PHYS465: Statistical Data Analysis in Physics

Week 1: Introduction, Model fitting

Dr. Mathew Smith

mat.smith@lancaster.ac.uk

Physics Building; C46

General information

This is the second year that this module has been taught

Assessment is 100% by coursework, which is built on workshop exercises

Comments/suggestions are very welcome

Office hour: C46 Physics

Thursday @ 1-2pm

Open door policy

In person in my office (Physics: C46)

Online via Teams/email (mat.smith@lancaster.ac.uk)

Course Structure

Weekly lectures

Thursday @ 5pm

Introduce key statistical concepts

Weekly workshops

Friday @ 9am

Problem sheets introducing key python libraries, with practical examples

Coursework problem sheets will extend this knowledge

Feedback sessions

Work through coursework solutions

Moodle quizzes based on lecture content

Assessment: key dates

Coursework deadlines:

Tuesday 20th Jan @ 4pm : week 11 content

20% of overall grade; feedback session on Fri 23rd Jan

Tuesday 27th Jan @ 4pm : week 12 content

20% ; feedback session Fri 30th Jan

Tuesday 3rd Feb @ 4pm : week 13 content

20% ; feedback session Fri 6th Feb

Tuesday 24th Feb @ 4pm : week 14 and 15 content

40% ; summative assessment.

NB: There is an additional week to complete this worksheet

Submission through Moodle:

Computer code *and* interpretative summary : see Friday / Moodle for details

Additional independent, investigative work is expected

What is the aim of this module?

This course aims to introduce and provide you with experience in using the key techniques used to analyse datasets in physics.

- All of these techniques are transferable!
- The focus of the module is to develop *practical* skills.
  - the theory behind the statistical concepts are complex
  - key concepts will be introduced.
  - additional reading is available to develop fundamental understanding

Module Structure

'*model testing*': does our data agree with our model?

Week 11: Fitting a model to data; estimating parameters

Week 12: Hypothesis testing; the likelihood; estimating uncertainties

Week 13: Posterior sampling; Bayesian statistics

'*data driven*': what does our data tell us?

Week 14: Clustering and Classification algorithms

Week 15: Machine learning techniques

Setting the scene : what is data analysis?

The process of analysing experimental data to validate (or disfavour) a hypothesis or theory

- The experimental data, and its uncertainties, have already been collected

Requires the application of statistical tools

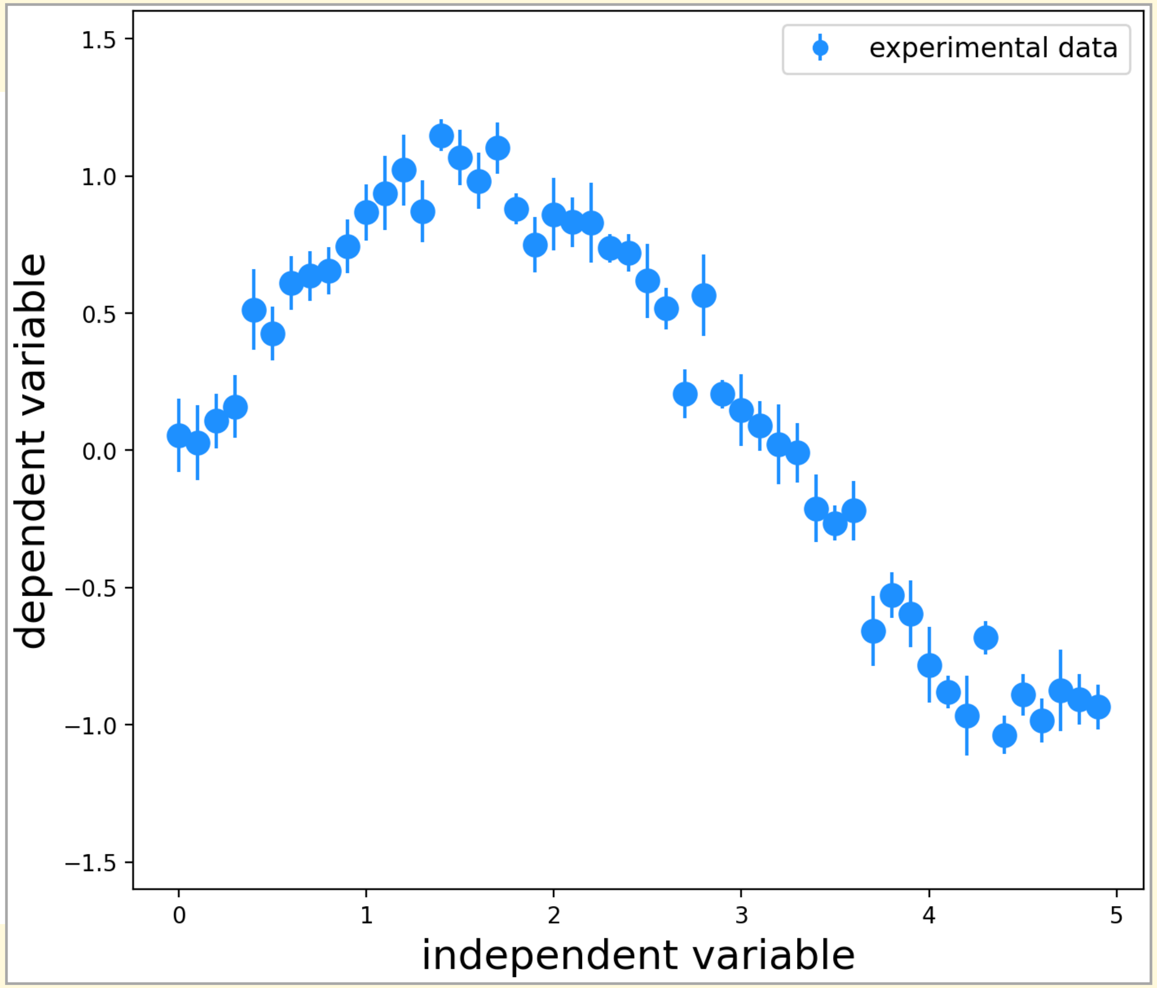
- This course will introduce the main concepts and statistical tests

