

LC Introduction to Probability and Statistics Lecture Notes

MSci Physics w/ Particle Physics and Cosmology
University of Birmingham

Year 1, Semester 2
Ash Stewart

Lectures Index

Lecture 1: Start of Stats: Introduction and Descriptive Statistics	1
Lecture 2: Population Statistics	2
Lecture 3: Error Propagation and Combinations of Variables	3
Lecture 4: Covariance and Correlation	5
Lecture 5: Distributions	6
Lecture 6: Likelihood and Log Likelihood	10
Lecture 7: Fitting a Straight Line 1	16
Lecture 8: Fitting a Straight Line 2	19
Lecture 9: Linear Regression	23
Lecture 10: Goodness of Fit	28
Lecture 11: End of Stats: Revision	32
Lecture 1: Start of Probability: Introduction	33
Lecture 2: Combinatorics	34
Lecture 3: Combining Probabilities	35
Lecture 4: Conditional Probability	38

Wed 01 Oct 2025 12:00

Lecture 1 - Start of Stats: Introduction and Descriptive Statistics

0.1 Course Welcome

- First half of the semester: Statistics
- Second half the of semester: Probability
- All slides and notes on Canvas.

Why Descriptive Statistics? If we want to share an interesting bit of data, sharing the whole data is going to be confusing. Instead, we can share a small number of stats which describe and summarise the data.

0.2 Sample Statistics

One of the most simple is the number of samples (N), and the sample mean:

$$\text{Sample Mean: } \bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$$

We can also calculate the sample standard deviation as the average of mean squared error across the points in the sample:

$$\text{Sample STDev: } s_n^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2$$

We can also use median or mode as measures of central tendency. The mode is a poor estimator however (as it massively depends on how binning is done, for a continuous measurement), while the median is more resistant to outliers.

Thu 02 Oct 2025 09:00

Lecture 2 - Population Statistics

0.1 Accuracy and Precision

We usually take measurements to determine some kind of true value. Usually, we can't actually know what this true value is, but if we could there are two bits of terminology that is particularly important:

Accuracy: Accuracy is the 'closeness' between our value and the 'true' value.

Precision: Precision is the 'closeness' between our measurements, i.e. how spread out are our various measurements.

0.2 Error

Random Error: is uncertainty related to the fact that our measurements are only a finite sample, so is not going to be immediately representative of the true value. The smaller this error, the more precise the measurement is.

Systematic Error: is related to some kind of issue with the measurement or the equipment. This shifts all values, and negatively affects accuracy (but leaves precision unchanged)

Taking many repeat measurements decreases the effects of random error, but the effects of systematic error are much harder to combat.

Ideally, we want to be both precise and accurate, however accuracy is arguably more important. This is because a value which is precise, but not accurate may lead to false conclusions around the inaccurate value.

Wed 08 Oct 2025 12:02

Lecture 3 - Error Propagation and Combinations of Variables

Office Hours: 11:00 to 13:00 Thursdays, Physics West Rm 122

1 Types of Error

Broadly two types of error: Statistical/Random Error (resulting from low precision) and Systematic Error (from Low Accuracy).

Random error widens the distribution, while systematic error shifts the whole distribution up or down, meaning no matter how many repeats you take and how precise you think you are, the value is still nonsense as all datapoints have been equally shifted (i.e. by a poor experimental setup).

For example, you are trying to measure the length of an object using a ruler that has been unknowingly stretched. You cannot get a true value no matter the number of repeats or degree of precision.

1.1 Accuracy vs Precision

High accuracy is preferable to high precision - having high precision but low accuracy can lead to false conclusions (as an incorrect value appears confidently correct). Accuracy is more difficult to improve - precision can be improved by gathering more data, while higher accuracy can only be improved by a better experimental design.

2 Error Propagation

If we take a distribution, and add a constant value to all points, the distribution is shifted up/down without changing the variance.

$$\langle x + k \rangle = \langle x \rangle + k$$

$$Var(x + k) = Var(x)$$

If we multiply by a constant value, the mean is multiplied by this value, but the distribution becomes stretched and the variance grows:

$$\langle xk \rangle = k\langle x \rangle$$

$$Var(kx) = k^2 Var(x)$$

Or taking the natural log:

$$\langle \ln x \rangle \approx \ln \langle x \rangle$$

$$Var(\ln x) \approx \frac{Var(x)}{x^2}$$

As this is a non-linear operator, these become good approximations rather than strict rules of equivalence.

And another example:

$$\langle e^x \rangle \approx e^{\langle x \rangle}$$

$$\text{Var}(e^x) \approx (e^x)^2 \text{Var}(x)$$

Note here, even though our underlying distribution is Normal and symmetric, the new distribution after e^x is neither, and these are an even worse approximation than before.

2.1 Combining Operators

We can apply some linear transformation $mx + c$, we can chain these rules together by doing the multiplicative transformation m first, then the linear scale c .

$$\langle mx + c \rangle = m\langle x \rangle + c$$

$$\text{Var}(mx + c) = m^2 \text{Var}(x)$$

2.2 Multiple Variables

What if we have multiple distributed variables we want to add?

$$\langle A + B \rangle = \langle A \rangle + \langle B \rangle$$

$$\text{Var}(A + B) = \text{Var}(A) + \text{Var}(B)$$

And multiplying them (again this are now approximations)?

$$\langle AB \rangle \approx \langle A \rangle \langle B \rangle$$

$$\text{Var}(AB) \approx \langle B \rangle^2 \text{Var}(A) + \langle A \rangle^2 \text{Var}(B)$$

$$\frac{\text{Var}(AB)}{\langle AB \rangle^2} \approx \frac{\text{Var}(A)}{\langle A \rangle^2} + \frac{\text{Var}(B)}{\langle B \rangle^2}$$

Or division?

$$\left\langle \frac{A}{B} \right\rangle \approx \frac{\langle A \rangle}{\langle B \rangle}$$

$$\text{Var}\left(\frac{A}{B}\right) = \frac{\text{Var}(A)}{\langle B \rangle^2} + \frac{\text{Var}(B)}{\langle A \rangle^2}$$

3 One Rule to Rule Them All

This single rule allows us to propagate error in any situation, assuming the two variables are uncorrelated:

$$\text{Var}(f) \approx \left(\left. \frac{\partial f}{\partial A} \right|_{A=\langle A \rangle, B=\langle B \rangle} \right)^2 \text{Var}(A) + \left(\left. \frac{\partial f}{\partial B} \right|_{A=\langle A \rangle, B=\langle B \rangle} \right)^2 \text{Var}(B)$$

Thu 09 Oct 2025 09:00

Lecture 4 - Covariance and Correlation

Office Hours: Thursday 11am - 1pm, Physics West Rm 222 (b.becsy@bham.ac.uk)

Previously, when looking at two or more variables for error propagation/combinations etc, we assumed that they were independent of one another. Today we look at how to handle multiple variables which may be correlated.

1 Covariance

Covariance is a measure that indicates how much two variables fluctuate together:

$$\text{Cov}(x, y) = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})$$

Covariance matrices represent all combinations of covariance (noting $\text{Cov}(x, y) = \text{Cov}(y, x)$ and $\text{Cov}(x, x) = \text{Var}(x)$)

$$\Sigma = \begin{pmatrix} \text{Cov}(x, x) & \text{Cov}(x, y) \\ \text{Cov}(y, x) & \text{Cov}(y, y) \end{pmatrix}$$

We can then define correlation:

$$\text{Corr}(x, y) = \frac{\text{Cov}(x, y)}{\sqrt{\text{Var}(x)\text{Var}(y)}} = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (x_i - \bar{x})^2 \sum_{i=1}^N (y_i - \bar{y})^2}}$$

This is bounded between -1 ($x = -y$), 1 ($x = y$) and zero for no correlation. We can again put this in a matrix, noting it is symmetrical:

$$\begin{pmatrix} 1 & \text{Corr}(x, y) \\ \text{Corr}(y, x) & 1 \end{pmatrix}$$

1.1 Variable Combinations

Now, with correlated variables, we can say:

$$\langle x + y \rangle = \langle x \rangle + \langle y \rangle$$

$$\text{Var}(x, y) = \text{Var}(x) + \text{Var}(y) + 2\text{Cov}(x, y)$$

And (noting the mean slightly increases with correlated variables):

$$\langle xy \rangle = \langle x \rangle \langle y \rangle + \text{Cov}(x, y)$$

And the one formula to rule them all, taking correlation into account:

$$\text{Var}(f) \approx \frac{\partial f}{\partial A}^2 \text{Var}(A) + \frac{\partial f}{\partial B}^2 \text{Var}(B) + 2 \frac{\partial f}{\partial A} \frac{\partial f}{\partial B} \text{Cov}(A, B)$$

Wed 15 Oct 2025 12:00

Lecture 5 - Distributions

1 Coin Flips and Probability Recap

Flipping a coin is one of the simplest distributions we can create.

Given a:

$$\begin{aligned}P(H) &= 0.5 \\P(T) &= 0.5\end{aligned}$$

We know that $P(HHHH) = 0.5^4$.

And $P(\text{Three Heads and One Tail}) = P(\text{HHHT or HHTH or THHH or HTHH}) = 4 \times 0.5^4$. We will take 4 coins, A, B, C, D. We denote a single result as A_{Heads} or C_{Tails} etc.

We can also say that the coins are independent, i.e. the probability of one result given another result is equal to just the probability of the first result:

$$P(A_h|B_h) = 0.5$$

The chance of A and B being heads is:

$$P(A_{\text{h and B}_h}) = P(A_h) \times P(B_h) = P(HH)$$

The chance of A *or* B being heads is (noting *or* excludes the case where both are true):

$$P(A_h \text{ or } B_h) = P(A_h) + P(B_h) - P(A_h \text{ and } B_h)$$

1.1 Discrete Distribution

Lets consider flipping 4 coins and counting the number of heads. This forms a discrete distribution (where only 5 possible values are possible, 0, 1, 2, 3, 4). This distribution must be normalised (sum to 1), so:

$$\sum_r P(r) = 1$$

We can also consider the mean (expected) number of heads:

$$\langle r \rangle = \sum_r rP(r)$$

This function, $P(x)$ is called a *probability mass function*, and the sum of all values must be 1.

1.2 Continuous Distributions

Continuous distributions have similar conditions:

$$\int_{-\infty}^{\infty} P(x) dx = 1$$

$$\langle x \rangle = \int_{-\infty}^{\infty} xP(x) dx$$

And for the probability of the result lying between a and b :

$$\int_a^b P(x) dx$$

We cannot, in a continuous distribution consider the probability of an exact result, i.e. $P(x = a)$, $a \in \mathbb{R}$. As there are infinitely many possible values, the probability of any precise one is not meaningful (always zero). We therefore must always consider the probability of the result lying in some non-zero range.

$P(x)$ in this case is called a *probability density function* and the area under the PDF curve must sum to one. Note that this means that $P(x)$ at any point may exceed one, so long as the overall area is equal to 1. For some probability $a < P(x) < b$ (noting that since the $P(a)$ for any precise a is zero, the equalities being strict or not is meaningless), the probability is the area of the curve between a and b .

2 Binomial Distribution

The Binomial Distribution represents a scenario where we conduct some number of identical trials, where each trial has two possible outcomes (which we denote success and failure). For example, flipping a coin. Here:

- n - The number of trials.
- p - The probability of success.
- q - The probability of failure ($q = 1 - p$).
- r - The number of successes.

