Bayesian Analysis with Stata

John Thompson
University of Leicester

A Stata Press Publication
StataCorp LP
College Station, Texas

Published by Stata Press, 4905 Lakeway Drive, College Station, Texas 77845
Typeset in LaTeX 2ε
Printed in the United States of America
10 9 8 7 6 5 4 3 2 1

ISBN-10: 1-59718-141-2
ISBN-13: 978-1-59718-141-9

Library of Congress Control Number: 2014930482

(Pages omitted)

Contents

(Pages omitted)

Preface

Bayesian analysis has become increasingly popular in recent years and proved useful in many areas, including economics, medicine, science, and the social sciences. Much of this growth has been fueled not by philosophical concerns about the nature of statistical analysis or by a genuine desire to use prior information but by very practical considerations; the fact is that many complex models are much easier to handle within a Bayesian framework.

The chief limiting factor in adopting Bayesian methods is the availability of software. Thanks to programs such as Stata, most frequentist statistical methods are easy to implement, but Bayesian analyses require their own specialist software. Following theoretical developments in the 1980s, a major step forward in Bayesian computing came about with the creation of the BUGS project, which provided flexible, free software for fitting Bayesian models. The original BUGS program was eventually replaced by WinBUGS, which in turn has given way to an open-source version called OpenBUGS. Unfortunately, while these programs will fit a model, they are not particularly user friendly, and they do not offer the many facilities for data handling, exploratory statistics, and graphics that Stata users take for granted. Consequently, even investigators familiar with WinBUGS tend to do most of their work in traditional statistics packages and only switch to WinBUGS to fit the model.

Thompson, Palmer, and Moreno (2006) describe a suite of Stata ado-files, all beginning with the prefix `wb`, that help Stata users integrate WinBUGS into their analyses. The idea is that users should be able to store their data in Stata and analyze them in the usual way. Then when they want to fit a Bayesian model, they should be able to prepare it, send it to WinBUGS, and read back the results without ever leaving Stata. The results provided by WinBUGS or OpenBUGS consist of a very large file of simulations, so other programs are provided to help the user read those results into Stata, check the analysis, and summarize the findings. The commands that operate on the results of the WinBUGS run work equally well with Markov chain Monte Carlo (MCMC) simulations produced by any other software, so they are not restricted to use with output from WinBUGS.

Since 2006, OpenBUGS has become available, the original `wb` ado-files have been extended by adding extra options, and further programs have been added to the collection. New and old versions of the commands can be distinguished between those required to run WinBUGS or OpenBUGS and those used to investigate the resulting MCMC simulations; the former have been collected together as the new commands beginning with the letters `wbs`, and the latter now form the new commands beginning with the letters mcmc. Thus the updated version of wbrun that runs WinBUGS from within Stata is called wbsrun, and the updated version of wbtrace, the program for drawing a trace or history plot, is called mcmctrace.

In this book, I describe the updated commands and introduce other programs for running Bayesian analyses that do not need WinBUGS or OpenBUGS. Chapter 3 presents a set of new programs collected under the names beginning with the letters mhs for running Metropolis–Hastings samplers, and chapter 4 introduces programs beginning with the letters gbs for running Gibbs samplers. These programs can create MCMC simulations and fit Bayesian models without using WinBUGS. For multiparameter models, it is convenient to have a housekeeping program that cycles through the parameters and stores the MCMC updates. This job is performed by the program mcmcrun, which oversees the fitting process by calling any user-specified combination of the mhs and gbs samplers or even the user's own samplers.

Creating MCMC simulations independently of WinBUGS serves two purposes: first, it can be used as a teaching aid to demonstrate what happens when a Bayesian model is fit; second, it offers an alternative practical method by which users can tackle real problems. The only limitation of using Stata on its own is the time taken by simulation-based Bayesian model-fitting algorithms. While the samplers with command names beginning mhs and gbs work well on small- or moderately sized problems, efficient programming is essential when there are more than about 10 parameters in the model or when the dataset is very large. This means either using Stata's matrix language, Mata, or exporting the problem to WinBUGS; these two approaches are broadly equivalent in terms of speed and are introduced in chapters 7 and 8. Users who are only interested in using WinBUGS or OpenBUGS can skip chapters 2, 3, 4, 7, and 12, though in doing so, they will miss out on the explanation of how MCMC samplers work.

The author's blog can be found at http://staffblogs.le.ac.uk/bayeswithstata/. It is dedicated to the discussion of the use of Stata for Bayesian analysis and to describing future developments of the ado-files introduced in this book.

