1 Inference

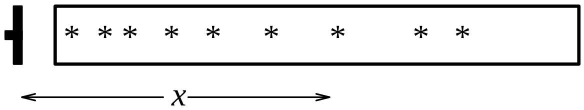
Fortunately, it is not a controversial statement that Bayes' theorem provides the correct language for describing communication over a noisy channel. But let's take a little tour of other applications of probabilistic inference.

Coherent inference can be mapped onto probabilities (Cox 1946). Many textbooks on statistics do not mention this fact, so maybe it is worth using an example to emphasize the contrast between Bayesian inference and the orthodox methods of statistical inference involving estimators, confidence intervals, hypothesis testing, etc.

A first example of probability theory

When I was an undergraduate in Cambridge, I was privileged to receive supervisions from Steve Gull. Sitting at his desk in a dishevelled office in St. John's College, I asked him how one ought to answer an old Tripos question:

> Unstable particles are emitted from a source and decay at a distance x that has an exponential probability distribution with characteristic length λ. Decay events can only be observed if they occur in a window extending from $x = 1$cm to $x = 20$cm. N decays are observed at locations $\{x_1 \ldots x_N\}$. What is λ?



I had scratched my head over this for some time. It was easy to invent an 'estimator' $\hat{\lambda} = \bar{x} - 1$ that was appropriate for $\lambda \ll 20$cm; with a little ingenuity and the introduction of ad hoc bins, promising estimators for $\lambda \gg 20$ cm could be constructed. But there was no obvious estimator that would work under all conditions.

Please stop and think about this problem for a moment.

. . . .

Steve wrote:

$$P(x|\lambda) = \begin{cases} \frac{1}{\lambda} e^{-x/\lambda}/Z(\lambda) & 1 < x < 20 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where

$$Z(\lambda) = \int_1^{20} dx\, \frac{1}{\lambda} e^{-x/\lambda} = \left(e^{-1/\lambda} - e^{-20/\lambda}\right). \tag{2}$$

This seemed obvious enough. Then he wrote:

$$P(\lambda|\{x_1 \ldots x_N\}) = \frac{P(\{x\}|\lambda)P(\lambda)}{P(\{x\})} \tag{3}$$

$$\propto \frac{1}{(\lambda Z(\lambda))^N} \exp\left(-\textstyle\sum_1^N x_i/\lambda\right) P(\lambda). \tag{4}$$

Suddenly, the straightforward distribution $P(\{x_1 \ldots x_N\}|\lambda)$, defining the probability of the data given the hypothesis λ, was being turned on its head so as to define the probability of a hypothesis given the data. A simple figure showed the probability of a single data point $P(x|\lambda)$ as a familiar function of x, for different values of λ (figure 1). Each curve was an innocent exponential, normalized to have area 1. Plotting the same function as a function of λ for a fixed value of x, something remarkable happened: a peak emerges (figure 2).

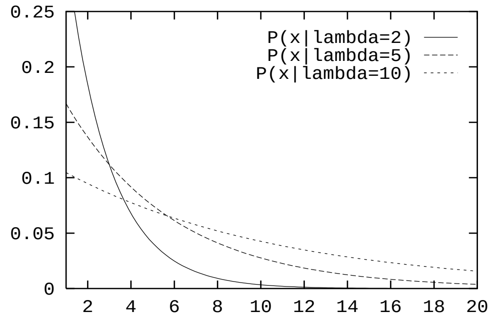


Figure 1: **The probability density $P(x|\lambda)$ as a function of x.**

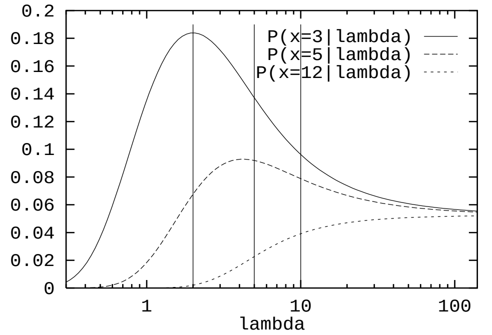


Figure 2: **The probability density $P(x|\lambda)$ as a function of λ.**
When plotted this way round, the function is known as the 'likelihood'.

Steve summarised Bayes' theorem (equation 3) as embodying the fact that 'What you know about λ after the data arrive is what you knew before [$P(\lambda)$], and what the data told you [$P(\{x\}|\lambda)$]'. Probabilities are used here to quantify degrees of belief. To nip possible confusion in the bud, it must be emphasized that the hypothesis λ which correctly describes the situation is *not* a stochastic variable, and the fact that the Bayesian uses a probability distribution P does *not* mean that he thinks of the world as stochastically changing its nature between the states described by the different hypotheses. He uses the notation of probabilities to represent his beliefs about the mutually exclusive micro-hypotheses, of which only one is actually true. That probabilities can denote degrees of belief, given assumptions, seemed intuitive to me, and is proved by Cox (1946).

