# Bayes' Theorem

Bayes' theorem concerns *conditional probabilities.* These are denoted $P(A|B)$ which is read as “probability of $A$ given that $B$ is true”.

The *product rule* (sometimes known as the *chain rule*) of probability states that

$$P(A\,\mathrm{and}\,B) = P(A|B)P(B)\,. \tag{1}$$

This can either be considered as one of the axioms of probability or (rearranged for $P(A|B)$) as the definition of the conditional probability.

Two events $A$ and $B$ are said to be *independent* if and only if $P(A\,\mathrm{and}\,B) = P(A)(B)$. Two events are independent if one of them occurring does not affect the probability of the other occurring; from equation 1 we see that

$$P(A\,\mathrm{and}\,B) = P(A)P(B) \quad\Longleftrightarrow\quad P(A|B) = P(A)\,, \tag{2}$$
$$P(B\,\mathrm{and}\,A) = P(B)P(A) \quad\Longleftrightarrow\quad P(B|A) = P(B)\,. \tag{3}$$

**Theorem 0.0.1.** *Bayes' theorem;*

$$P(A|B) = P(B|A)\frac{P(A)}{P(B)}\,. \tag{4}$$

*Proof.* Follows from the product rule (Eq. 1) and the symmetry of “and”; $P(A\,\mathrm{and}\,B) = P(B\,\mathrm{and}\,A)$. □

It is a common notation to abbreviate the “and” in formulas with just a comma; i.e. we write $P(A\,\mathrm{and}\,B) = P(A,B)$.

For inference, we will usually work with Bayes' theorem in the following alternative form;

$$P(M|D,I) = \frac{P(D|M,I)P(M|I)}{P(D|I)}\,. \tag{5}$$

Here $M$ denotes our model (with particular values for all of its parameters) and $D$ denotes the data. Every probability is conditional on $I$, this denotes any/all other prior information that is assumed throughout the analysis; although conceptually important and sometimes containing subtle assumptions that profoundly affect the results of our inference, $I$ is usually omitted from our notation in all terms in equation 7.

Equation 7 is so important and occurs so often that special names and symbols are given to each of the individual terms:

1. $P(D|M,I) = \mathcal{L}(D|M)$ is called the **likelihood**, it is the probability of obtaining the data given our model;
2. $P(M|I) = \pi(M)$ is called the **prior**, it the probability we assign to our model being correct *before* performing the experiment;
3. $P(M|D,I)$ is called the **posterior**, it is the probability that our model is correct *after* performing the experiment;
4. $P(D|I) = Z$ is called the **evidence**, it acts as a normalisation constant in Bayes' theorem. (The evidence isn't usually very important in parameter estimation problems, but is plays a central role in model selection). If the model has some free parameters, $M =$ parameters, then from Bayes' theorem (7) we see that the evidence plays the role of a normalisation constant for the posterior probability distribution and is equal to the following integral;

$$Z = \int \mathrm{d}(\text{parameters})\ \pi(\text{parameters})\mathcal{L}(D|\text{parameters})\ . \tag{6}$$

Using this notation, Bayes' theorem is written as

$$P(M|D) = \frac{\mathcal{L}(D|M)\pi(M)}{Z}\ . \tag{7}$$

In a Bayesian inference problem we are generally given the likelihood and the prior as inputs and it our task is to study the posterior (and sometimes compute the evidence).

**Box 0.1: Bayesian versus frequentist probabilities**

Historically, there has been debate between the *Bayesian* and *frequentist* views of probability. The debate does *not* concern the validity of Bayes' theorem itself (Eq. 4) but rather the way it is applied to a model and its parameters (in Eq. 7).

In the frequentist view, probability is defined as the relative frequency of an event occuring in the limit of many trials. This is the way most of us are first taught to think about probability as young children. This means that we must (at least in principle) be able to at imagine repeating some kind of trial (or experiment) many times in order to assign a probablity to the outcome. For this reason, the likelihood, $\mathcal{L}(D|M)$, is a valid frequentist probability because it assigns a probability to the data, $D$, resulting from an experiment which can be repeated.

In the Bayesian view, probabilities are instead interpretted as representing our state of knowledge or as quantification of a personal belief about a certain quantity. Anything we don't know with certainty can be assigned a probability. What's the probability that the $10^{100}$th digit of Pi is 3? For this reason, the prior and posterior, $\pi(M)$ and $P(M|D)$, are valid Bayesian probabilities but not frequentist ones. (How do you repeat a model?) The Bayesian view allows us apply the tools of probability to a wider range of problems and this has proved very fruitful.