

Bayesian Estimation – An Informal Introduction

Example:

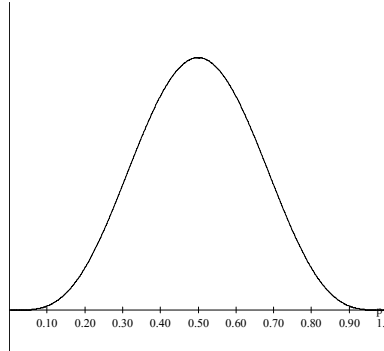
I take a coin out of my pocket and I want to estimate the probability of heads when it is tossed. I am only able to toss it 10 times. When I do that, I get seven heads. I ask three statisticians to help me decide on an estimator of p , the probability of heads for that coin.

Case 1. Sue, a frequentist statistician, used $\hat{p} = \frac{X}{10} = 0.7$.

Case 2. Jose, who doesn't feel comfortable with this estimator, says that he already has an idea that p is close to 0.5, but he also wants to use the data to help estimate it. How can he blend his prior ideas and the data to get an estimate? Answer: Bayesian statistics.

Jose makes a sketch of his prior belief about p . He thinks it is very unlikely that p is 0 or 1, and quite likely that it is somewhere pretty close to 0.5. He graphs his belief.

Jose's drawing:



Then he notices that the graph corresponds to a particular probability distribution, which is Beta(5,5). So this is called his prior distribution for p .

Recall that, for a beta distribution with parameters a and b , we have these formulas for the mean and variance.

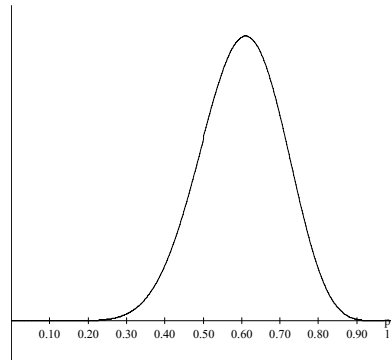
$$\text{Mean} = \frac{a}{a+b} \quad \text{Variance} = \frac{ab}{(a+b)^2(a+b+1)}$$

Thus the mean is $\frac{5}{5+5} = 0.5$ and the variance is $\frac{5 \cdot 5}{10^2 \cdot 11} = 0.022727$, so the standard deviation is 0.15.

Then, by some magic! (i.e. Bayes Theorem), he combines the data and his prior distribution and gets that the distribution of p , given the data, is a Beta(12,8). This is called the posterior distribution of p . Using those same beta distribution formulas, this mean now is 0.6 and the variance is 0.01143, so the standard deviation is 0.107.

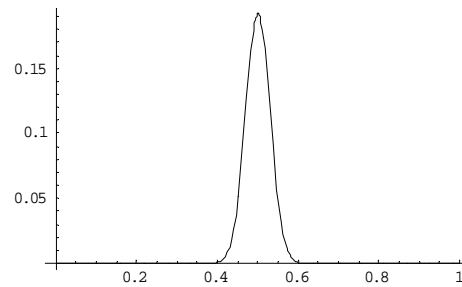
Jose's posterior distribution

Beta(12,8)



Case 3: Vicki, who is very sure that coins are unbiased, has a prior distribution like Jose's, but much narrower. There's a much higher probability on values very close to 0.5. She graphs her belief.

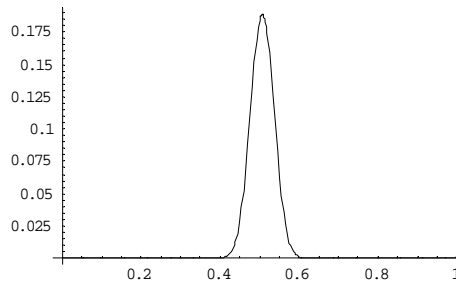
Vicki's prior distribution:



She notices that this corresponds to a particular probability distribution, which is Beta(138,138), so that is her prior distribution of p . So the mean is 0.5, the variance is 0.0009, and the standard deviation is 0.03. Notice that her standard deviation is much smaller than Jose's.

