best to organize a straight-line plot:
  - $Log(p) = log(a)+2log(q)$.
  - Easy to see if there are deviations from a straight line.
- In these cases, you need to "transform" your errors (in p, what you are measuring) using the previous results.

- But suppose instead $p=aq+bq^2$:
  - $p/q = a + bq$, so plot (p/q) vs q.
  - Now errors in x and y axis values are no-longer independent.
  - So, e.g., treatment of errors in gradient and intercept are non-trivial.

Reducing systematic errors (i)

- No general prescription – but use the physics and experience you know. Some common points:

1. Is the instrument accurate? Calibrate it.
2. Is the apparatus what you think it is?



Exploit the symmetry of the situation. Reverse the voltage source to your measurement device and the magnitude of *V* should be unchanged.

- But it <u>will</u> change – because the voltages arising from the different metals touching at the contacts have the same sign, regardless of the sign of the experimental voltage source.

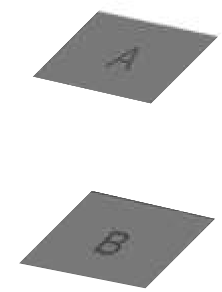
Reducing systematic errors (ii)

3. Can you believe what you see?
   This is a fundamental issue:

- The squares marked A and B are the same shade of grey.

- How do we proceed here?
  - Appreciate the extent to which your brain will imprint its prejudice onto the interpretation of data.

  - Try alternative methods in any given experiment.

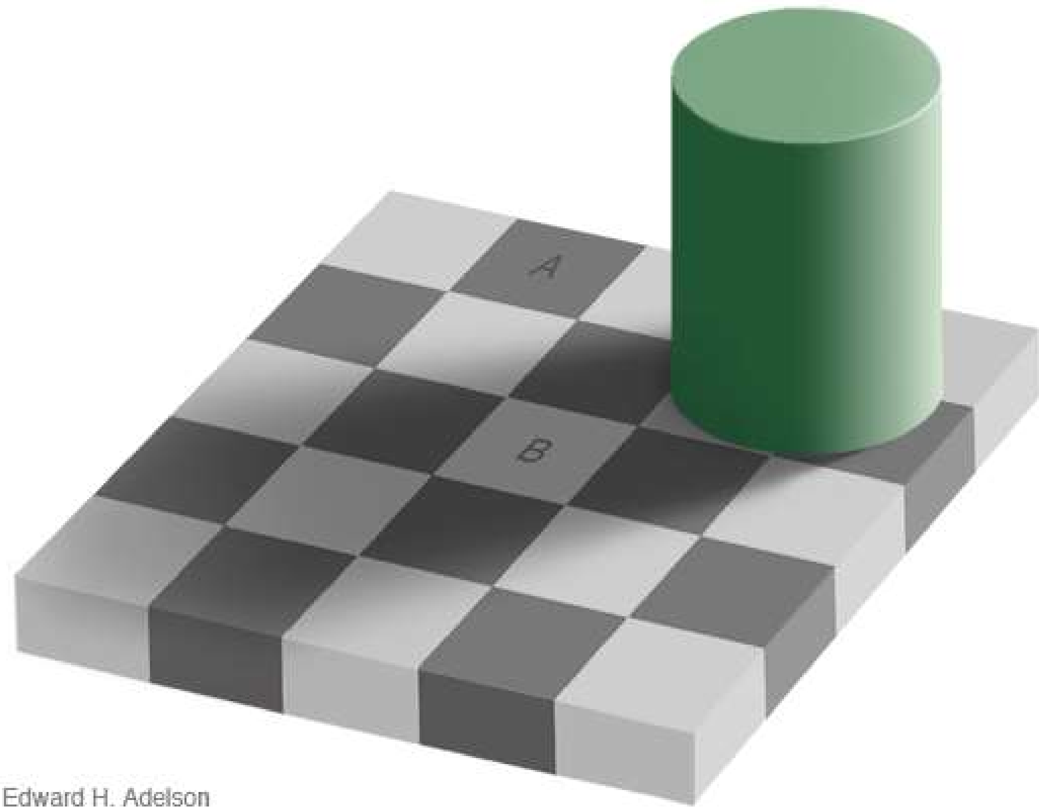


Edward H. Adelson

Reducing systematic errors (ii)

