LC Introduction to Probability and Statistics Lecture Notes

MSci Physics w/ Particle Physics and Cosmology
University of Birmingham

Year 1, Semester 1
Ash Stewart

Lectures Index

Wed 01 Oct 2025 12:00

Lecture 1 - Introdution 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 Propogation 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:

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

$$Var(e^x) \approx (e^x)^2 Var(x)$$

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.

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

$$Var(mx + c) = m^2 Var(x)$$

2.2 Multiple Variables

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

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

$$Var(A + B) = Var(A) + Var(B)$$

And multiplying them (again this are now approximations)?

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

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

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

Or division?

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

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

3 One Rule to Rule Them All

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

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

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 propogation/combinations etc, we assumed that they were independant 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 Cov(x, y) = Cov(y, x) and Cov(x, y) = Var(x))

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

We can then define correlation:

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

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

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

1.1 Variable Combinations

Now, with correlated variables, we can say:

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

$$Var(x,y) = Var(x) + Var(y) + 2Cov(x,y)$$

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

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

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

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

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:

$$P(H) = 0.5$$
$$P(T) = 0.5$$

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:

$$P(A_h \text{and} B_h) = P(A_h) \times P(B_h) = P(HH)$$

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:

$$\langle r \rangle = np$$
$$Var(r) = np(1-p)$$

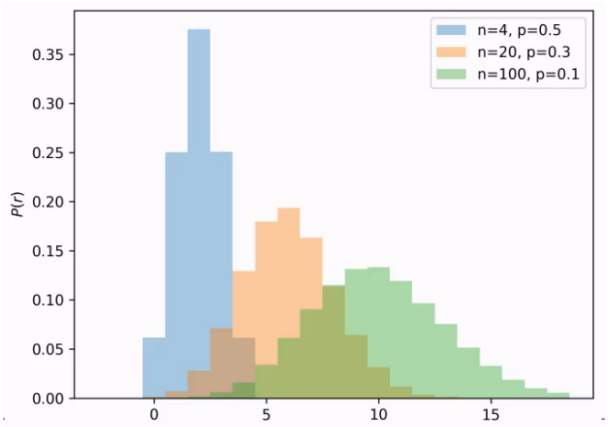

Figure 5.1: The Binomial Distribution

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 ($n \to \infty$)
- 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:

$$\langle r \rangle = \lambda$$
$$Var(r) = \lambda$$

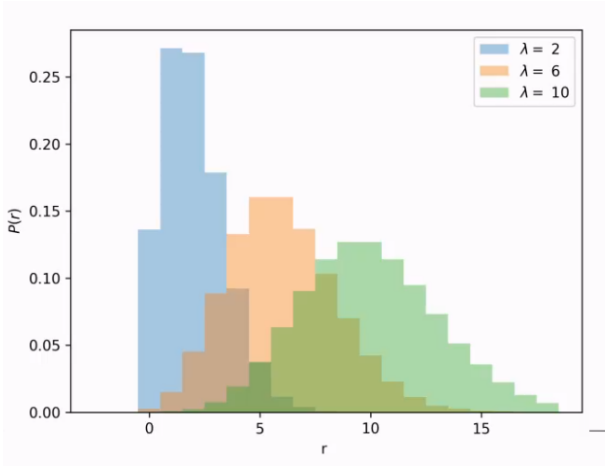

Figure 5.2: The Poisson Distribution

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":

