Data Analysis and Statistics



Prof Adrian Oeftiger

(adrian.oeftiger@physics.ox.ac.uk)

Based on lectures given by Dr Jenny Barnes

The Scientific Method

Feynman on the Scientific Method:

> *"It doesn't make a difference how beautiful your guess [theory] is,*
> *it doesn't make a difference how smart you are,*
> *who made the guess, or what his name is...*
>
> ***If it disagrees with experiment, it's wrong.**"*

$\Longrightarrow$ How can we be certain that an experiment and theory agree or disagree?

This is what we hope to teach you with this course and throughout your time in the practical course.

This course

4 lecture slots:

- 3 lectures
- problem sheet: go through questions in lecture 4

Syllabus (in Year 1 handbook)

- *A-level Maths plus the bridging material in (a)* ***data analysis and statistics*** *and (b)* ***computing*** *are assumed*
- *Qualitative introduction explaining what experiments are, what they are used for in science, and other relevant concepts such as hypothesis or model testing*
- *The fundamental origins of experimental uncertainty, e.g. shot and Johnson noise*
- *Examples throughout the course, from contemporary physics or famous discoveries, discussing experimental design, technique, and instrumentation as well as data analysis*

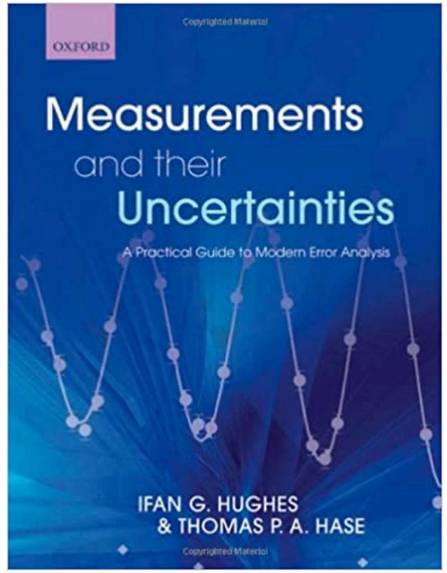
- *Uncertainties in measurements including dominant error, random and systematic errors, accuracy and precision, derivation and use of error propagation formulae assuming independent, normally distributed errors (i.e. no covariance matrices)*
- *Gaussian and Poisson distributions (as the distributions found in first year practicals), central limit theorem, least squares fitting (including weighted), meaning of residuals, confidence limits, simple statistical tests*

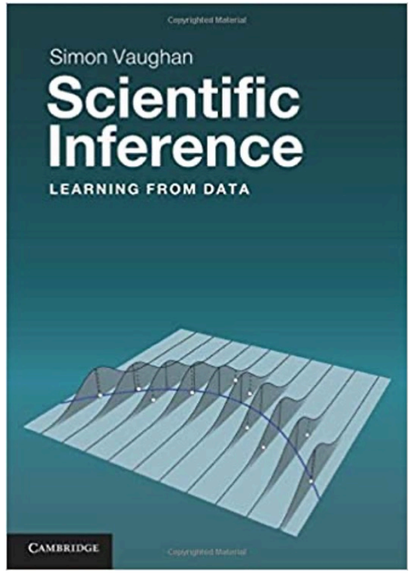
- *Practical implementation of the above in software and in data presentation, including*
    - *How to produce a plot (e.g. labels, units, title, expression of uncertainties etc.): see DA01 – Data Analysis practical*
    - *How to describe and present a numerical answer sensibly, with appropriate precision: used in practical course*
    - *Other useful skills in physics such as estimation. A pragmatic approach to data analysis should be stressed*

This first year material is taught through (this) short lecture course and the practical work. It is continually assessed through practical work (i.e. through log book assessment, lab reports, and discussion with demonstrators).

As is usual, knowledge of the first year syllabus will be assumed in the second year, where it can be examined and/or encountered in practicals.