This has probability mass function:

$$P(r; n, p) = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r}$$

And has the following properties:

$$\langle r \rangle = np$$

$$Var(r) = np(1-p)$$

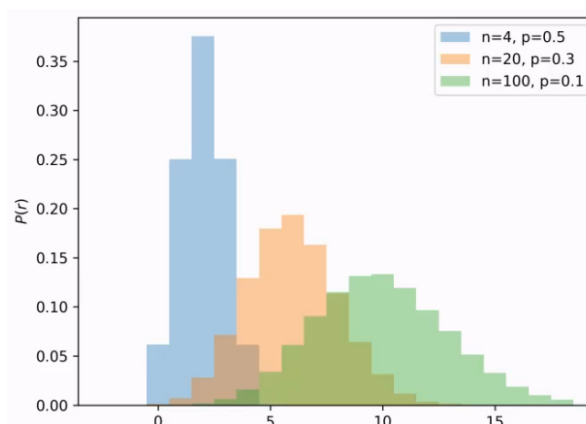


Figure 5.1: The Binomial Distribution

3 Poisson Distribution

This describes the number of events (i.e. the number of neutrinos detected by a neutrino detector) occurring in some time interval, given:

- The mean rate of events is constant.
- Each event occurs independently from the last.

This is created by taking the limit of a Binomial distribution, as:

- The number of trials tends to infinity ($n \rightarrow \infty$)
- The mean number of successes remains fixed ($np = \lambda = \text{constant}$)

Given λ as the mean number of expected events (per unit time) and r as the number of events occurring in that time, it has PMF:

$$P(r; \lambda) = \frac{\exp(-\lambda)\lambda^r}{r!}$$

And has the following properties:

$$\langle r \rangle = \lambda$$

$$\text{Var}(r) = \lambda$$

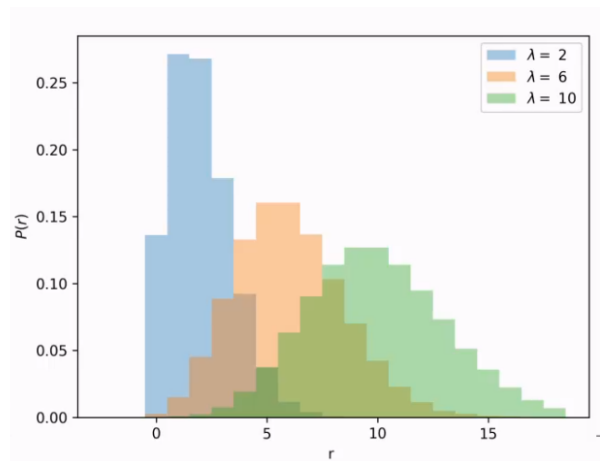


Figure 5.2: The Poisson Distribution

4 Normal Distribution

A.K.A. The Gaussian distribution. This is the most well known and most useful distribution. Given a mean μ and a standard deviation σ , the probability density function is:

$$P(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x - \mu)^2}{2\sigma^2}\right)$$

It is a very important distribution as it arises as a result of the Central Limit Theorem, which we will cover properly in the probability section of the course. It looks like this, noting it is symmetric and forms a “bell-shaped curve”:

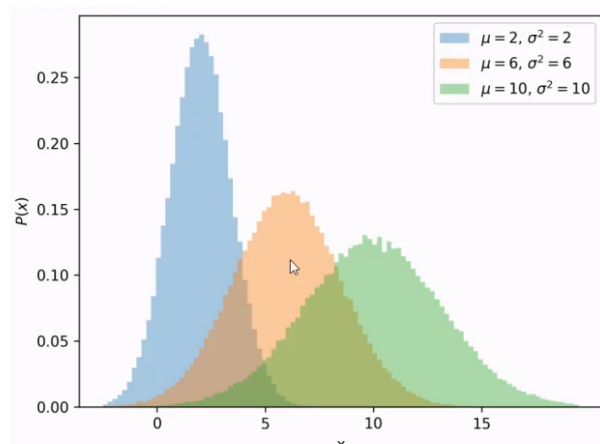


Figure 5.3: The Normal Distribution

Note that since the tails are logarithmic, they tend to zero, but never reach it truly. A Poisson distribution approaches a Normal distribution as $\lambda \rightarrow \infty$. It is generally a good approximation for $\lambda > 30$ but this depends on the application being used.

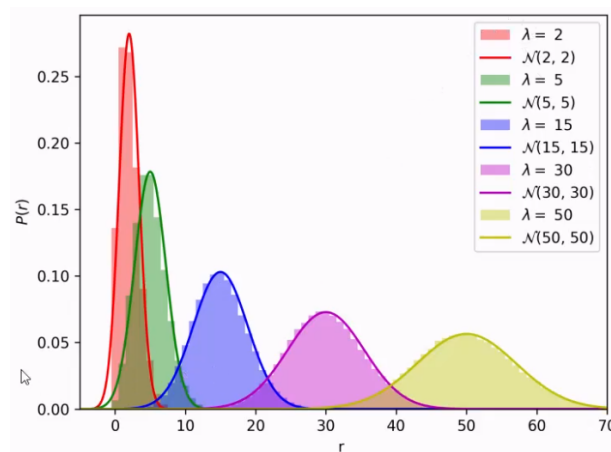


Figure 5.4: Poisson approximations to a Normal

Thu 16 Oct 2025 09:00

Lecture 6 - Likelihood and Log Likelihood

1 Likelihood

We want to fit a model to our data. We want some kind of function to specify how well this model fits the data, so that we can optimise to find the best. This is the likelihood function. There are many different ways to formulate it, but we denote it:

$$P(D \mid \theta)$$

Where D is our data, and θ is our model parameters. This is the probability of the data, given some parameters.

2 An Example

Lets say we have this model:

$$T(t) = T_{\text{env}} + (T_0 - T_{\text{env}}) \exp(-t/\tau)$$

Which represents the cooling of an object, where τ is a constant of cooling. We think we will observe some additive, normally distributed noise on these measurements, giving us:

$$T_{\text{obs}}(t) = T(t) + \epsilon$$

We may observe something like this, where the blue dots are the model-predicted values and the observed data with error noise is in black:

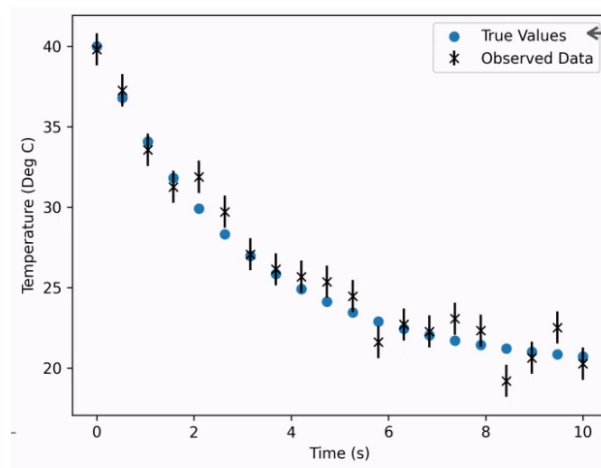


Figure 6.1: Simulated Data

Given the noise is normally distributed, we would expect the true values to lie within the error bars of our observation about 68% of the time. We see this approximately here. How can we then fit a model to this data? We need to:

1. Formulate a model.
2. Estimate a probability that the model is correct.

Given we now have data, and some model we would like to try to fit the data to (we want to fit it to Newton's Law of Cooling, the model previously, and determine an appropriate value of parameters and τ). We therefore want to find a 'merit function' to describe how good a fit any model we might create is. We start from the probability of getting some value of the noise.

As a reminder, our model is, noting we are treating time as a discrete set of times, indexed by i :

$$M(t_i, \theta) = T_{\text{env}} + (T_0 - T_{\text{env}}) \exp(-t_i/\tau)$$

And the probability of getting some value of the noise on the i th measurement is (note the first equality, where we can also write it ignoring theta, because noise is independent of the parameters):

$$P(\epsilon_i | \theta) = P(\epsilon_i) = \frac{1}{\sigma_{D_i} \sqrt{2\pi}} \exp\left(\frac{-\epsilon_i^2}{2\sigma_{D_i}^2}\right)$$

This is a Normal distribution with a mean of zero, and a standard deviation of σ_{D_i} .

We cannot directly measure ϵ_i , but we know it is the difference between the measured value in the data and the 'true' value predicted by our model:

$$\epsilon_i = D_i - M(t_i, \theta)$$

Since the noise is additive and Normal, we can combine these two equations to get our merit function - the probability of a single observed data point given the parameters as:

$$P(D_i | \theta) = \frac{1}{\sigma_{D_i} \sqrt{2\pi}} \exp\left(\frac{-(D_i - M(t_i, \theta))^2}{2\sigma_{D_i}^2}\right)$$

Where θ is our list of parameters $\theta = [T_0, T_{\text{env}}, \tau]$. Crucially, this is a Normal distribution where the mean is our model's prediction given the parameters, and the standard deviation is the uncertainty on the error point. Assuming we have multiple uncorrelated data points, the total likelihood function is:

$$P(D | \theta) = \prod_{i=1}^n P(D_i | \theta)$$

Considering τ as the variable we actually change, we get:

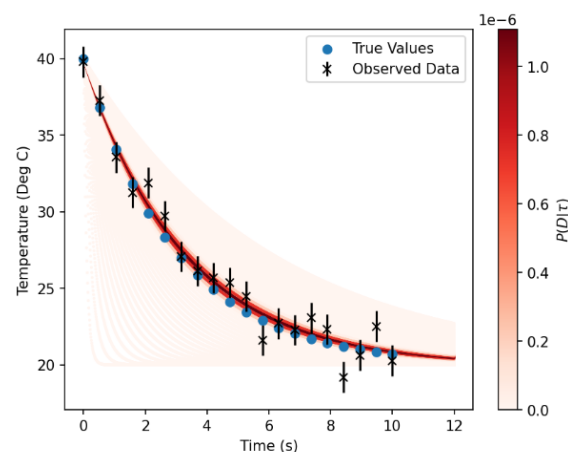


Figure 6.2: A heatmap of the likelihood function overlaid on the data.

The high likelihood values of $P(D | \tau)$ are the models which are most likely to generate the observed data, given the parameters. This, therefore, means that they are the models which best fit the data.

If we plot $P(D, \tau)$ against τ , we can see that the likelihood does a reasonable job of giving us a value which is close to true:

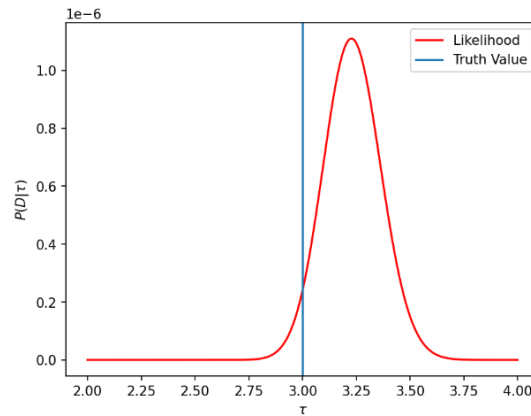


Figure 6.3

We can use Maximum Likelihood Estimation (note that it is just an estimate, the value of tau given by the maximum likelihood and the true value are **not** the same) to estimate the best value of tau for the model. We chose the value of tau that gives the maximum likelihood:

$$\hat{\tau} = \arg \max_{\tau} P(D | \tau)$$

In general, given a set of multiple parameters, we say:

$$\hat{\theta} = \arg \max_{\theta} P(D | \theta)$$

3 Log Likelihood

We still need to estimate the uncertainty on this predicted best value of tau. It turns out that a good way to do this is by taking the log likelihood instead of just the likelihood. We take the natural log of the normal probability density function for a single data point:

$$\begin{aligned} P(D_i | \theta) &= \frac{1}{\sigma_{D_i} \sqrt{2\pi}} \exp \left(\frac{-(D_i - M(t_i, \theta))^2}{2\sigma_{D_i}^2} \right) \\ \ln P(D_i | \theta) &= \ln \left(\frac{1}{\sigma_{D_i} \sqrt{2\pi}} \exp \left(\frac{-(D_i - M(t_i, \theta))^2}{2\sigma_{D_i}^2} \right) \right) \\ \ln P(D_i | \theta) &= \ln \left(\frac{1}{\sigma_{D_i} \sqrt{2\pi}} \right) + \ln \left(\exp \left(\frac{-(D_i - M(t_i, \theta))^2}{2\sigma_{D_i}^2} \right) \right) \\ \ln P(D_i | \theta) &= \left(\exp \left(\frac{-(D_i - M(t_i, \theta))^2}{2\sigma_{D_i}^2} \right) \right) - \ln(\sigma_{D_i}) - \ln(\sqrt{2\pi}) \end{aligned}$$

