PH3010 / Advanced Skills Miniproject

Introduction to Statistical Data Analysis

Week 2: Goodness-of-fit, LS fitting with correlated data,

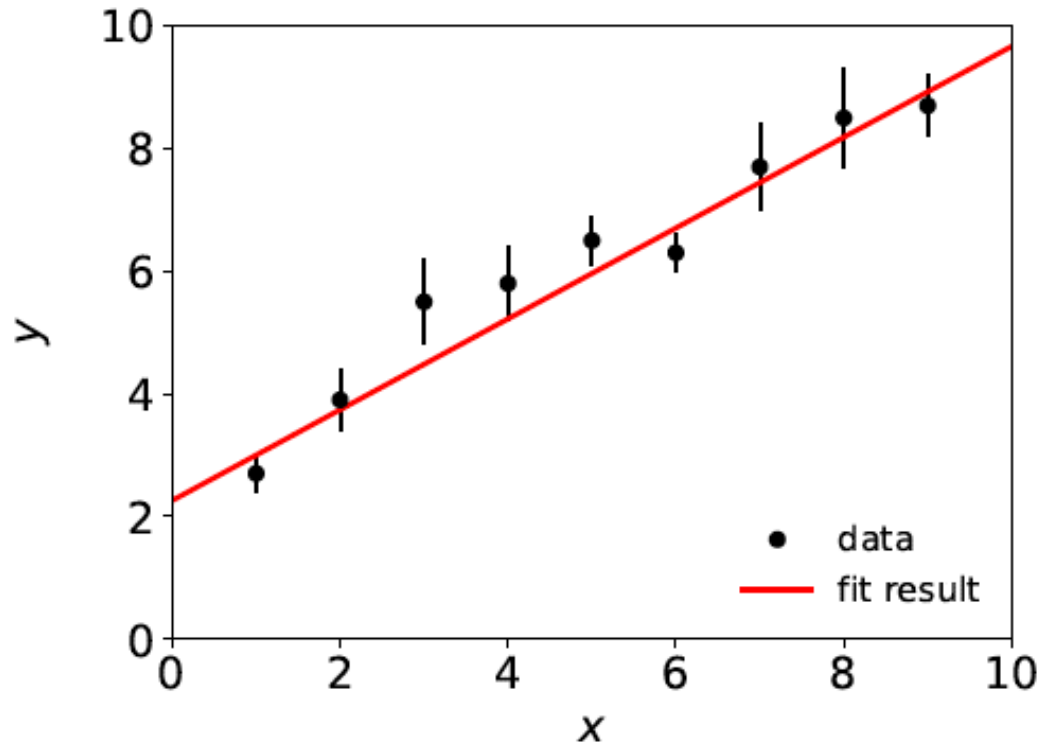## Autumn term 2022

Glen D. Cowan

RHUL Physics

Statistical Data Analysis

Lecture 2-1

- General idea of goodness-of-fit
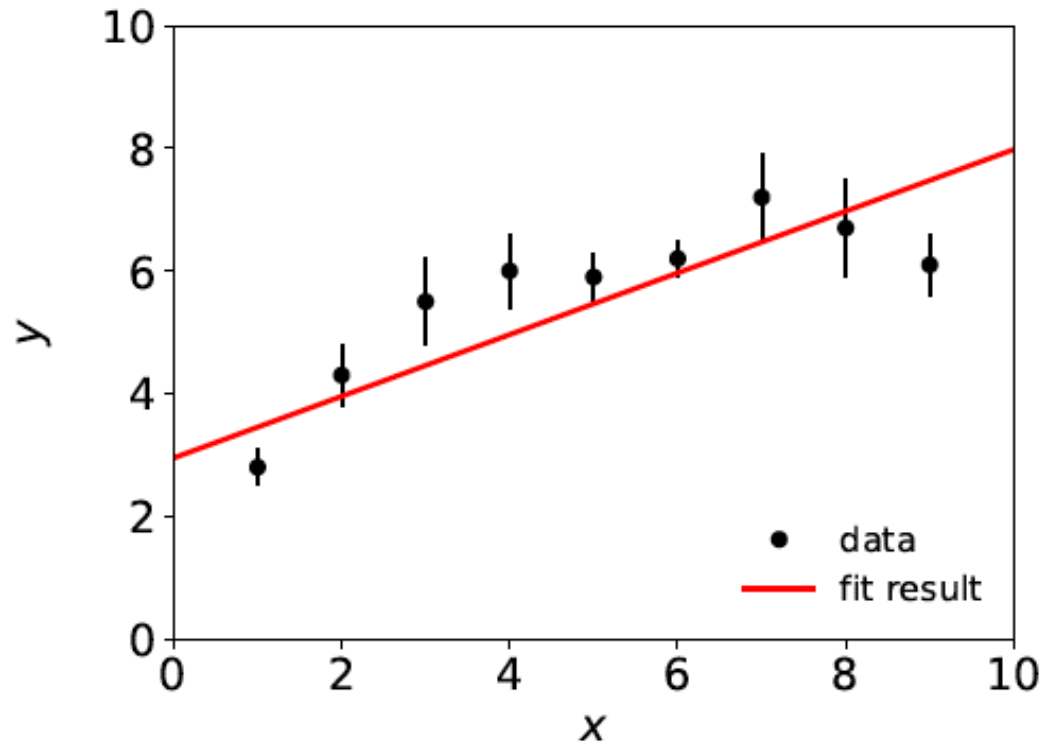
- Distribution of minimum of chi-squared

A "good" fit

Last week we fitted data that were reasonably well described by a straight line:

A "bad" fit

But what if a straight-line fit looks like this:

Maybe here we should fit a higher-order polynomial?

