Bayesian Estimation and Cricket

By Brendon Brewer

The aim of this article is to give a simple demonstration of the application of Bayes' theorem to a parameter estimation problem that has significant prior information, and also censored data. I also apply one method from orthodox statistics (maximum likelihood), and show that the two approaches give the same sensible results for large datasets, but that the Bayesian method has superior performance for small data sets. Orthodox methods fail to give a sensible uncertainty measurement for small data sets. The article is aimed at the level of someone with some exposure to university mathematics and statistics at a first or second year level.

Introduction

Suppose you are an Australian cricket fan, and you sit down to watch the opening test of the 2005 Ashes series. For months, there have been hundreds of articles in newspapers by sports journalists and cricketers, excitedly speculating on whether England's recent successes will translate into a close Ashes series, or even a long awaited series win for England.

Michael Vaughan, England's captain, wins the toss and elects to bat first. The opening batsmen for England, Marcus Trescothick and Andrew Strauss, make their way to the crease, nervous but pretending not to be as they practice some defensive strokes. Being an avid fan of Australian cricket, you are familiar with Marcus Treschothick from previous series. His career statistics are flashed up on the screen:

Name	**Matches**	**Innings**	**Not outs**	**Runs**	**Average**
Trescothick	61	115	10	4775	45.48

His average (runs per dismissal) of 45.47 would generally be considered very good for an international batsman, confirming your feelings that he is a threat, but not a major one when compared to a handful of other players going around (and also considering some previous mediocre performances against Australia). Next, Andrew Strauss's career statistics appear on the screen:

Name	Matches	Innings	Not outs	Runs	Average
Strauss	14	26	2	1323	55.13

prompting the following predictable comment from one of the commentators: "excellent figures, but it's early days". This is certainly true - it doesn't take much (or any) mathematics to realise that an early career average is not so reliable as an indicator of a player's ability. Yet we readily accept that, after a while, a player's average becomes more useful as a summary of their batting ability. This prompts the following question:

- How early is "early days" really? Presumably the data of the career statistics provides some evidence that Strauss is a very good batsman, but how much evidence? How can we quantify this?

Even if someone has only ever played one innings, this should constitute some (however weak) evidence about the player's ability. For example, on learning that a player has scored a century on debut, we feel very confident in ruling out the possibility that the person is a worse batsman than Glenn McGrath. Yet we are still cautious about overrating the player, and would be very hesitant to read too much into the initial good performance. How confident should we be?

This question can be answered by both Bayesian statistics, and also "orthodox" statistics. The Bayesian solution is mathematically simple, intuitive, and continues to yield sensible answers even as the amount of data decreases to zero ("very early days" indeed!). In addition, it gives a superior estimate of a player's average (the Bayes estimate).

In contrast, the orthodox methods involve lengthier and more difficult calculations, and, by ignoring prior information, have difficulties with small data sets. However, the two approaches become equivalent as the amount of data increases.

The Sampling Distribution

Consider a single innings, and let x be the amount of runs the batsman scores in that innings. The sample space (set of possible values for x) is simply the natural numbers {0,1,2,…}. We would like to measure a player's ability by the quantity μ, which we will call the "true average". If a player has a true average μ, then the probability distribution of x is then[1]

$$p(x \mid \mu I) = \frac{1}{\mu + 1}\left(\frac{\mu}{\mu + 1}\right)^{x} \qquad (1)$$

This is a geometric distribution (see the figure) with expectation value <x> = μ.

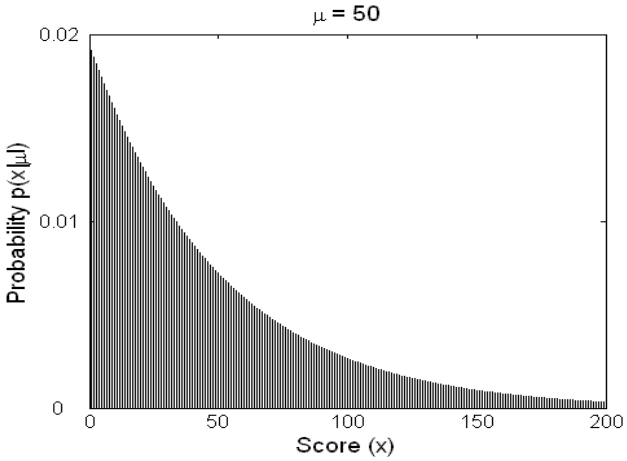


Figure 1 – Distribution of the score in a single innings, for an excellent batsman with a "true average" μ=50.