If we ignore the constant term, as it does not change the results (as we care about the results comparative to each other to find the maximum), and if we assume that the uncertainty is the same on each data point (which we must be careful about, in case uncertainties are variables too):

$$\ln P(D_i | \theta) = \left(\exp \left(\frac{-(D_i - M(t_i, \theta))^2}{2\sigma_{D_i}^2} \right) \right) + \text{constant}$$

This gives the final log likelihood as:

$$\mathcal{L} = \ln P(D | \theta) = \sum_{i=1}^n \ln P(D_i | \theta) = \sum_{i=1}^n \left(\exp \left(\frac{-(D_i - M(t_i, \theta))^2}{2\sigma_{D_i}^2} \right) \right)$$

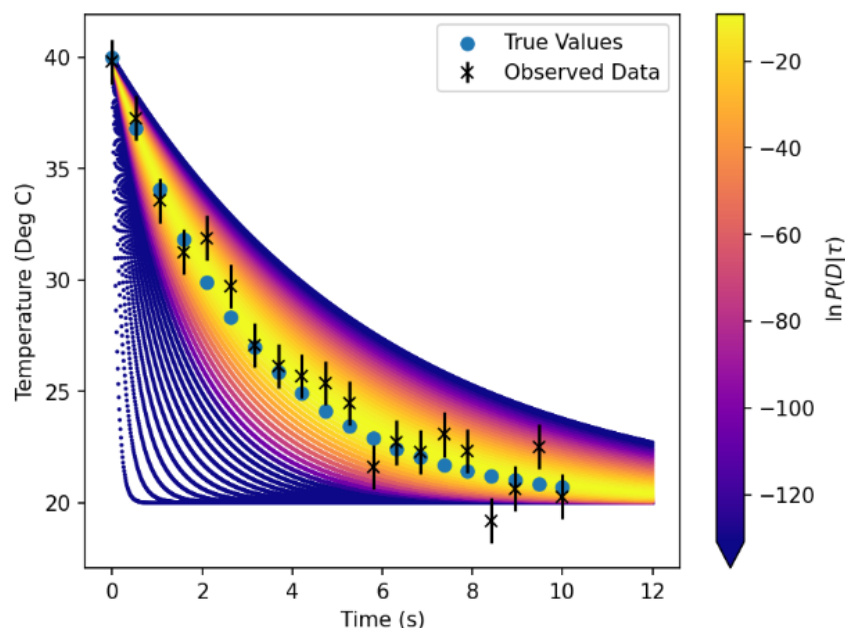


Figure 6.4: The likelihood plot, repeated with log likelihood

We can see this generates similar values to the standard likelihood, but with much friendlier values (-20 to -120, rather than very small numbers). The equation is also nicer to calculate as we're able to get rid of the constant terms. If we again consider this as a function of tau (the parameter we're actually changing to produce a fit):

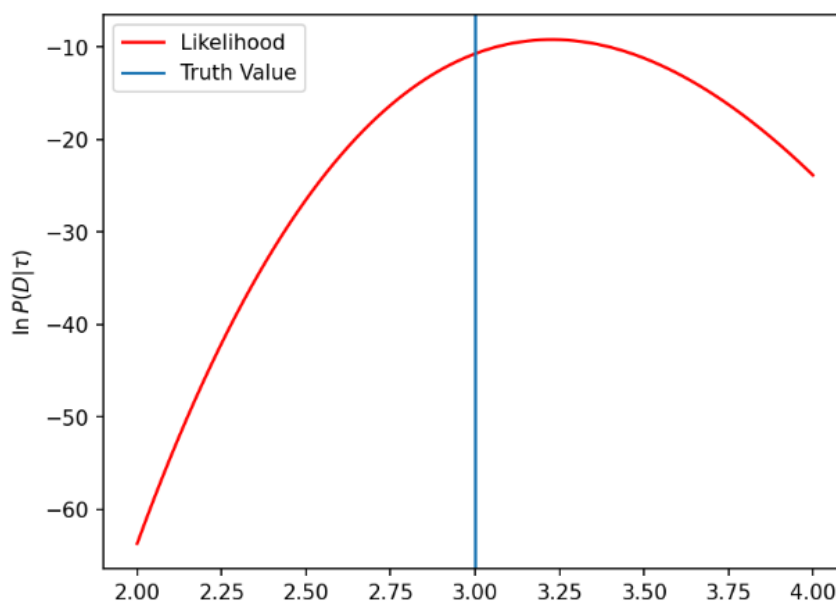


Figure 6.5

This is similar again, but with a very different order of magnitude. We can calculate the predicted optimum τ by determining the maximum point where:

$$\frac{\partial}{\partial \tau} \ln P(D | \tau) = 0$$

Note the swap from theta to tau, this is because tau is the parameter we're actually using to fit, while the other parameters bundled into theta are constant. This gives $\tau \approx 3.23$. Yes we could have done this with traditional likelihood, but the differentiation is nicer in log form and it becomes relevant later.

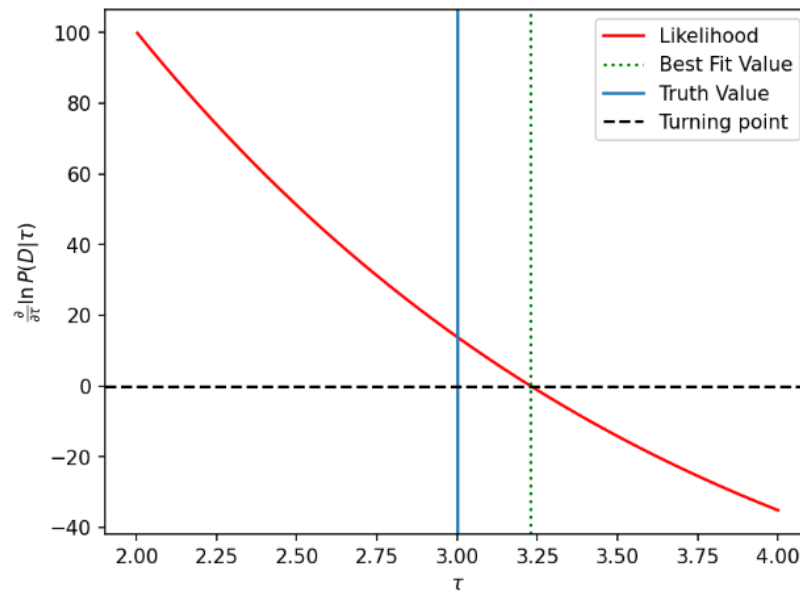


Figure 6.6

4 Uncertainties on Log Likelihood

In our single dimension problem, we'll create a new value called the Hessian (H), we define this as:

$$H = \left. \frac{\partial^2 \mathcal{L}}{\partial \tau^2} \right|_{\tau=\hat{\tau}}$$

Why is this (and the log likelihood) actually useful? It turns out that the curvature of the log likelihood around the maximum point tells us the uncertainty. We can use the Hessian and log likelihood to estimate the uncertainty on the parameter τ , using the inverse of H , H^{-1} where $HH^{-1} = 1$.

$$\sigma_{\hat{\tau}} \approx \sqrt{|H^{-1}|}$$

Since we're in 1D, this value of the Hessian is a scalar, and we can estimate it by either differentiating twice or estimating using two points:

$$f''(x) = \lim_{h \rightarrow 0} \frac{f(x+h) - 2f(x) + f(x-h)}{h^2}$$

And calculating:

$$\sigma_{\hat{\tau}} = \sqrt{\frac{1}{|H|}}$$

But note that this works differently in higher dimensions (more parameters) as H becomes a matrix. This gives us $\hat{\tau} \approx 3.23$ from the previous result, and now $\sigma_{\hat{\tau}} \approx 0.133$. Note that the true value is 1.8 standard deviations away from the best estimate, which is okay - anything larger than 3 sigma away would start to become worrying.

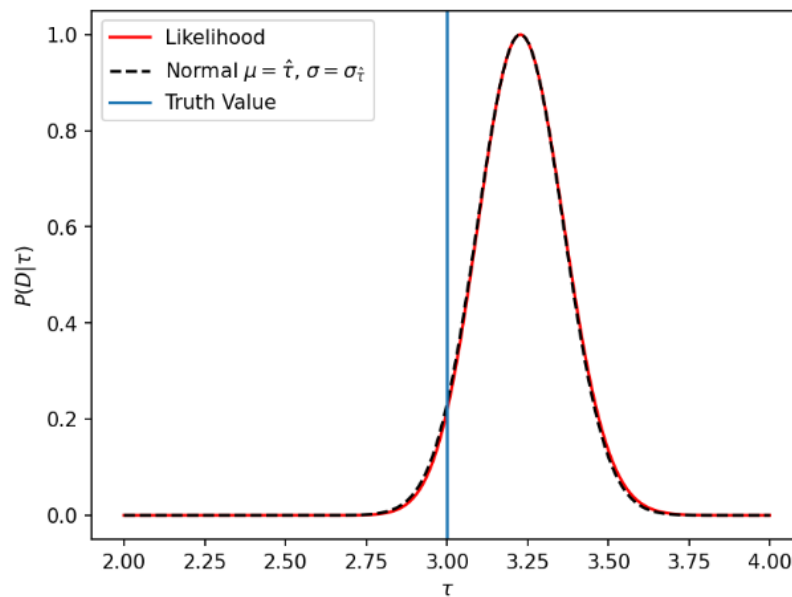


Figure 6.7

Using the log likelihood has:

- Been a nicer calculation, which is computationally easier, as we can strip out all the constant terms.
- Allowed us to calculate the uncertainties on our predicted values.
- Allowed us to get back to the standard likelihood (as the black and red curves above are approximately the same) anyways.

4.1 Multidimensional Generalisation

This works in 1D, but we can generalise to a higher number of parameters. We assume here however that $P(D | \theta)$ is going to be a normal distribution. If this is not the case, we cannot use the approximations for uncertainties, and the problem becomes too complex for first year stats. It is generally the case that it either will approximate a normal as many datapoints are taken. If this isn't true, it becomes a problem for Y4 Bayesian Stats.

We find the maximum likelihood (given multiple parameters $\hat{\theta}$) with:

$$\frac{\partial \mathcal{L}}{\partial \theta_1} = \frac{\partial \mathcal{L}}{\partial \theta_2} = \dots = 0$$

And the Hessian is given by:

$$\mathbf{H}(\hat{\theta}) = \begin{bmatrix} \frac{\partial^2 \mathcal{L}}{\partial \theta_1^2} \Big|_{\theta=\hat{\theta}} & \frac{\partial^2 \mathcal{L}}{\partial \theta_1 \partial \theta_2} \Big|_{\theta=\hat{\theta}} & \cdots & \frac{\partial^2 \mathcal{L}}{\partial \theta_1 \partial \theta_k} \Big|_{\theta=\hat{\theta}} \\ \frac{\partial^2 \mathcal{L}}{\partial \theta_2 \partial \theta_1} \Big|_{\theta=\hat{\theta}} & \frac{\partial^2 \mathcal{L}}{\partial \theta_2^2} \Big|_{\theta=\hat{\theta}} & \cdots & \frac{\partial^2 \mathcal{L}}{\partial \theta_2 \partial \theta_k} \Big|_{\theta=\hat{\theta}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \mathcal{L}}{\partial \theta_k \partial \theta_1} \Big|_{\theta=\hat{\theta}} & \frac{\partial^2 \mathcal{L}}{\partial \theta_k \partial \theta_2} \Big|_{\theta=\hat{\theta}} & \cdots & \frac{\partial^2 \mathcal{L}}{\partial \theta_k^2} \Big|_{\theta=\hat{\theta}} \end{bmatrix}.$$

Wed 22 Oct 2025 12:00

Lecture 7 - Fitting a Straight Line 1

We want to create a model for a straight line:

$$M(x, \theta) = mx + c$$

Where datapoints are given by this model and some additive noise:

$$D = M(x, \theta) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma_i)$$

The general recipe for line fitting is given by:

1. A generative model for the data, with knowledge of how the noise is distributed.
2. Likelihood function.
3. A method for finding the maximum likelihood.
4. Method for finding the uncertainties on best fit parameters.
5. A method for checking how good the fit is.