The relevance of Physics:

- Physics (in particular astro and particle) involve the collection of extremely large sets of data
- Requires complex analysis techniques

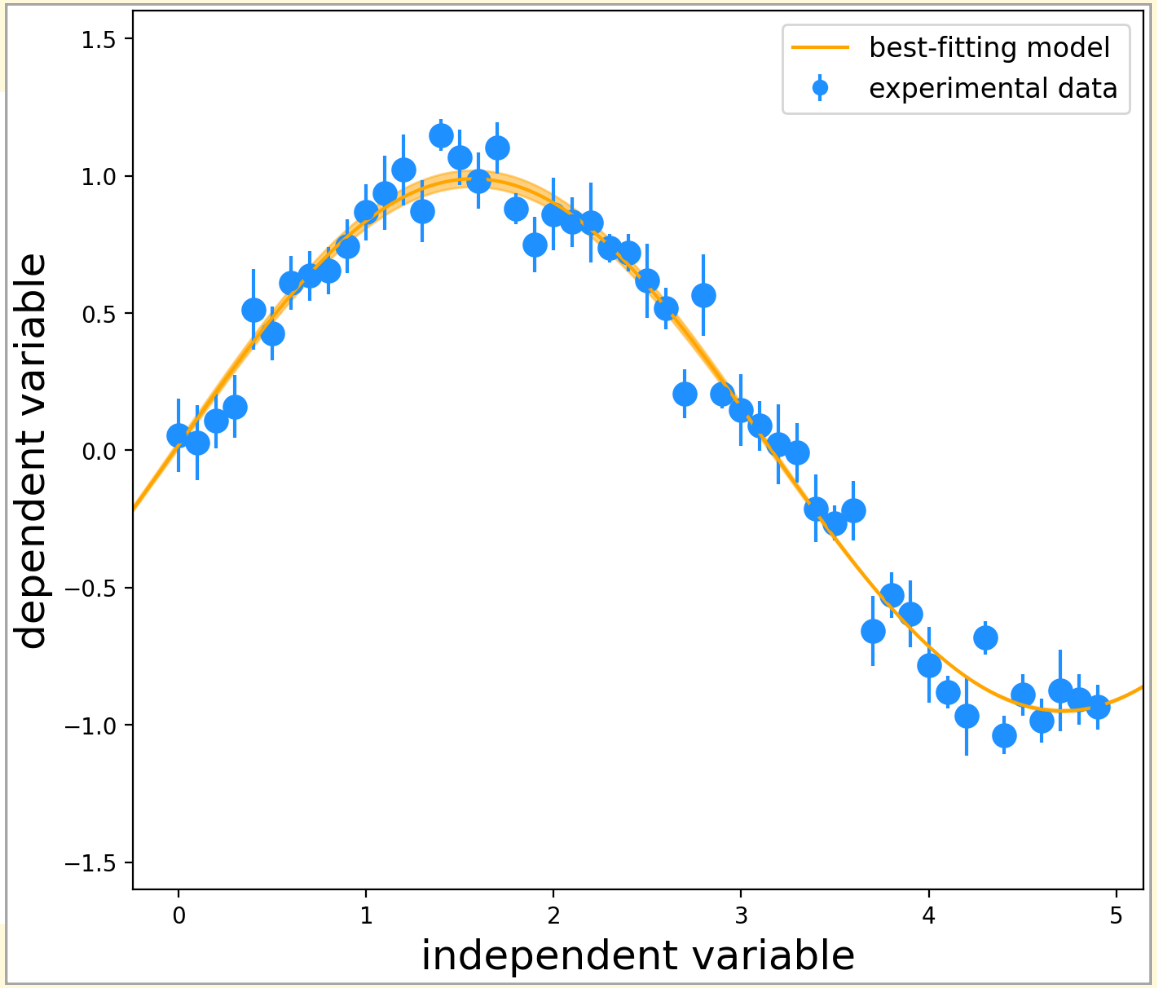
Week 11: Explaining data with models

(1) Model fitting, (2) parameter estimation and (3) hypothesis testing



Week 11: Explaining data with models

(1) Model fitting, (2) parameter estimation and (3) hypothesis testing



(1) Model fitting

Given some experimental data (and measured uncertainties) compare to a defined model

Use data to deduce the relevant laws (parameters) for our experiment.

Two main foci:

- **Parameter estimation:** Determine the numerical value of a physical quantity
- **Hypothesis testing:** Test whether a theory is consistent with our measured data

(2) Parameter estimation

Use the data to determine a model parameter AND its associated uncertainty in an efficient and unbiased manner.

i.e. Use data to calculate/obtain a value for a free (unknown) parameter

- e.g. given a set of astrophysical distances what is the amount of matter in the Universe?

—

Unbiased = the planned method will, on average, give the correct result

Efficient = Matching the experimental data and model to the analysis method.

Intricate and expensive methods are only necessary for complex models and datasets.

(3) Hypothesis testing

Determine whether our data is consistent with a specific hypothesis

Is the data we obtain in our experiment consistent with a given theory?

- e.g. how does the measured energy spectrum compare with the prediction

—

Does not take the form of a simple “yes/no” answer.

Answer will be yes or no accompanied by a statement of confidence.

Given multiple models, determine the model that best describes our measurements

In the coming weeks we will explore several methods for hypothesis testing and parameter estimation.

Getting started: What is a measurement?

