Data Analysis and Statistics



Prof Adrian Oeftiger

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

Based on lectures given by Dr Jenny Barnes

Lecture 3

- Weighted mean
- So I have some data…
- Method of least squares
- Residuals -- what do they tell us?
- What is a good fit?
- χ^2 (Chi squared) distribution
- Estimation

Weighted mean (1)

How do you **combine several different measurements** of the same thing, each with their own value and error?

- Use a weighted mean, so each value takes into account the error in it: those with a small error are given more weight than those with a larger error.
- We do need to use a bit of judgement -- perhaps ignore measurements which are too far away / outliers (but be prepared to justify that decision!)

Weighted **mean:** $\bar{x} = \sum_{i=1}^{N} \omega_i \, x_i$, introducing weights ω_i with

$$\sum_{i=1}^{N} \omega_i = 1$$

Weighted mean (2)

- Suppose we have two measurements, $x_1 \pm \alpha_1$ and $x_2 \pm \alpha_2$
- We can use a weight factor related to the error:

$$\omega_i = \frac{1/\alpha_i^2}{\sum_{i=1}^{N} 1/\alpha_i^2}$$

- Best estimate in this case is then:

$$\bar{x}_\omega = \frac{\frac{x_1}{\alpha_1^2} + \frac{x_2}{\alpha_2^2}}{\frac{1}{\alpha_1^2} + \frac{1}{\alpha_2^2}}$$

- error α_ω in this new weighted mean:

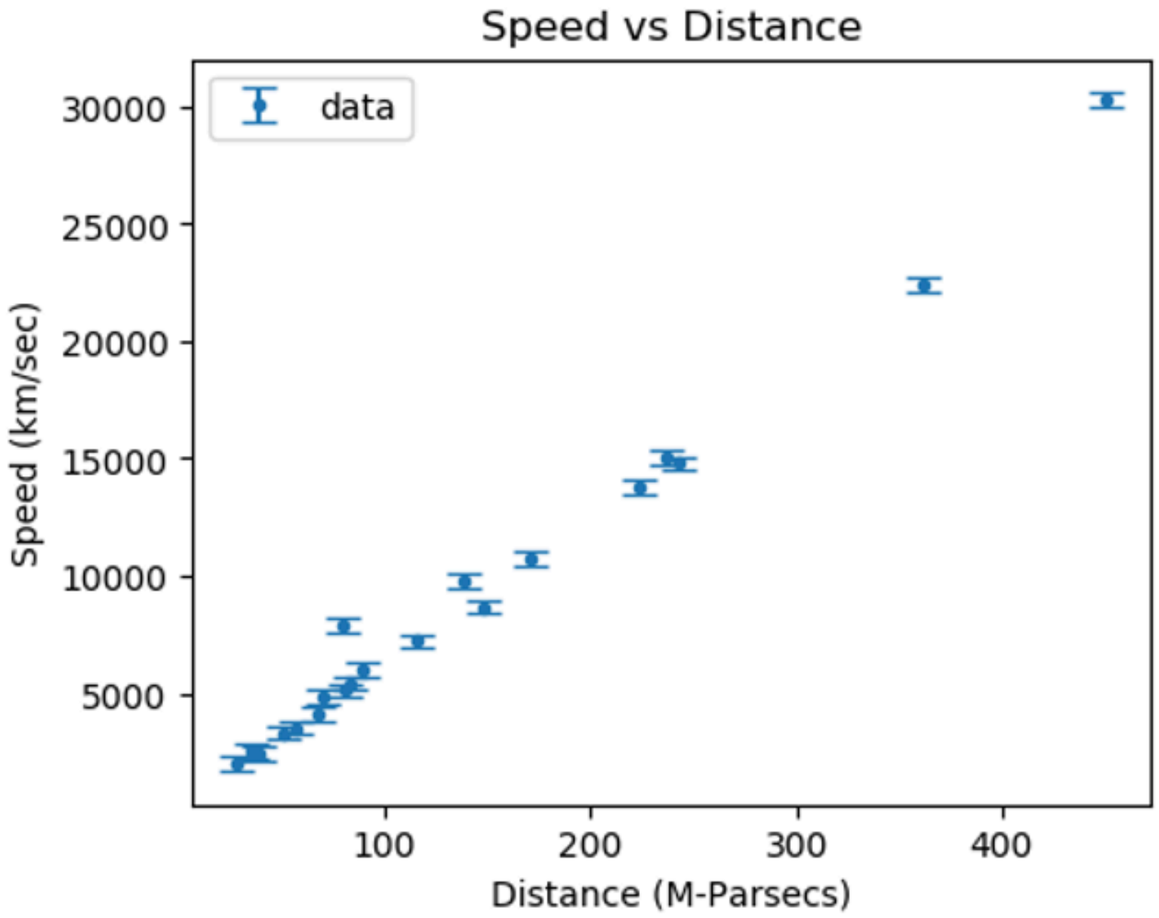
$$\frac{1}{\alpha_\omega} = \sqrt{\sum_{i=1}^{N} 1/\alpha_i^2}$$