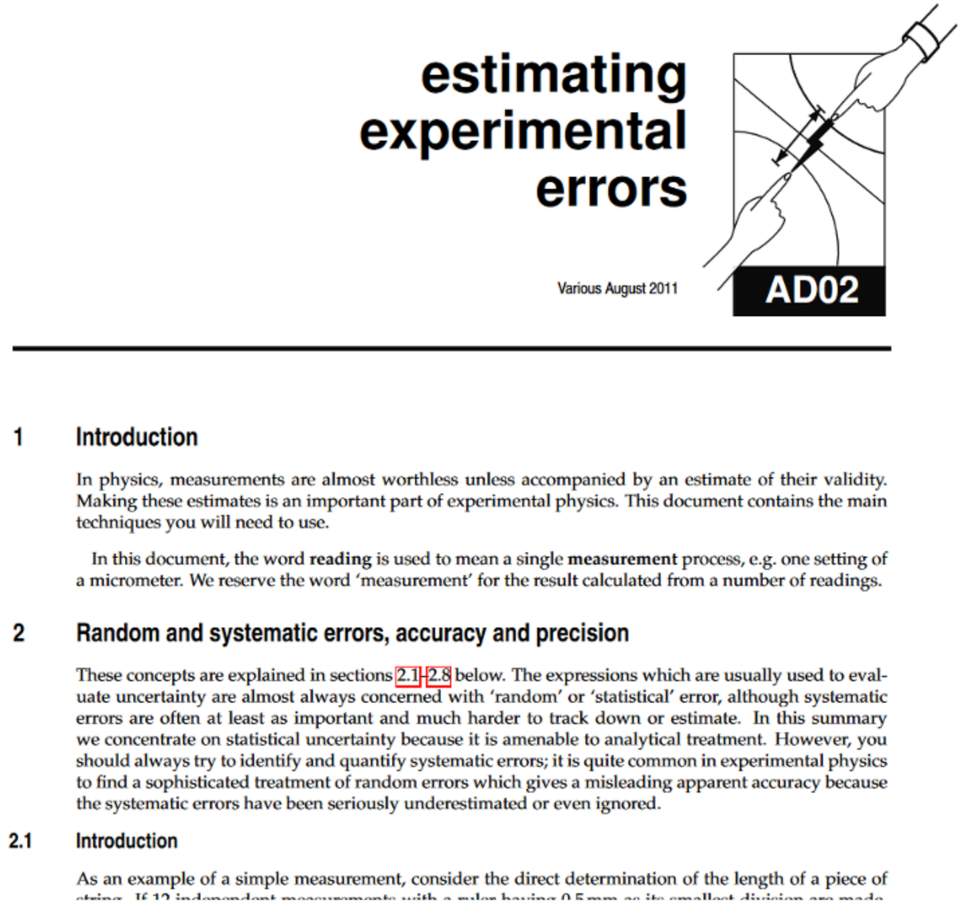
Recommended Books





Useful documents

AD02 and AD26



estimating experimental errors

Various August 2011

AD02

1 Introduction

In physics, measurements are almost worthless unless accompanied by an estimate of their validity. Making these estimates is an important part of experimental physics. This document contains the main techniques you will need to use.

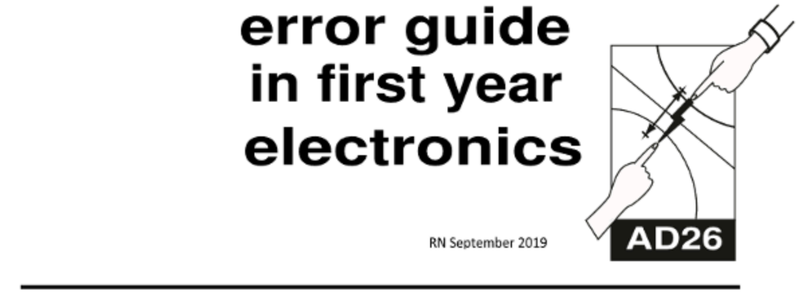
In this document, the word **reading** is used to mean a single **measurement** process, e.g. one setting of a micrometer. We reserve the word 'measurement' for the result calculated from a number of readings.

2 Random and systematic errors, accuracy and precision

These concepts are explained in sections 2.1–2.8 below. The expressions which are usually used to evaluate uncertainty are almost always concerned with 'random' or 'statistical' error, although systematic errors are often at least as important and much harder to track down or estimate. In this summary we concentrate on statistical uncertainty because it is amenable to analytical treatment. However, you should always try to identify and quantify systematic errors; it is quite common in experimental physics to find a sophisticated treatment of random errors which gives a misleading apparent accuracy because the systematic errors have been seriously underestimated or even ignored.

2.1 Introduction

As an example of a simple measurement, consider the direct determination of the length of a piece of



error guide in first year electronics

RN September 2019

AD26

AD02 gives a good description of error handling for labs and should be studied. This simplified treatment is for first year students with little or no experience using errors. It is a 'minimum knowledge' on errors and combining errors for first year electronics labs.

Summary

Nomenclature: True value A_0 is measured with result A and error $\pm\sigma_A$

Error types: Random (statistical, reading), Systematic (Calibration), Tolerance

Combining: $C = A \pm B \Rightarrow \quad \sigma_C^2 = \sigma_A^2 + \sigma_B^2$ these for random errors

$C = A \times or \div B \Rightarrow \quad \left(\frac{\sigma_C}{C}\right)^2 = \left(\frac{\sigma_A}{A}\right)^2 + \left(\frac{\sigma_B}{B}\right)^2$

$y = F(x) \Rightarrow \quad \sigma_y = \left.\frac{\partial F}{\partial x}\right|_{x_0} \sigma_x$

