LC Introduction to Probability and Statistics Lecture Notes

Year 1 Semester 1
MSci Physics with Particle Physics and Cosmology

Ash Stewart
University of Birmingham

Lectures Index

Wed 01 Oct 2025 12:00

Lecture S1 - 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 S2 - 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 S3 - 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 S4 - 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 S5 - 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(Three Heads and One Tail) $= P$(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:

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

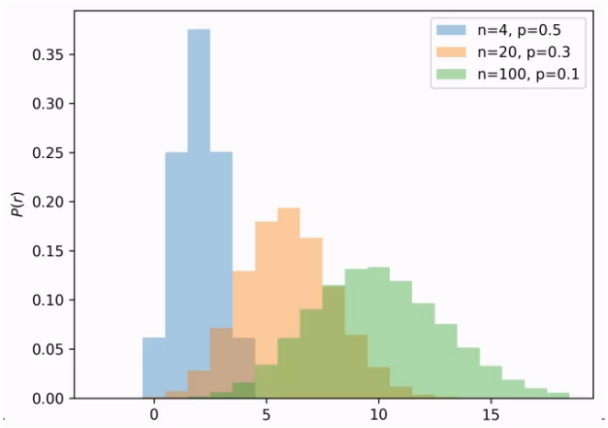

Figure 5.1: The Binomial Distribution

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 ($np = \lambda =$ 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:

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

$$Var(r) = \lambda$$

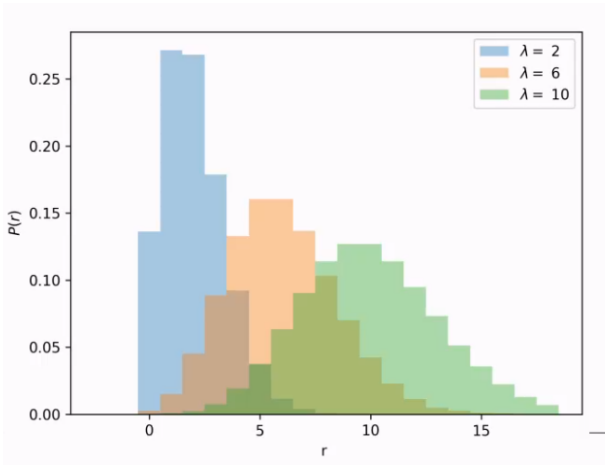

Figure 5.2: The Poisson Distribution

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":

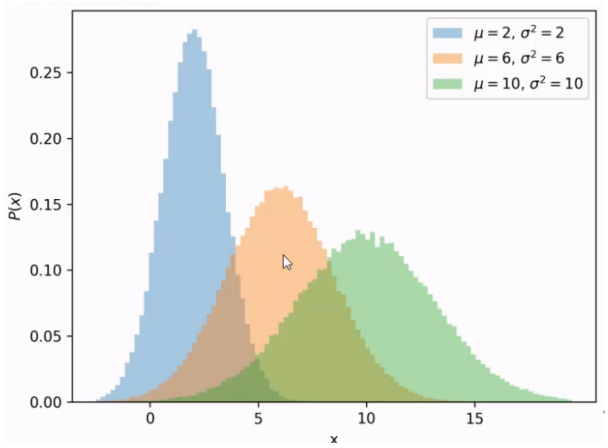

Figure 5.3: The Normal Distribution

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.

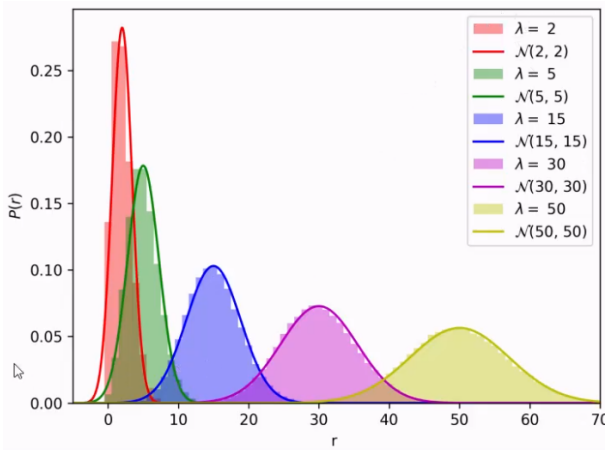

Figure 5.4: Poisson approximations to a Normal

Thu 16 Oct 2025 09:00

Lecture S6 - 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:

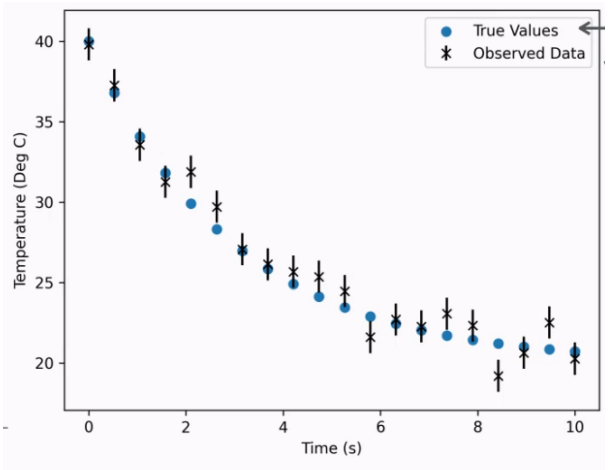

Figure 6.1: Simulated Data

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):

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

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:

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

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:

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

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

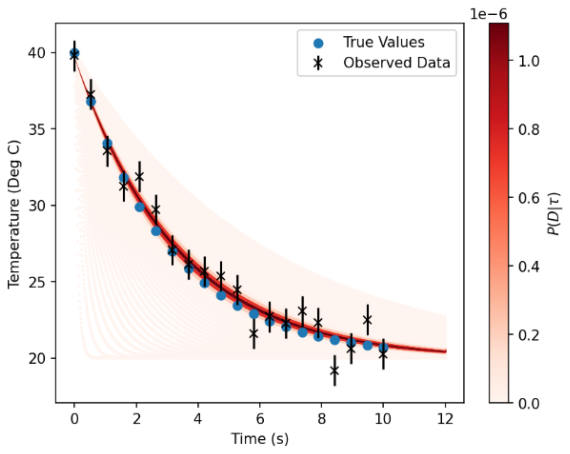

Figure 6.2: A heatmap of the likelihood function overlaid on the data.

The high likelihood values of $P(D \mid \tau)$ 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:

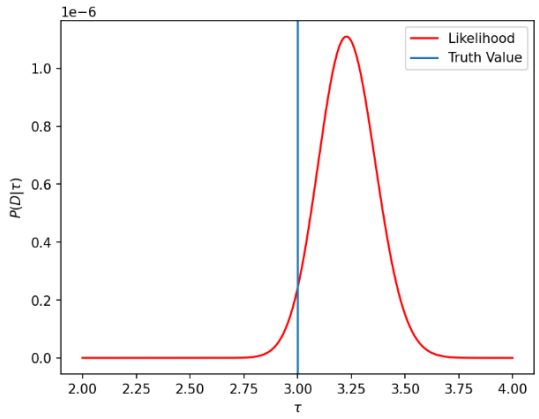

Figure 6.3

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:

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

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

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

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:

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

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

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

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

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):

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

This gives the final log likelihood as:

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

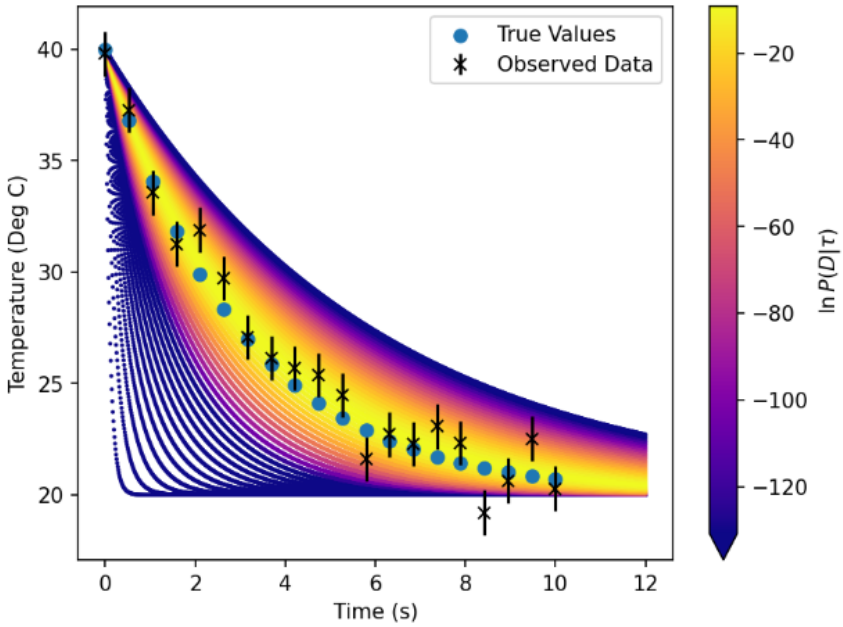

Figure 6.4: The likelihood plot, repeated with log likelihood

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):

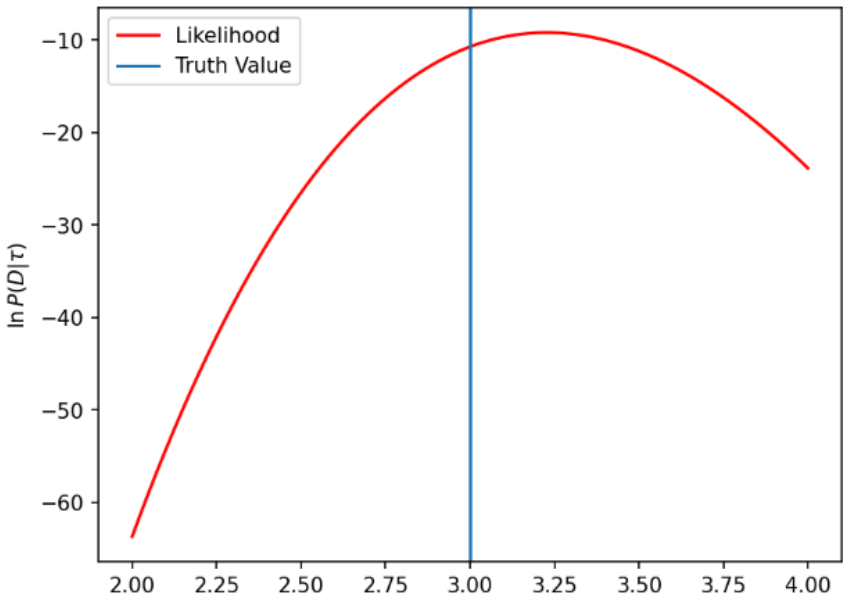

Figure 6.5

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:

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

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.