Goodness-of-fit: the questions

How do we quantify the level of agreement between the data and the hypothesized form of the fit function?

How do we decide whether to try a different fit function?

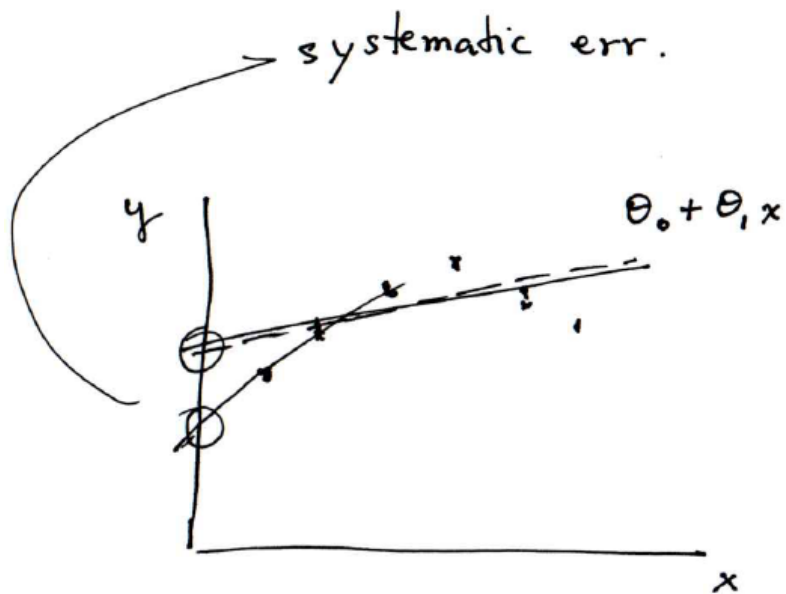⚠ Note first the following common misunderstanding:

If the fit is "bad", you may expect large statistical errors for the fitted parameters. This is not the case.

The statistical errors say how much the parameter estimates will fluctuate under repetition of the experiment. This is not the same as the degree to which the function is able to describe the data.
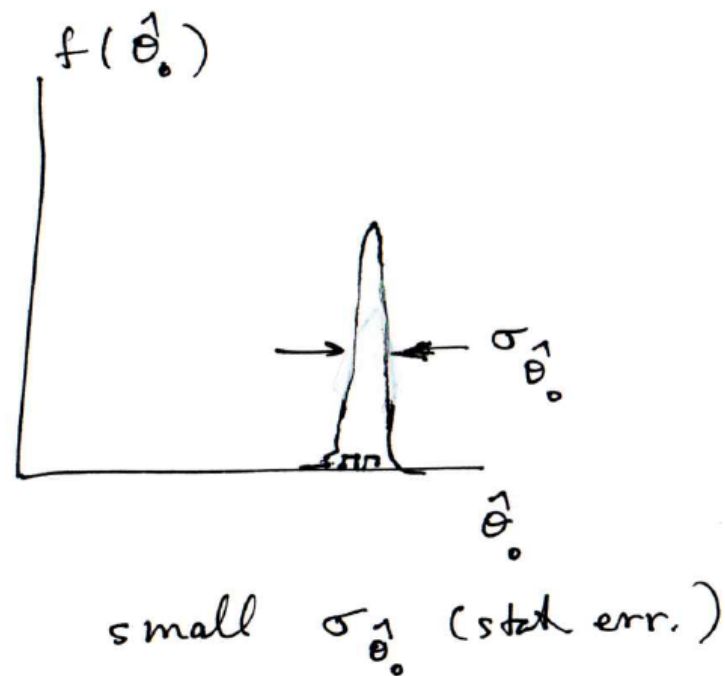
If the hypothesized $f(x; \boldsymbol{\theta})$ is not correct, the fitted parameters will have some systematic uncertainty – a more complex question that we will not take up here.

Statistical errors versus goodness-of-fit

Statistical errors versus goodness-of-fit

systematic err.

$\theta_0 + \theta_1 x$

bad fit, but

$f(\hat{\theta}_0)$

$\sigma_{\hat{\theta}_0}$

$\hat{\theta}_0$

small $\sigma_{\hat{\theta}_0}$ (stat err.)

Quantifying goodness-of-fit

We can quantify the goodness-of-fit directly from the value of $\chi^2(\boldsymbol{\theta})$ evaluated at its minimum, i.e., at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$:

$$\chi^2_{\min} = \sum_{i=1}^{N} \frac{(y_i - f(x_i; \hat{\boldsymbol{\theta}}))^2}{\sigma_i^2}$$

If the fitted function is in good agreement with the data, then the numerator of each term in the sum should be small.

If $f(x_i; \hat{\boldsymbol{\theta}})$ were equal to the true mean of y_i, then we would expect the residual $y_i - f(x_i; \hat{\boldsymbol{\theta}})$ to have an rms value of σ_i. Each term in the sum would contribute 1, and we'd have $\chi^2_{\min} = N$.

This is not quite true: if we have fitted m parameters and the hypothesized function is correct, the *expected* value of $\chi^2_{\min}$ is $N - m$ (called the *number of degrees of freedom* of the fit.)