Results should take the form: $A \pm \sigma_{A1}(stat) \pm \sigma_{A2}(sys)$

Truncation Errors are: $\pm\sigma = \pm LSB/\sqrt{12}$

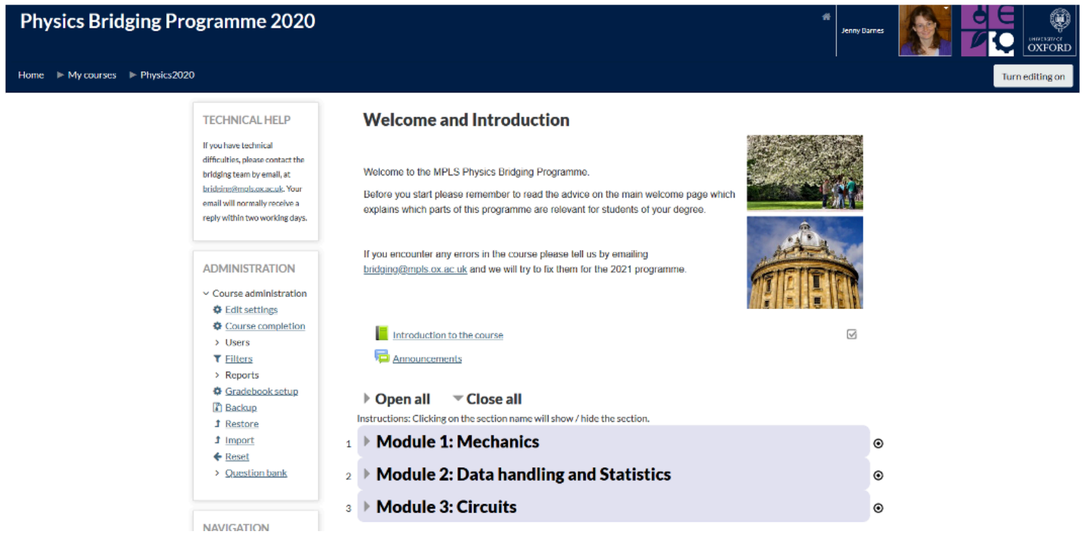
Content

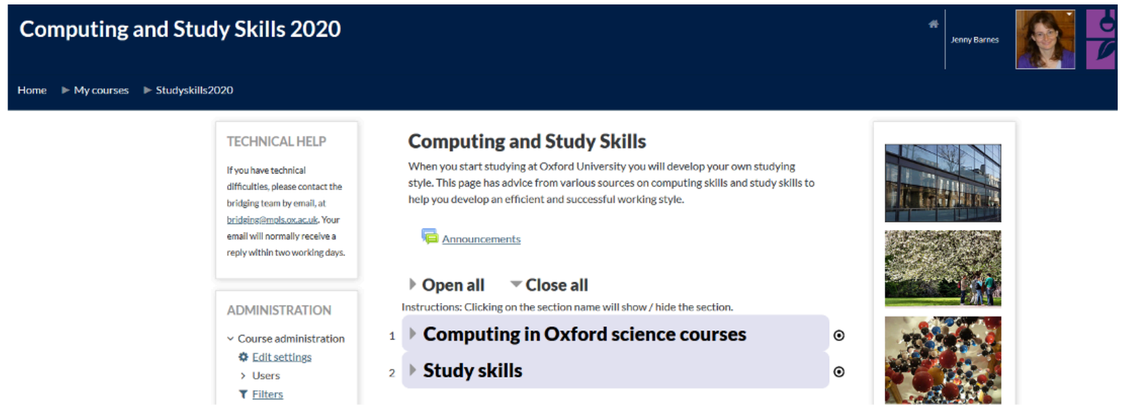
Material on MPLS bridging programme

Web address: https://mplsbridging.conted.ox.ac.uk/

You should have been sent a username and password over the summer, if not contact bridging@mpls.ox.ac.uk

.





Other resources

- Try the “More or Less” podcasts on Radio 4, available through BBC sounds app or at: https://www.bbc.co.uk/sounds/brand/b006qshd
- Online statistics resources can be found at: http://www.statstutor.ac.uk/
- A very helpful leaflet summarising lots of statistics formulae can be found at: https://www.statstutor.ac.uk/resources/uploaded/43799-prob-stats-ff-for-web.pdf

DA01 – Basic Data Analysis in Python

This is a practical which teaches you:

a) how to plot a graph using Python

b) how to fit some data and extract the fit parameters (including their errors)

c) how to record experimental things in your logbook

You need all of these skills in the other practicals this year.

It does not teach you everything to do with Python! You will learn the commands in Python you need for doing a) and b), but it isn't a full course in Python.

Other packages exist to plot and fit data (Excel, Origin, R etc.) but we think Python can go above and beyond some of these.