Now, she also uses Bayes Theorem to combine the data and her prior distribution and finds that her posterior distribution of p is a Beta(145,141). So her posterior mean is 0.507, variance is 0.0008709, and standard deviation is 0.0295.

Vicki's posterior distribution. Beta(145, 141)



Jose and Vicki are both doing Bayesian estimation. Both of them decide to use the mean of the posterior distribution of the parameter as their estimator.

Summary:

Sue's estimate of the probability of heads: 0.700

Jose's estimate of the probability of heads: 0.600

Vicki's estimate of the probability of heads: 0.507

Now, Jennifer offers you a bet. You pick one of these values. She chooses one of the other two. An impartial person tosses the coin 1000 times, and get a sample proportion of heads. If that sample proportion is closer to Jennifer's value, you pay her \$25. If it is closer to yours, she pays you \$25. Which value would you choose?

Overview:

Jose and Vicki are both doing Bayesian estimation. But they get different answers. This is one of the reasons that frequentist statisticians criticize Bayesian methods. It doesn't seem objective enough. Different people can get different estimates based on the same data, they say. And of course, Jose and Vicki did get different estimators. But, were they really using the same data? Well, not if we consider data in the broad sense. If their prior beliefs are considered data, then these are not based on the same data. And, if their prior beliefs are not considered data, then we are back to frequentist statistics. Would any of you be willing to take up Jennifer's bet and choose Sue's frequentist statistics estimator of 0.7? (If so, I'll be willing to bet with you.)

What are some of the difficulties in the Bayesian approach?

1. Quantifying prior beliefs into probability distributions is not simple. First, we haven't all thought much about our prior beliefs about most things, and, even if we have some beliefs, those aren't usually condensed into a probability distribution on a parameter.
2. We might not agree with colleagues on the prior distribution.
3. Even if we can find a formula for the distribution describing our prior beliefs about the parameter, actually doing the probability calculations to find the posterior distribution using Bayes Theorem may be more complex than we can do in closed form. Until people had powerful computing easily available, this was a major obstacle to using Bayesian analysis.

The Bayesian statisticians have some answers for most of this, which will become clearer as we develop the theory.

Mathematical Theory:

How do we do the mathematics to get the posterior distribution?

Use Bayes Theorem, which is a theorem to "turn around" conditional probabilities.

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

From our data, we get $f(\text{data}|p)$, which can also be denoted as $L(p)$. We want $h(p|\text{data})$, which is the posterior distribution of p . We also need $g(p)$, which is our prior distribution of p .

For theoretical purposes, (but we often manage to avoid using it), we also need the distribution of the data, not conditioned on p . That means $k(\text{data}) = \int_{\text{domain}} f(\text{data}|p)dp$, or, if the distribution of p is

discrete, $k(data) = \sum_p f(data|p)$. In these, we are integrating or summing over all possible values of p .

So, using Bayes Theorem, we get

$$h(p|data) = \frac{f(data|p) \cdot g(p)}{k(data)}$$

Now, since $k(data)$ is not a function of p , and since we are going to fix our data, then the denominator is a constant, as we calculate $h(p|data)$. Since the only difference that a constant multiple makes in specifying a probability distribution is to make the density function integrate (or sum) to 1 over the entire domain, then we can know the form of the distribution even before we bother with computing the denominator. And, once we know the form of the distribution, we could figure out the constant multiple just by determining what it would take to have the integral over the entire domain be 1. So, in practice, people use this form of Bayes Theorem, where the symbol \propto means “proportional to”, i.e. “is a constant multiple of” and we call the right hand side of this the “unnormalized posterior density”.

$$h(p|data) \propto f(data|p) \cdot g(p)$$

In fact, when we do problems using this, we often neglect the constant multiples (considering p as the variable and the rest as constants) on both $f(data|p)$ and $g(p)$, since if we are going to gloss over one constant multiple, then there is little point in keeping track of the rest until we start trying to really get the constant multiples figured out. And then, when we do need to find the constant multiple, if we recognize the kernel of the distribution (the factors which include the variable) as one of our standard distributions, we can use that to determine the constant multiple.