Downloading the user-written commands

The user-written commands discussed in this book can be downloaded from within Stata using the following commands:

```
. net from http://www.stata-press.com/data/bas/
. net install bas
```

To download the do-file to install the user-written Mata code, type

```
. net get bas
```

and then run the mcmclibrarymata.do file.

(Pages omitted)

1 The problem of priors

1.1 Case study 1: An early phase vaccine trial

A researcher tests a new vaccine in an early phase clinical trial by giving it to 10 healthy volunteers, 4 women and 6 men. The researcher is interested in the safety of the vaccine rather than its efficacy, so amongst the many recorded measurements is whether the site of the vaccination becomes tender. In the trial, five of the volunteers report tenderness, one woman and four men. Using these data, the researcher wishes to estimate how many people would experience tenderness if the vaccine were to go forward into an efficacy trial involving 3,000 subjects.

Even this simple problem has more than one plausible solution. If the researcher believes that men and women react differently to the vaccine, then the researcher might argue that if he or she were to recruit 1,500 men and 1,500 women, then $1{,}500 \times 4/6 + 1{,}500 \times 1/4 = 1{,}375$ people who would be expected to experience tenderness. However, if men and women react in the same way, $3{,}000 \times 5/10 = 1{,}500$ people who would be expected to report tenderness.

To decide which estimate to use, the researcher might discuss the problem with experts who have wide experience in vaccine trials or might look back over the records of previous trials of similar vaccines to discover whether men and women reacted differently. Suppose that an expert tells the researcher that there is no biological reason why men and women should react differently and that in the expert's experience of similar vaccines, about 40% of both men and women report tenderness. This seems to tip the balance in favor of the estimate of 1,500, but it also suggests a third estimate based on the expert's judgment: 40% of 3,000 is 1,200. If the researcher is willing to accept the expert's opinion on the difference between men and women, why not also make use of the estimate of the proportion experiencing tenderness? Perhaps the researcher could use the estimate in combination with the data from the early phase trial.

In practice, most statisticians would accept the advice of the expert on whether men and women react differently, but fewer would be willing to use the expert's opinion on the proportion who would report tenderness, even though both types of information affect the final answer. Those statisticians willing to use numerical information that does not come from the study being analyzed are referred to as Bayesians; this book considers the methods of statistical analysis that flow from that willingness and, in particular, how those methods might be implemented in Stata. However, before we get into those technicalities, it is helpful to consider why some people are uneasy about using external numerical information.

The most common objection to the use of an expert's numerical estimate regards subjectivity. The expert's opinion is specific to that individual; had we asked a different expert, we might well have had a different response. When a problem can have different answers depending on which expert we consult, can any answer be relied on? Some would argue that the statistician's job is to summarize the evidence from the current study as objectively as possible. Unfortunately, this argument is undermined by the fact that no analysis is truly objective: subjective judgments are made throughout the planning, conduct, and analysis of any study, and those judgments have an influence on the final answer. For instance, our researcher had to decide whether to adjust for gender and even whether it is sensible to extrapolate from healthy volunteers to the people recruited into an efficacy trial.

Other statisticians argue that although the use of subjective estimates is acceptable in principle, it is impossibly hard to do in practice. To use the expert's opinion, we must decide how much weight to give it relative to the weight that we give to the actual data. In the case of the vaccine trial, this weight will depend on how certain the expert is that 40% of future subjects will experience tenderness. Although 40% is the best guess based on past experience, does the expert think that the percentage for the new vaccine will be something between 35% and 45% or perhaps something between 20% and 60%? Such certainty is difficult to quantify, yet it changes the relative weight given to the expert's opinion compared with the actual data and can greatly influence the final answer. This difficulty is magnified in complex problems in which many factors have to be taken into account because the expert will need to give his or her opinion—with uncertainty—on all of those potentially interrelated factors.

The distinction between Bayesian and non-Bayesian statisticians used to be clear, but in recent years, the barriers between the two traditions have largely disappeared with most statisticians adopting a much more open approach. Experience shows that Bayesian analyses are good for some problems and not so good for others, and unless the expert's opinion is held with great certainty, the actual data will carry more weight in the analysis, and the Bayesian and non-Bayesian answers will be similar.

1.2 Bayesian calculations

Before we consider the benefits of the Bayesian approach, it is helpful to set out a typical statistical analysis in more general terms. To analyze a set of data, y, we must first select a model that describes the trend and natural variation in the data. This model is usually specified in terms of a set of parameters, θ, that allow the model to be tuned to fit the observations. So the model describing the probability of observing a particular set of data can be written as $p(y|\theta)$. Once the general form of the model has been selected, we need a method for using the data to guide us in selecting sensible values for the parameters.