Repeating an experiment doesn't always give the same result. Variation in the experiment will produce a distribution of answers.

$$x = [26, 24, 26, 28, 23, 24, 25, 24, 26, 25]$$

In previous years, you have learnt that multiple measurements can be summarised through:

$$\hat{x} = \frac{\sum_i^N x_i}{N} \qquad \sigma_x = \sqrt{\frac{\sum_i^N (x_i - \hat{x})^2}{N}} \qquad \text{se} = \frac{\sigma_x}{\sqrt{N}}$$

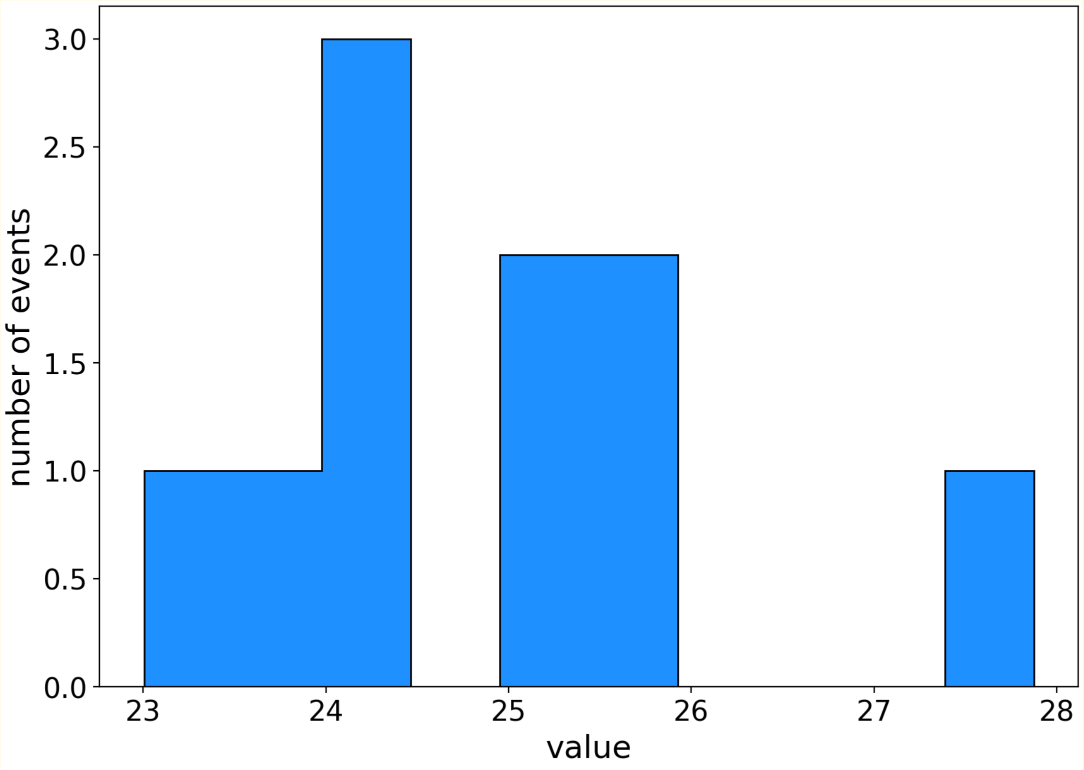
mean: 'most likely value' — standard deviation: 'dispersion' — standard error on mean

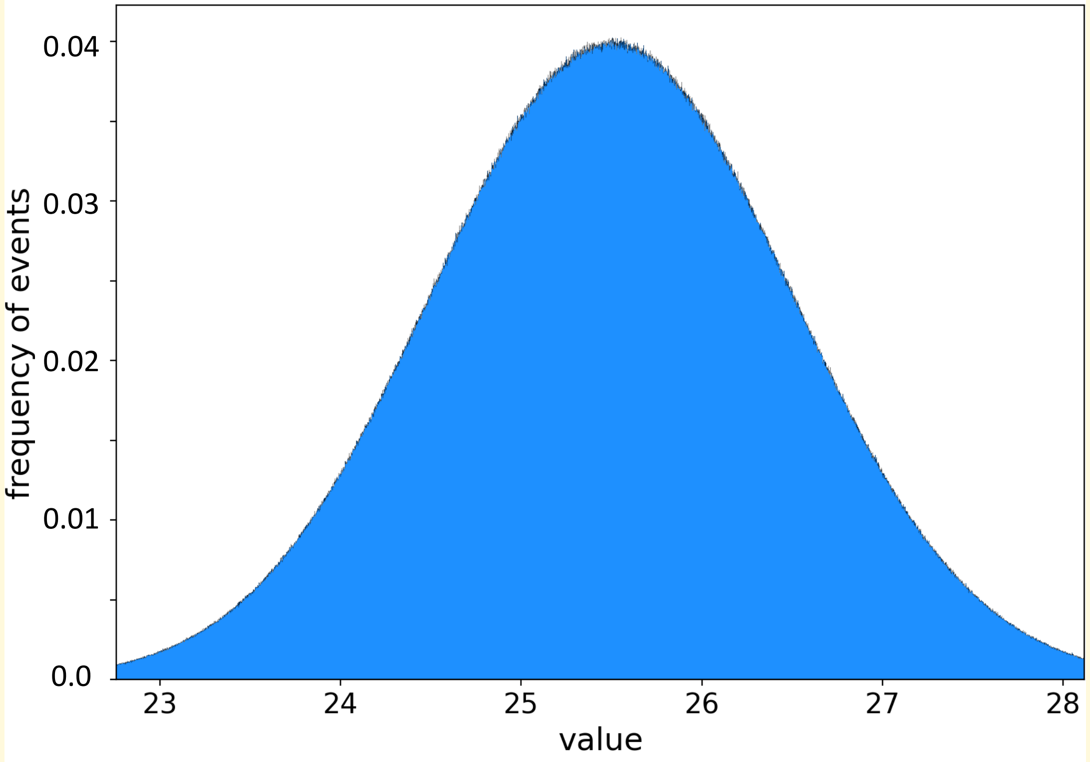
As we are testing physical systems there is a ground-truth. The values we measure are drawn from an underlying **distribution.**

There are many different distributions (see next week) depending on what is being tested and how

Distributions

In physics, values and functions are rarely discrete, they are continuous





counts = number of measurements in each bin

area under the curve sums to total(counts)

frequency = fraction of measurements in each bin

area under the curve sums to 1

Probability density function (pdf)

Consider a continuous function $f(x)$

The probability density function is defined as the probability that the variate has a given value x .