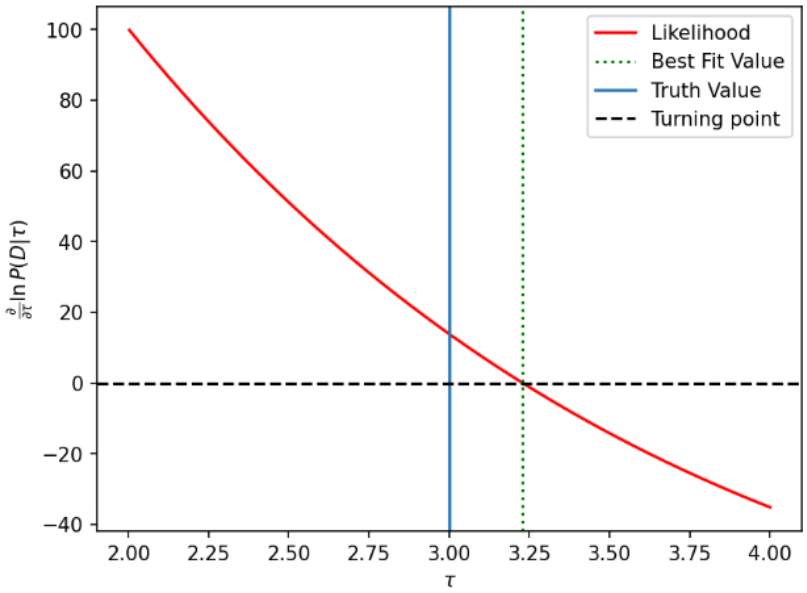

Figure 6.6

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.

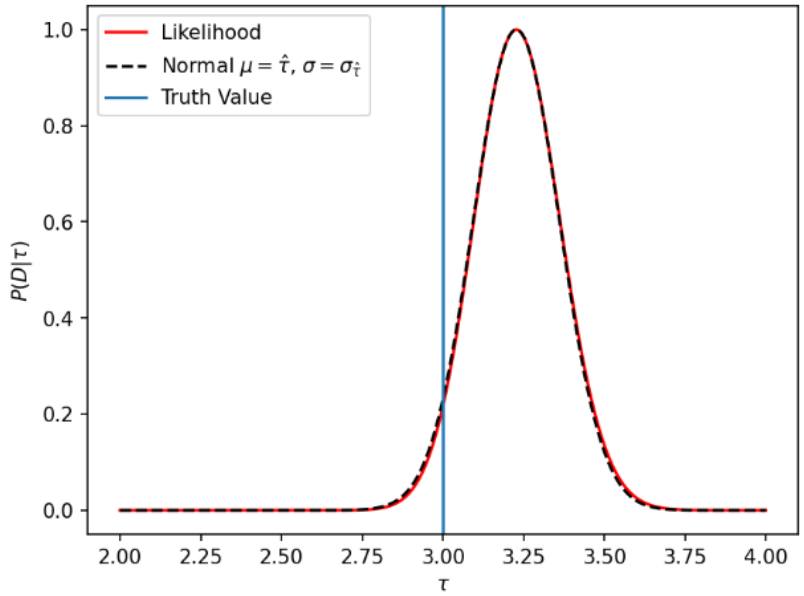

Figure 6.7

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 $P(D \mid \theta)$ 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 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 S7 - 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:

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

And again:

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

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

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

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):

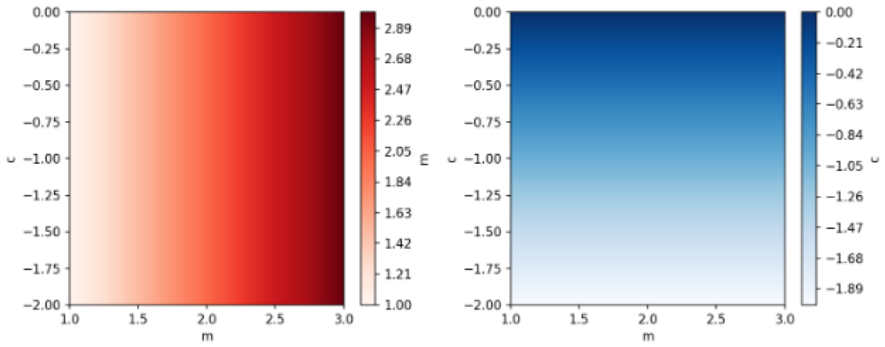

Figure 7.1

And the likelihood:

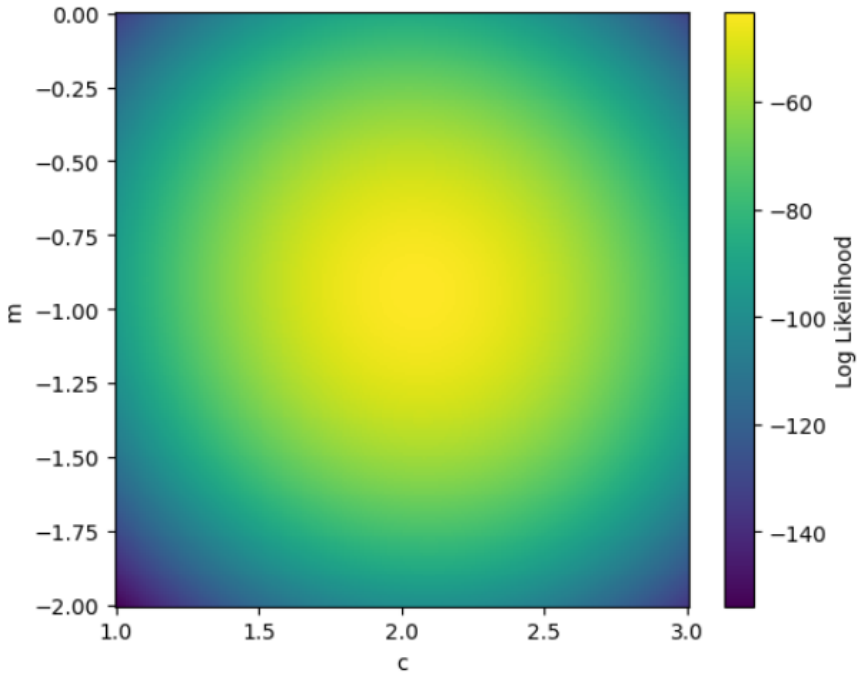

Figure 7.2

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:

$$m = 2.0552763819095476$$
$$c = -0.9447236180904524$$

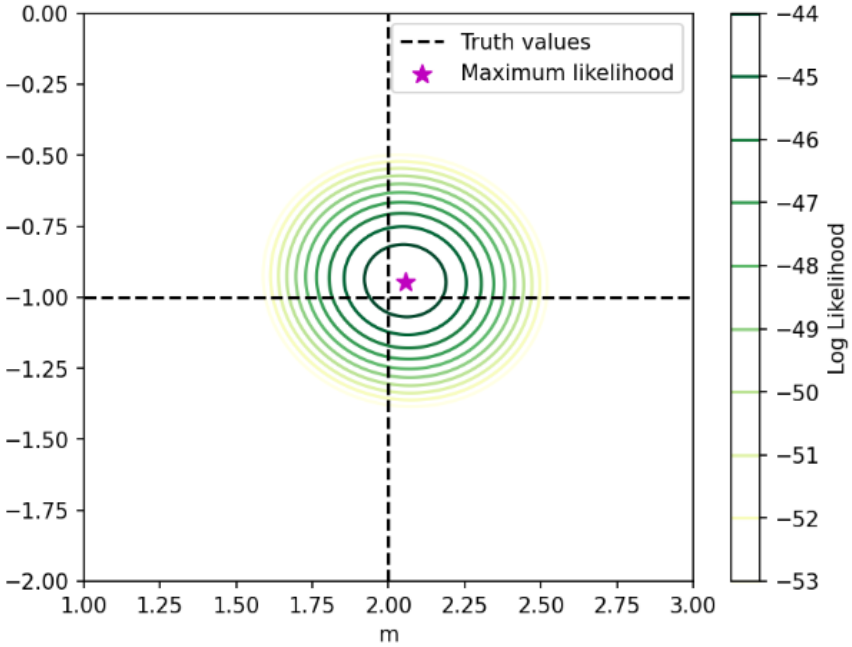

Figure 7.3

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 S8 - 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.

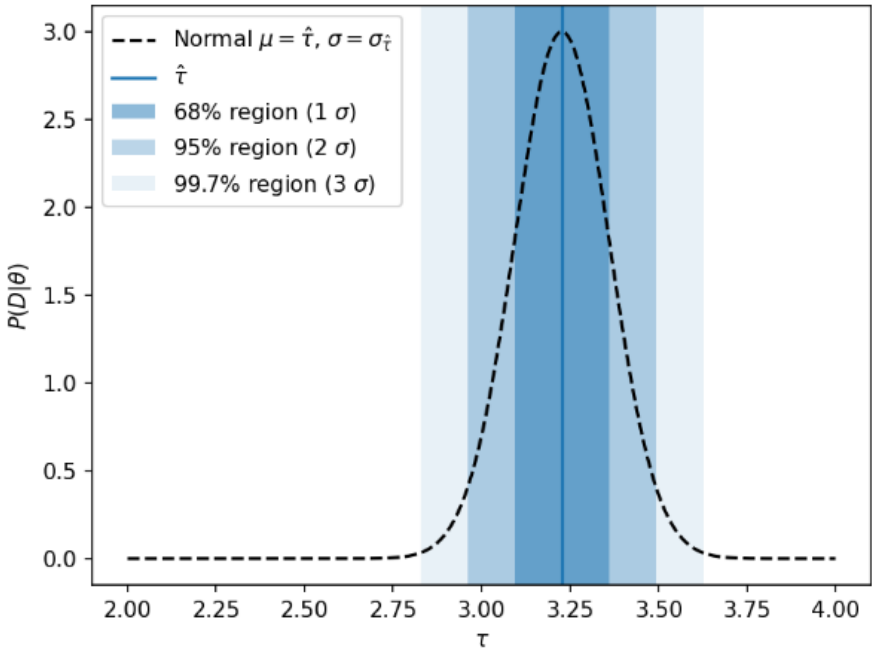

Figure 8.1

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:

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

And the same for 2 sigma uncertainty with 0.95, and 2 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.