[1] In orthodox notation, X would be the "random variable" which is the amount of runs, and the probability distribution would be written as P(X=x), so x is a dummy variable. Also, the conditioning on μ would not be present because μ is not a "random variable", but a "fixed unknown constant". In Bayesian statistics, there is no such distinction, and the "given μ" part of the statement is crucial. The shorthand "p" notation (with its ambiguity between the quantity and the dummy variable, and also use of the same letter p to stand for different functions depending on the argument) is actually standard notation in the Bayesian literature. Sorry if you're not used to this, but it's really convenient once you get the hang of it. Just read p() as "the probability distribution of…" and everything should be fine.

There are several different lines of argument which lead to this expression as an appropriate choice for the sampling distribution. If the probability of the player getting out before scoring their next run is independent of their current score, then the geometric distribution will result. While this is not completely accurate (due to “getting one’s eye in” and the “nervous nineties” and other such effects), it is a reasonable approximation, which can be seen by viewing a frequency histogram of some player’s scores. It should probably be pointed out that, to a Bayesian, there is no such thing as a “completely accurate” probability distribution because the point of probability distributions is to represent incomplete knowledge. The purpose of the I (for “information”) in the right hand side of a Bayesian’s probability symbols is to acknowledge the fact that all probabilities are conditional on some background information and assumptions. Another person with different information (such as a corrupt bookmaker!) I_2 may assign a different sampling distribution $p(x|\mu I_2)$.

Another way of justifying the geometric distribution is that it is the probability distribution that has the maximum entropy, for a given expectation value. A frequency interpretation is irrelevant in this case, and the results of any calculations will still be valid, but any estimates obtained would be more conservative than if we had used some other sampling distribution.

I have stressed the assumptions behind the sampling distribution because in my experience, people are happy to take these for granted, but start worrying about philosophy when it comes time to assign prior distributions. If you are worried about philosophy when it comes to priors, you should also have hangups about the sampling distribution. However, probability theory is kind to us: if we don’t know what probability distribution to assign, we just assign something suitably broad (the well known maximum entropy distributions) and we get conservative results. Probability theory is a mathematical model of uncertainty, and provided we don’t put absurdities into the calculation, the results will always be reasonable.

We denote the career performances of a player by two sets of quantities, $\{x_1,x_2,\ldots,x_n\}$, which are their scores in each innings where they were dismissed, and $\{y_1,y_2,\ldots,y_m\}$, their scores in each innings where they were not dismissed (the "not outs"). If each innings is independent (for a given μ), then the probability of obtaining a "career list of scores" $C = \{x_1,x_2,\ldots,x_n, y_1,y_2,\ldots,y_m\}$ is, after some algebra,

$$p(\mathbf{C} \mid \mu I) = \left(\prod_{i=1}^{n} p(x_i \mid \mu) \right) \left(\prod_{j=1}^{m} p(y_i \mid \mu) \right)$$

$$= \left(\prod_{i=1}^{n} \frac{1}{\mu+1} \left(\frac{\mu}{\mu+1} \right)^{x_i} \right) \left(\prod_{j=1}^{m} \left(\frac{\mu}{\mu+1} \right)^{y_i} \right)$$

$$= \mu^{R} (\mu+1)^{-(R+n)} \qquad (2)$$

where $R = \sum_{i=1}^{n} x_i + \sum_{i=1}^{m} y_i$ is the total number of runs in the career, and the probabilities for the "not outs" $p(y_i|\mu)$ were calculated using the fact that, for the geometric distribution (1), the probability of a value greater than or equal to y is given by $(\mu/(\mu+1))^y$.