Guidelines for graphs

- Plot independent variable on x axis, dependent on y axis
- Use appropriate scales on axes
    - Consider whether the origin should be included
- Label axes including units
- Label the title of the graph
- Plot data points with their errors if you have them
- Fit the data using a computer package to extract the best fit parameters and their errors so you can compare to e.g. theory

Lecture 1

- What are experiments for?
- A result without an error is meaningless
- Definitions of accuracy and precision
- Types of errors and fundamental limits
- What do we quote after we take a set of measurements with random errors: **mean**, **standard deviation** and **standard error**

What are experiments for?

- Prove or disprove a theory
    - Michelson-Morley experiment disproved the existence of aether and led to special relativity being accepted
    - Gravitational waves detected at LIGO proved general relativity

What are experiments for?

- Prove or disprove a theory
    - Michelson-Morley experiment disproved the existence of aether and led to special relativity being accepted
    - Gravitational waves detected at LIGO proved general relativity

- Can we reproduce a result?
    - failure by others to repeat cold fusion measurements of Fleischmann and Pons (1989)
    - Do neutrinos travel faster than speed of light? (OPERA experiment, 2011)

- Discover new things
    - discovery of graphene
    - discovery of radio pulsars

- Discover new things
    - discovery of graphene
    - discovery of radio pulsars

- Find values of fundamental constants
    - measure speed of light c, gravitational constant G, charge e of an electron

Data sheet for Physics exams

(link from Physics Undergraduate Degree Canvas course, click on Examination papers)

7D (Version: 2018)
F3100 (Version: 2018)

HONOUR SCHOOL OF PHYSICS

HONOUR SCHOOL OF PHYSICS AND PHILOSOPHY

PRELIMINARY EXAMINATION IN PHYSICS

PRELIMINARY EXAMINATION IN PHYSICS AND PHILOSOPHY

SOME USEFUL DATA AND MATHEMATICAL FORMULAE

Fundamental constants

Constants are given to 4 or 5 significant figures (data from NIST CODATA 2014).

Electron rest mass	m_e	$9{\cdot}109 \times 10^{-31}\,\mathrm{kg}$
Proton rest mass	M_p	$1{\cdot}6726 \times 10^{-27}\,\mathrm{kg}$
Electronic charge	e	$1{\cdot}6022 \times 10^{-19}\,\mathrm{C}$
Speed of light in free space	c	$2{\cdot}9979 \times 10^{8}\,\mathrm{m\,s^{-1}}$
Planck's constant	h	$6{\cdot}626 \times 10^{-34}\,\mathrm{J\,s}$
	$h/2\pi = \hbar$	$1{\cdot}0546 \times 10^{-34}\,\mathrm{J\,s}$
	$\hbar c$	$197{\cdot}33\,\mathrm{MeV\,fm}$
Boltzmann's constant	k_B	$1{\cdot}3806 \times 10^{-23}\,\mathrm{J\,K^{-1}}$
Molar gas constant	R	$8{\cdot}314\,\mathrm{J\,mol^{-1}\,K^{-1}}$
Avogadro's number	N_A	$6{\cdot}022 \times 10^{23}\,\mathrm{mol^{-1}}$
Standard molar volume		$22{\cdot}414 \times 10^{-3}\,\mathrm{m^3\,mol^{-1}}$
Bohr magneton	μ_B	$9{\cdot}274 \times 10^{-24}\,\mathrm{A\,m^2}$ *or* $\mathrm{J\,T^{-1}}$
Nuclear magneton	μ_N	$5{\cdot}051 \times 10^{-27}\,\mathrm{A\,m^2}$ *or* $\mathrm{J\,T^{-1}}$
Bohr radius	a_0	$5{\cdot}292 \times 10^{-11}\,\mathrm{m}$
Fine structure constant	$e^2/(4\pi\varepsilon_0\hbar c) = \alpha$	$(137{\cdot}04)^{-1}$
Thomson cross section	σ_T	$6{\cdot}6525 \times 10^{-29}\,\mathrm{m^2}$
Compton wavelength of electron	$h/(m_e c) = \lambda_C$	$2{\cdot}4263 \times 10^{-12}\,\mathrm{m}$
Rydberg's constant	R_∞	$1{\cdot}0974 \times 10^{7}\,\mathrm{m^{-1}}$
	$R_\infty hc$	$13{\cdot}606\,\mathrm{eV}$
Stefan's constant	σ	$5{\cdot}6704 \times 10^{-8}\,\mathrm{W\,m^{-2}\,K^{-4}}$
Gravitational constant	G	$6{\cdot}6741 \times 10^{-11}\,\mathrm{N\,m^2\,kg^{-2}}$

A result without an error is meaningless

- All measured fundamental constants have an error in them (e.g. from [http://physics.nist.gov/cuu/Constants]):