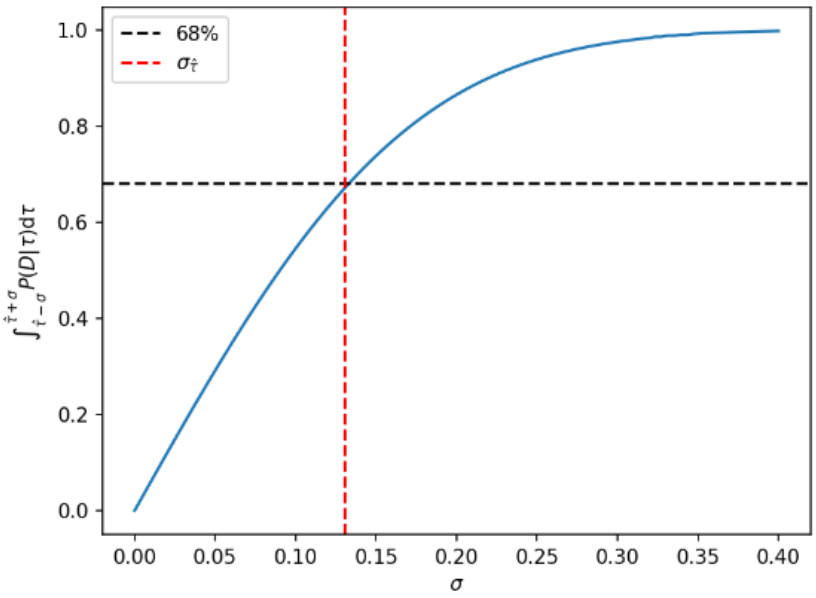

Figure 8.2

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:

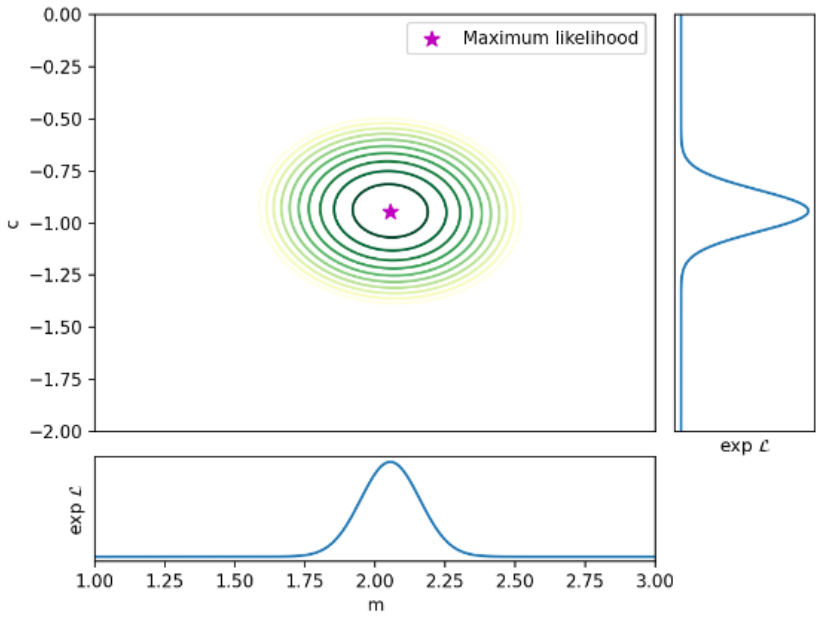

Figure 8.3

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:

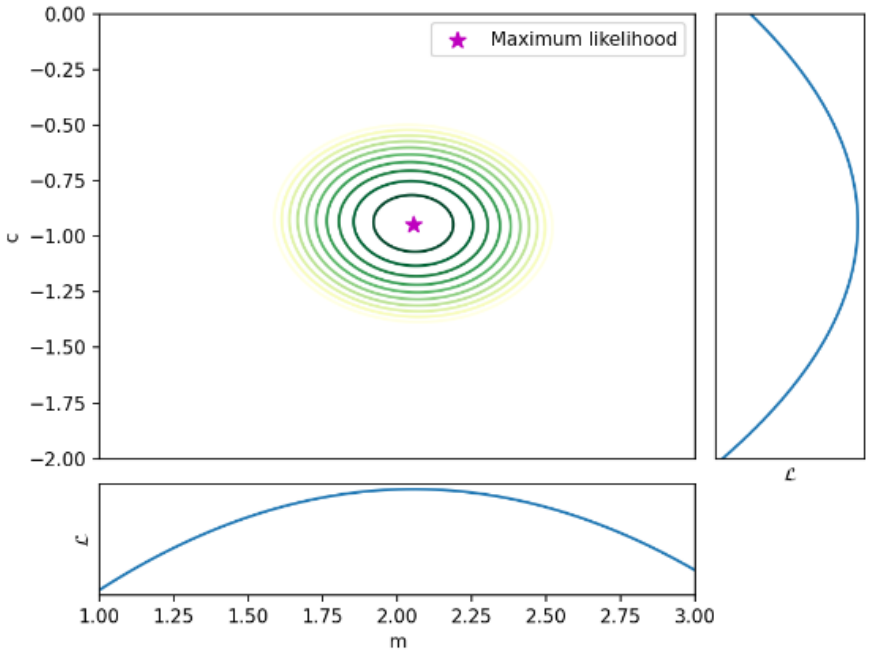

Figure 8.4

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

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

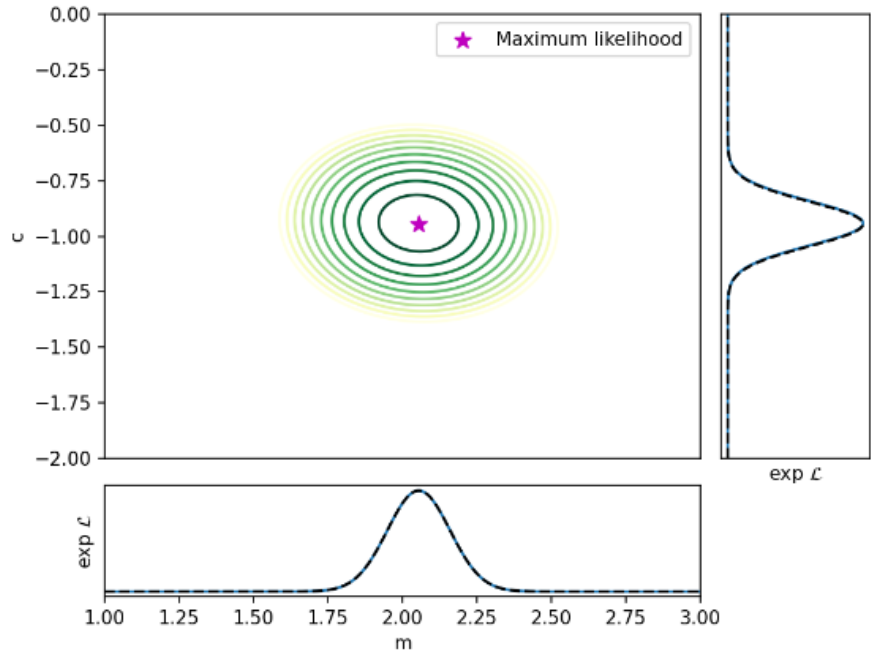

Figure 8.5

Calculating our summary stats and plotting them:

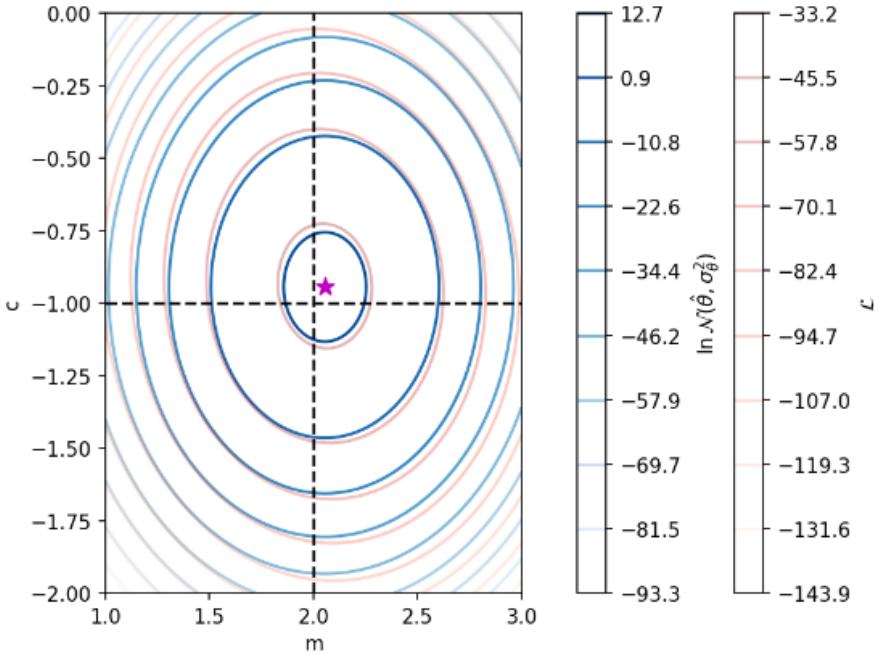

Figure 8.6

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 S9 - 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 has 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:

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

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

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

$$= \sum_{i=1}^{n}\frac{\partial u_i}{\partial m}\frac{\partial}{\partial u_i}u_i^2$$

$$= \sum_{i=1}^{n}\frac{\partial u_i}{\partial m}(2u_i)$$

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

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

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:

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

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

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

$$= \sum_{i=1}^{n}\frac{\partial u_i}{\partial c}\frac{\partial}{\partial u_i}u_i^2$$

$$= \sum_{i=1}^{n}\frac{\partial u_i}{\partial c}(2u_i)$$

$$= \sum_{i=1}^{n}\left(\frac{-1}{\sigma_{D_i}}\right)(2u_i)$$

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

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:

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

Rearranging (1) gives:

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

And (3) into (2):

$$S_{Dx} = mS_{xx} + \frac{S_d - mS_x}{S} S_x$$

$$S_{Dx} = mS_{xx} + \frac{S_D S_x}{S} - \frac{m(S_x)^2}{S}$$

$$S_{Dx} - \frac{S_D S_x}{S} = mS_{xx} - \frac{m(S_x)^2}{S}$$

$$S_{Dx} - \frac{S_D S_x}{S} = m\left(S_{xx} - \frac{(S_x)^2}{S}\right)$$

$$m = \left(S_{Dx} - \frac{S_D S_x}{S}\right) / \left(S_{xx} - \frac{(S_x)^2}{S}\right)$$

Simplifying to:

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

And for c, rearranging (2) gives:

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

(4) into (2)

$$S_{Dx} = \frac{S_D - cS}{S_x} S_{xx} + cS_x$$

$$S_{Dx} = \frac{S_D S_{xx}}{S_x} - \frac{cSS_{xx}}{S_x} + cS_x$$

$$S_{Dx} - \frac{S_D S_{xx}}{S_x} = cS_x - \frac{cSS_{xx}}{S_x}$$

$$S_{Dx} - \frac{S_D S_{xx}}{S_x} = c\left(S_x - \frac{SS_{xx}}{S_x}\right)$$

$$c = \left(S_{Dx} - \frac{S_D S_{xx}}{S_x}\right) / \left(S_x - \frac{SS_{xx}}{S_x}\right)$$

Simplifying to:

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

2.4 And Finally...

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

$$\hat{c} \equiv \langle c \rangle = \frac{S_D S_{xx} - S_x S_{Dx}}{\Delta}$$

$$\hat{m} \equiv \langle m \rangle = \frac{SS_{Dx} - S_D S_x}{\Delta}$$

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:

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

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

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 S10 - 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:

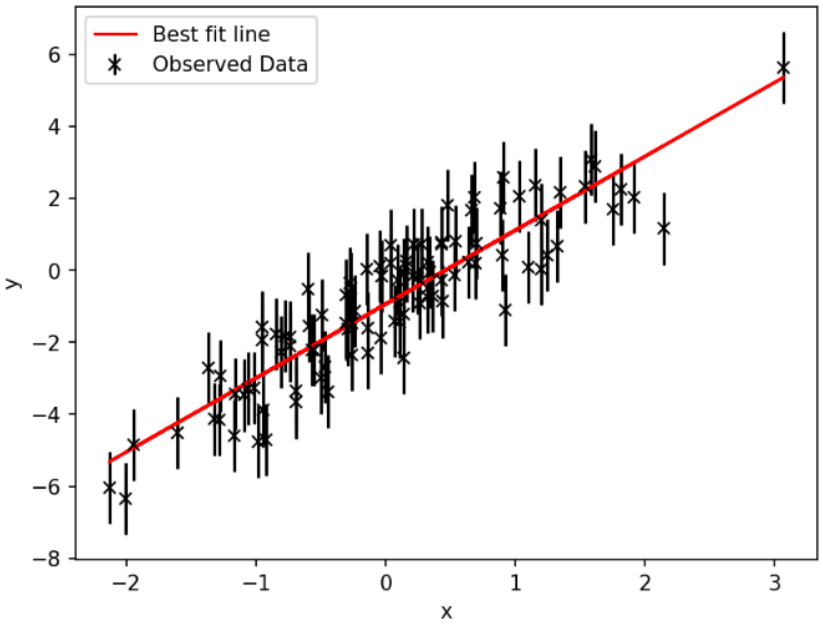

Figure 10.1

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:

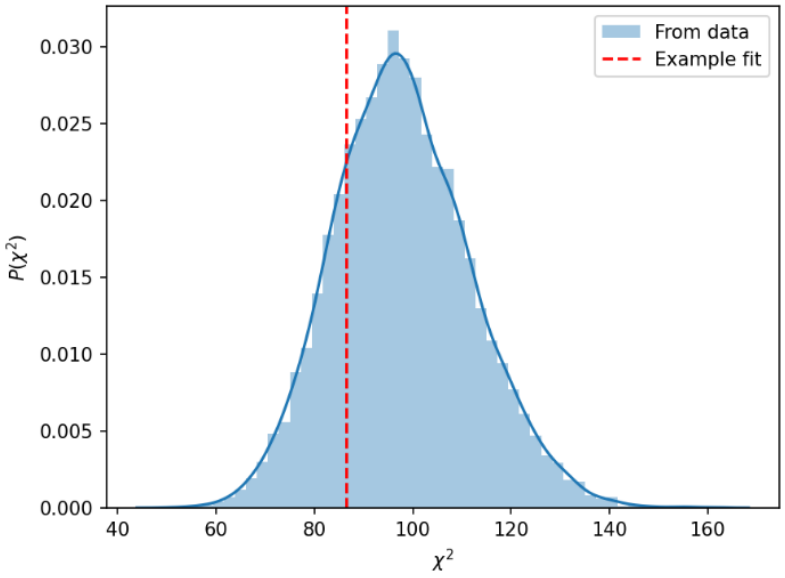

Figure 10.2

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^2_k. 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:

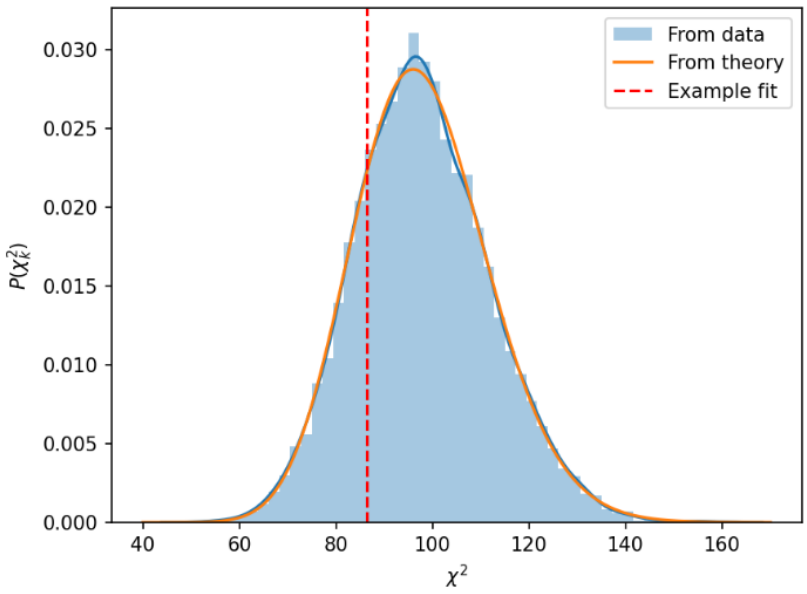

Figure 10.3

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:

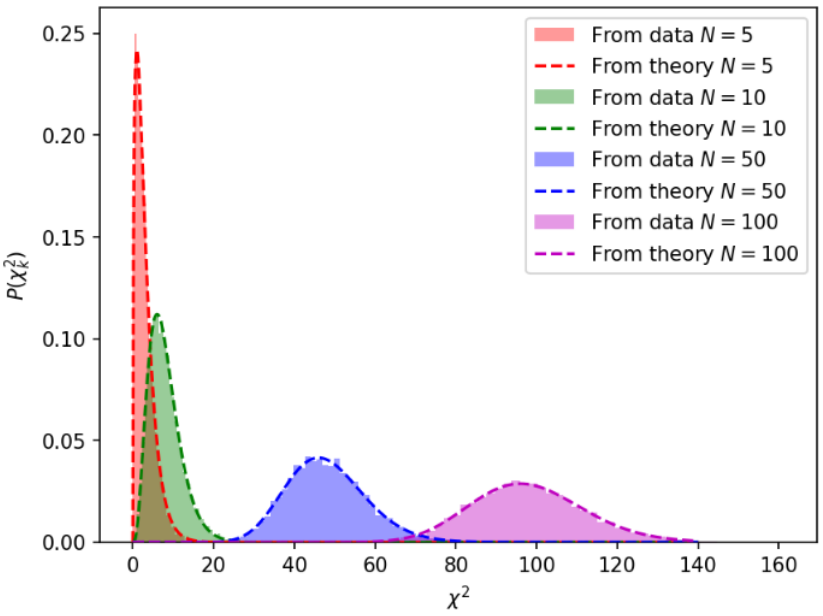

Figure 10.4

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:

$$P(\chi^2 \geq 86.37) = \int_{86.37}^\infty P(\chi^2; 86.36)\, d\chi^2$$

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:

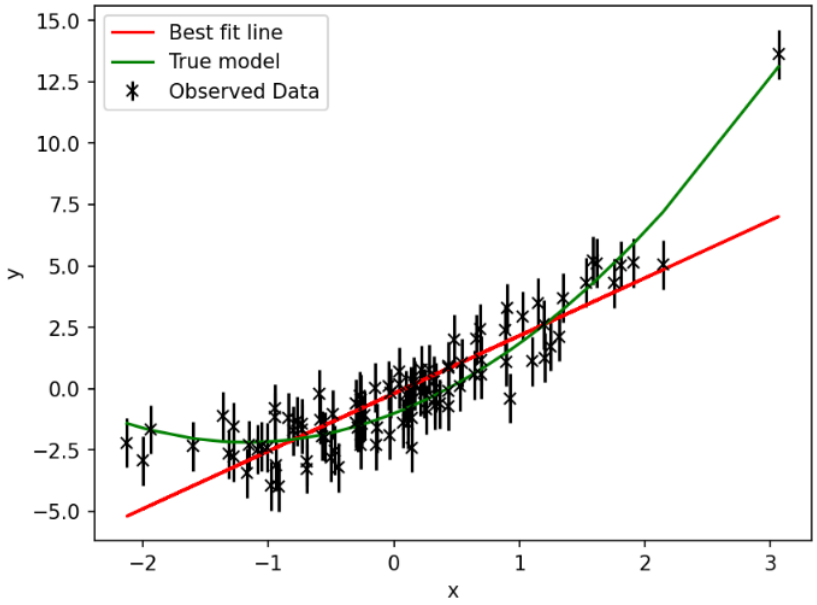

Figure 10.5

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

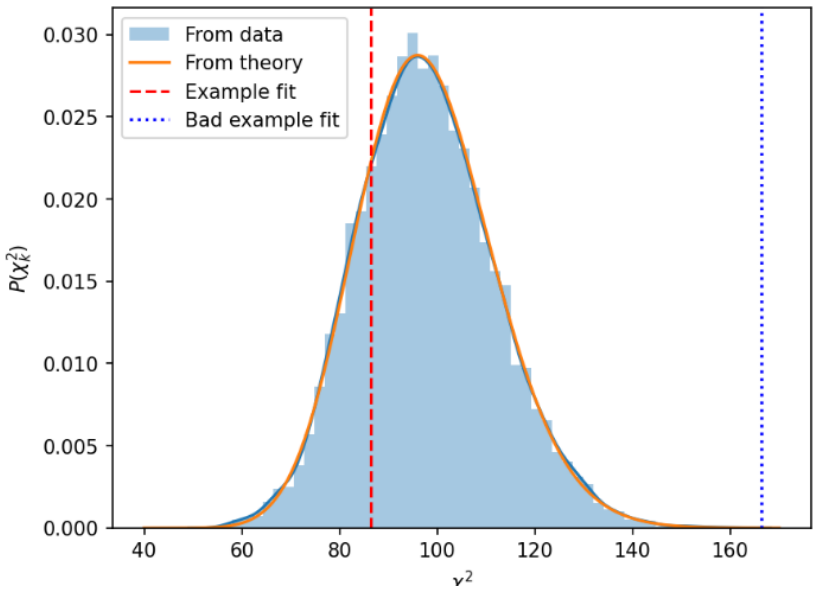

Figure 10.6: Probability is $1.98e-05$

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.

Thu 06 Nov 2025 09:00

Lecture S11 - End of Stats: Revision

Revision lecture. No new content.

Fri 07 Nov 2025 11:00

Lecture P1 - Start of Probability: Introduction

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

1 Empirical Probability

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:

	Tall	Not Tall	Total
Variegated	**20**	**41**	61
Not Variegated	**72**	**9**	81
Total	92	50	142

Figure 12.1

We denote Tall as T, Variegated as V and not Tall/not Variegated as $\bar{T}$ and $\bar{V}$. We use n as the number of number of trees that satisfy the parameters, i.e $n(T, \bar{V}) = 72$. $N = 142$ is the total number of trees, so:

$$N = n(T, V) + n(T, \bar{V}) + n(\bar{T}, \bar{V}) + n(\bar{T}, V)$$

Similarly, we define the fraction of trees that meet the provided criteria as f, i.e:

$$f(T, V) \equiv \frac{n(T, V)}{N}$$

Hence:

$$1 = f(T, V) + f(T, \bar{V}) + f(\bar{T}, \bar{V}) + f(\bar{T}, V)$$

We define probability as the limit of this as $N \to \infty$:

$$\boxed{P(T, V) = \lim_{N \to \infty} f(T, V) = \lim_{N \to \infty} \frac{n(T, V)}{N}}$$

And discrete probability as the limiting fraction of the times an event will occur. Probability is bounded between 0 (the event **never** occurs) and 1 (the event **always** occurs). In a particular experiment, the total probability is one, therefore something must happen (we just may not know what exactly).

2 Set Theory

A set is a collection of elements, i.e.:

$$A = \{1, 2, 3\}$$

A deck of cards is a set, with 52 elements. Elements can be anything, including other sets.

2.1 Subsets

Consider some set A. Another set, B is a "subset" of A if it contains (solely) some of the elements of A. We denote this $B \subset A$.

For example:

$$A = \{1, 2, 3, 4, 5, 6\}$$
$$B = \{1, 2, 3\}$$
$$C = \{4, 5, 6\}$$
$$D = \{1, 5, 7\}$$
$$B \subset A$$
$$C \subset A$$
$$D \not\subset A$$

2.2 Sample Space

We observe a particular outcome, and the set of all possible outcomes we could observe is called the sample space Ω. Any event (that can occur) is a subset of Ω.

If drawing 5 cards from a deck, all possible combinations of 5 cards form the sample space. Drawing an ace is an event.