Distribution of χ^2_{min}

Distribution of χ^2_{min}
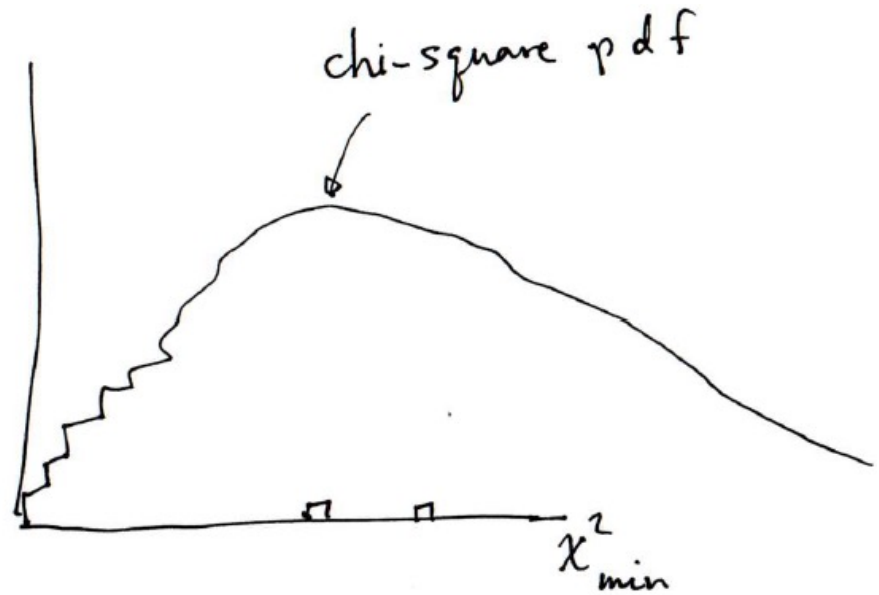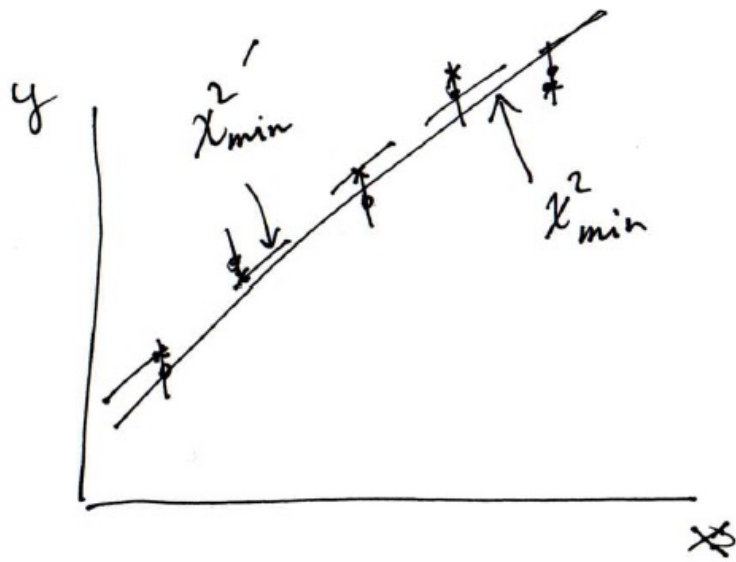
Distribution of $\chi^2_{\min}$

$\chi^2_{\min}$ is a function of the data so is itself a random variable.

If the hypothesized fit function is correct and the data are Gaussian distributed, one can show $\chi^2_{\min}$ follows a chi-square distribution with $n = N - m$ degrees of freedom (here let $\chi^2_{\min} = z$):

$$f_{\chi^2}(z; n) = \frac{1}{2^{n/2}\Gamma(n/2)} z^{n/2-1} e^{-z/2}$$

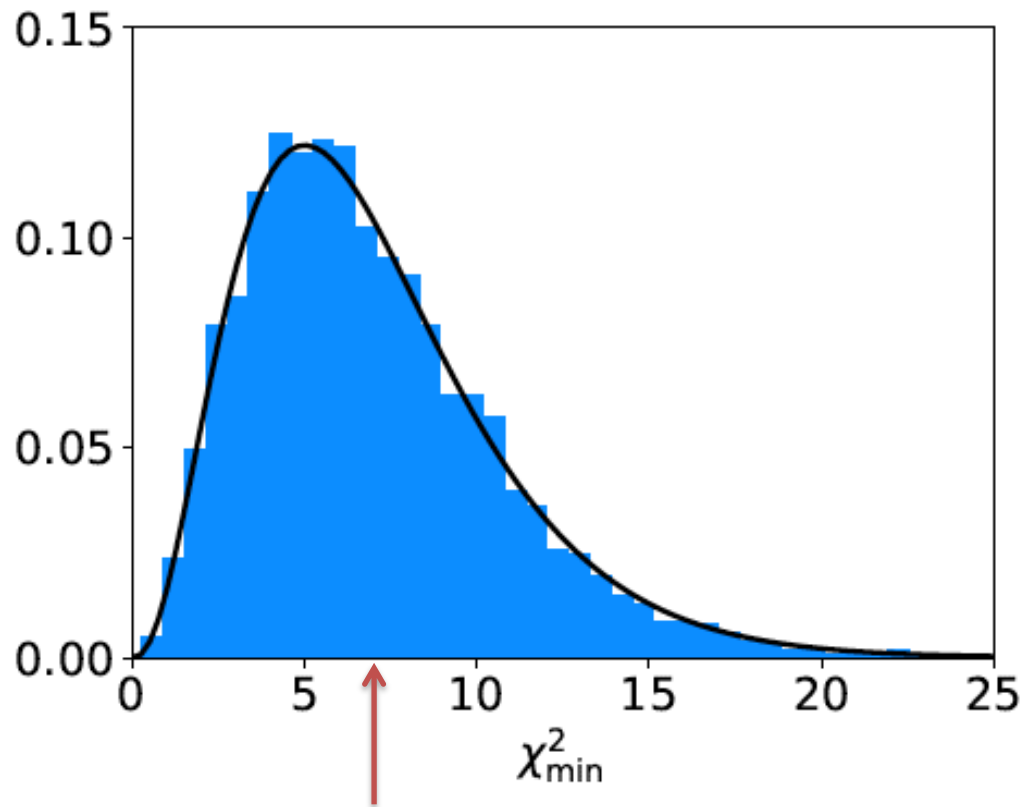Mean and standard deviation of the chi-square distribution:

$$\langle z \rangle = n$$

$$\sigma_z = \sqrt{2n}$$

If the hypothesized fit function is not correct, then one would obtain a distribution of $\chi^2_{\min}$ shifted to higher values.

Distribution of $\chi^2_{\min}$ from straight-line fit

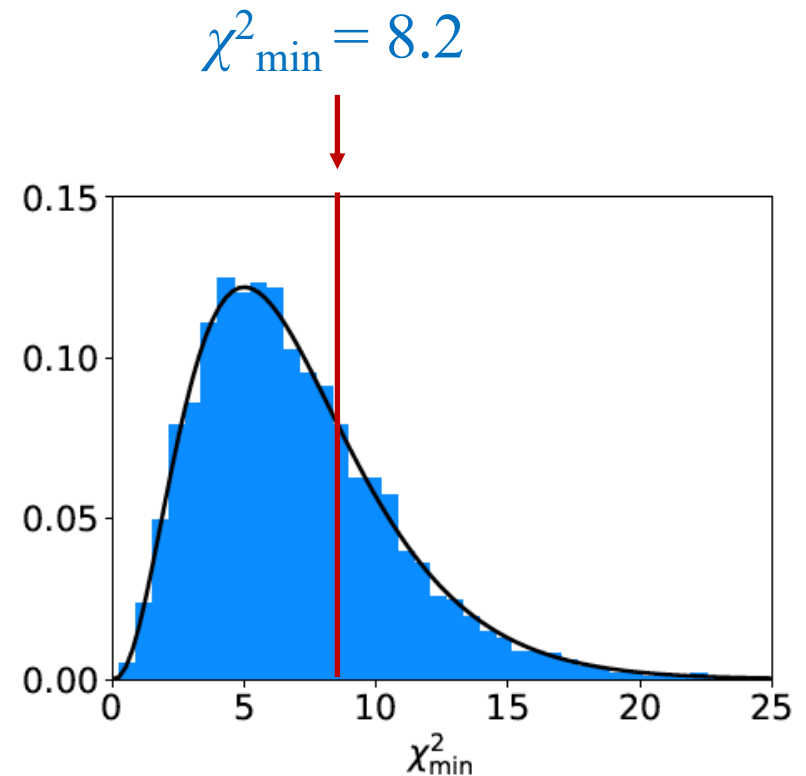$\chi^2_{\min}$ from straight-line fit with $N = 9$ data points, $m = 2$ fitted parameters.

Curve: chi-square pdf for $n = 7$ degrees of freedom.

Histogram: values of $\chi^2_{\min}$ from straight-line fit from repeated Monte Carlo simulation of the experiment.

mean $= N - m = 7$

$\chi^2_{\min}$ from the "good" fit

9 data points, 2 fitted parameters

$\rightarrow$ 7 degrees of freedom

$\chi^2_{\min} = 8.2$
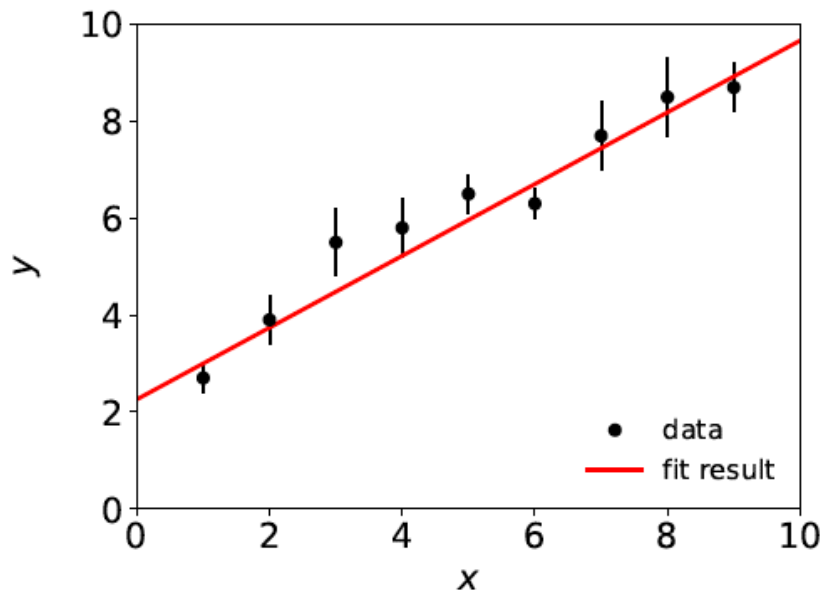
By eye one would say that under the assumption of the straight-line hypothesis, the goodness-of-fit was "typical".