So, now we have calculated the probability of getting a particular career record C, if only we knew μ. But in practice, we actually know C, and want to use this to estimate μ. Presumably we want to get some measure of the uncertainty in our estimate, as well. This is the aim of the game, now we'll take a look at some of the methods for doing this.

Orthodox Methods - Maximum Likelihood

For a particular value of μ, equation (3) tells us the probability of obtaining a particular career record C, it is $p(C|\mu I)$ and is a function of μ and C. If we consider the same function $p(C|\mu I)$, but fix C at the observed career record, and consider the dependence on μ, then $L(\mu) = p(C|\mu I)$ is referred to as the likelihood function. The maximum likelihood method says that our best estimate of μ should be the value which maximises the likelihood function (3). Intuitively, this means we estimate μ by choosing the value of μ that would

have made the observed data C the “least surprising”. Let’s see what this does. The likelihood function is

$$L(\mu) = \mu^{R}(\mu+1)^{-(R+n)}$$

We want to maximise L, but we may as well maximise log(L):

$$\log L(\mu) = R\log\mu - (R+n)\log(\mu+1)$$

The maximum occurs when d/dμ[log L] = 0. ie when

$$\frac{R}{\mu} - \frac{R+n}{\mu+1} = 0$$

Solving for μ gives our maximum likelihood estimate:

$$\hat{\mu}_{MLE} = \frac{R}{n}$$

This is a lovely result. It says that our best estimate (by the criterion of maximum likelihood) of the player’s “true average” is simply the number of runs divided by the number of dismissals – just what is conventionally given as a “batting average”!

However, a little bit of thought suggests that something’s not quite right here. Consider our example of a player who scores 100 on debut. Do we really believe that the best estimate possible is that the player is as good as Bradman? What if the player scores 100 not out (or even 1 not out)? The best estimate is that they are infinitely good! Of course, we don’t actually think that. The Bayesian result derived later does not have these properties.

So, we have our maximum likelihood estimate of μ, but what is our uncertainty? The way (actually, one of the ways) that orthodox statistics

answers this question is by considering the sampling distribution of the estimator.

Sampling Distribution of the Estimator

In orthodox statistics, the observed career record C is regarded as having been "drawn at random" from the sampling distribution (2). In that case, if we knew the true value of μ, we can calculate the probability distribution of the estimator $\hat{\mu}_{MLE} = R/n$. If this is highly concentrated, then the estimate will rarely be far from the true value. So it seems reasonable to take the width (the standard deviation, say) of the sampling distribution of the estimator as an indication of the uncertainty in the estimate.

Calculating the sampling distribution of $\hat{\mu}_{MLE}$ for the cricket example is complicated by the presence of "not outs", as it is hard to define the sampling distribution for them. However, the problem becomes much simpler if we note that the sampling distribution (3) is the same as for a sample of size n from the geometric distribution.

Hence, we can regard the career C as being a sample of n numbers drawn independently from a geometric distribution with mean μ. We will suppose *n* is *n*ot very small (this assumptio*n* will be relaxed i*n* the *n*ext sectio*n*), so that we ca*n* use the ce*n*tral limit theorem. I*n* that case, $\hat{\mu}_{MLE}$ will be approximately normally distributed with expected value $< \hat{\mu}_{MLE} >= \mu$ and variance which is 1/ n times the variance of the geometric distribution, so $Var(\hat{\mu}_{MLE}) = \frac{\mu(\mu+1)}{n}$.

If we repeated the career many times with the same μ, so the argument goes, our estimates would be scattered around the true value of μ, but with the given variance. But we actually have one "instance" of the "random experiment" and we need to give our uncertainty from this. It would be nice if we could give the square root of the variance as our error estimate – but it depends on the true value of μ, which we don't have! It seems as though the

best we can do is give an “estimated uncertainty” where we plug the estimated value of μ into the formula for the sampling variance of the estimate. Using this, the maximum likelihood method gives us the result that, if n is not too small, our estimate and “1-sigma” uncertainty is

$$\mu = \hat{\mu}_{MLE} \pm \sqrt{\frac{\hat{\mu}_{MLE}(\hat{\mu}_{MLE}+1)}{n}}$$

$$\mu = \frac{R}{n} \pm \frac{1}{\sqrt{n}} \sqrt{\frac{R}{n}\left(1+\frac{R}{n}\right)}$$