Thu 13 Nov 2025 09:00

Lecture P2 - Combinatorics

Thu 14 Nov 2025 11:00

Lecture P3 - 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'es 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:

$$(A \cup B)^C = A^C \cap B^C \tag{14.1}$$

$$(A \cap B)^C = A^C \cup B^C \tag{14.2}$$

This can be illustrated visually as follows:

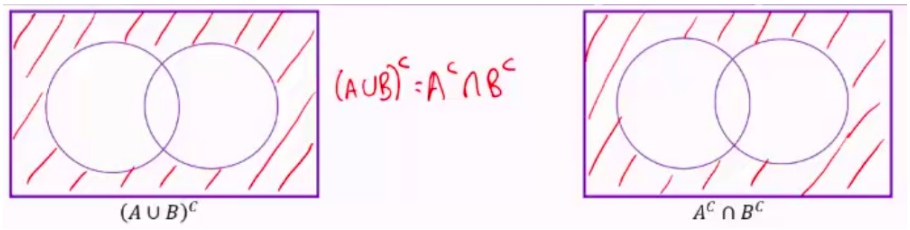

Figure 14.1

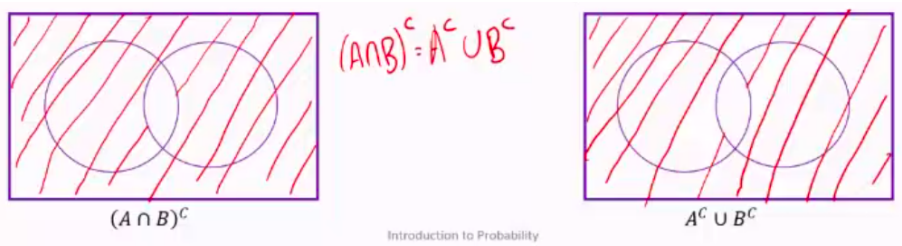

Figure 14.2

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:

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

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

$$P(A \cup A^C) = P(\emptyset) = 0$$

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

4 Multiple Events

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

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

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?

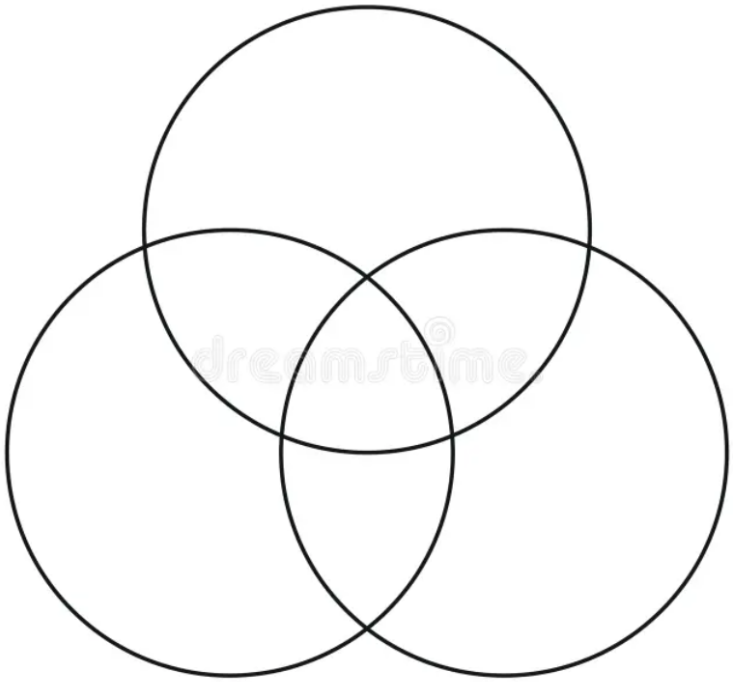

Figure 14.3

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:

$$P(e_1 \cup e_2 \cup \cdots \cup e_3) = P(e_1) + P(e_2) + \cdots + P(e_n)$$

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:

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

Thu 20 Nov 2025 09:00

Lecture P4 - 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. $P(e_1 \cup e_2 \cup + \cdots \cup e_n) = P(e_1) + P(e_2) + \cdots + P(e_n)$
   - 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.

$$\Omega = \begin{matrix} (1,1) & (2,1) & (3,1) & (4,1) & (5,1) & (6,1) \\ (1,2) & (2,2) & (3,2) & (4,2) & (5,2) & (6,2) \\ (1,3) & (2,3) & (3,3) & (4,3) & (5,3) & (6,3) \\ (1,4) & (2,4) & (3,4) & (4,4) & (5,4) & (6,4) \\ (1,5) & (2,5) & (3,5) & (4,5) & (5,5) & (6,5) \\ (1,6) & (2,6) & (3,6) & (4,6) & (5,6) & (6,6) \end{matrix}$$

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 $P(A \mid B)$ and is defined by:

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

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 $Q(A \mid B) \equiv \dfrac{P(A \cap B)}{P(B)}$

Consider the first axiom of probability:

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

Consider the second:

$Q(\Omega \mid B) = \dfrac{P(\Omega \cap B)}{P(B)} = \dfrac{P(B)}{P(B)} = 1$ so satisfied.

Consider the third:

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

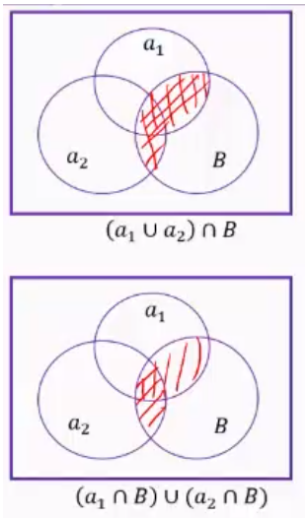

Figure 15.2

The definition of Q gives:

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

So yes, the third axiom is satisfied.

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

3 Reconditioning

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

$$(1) \quad P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

$$(2) \quad P(B \mid A) = \frac{P(B \cap A)}{P(A)}$$

Using the fact that $A \cap B = B \cap A$:

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

$$\implies P(A \cap B) = P(A \mid B)P(B) = P(B \mid A)P(A)$$

$$\boxed{\implies P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)}}$$

3.1 Example

A rare cold-like disease has symptoms with probability 0.95. The probability in the population of having the disease is 0.0001. The probability of having the cold-like symptoms in the population is 0.4, either from the disease or any other cold.

Given someone has symptoms, what is the probability they have the disease? Let d be having the disease, and let s be having symptoms. Then:

$$P(d) = 0.0001$$

$$P(s) = 0.4$$

$$P(s \mid d) = 0.95$$

Therefore:

$$P(d \mid s) = \frac{P(s \mid d)P(d)}{P(s)}$$

$$= \frac{0.95 \times 0.0001}{0.4} = 0.002$$

So, as expected, it is very unlikely for someone to have the disease even if they present with symptoms.

4 Statistical Independence

If knowing B happened has no impact on whether A happens or not, then:

$$P(A \mid B) = P(A)$$

In this case:

$$P(A \cap B) = P(A \mid B)P(B) = P(A)P(B)$$

$$P(x_1 \cap x_2 \cap \cdots \cap x_n) = P(x_1)P(x_2) \cdots P(x_n)$$

This (if it is true) is called *statistical independence*.

Fri 21 Nov 2025 11:00

Lecture P5 - Law of Total Probability and Bayes Theorem

1 Total Probability

Given some sets A and B that are not mutually exclusive, we can take an event in A and break it down into:

1. The piece of A which is also in B.
2. The piece of A which is not also in B.

This means that:

$$P(A) = P(A \cap B) + P(A \cap B^C)$$

1.1 The Law of Total Probability

This becomes more useful when we consider multiple events. Say we have a sequence of disjoint sets $B_1, B_2, \cdots, B_n$. These events tile A. We then have:

$$P(A) = P(A \cap B_1) + P(A \cap B_2) + \cdots + P(A \cap B_n)$$

$$P(A) = \sum_{n=1}^{N} P(A \cap B_n)$$

This distribution $P(A)$ is called the marginal distribution.

2 Change of Notation

We now abstract ourselves away from sets, so change our notation:

$$P(A \cap B) \mapsto P(A, B)$$

This is called the "joint distribution". Our previous equations now become:

$$\text{Marginalisation:} \qquad P(A) = \sum_{n=1}^{N} P(A, B_n)$$

$$\text{Conditional Probability:} \qquad P(A \mid B) = \frac{P(A, B)}{P(B)}$$

$$\text{Statistical Independence:} \qquad P(A, B) = P(A)P(B)$$

2.1 Example 1

If the joint distribution $P(x, y)$ is given by:

$P(x,y)$	$x=0$	$x=1$	$x=2$
$y=0$	$2/9$	$2/9$	$2/9$
$y=1$	$1/9$	$1/9$	$1/9$

Figure 16.1

What is $P(y)$? We marginalise over x so:

$$P(y=1) = \sum_{x=0}^{2} P(x, y=1)$$
$$= 1/9 + 1/9 + 1/9 = 1/3$$

$$P(y=0) = \sum_{x=0}^{2} P(x, y=0)$$
$$= 2/9 + 2/9 + 2/9 = 2/3$$

3 Conditional Probability with the Law of Total Probability

We have that:

$$P(A, B) = P(A \mid B)P(B)$$

So we can adapt our law of total probability to use this:

$$P(A) = \sum_{n=1}^{N} P(A, B_n) = \sum_{n=1}^{N} P(A \mid B_n)P(B_n)$$

3.1 Example

If:

$$P(A \mid B) = 0.2$$
$$P(A \mid \bar{B}) = 0.4$$
$$P(B) = 0.1$$

What is $P(A)$? We marginalise (taking the two possible outcomes as B happening or not):

$$P(A) = P(A \mid B)P(B) + P(A \mid \bar{B})P(\bar{B})$$
$$= 0.2 \times 0.1 + 0.4 \times (1 - 0.1) = 0.38$$

4 Bayes' Theorem

We can now derive the general form of Bayes' Theorem by combining the definition of conditional probability with the Law of Total Probability.

The definition of conditional probability is

$$P(A \mid B) = \frac{P(A, B)}{P(B)}$$

We can recondition to get:

$$P(A, B) = P(B \mid A)P(A)$$

Substituting this into the definition gives:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

However, we often do not know $P(B)$ explicitly. We need to marginalise over all possible states of A to get this:

$$P(B) = \sum_A P(B, A) = \sum_A P(B \mid A)P(A)$$

$$\boxed{P(A \mid B) = \frac{P(B \mid A)P(A)}{\sum_A P(B \mid A)P(A)}}$$

5 Bayes' Examples

5.1 Example 1

Likely to come up on exam.

There are two boxes:

- Box A has 5 gold and 5 silver.
- Box B has 5 gold coins and 10 silver.

If a box is picked at random, and a coin picked at random from this box, what is the probability that the box chosen was A, given the coin was silver.

We want $P(A \mid s)$:

$$P(A \mid s) = \frac{P(s \mid A)P(A)}{P(S \mid A)P(A) + P(S \mid B)P(B)}$$
$$= \frac{1/2}{1/2 + 2/3} = 3/7$$

5.2 Example 2

Polygraphs are used to screen people. Either an individual tells the truth, or they lie. The (imperfect) polygraph returns 0 if it thinks the person lies, or 1 if being truthful, with the following probabilities:

$$P(0 \mid \text{lies}) = 0.88$$

$$P(1 \mid \text{truth}) = 0.86$$

People lie rarely, about 1% of the time. What is the probability that a person is actually lying if the polygraph gives a zero? I.e. we want $P(\text{lie} \mid 0)$

$$P(L \mid 0) = \frac{P(0 \mid L)P(L)}{P(0 \mid L)P(L) + P(0 \mid T)P(T)}$$
$$= \frac{0.88 \times 0.01}{0.88 \times 0.01 + (1 - 0.86) \times (1 - 0.01)}$$
$$= 0.06$$

We therefore also know: $P(T \mid 0) = 0.94$. Even though the polygraph is "accurate", so assume that so few people will lie which makes the results unreliable. This means it is inherently difficult to conduct large-scale testing for rare events.