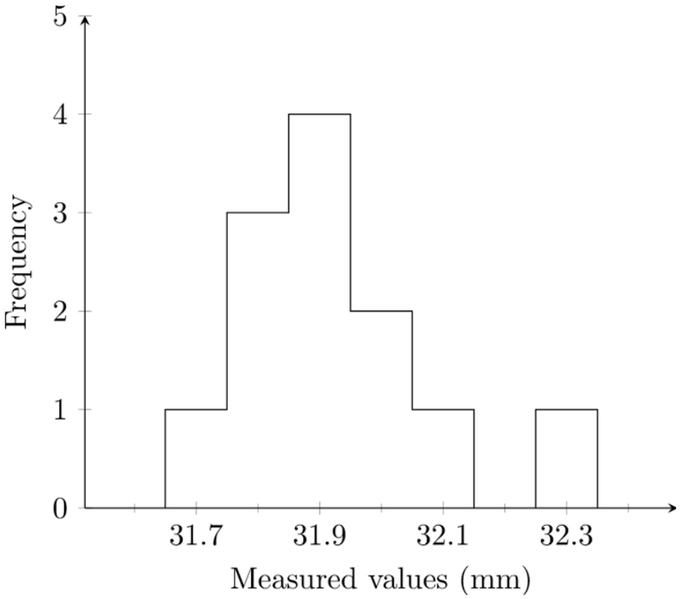
$13.605693122994 \pm 0.000000000026$ eV or

$13.605693122994(26)$ eV

- This means that we expect the value to lie between 13.605693122968 eV and 13.605693123020 eV
- This range is related to the **confidence** level of the parameter value; more on that later.

Measuring the length of a piece of string

- Suppose you repeatedly measured a piece of string with a ruler and plotted the results
- Ruler has 0.5 mm divisions
- You estimate by eye to 0.1 mm
- Resulting numbers all different
- Observe peak around one value



Some definitions

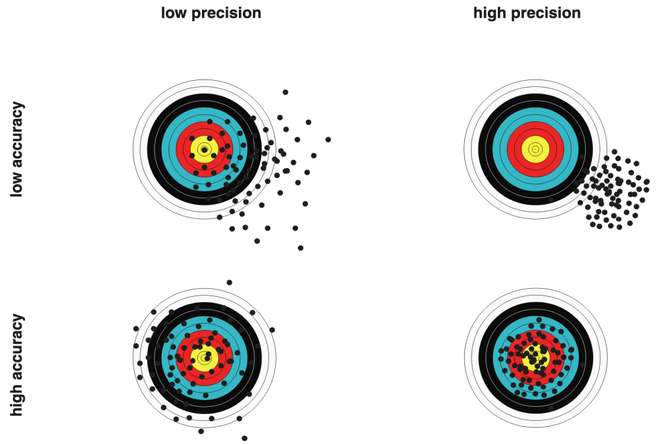
Accuracy: The difference between the result and the true value; an accurate measurement is in agreement with the "accepted" value

Precision: The least distinguishable change, given by the number of digits or the spread in the measurements (precise = small spread)

Some definitions

Accuracy: The difference between the result and the true value; an accurate measurement is in agreement with the "accepted" value

Precision: The least distinguishable change, given by the number of digits or the spread in the measurements (precise = small spread)



Types of error (1)

Random errors: error associated with limitations of the measuring device, they prevent observer from obtaining same value each time they measure. Measurements usually look like a peak centered around a value. Example: see previous graph. $\Longrightarrow$ Reduce by taking more measurements.

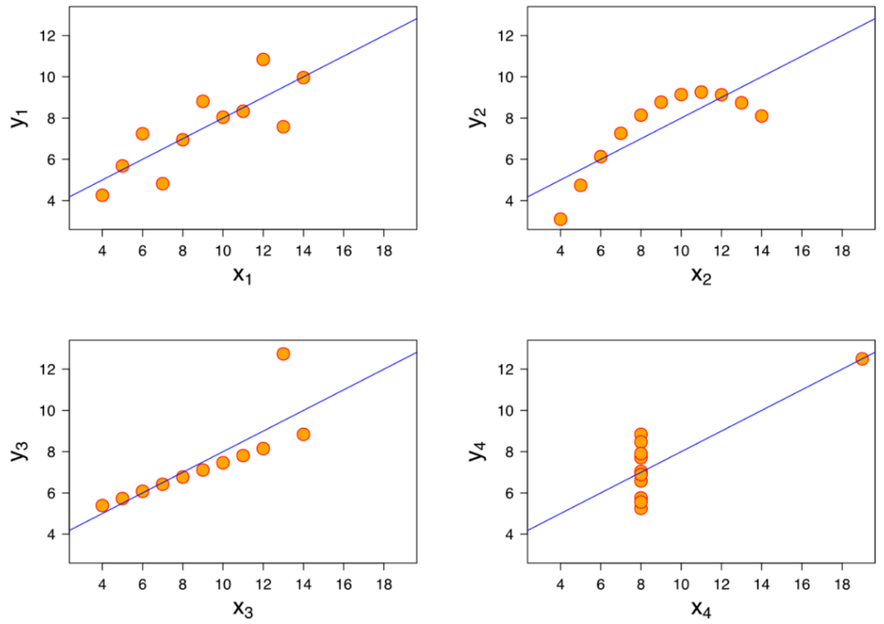
Mistakes: e.g. wrote the number 3.75 as 3.57, typed number in wrong to computer, misread scale. Plotting data can show up points which “don’t fit” and these can then be discarded or perhaps measured again. Example: the point at 32.3mm

Anscombe’s quartet

Anscombe's quartet (1973): frequently quoted standard examples for simple regression

- Four different datasets
- Fitting a line: identical regression results
- Graph 1: ideal -- Graph 2: a line is a bad choice
- Graphs 3 and 4 have obvious outliers

$\Longrightarrow$ how meaningful are the fitted lines?



To quote Anscombe: "Most of these [statistical packages for regression] originated in the pre-visual era. The user is not showered with graphical displays. He can get them only with trouble, cunning and a fighting spirit. **It's time that was changed.**"

$\Longrightarrow$ **really helpful to plot data roughly as you go along!**

Types of error (2)

Systematic errors: error which always has the same sign so shifts the centre of the peak one way or another but not both. Example: ruler calibrated at one temperature but we are using it at a higher temperature and ruler has expanded. $\Longrightarrow$ Difficult to detect, have to try and design experiment to avoid them.

Tolerances: Engineering tolerances on mechanical and electrical components. Don't confuse with random error. For discussion, see aforementioned AD02 and AD26 scripts.

Which types of errors dominate?

- The error which has the largest influence on your result is called the **dominant error**
- Part of your job as a physicist (especially when carrying out experiments) is to determine which is the dominant error
    - $\Longrightarrow$ reduce it as much as possible!
    - Caveat: what happens when this error becomes subdominant?
- Important when propagating errors in formulae -- if one value has an error much bigger than all the rest or appears as a power in a formula it may dominate any error calculation.

More about propagating errors later!

Fundamental limits

Some uncertainties can never be eliminated:

Shot noise: discrete nature of certain quantities, e.g. measuring a feeble light beam start to get random fluctuations of intensity because light is made up of photons (finite energy bundles)

Johnson noise: arises in a resistor as a consequence of the thermal agitation of the charge carriers. Becomes important in experiments measuring low voltages -- so might cool down experimental setup to reduce this but cannot eliminate.

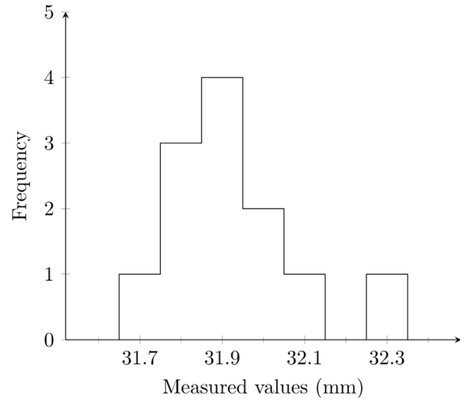
$\Longrightarrow$ The best you will ever be able to do is get down to these limits, but you are often way above this.

Going back to our piece of string...

So what answer would we quote?

Values were:

31.7mm	31.8mm (x 3)	31.9mm (x 4)	32.0mm (x 2)	32.1mm	32.3mm (discard this point)

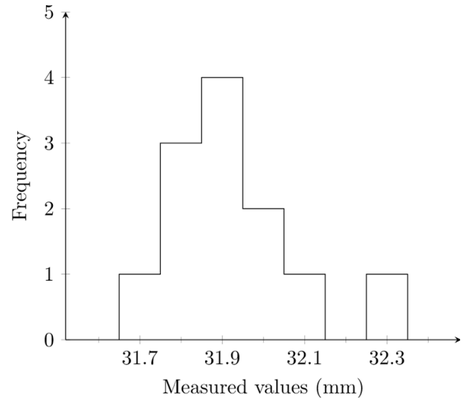


Going back to our piece of string...

So what answer would we quote?

Values were:

31.7mm	31.8mm (x 3)	31.9mm (x 4)	32.0mm (x 2)	32.1mm	32.3mm (discard this point)



Mean: for a set of N measurements, each with value x_i, the mean $\bar{x}$ is given by:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_N}{N}$$

$\Longrightarrow$ This is the result!

$$\implies \text{so } \bar{x} = \frac{31.7 + 31.8 \cdot 3 + 31.9 \cdot 4 + 32.0 \cdot 2 + 32.1}{11} \text{ mm} = 31.89090909' \text{ mm}$$

- How many significant figures should we quote? Depends on the error in this value.

