# Bayesian Inference and Prediction

A typical situation in statistics is that we have a sequence of identical experiments and we have to extract information from the resulting data. The numerical results of the experiments are modelled as a sequence $X_1, X_2, \ldots, X_n$ of independent random variables following a distribution $f_\theta$ where $\theta$ is the parameter that we wish to determine using the data.

So far, we have been assuming that all values of $\theta$ are permissible (or that all values from a particular set of values are equally likely). However, this need not always be the case. We may have some information "before-the-experiment" (*a priori* knowledge) that may tell us the probability distribution of $\theta$. Another possibility is that we already did some experiments. As seen earlier, we need not see the result of the experiment as a *definite* value of $\theta$, rather we can see the result as a probability distribution for $\theta$.

Bayesian statistics takes the approach that data collected allows us to *modify* our "before-the-experiment" distribution into an "after-the-experiment" distribution. In other words, it allows us to use the information gathered to alter our perception of the probability distribution of $\theta$.

## Bayesian Inference

Let us start with a simple example of two different types of coins. The first type $\theta_1$ is a fair coin and has a probability $1/2$ of showing a head. The second type $\theta_2$ is a biased coin as has a probability $2/3$ of showing a head. We also assume that the coin is being taken out a box that has 3 coins of the first type and 5 coins of the second type. If we assume that each coin is equally likely to be picked then $P(\theta = \theta_1) = 3/8$ and $P(\theta = \theta_2) = 5/8$.

Now we pick a coin of the box as above and flip it 100 times to get 60 heads. Is the coin of type 1 or of type 2?

If the coin is of type 1, the probability of getting the above result is

$$L(\theta_1) = P(S_{100} = 60 | \theta = \theta_1) = \binom{100}{60} \frac{1}{2^{100}}$$

Similarly, if the coin is of type 2, the probability of getting the above result is

$$L(\theta_2) = P(S_{100} = 60 | \theta = \theta_2) = \binom{100}{60} \left(\frac{2}{3}\right)^{60} \frac{1}{3^{40}}$$

Note that these are the same as the likelihoods that we calculated earlier. We can estimate

$$L(\theta_1)/L(\theta_2) = \frac{3^{100}}{2^{160}} \approx \frac{(2^3 \cdot 10)^{25}}{2^{160}} \approx \frac{10^{25}}{2^{85}} \approx \frac{10^{25}}{3 \cdot 10 \cdot 10^{24}} = 1/3$$

(Note that the approximate value from a calculator is 0.35 so our rapid estimates are quite good!)

So we see that $L(\theta_2)$ is about 3 times $L(\theta_1)$. By our earlier discussion this is not enough to rule out either of these possibilities! Moreover, $\theta_1$ and $\theta_2$ were *a priori* not equally likely. We use Bayes rule to calculate the probability of the event that occurred

$$\begin{aligned} &P(S_{100} = 60) = \\ &\qquad P(S_{100} = 60 | \theta = \theta_1) P(\theta_1) + P(S_{100} = 60 | \theta = \theta_2) P(\theta_2) \\ &\qquad\qquad = (1/8) \binom{100}{60} \left( 3 \frac{1}{2^{100}} + 5 \frac{2^{60}}{3^{100}} \right) \end{aligned}$$

Note that the first term is $P(S_{100} = 60 \wedge \theta = \theta_1)$. Using Bayes rule again we can then calculate

$$\begin{aligned} &P(\theta = \theta_1 | S_{100} = 60) = \\ &\qquad \frac{P(\theta = \theta_1 \wedge S_{100} = 600)}{P(S_{100} = 60)} = \frac{3 \cdot 3^{300}}{3 \cdot 3^{300} + 5 \cdot 2^{160}} \\ &\qquad\qquad \approx \frac{3(1/3)}{3(1/3) + 5} = 1/6 \end{aligned}$$

We similarly estimate that $P(\theta = \theta_2 | S_{100} = 60) = 5/6$. (Note that the approximate values obtained using a calculator are 0.17 and 0.83, so our estimates are not badly off!)

We see that $\theta_2$ is 5 times as likely as $\theta_1$ once we take into account the information that before-the-experiment $\theta_2$ was $5/3$ as likely as $\theta_1$. This follows from

$$P(\theta = \theta_1 | \text{Result}) \text{ is proportional to } P(\theta = \theta_1) L(\theta_1)$$

Since we *have* the Result of the experiment, we can now use the probability $P(\theta = \theta_1 | \text{Result})$ as the distribution after-the-experiment.

The above situation can be seen to hold in general and is the cornerstone of calculations involving Bayesian Inference. The distribution for $(\theta | \text{Result})$, or equivalently, the distribution of $\theta$ after-the-experiment is proportional to the product of the distribution before-the-experiment with the likelihood of $\theta$ given the result of the experiment. The constant of proportionality can be calculated by using the fact that the sum of the probabilities over all possible values of $\theta$ is 1.

In the case of a discrete probability distribution, this becomes

$$P(\theta = \theta_i | \text{Result}) = \frac{P(\theta_i)L(\theta_i)}{\sum_j P(\theta_j)L(\theta_j)}$$

In the (absolutely) continuous distribution case, the statement is about probability densities and the sum in the denominator gets replaced by an integral.

## Bayesian Prediction

One may want to continue the experiment as described above and use the data gathered to make a prediction of the outcome.

