The Likelihood

The *likelihood function* (or simply the *likelihood*) is the probability of observed data conditioned on a particular choice of the model and any model parameters.

In this course, my chosen notation for the likelihood is $\mathcal{L}(D|M)$, where D is the data and M is the model. This is not standard, but the reason for this choice is explained below.

Up until now we have been agnostic about what exactly D and M are. The data D might include text, time series measurements or images. The model M might involve multiple parameters, both continuous and discrete (e.g. labels for data). For the specific case when the data is continuous random variable, $D = x$, and when the model depends only on continuous parameters, $M = \theta$, we will use the notation $\mathcal{L}(x|\theta)$.

The likelihood is a probability distribution for the data conditioned, which is why you will often see the likelihood written in the form $P(x|\theta)$. However, it is regarded as being function of the model parameters (this is the reason for the name *likelihood function*) and this is why you will also often see it written as something like $L(\theta)$. These two guises of the likelihood are illustrated further in Box. 0.1 To avoid the confusion of multiple notations, in these notes we will sick to the hybrid notation $\mathcal{L}(x|\theta)$.

Box 0.1: The two different guises of the likelihood

A source emits unstable particles that decay after travelling a distance x. A number of decay locations are observed, $\{x_1, x_2, \ldots, x_N\}$.

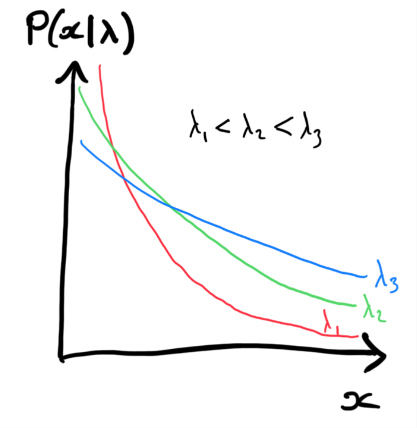
For a single particle, the distance travelled, $x > 0$, has an exponential distribution with a characteristic decay length λ,

$$P(x|\lambda) = \frac{\exp(-x/\lambda)}{\lambda}. \tag{i}$$

Or, for a number of independent particles

$$P(\{x\}|\lambda) = \frac{\exp(-\sum_i x_i/\lambda)}{\lambda^N}. \tag{ii}$$

This is the probability distribution (or likelihood) of the data, conditioned on a specific value for λ. If the decay length is known, this allows us the predict decay locations. (Forward model.)



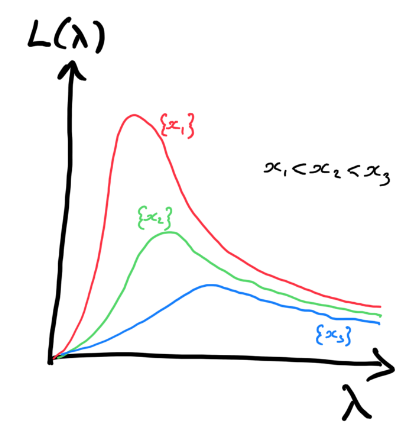
For different values of λ, this describes a set of exponential curves. These curves are all normalised with unit area.

But we've already measured $\{x_i\}$ and want to find the decay length? (Inference, or inverse problem.) The same expression can be used to answer this, when viewed as a function of λ. This is the likelihood function,

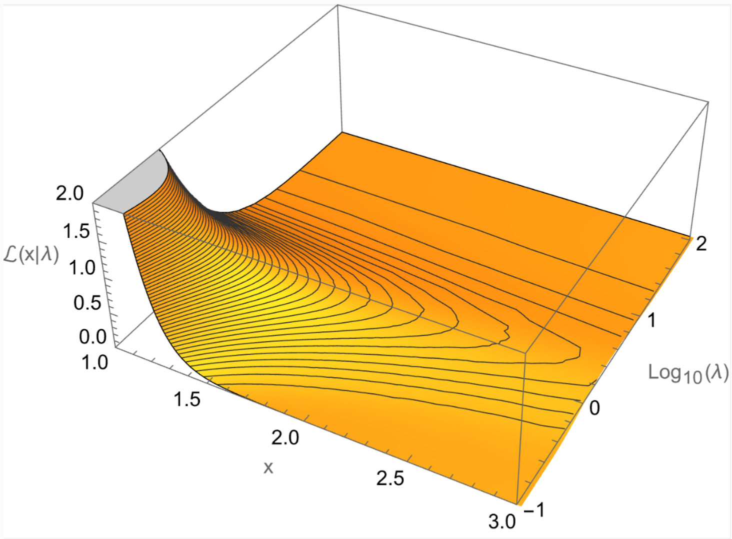
$$L(\lambda) = \frac{\exp(-\sum_i x_i/\lambda)}{\lambda^N}. \tag{iii}$$