How do we find the error in the mean? (1)

We will need the **standard error** which we derive from the standard deviation:

Sample standard deviation:

$$\sigma_{N-1} = \sqrt{\frac{\sum(x_i - \bar{x})^2}{N-1}}$$

- We are squaring the differences between each point and the mean -- if we didn't square it some of the terms would be positive, some negative, and they cancel out
- We use $N-1$ as we use N to calculate the mean, but are now using the mean in this formula (we have reduced the number of independent values or degrees of freedom)

How do we find the error in the mean? (2)

$$\sigma_{N-1} = \sqrt{\frac{\sum(x_i - \bar{x})^2}{N-1}}$$

- Definitions of standard deviation (sample, population, ...) can vary! -- see lecture 2
- The square of the standard deviation (σ^2) is called the **variance**.

Standard error: the standard deviation of the mean. $\implies$ This represents the random error!

$$\alpha = \frac{\sigma_{N-1}}{\sqrt{N}} = \sqrt{\frac{\sum(x_i - \bar{x})^2}{N(N-1)}}$$

$\Longrightarrow$ As N increases, α decreases! So take more readings to reduce your standard error α!

NB: the estimated standard deviation σ_{N-1} will typically **not** reduce with increasing N!

Going back to our piece of string...

$$\alpha = \sqrt{\frac{\sum(x_i - \bar{x})^2}{N(N-1)}} = \sqrt{\frac{(31.7 - 31.89090909')^2 + \ldots + (32.1 - 31.89090909')^2}{11 \cdot (11-1)}}\ \text{mm}$$

$$\implies \alpha = 0.03425717\ \text{mm}$$

So what do we quote as the result for the length of the string?

$$\implies \bar{x} \pm \alpha = (31.89090909 \pm 0.03425717)\ \text{mm}$$

Going back to our piece of string...

$$\alpha = \sqrt{\frac{\sum(x_i - \bar{x})^2}{N(N-1)}} = \sqrt{\frac{(31.7 - 31.89090909')^2 + \ldots + (32.1 - 31.89090909')^2}{11 \cdot (11-1)}} \text{ mm}$$

$$\implies \alpha = 0.03425717 \text{ mm}$$

So what do we quote as the result for the length of the string?

$$\implies \bar{x} \pm \alpha = (31.89090909 \pm 0.03425717) \text{ mm}$$

What about the number of significant figures? Only quote as many in the mean as is sensible given the error (typically quoted to **1 significant figure**):

$$(31.89 \pm 0.03) \text{ mm} \qquad \text{or} \qquad 31.89(3) \text{ mm}$$

Don't forget the units!

A slight caveat

- Sometimes you don't get such a spread of numbers, e.g. when always getting the same value while measuring with some digital calipers or a digital multimeter
- The error in the value is then not given by the formulae on previous slides but related to the resolution of the measuring device (the **tolerance!**)
    - $\Longrightarrow$ Quote this as the error instead!

Standard error visualisation (1)

Python programme for example from p. 14-16 Hughes & Hase:

In [1]:
```
# Programme illustrating the standard error and standard deviation
import numpy as np

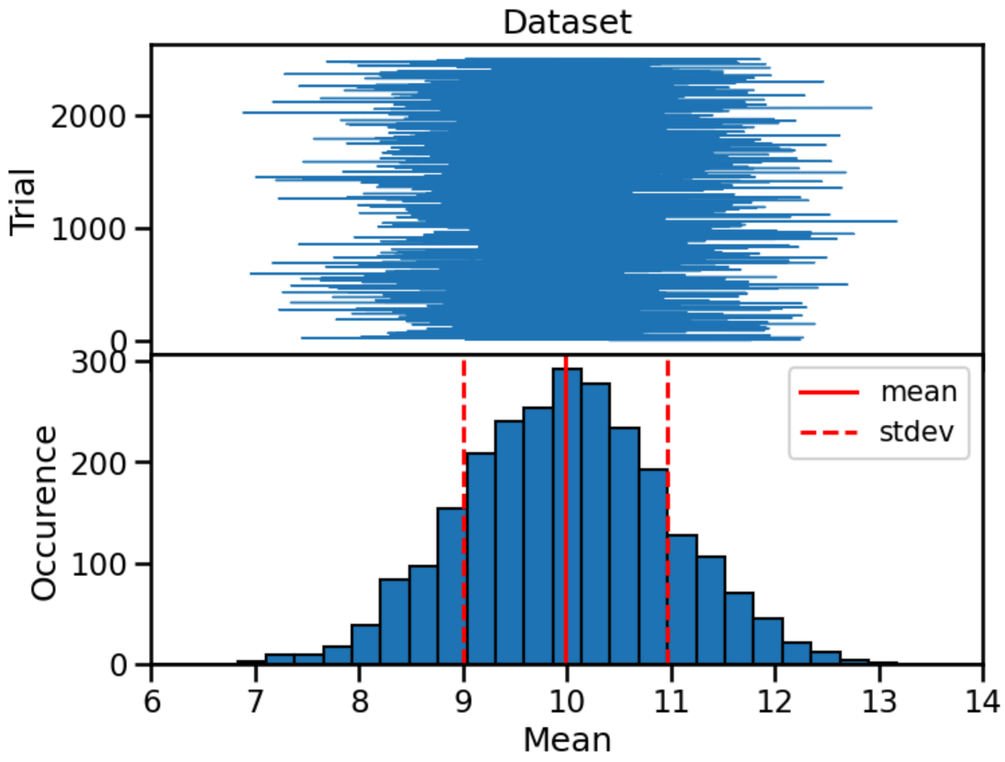
mu = 10
sig = 1

# initialise random number generator:
np.random.seed(0)
# create dataset by drawing from a random distribution:
dataset = np.random.normal(mu, sig, 2500)
```