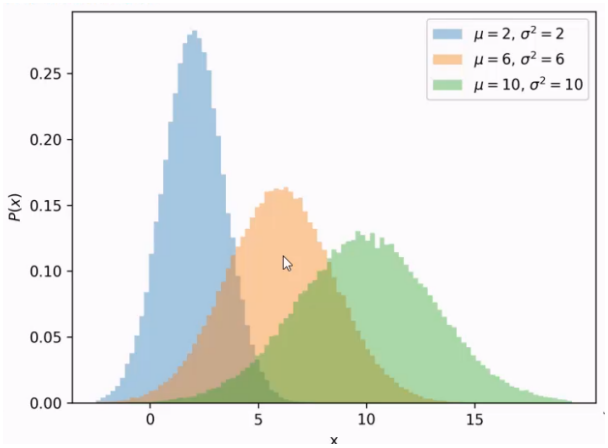

Figure 5.3: The Normal Distribution

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

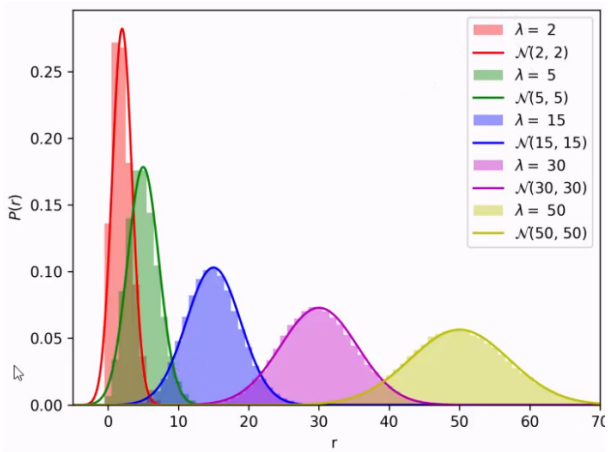

Figure 5.4: Poisson approximations to a Normal

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:

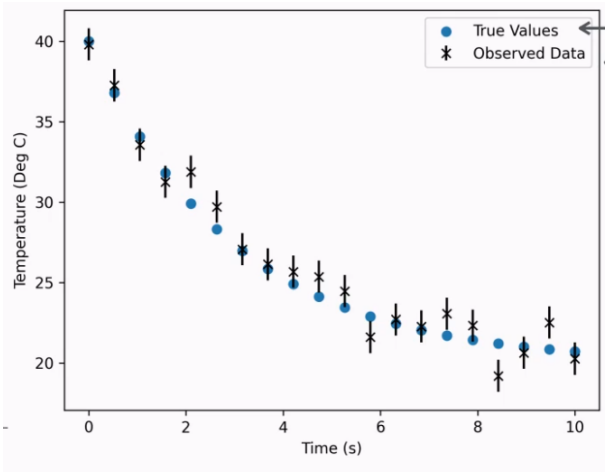

Figure 6.1: Simulated Data

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 and 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 ith 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:

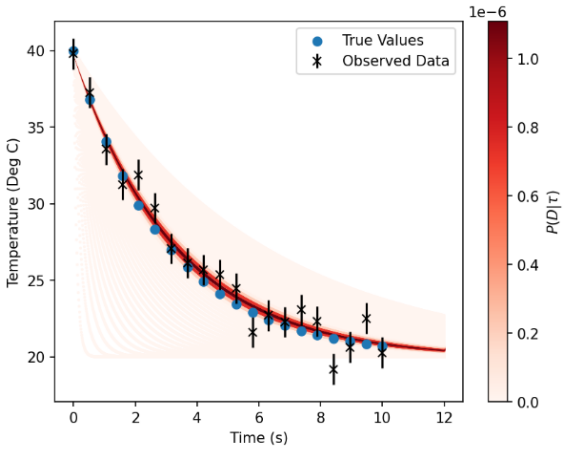

Figure 6.2: A heatmap of the likelihood function overlaid on the data.