When you read about Bayesian statistics, you’ll see the statement that Bayesians use the data only through the likelihood function. Recall that the likelihood function is just the same as the joint density function of the data, given p , reinterpreted as a function of p . $L(p) = f(data|p)$

$$h(p|data) \propto f(data|p) \cdot g(p) = L(p) \cdot g(p)$$

Back to the example:

Here, $L(p) = f(data|p) \propto \prod_{i=1}^{10} p^{x_i} (1-p)^{(1-x_i)} = p^7 (1-p)^3$

which is from a binomial distribution, in “unnormalized form”, i.e. without the constant multiple, considering it as a function of p and defined on $p \in [0,1]$.

Also, for Jose, his prior distribution is $g(p) = \frac{\Gamma(10)}{\Gamma(5) \cdot \Gamma(5)} p^{5-1} (1-p)^{5-1} \propto p^{5-1} (1-p)^{5-1}$

so $h(p|data) = f(data|p) \cdot g(p) \propto p^{7+5-1} (1-p)^{3+5-1}$

Since the domain of h is all real numbers between 0 and 1, it is a continuous distribution, so it fits the form of a Beta distribution, with parameters 12 and 8. Thus, Jose has a posterior distribution on p which is Beta(12,8), as we said before, and is used to calculate his estimator, which is the mean of the posterior distribution on p .

We calculate Vicki’s posterior distribution on p in exactly the same manner.

We can use these posterior distributions on p to make probability statements about p as well. Typically Bayesians would compute a HPD (highest posterior density) interval with the required probability.

Back to the Theory:

Since we are using this theory to find a posterior distribution on the parameter, we can use that distribution to find probability intervals for the parameter. In a frequentist statistics course, we learn that parameters aren't random variables, however, so when we find intervals for parameters, we called them confidence intervals to distinguish them from probability intervals. The major conceptual difference between Bayesian statistics and frequentist statistics is that, in Bayesian statistics, we consider the parameters to be random, so one consequence of that is that we can write probability intervals for parameters.

Free the Parameters!

-- Bayesian Liberation Front

Back to the Overview: (How they chose their prior distributions)

No doubt it will have occurred to you that it was very convenient that both Jose's and Vicki's prior beliefs about the parameter fit into a beta distribution so nicely. It was particularly convenient since the data, given the parameter, fit a binomial distribution and so the product of those had such a simple form. Surely, you say, that doesn't always happen in real life! Probably not.

But, in real life, most people don't have fully developed prior beliefs about most of the interesting questions. In fact, you can usually elicit from a person what the average (mean) of their prior belief is, and, after a fair amount of coaxing, you might be able to get their standard deviation. There are convenient forms for prior distributions in a number of cases, based on the distribution of the data. These are called "conjugate priors". If the data is binomial, the conjugate prior is a beta distribution. If the data is normal, the conjugate prior is a normal distribution. So, in practice, you often elicit the mean and standard deviation of a person's prior belief, notice what form the distribution of the data has, and decide to use a conjugate prior. You then use the mean and standard deviation to solve for the parameters in your conjugate prior.