$\Longrightarrow$ Simulate 2500 measurements subject to random errors. Observe their distribution!

Standard error visualisation (2)

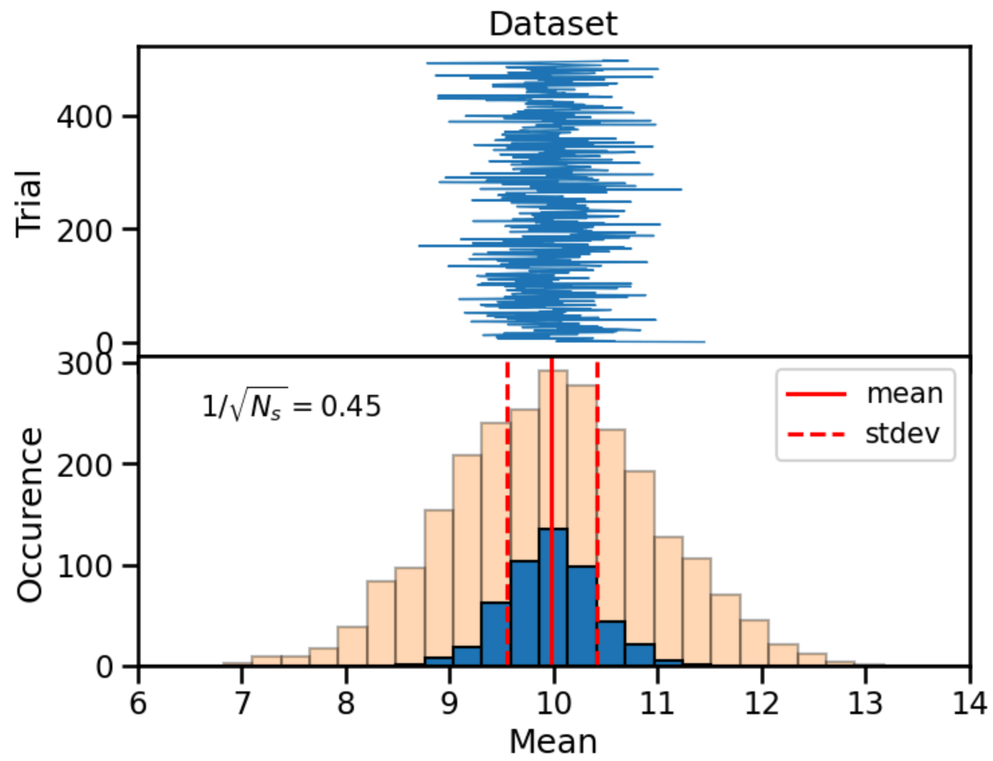
In [2]:
```
from config import plot_dataset
plot_dataset(dataset);
```



Standard error visualisation (3)

$\Longrightarrow$ Partition into samples of size `N_per_sample`. Observe the distribution of means.

In [3]:
```
plot_dataset(dataset, N_per_sample=5);
```



Standard error visualisation (4)

$\Longrightarrow$ The distribution of means becomes narrower with increasing sample size N as expected (increased precision!):

$$\alpha = \frac{\sigma_{N-1}}{\sqrt{N}}$$

I.e. taking more measurements for a sample gives more confidence in the mean, the confidence interval $\bar{x} \pm \alpha$ shrinks!

NB: The estimated standard deviation σ_{N-1} of the original raw dataset would not change with more measurements!

Summary of lecture 1

- What are experiments for?
- A result without an error is meaningless
- Definitions of accuracy and precision
- Types of errors and fundamental limits
- Quote result of a set of measurements with random errors as: $\bar{x} \pm \alpha$, where
  - **Mean**: $\bar{x} = \dfrac{x_1 + x_2 + \ldots + x_N}{N}$
  - **Standard error**: $\alpha = \sqrt{\dfrac{\sum (x_i - \bar{x})^2}{N(N-1)}}$

Questions?