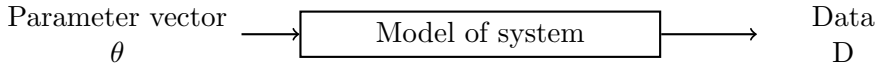
# 12 Inference

Let's consider how we pose and tackle problems in a statistical framework. Suppose we have a statistical model for some real process (from biology, physics, sociology, psychology etc...). By having a model, we mean that given a set of control parameters, $\theta$, we can predict the outcome of the system, $D$. For example, our model may be that each element of $D$ is a draw from one of the distributions we described above so the control parameters $\theta$ are just the parameter(s) of that distribution. Note that the model of the process may include our (imperfect or incomplete) method of measuring the outcome.

In an abstract sense, then, we can consider the model as a black box with input vector $\theta$ (the parameters) and output vector $D$, the data.




The model gives us the forward probability density of the outcome given the parameter, that is, $P(D|\theta)$. This density is the called the likelihood, although, as we see below, we don't usually consider it as a density in the usual way.

This model is not deterministic. The data $D$ can be seen as a random sample from the probability distribution defined by the model (and parameters). Changing the value of the parameters typically does not change the possible outcomes of the model but it will change the shape of the probability distribution, making some outcomes more likely, others less likely.

**Example**: Suppose we are interested the number of buses stopping at a bus stop over the course of an hour. We watch for the hour between 8am and 9am every weekday morning for 2 weeks. We observe the outcomes $D = (10, 7, 5, 6, 12, 9, 10, 5, 14, 7)$. A sensible model here might be the Poisson distribution where we say that the number of bus arrivals in an interval is Poisson distributed with parameter $\lambda$. Our parameter vector contains just the single parameter $\theta = (\lambda)$ and our data vector contains the 10 observed outcomes $D = (D_1, D_2, \ldots, D_{10}) = (10, 7, 5, 6, 12, 9, 10, 5, 14, 7)$.

We derive the likelihood as follows.

The probability of observing the data $D$ for a given value of $\lambda$ is $P(D|\lambda)$. Let's assume that each observation is independent of others then $P(D|\lambda) = \prod_i P(D_i|\lambda)$. That is, the probability of observing this series of outcomes is just the product of the probabilities of observing each particular outcome.

The likelihood of a single observation is given by the probability distribution function for the Poisson since $D_i \sim Poiss(\lambda)$ so:

$$P(D_i|\lambda) = \frac{\lambda^{D_i}}{D_i!} e^{-\lambda}.$$

And so the likelihood of observing the full data $D$ is just

$$P(D|\lambda) = \prod_i P(D_i|\lambda) = \prod_i \frac{\lambda^{D_i}}{D_i!} e^{-\lambda}.$$

□

Note that the likelihood is a probability density function for $D$. But $D$ is typically fixed in the sense that we make the observations which remain fixed through-out the analysis. We will be interested in considering the likelihood as a function of the parameters $\theta$. The likelihood is *not* a probability density function for $\theta$ since, in general $\int_{\theta\in\Theta} P(D|\theta)\, d\theta \neq 1$.

## 12.1 Bayesian inference

The statistical problem essentially comes down to one of observing the outcome, $D$ and wanting to recover the parameters $\theta$.

That is, we want to estimate $\theta$ given $D$. We summarise our estimate of $\theta$ as a probability distribution, conditional on having observed $D$: $P(\theta|D)$. This is called the **posterior distribution** of $\theta$.

From Bayes' theorem, we can express the posterior in terms of the likelihood:

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)},$$

where $P(D|\theta)$ is the **likelihood**, $P(\theta)$ is the **prior distribution** of $\theta$ and $P(D)$ is a normalisation constant.

The prior $p(\theta)$ summarises what we know about a parameter before making any observations.

The posterior, $p(\theta|D)$ summarises what we know about $\theta$ after observing the data.

The likelihood tells us about the likeliness of the data under the model for any value of $\theta$. Recall that we consider the likelihood a function of $\theta$ rather than a probability density for $D$; to emphasise this fact, people often write it explicitly as a function of $\theta$: $L(\theta) = P(D|\theta)$.

Bayes' theorem tells us how we update our beliefs given new data. Our updated beliefs about $\theta$ are encapsulated in the posterior, while are initial beliefs are encapsulated in the prior. Bayes' theorem simply tells us that that we obtain the posterior by multiplying the prior by the likelihood (and dividing by $P(D)$ which we can think of as a normalisation constant).

Note that we need the normalisation constant as the posterior is a probability distribution for $\theta$, so its density must integrate to 1, i.e., $\int_{\theta\in\Theta} f(D|\theta)\, d\theta = 1$. Thus the normalisation constant is $P(D) = \int_{\theta\in\Theta} P(D|\theta)P(\theta)\, d\theta$. Typically this integral is hard to calculate so we try to find that will avoid having to calculate it.

**Example**: In the example above, we found an expression for the likelihood. To find an expression for the posterior, we need to decide on a prior distribution. Suppose we

had looked up general info about bus stops in the city and found that the busiest stop had an average of 30 buses an hour while the quietest had an average of less than 1 bus per hour. We use this prior information to say that any rate parameter $\lambda$ producing an average of between 0 ($\lambda = 0$) and 30 ($\lambda = 30$) buses an hour is equally likely. This leads us to the prior $\lambda \sim U(0, 30)$. The density of this prior is $f(\lambda) = 1/30$ for $0 \leq \lambda \leq 30$.

To get the posterior density, we use the formula above:

$$f(\lambda|D) = \frac{f(D|\lambda)f(\lambda)}{P(D)} = \frac{\prod_i \frac{\lambda^{D_i}}{D_i!}e^{-\lambda}\frac{1}{30}}{P(D)}.$$

The normalisation constant $P(D)$ is the integral of the numerator over all possible values of $\lambda$:

$$P(D) = \int_0^{30} \prod_i \frac{\lambda^{D_i}}{D_i!}e^{-\lambda}\frac{1}{30}\, d\lambda.$$

□

While it is possible to calculate this particular integral analytically, for most posterior distributions analytical integration is either very difficult or impossible. We'll investigate methods for avoiding calculating difficult integrals like this in later sections.

## 12.2 Maximum likelihood

It is often difficult or inconvenient to deal with the posterior distribution (when the prior is hard to specify or the normalisation constant is impossible to calculate). In these cases, we can still use our probabilistic model by concentrating solely on the likelihood function. The aim here is typically to find the parameters that maximise the likelihood function. That is, those parameters under which the observed data is most likely. We call this parameter estimate the maximum likelihood estimate and write it as

$$\hat{\theta} = \arg\max_\theta f(D|\theta) = \arg\max_\theta L(\theta; D)$$

This function can be maximise using standard tools from calculus (taking the derivative and setting it to zero – it is often easier to work with the log of the likelihood function as they both share a maximum) or using numerical techniques such as hill-climbing.

Many methods in statistics are based on maximum likelihood including regression, $\chi^2-$tests, $t-$tests, ANOVA and so on.

**Example**: In the bus example above, we could find the maximum likelihood estimator for $\lambda$ by differentiating the log-likelihood, $\log(L(\lambda; D))$ with respect to $\lambda$, setting the result to zero and solving. Note that we often work with the log-likelihood rather than the likelihood for a couple of reasons: it is often easier algebraically and it helps avoid numerical under-flow when the likelihood itself is very small.