Thu 27 Nov 2025 09:00

Lecture P6 - Ordered Events and Expectation Values

1 Ordering

So far we've just had fairly abstract events. We now want to add a structure into these events, by considering what happens when the events are ordered. This (rather than physical order) means what if each event is associated with an actual number.

For example, events being getting a head or getting a tail is abstract, while counting the number of heads when tossing a coin is numerical and ordered.

1.1 Probability Mass Functions (PMFs)

If we have a variable x, the probability of observing x is $P(x)$. Since we now consider cases where events are numerical values, we can graph $P(x)$.

We previously had a probability function that assigns probabilities to events. Now we are assigning probabilities to numbers, this is called a probability mass function (PMF).

1.2 Normalisation

The events in a PMF are disjoint from the rest. We can therefore normalise through summation to 1:

$$\sum_x P(x) = 1$$

1.3 Cumulative Distributions

This is the probability that a variable is less than or equal to a certain value.

$$C(x) \equiv P(X \leq x) = \sum_{X \leq x} P(x)$$

x	1	2	3	4	5	6
P(x)	1/6	1/6	1/6	1/6	1/6	1/6
C(x)	1/6	1/3	1/2	2/3	5/6	1

Figure 17.1

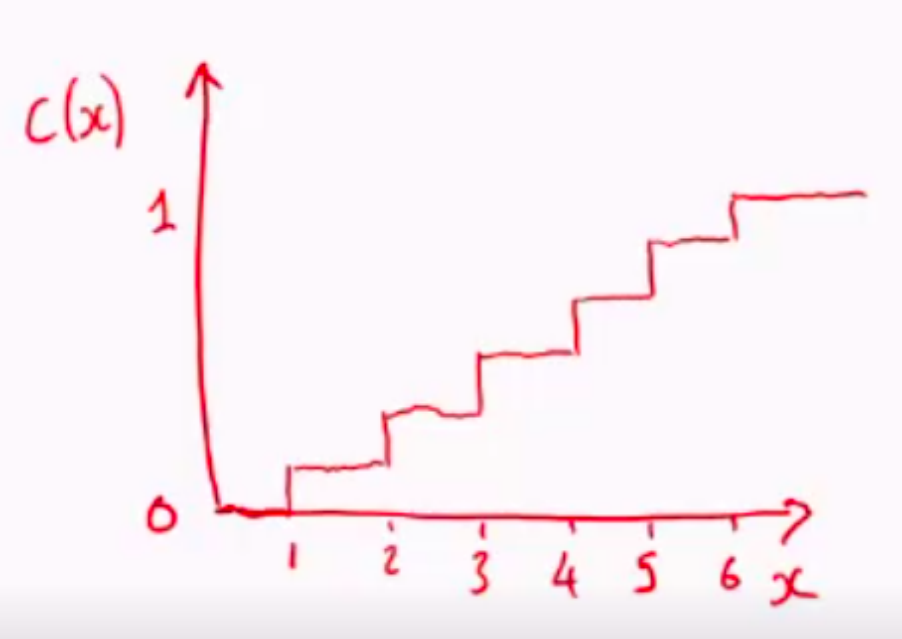

Figure 17.2

2 Expectation Values

The expectation value for a probability mass function is defined by:

$$\langle x \rangle = \sum_x xP(x)$$

It's the weighted sum of all the outcomes (weighted by their probability of occurring), and is a measure of location. Note that the expectation value may not be an actual value that our discrete distribution can obtain, it may be between two.

The best way to describe what it actually represents is the long-term average of a distribution (if you take many draws from the PMF and average them).

2.1 Expectation Value of a Function

$$\langle f \rangle = \sum_x f(x)P(x)$$

This is again a weighted sum over the values of x.

Note: For some constant c:

$$\langle c \rangle = \sum_x xP(x) = c\sum_x P(x) = c$$

$$\langle c \rangle = c$$

This implies that:

$$\langle \langle x \rangle \rangle = \langle x \rangle$$

2.2 Linear Combinations

We stated these as fact earlier in the stats portion of the module. We can now derive them.

$\langle \mathbf{ax + b} \rangle$:

$$\begin{aligned}\langle ax+b\rangle &= \sum_x (ax+b)P(x) \\ &= \sum_x axP(x) = \sum_x bP(x) \\ &= a\sum_x xP(x) + b\sum_x P(x) \\ &= a\langle x\rangle + b\end{aligned}$$

3 Measure of Dispersion

Now we know $\langle x \rangle$, we can ask the question how far, on average, does the value of x get away from this.

In other words, what is the expectation value of the difference between x and $\langle x \rangle$? Lets try:

$$\langle x - \langle x\rangle\rangle = \langle x\rangle = \langle\langle x\rangle\rangle = 0$$

Ah, this is always 0, so doesn't work.... What else can we try?

$$\langle |x - \langle x\rangle|\rangle = \text{MAD}(x)$$

This is the Mean Absolute Deviation of x. This is perfectly valid, but not really very common anymore. More commonly, we use the variance:

$$\langle |x - \langle x\rangle|^2\rangle = \sum_x (x - \langle x\rangle)^2 P(x) = \text{Var}(x)$$

3.1 Simplifying Variance

$$\begin{aligned}\text{Var}(x) &\equiv \langle (x - \langle x\rangle)^2\rangle \\ &= \langle x^2 - 2\langle x\rangle x + \langle x\rangle^2\rangle \\ &= \langle x^2\rangle - \langle 2\langle x\rangle x\rangle + \langle \langle x\rangle^2\rangle \\ &= \langle x^2\rangle - 2\langle x\rangle^2 + \langle x\rangle^2 \\ \text{Var}(x) &= \langle x^2\rangle - \langle x\rangle^2\end{aligned}$$

We are further going to define *standard deviation* as the square root of variance:

$$\text{std}(x) = \sqrt{\langle x^2\rangle - \langle x\rangle^2}$$

3.2 Linear Variance Combinations

We now want to consider the same linear combinations but for variance this time:

$$\begin{aligned}\text{var}(ax+b) &= \langle (ax)^2\rangle - \langle ax\rangle^2 \\ &= \langle a^2x^2\rangle - a^2\langle x\rangle^2 \\ &= a^2\left\{\langle x^2\rangle - \langle x\rangle^2\right\} \\ \boxed{\text{var}(ax+b) = a^2\text{var}(x)}\end{aligned}$$

Fri 28 Nov 2025 11:00

Lecture P7 - Discrete Distributions

1 Parametric Distributions

So far, we've treated $P(x)$ very arbitrarily. We now want to define it a bit more tightly.

So far, we've explicitly defined values for every combination (i.e. $P(x=0)=a, P(x=1)=b$) etc. Instead of doing this, we can define then *parametrically*. We add some number of parameters θ in:

$$P(x \mid \theta)$$

We can have more than one parameter:

$$P(x \mid \theta_1, \theta_2, \cdots, \theta_n) = P(x \mid \boldsymbol{\theta})$$

1.1 Bernoulli Distribution

The Bernoulli Distribution is the simplest distribution we can define. It underpins all of the rest. We define two events, 0 and 1. 1 happens with probability p and 0 with probability $1-p$. We then run a **single** trial.

If:

$$P(x=0 \mid p) = 1-p \qquad P(x=1 \mid p) = p$$

Then:

$$P(x \mid p) = p^x(1-p)^{1-x} \qquad x = 0, 1$$

Or:

$$P(x \mid p) = (1-x)(1-p) + xp \qquad x = 0, 1$$

Properties

$$\langle x \rangle \equiv \sum_x xP(x \mid p) = \sum_{x=0}^{1} xp^x(1-p)^{1-x} = 0p^0(1-p)^1 + 1p^1(1-p)^0 = p$$

$$\langle x^2 \rangle \equiv \sum_x x^2 P(x \mid P) = p$$

$$\text{var}(x) = p - p^2 = p(1-p)$$

1.2 Binomial Distribution

The binomial distribution is the sum of Bernoulli distributions. If we toss N coins and count the number of heads k, what is the distribution of k?

Let $P(H) = p$, so $P(T) = 1-p$.

If $n = 3$:

$$\Omega \{\text{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}\}$$

We assume that each throw of the coin is independent of the previous throw. Therefore $P(\text{HHH}) = P(\text{H})P(\text{H})P(\text{H}) = P(\text{H})^3$

For k heads, we use the union. For $k = 1$, we have:

$$P(k = 1) = P(\text{TTH} \cup \text{THT} \cup \text{HTT})$$

Each one of these options has probability $p(1-p)^2$, so $P(k = 1) = 3p(1-p)^2 = {}^3C_1 p(1-p)^2$ (as there are 3 possible outcomes that give us that combination of H and T).

This gives us:

$$P(k = 0) = \binom{3}{0}(1-p)^3$$

$$P(k = 1) = \binom{3}{1}p(1-p)^2$$

$$P(k = 2) = \binom{3}{2}p^2(1-p)$$

$$P(k = 3) = \binom{3}{3}p^3$$

We can get the final probability mass function:

$$P(k \mid N, p) = \binom{N}{k}p^k(1-p)^{N-k}$$

And therefore the expectation value:

$$\begin{aligned}
\langle k \rangle &\equiv \sum_{k=0}^{N} kP(k \mid N, p) \\
&= \sum_{k=0}^{N} k\frac{N!}{k!(N-k)!}p^k(1-p)^{N-k} \\
&= 0 + \sum_{k=1}^{N} k\frac{N!}{k!(N-k)!}p^k(1-p)^{N-k} \\
&= \sum_{k=1}^{N} \frac{N!}{(k-1)!(N-k)!}p^k(1-p)^{N-k} \\
&= Np\sum_{k=1}^{N} \frac{(N-1)!}{(k-1)!(N-k)!}(p^{k-1})(1-p)^{N-k} \\
&\text{Let: } t = k-1 \implies k = t+1: \\
&= Np\sum_{t=0}^{N-1} \frac{N!}{(t)!(N-1-t)!}(p^r)(1-p)^{N-1-t} \\
&= Np\sum_{t=0}^{N-1} P(t \mid N-1, p) \\
&= Np \times 1 \\
\langle x \rangle &= Np
\end{aligned}$$

We can do the same thing for $\langle k^2 \rangle$ to get:

$$\langle k^2 \rangle = \sum_{k=0}^{N} k^2 P(k \mid N-1, p)$$

$$\langle k^2 \rangle = N^2p^2 - Np^2 + Np$$

Hence:

$$\text{var}(k) = \langle k^2 \rangle - \langle k \rangle^2 = N^2p^2 - Np^2 + Np - N^2p^2 = Np - Np^2 = \boxed{Np(1-p)}$$

1.3 Poisson Distribution

What if p was very small, but N was very large? Consider N people living in a town, each of whom (rarely) goes to a shop, independently with probability p. We use $N \to \infty$ and $p \to 0$, fixed by $Np = \lambda$.

$$P(k \mid N, p) = \frac{N!}{k!(N-k)!} p^x (1-p)^{N-k}$$

If $k \ll N$, then:

$$\frac{N!}{(N-k)!} = \frac{N \times (N-1) \times \cdots \times (N-k+1) \times \overbrace{(N-k) \times \cdots \times 1}^{(N-k)!}}{\underbrace{(N-k) \times \cdots \times 1}_{(N-k)!}}$$

$$N(N-1)(N-2)\ldots(N-k+1) \approx \underbrace{N \cdot N \cdot \ldots \cdot N}_{k \text{ times}} = N^k$$

Hence $N!(N-k)! \approx N^k$ (better and better approximation as N gets larger and larger):

$$P(k \mid N, p) = \frac{N^k}{k!} p^x (1-p)^{N-k}$$

And using $\lambda = np$:

$$P(k) = \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{N}^{N-k}\right)$$

Taking limits as $N \to \infty$:

$$\lim_{N\to\infty} \left(1 - \frac{x}{N}^{N}\right) = e^{-x} \qquad \lim_{N\to\infty} \left(1 - \frac{x}{N}^{a}\right) = 1$$

Hence:

$$\boxed{P(k \mid \lambda) \equiv \frac{\lambda^k}{k!} e^{-\lambda}}$$

This gives us the probability mass function for the Poisson Distribution. It is an approximation for the binomial distribution for very large N. It has expectation value $\langle k \rangle = \lambda$, as the underlying binomial has expectation Np and $\lambda = Np$.