So I have some data...

- Imagine you have taken some data and plotted it on a graph
- You could perhaps eliminate any points which don't look right or do a fit which doesn't include all the points -- **use your judgement but don't just throw points away without a really good reason and always state this reason in your logbook.** Maybe retake the point rather than throw it away.
- Try to include errors on all your points if possible -- for random errors the error bar is α_i, but think about the probability distribution which is most relevant for your data.

Typical data

DA01



So I want to fit a graph...

- Having plotted data, I want to fit a line or function to it
- Many different methods exist for fitting a graph and for deciding if you have a good fit between theory and experiment
- Here we cover the **method of least squares**

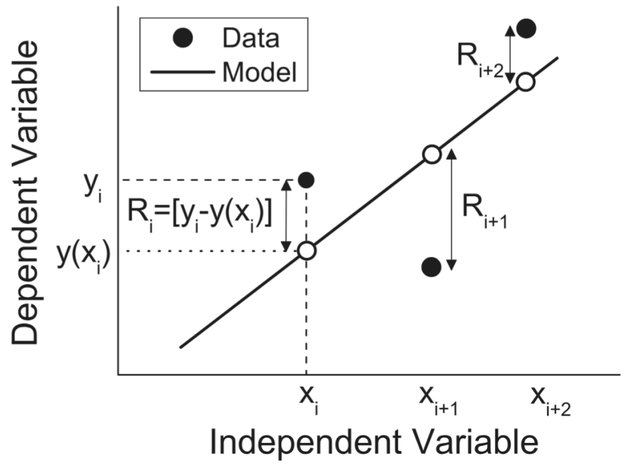
Fitting data – method of least squares

- We have a set of points (x_i, y_i) on a graph
- For simplicity here, we assume best fit given by line:

  $y = mx + c$ with slope m and intercept c
- For each data point x_i the predicted y value would be $y_i = mx_i + c$
- Want to minimise the differences between the best fit line and the data points:

  the residual: $(y_i - y(x_i))$
- This residual might be positive or negative. We really need to always have a positive quantity: use the modulus -- or square the residual