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:

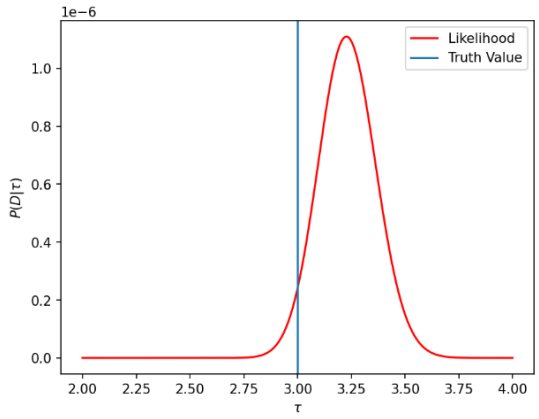

Figure 6.3

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:

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

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

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

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

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:

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

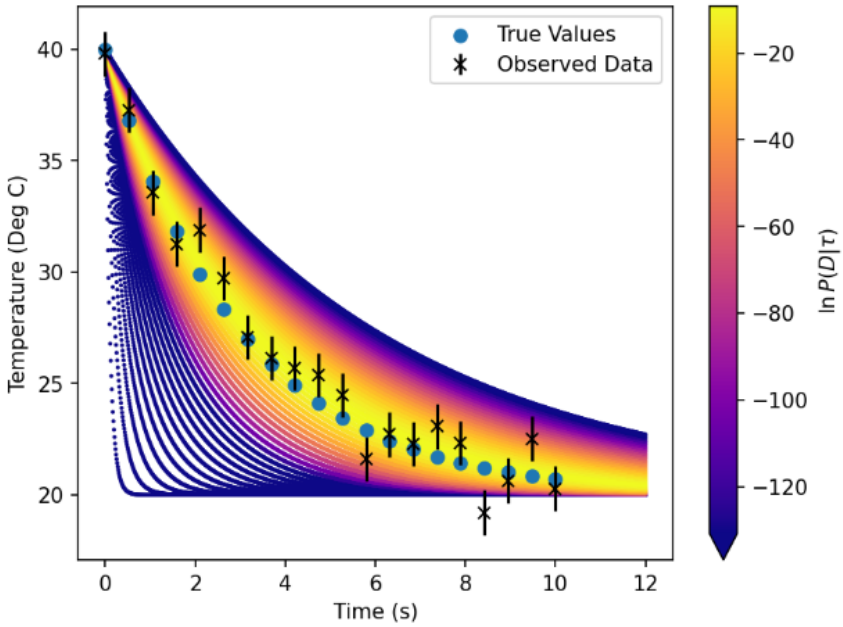

Figure 6.4: The likelihood plot, repeated with log likelihood

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):

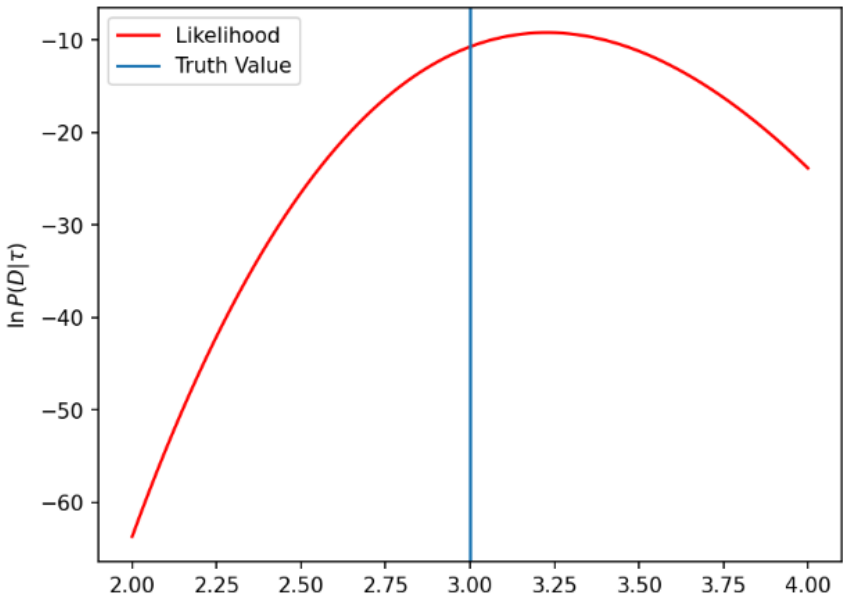

Figure 6.5

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.