By the same trick, for the underlying binomial var$(k) = Np(1-p)$, so for very large N and very small p, var$(k) = \lambda$ too.

Thu 04 Dec 2025 09:00

Lecture P8 - Multivariate Distributions

For one variable, we can get $P(x)$:

$$\langle x \rangle = \sum_x xP(x)$$

For multiple variables, we get the same thing a slightly different way by marginalising to find $P(x)$ and using that:

$$\langle x \rangle \equiv \sum_x \sum_y xP(x,y) = \sum_x xP(x)$$

The two methods (direct sum including y or marginalising) are equivalent, they're just conceptually different.

1 Sums of Random Variables

We say x was drawn from some distribution $P(x)$ by writing:

$$x \sim P(x)$$

This tells us that x is a random variable, and follows the distribution given after the $\sim$.

Consider N random variables, all drawn from some distribution $P(x)$

$$x_1 \sim P(x) \quad x_2 \sim P(x) \ldots \quad x_n \sim P(x)$$

We define the total:

$$t = x_1 + x_2 + x_3 + \cdots + x_n$$

Working out $P(t)$ directly at this point is too difficult for us. We can however ask about the expected total $\langle t \rangle$ or var(t). Recall the sample mean is:

$$\bar{x} = \frac{1}{N}(x_1 + \cdots + x_n) = \frac{t}{N}$$

We already know expectation is linear, so $\langle x_1 + x + 2 + \cdots + x_n \rangle = \langle x_1 \rangle + \langle x_2 \rangle + \cdots + \langle x_n \rangle$. The expectation value is given easily by:

$$\langle t \rangle = \langle x_1 + x + 2 + \cdots + x_n \rangle$$

For the variance, let's consider $N = 2$:

$$\text{var}(x_1 + x_2) = \langle (x_1 + x_2)^2 \rangle + \langle x_1 + x_2 \rangle^2$$

$$= \langle x_1^2 + x_2^2 + 2x_1x_2 \rangle - (\langle x_1 + x_2 \rangle)^2$$

$$\langle x_1^2 \rangle + \langle x_2^2 \rangle + 2\langle x_1x_2 \rangle - \langle x_1 \rangle^2 - \langle x_2 \rangle^2 - 2\langle x_1 \rangle \langle x_2 \rangle$$

$$\langle x_1^2 \rangle - \langle x_1 \rangle^2 + \langle x_2^2 \rangle - \langle x_2 \rangle^2 + 2\left(\langle x_1x_2 \rangle - \langle x_1 \rangle - \langle x_2 \rangle\right)$$

$$\text{var}(x_1) + \text{var}(x_2) + 2\text{cov}(x_1, x_2)$$

Where $cov(x, y)$ is the covariance.

2 Covariance

Recall that:

$$\text{var}(x) = \sum_x (x - \langle x \rangle)^2 P(x)$$

We now define:

$$\text{cov}(x, y) = \sum_{xy} (x - \langle x \rangle)(y - \langle y \rangle) P(x, y)$$

$$= \langle (x - \langle x \rangle)(y - \langle y \rangle) \rangle = \langle xy - x\langle y \rangle - y\langle x \rangle + \langle x \rangle \langle y \rangle \rangle$$

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

$$\langle xy \rangle - \langle x \rangle \langle y \rangle$$

This is a generalisation for variance for multiple variables, so note that:

$$\text{var}(x) = \text{cov}(x, x)$$

Covariance measures linear association between two variables:

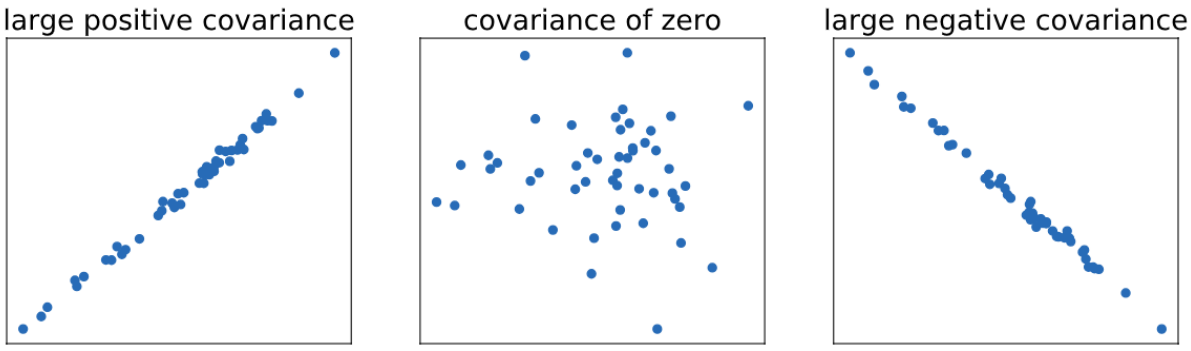

Figure 19.1

The caveat it's only linear is important - there may well be an association between two variables, but if it's not linear covariance won't test for it. For example, a true n-wavelengths of a sine wave will have zero covariance due to symmetry, but there is obviously an association (sinusoidal).

We define an additional measure called correlation using covariance (again, this is linear):

$$\text{corr}(x, y) = \frac{\text{cov(x, y)}}{\text{std}(x)\text{std}(y)}$$

This is useful, because it standardises, i.e the maximum value for perfect linear association from the correlation function is 1. This helps us to standardise regardless of units (i.e. if we were testing the correlation of two measurements with two different units, changing one of the units from km to mm would change the covariance. This is not helpful, as the actual underlying correlation stays the same. The corr function avoids this)

3 Covariance and Independence

If x and y are independent, then we know:

$$P(x, y) = P(x, y)$$

How does this impact the covariance however?

$$\text{cov}(x, y) = \langle xy \rangle - \langle x \rangle \langle y \rangle$$

Assume they are independent, then:

$$\langle xy \rangle = \sum_x \sum_y xyP(x)P(y)$$

$$= \sum_x xP(x) \sum_y xP(y)$$

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

Hence *if* they are independent, *then* the covariance is zero (this does not go backwards).

3.1 Variance of Sum

If the variables are independent, the variance of the sum is the sum of the variances.

If they are not independent, we must include the covariance:

$$\text{var}(x_1 + x_2 + \cdots + x_n) = \sum_n + 2 \sum_{m>n} \text{cov}(x_n, x_m)$$

As an example, what is the expectation value and the variance if we sum N independent Bernoulli variables, each with the same parameter p.

$$x_1 \sim \text{Bern}(p) \qquad x_2 \sim \text{Bern}(p) \qquad \cdots \qquad x_n \sim \text{Bern}(p)$$

For expectation values:

$$\langle x_1 + x_2 + \cdots + x_n \rangle = \sum_{i=1}^{N} \langle x_n \rangle = Np$$

And for variance:

$$\text{var}(x_1 + \cdots + x_n) = \sum_{i=1}^{N} \text{var}(x_n) = Np(1-p)$$

This is our results for the binomial distribution, as expected. Since we assume independence here, this is why the binomial distribution requires independent trials.

4 Change of Variables (Discrete)

We have some $P_x(x)$, in a sample space Ω_x. We make a transformation of x, $y = f(x)$.

For example, we have a particle moving with random velocity, and we want to the random variable that models kinetic energy. Kinetic energy is the new post-transformation random variable - the "induced distribution".

What then is $P_y(y)$ or Ω_y?

Consider a fair six-sided die:

$$P_x(x) = 1/6, \quad x = 1, 2, 3, 4, 5, 6$$

Example 1: $y = x - 2$ This changes the sample space but not the distribution shape itself. This is simple, as it's a bijection:

$$P_y(y) = \frac{1}{6} \qquad y = -1, 0, 1, 2, 3, 4$$

Example 2: $z = |x - 2|$ This is a bit tricker, because it is not a bijection. Two different values of x ($x = 1, 3$) both map to $z = 1$. We therefore need to consider both cases to ensure the new PMF is still normalised:

$$P_z(z) = \begin{cases} 2/6 & z = 1 \ (x = 1, 3) \\ 1/6 & z = 0, 2, 3, 4 \ (x = 2, 4, 5, 6) \end{cases}$$

The general formula for a particular case of $P_y(y)$ is:

$$P_y(y) = \sum_{x:y=f(x)} P_x(x)$$

The $x : y = f(x)$ notation means that we are summing over values of x such that $y = f(x)$. In english, the formula is saying "to find the probability of the new variable having value y, sum the probabilities for all the $x$es that transform into that value of y".

Ω_y is the unique set of values that arise from $y = f(x),\ \forall x \in \Omega_x$.

Fri 05 Dec 2025 11:00

Lecture P9 - Continuous Probability

So far, we have only discussed discrete probability. In this, $P(x)$ means the probability of x happening and we've ignored the possibility of an infinite number of possible events, because it never caused a problem.

1 Continuous Distributions

1.1 Continuous Random Walk

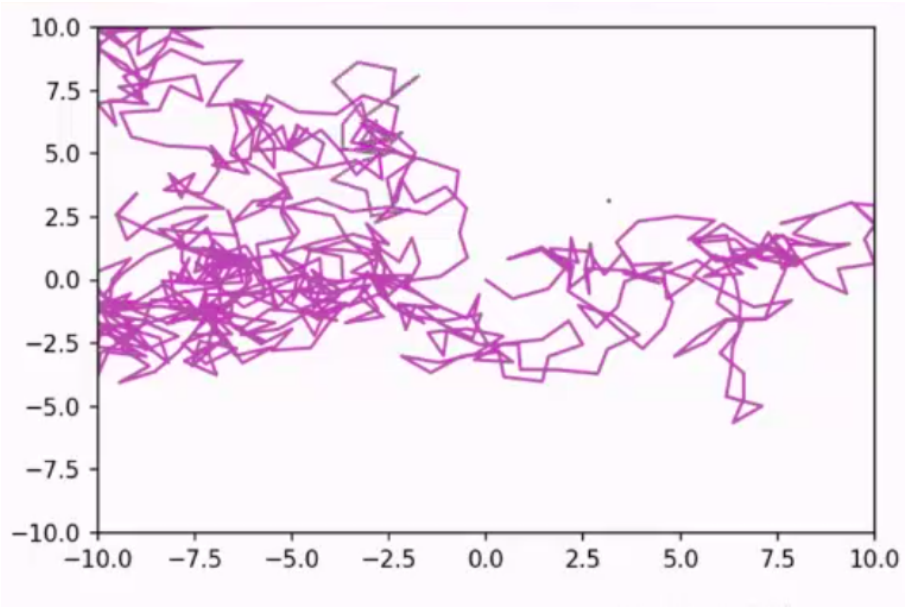

Figure 20.1: The path taken by the particle in a random walk.