  $(y_i - y)^2$ ⇝ method of least squares

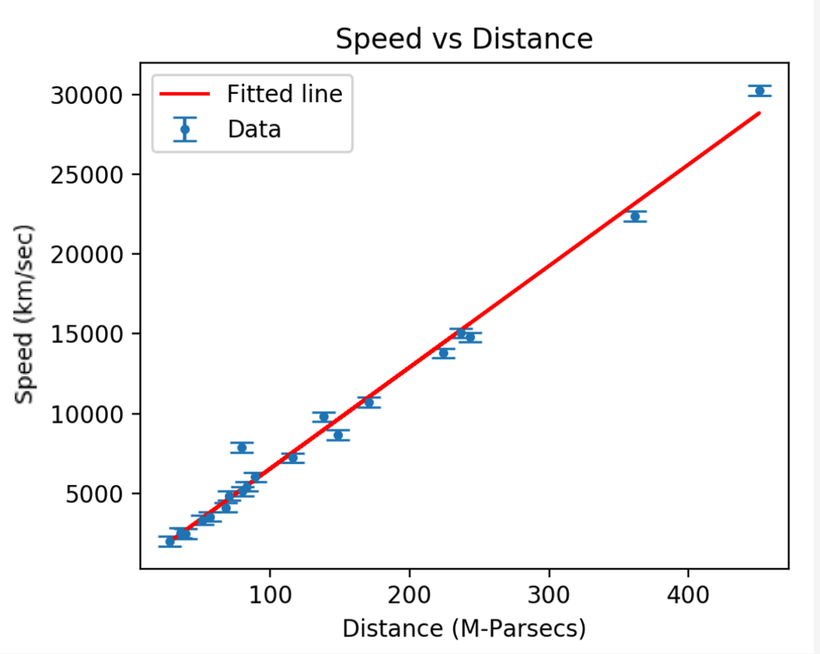


An example: DA01 exercise 1

- Try to minimise the square of the residuals to get the best fit
- Use a so-called "goodness of fit" parameter
- A common one is the "chi-squared" parameter which includes the errors α_i:

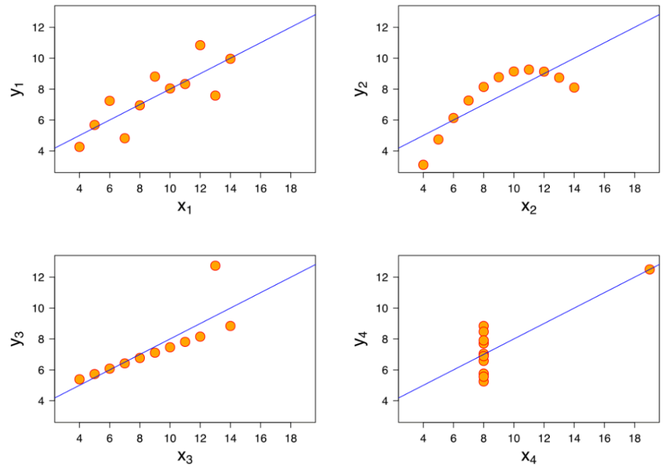
$$\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - y(x_i)}{\alpha_i} \right)^2$$

- NB: If the error for each point is the same, i.e. constant errors, then they don't affect the values of gradient, intercept, or their errors.



What is a good fit?

- So how do we know if it is a good fit?
- At first, look by eye: does it look reasonable?
  E.g. just as many points above the line as below it?
- Remember you can fit a straight line to anything...



For a straight line...

Straight line fit: explicit formulae to find m and c (with or without including weighting due to the errors), see e.g.

AD02 script

Unweighted straight line fit

Assume the data consist of n sets of measured quantities (x_i, y_i) and that only the dependent variables y_i are subject to experimental error. If the precision of individual data points is unknown, assume that all the weights are equal, $w_i = 1/n$, and use an unweighted fit to a straight line

$$y = a + bx \tag{31}$$

Best fit parameters

The best fit is taken to be the function which minimises the equally weighted sum S of the squares of the deviations d_i of the observed points from the straight line. Thus we have

$$S = \frac{1}{n}\sum_{i=1}^{n} d_i^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - a - bx_i)^2 \tag{32}$$

$$\text{subject to the constraints} \quad \frac{\partial S}{\partial a} = 0 \quad \text{and} \quad \frac{\partial S}{\partial b} = 0 \tag{33}$$

The following results can be derived:

$$\begin{aligned} &\text{gradient} && b = \frac{[\xi\eta]}{[\xi^2]} \\ &\text{y-intercept} && a = \bar{y} - b\bar{x} \\ &\text{x-intercept} && x_{\text{int}} = -\frac{a}{b} \end{aligned} \tag{34}$$

where

$$\begin{aligned} \bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i & \bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ \xi_i &= x_i - \bar{x} & \eta_i &= y_i - \bar{y} \\ [\xi^2] &= \frac{1}{n}\sum_{i=1}^{n} \xi_i^2 & [\xi\eta] &= \frac{1}{n}\sum_{i=1}^{n} \xi_i\eta_i \end{aligned} \tag{35}$$

This line goes through the centroid of the points.