This is often expressed as the integral between two points

$$\boxed{\int_a^b f(x)\,dx = P\left[a \le x \le b\right]}$$

A normalised histogram approaches the PDF

when the variable is continuous

The function must be normalised such that

$$\boxed{\int_{-\infty}^{\infty} f(x)\,dx = 1}$$

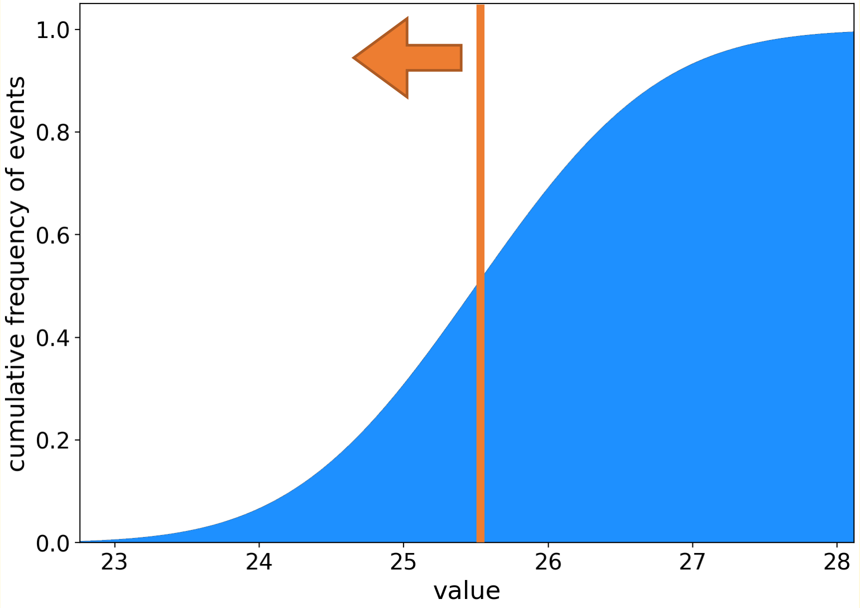
We use the PDF to estimate the probability

that a variable falls in a given range

Cumulative density function (cdf)

Tells us the percentile that a parameter value represents

$$F(x') = \int_{-\infty}^{x'} f(x)\,dx$$



i.e. $F(x') = 0.6$ means that 60% of the probability lies $\leq x'$

So the CDF returns the expected probability of observing a value less than or equal to the given value

Expectation values

Discrete distribution of variable x:

$$E(x) = \sum_{i=0}^{n} x_i P(x_i)$$

where $P(x_i)$ is the probability that x has the value x_i

Continuous distribution of variable x:

$$E(x) = \int_{-\infty}^{+\infty} x f(x)\, \mathrm{dx}$$

where $f(x)$ is the probability density function

This is the formal definition for the mean as $N \to \infty$

Gaussian distribution

aka a normal distribution

$$P(x) = \frac{1}{\sigma\sqrt{(2\pi)}} \exp\{-(x-\mu)^2/2\sigma^2\}$$

Expectation value: $E(x) = \mu$

Standard deviation: $\sigma_x = \sigma$

A Gaussian distribution is the "high-N" limit for the Binomial and Poisson distributions (see week 12).

<u>The central limit theorem:</u> states if the average is taken of variables drawn many times for ANY probability distributions, the resulting average will follow a Gaussian.

In practice, if we take a lot of data (N>30), our sample will resemble a normal distribution.

(2). Parameter estimation : *least squares*

NB: for straight-line model ($y = A + Bx$), this is known as 'linear regression'

For one measured data point (x_i, y_i) that is drawn from a Gaussian distribution:

$$p_{A,B}(y_i) \propto \frac{1}{\sigma_y} \exp\{-(y_i - A - Bx_i)^2/2\sigma_y^2\}$$

Over the entire data set:

$$p_{A,B}(y_1, y_2, \ldots, y_N) \propto \frac{1}{\sigma_y^N} \exp\{-\chi^2/2\} \qquad \chi^2 = \sum_{i=1}^{N} \frac{(y_i - A - Bx_i)^2}{\sigma_y^2}$$

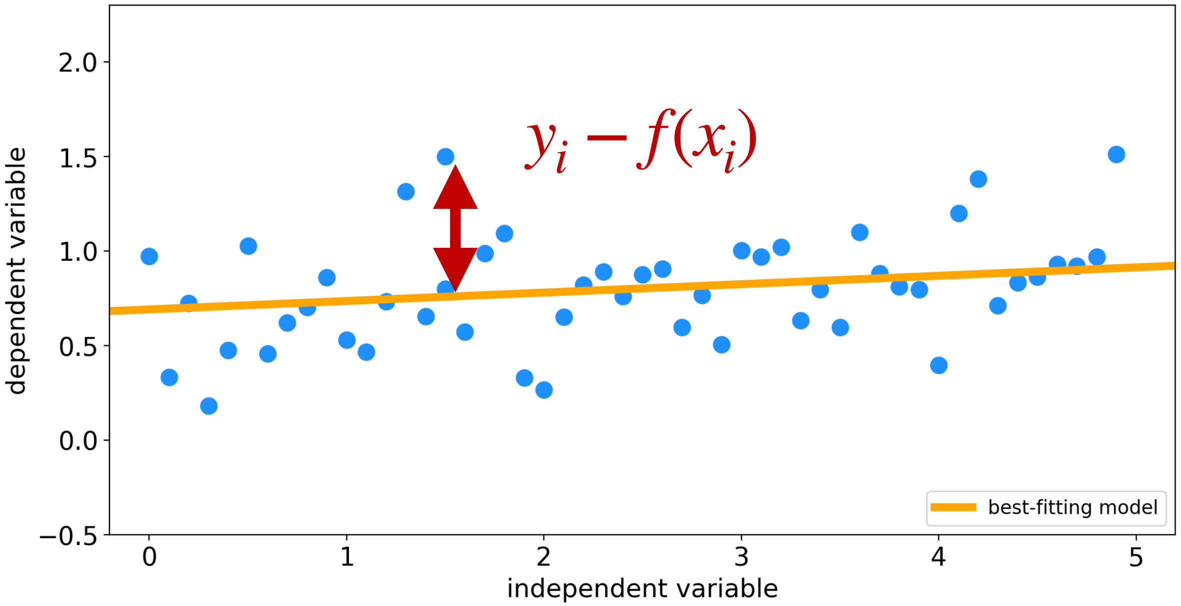
'Least-squares' : $p_{A,B}$ is maximised when χ^2 is smallest