We have a particle conducting a random walk. What is the probability of the particle landing exactly on (π, π)? This is zero, and indeed it will be zero for every specific point we can name.

We have to think about probability a bit differently, and we have to ask the question about the particle being *near* (π, π). We cannot consider exact values.

Continuous probability allows possible values to be real numbers (uncountably infinite) whereas before we were limited to discrete integers (countably infinite). Ω is therefore $\mathbb{R}$ or a subset thereof.

We have to set our question to be "What is the probability that x lies in some interval". Effectively, sums become integrals. $P(x)$ is now a probability density function, and the area under $P(x)$ is what gives us probability, rather than values of $P(x)$ alone.

1.2 Properties of a PDF

- $P(x) \geq 0$:
  - $P(x) = 0,\ \forall x \notin \Omega$

  - $P(x) > 0,\ \forall x \in \Omega$

- It is normalised, hence:

$$\int_\Omega P(x)\,dx = 1$$

- Since we care about the area under $P(x)$ to get probabilities, $P(x)$ itself may be bigger than 1, provided the integral is never bigger than 1.

All the old formulae still hold, but with integration instead of summation:

- Expectation Values:

$$\langle x \rangle = \int_\Omega xP(x)dx$$

- Expectation of a Function:

$$\langle f \rangle = \int_\Omega f(x)P(x)dx$$

- Variance:

$$\begin{aligned}\text{var}(x) &= \langle (x - \langle x \rangle)^2 \rangle \\ &= \int_\Omega (x - \langle x \rangle)^2 P(x) \\ &= \langle x^2 \rangle - \langle x \rangle^2\end{aligned}$$

1.3 Example

$$C(x) \equiv$$

Thu 10 Dec 2025 11:00

Lecture P10 - Example Distributions and Central Limit Theorem

1 Uniform Distribution

The continuous uniform distribution is a flat distribution between two points, a and b. The PDF is given by:

$$P(x \mid a, b) = \begin{cases} \dfrac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

1.1 Expectation Value

$$\begin{aligned} \langle x \rangle &= \int_\Omega x P(x)\, dx \\ &= \frac{1}{b-a} \int_a^b x\, dx \\ = \frac{1}{2} \frac{b^2 - a^2}{b-a} &= \frac{1}{2} \frac{(b-a)(b+a)}{(b-a)} \\ \langle x \rangle &= \frac{b+a}{2} \end{aligned}$$

2 Exponential Distribution

The exponential distribution models exponential growth or decay. There are two different forms, depending on what units we give the parameter μ or λ:

$$P(x \mid \mu) = \begin{cases} \mu e^{-\mu x} & \text{if } 0 \leq x < \infty \\ 0 & \text{otherwise.} \end{cases}$$

$$P(x \mid \lambda) = \begin{cases} \dfrac{1}{\lambda} e^{-x/\lambda} & \text{if } 0 \leq x < \infty \\ 0 & \text{otherwise.} \end{cases}$$

2.1 Expectation Value

Considering the first form:

$$\langle x \rangle = \int_0^\infty \mu e^{-\mu x}\, dx$$

Letting $t = \mu x \implies dt = \mu dx$:

$$= \frac{1}{\mu} \int_0^\infty t e^{-t}\, dt$$

Solving by parts:

$$= \frac{1}{\mu} \int_0^\infty \underbrace{t}_{u} \underbrace{e^{-t}}_{dv}\, dt$$

$$= \frac{1}{\mu}\left[-te^{-t}\right]_0^\infty + \frac{1}{\mu}\int_0^\infty e^{-t}\,dt$$

And, as boundary conditions (can do more formally via L'Hopital's rule) cause the first term to equal zero:

$$= \int_0^\infty e^{-t}\,dt$$

$$\langle x \rangle = \left[\frac{1}{\mu} - e^{-}t\right]_0^\infty = \frac{1}{\mu}$$

3 Normal Distribution and the Central Limit Theorem

The Normal distribution (a.k.a Gaussian Distribution)

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

Properties:

- Expectation: $\langle x \rangle = \mu$,
- Variance: $\text{var}(x) = \sigma^2$.

To say that the random variable x follows a normal distribution we write:

$$x \sim \mathcal{N}(\mu, \sigma^2)$$

Or:

$$x \sim \mathcal{N}(x \mid \mu, \sigma^2)$$

3.1 Standard Normal

If $\mu = 0$ and $\sigma = \sigma^2 = 1$, we call the distribution the standard normal distribution.

$$P(x \mid 0, 1) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$$

We can rescale any normal distribution onto the standard normal.

$$\text{let: } z = \frac{x-\mu}{\sigma} \implies x = \sigma z + \mu$$

$$P_z(z) = \left|\frac{d}{dz}f^{-1}(z)\right| P_x(f^{-1}(z))$$

$$= \sigma\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\sigma z + \mu - \mu}{2\sigma^2}\right)$$

$$= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\right)$$

3.2 CLT

The Central Limit Theorem states that rescaled sums of standard variables appear to be Normally distributed. We will not prove it formally.

Let:

$$x_1 \sim P(x) \qquad x_2 \sim P(x) \qquad \cdots \qquad x_n \sim P(x)$$

We define:

$$X = x_1 + x_2 + \cdots + x_n$$

We now form the **rescaled** variable:

$$z = \frac{X - \langle X \rangle}{\text{std}(X)}$$

Assuming independence, we have already shown that:

$$\langle X \rangle = N\langle x \rangle$$

$$\text{var}(X) = N\text{var}(x)$$

Hence:

$$z = \frac{X - N\langle x \rangle}{\sqrt{N}\text{std}(x)}$$

The CTL states that:

$$\lim_{N\to\infty} z = \mathcal{N}(0,1) \qquad \text{if std}(x) < \infty$$

This is regardless of what $P(x)$ is, even if it is not normal.

It also says that averages appear Normal. Consider the sample mean:

$$\bar{x} = \frac{X}{N} = \frac{1}{N}(x_1 + x_2 + \cdots + x_n)$$

$$\langle \bar{x} \rangle = \langle \frac{X}{n} \rangle = \frac{1}{N}\langle X \rangle = \frac{1}{N} N \langle x \rangle = \langle x \rangle$$

And:

$$\text{var}(\bar{x}) = \frac{1}{N^2}\text{var}(X) = \frac{1}{N}\text{var}(x)$$

Again lets define a standard variable:

$$z = \frac{\bar{x} - \langle x \rangle}{\text{std}(\bar{x})/\sqrt{N}} \to \mathcal{N}(0,1)$$

We can invert this to get:

$$\bar{x} = \frac{\text{std}(x)}{\sqrt{N}} z + \langle x \rangle$$

$$\bar{x} \sim \mathcal{N}\left(\langle x \rangle, \frac{\text{var}(x)}{N}\right)$$

$$\bar{x} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{N}\right)$$

This tells us that when we construct a sample mean from something, we can treat it as a Normal distribution (regardless of the source distribution). The variance is a decreasing function of N, so taking more data creates a distribution better centred on the sample mean.

3.3 Example

The Erlang Distribution has PDF:

$$P(x \mid \lambda, k) = \frac{\lambda^x x^{k-1} e^{-\lambda x}}{(k-1)!}$$

With:

$$\langle x \rangle = \frac{k}{\lambda}$$

$$\text{var}(x) = \frac{k}{\lambda^2}$$

What is the limiting distribution of the sum of N Erlang distributed random variables, if N is large?

Let:

$$X = x_1 + x_2 + \cdots + x_n \qquad \text{where all are Erlang distributed random variables}$$

We have:

$$\langle x \rangle = N\frac{k}{\lambda} \qquad \text{var}(x) = N\frac{k}{\lambda^2}$$

Creating our rescaled variable:

$$z = \frac{X - \langle X \rangle}{\text{std}(x)} = \frac{X - (Nk/\lambda)}{\sqrt{\frac{Nk}{\lambda^2}}} \xrightarrow{\text{C.L.T.}} \mathcal{N}(0,1)$$

And inverting and solving for X:

$$X = \sqrt{\frac{Nk}{\lambda^2}} z + \frac{Nk}{\lambda}$$

So:

$$X \sim \mathcal{N}\left(\frac{Nk}{\lambda}, \frac{Nk}{\lambda^2}\right)$$

Fri 12 Dec 2025 11:00

Lecture P11 - End of Probability: Variance Propagation

1 Variance Propagation

Say we have $x \sim P_x(x)$, but we don't know what the actual distribution $P_x(x)$ is. We have however been able to determine (or estimate) values for $\langle x \rangle$ and var(x). We set $y = f(x)$, and what is $P_y(y)$? This is a common scenario in e.g. labs, where we can measure a sample mean and estimate a standard deviation, but we cannot necessarily actually determine the underlying distribution.

We would formally use:

$$P_y(y) = \left| \frac{d}{dy} f^{-1}(y) \right| P_x(x)(f^{-1}(y))$$

But that is impossible here, as we do not know $P_x(x)$. We can however calculated values for $\langle y \rangle$ and var(y).

1.1 Example

Imagine an experiment where we want to determine kinetic energy. This is difficult (as we would have to transfer this energy into another easier to measure form first). Instead, we measure velocity and use:

$$E = \frac{1}{2}mv^2 \quad \text{where } v \text{ is a random variable.}$$

What is $\langle E \rangle$ and var(E)?

We use Taylor Series:

$$f(x) = f(x_0) + (x - x_0)f'(x_0) + \frac{(x - x_0)^2}{2!}f''(x_0) + \cdots$$

This generates a polynomial approximation for $f(x)$ about the point x_0. What if x is a random variable? Lets assume $\langle x \rangle = \mu$ and var$(x) = \sigma^2$.

$$f(x) = f(\mu) + (x - \mu)f'(\mu) + \frac{(x - \mu)^2}{2}f''(\mu) + \cdots$$

We have effectively decoupled the f and the x, allowing us to work with it more easily. We take expectations:

$$\langle f(x) \rangle \approx \langle f(\mu) \rangle + \langle (x - \mu)f'(\mu) \rangle + \frac{1}{2}\langle (x - \mu)^2 f''(\mu) \rangle + \cdots$$

$$\langle f(\mu) \rangle = f(\mu)$$
$$\langle (x - \mu) \rangle f'(\mu) = 0$$
$$\frac{1}{2}\langle (x - \mu)f''(\mu) \rangle = \frac{\sigma^2}{2}f''(\mu)$$

We know the second term is zero, and if we assume the third term and onwards are 'small':

$$\langle f \rangle \approx f(\mu)$$

We therefore use the sample mean as an estimator for μ.

End of Module.