In the case of the vaccine trial, 5 out of 10 subjects reported tenderness, so $y = 5$, and a possible model for the data is the binomial, which depends on a single parameter, θ, representing the probability that an individual reports tenderness. Under this model,

$$p(y|\theta) = \begin{pmatrix} 10 \\ y \end{pmatrix} \theta^y (1-\theta)^{10-y}$$

In the Bayesian approach, the expert's beliefs about all possible values of the parameter must be specified as a *prior distribution*, $p(\theta)$; in this way, we capture both what we expect to find and our uncertainty. In specifying this prior, we must not be influenced in any way by the data that we are about to analyze, so it is best to specify the prior before seeing the data. The prior could take any form, but for illustration, suppose that the expert believes that the true value of θ could be anything between 0.3 and 0.5 and that within those limits, any value is just as likely as any other. This prior would be represented by a uniform distribution

$$p(\theta) = \begin{cases} 5 & 0.3 < \theta < 0.5 \\ 0 & \text{otherwise} \end{cases}$$

Combining the data and the prior opinion is now a noncontroversial exercise in probability and relies on Bayes' theorem,

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

Here $p(\theta|y)$ is referred to as the *posterior distribution* of θ because it describes what we believe about θ after combining our prior opinion with the data. $p(y|\theta)$ is the same likelihood that forms the basis of many traditional analyses, and the denominator, $p(y)$, is the probability of the data averaged over all possible values of θ, so $p(y) = \int p(y|\theta)p(\theta)d\theta$.

Figure 1.1 shows our prior and posterior beliefs about θ for the vaccination data. Notice how the observation of 5 out of 10 has shifted our beliefs toward $\theta = 0.5$ and away from $\theta = 0.3$ and also how values totally excluded by the prior will have no posterior support whatever the data might suggest. This plot was created in Stata, but because our interest here is in the principles of Bayesian analysis, the description of the code that was used is postponed until the start of chapter 2.

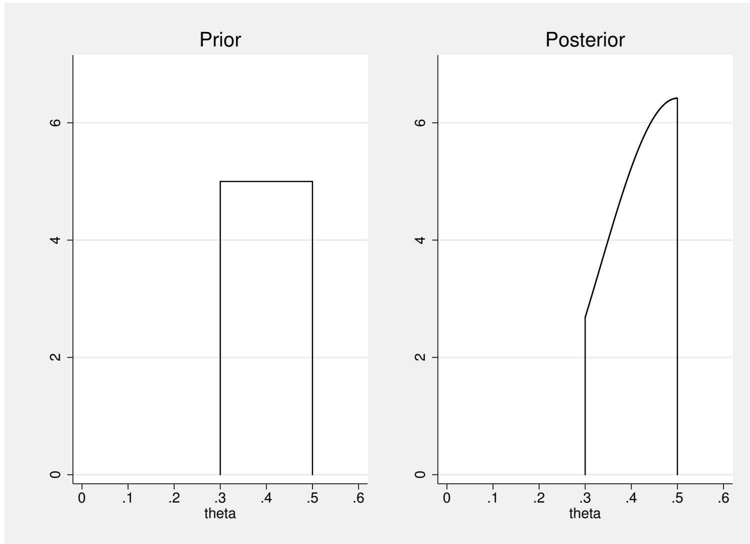

Figure 1.1. Prior and posterior distributions of beliefs about θ

Often the results of an analysis will be used to make predictions about data that will be seen in the future; here we will denote such future data by y^* and its *predictive distribution* by $p(y^*)$ and denote predictions informed by past data, y, by $p(y^*|y)$. Once again the formula for calculating the predictive distribution follows from standard probability theory; it is either

$$p(y^*) = \int p(y^*|\theta)p(\theta)d\theta$$

or, if we have past data available to refine our estimates of the parameters,

$$p(y^*|y) = \int p(y^*|\theta)p(\theta|y)d\theta$$

That is, the predictive distribution based on previous data is derived by averaging the data-generating model over our posterior beliefs about the parameters.

1.3 Benefits of a Bayesian analysis

Once we have calculated the posterior distribution, $p(\theta|y)$, we can make probability statements about θ, for instance, by finding the probability that θ lies within some specified range. Such probabilities have to be interpreted in a similar way to the prior (that is, as subjective expressions of our posterior beliefs about θ), but at least they refer to the actual problem before us. This is in marked contrast to more traditional measures, such as the 95% confidence interval, that rely on repeated sampling for their meaning. Non-Bayesians have to imagine repeated samples of data drawn under the same model so that they can say 95% of such samples would produce confidence intervals that include the true value of θ. Of course, that statement tells us nothing about the properties of our particular confidence interval. The Bayesian analysis is more directly relevant to our study but at the expense of making our statements conditional on both the model and the prior.