3. Can you believe what you see?
   This is a fundamental issue:

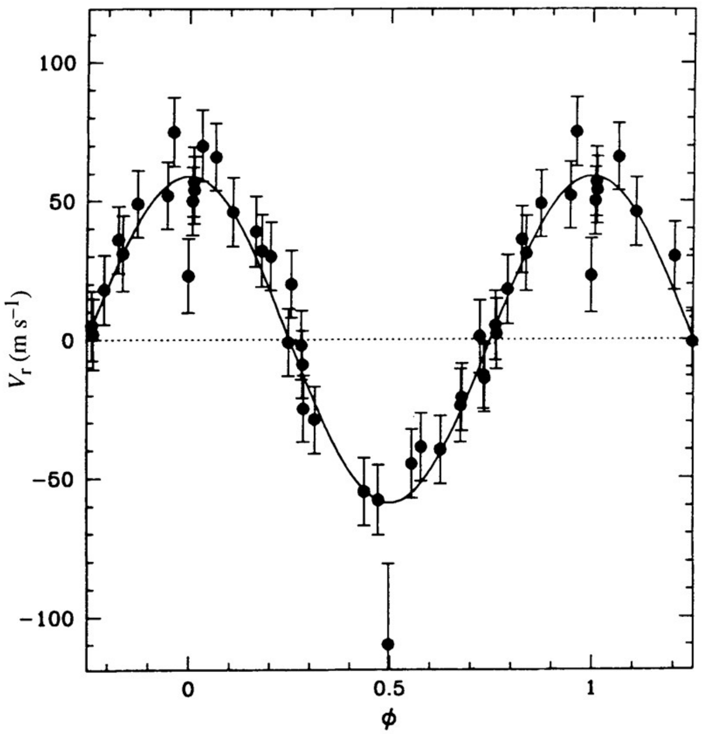
- The squares marked A and B are the same shade of grey.

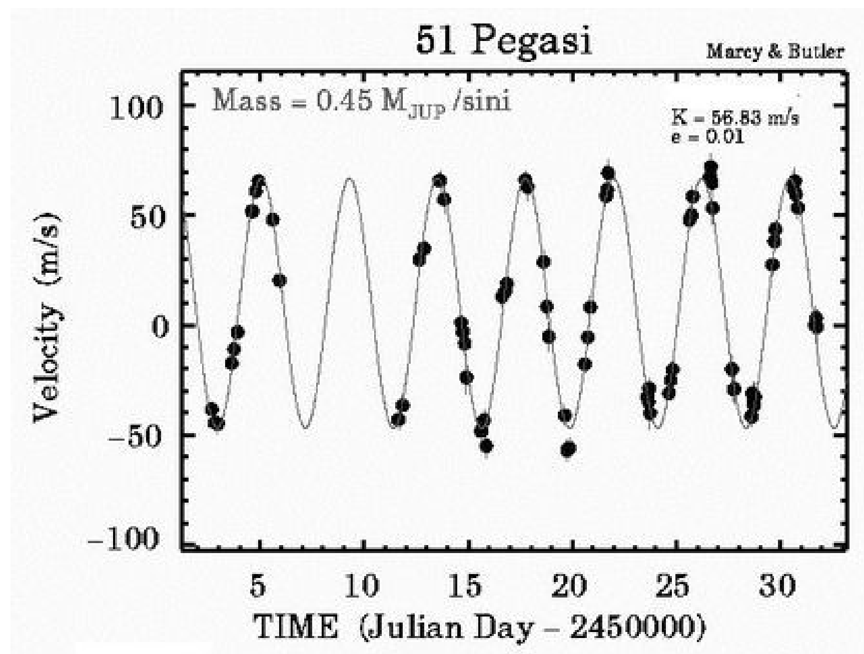


- How do we proceed here?
  - Appreciate the extent to which your brain will imprint its prejudice onto the interpretation of data.
  - Try alternative methods in any given experiment.

Prejudice can limit your success considerably

- For a Jupiter orbiting at 5AU the radial velocity signature has an amplitude of 12 ms^{-1} over 12 years.