Statistical Data Analysis

Lecture 2-2

- How to interpret $\chi^2_{\min}$

- p-value from $\chi^2_{\min}$

- Least squares with correlated data

How to interpret $\chi^2_{\min}$

A simple way to assess the goodness-of-fit is simply to compare $\chi^2_{\min}$ to the number of degrees of freedom, $n_{\mathrm{dof}} = N - m$.

$\chi^2_{\min} \sim n_{\mathrm{dof}}$ $\rightarrow$ fit is "good"

$\chi^2_{\min} \gg n_{\mathrm{dof}}$ $\rightarrow$ fit is "bad"

$\chi^2_{\min} \ll n_{\mathrm{dof}}$ $\rightarrow$ fit is better than what one would expect given fluctuations that should be present in the data.

Often this is done using the ratio $\chi^2_{\min}/n_{\mathrm{dof}}$, i.e. fit is good if the "chi-square per degree of freedom" comes out not much greater than 1.

Often report as, e.g., $\chi^2_{\min}/n_{\mathrm{dof}} = 8.2/7$. It is best to communicate both $\chi^2_{\min}$ and n_{dof}, not just their ratio.

p-value from χ^2_{min}

Another way to assess the goodness-of-fit is to give the probability, assuming the fit function is correct, to obtain a χ^2_{min} value as high as the one we got or higher:

$$p = \int_{\chi^2_{\mathrm{min}}}^{\infty} f_{\chi^2}(z; n)\, dz$$

This is an example of what is called a p-value of the hypothesis (here the hypothesized form of the fit function).

p-value is not the same as the probability that the hypothesis is true!

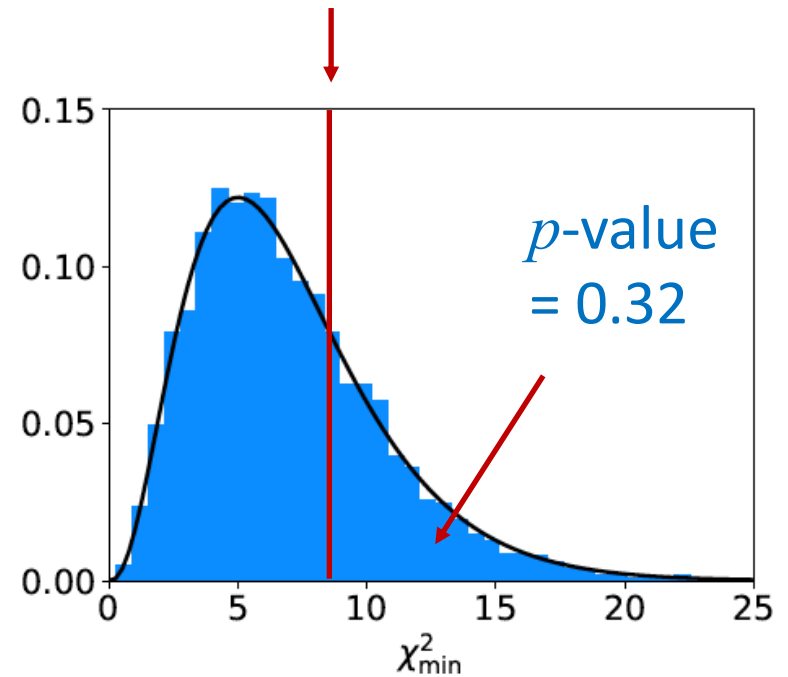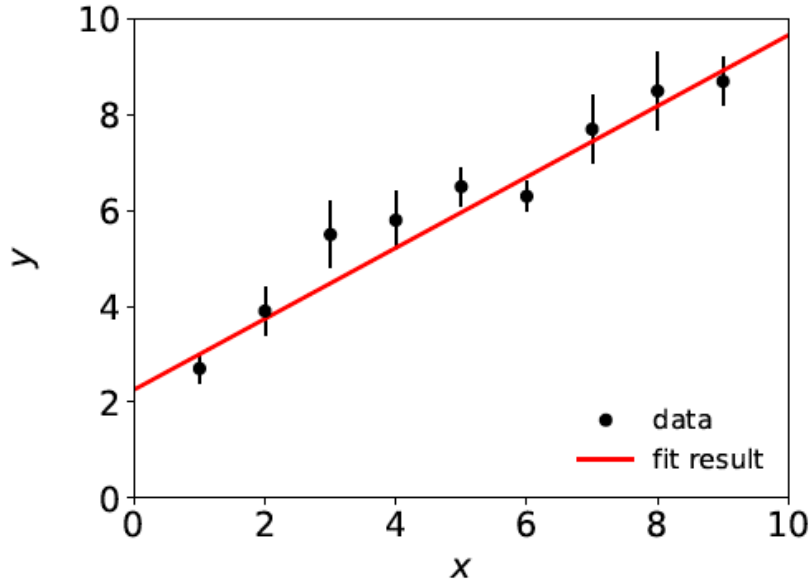Nevertheless, a small p-value indicates that the hypothesis is disfavoured.