Returning to the example of the coins. One can conclude from the experiment that $\theta = \theta_2$ is more likely, and thus predict that a new coin flip will result in Head with a probability of 2/3. This is the "plug-in" method which uses Bayesian Inference to make a prediction.

A different approach is to continue to use Bayes rule to make a prediction. Let $R$ be the event $S_{100} = 60$, or more generally the event representing the Result of the experiment. Let $H$ denote the event that represents a Head occurring in the new coin flip. We apply Bayes' rule to get

$$P(H \wedge (\theta = \theta_1) \wedge R) = P(H| \ (\theta = \theta_1) \wedge R) P((\theta = \theta_1) \wedge R)$$

Now, the new experiment is *independent* of the earlier experiments. So

$$P(H| \ (\theta = \theta_1) \wedge R) = P(H| \ (\theta = \theta_1))$$

Moreover, we can also apply Bayes rule to the second factor to get

$$P(H \wedge (\theta = \theta_1) \wedge R) = P(H| \ (\theta = \theta_1)) P((\theta = \theta_1)| \ R) P(R)$$

We obtain a similar formula with $\theta_2$ instead of $\theta_1$. It follows that

$$P(H \wedge R) = \sum_i P(H \wedge (\theta = \theta_i) \wedge R) = \sum_i P(H| \ (\theta = \theta_i)) P((\theta = \theta_i)| \ R) P(R)$$

Dividing both sides by $P(R)$ and applying Bayes Rule once again, we get

$$P(H| \ R) = \sum_i P(H| \ (\theta = \theta_i)) P((\theta = \theta_i)| \ R)$$

In other words, the probability of getting a Head given the result already obtained is the convex combination of the probability of getting Head with each coin with weight as the probability of that particular coin being the chosen one given the Result. Write out these words as a mathematical formula (and using the previous section) we have

$$P(H| \ R) = \frac{\sum_i P(H| \ (\theta = \theta_i)) P(\theta = \theta_i) L(\theta_i)}{\sum_i P(\theta = \theta_i) L(\theta_i)}$$

In our specific example, we see that (using the approximations earlier)

$$P(H|S_{100} = 60) \approx (1/2)(1/6) + (2/3)(5/6) = 23/36$$

This is very close to the answer 2/3 which we got by the plug-in method but is actually a bit smaller. (Using a calculator we get this to be approximately 0.64.)

## A Healthy Counter-example

Even though we may feel happy with 95% confidence in our procedures, we should know that this is not always adequate. To see why let us consider the following example (taken from Lavine's book on Statistics).

Suppose that 1 out of every thousand persons has a certain disease.

Suppose that there is test for the presence of the disease which is 95% accurate.

We formulate this information as follows. Let $D$ be the random variable representing which takes the value 1 if the chosen person has the disease and 0 otherwise. We have $P(D = 1) = 1/1000$ and $P(D = 0) = 999/1000$.

Let $T$ be the random variable representing the *result of the* “test* on a randomly chosen member of the population. We are given that the test is 95% successful. In other words:

- Given that the person has the disease ($D = 1$) the probability of the test showing the disease ($T = 1$) is 0.95; in terms of probability theory we write this as $P(T = 1|D = 1) = 0.95$.
- Given that the person does not have the disease ($D = 0$) the probability that the test showing the disease ($T = 1$) is 0.05; in terms of probability theory we write this as $P(T = 1|D = 0) = 0.05$.

We want to calculate the probability that the person has the disease *given* that the test shows its presence. In other words, we want $P(D = 1|T = 1)$. We calculate

$$P(D = 1|T = 1) = \frac{P(D = 1 \cap T = 1)}{P(T = 1)} =$$

$$\frac{P(T = 1|D = 1)P(D = 1)}{P(T = 1|D = 1)P(D = 1) + P(T = 1|D = 0)P(D = 0)}$$

$$= \frac{0.95 \times 0.001}{0.95 \times 0.001 + 0.05 \times 0.999} \simeq 0.02$$

In other words, even though the test is 95% accurate, there is only a 2% chance of a person having the disease when the test says that the disease is present!

To understand what is happening, let us replace 95% by a parameter $p$. So

$$P(T = 1|D = 1) = p \text{ and } P(T = 0|D = 0) = p$$

It follows that

$$P(T = 0|D = 1) = (1 - p) \text{ and } P(T = 1|D = 0) = (1 - p)$$

Repeating the above calculation we have

$$P(D = 1|T = 1) = \frac{p \cdot 0.001}{0.001 + (1 - p) \cdot 0.999} = \frac{p}{1 + (1 - p) \cdot 999}$$

It follows easily that if $p > 0.999$ then this is $> 1/2$. In other words, in order for one to have more than even chance of the test indicating the presence of the disease, the effectivity of the test should be at least 99.9%!

In other words, the efficiency of the test should surpass the percentage of the population that *does not* have the disease for the test to be a worthwhile indicator of the presence of the disease. Otherwise, statistical errors in the test would generate “false positives”.

This problem of “false positives” is a significant one in many contexts since it is difficult to decide *a priori* how large a proportion of the population has a certain characteristic. The rarer the characteristic, the more stringent the requirements on correct-ness of the test. At least *that* aspect is intuitively more obvious than that the other aspects of the above calculations!