We can write down the likelihood function for this model as:

$$P(D_i | \theta) = \frac{1}{\sigma_{D_i} \sqrt{2\pi}} \exp \left(\frac{-(D_i - M(x_i, \theta))^2}{2\sigma_{D_i}^2} \right)$$

$$P(D | \theta) = \prod_{i=1}^n P(D_i | \theta)$$

And again:

$$\mathcal{L} = \ln P(D | \theta) = \sum_{i=1}^n \ln P(D_i | \theta) \propto \sum_{i=1}^n \left(\frac{-(D_i - M(x_i, \theta))^2}{2\sigma_{D_i}^2} \right)$$

We want to find the parameters of distribution that maximise the (log)likelihood:

$$\hat{\theta} = \arg \max_{\theta} P(D | \theta)$$

Or:

$$\hat{\theta} = \arg \max_{\theta} \mathcal{L}$$

There are a number of different approaches to do this:

- Find where all first derivatives equal zero (as last lecture, various clever algorithms to do so).
- Brute force on a grid.
- Iterative or stochastic methods.
- Analytic maximisation for a simple linear model - see next lecture.

1 Finding Maximum Likelihood

The most crude way to do this is to build a grid of all values of m and c , and iterate through over all points (with some resolution) to find the maximum likelihood generated by them. Plotting M (as colour):

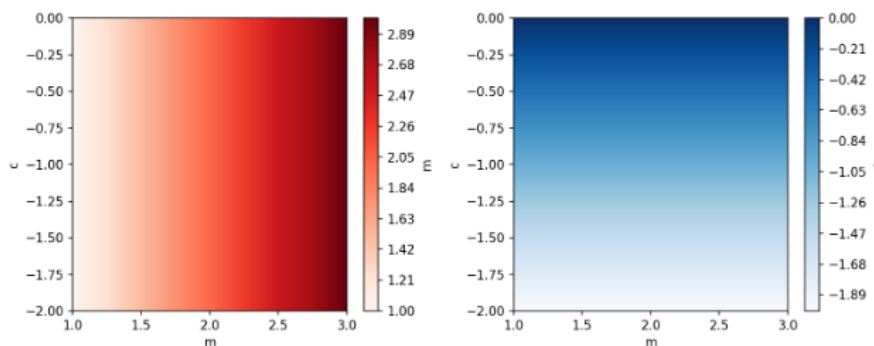


Figure 7.1

And the likelihood:

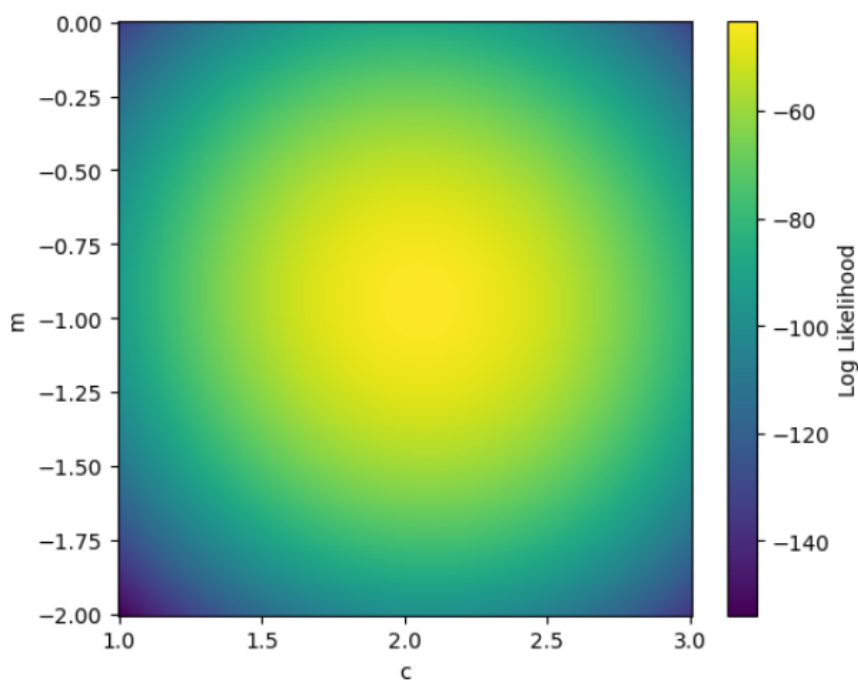


Figure 7.2

We assume that the grid point with the highest likelihood and the true point with the highest likelihood are the same. In this case, the grid resolution is small enough that this is true, but it may not always be. This gives:

$$m = 2.0552763819095476$$

$$c = -0.9447236180904524$$

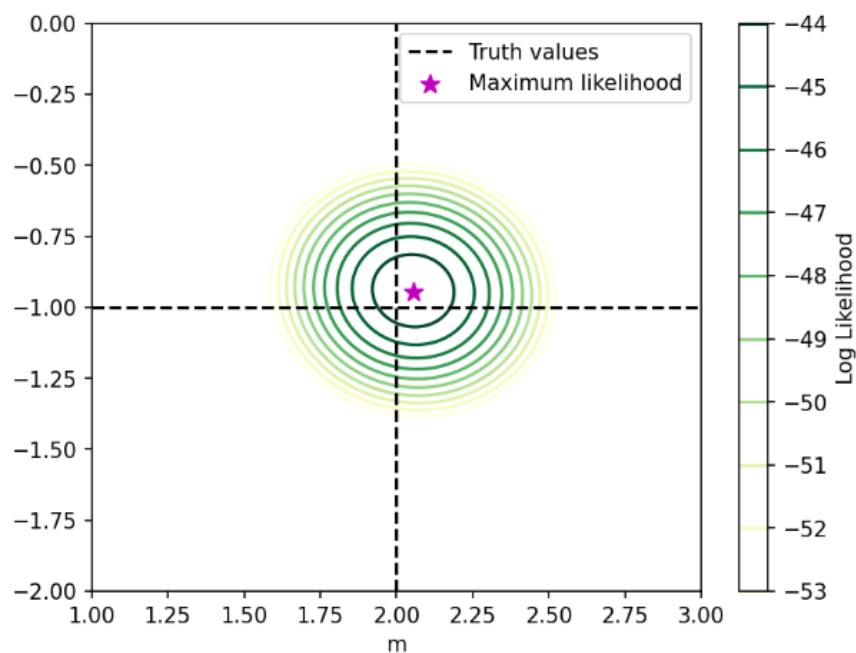


Figure 7.3

We see that this does a good, but not perfect job, of fitting the data. This is due to noise in the data, and is acceptable, provided it's within a reasonable uncertainty.

Thu 23 Oct 2025 09:00

Lecture 8 - Fitting a Straight Line 2

1 Uncertainties on Best Fit Parameters

1.1 What do these uncertainties actually mean?

Previously, we found that the likelihood of a single value could be described by a Normal distribution.

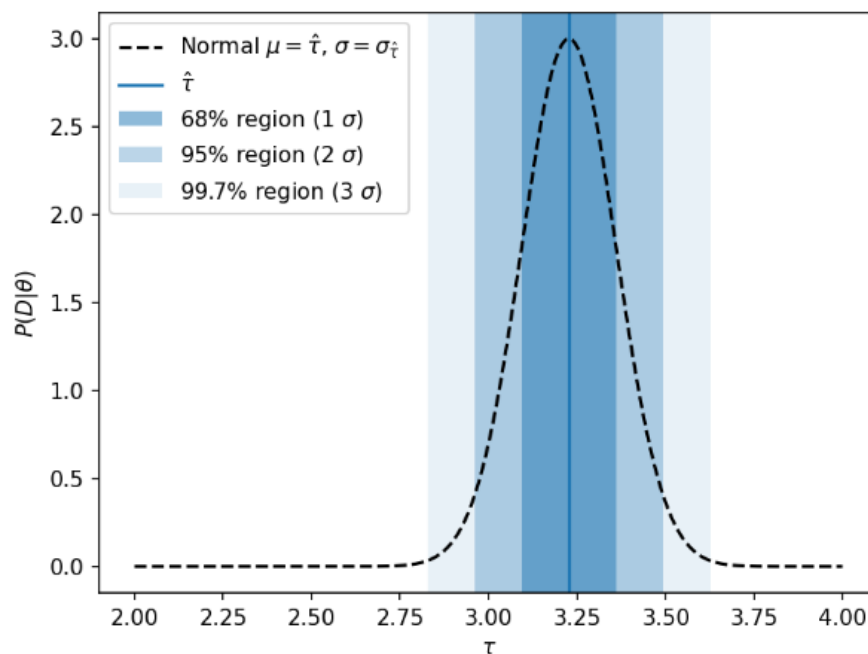


Figure 8.1

We can therefore quote the best fit value as $\hat{\tau} = 3.227 \pm 0.133$. What we're effectively saying is that our 1 sigma uncertainty encompasses 68% of the probability density such that:

$$\int_{\hat{\tau} - \sigma_{\hat{\tau}}}^{\hat{\tau} + \sigma_{\hat{\tau}}} P(D | \tau) d\tau \approx 0.68$$

And the same for 2 sigma uncertainty with 0.95, and 3 sigma uncertainty with 0.997. We can therefore say that while it may be common for the value to lie outside the 1 sigma uncertainty, it is rare for it to lie outside the 3 sigma uncertainty and if this happens (including error), something probably went wrong with our measurement.

Here, where we're attempting to determine an uncertainty, we instead increase the value of $\sigma_{\hat{\tau}}$ until this first integral is satisfied.

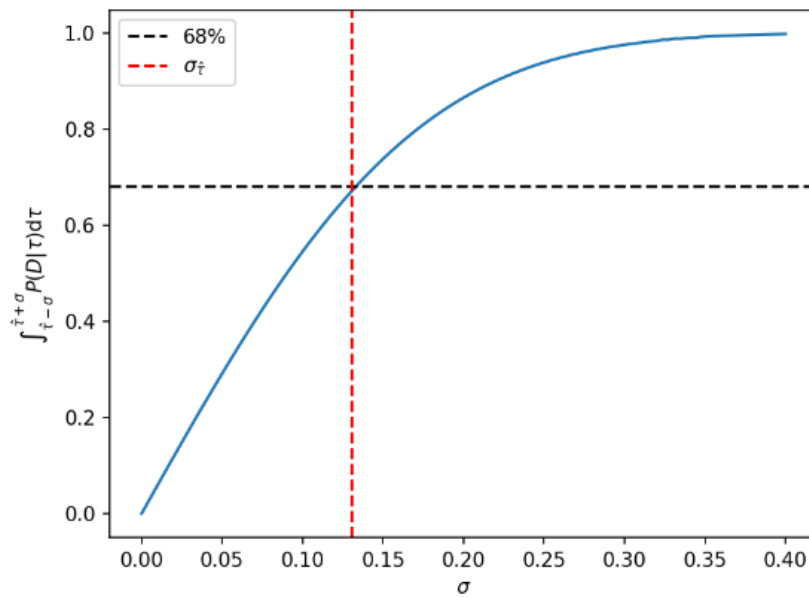


Figure 8.2

However, with a line of best fit, we want to consider uncertainties in two dimensions. This makes life a little bit more difficult, we want some boundary on the parameter space that is centred on the best fit parameters and encompasses 68% of the whole probability space.