Compute using: `scipy.stats.chi2.sf`

p-value for the "good" fit

$N = 9$ data points, $m = 2$ fitted parameters,

$\chi^2_{\min}/n_{\mathrm{dof}} = 8.2/7 = 1.2$

$\chi^2_{\min} = 8.2$
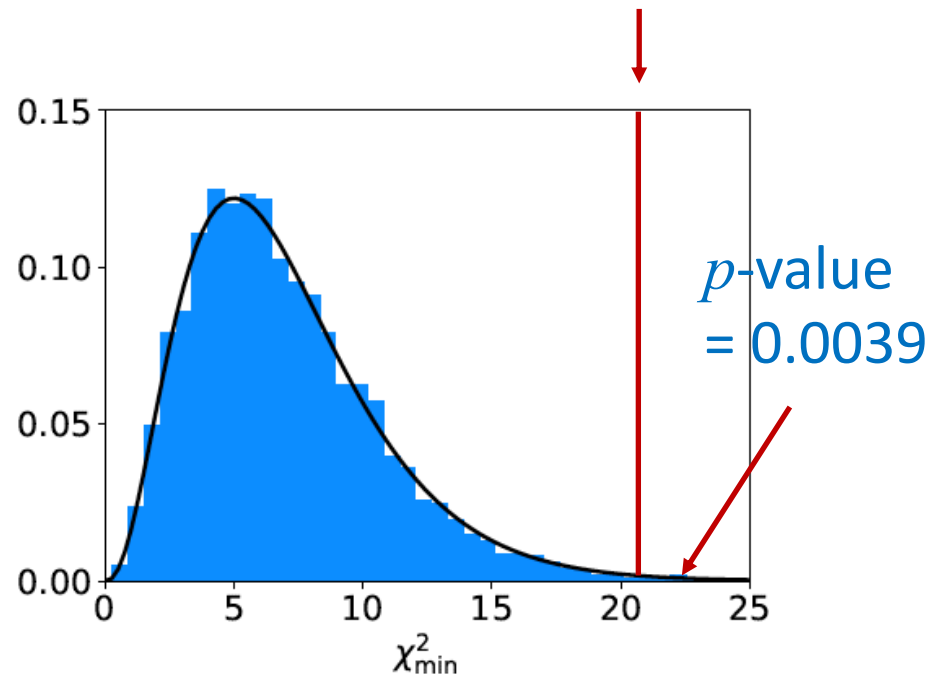

p-value
= 0.32

If the straight-line hypothesis is true, expect equal or worse agreement almost 1/3 of the time (i.e. our result is not unusual).

p-value for the "bad" fit

$N = 9$ data points, $m = 2$ fitted parameters,

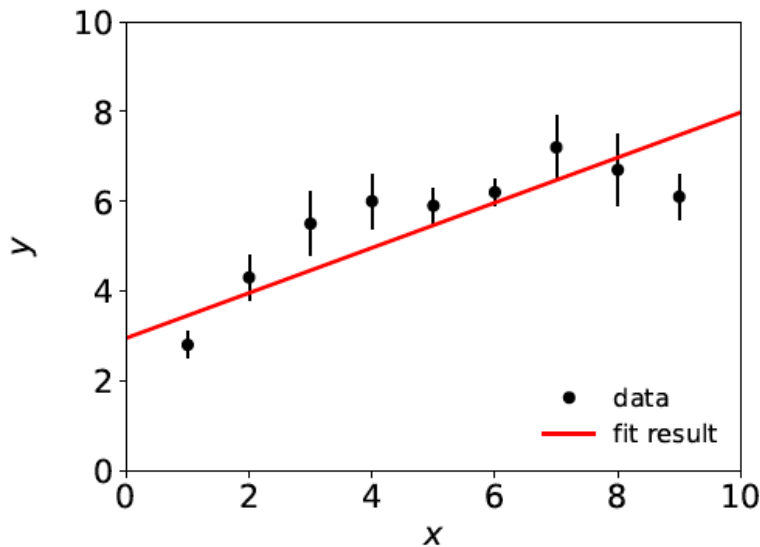$\chi^2_{\min}/n_{\text{dof}} = 20.9/7 = 3.0$

$\chi^2_{\min} = 20.9$

p-value
= 0.0039

So is the straight-line hypothesis correct? It could be, but if so we would expect a $\chi^2_{\min}$ as high as observed or higher only 4 times out of a thousand.

A better fit

If we decide the agreement between data and hypothesis is not good enough (exact threshold is a subjective choice), we can try a different model, e.g., a 2nd order polynomial:

$$f(x; \boldsymbol{\theta}) = \theta_0 + \theta_1 x + \theta_2 x^2$$
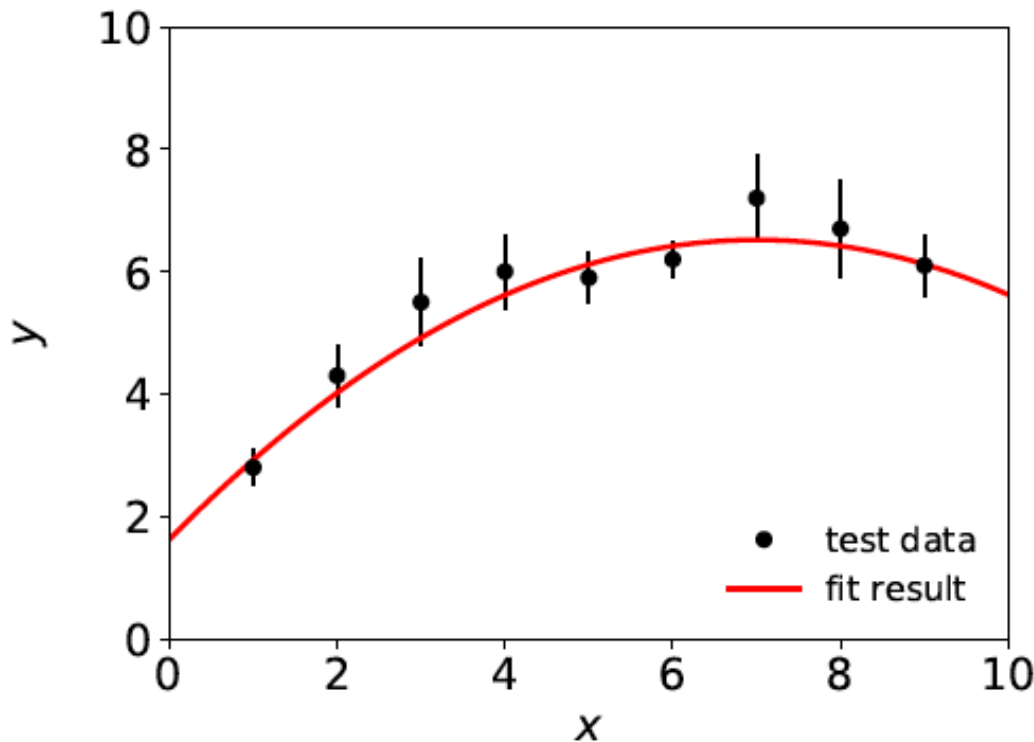
$\chi^2_{\min} = 3.5$ for $n_{\mathrm{dof}} = 6$

$\chi^2_{\min} / n_{\mathrm{dof}} = 0.58$

p-value $= 0.75$

Least squares with correlated measurements

Up to now we have assumed that the measurements $y_1, ..., y_N$ are all independent.

This means that if one value fluctuates, say, high, then this has no influence on whether one of the others will fluctuate high or low.

But there could be cases where the y_i are correlated, i.e., they have nonzero covariances

$$\text{cov}[y_i, y_j] = V_{ij} = \rho_{ij}\sigma_i\sigma_j$$

In this case, the formula for $\chi^2(\boldsymbol{\theta})$ becomes

$$\chi^2(\boldsymbol{\theta}) = \sum_{i,j=1}^{N} (y_i - f(x_i; \boldsymbol{\theta}))V_{ij}^{-1}(y_j - f(x_j; \boldsymbol{\theta}))$$

where V^{-1} is the inverse of the covariance matrix of the data V.

LS with correlated measurements (2)

As a check, suppose we have the special case of a diagonal covariance matrix, i.e., uncorrelated measurements. Then

$$
V = \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & 0 & \dots \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n^2 \end{pmatrix} \qquad \longrightarrow \qquad V^{-1} = \begin{pmatrix} 1/\sigma_1^2 & 0 & \dots & 0 \\ 0 & 1/\sigma_2^2 & 0 & \dots \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & 1/\sigma_n^2 \end{pmatrix}
$$

$V^{-1}{}_{ij} = \delta_{ij}/\sigma_i^2$, carry out one of the sums, $\chi^2(\boldsymbol{\theta})$ same as before:

$$
\chi^2(\boldsymbol{\theta}) = \sum_{i,j=1}^{N} (y_i - f(x_i; \boldsymbol{\theta})) \frac{\delta_{ij}}{\sigma_i^2} (y_j - f(x_j; \boldsymbol{\theta})) = \sum_{i=1}^{N} \frac{(y_i - f(x_i; \boldsymbol{\theta}))^2}{\sigma_i^2}
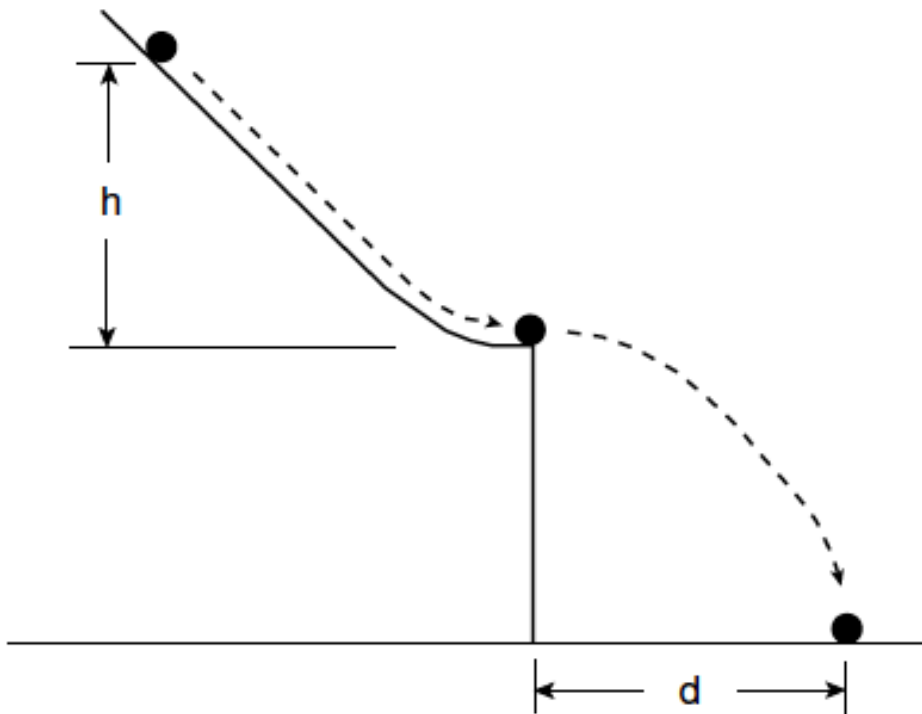$$