Linear equation:

$$A = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{N \sum x_i^2 - (\sum x_i)^2} \qquad B = \frac{N \sum x_i y_i - \sum x_i \sum y_i}{N \sum x_i^2 - (\sum x_i)^2}$$

(2). Parameter estimation : *least squares*

In practice:



Over the entire data set:

$$p_{A,B}(y_1, y_2, \ldots, y_N) \propto \frac{1}{\sigma_y^N} \exp\{-\chi^2/2\} \qquad \chi^2 = \sum_{i=1}^{N} \frac{(y_i - A - Bx_i)^2}{\sigma_y^2}$$

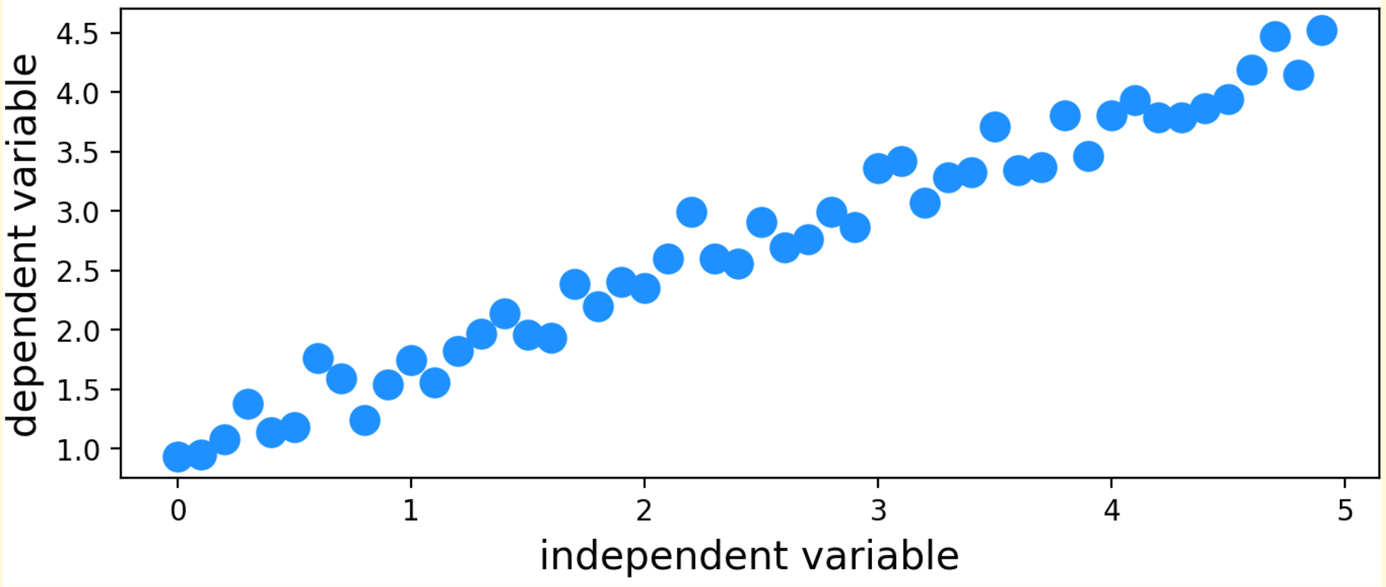
'Least-squares' : $p_{A,B}$ is maximised when χ^2 is smallest

Considering uncertainties

The least-squares formula does not consider uncertainties

Most sciences do not measure uncertainties : e.g. population studies, medical diagnoses, clime science

uncertainties can be inferred through the variance of the data



Most physics experiments do have uncertainties: Identical for every point (e.g. systematic): *homoscedastic*

(see Week 12) Different for every point (e.g. statistical): *heteroscedastic*

Including uncertainties

Uncertainties can be re-purposed as weights: $w_i = \dfrac{1}{\sigma_i^2}$ "inverse variance"

Minimising the χ^2 : $\chi^2 = \sum_{i=1}^{N} \dfrac{(y_i - A - Bx_i)^2}{\sigma_i^2}$ or $\boxed{\chi^2 = \sum_{i=1}^{N} \dfrac{(y_i - f(x_i))^2}{\sigma_i^2}}$

This is a solved problem for linear relationships : for complex functions we need to use minimising techniques

$$A = \frac{\sum w_i x_i^2 \sum w_i y_i - \sum w_i x_i \sum w_i x_i y_i}{\sum w_i \sum w_i x_i^2 - (\sum w_i x_i)^2} \qquad B = \frac{\sum w_i \sum w_i x_i y_i - \sum w_i x_i \sum w_i y_i}{\sum w_i \sum w_i x_i^2 - (\sum w_i x_i)^2}$$

$$\sigma_A = \sqrt{\frac{\sum w_i x_i^2}{\sum w_i \sum w_i x_i^2 - (\sum w_i x_i)^2}} \qquad \sigma_B = \sqrt{\frac{\sum w_i}{\sum w_i \sum w_i x_i^2 - (\sum w_i x_i)^2}}$$

NB: Plenty of minimisers are available in Python

(3) Hypothesis testing

How good is the model?