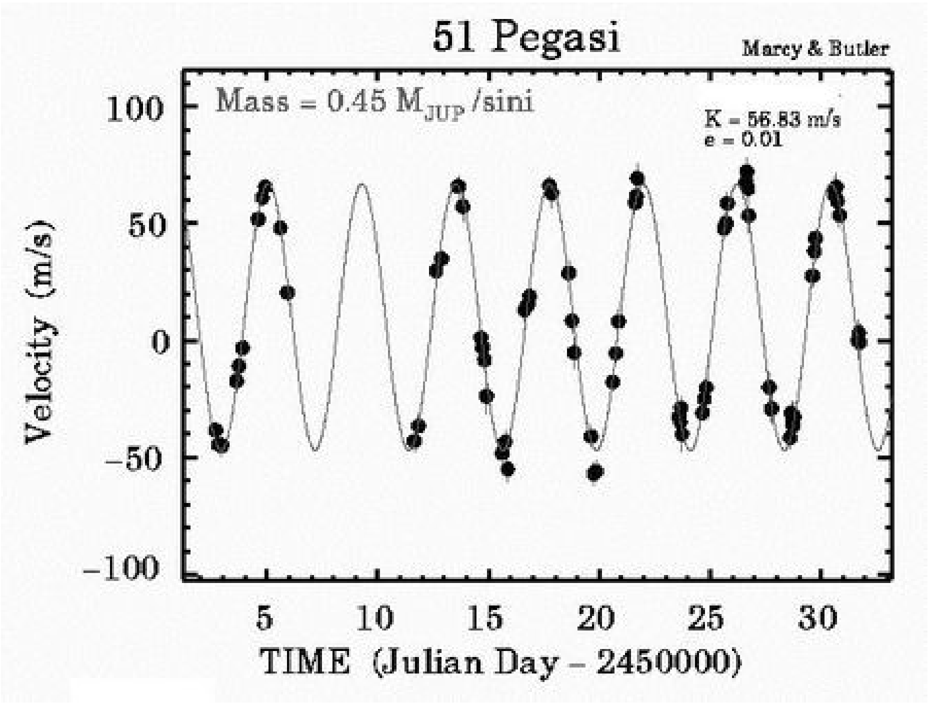
[First detection of 51 Peg by Mayor & Queloz (Nature, 378,355, 1995) together with additional data from Marcy and Butler.]

How systematic effects hindered the detection of extra-solar planets for many years

- For a Jupiter orbiting at 5AU the radial velocity signature has an amplitude of 12 ms^{-1} over 12 years.

- The experiment was designed to search for a temporal signal with a long period

- The experiment needed very well calibrated data

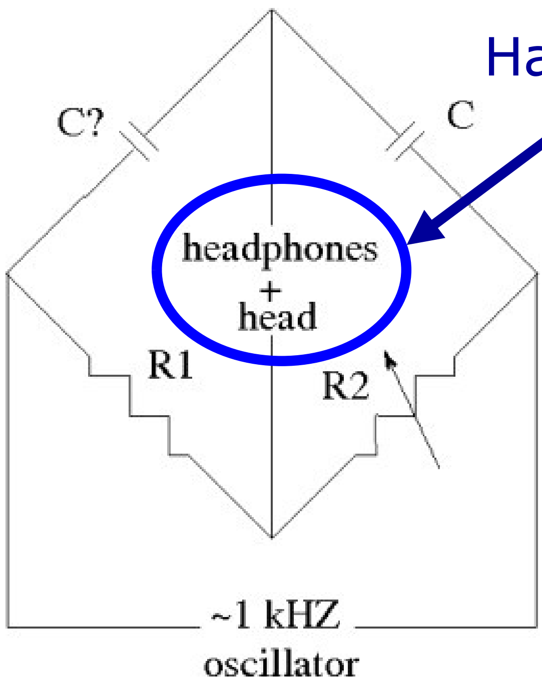


[Radial velocity data courtesy of Marcy and Butler.]

Reducing systematic errors – cont^d.

4. Use a null method:
   - Quantity X being measured is opposed by an adjustable Y until the indicating device shows balance.
   - The benefit is that the indicating device does not need to be linear or well calibrated.

Hardly linear or calibrated



No sound => $C_?/C = R_2/R_1$

But R_1, R_2 and C do need to be accurately known.

Reducing systematic errors – cont$^{d.}$

5. Be attentive to changes with time (the experiment may be warming up, etc.):
   - To at least randomize this effect in measuring the quantities A, B & C, do not measure, e.g., in order AAABBBCCC but rather something like ABCCBA… etc.

6. Make a differential measurement:
   - Measure the difference in the quantity with the system in the desired state and then in a standard – and if possible – similar states.
     - E.g. thermocouple output in unknown temperature ≈100 C compared with output in boiling water.

Skip in lecture

Reducing systematic errors – cont$^{d.}$

- A micrometer (which expands and contracts with temp) is such an example

The difference in the location of the spindle relative to the anvil measures the "width" of the sample. The instrument relies upon the frame being rigid – but only on short timescales.

Importantly, only the thread on the spindle and its bushing need to be made of low CTE material.

- NB – a micrometer (like many mechanical devices) exhibits "backlash": play in its mechanical parts such that you get a different systematic error as you move the spindle to left or right. => always approach from the *same* side.

Reducing systematic errors – cont$^{d.}$

7. Selection effects:

   Example 1: "You catch fish in a river with a net. You determine (a) all fish have scales, (b) all fish are longer than 4 inches."