The posterior probability distribution of equation (4) represents the unique and complete solution to the problem. There is no need to invent estimators; nor do we need to invent criteria for comparing alternative estimators with each other. Whereas orthodox statisticians offer twenty-seven ways of solving a problem, and another twenty different criteria for deciding which of these solutions is the best, Bayesian statistics only offers one answer to a well-posed problem.

Our inference is conditional on our assumptions [for example, the prior $P(\lambda)$]. Critics view such priors as a difficulty because they are 'subjective', but I don't see how it could be otherwise. How can one perform inference without making assumptions? I believe that it is of great value that Bayesian methods force one to make these tacit assumptions explicit. First, once assumptions are made, the inferences are objective and unique, reproduceable with complete agreement by anyone who has the same information and makes the same assumptions. For example, given the assumptions listed above $\mathcal{H}$ and the data D, everyone will agree about the posterior probability of the decay length λ:

$$P(\lambda|D,\mathcal{H}) = \frac{P(D|\lambda,\mathcal{H})P(\lambda|\mathcal{H})}{P(D|\mathcal{H})} \tag{5}$$

Second, when the assumptions are explicit, they are easier to criticize, and we can quantify the sensitivity of our inferences to the details of the assumptions. We can note from the likelihood curves in figure 2 that in the case of a single data point at $x=5$, the likelihood function is less strongly peaked than in the case $x=3$; the details of the prior $P(\lambda)$ become more important if the sample mean $\bar{x}$ is close to 10.5. In the case $x=12$, the likelihood function doesn't have a peak at all. Such data merely rule out small values of λ, and don't give any information about the relative probabilities of large values of λ. So in this case, the details of the prior at the small λ end of things are not important, but at the large λ end, our prior is important.

Third, when we are not sure which of various alternative assumptions is the most appropriate for a problem, we can treat this question as another inference task. Thus, given data D, we can compare alternative assumptions $\mathcal{H}$ using Bayes' theorem:

$$P(\mathcal{H}|D,\mathcal{I}) = \frac{P(D|H,\mathcal{I})P(\mathcal{H}|\mathcal{I})}{P(D|\mathcal{I})}, \tag{6}$$

where $\mathcal{I}$ denotes the highest assumptions of all, which we are not questioning.

Fourth, we can take into account our uncertainty regarding such assumptions when we make subsequent predictions. Rather than choosing one particular assumption $\mathcal{H}^*$, and working out our predictions about some quantity $\mathbf{t}$, $P(\mathbf{t}|D,\mathcal{H}^*,\mathcal{I})$, we obtain predictions that take into account our uncertainty about $\mathcal{H}$ by using the sum rule:

$$P(\mathbf{t}|D,\mathcal{I}) = \sum_{\mathcal{H}} P(\mathbf{t}|D,\mathcal{H},\mathcal{I})P(\mathcal{H}|D,\mathcal{I}). \tag{7}$$

(This is another contrast with orthodox statistics, in which it is conventional to 'test' a default model, and then, if the test 'accepts' the model, to use that model exclusively to make predictions.)

Steve thus persuaded me that

> Probability theory reaches parts that ad hoc methods cannot reach.

Let's look at a few more examples of simple inference problems. The following example illustrates that there is more to Bayesianism than the priors.

An example of legal evidence

Two people have left traces of their own blood at the scene of a crime. Their blood groups can be reliably identified from these traces and are found to be of type 'O' (a common type in the local population, having frequency 60%) and of type 'AB' (a rare type, with frequency 1%). A suspect is tested and found to have type 'O' blood. Do these data $D=$ (type 'O' and 'AB' blood were found at scene) make it more probable that this suspect was one of the two people present at the crime? A careless lawyer might claim that the fact that the suspect's blood type was found at the scene is positive evidence for the theory that he was present.

Denote the proposition 'the suspect and one unknown person were present' by S. The alternative, $\bar{S}$, states 'two unknown people from the population were present'. The prior in this problem is the prior probability ratio between the propositions S and $\bar{S}$. This quantity is important to the final verdict and would be based on all other available information in the case. Our task here is just to evaluate the contribution made by the data D, that is, the likelihood ratio, $P(D|S,\mathcal{H})/P(D|\bar{S},\mathcal{H})$. In general, a jury's task should be to multiply together carefully evaluated likelihood ratios from each independent piece of admissible evidence.