Fundamental question: Is the model that we are testing a good fit to our measured data?

Is there a more likely alternative?

Null Hypothesis: Our starting assumption

Can be either parameter values or choice of model

In practice: Compares our assumption (or prediction) with experimental measurements
e.g. does the luminosity distribution of galaxies match a Schechter function?

Outcome: Make a statement on the probability of obtaining our result (see confidence level slides).

Plenty more on hypothesis testing next week

Goodness-of-fit: Reduced χ^2 test

To determine if our fitted model is a good match to the data we apply a χ^2-test (NB: slide 23):

$$\bar{\chi}^2_{\text{red}} = \frac{1}{N_{\text{dof}}}\chi^2 = \frac{1}{N_{\text{dof}}}\sum_{i=1}^{N}\frac{(y_i - f(x_i))^2}{\sigma_y^2}$$

$$N_{\text{dof}} = N_{\text{data}} - N_{\text{params}}$$

Recall (slide 21) that to determine the best-fit in the least-squares process we minimised χ^2

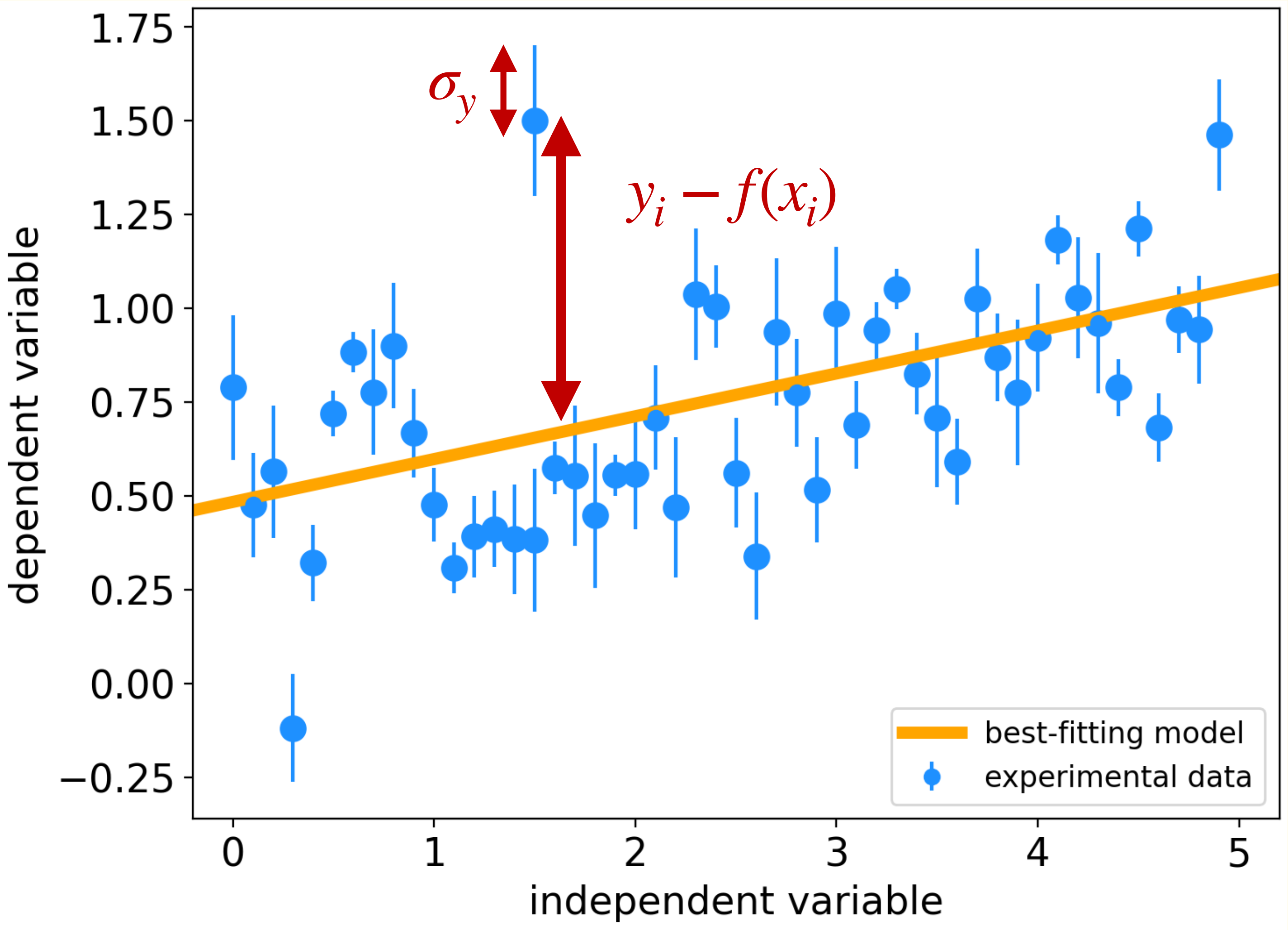
Interpretation: *If* our data points are drawn from the model and follow a normal distribution

Expect: $\chi^2 \sim 1$

$\bar{\chi}^2 > 1$ evidence that our data are NOT drawn from the model (see lookup tables for probabilities)

$\bar{\chi}^2 < 1$ evidence that the model has too much freedom

The χ^2 test : practical implications



$$\bar{\chi}^2_{\text{red}} = \frac{1}{N_{\text{dof}}} \sum_{i=1}^{N} \frac{(y_i - f(x_i))^2}{\sigma_y^2}$$

A measure of the averaged normalised distance to the best-fit model

Parameter Estimation: *Confidence interval*

The range of parameter values that are plausible given our dataset

Confidence interval:

For a value of a parameter θ estimated from a continuous (discrete) random variable x, the confidence interval is a member of a set of intervals $[\theta_1, \theta_2]$ such that (at least) a fraction $1-\alpha$ of them contains the true value of θ.

If we repeat an experiment millions of times, this will result in X% CI, X% of the time

Note:

$1-\alpha$ is the confidence level. Typical values are 68%, 90%, 95%.