That's how Jose and Vicki did the problem here. The data was binomial, so they needed a beta prior distribution. Recall that, for a beta distribution with parameters a and b , we have these formulas for the mean and variance.

$$\text{Mean} = \frac{a}{a+b} \quad \text{Variance} = \frac{ab}{(a+b)^2(a+b+1)}$$

Each of them wanted a mean of 0.5, so they needed the two parameters to be the same. Then, the smaller that they wanted the standard deviation, the larger the parameters had to be. So each decided on a standard deviation that seemed to reflect his(her) beliefs. Jose chose 0.15 and then wrote

$$(0.15)^2 = \frac{a \cdot a}{(2a)^2 \cdot (2a + 1)} \text{ and solved for } a. \text{ He reduced this to a linear equation and found } a = 5.0556.$$

Since his standard deviation was only an approximation, it seemed most reasonable to just round this off

to use integer parameters. Vicki did a similar calculation, starting from her prior belief, which she characterized as mean 0.5 and standard deviation 0.03.

Notice that, when we look at the forms of the data distribution and the prior distribution, they are similar, and so Jose's prior distribution was the equivalent to starting your problem with 4 heads and 4 tails. Vicki's was equivalent to starting with 137 heads and 137 tails.

Exercise:

How would your prior belief compare with Jose's or Vicki's? Decide on a mean and standard deviation that reflect your beliefs. Then compute which beta distribution reflects those beliefs. Then compute your posterior distribution, based on your prior and the data. What is the mean of your posterior distribution? That is your Bayesian estimate for p , using a conjugate prior distribution that best fits your beliefs. Would you rather make a bet on that, or on one of the three estimates given here?

Perhaps you fall at one of the extremes – either that you believe $p = 0.5$ with a standard deviation of 0, or that you have no prior belief about p at all, and so want to go entirely by the data. The Bayesian method can accommodate both of those. In the former case, Bayes Theorem will give you the same posterior distribution as the prior distribution – the probability of zero elsewhere than 0.5 wipes out any other possibility, even from the data. For problems of the latter sort, Bayesians might use a “flat prior” which gives equal probability to all possible values of the parameter.

Back to the Overview:

Everyone is a Bayesian in some situations. I don't think I could find anyone to actually place a bet on the estimate of 0.7 in the example. So does that mean that we are all really Bayesians, and there is no point to learning frequentist statistics?

Of course, the basic issue in this example is that we had a quite small set of additional data and most of us feel that we have quite a lot of prior information about whether a coin is likely to be fair. It is an example “made to order” to make frequentist analysis look bad. In other situations where prior information is not agreed upon, then Bayesian analysis looks less attractive.

How do Bayesians deal with these issues?

The main criticism of Bayesian analysis is that it is subjective because the conclusion depends on the prior distribution and different people will have different prior distributions.

One approach to counter the criticism of subjectivity is to use “flat priors” which give equal weight to all possible values of the parameter. This is a very common approach in contemporary Bayesian methods. There are some complications here. One is that, when the parameter space is infinite (e. g. the whole real line) then a flat prior is not actually a distribution because it has an infinite integral rather than integrating to 1. Another complication is that we may be interested in a function of the parameter rather than the parameter itself, and whether you take a flat prior on the parameter or on the function of the parameter makes a difference in the result. Nevertheless, much work has been and is being done in this area.

Some people call this “objective Bayesian analysis.” The general term is “noninformative priors” and some names of these types of prior distributions are default priors, Jeffrey's prior, maximum entropy priors, and reference priors.

Many Bayesian statisticians consider “**subjective Bayesian analysis**” to be the heart of the Bayesian method and work in areas where it is feasible to specify prior distributions reasonably well. Among

other things, they have developed techniques for eliciting such priors. (That is a harder task than it might appear at first glance!)

Another approach, called **“robust Bayesian analysis”** works with **“informative” priors but allows for a family of prior distributions** and a range of possible priors in the analysis, and so avoids some of the problems of trying to be quite specific. It is not unusual for a group of researchers to be able to agree on a reasonable range of possible values for the parameters in the prior distribution and so this removes some of the objection to the subjectivity.

“Empirical Bayes” is a technique that allows one to use data from somewhat similar situations to “borrow strength” and produce better estimators in the individual situations.