We can find the value of the decay length parameter that maximises the likelihood and use this as an estimator of λ. (In practice, it is almost always easier to minimise the log-likelihood, rather than working with the likelihood itself; minimising the log-likelihood is the same as minimising the likelihood because the logarithm function is a monotonic.) The maximum likelihood estimator $\hat{\lambda}$ is defined as the solution of

$$\left.\frac{\partial}{\partial\lambda}\right|_{\lambda=\hat{\lambda}} \log L(\lambda) = 0 \quad \Rightarrow \quad \hat{\lambda} = \frac{\sum_i x_i}{N}. \tag{iv}$$



For different sets of measured decay locations $\{x\}$ this describes a very different set of curves. These curves are NOT normalised. The relationship between the two sets of curves sketched above can be understood by thinking about slices through the surface $\mathcal{L}(x|\lambda) = \exp(-x/\lambda)/\lambda$ taken parallel to both axes.



The likelihood can be thought of in two ways: the *probability distribution* of the data conditioned on the model parameters $P(d|\lambda)$, or as a *function* of the model parameters $L(\lambda)$. But it's always the same expression which we denote here using the hybrid notation $\mathcal{L}(x|\lambda)$.

Constructing a likelihood for a specific problem requires us to know something about the experiment that produced the data. It is also often necessary to make some simplifying assumptions to obtain a useful expression (such as assuming different measurements are perfectly independent, or that errors are exactly Gaussian distributed). These assumptions need to be tested if any subsequent Bayesian inference that uses the likelihood is to be trusted.

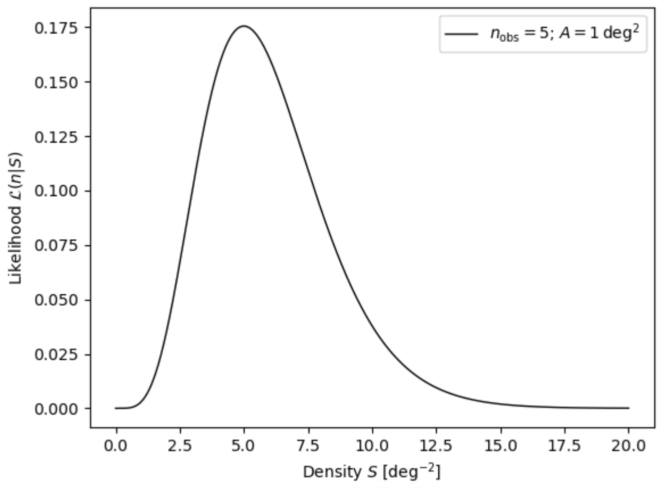
Box 0.2: Number density of stars

We want to measure the density (i.e. number per square degree) of bright stars above a particular threshold magnitude, S, on a particular patch of sky. We perform a survey of the area and find $n = 5$ stars above the threshold in an area $A = 1$ square degree. For simplicity, we will assume that stars occur independently (no binaries!) and uniformly across this patch of the sky at a constant average density S.

Suppose we knew exactly the number density S. (Obviously we don't! That's the whole point, we are going to try to infer S from the count n. But just suppose that we did know S.) Then the likelihood is a Poisson distribution with expectation parameter $\lambda = AS$;

$$\mathcal{L}(n|S) = \frac{(AS)^n \exp(-AS)}{n!}. \tag{i}$$

The likelihood is a conditional probability; given a particular value of S, how likely is it that I found this number of stars?



Think about the area under the curve; only one of these statements is true:

$$\sum_{n=0}^{\infty} \mathcal{L}(n|S) \overset{?}{=} 1 \quad , \quad \int \mathrm{d}S \; \mathcal{L}(n|S) \overset{?}{=} 1. \tag{ii}$$

We'll return to this example several times in the next few lectures.