Statistical Data Analysis

Lecture 2-3

- Further discussion of the exercises

- Structure of the report

Goodness-of-fit with Galileo's data

Last week we used data from Galileo...

h	d
1000	1500
828	1340
800	1328
600	1172
300	800

Goodness-of-fit with Galileo's data

...to fit several hypotheses for the functional relation between the initial height h and flight distance d:

$$d = \alpha h$$

$$d = \alpha h + \beta h^2$$

$$d = \alpha h^\beta$$

So now we can put ourselves in Galileo's position and, without knowledge of Newton's laws, find which hypotheses are compatible with or disfavoured by the data (find $\chi^2_{\text{min}}/n_{\text{dof}}$ and p-value).

Extended analysis of Galileo's data

You can also use your knowledge of Newtonian mechanics to work out the predicted law and compare to what you find empirically.

Hints: apply conservation of energy to the rolling ball; do not forget that some of the gravitational potential energy goes into rotational energy.

Bonus question: work out the height of the table in terms of the measured data and compare to the value recorded in Galileo's notebook.   Are the values consistent? You will need to use error propagation to work out the uncertainty on the predicted height of the table.

Exercise 3: refraction data from Ptolemy

Astronomer Claudius Ptolemy obtained data on refraction of light by water in around 140 A.D.:
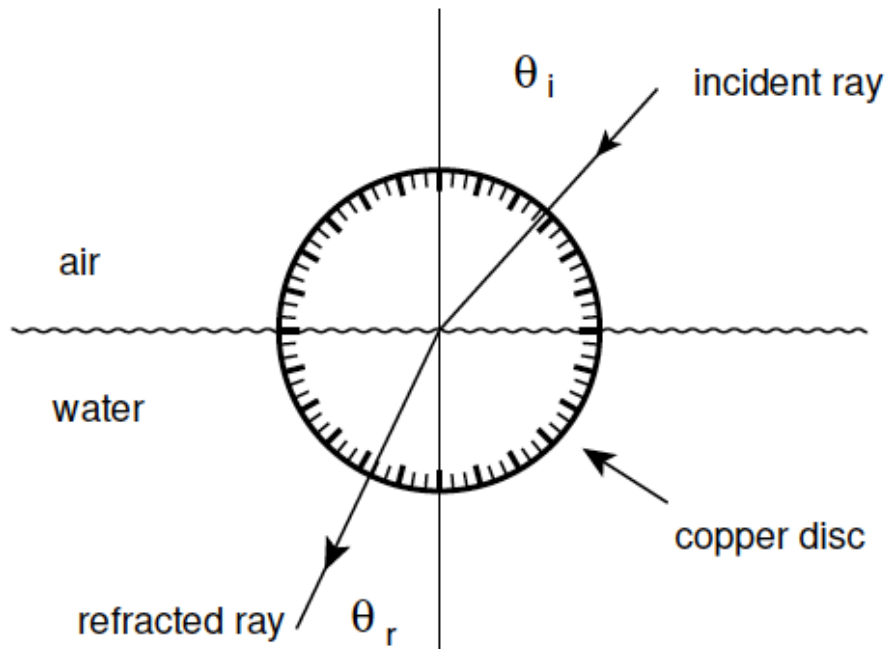
Angles of incidence and refraction (degrees)

θ_i	θ_r
10	8
20	$15\frac{1}{2}$
30	$22\frac{1}{2}$
40	29
50	35
60	$40\frac{1}{2}$
70	$45\frac{1}{2}$
80	50

Suppose the angle of incidence is set with negligible error, and the measured angle of refraction has a standard deviation of ½° .

Laws of refraction

A commonly used law of refraction was

$$\theta_{\mathrm{r}} = \alpha\theta_{\mathrm{i}} \ ,$$

although it is reported that Ptolemy preferred

$$\theta_{\mathrm{r}} = \alpha\theta_{\mathrm{i}} - \beta\theta_{\mathrm{i}}^2 \ .$$

The law of refraction discovered by Ibn Sahl in 984 (and rediscovered by Snell in 1621) is

$$\theta_{\mathrm{r}} = \sin^{-1}\left(\frac{\sin\theta_{\mathrm{i}}}{r}\right) \ .$$

where $r = n_{\mathrm{r}}/n_{\mathrm{i}}$ is the ratio of indices of refraction of the two media.

Analysis of refraction data

Fit the parameters and find their statistical errors and (where relevant) covariance matrix.

Assess goodness-of-fit of each hypothesis ($\chi^2_{\min}/n_{\text{dof}}$ and p-value).

What can you conclude?

(See The Feynman Lectures on Physics, Vol I., Addison-Wesley, 1963, Section 26-2, `www.feynmanlectures.caltech.edu` .)

Discussion of project report

Your report should have

Title, abstract on separate page

An introduction

A brief description of the statistical methods (can be a separate section or combined with others)

A section for each exercise

Brief conclusions

Bibliography

Appendices (including all code you've written)

Further details on this next week.

Discussion of project report (2)

Guidelines for report writing are on the moodle page, with LaTeX template, etc.

Use the relevant tools in LaTeX for the components of the report (sections, figures, bibliography, etc.).

For the code listings you can use e.g. LaTeX's **minted** package.

Deadlines and instructions for submission are on moodle.

Do not put off writing to the last minute (start ~now); more info on writing will be given in week 3.