Computation:

Perhaps it has occurred to you that when we use a prior distribution other than a beta distribution on this problem, we can certainly write an expression for the posterior density as we did here, but actually doing any calculations with this density function, such as finding the mean, variance, or any probability calculations could be very difficult in general if the posterior density wasn't one of our usual distributions.

There are many functions which can't be integrated in closed form and so numerical techniques are needed. In high-dimensional problems, numerical integration is an even more complicated problem than in low-dimensional problems. Markov Chain Monte Carlo (MCMC) techniques were introduced in the late 1980's and use a computationally-intensive simulation method to replace exact integrals. The aim of MCMC is to use simulation to draw a large sample from the full posterior distribution and, using that sample, estimate the needed values from the posterior distribution. Some of those are means, medians, probability intervals, etc. There are several approaches to MCMC.

What does the statistics community say?

When I was in graduate school in the 1970's and 1980's, we discussed the “Bayesian-frequentist” controversy. People took sides and made impassioned arguments. Most graduate schools had no courses involving Bayesian statistics and few courses mentioned it. My advisor was working in empirical Bayes areas and I learned various interesting things, such that insurance companies had been using some of these estimators since the early 1900's without any formal proof that they were good estimators. Why do you suppose it was the insurance companies in particular?

In the 1990's to now, most of the graduate courses don't mention Bayesian statistics. It is becoming more common for graduate schools to have one or two courses in Bayesian methods. From the leading theoretical statisticians, I hear “one should have many techniques in your ‘toolkit’ and use whatever ones are appropriate for the particular problem.” One such person is Bradley Efron, the current President of the American Statistical Association. I found a couple of interesting articles of his on the web and those are listed at the end of this handout.

Summary:

In frequentist statistics, the parameters are fixed, although maybe unknown. The data are random.

In frequentist statistical analysis, we find confidence intervals for the parameters. In Bayesian analysis, the parameters are random and the data are fixed. We find probability intervals for the parameters.

In hypothesis testing, the frequentist statistician computes a p value, which is $p\text{-value} = \Pr(\text{data} | H_0)$. But the Bayesian statistician computes $\Pr(H_0 | \text{data})$.

Additional Resources:

These are generally organized, within categories, by the order in which I think you might want to pursue them, with the easiest first!

Overview of Bayesian statistics:

Stangl, Dalene, *An Introduction to Bayesian Methods for Teachers in Secondary Education*. 1999.

<http://www.isds.duke.edu/~dalene/talks/ncssm/sld001.htm>

Yudkowsky, E. *An Intuitive Explanation of Bayesian Reasoning*.

<http://yudkowsky.net/bayes/bayes.html>

Lee, P. M. *Bayesian Statistics, an Introduction*. 1989. Edward Arnold. Distributed in US by Oxford U Press. (second ed. 1997)

Berger, J. O. “Bayesian Analysis: A Look at Today and Thoughts of Tomorrow.” p. 275-290 *Statistics in the 21st Century*, 2002. Chapman and Hall.

Gelman, A., Carlin, J. b., Stern, H. S., and Rubin, D. B. *Bayesian Data Analysis, Second Edition*. 2003. Chapman and Hall.

Use a Bayesian application yourself!

Bayesian Spam Filtering. http://email.about.com/cs/bayesianfilters/a/bayesian_filter.htm

Elementary Statistics with Bayesian methods:

Albert, J, Rossman, A. *Workshop Statistics: Discovery with Data. A Bayesian Approach*. 2001 Key College Publishing.

Berry, Donald. *Statistics, A Bayesian Perspective*, 1996 Wadsworth.

Comparison of Bayesian and Frequentist methods:

Efron, Bradley, Interview. <http://www.mhhe.com/business/opsci/bstat/efron.mhtml>

Efron, Bradley, *Bayesian, Frequentists, and Physicists*. <http://www-stat.stanford.edu/~brad/papers/physics.pdf>

Barnett, Vic. *Comparative Statistical Inference*. 1999 Wiley.