   Example 2: You have 4 filament lamps. Which radiates the most power?

   Look at them – surely it's the brightest?

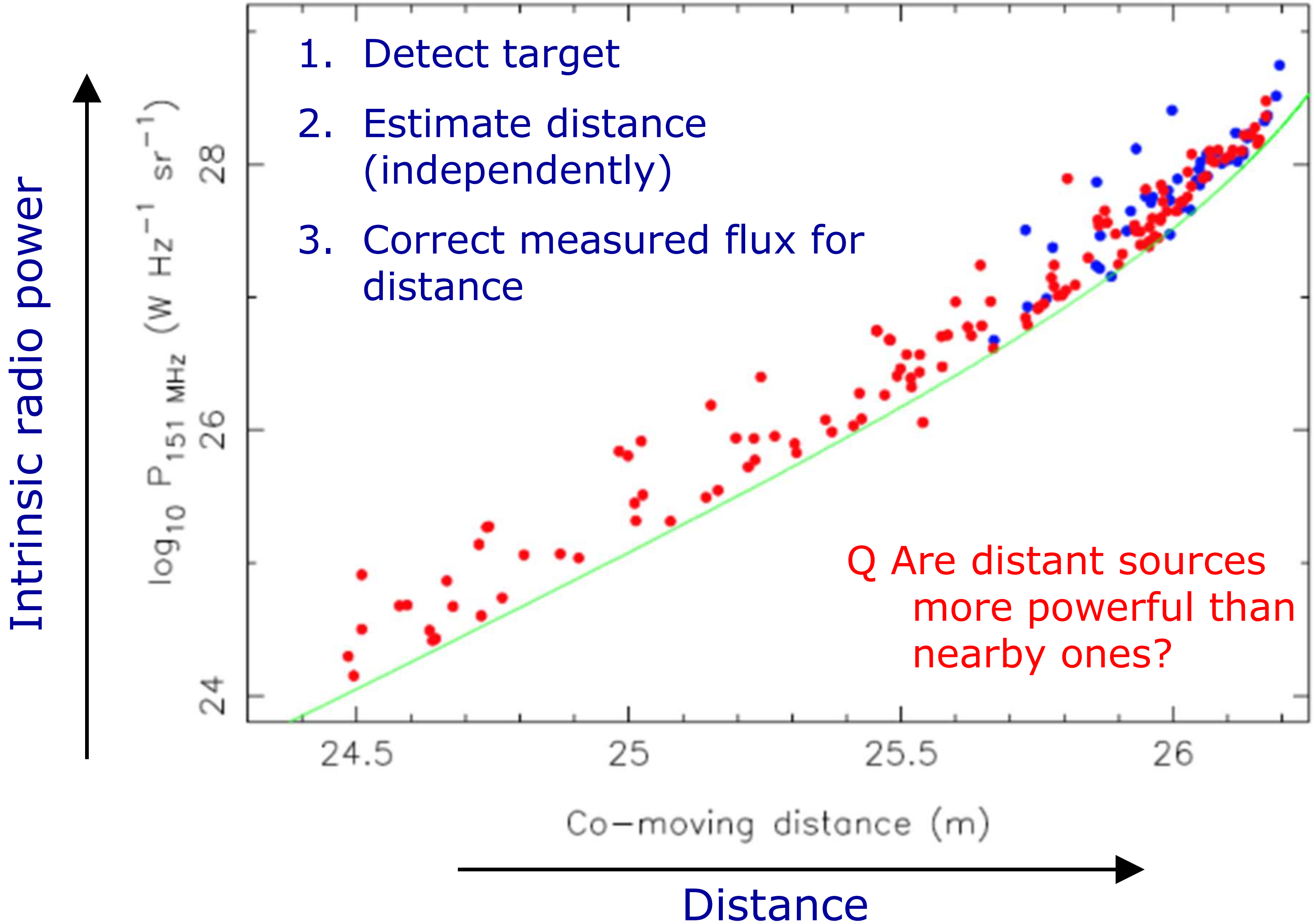
   Yes but a visibly dim one may be still be putting out say 100 W in the infrared.

   Selection effects can be subtle and all too easy to miss. You have to make sure you measure what you want to measure.

   Be extra careful about the result you would like to measure.

Skip in lecture

Selection effects – radio power vs distance



Summary so far

- The estimation of uncertainties is crucial in undertaking real science.
- We have distinguished random and systematic errors:
  - Random:
    - Error in mean of X = error in each value/$\sqrt{N}$.
    - Quadrature formula for combination (if independent).
  - Systematic:
    - Not reduced by averaging.
    - Eliminated through careful design.
    - Usually the limiting factor in all precision experiments.