Applying this formula to the careers of Trescothick and Strauss, we get the following results:

Trescothick

$$\mu = 45.5 \pm 4.5$$

Strauss

$$\mu = 55.1 \pm 11.4$$

These results seem reasonable at first glance, the estimates are just the conventional batting averages, and the uncertainty for Strauss is higher (it’s early days!). This is just as we anticipated, but the statistics has helped us by telling us how much higher. However, prior knowledge about cricket casts a small amount of doubt on these figures. The figures seem to imply that Andrew Strauss is just as likely to have μ=65 as he is to have μ=45. Yet anyone who knows anything about cricket would think that μ=45 is far more plausible, because μ=65 would make Strauss the 2nd best batsman ever, and (with all due respect to the man!) that is not very plausible.

This sort of situation can occasionally arise in scientific practice. If we ignore common sense and prior knowledge about some phenomenon that we are studying, we can get a result from statistics that conflicts with it. Some people seem to think that statistics is magical and objective, and if it disagrees with

prior knowledge, well, the prior knowledge must be wrong. Bayesian methods actually give us a way of taking prior information into account, if we have it.

Note that I have avoided using "confidence intervals" in this analysis. I highly recommend this practice – confidence intervals are often difficult to calculate, conceptually difficult and answer the wrong question[1].

Bayesian Approach

In Bayesian statistics, instead of inventing methods like maximum likelihood, we just use probability theory. The correct procedure to estimate any unknown quantity from some data (in our cricket example, estimating μ from C) is to calculate the probability distribution of the quantity of interest, given the data, in other words, calculate $p(\mu|CI)$. This probability distribution (called the posterior distribution for μ, given the data C) gives us all the information that we have about μ – if there is a lot of data, then the posterior distribution will be sharply peaked around some value, indicating an estimate of high accuracy. If there is not much data, then the distribution $p(\mu|CI)$ will be wide, indicating that we still don't know all that much about μ. Bayes' theorem gives the relationship between this probability density function, and some others:

$$p(\mu \mid \mathbf{C}I) = p(\mu \mid I)\frac{p(\mathbf{C} \mid \mu I)}{p(\mathbf{C} \mid I)}$$

The p(C|I) term in the denominator does not depend on μ, so we can just regard it as part of the normalising constant for the PDF that we are interested in.

[1] Confidence intervals allow you to calculate an interval which, on repeated sampling from the sampling distribution, would contain the true value of μ 95% (or some other percentage) of the time. It is tempting to think of this as saying that there is a 95% probability that μ is in the specified interval. Frequentists tell us not to think of confidence intervals in that way, and they are correct, because confidence intervals refer to long run performance of the estimator, and not the accuracy for our specific data set. In our example though, there is no such thing as the long run performance, each player only has one career! Why should we even consider an imaginary set of careers that the player could have had, but didn't? How could it possibly be relevant?

$$p(\mu \mid \mathbf{C}I) \propto p(\mu \mid I)p(\mathbf{C} \mid \mu I)$$

The second term on the right hand side is just the likelihood function L(μ). But what is the other term on the right hand side? Evidently, p(μ|I) is the probability distribution that we assign to μ before we know anything about the data. This is referred to as a prior distribution, and it is not there because μ is "random", but because it is unknown – probabilities are just a mathematical description of uncertainty.

We must assign a prior distribution before we can get an answer. Note, however, that if we assign a "noninformative" uniform prior distribution, p(μ|I) = constant, over a wide range of possible μ, then the posterior PDF for μ is proportional to the likelihood function – and its peak will be the maximum likelihood estimate! Hence, a maximum likelihood estimate corresponds to the most probable μ, for a flat ("noninformative") prior distribution. This is quite a general phenomenon; when someone wants to pretend prior probabilities don't exist, and invents methods that don't use them (usually because it makes them feel more "objective"), it is usually just equivalent to a specific choice of prior!