Plotting the same distribution again, but including the 1D likelihood for each parameter individually:

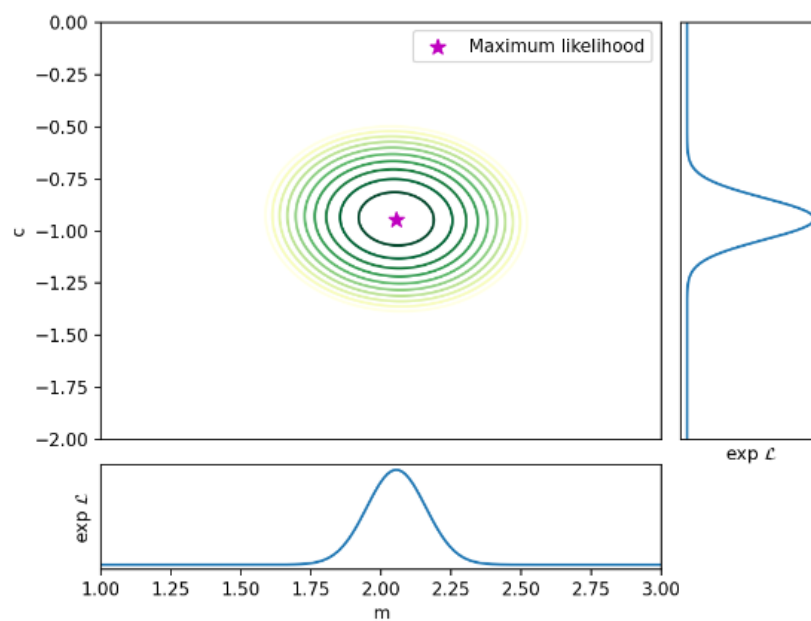


Figure 8.3

In general, we shouldn't just separately calculate the 1 dimensional approach per parameter and combine them (as the parameters may be correlated), but it's a useful starting point. Swapping to log likelihood:

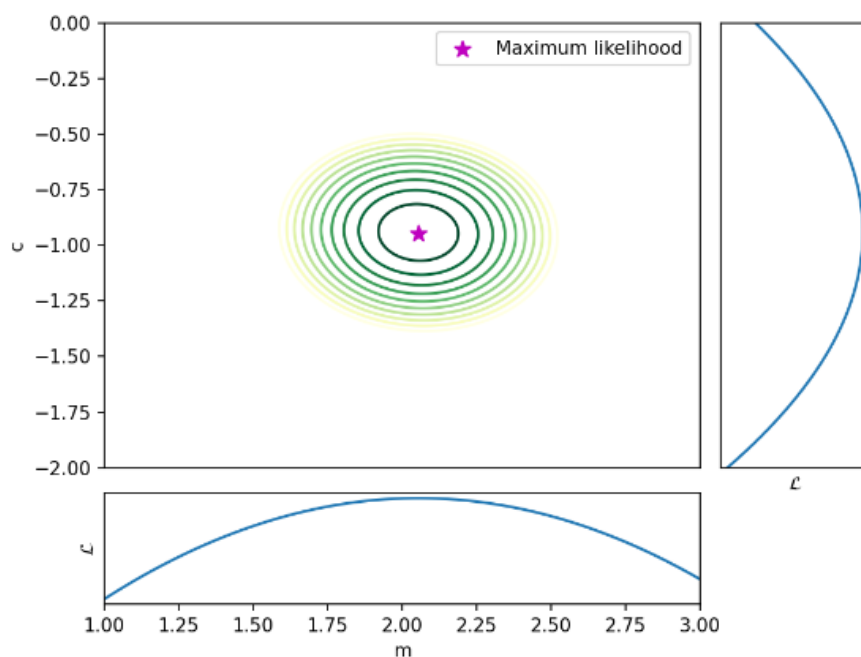


Figure 8.4

We want to estimate:

$$\left. \frac{\partial^2 \mathcal{L}}{\partial m^2} \right|_{\theta=\hat{\theta}}$$

$$\left. \frac{\partial^2 \mathcal{L}}{\partial c^2} \right|_{\theta=\hat{\theta}}$$

While we shouldn't use these 1D distributions to estimate the best fit parameters, this problem has deliberately been created to minimise the correlation between m and c and creating something with them that describes $P(D | \theta)$ is still instructive. We can however create a grid of the values of m and c , calculate the likelihood for all points with some resolution and pick the maximum. This is slow, but crucially does find the maximum point of the 3D distribution surface and is not the same as finding the max for each variable individually.

We assume that $P(D | \theta)$ can be described by a two dimensional normal distribution, and we can build this from two discrete normal distributions of two independent variables (with the caveats above). We say that the mean value of $P(D | m)$ is \hat{m} and the uncertainty is therefore given by:

$$\sigma_m^2 = \left(- \left. \frac{\partial^2 \mathcal{L}}{\partial m^2} \right|_{\theta=\hat{\theta}} \right)^{-1}$$

And likewise for c , \hat{c} , σ_c^2

$$\sigma_c^2 = \left(- \left. \frac{\partial^2 \mathcal{L}}{\partial c^2} \right|_{\theta=\hat{\theta}} \right)^{-1}$$

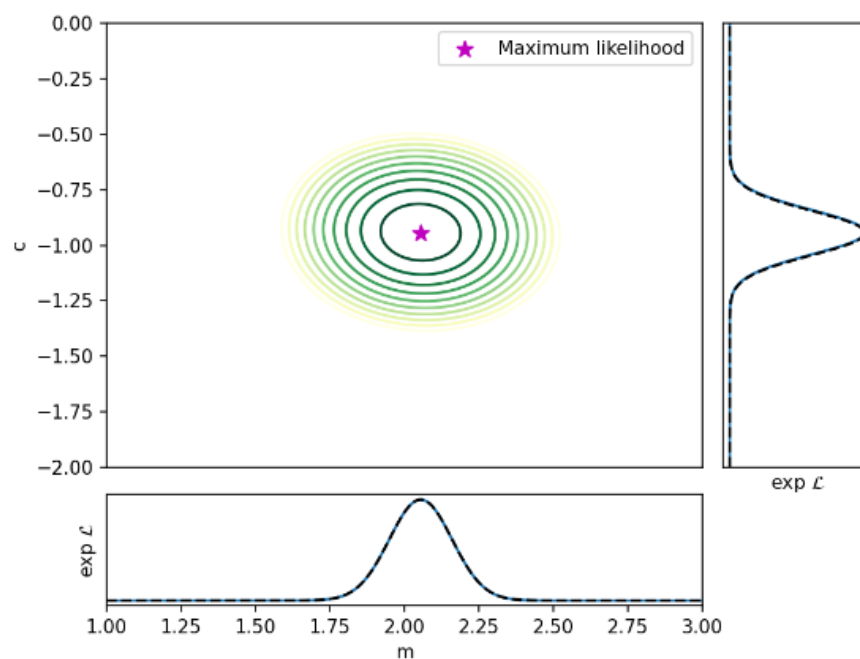


Figure 8.5

Calculating our summary stats and plotting them:

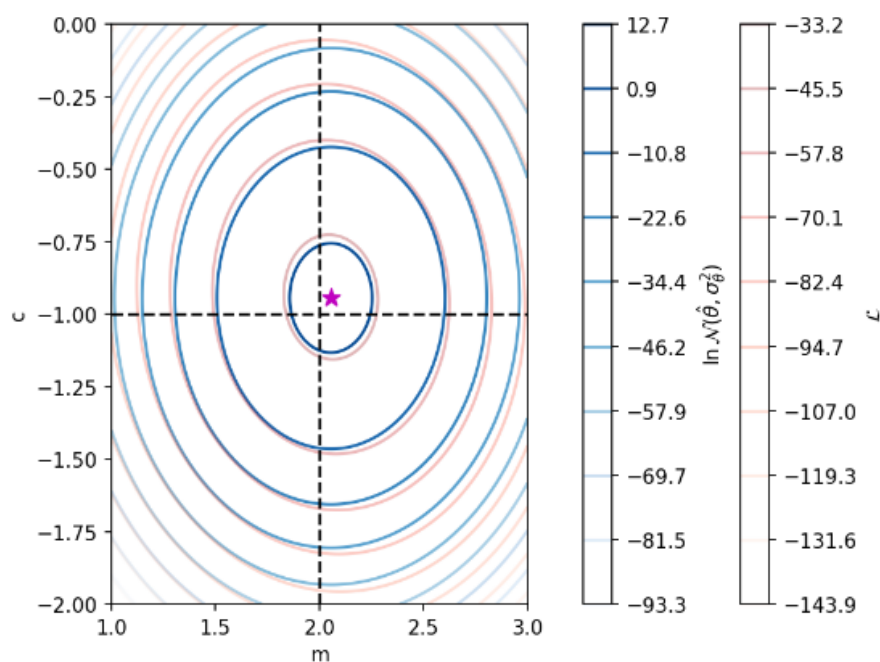


Figure 8.6

We can see that this is pretty good agreement between our estimate and the actual log likelihood function. We now have everything we need to quote the best fit parameters and (crucially) their uncertainty.

Wed 29 Oct 2025 12:00

Lecture 9 - Linear Regression

Previously, we've used numerical methods to determine the best fit parameters for a line of best fit to some data. This is good, because it easily generalises to more complex problems, however lines of best fit have a more specific (and easier to implement) algebraic method.

1 Method

And, we have to perform the same broad steps:

1. A generative model for the data, with knowledge of how the noise is distributed.
2. Likelihood function.
3. A method for finding the maximum likelihood.
4. Method for finding the uncertainties on best fit parameters.
5. A method for checking how good the fit is.

1.1 Generative Model

We use the same generative model as before, with a straight line fit with some additively generated noise (with a standard deviation that may differ from point to point, σ_{D_i}):

$$M(x, \theta) = mx + c$$

1.2 Likelihood Function

Using the same likelihood and log likelihood formulae as before, we can take this one step further by defining (as the extra factor of -2 does not matter when calculating the maxima/minima and ignoring it will make the algebra nicer):

$$\chi^2 = -2\mathcal{L} = \sum_{i=1}^n \left(\frac{D_i - M(x_i, \theta)}{\sigma_{D_i}} \right)^2$$

1.3 Finding Maximum Likelihood (Minimum χ^2)

Since our model is linear, there will only be one turning point for χ^2 , so we can be sure that the minimum of χ^2 will be at the point where the first derivatives (wrt m and c) are zero.

$$\frac{\partial(\chi^2)}{\partial m} = \frac{\partial(\chi^2)}{\partial c} = 0$$

Rather than doing this numerically, as last lecture, we can do it algebraically:

$$\chi^2 = \sum_{i=1}^n \left(\frac{D_i - M(x_i, \theta)}{\sigma_{D_i}} \right)^2 = \sum_{i=1}^n \left(\frac{D_i - mx_i - c}{\sigma_{D_i}} \right)^2$$

w.r.t m:

$$\begin{aligned}\frac{\partial(\chi^2)}{\partial m} &= \frac{\partial}{\partial m} \sum_{i=1}^n \left(\frac{D_i - mx_i - c}{\sigma_{D_i}} \right)^2 = 0 \\ &= \sum_{i=1}^n \frac{\partial}{\partial m} \left(\frac{D_i - mx_i - c}{\sigma_{D_i}} \right)^2\end{aligned}$$