Chapters 5 and 6 in Hughes and Hase

Chapter summary

- The overriding considerations when producing a good graph are simplicity and clarity.
- The error bar on a point on a graph is the two-thirds confidence limit.
- The error bar is drawn between $\left(\overline{Z} - \alpha_Z^-\right)$ and $\left(\overline{Z} + \alpha_Z^+\right)$, where α is the standard error (standard deviation of the mean).
- For Poisson statistics with N counts the error bar is drawn between $\left(N - \sqrt{N}\right)$ and $\left(N + \sqrt{N}\right)$.
- The method of least squares can be used to deduce the best-fit parameters for linear regression $y = mx + c$ where the data have constant error bars.
- The best estimate of the intercept and slope are

  $c = \dfrac{\sum_i x_i^2 \sum_i y_i - \sum_i x_i \sum_i x_i y_i}{\Delta}$, and

  $m = \dfrac{N\sum_i x_i y_i - \sum_i x_i \sum_i y_i}{\Delta}$,

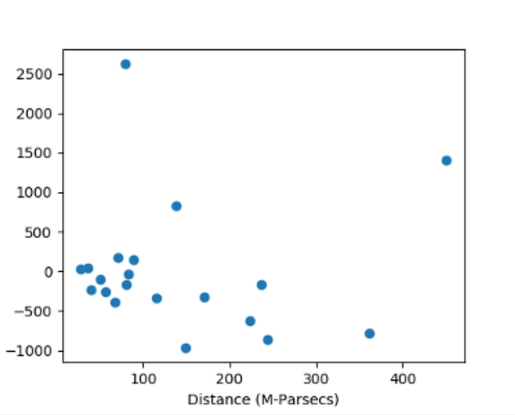
  where $\Delta = N\sum_i x_i^2 - \left(\sum_i x_i\right)^2$.
- The best estimate of the error in the intercept and slope are

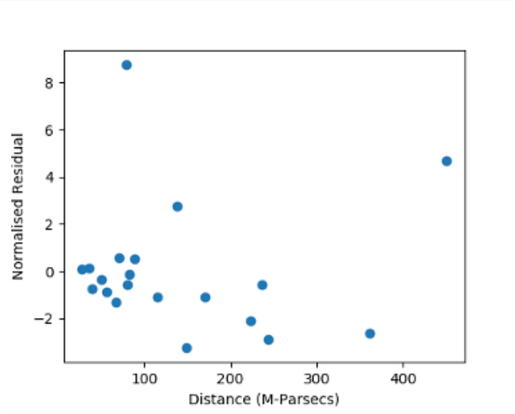
  $\alpha_c = \alpha_{\text{CU}}\sqrt{\dfrac{\sum_i x_i^2}{\Delta}}$, and $\alpha_m = \alpha_{\text{CU}}\sqrt{\dfrac{N}{\Delta}}$,

  where $\alpha_{\text{CU}} = \sqrt{\dfrac{1}{N-2}\sum_i (y_i - mx_i - c)^2}$.
- Having plotted a straight-line graph the residuals should be analysed—two-thirds of the data points should be consistent with the line within their error bars.

Next: plot out residuals vs x

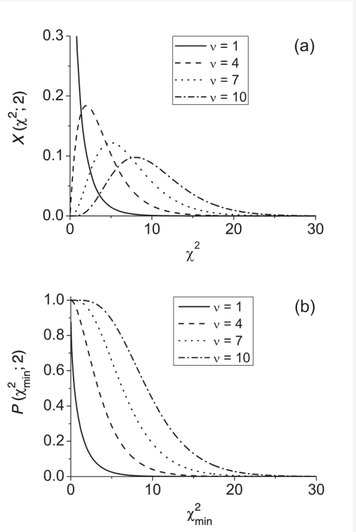
- If the residuals are random values (just as many + as −) then picking a straight line to fit the data will be sensible
- Plot the normalised residuals $\frac{y_i - y(x_i)}{\alpha_i}$ as we ideally want 95% of normalised residuals within -2 to +2 so within $\pm 2\alpha$
- If there is structure to the residual, you might need to modify your straight line to include a quadratic term or another function