Modern simulation methods for calculating posterior distributions have proved so powerful that they have completely turned the tables on more traditional approaches to statistics. Previously, Bayesian statisticians could tackle only relatively simple problems in which the mathematics was tractable or the integrals could be evaluated numerically, while non-Bayesians could tackle many more problems before encountering this type of constraint. With simulation methods, Bayesians can now fit almost any model that they choose, while non-Bayesians still run into computational problems when models contain large numbers of random effects or awkward hierarchical structures. This new freedom has made Bayesian analysis much more attractive and is probably the driving force behind the wider acceptance of the approach.

Now that Bayesian statisticians can choose more or less any model that they think describes the data, the quality of the analyses that they produce ought to have improved. Unfortunately, this new freedom has sometimes led to the creation of needlessly complex models that are almost impossible to verify and that have so many potentially interacting key assumptions that a thorough sensitivity analysis is ruled out. Simplicity is still a virtue, even in Bayesian analysis.

1.4 Selecting a good prior

Even experienced researchers analyzing simple problems find it difficult to specify their prior distribution. Experts unfamiliar with Bayesian analysis find it particularly difficult, and considerable effort has gone into devising methods of eliciting good prior information from individuals and groups. Such methods will always have their limitations, so just as statisticians are used to thinking of their models as being useful approximations rather than as true data-generating mechanisms, researchers should also think of their priors as acceptable approximations.

When the results of an analysis are intended for use by a small number of individuals, then those people can agree among themselves about the form of the prior distribution and so obtain an acceptable Bayesian analysis. However, when research is intended for general consumption, as in the case of a scientific study reported in an academic journal, then the grounds for choosing an appropriate prior are no longer clear. If the authors of the article are allowed to use their own priors, then the possibility exists that some unscrupulous authors will select the priors to skew the results and influence the conclusions. To avoid this, either the authors would have to exert a great deal of effort to explain whether the result is sensitive to the choice of a prior or the authors would have to use some standard or reference prior.

The use of a prior that is not chosen to reflect the views of a particular person or group of people is sometimes referred to as objective Bayesian analysis, although this name is not really appropriate. Usually the standard prior is chosen to have a small influence on the results so that the data dominate. So far the search for a general definition of a minimally informative prior has not been successful; it is common to choose vague priors that are relatively flat over a wide range of values in the hope that they will be noninformative in the sense of having little impact on the final answer.

Early Bayesian analyses were dominated by completely flat priors that extended over the full range of the parameters. When the parameter has an infinite range, a flat prior will not integrate to one and is described as improper. An improper distribution cannot really be said to represent probabilities and at best is an approximation to a uniform distribution over a very large but finite range. Even though improper priors do sometimes lead to proper posterior distributions, their use has fallen out of favor, and they are generally avoided in modern Bayesian analyses.

Before simulation methods were available for calculations in a Bayesian analysis, it was common to choose priors that simplified the mathematics and made the integrals tractable. It became popular to seek out priors that lead to posteriors from the same family. Thus, if we are using a binomial model with parameter θ and choose a beta distribution for the prior, say, beta(a,b), then the posterior distribution of θ will also be a beta distribution but with parameters beta($a+y$,$b+n-y$). Such priors are said to be conjugate. Unfortunately, conjugate priors only exist for a few simple problems, and it is questionable whether the form of the prior should be dictated by mathematical convenience, even if we accept the idea that the prior is only an approximation to our beliefs.

The search for proper noninformative priors is made more complex by the fact that the properties of a prior distribution may be altered by a change of scale. For example, take the problem of setting a prior for a standard deviation. One vague distribution that might be used is a uniform between two limits, say, uniform(0,10). Under this prior, all standard deviations between 0 and 10 are thought equally likely, so the prior probability that the standard deviation is over 6 is 0.4. Now suppose instead that we had decided to work in terms of the variance. We might have opted for a flat distribution over an equivalent range, uniform(0,100); in this case, the probability that the variance is over 36 (standard deviation over 6) is 0.64. A flat distribution for the standard deviation does not give the same results as a flat distribution for the variance.