Let $u_i = (D_i - mx_i - c)/\sigma_{D_i}$, so $\partial u_i/\partial m = -x_i/\sigma_{D_i}$:

$$\begin{aligned}&= \sum_{i=1}^n \frac{\partial u_i}{\partial m} \frac{\partial}{\partial u_i} u_i^2 \\ &= \sum_{i=1}^n \frac{\partial u_i}{\partial m} (2u_i) \\ &= \sum_{i=1}^n \left(\frac{-x_i}{\sigma_{D_i}} \right) (2u_i) \\ &\quad -2 \sum_{i=1}^n \frac{x_i u_i}{\sigma_{D_i}}\end{aligned}$$

So:

$$\frac{\partial(\chi^2)}{\partial m} = -2 \sum_{i=1}^n \frac{x_i}{\sigma_{D_i}} \frac{D_i - mx_i - c}{\sigma_{D_i}}$$

$$\boxed{\frac{\partial(\chi^2)}{\partial m} = -2 \sum_{i=1}^n \left(\frac{x_i(D_i - mx_i - c)}{\sigma_{D_i}^2} \right) = 0}$$

w.r.t c:

$$\begin{aligned}\frac{\partial(\chi^2)}{\partial c} &= \frac{\partial}{\partial c} \sum_{i=1}^n \left(\frac{D_i - mx_i - c}{\sigma_{D_i}} \right)^2 = 0 \\ &= \sum_{i=1}^n \frac{\partial}{\partial c} \left(\frac{D_i - mx_i - c}{\sigma_{D_i}} \right)^2\end{aligned}$$

Let $u_i = (D_i - mx_i - c)/\sigma_{D_i}$, so $\partial u_i/\partial c = -1/\sigma_{D_i}$:

$$\begin{aligned}&= \sum_{i=1}^n \frac{\partial u_i}{\partial c} \frac{\partial}{\partial u_i} u_i^2 \\ &= \sum_{i=1}^n \frac{\partial u_i}{\partial c} (2u_i) \\ &= \sum_{i=1}^n \left(\frac{-1}{\sigma_{D_i}} \right) (2u_i) \\ &= -2 \sum_{i=1}^n \frac{u_i}{\sigma_{D_i}}\end{aligned}$$

So:

$$\boxed{\frac{\partial(\chi^2)}{\partial c} = -2 \sum_{i=1}^n \frac{(D_i - mx_i - c)}{\sigma_{D_i}^2} = 0}$$

2 Putting it All Together

We now have two simultaneous equations with two variables (m, c) , so we can solve for the optimum values:

$$\frac{\partial(\chi^2)}{\partial c} = -2 \sum_{i=1}^n \frac{(D_i - mx_i - c)}{\sigma_{D_i}^2} = 0$$

$$\frac{\partial(\chi^2)}{\partial m} = -2 \sum_{i=1}^n \frac{x_i(D_i - mx_i - c)}{\sigma_{D_i}^2} = 0$$

We can make some substitutions for ease:

$$S = \sum_{i=1}^n \frac{1}{\sigma_{D_i}^2}$$

$$S_x = \sum_{i=1}^n \frac{x_i}{\sigma_{D_i}^2}$$

$$S_{xx} = \sum_{i=1}^n \frac{x_i^2}{\sigma_{D_i}^2}$$

$$S_D = \sum_{i=1}^n \frac{D_i}{\sigma_{D_i}^2}$$

$$S_{Dx} = \sum_{i=1}^n \frac{x_i D_i}{\sigma_{D_i}^2}$$

2.1 Subbing into $\partial(\chi^2)/\partial c$

$$-2 \sum_{i=1}^n \frac{(D_i - mx_i - c)}{\sigma_{D_i}^2} = 0$$

$$\sum_{i=1}^n \frac{(D_i - mx_i - c)}{\sigma_{D_i}^2} = 0$$

$$S_D - mS_x - cS = 0 \implies S_D = mS_x + cS$$

2.2 Subbing into $\partial(\chi^2)/\partial m$

$$-2 \sum_{i=1}^n \frac{x_i(D_i - mx_i - c)}{\sigma_{D_i}^2} = 0$$

$$\sum_{i=1}^n \frac{x_i(D_i - mx_i - c)}{\sigma_{D_i}^2} = 0$$

$$S_{Dx} - mS_{xx} - cS_x = 0 \implies S_{Dx} = mS_{xx} + cS_x$$

2.3 Combining

We now have two simultaneous equations to solve:

$$\begin{cases} S_D = mS_x + cS & (1) \\ S_{Dx} = mS_{xx} + cS_x & (2) \end{cases}$$

Rearranging (1) gives:

$$c = \frac{S_D - mS_x}{S} \quad (3)$$

And (3) into (2):

$$\begin{aligned}
 S_{Dx} &= mS_{xx} + \frac{S_d - mS_x}{S} S_x \\
 S_{Dx} &= mS_{xx} + \frac{S_D S_x}{S} - \frac{m(S_x)^2}{S} \\
 S_{Dx} - \frac{S_D S_x}{S} &= mS_{xx} - \frac{m(S_x)^2}{S} \\
 S_{Dx} - \frac{S_D S_x}{S} &= m \left(S_{xx} - \frac{(S_x)^2}{S} \right) \\
 m &= \left(S_{Dx} - \frac{S_D S_x}{S} \right) / \left(S_{xx} - \frac{(S_x)^2}{S} \right)
 \end{aligned}$$

Simplifying to:

$$m = \frac{SS_{Dx} - S_D S_x}{SS_{xx} - S_x^2}$$

And for c , rearranging (2) gives:

$$m = \frac{S_D - cS}{S_x} \quad (4)$$

(4) into (2)

$$\begin{aligned}
 S_{Dx} &= \frac{S_D - cS}{S_x} S_{xx} + cS_x \\
 S_{Dx} &= \frac{S_D S_{xx}}{S_x} - \frac{cSS_{xx}}{S_x} + cS_x \\
 S_{Dx} - \frac{S_D S_{xx}}{S_x} &= cS_x - \frac{cSS_{xx}}{S_x} \\
 S_{Dx} - \frac{S_D S_{xx}}{S_x} &= c \left(S_x - \frac{SS_{xx}}{S_x} \right) \\
 c &= \left(S_{Dx} - \frac{S_D S_{xx}}{S_x} \right) / \left(S_x - \frac{SS_{xx}}{S_x} \right)
 \end{aligned}$$

Simplifying to:

$$c = \frac{S_D S_{xx} - S_x S_{Dx}}{SS_{xx} - S_x^2}$$

2.4 And Finally...

To simplify, let $\Delta = SS_{xx} - S_x^2$:

$$\begin{aligned}
 \hat{c} \equiv \langle c \rangle &= \frac{S_D S_{xx} - S_x S_{Dx}}{\Delta} \\
 \hat{m} \equiv \langle m \rangle &= \frac{SS_{Dx} - S_D S_x}{\Delta}
 \end{aligned}$$

We have therefore managed to calculate the best fit parameters \hat{m} and \hat{c} in closed form without any numerical methods.

3 Uncertainties on Best Fit Parameters

Given we're now in 2D, the Hessian matrix is given as:

$$\mathbf{H} = \begin{bmatrix} \frac{\partial^2 \mathcal{L}}{\partial \theta_1^2} \big|_{\theta=\hat{\theta}} & \frac{\partial^2 \mathcal{L}}{\partial \theta_1 \partial \theta_2} \big|_{\theta=\hat{\theta}} \\ \frac{\partial^2 \mathcal{L}}{\partial \theta_2 \partial \theta_1} \big|_{\theta=\hat{\theta}} & \frac{\partial^2 \mathcal{L}}{\partial \theta_2^2} \big|_{\theta=\hat{\theta}} \end{bmatrix}$$

And the covariance matrix Σ is given as:

$$\Sigma = -\mathbf{H}^{-1}$$

We therefore need to calculate the relevant second derivatives. Note that while we could ignore the -2 term before, as it did not matter for finding the location of the maximum, it does matter for errors and cannot be left off. We therefore go back to working in \mathcal{L} and not χ^2

We know the first derivatives of the log likelihood are:

$$\frac{\partial(\chi^2)}{\partial m} = -2(S_{Dx} - mS_{xx} - cS_x) \implies \frac{\partial \mathcal{L}}{\partial m} = S_{Dx} - mS_{xx} - cS_x$$

$$\frac{\partial(\chi^2)}{\partial c} = -2(S_D - mS_x - cS) \implies \frac{\partial \mathcal{L}}{\partial c} = S_D - mS_x - cS$$

Taking second derivatives:

$$\frac{\partial^2 \mathcal{L}}{\partial m^2} = \frac{\partial}{\partial m} (S_{Dx} - mS_{xx} - cS_x) = -S_{xx}$$

$$\frac{\partial^2 \mathcal{L}}{\partial c^2} = \frac{\partial}{\partial c} (S_D - mS_x - cS) = -S$$

And for the term wrt both variables:

$$\frac{\partial^2 \mathcal{L}}{\partial m \partial c} = \frac{\partial^2 \mathcal{L}}{\partial c \partial m} = \frac{\partial}{\partial c} (S_{Dx} - mS_{xx} - cS_x) = -S_x$$

Hence (taking $\theta_1 = m$, $\theta_2 = c$):

$$\mathbf{H} = \begin{bmatrix} -S_{xx} & -S_x \\ -S_x & -S \end{bmatrix}$$

And finally:

$$\Sigma = -\mathbf{H}^{-1} = \frac{1}{SS_{xx} - S_x^2} \begin{bmatrix} S & -S_x \\ -S_x & S_{xx} \end{bmatrix} = \frac{1}{\Delta} \begin{bmatrix} S & -S_x \\ -S_x & S_{xx} \end{bmatrix},$$

Therefore (from the definitions of the covariance matrix):

$$Var(m) = \Sigma_{11} = \frac{S}{\Delta}$$

$$Var(c) = \Sigma_{22} = \frac{S_{xx}}{\Delta}$$

$$Cov(m, c) = \Sigma_{12} = \Sigma_{21} = -\frac{S_x}{\Delta}$$

And using the definition of correlation:

$$Cor(m, c) = \frac{Cov(m, c)}{\sqrt{Var(m)Var(c)}} = \frac{-S_x}{\sqrt{SS_{xx}}}$$

Thu 30 Oct 2025 09:00

Lecture 10 - Goodness of Fit

Using likelihood, we can quantify how close our model is to the data, but how do we know if we've got the right model in the first place? We can use a *goodness of fit statistic* to quantify this probabilistically.