The probability of the data given S is the probability that one unknown person drawn from the population has blood type AB.

$$P(D|S,\mathcal{H}) = p_{AB} \tag{8}$$

The probability of the data given $\bar{S}$ is the probability that two unknown people drawn from the population have types O and AB.

$$P(D|\bar{S},\mathcal{H}) = 2\,p_O\,p_{AB} \tag{9}$$

In these equations $\mathcal{H}$ denotes the assumptions that two people were present and left blood there, and that the probability distribution of the blood groups of unknown people in an explanation is the same as the population frequencies.

Dividing, we obtain the likelihood ratio:

$$\frac{P(D|S,\mathcal{H})}{P(D|\bar{S},\mathcal{H})} = \frac{1}{2p_0} = \frac{1}{2\times 0.6} = 0.83 \tag{10}$$

Thus the data in fact provide weak evidence *against* the supposition that this suspect was present.

This result may be found surprising, so let us examine it from various points of view. First consider the case of another suspect who has type AB. Intuitively, the data do provide evidence in favour of the theory S' that this suspect was present, relative to the null hypothesis $\bar{S}$. And indeed the likelihood ratio in this case is:

$$\frac{P(D|S',\mathcal{H})}{P(D|\bar{S},\mathcal{H})} = \frac{1}{2\,p_{AB}} = 50. \tag{11}$$

Now let us change the situation slightly; imagine that 99% of people are of blood type O, and the rest are of type AB. The data at the scene are the same as before. Consider again how these data influence our beliefs about a particular suspect of type O and another of type AB. Intuitively, we still believe that the presence of the rare AB blood provides positive evidence that the suspect of type AB was there. But do we still have the feeling that the fact that type O blood was detected at the scene favours the hypothesis that the type O suspect was present? If this were the case, that would mean that regardless of who the suspect is, the data make it more probable they were present, which would be absurd. The data may be *compatible* with any suspect of either blood type being present, but if they provide positive evidence for some theories, they must also provide evidence against other theories.

Here is another way of thinking about this: imagine that instead of two people's blood stains there are ten, and that in the entire local population of one hundred, there are ninety type O suspects and ten type AB suspects. Consider a particular type O suspect: without any other information, there is a one in 10 chance that he was at the scene. We now get the results of blood tests, and find that nine of the ten stains are of type AB, and one of the stains is of type O. Does this make it more likely that the type O suspect was there? No, although he could have been, there is now only a one in ninety chance that he was, since we know that only one person present was of type O.

Maybe the intuition is aided finally by writing down the formulae for the general case where n_O blood stains of individuals of type O are found, and n_{AB} of type AB, a total of N individuals in all, and unknown people come from a large population with fractions p_O, p_{AB}. The task is to evaluate the likelihood ratio for the two hypotheses S, 'the type O suspect and $N-1$ unknown others left N stains', and $\bar{S}$, 'N unknowns left N stains'. The probability of the data under hypothesis $\bar{S}$ is just the probability of getting n_O, n_{AB} individuals of the two types when N individuals are drawn at random from the population:

$$P(n_O, n_{AB}|\bar{S}) = \frac{N!}{n_O! n_{AB}!} p_O^{n_O} p_{AB}^{n_{AB}}. \tag{12}$$

In the case of hypothesis S, we need to predict the distribution of the $N-1$ other individuals:

$$P(n_O, n_{AB}|S) = \frac{(N-1)!}{(n_O-1)! n_{AB}!} p_O^{n_O-1} p_{AB}^{n_{AB}}. \tag{13}$$

The likelihood ratio is:

$$\frac{P(n_O, n_{AB}|S)}{P(n_O, n_{AB}|\bar{S})} = \frac{n_O/N}{p_O}. \tag{14}$$

This is a very instructive result. The likelihood ratio, *i.e.* the contribution of this data to the question of whether the type O suspect was present, depends simply on a comparison of the frequency of type O blood in the observed data with the background frequency of type O blood in the population. There is no dependence on the counts of the other types found at the scene, or their frequencies in the population. If there are more type O stains than the average expected by chance (hypothesis $\bar{S}$), then the data gives evidence in favour of the presence of this type O suspect. Conversely, if there are fewer type O stains than the expected number under $\bar{S}$, then the data reduce the probability of the hypothesis that he was there. In the special case $n_O/N = p_O$, the data contribute no evidence either way, regardless of the fact that the data are compatible with the hypothesis S.

References

Cox, R. (1946) Probability, frequency, and reasonable expectation. *Am. J. Physics* **14**: 1–13.