The set of intervals is ideally obtained by repeating the same experiment.

θ_1, θ_2 are functions of x.

The interval may not contain the true value of the parameter: the probability $1-\alpha$ refers to the estimation procedure, not the specific interval.

Confidence Interval

Conventional choice: 68%, which is that defined by $\pm\theta$. This corresponds to

$$\hat{p} - \delta p \leq P_0 \leq \hat{p} + \delta p$$

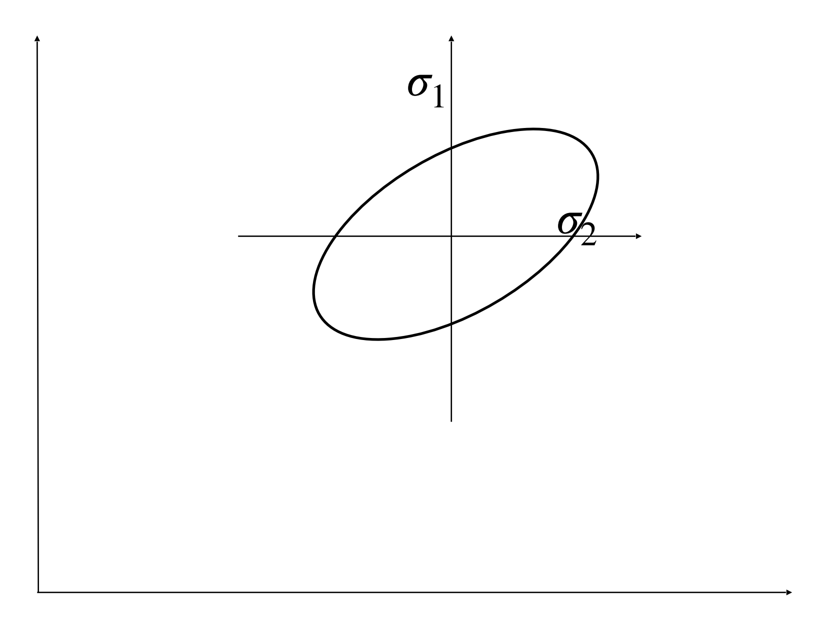
Interpretation: This is a confidence interval (range) for P_0. It indicates how often we expect to include P_0 within our quoted range for a repeated series of experiments.

Confidence levels – more than one variable

Consider a function of two variables A and B

Errors on variables A and B can be used to define an error ellipse.

Confidence region is calculated such that if a set of measurements were repeated many times and the confidence region calculated in the manner for each set of measurements, then a certain percentage of the time the confidence region would include the point representing the true values of the set of variables being estimated.



Confidence levels and sigma

Commonly used values:

1σ : area bounded from -1σ to $+1\sigma$
contains **68.3%** of the probability

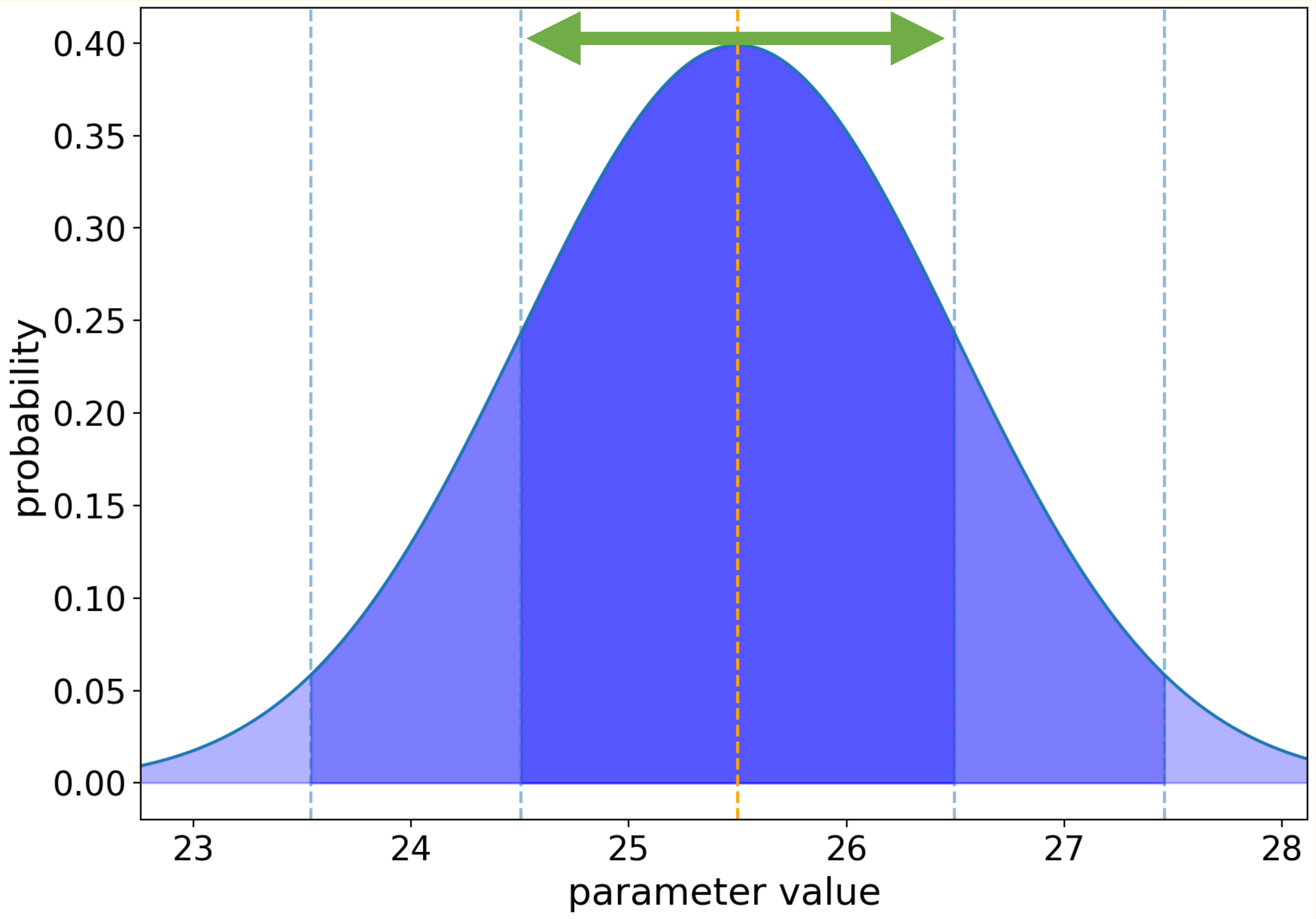
2σ : area bounded from -2σ to $+2\sigma$
contains **95.5%** of the probability