1 Straight Line Example

Lets start by taking some data and fitting a straight line to it, using last lectures content:

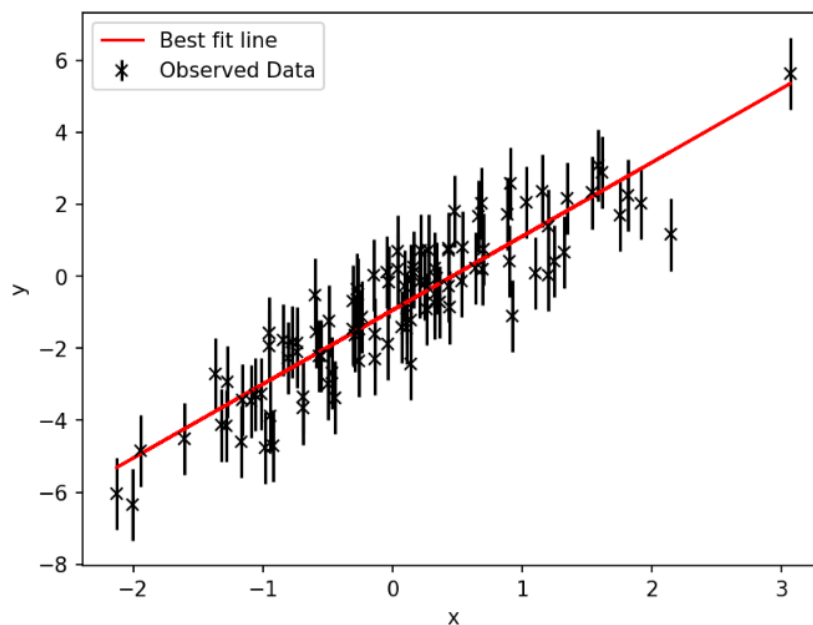


Figure 10.1

We can calculate the χ^2 value for this fit with:

$$\chi^2 = \sum_{i=1}^n \left(\frac{D_i - M(x_i, \theta)}{\sigma_{D_i}} \right)^2$$

Where $M(x_i, \theta) = \hat{m}x_i + \hat{c}$. This gives $\chi^2 \approx 86.37$. We can generate many different datasets, and create lines of best fit for them, and calculate each fit's χ^2 . If we plot these values as a histogram we get:

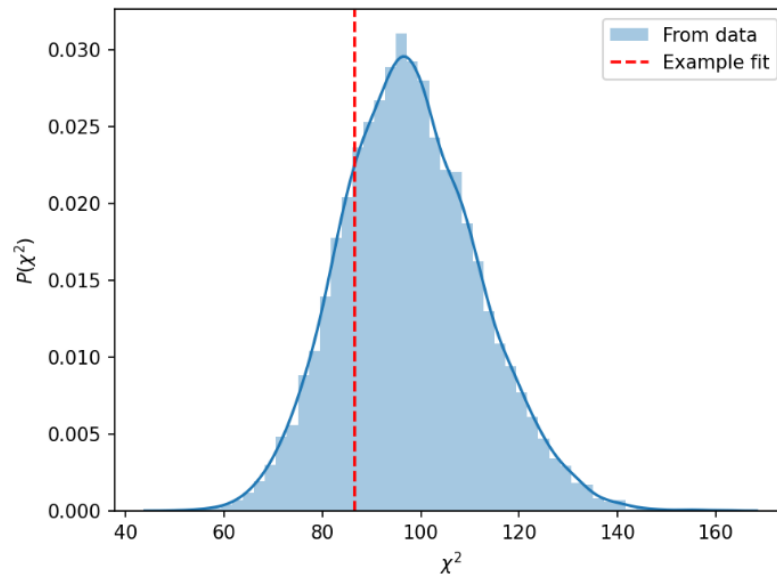


Figure 10.2

This is called a “ χ^2 distribution” and is generated by the frequencies of χ^2 values across many realisations of the data, if the model fitted to the data is the same model actually used to generate the data. A chi squared distribution can be described using the number of “degrees of freedom”, k , and we denote a chi squared distribution with k degrees as χ_k^2 . The PDF is given by:

$$P(x; k) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}$$

Where $\Gamma(x)$ is the gamma function, an interpolation of the factorial function across all reals. The number of degrees of freedom is given by the number of data points minus the number of fitted parameters, i.e. $k = N - M$. Comparing out calculated distribution to the theoretical distribution given by the PDF gives:

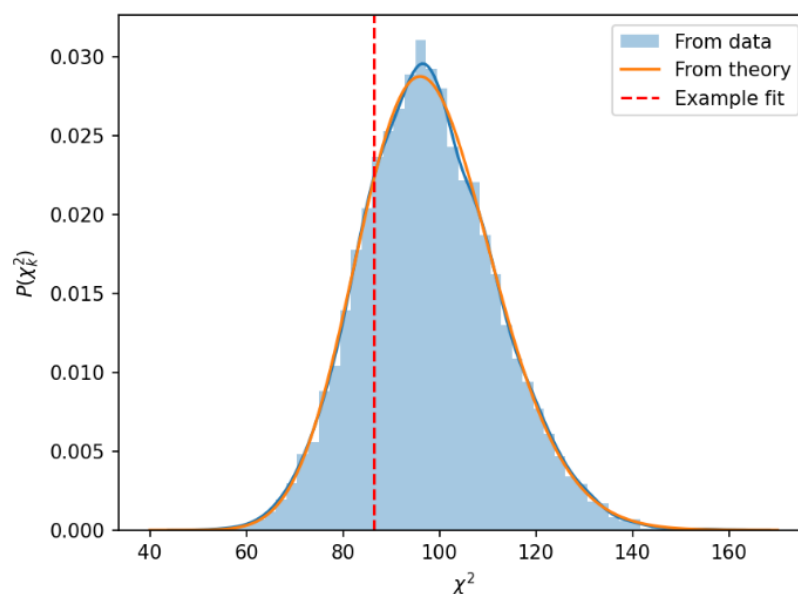


Figure 10.3

Which is a pretty good fit. We can also look at this distribution for different numbers of data points,

where a larger number of data points gives a larger number of degrees of freedom:

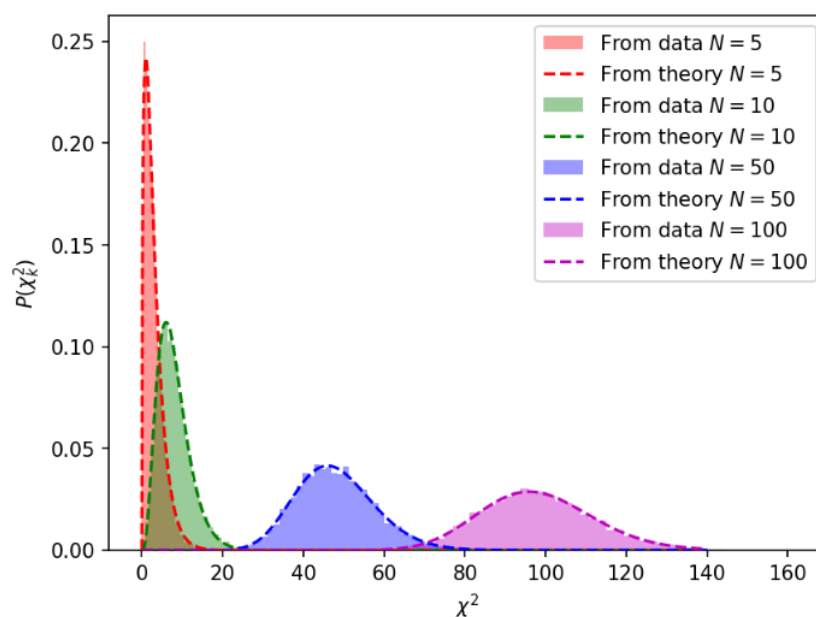


Figure 10.4

If we are fitting the correct model to our data, then we know the distribution of expected χ^2 values is (a χ^2_k distribution). This provides a goodness of fit statistic, by checking to see if the χ^2 value from the best fit parameters is consistent with what we'd expect to see if our model was correctly chosen.

If assume our model is correct, we can ask how likely it is that we'd get the data χ^2 value from the χ^2_k likelihood. If this is reasonably likely, we say the goodness of fit is acceptable. If not, then we have a problem. This problem could be many things, including:

- The incorrect model is being fitted - *larger χ^2 than expected.*
- The uncertainties on the data are too small to account for the observed noise - *larger χ^2 than expected.*
- The uncertainties on the data are too large so account for more than the actual noise - *smaller χ^2 than expected.*
- Something else has gone wrong...

The test itself cannot tell us exactly which of these is true, we have to use scientific judgement.

2 Quantifying the Likelihood

Since the χ^2_k distribution is continuous, the likelihood of getting a specific χ^2 value is zero. Instead, we reframe and look at “what is the chance of getting this value of χ^2 or larger?”. This is done by:

$$P(\chi^2 \geq a) = \int_a^\infty P(\chi^2; k) d\chi^2$$

Crucially, **the probability returned is only valid on the assumption that the model fitted to the data is correct.** If this value is above a certain threshold, we say this is evidence of a sensible fit. This is **not the chance that the model is correct, as the model is just that, a model, and is almost certainly never entirely correct.** It is the chance of getting this χ^2 value or larger *if* the model is correct.

Taking the previous example of $\chi^2 = 86.37$, with $k = 100 - 2 = 98$. This gives:

$$P(\chi^2 \geq 86.37) = \int_{86.37}^\infty P(\chi^2; 98) d\chi^2$$

This needs to be evaluated numerically, as it becomes unpleasant for not-trivial values of k . This gives a probability of 0.793. This is a reasonable value, not worryingly high (i.e. better than 99%) or worryingly low (i.e. less than 1%), so we say the fit is adequate.

To counterexample, say we have a poor fit. The data is from $y = 0.85x^2 + 2x - 1$, and we try to fit a linear model to it:

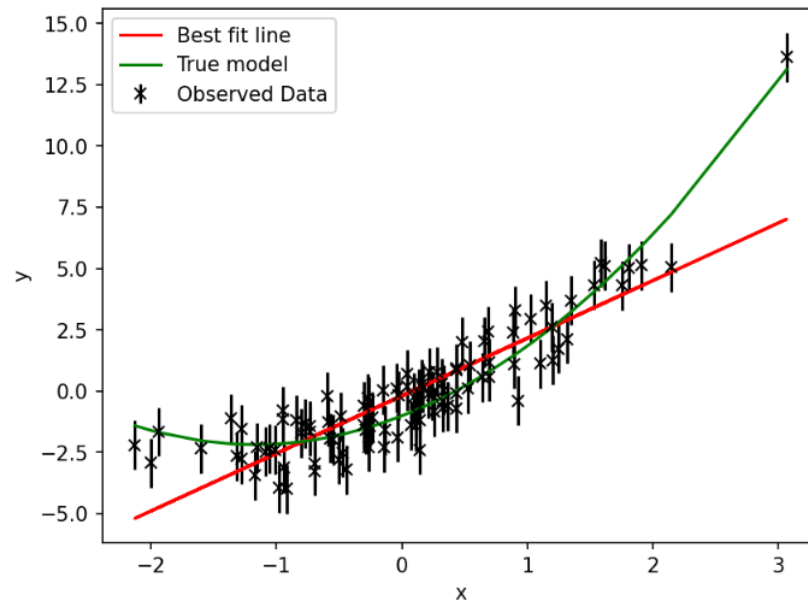
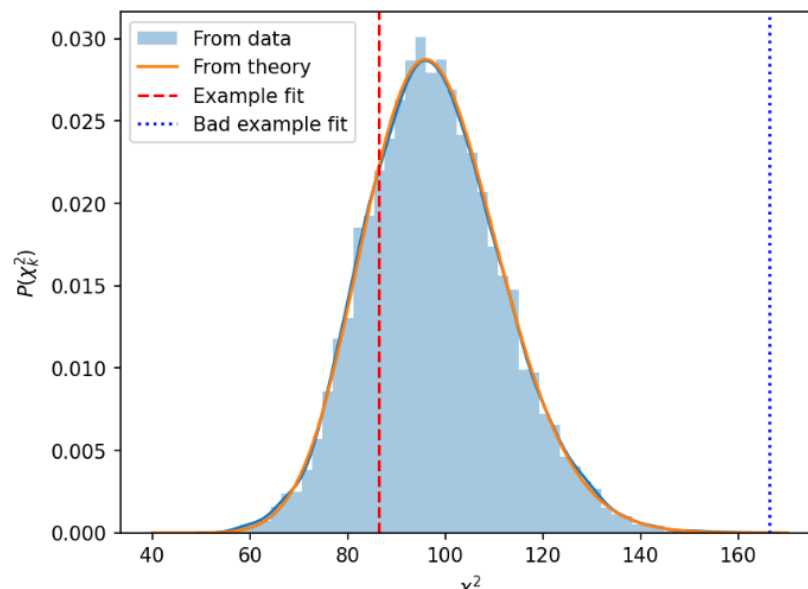


Figure 10.5

This gives a χ^2 value of 166.407. This is much larger than before, let's plot it on the distribution of the expected χ^2 :

Figure 10.6: Probability is $1.98e - 05$

So we can confidently say that the fit is not an acceptable quality fit. This is despite it not being a terrible fit by eye.