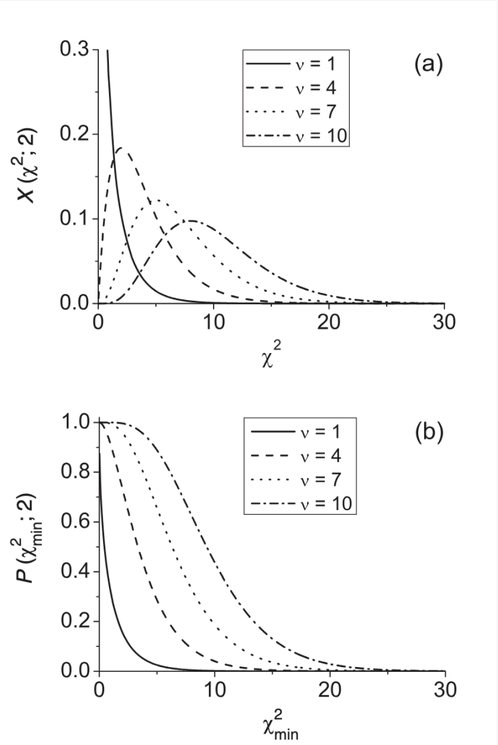
What is a good fit? – χ^2 distribution

- Assume the equation we are using to fit the data is called the **null hypothesis** and the minimised value of χ^2 is χ^2_{min}
- For a given number of degrees of freedom, ν, χ^2 has a normalised probability distribution and a cumulative probability function ($P =$ integral from χ^2_{min} to infinity of the χ^2 probability distribution)
- Form of χ^2 is non-trivial and usually use look up tables or statistical packages. The mean (where $P \approx 0.5$) is the number of degrees of freedom ν.



What is a good fit? – reject or question?

- We can only really say if we **reject**, **question**, or **do not reject the null hypothesis** -- we cannot say that we accept it
- For a good fit, expect P to be around 0.5 or $\chi^2_{\min} \approx \nu$
- Question null hypothesis if $P \lesssim 10^{-3}$
- Reject null hypothesis if $P \lesssim 10^{-4}$
  (in fact, often already from $P \leq 0.05$, "rejecting null hypothesis at 5% significance level")
- If $0.5 < P \to 1$ (or, equivalently, $\chi^2_{\min} \ll \nu$), check your calculations for the measurement uncertainties α_i (unreasonably large)



What is a good fit? – reduced Chi-squared

- Slightly easier -- use the reduced chi-squared statistic: $\chi^2_\nu = \dfrac{\chi^2_{\min}}{\nu}$
- For a reasonable fit $\chi^2_\nu \approx 1$
- If $\chi_\nu \ll 1$, check your calculations for the measurement uncertainties α_i
- If $\chi_\nu \gtrsim 3$, then it is a poor fit and the null hypothesis is questioned
    - (exact value at which you make this judgement depends on the number of degrees of freedom ν (e.g. $\nu = 20$ then poor fit if $\chi_\nu > 1.9$). See Table 8.1 on p. 107 in Chapter 8 in Hughes and Hase for table of numbers of χ_ν to reject at for different values of ν)

How many parameters to fit?

- Try to get as many data points as possible to fit through (large ν)
- Choose as few parameters to vary as possible

> *Occam's razor: if two competing theories give the same answer, the simpler one is better*

- For a straight line fit:
  - If you want to know the gradient accurately then choose a large range of x values and minimise the errors on the extreme points in the dataset
  - If you want to know the intercept accurately then get the errors close to the origin as small as possible

Estimation

- Sometimes doing a rough and ready calculation can be really helpful -- it doesn't always have to be precise.
- Getting a rough idea can also be informative, sometimes known as a "back of the envelope" calculation.
- Experiments are always a trade off between how precise a measurement is and cost (money or time).

Summary of lecture 3

- Weighted mean
- Method of least squares
- Residuals -- what do they tell us?
- What is a good fit? (Hypothesis rejection)
- Chi squared distribution
- Estimation

Summary of the course

- **A result without an error has no scientific meaning**
- If you have taken a set of readings with random errors quote $\bar{x} \pm \alpha$

  where $\bar{x} = \sum_{i=1}^{N} \frac{x_i}{N}$ and $\alpha = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N(N-1)}}$
- Always use your judgement -- a good physicist won't blindly put numbers in formulae
- Remember to propagate your errors
- Errors and error analysis can be fiddly but you can't do science without it!
- Error propagation (functional vs calculus approach)

Questions?