Even more disturbing than the problem of changing scale is the sensitivity that results can have seemingly small changes in the prior. Once again, take the problem of the prior for a variance. For mathematical convenience, it is common for Bayesians to work on the scale of the precision, which is equal to one over the variance. Because the precision must be positive, a popular vague prior is a gamma distribution with mean one and a large variance. Such a distribution will be relatively flat over a large range of positive values. Unfortunately, as Gelman (2006) has demonstrated, the posterior estimate of the precision can be sensitive to the arbitrary choice of the variance of the gamma prior; thus the gamma prior with large variance has declined in popularity.

When researchers select priors intended to be noninformative, it is tempting to play it safe and extend them over huge ranges, well beyond any value that might realistically occur. However, such impossibly wide priors can create problems when simulation methods are used to investigate the posterior distribution. Very wide priors might cause the simulation to investigate extreme parameter combinations, which could slow down the algorithm or even cause numeric overflow when the posterior is evaluated. Realistically vague priors are generally preferable to impossibly wide priors and have the advantage of making researchers think about their models. Seeking realistic priors will avoid the situation in which, for example, a normal prior with mean 0 and variance 10,000 is routinely used for all parameters that represent means, including perhaps one that represents the average human femur length measured in millimeters.

Three general statements can be made about the selection of noninformative priors: first, automatic choices are potentially dangerous, so it is much better to put some thought into the actual range of plausible values of the parameter; second, vague priors that completely exclude theoretically possible parameter values should be used only after careful thought; and third, whatever choice is made needs to be assessed in a sensitivity analysis to see whether small changes to the prior have a noticeable effect on the results.

1.5 Starting points

Much of the early literature criticizing Bayesian statistics has become outdated because of the enormous progress that has been made in the subject over the last decade or so, and a starting point is the recent article by Gelman (2008). Although the author accepts the Bayesian approach himself, he still manages to convey many of the objections that people have to the use of Bayesian methods, including doubts about subjectivity and the excessive complexity of some Bayesian analyses. The article is followed by a series of equally informative discussions and a rejoinder.

The distinction between objective and subjective Bayesian analysis is well summarized by Berger (2006), who puts the case for objective methods, and Goldstein (2006), who puts the case for subjective methods. These articles are also accompanied by discussion articles and rejoinders. An interview with Jose Bernardo conducted by Irony and Singpurwalla (1997) provides an interesting overview of many different types of priors seen from a subjective standpoint. Natarajan and McCulloch (1998) and Gelman (2006) consider practical problems that can arise when supposedly noninformative priors are used.

Eliciting subjective priors from an expert is a complex problem with its own extensive literature. Good starting places for reading on this topic are the Elicitation of Experts' Probabilities website, http://www.shef.ac.uk/chebs/research/themes/elicitation/, and a systematic review by Johnson et al. (2010).

1.6 Exercises

1. An expert tells you that the prior probability of an event occurring should be described as uniform between 0.3 and 0.4. You conduct an experiment and find that the event occurs only once in 20 occasions.

   a. Sketch the shapes of the prior and posterior distributions.
   b. In light of the data, was the prior reasonable?
   c. Would you be willing to use predictions based on this posterior?
   d. Would it be reasonable to ask the expert to reassess his or her prior?

2. A model is devised for predicting movements in the exchange rate between the euro and the dollar. The intention is to fit the model to data from the last 12 months and then use the model to make predictions.

   a. Is it possible for an expert to give informative priors for the parameters in the model that are not already influenced by the data you hope to analyze?
   b. How might you obtain priors for this problem?

3. An expert gives his or her prior distribution for the average daily sulfur dioxide level at a particular point in the center of town. You measure the sulfur dioxide level daily for a week and use a Bayesian analysis to obtain the posterior distribution. When you show the data and the posterior distribution to the expert, he or she refuses to accept your analysis and draws a completely different distribution, saying that this is what he or she now believes after seeing the data.

   a. Why might your calculation and the expert's sketch disagree?
   b. Which distribution better describes the expert's updated beliefs, your posterior or his sketch?
   c. If you reject the posterior, can you ever again be sure that a posterior distribution describes an updated opinion?
   d. If you reject the sketch, what confidence can you have in the prior that the expert drew?

4. A researcher wishes to estimate the proportions of different types of terrain using aerial photographs of sample areas. He or she agrees to classify the terrain as forest, grassland, or other and, for a particular region, gives prior estimates of the percentages as 60%, 30%, and 10%. When you ask about his or her uncertainty, the researcher says that all prior estimates are accurate to within ±10%; that is, the researcher believes the percentages to lie within the ranges (50%,70%), (20%,40%), and (0%,20%). Further, he or she believes that within those ranges, all percentages are equally likely.

   a. Why do these figures not describe a prior distribution?
   b. How would you get a nonstatistician to describe his or her prior for this problem?