Wed 05 Nov 2025 12:00

Lecture 11 - End of Stats: Revision

Revision lecture. No new content.

Fri 07 Nov 2025 11:00

Lecture 1 - Start of Probability: Introduction

What is probability? Probability is the pure mathematical description of randomness.

1 Set Theory

Say we want to group trees into four sets:

- Tall, or not.
- Variegated (has a lighter coloured leaf border) or not.

In a park of 142 trees, we observe:

Fri 07 Nov 2025 11:00

Lecture 2 - Combinatorics

Thu 13 Nov 2025 09:00

Lecture 3 - Combining Probabilities

Today we will arrive at:

- The formula for $P(A \cap B)$ (Probability of A and B).
- Summing mutually exclusive events.

1 More Set Theory

We have a sample space Ω , and a subset labelled A . We then have the remainder of Ω (the portion of Ω which is not in A), denoted "A Complement" - A^C or \bar{A} . We also have a subset labelled B .

We can define A using set builder notation, to slightly redundantly say "A is the set of all x's which are in A":

$$A = \{x \mid x \in A\}$$

There is some overlap between A and B. We denote this intersection as $A \cap B$.

$$A \cap B = \{x \mid x \in A \text{ and } x \in B\}$$

Everything written in A or B (including the intersection) is called the union, $A \cup B$:

$$A \cup B = \{x \mid x \in A \text{ or } x \in B\}$$

Note that "or" in standard language excludes both, i.e. you may have x or you may have y. In mathematics, we refer to this as XOR (exclusive or). "Or" by itself does allow for this case of both, so an item in A or B may be in A alone, B alone, or both (i.e. in the intersection).

We also have the empty set $\emptyset = \{\}$. If two sets have no common elements, the intersection is this empty set. We say that the events are mutually exclusive (they cannot both happen) and the sets are pairwise disjoint. The empty set is the complement of Ω , $\emptyset = \Omega^C$.

2 De Morgan's Laws

De Morgan's Laws give us these relations:

$$(A \cup B)^C = A^C \cap B^C \quad (14.1)$$

$$(A \cap B)^C = A^C \cup B^C \quad (14.2)$$

This can be illustrated visually as follows:

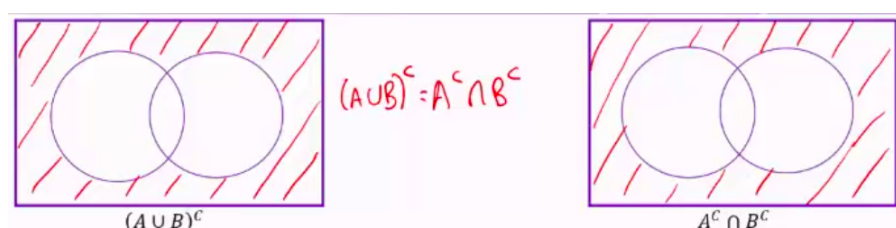


Figure 14.1

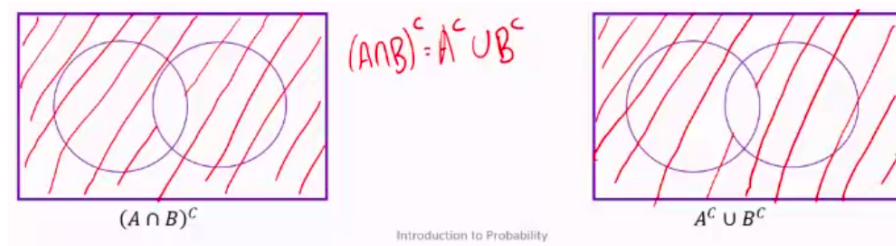


Figure 14.2

3 The Inclusion-Exclusion Principle

The number of elements in A and B is given by:

$$|A \cup B| = |A| + |B| - |A \cap B|$$

The last term is required to account for the intersection of A and B being included in A, and included in B. Therefore it double-counts the intersection, and we subtract it away.

The same is true of probability functions:

$$P(A) + P(B) = P(A \cup B) + P(A \cap B)$$

In other words, the probability of A + the probability of B is the probability of A or B + the probability of A + B.

This has the following consequences:

$$P(\emptyset) = 0$$

As:

$$P(A) = \frac{|A|}{|\Omega|} \implies P(\emptyset) = \frac{0}{|\Omega|} = 0$$

And:

$$P(A) = p \implies P(A^C) = 1 - p$$

As:

$$\Omega = A \cup A^C$$

$$P(\Omega) = P(A) + P(A^C) - P(A \cup A^C) \quad \text{By inclusion-exclusion principle.}$$

$$P(A \cup A^C) = P(\Omega) = 1$$

$$1 = P(A) + P(A^C) \quad \text{As: } P(\Omega) = 1$$

4 Multiple Events

Given some events e_n , the inclusion-exclusion principle says:

$$P(e_1 \cup e_2) = P(e_1) + P(e_2) - P(e_1 \cap e_2)$$

Hence for independent events ($e_1 \cap e_2 = \emptyset$), the probability of e_1 or e_2 occurring is:

$$P(e_1 \cup e_2) = P(e_1) + P(e_2)$$

4.1 What about 3 events?

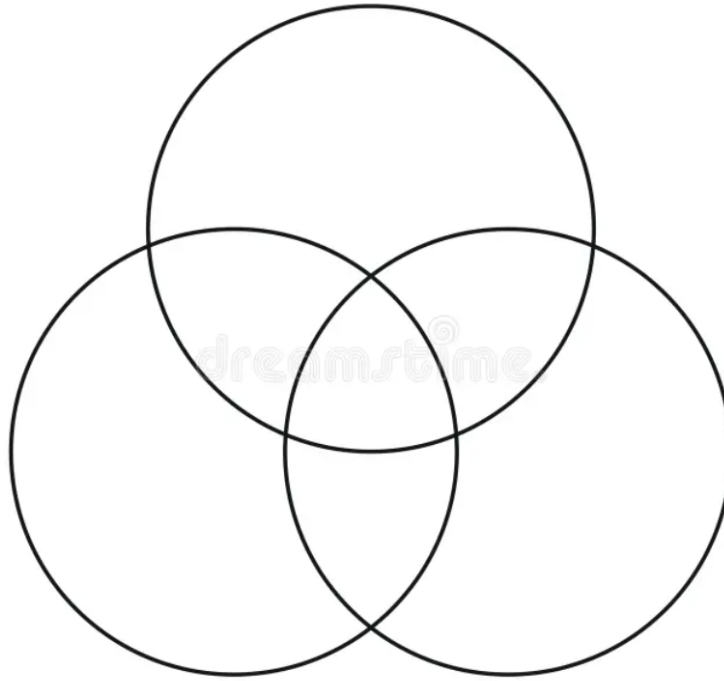


Figure 14.3

For three events, our venn diagram becomes more complex. We want to calculate $|e_1 \cup e_2 \cup e_3|$ (in this general example, which circle is which event is irrelevant).

This is given by:

$$|e_1 \cup e_2 \cup e_3| = |e_1| + |e_2| + |e_3| - |e_1 \cap e_2| - |e_1 \cap e_3| - |e_2 \cap e_3| + |e_1 \cap e_2 \cap e_3|$$

Note the final term, as the central portion is included three times when summing the whole event, but subtracted three times when removing intersections, so we must add it back.

If they are all pairwise disjoint, so $e_i \cup e_j = \emptyset$, then:

$$P(e_1 \cup e_2 \cup \dots \cup e_n) = P(e_1) + P(e_2) + \dots + P(e_n)$$

Or:

$$P\left(\bigcup_{n=1}^N e_n\right) = \sum_{n=1}^N P(e_n)$$

4.2 Normalisation

If the events are all mutually exclusive, and the sample space is “tiled” by the events (i.e. one of them must happen), then:

$$P(\Omega) = P(e_1) + P(e_2) + \dots + P(e_n) = 1$$

Thu 20 Nov 2025 09:00

Lecture 4 - Conditional Probability

1 Axioms of Probability

For a sample space Ω , a distribution $P(x)$ must satisfy:

1. $P(x) \geq 0$ for any $x \in \Omega$ (for discrete events)
2. $P(\Omega) = 1$
3. $P(e_1 \cup e_2 \cup \dots \cup e_n) = P(e_1) + P(e_2) + \dots + P(e_n)$
 - if the elements are pairwise disjoint ($e_i \cap e_j = \emptyset, i \neq j$).

2 Conditional Probability

2.1 Example

Throw two dice. What is the probability that we see a 4, given the total was 6?

This “given that” is the key. It provides us with an extra piece of information that the final probability depends on.

	(1,1)	(2,1)	(3,1)	(4,1)	<u>(5,1)</u>	(6,1)
	(1,2)	(2,2)	(3,2)	<u>(4,2)</u>	(5,2)	(6,2)
	(1,3)	(2,3)	<u>(3,3)</u>	(4,3)	(5,3)	(6,3)
$\Omega =$	(1,4)	<u>(2,4)</u>	(3,4)	(4,4)	(5,4)	(6,4)
	<u>(1,5)</u>	(2,5)	(3,5)	(4,5)	(5,5)	(6,5)
	(1,6)	(2,6)	(3,6)	(4,6)	(5,6)	(6,6)

Figure 15.1: The sample space.

We can now see that our probability is $\frac{2}{5}$. This is different than if we had not considered the extra information, and would have drastically changed our answer.

2.2 Definition

The conditional probability of A given B is written $P(A | B)$ and is defined by:

$$P(A | B) \equiv \frac{P(A \cap B)}{P(B)} = \frac{\text{number of events in A and B}}{\text{number of events in B}} \quad P(B) \neq 0$$

This is the fraction of events in B where both A and B happen.

2.3 Verifying This is Still a Probability

It may not be obvious that this is still a valid probability (i.e. that taking ratios of probabilities still yields a probability).

Proof. Assume that P is a valid probability function and let $Q(A | B) \equiv \frac{P(A \cap B)}{P(B)}$

Consider the first axiom of probability:

$Q(A | B) \geq 0$. This is satisfied, as $P(x) \geq 0$ for any $x \in \Omega$.

Consider the second:

$$Q(\Omega | B) = \frac{P(\Omega \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1 \text{ so satisfied.}$$

Consider the third:

If $a_1 \cap a_2 = \emptyset$, is $Q(a_1 \cup a_2 | B) = Q(a_1 | B) + Q(a_2 | B)$?

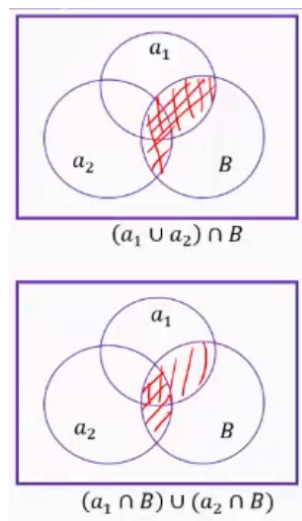


Figure 15.2

The definition of Q gives:

$$\begin{aligned} Q(a_1 \cup a_2 | B) &= \frac{P([a_1 \cup a_2] \cap B)}{P(B)} \\ &= \frac{P([a_1 \cap B] \cup [a_2 \cap B])}{P(B)} \\ &= \frac{P(a_1 \cap B) + P(a_2 \cap B)}{P(B)} \\ &= \frac{P(a_1 \cap B)}{P(B)} + \frac{P(a_2 \cap B)}{P(B)} \\ &= Q(a_1 | B) + Q(a_2 | B) \end{aligned}$$

So yes, the third axiom is satisfied.

Since all three axioms are satisfied, Q , hence P , is a valid probability distribution. \square

3 Reconditioning

Reconditioning is using $P(A | B)$ to determine $P(B | A)$.