3σ : area bounded from -3σ to $+3\sigma$
contains **99.7%** of the probability

5σ : "discovery threshold" : 99.99994%!

Area bounded from $-\infty$ to 1.28σ is 90%

area containing 68% of the probability



Confidence levels & χ^2

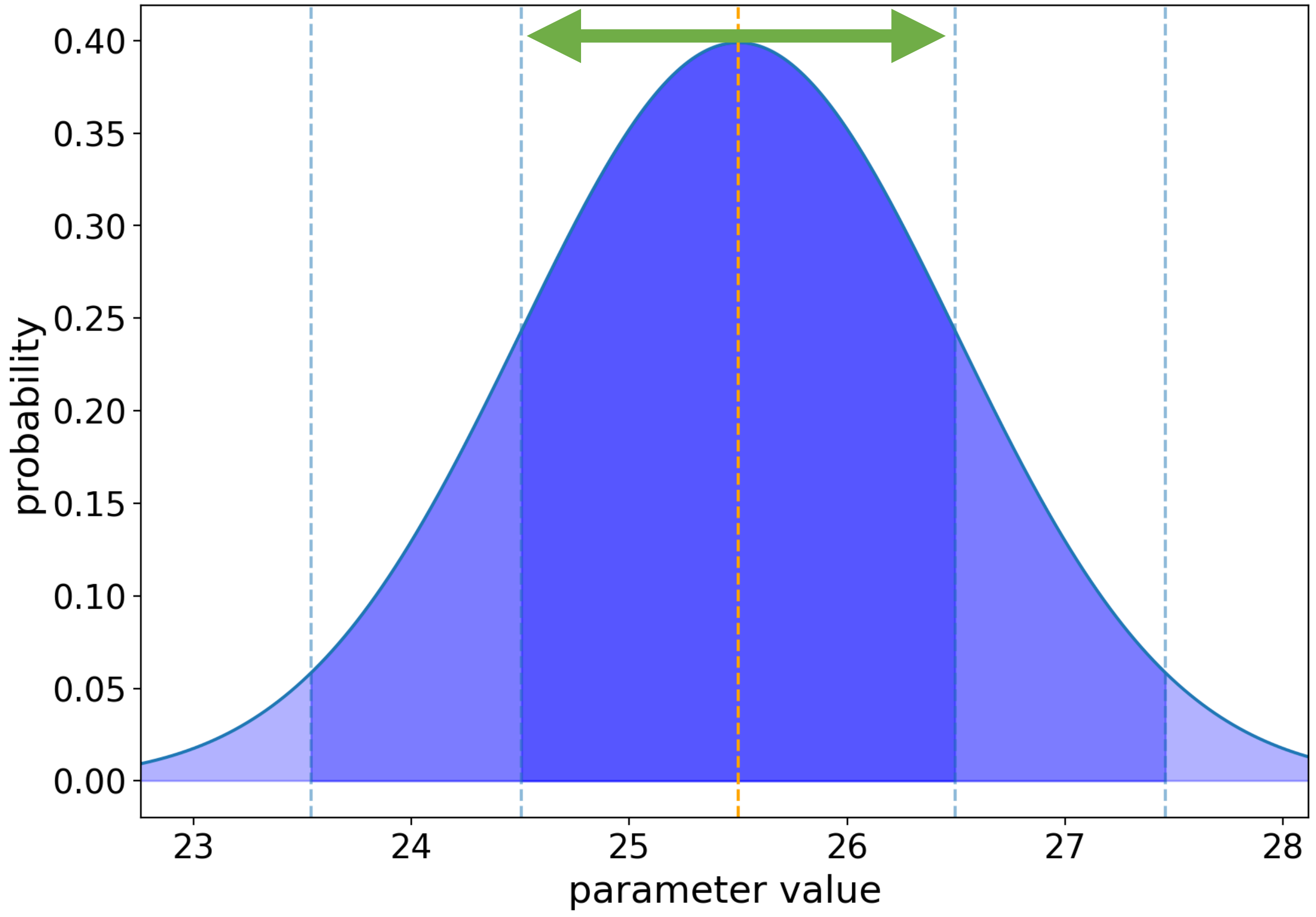
If our best-fit value has $\chi^2 = \chi^2_{min}$:

Contours of equal probability are defined by

$$\chi^2 = \chi^2_{\boldsymbol{min}} + \Delta$$

			N_{params}	
		1	2	3
CI 68%	1σ	1.0	2.30	3.53
CI 95.4%	2σ	4.0	6.17	8.02
CI 99.7%	3σ	9.0	11.8	14.2

$\chi^2 < \chi^2_{\text{min}} + 1$



Week 11: Learning outcomes

Today you have learnt

- The relevance of data analysis in experimentation
- Measurements are drawn from underlying distributions
  - The Gaussian distribution is a principal example
- How to estimate the value of a model parameter given a dataset
  - For a linear model, this is commonly known as linear regression
  - How least-squares fitting is related to the Gaussian distribution
  - Confidence Intervals : expressing our results
- Goodness-of-fit metrics: (dis)favouring a given model
  - The χ^2_{red} test

Practical examples on Friday!