Consider the batsman that has scored 100 on debut. The posterior PDF for μ, with a noninformative prior[1], is shown in Figure 2. This would be appropriate in the case that we knew absolutely nothing about cricket before finding out the score of 100. As expected, it gives an extremely spread out PDF, so that, while the peak is at μ=50 (the Jeffreys prior makes quite a difference to the location of the peak, compared to a flat prior, but the peak is not all that meaningful as the PDF is so spread out anyway), there is substantial

[1] Since μ is a *scale parameter* (a positive number that, if changed, just stretches out the sampling distribution, or squashes it), the appropriate noninformative prior is not a flat prior, but the Jeffreys Prior 1/μ. This can't be normalised, so is called an improper prior. In principle we should set upper and lower limits, calculate the posterior distribution and then take the limit as the upper and lower limits tend to infinity and zero respectively. However this would give us the same result as if we just blindly put 1/μ in as the prior. This is equivalent to assigning a uniform pdf for log(μ), and is appropriate because this prior does not change under a change of scale, ie if we let μ' = constant*μ, our prior for μ' is still the Jeffreys prior.

uncertainty about the value of μ. However, we actually have more prior information than this.

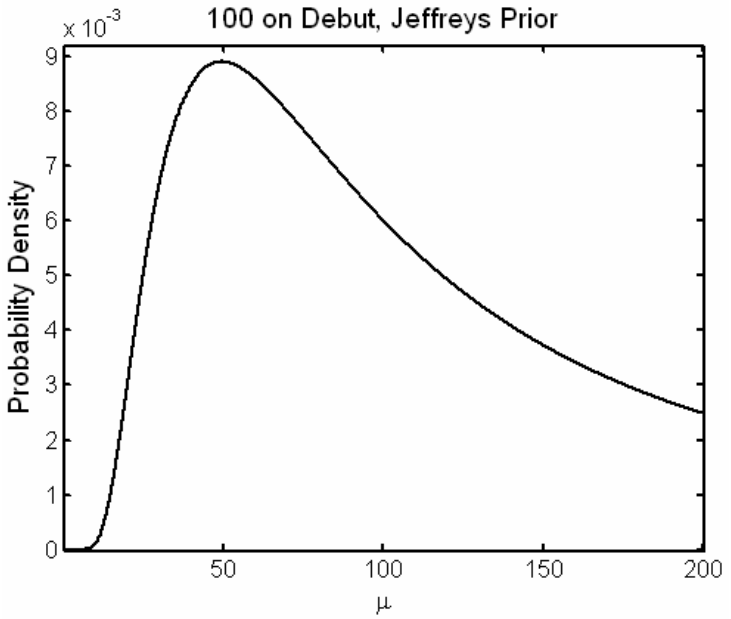


Figure 2 - Posterior PDF of μ, after learning of a score of 100 on debut. This used a noninformative "Jeffreys prior", so would describe the state of knowledge of someone who had no prior idea about cricket, and therefore no reason to be sceptical of really high values of μ.

Normal Prior

We saw in the previous section what our inference would be about a player, given their score in a single innings. This gave a very conservative estimate of μ, but used a noninformative Jeffreys prior distribution for μ. This prior distribution would be appropriate if we actually knew nothing about the significance of a score of 100. For example, for a person from the USA who's never heard of cricket, for all they know, 100 might be a typical score and would not hesitate to say that μ might possibly be really high, more than 100 (but they would tell us they still have a large uncertainty).

We do hesitate to estimate such high values, and that is because we have a different prior distribution. We are more informed about cricket history than our hypothetical American. What do we know about μ before getting any data?

Historically, we know that on average, about 30 runs are scored per wicket in test matches. In fact, I just made that number up, but it's in the right ballpark. This restricts the form of the prior pdf; we know that it should have an expectation value of 30.

We also have some idea about the fre*q*uencies with which various values of μ occur. Players with μ between 10 and 50 are very common, outside these ranges they are relatively rare. So our prior pdf for μ should assign most of its probability within these ranges. We are not completely ruling out higher and lower values than these, but we are initially fairly sceptical of them. Let's just pretend that this gives a constraint that the standard deviation of the prior pdf is 10. The maximum entropy (most conservative) probability distribution for a given mean and standard deviation is a Gaussian or Normal distribution.

This analysis of the prior information has not been very rigorous, indeed it is very hand waving! However, I am confident that a more detailed analysis of the prior information would not lead to anything substantially different to the prior pdf I am about to assign:

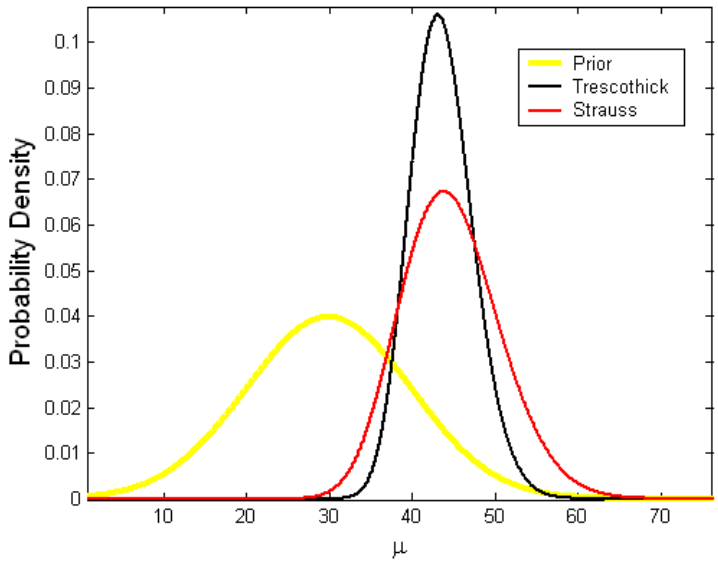
$$p(\mu \mid I) \propto e^{-\frac{1}{2}\left(\frac{\mu-30}{10}\right)^2}$$

This is a normal, or Gaussian, probability density with a mean of 30 and standard deviation of 10. To estimate μ now, we just need to multiply the likelihood function by this, normalise (if necessary), and that's it. We then have the posterior distribution of μ, which fully describes our state of knowledge about μ, so we can give estimates and uncertainties of the estimates.

If we knew in advance that the player we are analysing is a specialist batsman, it might be more appropriate to have the prior centred at around 40, and maybe lightly narrower, and if we knew they were a bowler we might like to centre it around 20 or so. We won't do this here.

Case Studies

Let's have a look at the posterior distribution for μ for the two England opening batsmen:



The yellow curve is the prior distribution, which describes what we knew about μ before getting any data. The black is the posterior PDF for Trescothick, the red is for Strauss. We see immediately that there are really quite large uncertainties, more so for Strauss than Trescothick, and that Strauss might prefer maximum likelihood, because the Bayesian analysis is careful not to overrate him!

The corresponding estimates (posterior mean +- standard deviation) are[1]

Trescothick

$$\mu = 43.6 \pm 3.8$$

Strauss

$$\mu = 44.7 \pm 6.0$$

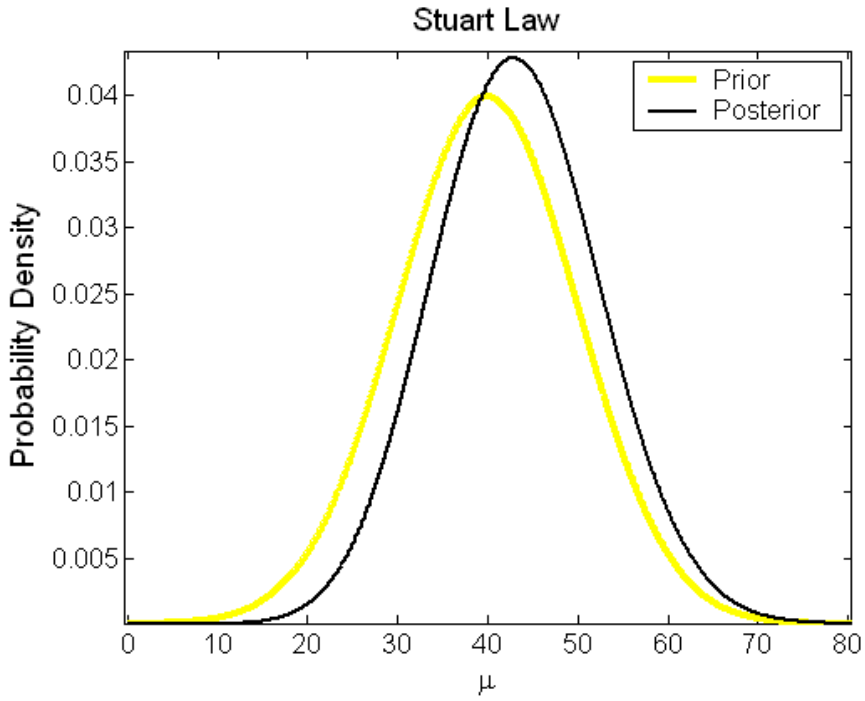
[1] I calculated these numerically. Analytically it would've been difficult. Sometimes people choose priors for analytical convenience, so that they combine with the likelihood to make a posterior PDF with nice easy properties. These are called conjugate priors. We didn't do that.

Compare these to the maximum likelihood results – for Trescothick they are similar. For Strauss, our estimate is significantly lower, and its uncertainty is also a fair bit lower. This is the effect of the prior information – it drags our estimate down to the prior estimate, and the strength of this effect is greatest when we have the lowest amount of data. This makes a lot of sense, and dispenses with the anomaly that we had previously, when we contemplated whether Andrew Strauss was the 2nd best batsman ever.

Incidentally, as n increases, the likelihood function becomes sharper peaked and tends to a narrow Gaussian, so that the prior is essentially constant within the sharply peaked region. I won't prove it here, but that is the regime in which both approaches (Bayesian vs. orthodox) give the same answer.

Stuart Law

Stuart Law is an Australian batsman who only ever played one test innings (against Sri Lanka in Perth), in which he scored 54 not out. Since he has never been dismissed, n=0, and the maximum likelihood estimate (conventional batting average) is infinite, so that analysis breaks down completely. The Bayesian one has no problems. Since we know Stuart Law is a batsman, let's shift the prior up to be centred at 40. The posterior distribution for μ for Stuart Law is plotted below:



I don't know about you, but I think this result is just wonderful. Our state of knowledge about μ is shifted upwards a little, and is also a little bit narrower – but not much – one uncompleted innings provides very little information! This one innings changes our estimate (mean +- sd) from $\mu = 40.0 \pm 10.0$ to $\mu = 43.3 \pm 9.2$. We can't do much more than that. Actually, we probably could – the prior could probably be made a bit narrower than that, substantially so if we devised a way of incorporating knowledge of his first class averages, which are presumably strongly correlated with batting ability in tests.

Conclusion

I have presented a Bayesian method for estimating the batting ability of a cricketer based on their career figures. Hopefully, I have shown that this method is superior to the ones given by non-Bayesian statistics, which are implicitly what is given in career summaries. It does not break down on small datasets (early days), it takes prior information (which was highly cogent in this example) into account, and the calculations are easier. It doesn't leap to wild conclusions based on a few good performances, like the conventional methods do.

Comments

At times I may have sounded a bit harsh in my judgement of non-Bayesian methods. Let me just point out that I don't think they're completely useless or anything, far from it. What I do believe is that, when orthodox methods are appropriate (which is a lot of the time, don't get me wrong), they are equivalent to the Bayesian ones anyway. So I just start from Bayes' theorem every time, and consider it a good thing that the two approaches usually give the same answers. However I find it more interesting (and fun) when they don't – and invariably in these cases, Bayes seems to win.

I am saddened that this approach is not emphasised much, if at all, in statistics courses at the University of Sydney. Bayesian methods are used in only one undergraduate course (as far as I know), and even then, they were referred to as "controversial" and "subjective". In addition, the "random variable" type notation was used, which made expressions look horrid. Worst of all, the lecturer tried to give the prior distribution a frequency interpretation.

References

I decided not to use references in the text, because then it would have sounded too academic. And I was a bit lazy. But here's a couple anyway.

I got the data from the Cricinfo website: http://www.cricinfo.com/

A highly entertaining book on Bayesian statistics.
Probability Theory: The Logic of Science, by E.T. Jaynes