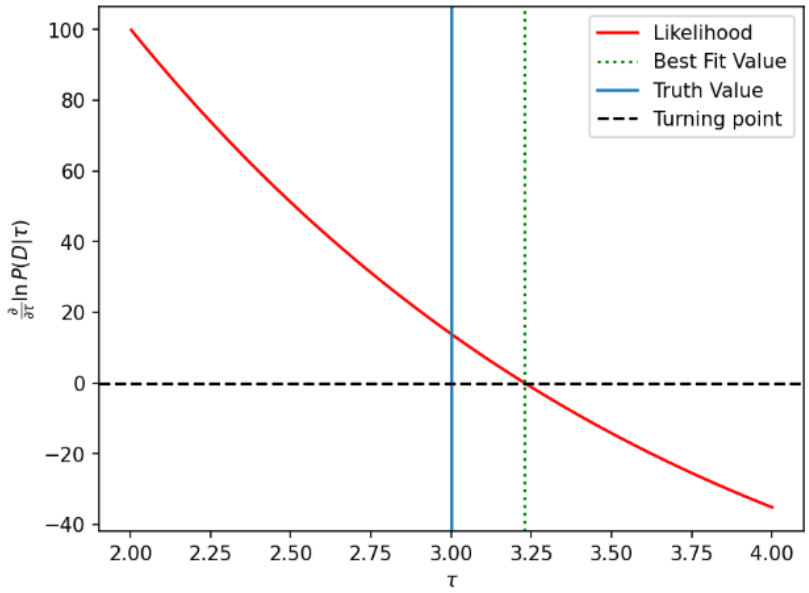

Figure 6.6

4 Uncertainties on Log Likelihood

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

$$\boxed{H = \frac{\partial^2 \mathcal{L}}{\partial \tau^2}\bigg|_{\tau=\hat{\tau}}}$$

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:

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

And calculating:

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

But note that this works differently in higher dimensions (more parameters) has 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.

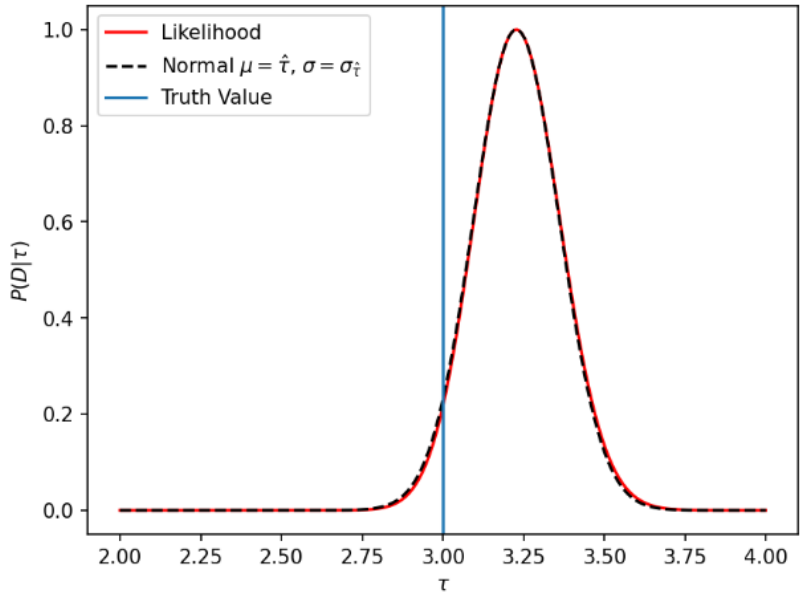

Figure 6.7

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:

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

And the Hessian is given by:

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

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 are 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.

Fri 07 Nov 2025 11:00

Lecture 10 - Introduction to Probability

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: