A One-Hour Training Seminar on Bayesian Statistics for Nursing Graduate Students

Byron J Gajewski, University of Kansas Medical Center
Stephen D Simon

Available at: https://works.bepress.com/byron_gajewski/16/
A One-Hour Training Seminar on Bayesian Statistics for Nursing Graduate Students

Byron J. GAJEWSKI and Stephen D. SIMON

In this article we discuss one-way ANOVA with the goal of introducing Bayesian statistics to nursing graduate students in a one-hour seminar. We use ANOVA because it is arguably one of the most widely used statistical models in practice and because its Bayesian treatment has interesting ties to the classical ANOVA, creating a smooth transition for students in a frequentist-based course to be introduced to Bayesian ideas. We discuss the limitations of this seminar in the context of our assumptions of what the students can or cannot do. An example from nursing research is used throughout the seminar.

KEY WORDS: ANOVA; Hierarchical model; Posterior; Prior; Shrinkage; Teaching.

1. INTRODUCTION

Is it possible to introduce Bayesian ideas in a graduate service course? Yes, but the timing, how much, and the context are important issues to be considered. We propose that after students have been introduced to analysis of variance (ANOVA), a one-hour lecture that introduces Bayesian analysis based on ANOVA would be practical and effective. ANOVA provides an interesting framework for our purpose since its Bayesian treatment has interesting ties to the classical ANOVA and arguably it is one of the most widely used statistical methods.

We have specific experience with the above proposal with first- and second-year Ph.D. students majoring in a variety of health sciences. The students in our one-hour seminar have taken, or are enrolled in, a course with ANOVA content. Therefore, they would be comfortable with writing and interpreting the cell means and effects and can calculate the classic one-way ANOVA model by hand. We assume that they have had a first course in biostatistics at the level of Rosner (2000) and that they can read and write statistical models at the level of Ott and Longnecker (2000).

After studying one-way ANOVA and having seen several applications of traditional hypothesis testing, students should be comfortable with branching out to new approaches for data analysis. We believe that this should be done in small steps.

A one-hour introduction to Bayesian data analysis is easier to incorporate than a wholesale revision of a curriculum.

To understand what type of Bayesian data analysis we can teach effectively, we must understand the background and motivation of the students. We teach service courses to nonstatisticians and we may have a hard time just getting them to like the concept of probability. These students may have the idea that probability is completely unnecessary, because they may take similar service courses from nonstatisticians and hardly ever hear the word “probability.” The stretch to the likelihood might be even more difficult for them, and the subsequent stretch to Bayesian ideas might be torturous. To teach Bayesian analysis effectively, we need to recognize what we can do with this audience and what we cannot do. Table 1 lists a set of effective and ineffective approaches for our one-hour seminar.

While computing is outside the scope of this article, code is available from the authors. Teachers might write their own code with help from Gelman, Carlin, Stern, and Rubin (2004, p. 81 and p. 134). While this code would not be shared with the student, instructors who have limited experience with Bayesian data analysis might gain extra confidence and comfort with this topic if they can run some simple examples on their own.

2. THE ONE-HOUR SEMINAR

In this section, we provide our version of the one-hour seminar. Other teachers may want to directly use our code to build on their approach. Relevant references for this one-hour seminar include Albert and Rossman (2001); Berry (1995); Albert (1997); Berry (1997); Gonen, Johnson, Lu, and Westfall (2005); Gelman, Carlin, Stern, and Rubin (2004); Gelman (2005); Neath and Cavanaugh (2006); and Lecoutre (2006).

Before this seminar we require a reading assignment. This is a good way to get the class “warmed up.” In our pilot studies we have asked students to read Albert (1995). Although Albert’s paper showcases a sports example, our pilot studies indicate that students enjoy this paper, particularly after re-emphasizing Albert’s points through our one-hour seminar. Other possible examples are in business (Kim and Nelson 1999); politics (Airoldi, Anderson, Fienberg, and Skinner 2006); genetics (Beaumont and Rannala 2004); and astronomy (Molina, Nunez, Cortijo, and Mateos 2001).

We now begin the lecture. The language is intended to be directed at the student for the rest of Section 2. In Section 3 we give discussion and conclusions intended for teachers.

2.1 Lecture: Introduction

What is Bayesian analysis? Bayesian methods allow us to formally combine prior information and observed information...
(data). The prior information can be obtained from a previous study or expert opinion. In the next hour we will explore this basic concept by introducing a dataset (bereavement example) which we examine using Bayesian analysis and an ANOVA model. The goals of today’s talk are (1) to provide an overview of Bayesian ideas, (2) to introduce a dataset that we will use in a Bayesian analysis (bereavement example), (3) to present a model without center effects that produces a compromise estimate between the prior beliefs and the observed data, and (4) to present a hierarchical model with center effects that produces a compromise estimate between individual center means and the overall mean. We also feel strongly about your feedback, so we will ask you to evaluate the one-hour seminar.

### 2.2 Lecture: Overview of Bayesian Ideas

First of all, the basic formula that comes out of Bayes’ Theorem is

\[
\text{Total information} = \text{Historical information} + \text{Data}. \quad (1)
\]

Much of the literature calls the historical information the “prior distribution,” the data the “likelihood,” and the total information the “posterior distribution.” The main point of Equation (1) is that the analysis (posterior) relies on a combination of the analyst’s knowledge before the experiment and what is learned from the collected data (or empirical information).

This equation is compared and contrasted to a classic analysis in the following formula

\[
\text{Total information} = \text{Data}. \quad (2)
\]

The analysis does not rely on the prior distribution. Only the data is used for analysis. Notice that if (historical information) = 0, then Equation (1) is the same as Equation (2). Therefore, conceptually, in the absence of prior information, the Bayesian has similar results as the classic analysis.

We will consider ANOVA ideas for the rest of this lecture by first introducing the data for the bereavement example. We will show how to incorporate prior beliefs into an ANOVA analysis.

### 2.3 Lecture: Bereavement Example

We consider an example in nursing research to illustrate Bayesian statistics. Clinicians are challenged to care for patients who may not be affected by a terminal illness but are dying after many years of a chronic disease. Patients should have a peaceful death, free from undue suffering and consistent with their wishes. Palliative care addresses these challenges. In nursing homes, it is vital that staff be skilled in applying palliative care principles. Therefore, nursing researchers desire to measure the state of palliative care in nursing homes.

A study was proposed to collect data from staff members in 24 Kansas nursing homes. The goal of this study was to measure perceived bereavement in nursing homes. A sample question was “When a resident dies, someone who works here will follow up with his or her friends.” Each of the nine questions in the perceived bereavement subscale is on a Likert-type scale (1–5): (1) never, (2) hardly ever, (3) sometimes, (4) almost always, and (5) always. Each nursing home staff member who completed the survey is assigned a perceived bereavement score by averaging their response to the set of questions, with higher scores corresponding to higher perceived bereavement practice skills for the individual staff members. It was necessary to ask several staff members because nursing home practice varies from staff member to staff member within the same nursing home.

There is one important question. How much does the perceived bereavement score change between different nursing homes? If you or a loved one are planning to live in a nursing home anytime soon, you would want to know which nursing homes are low in perceived bereavement scores and which are high.

We can model the variation across nursing homes (centers) and between nursing homes using the same ANOVA model you have already learned. Let \( y_{ij} \) represent the response of the \( i \)th subject within the \( j \)th nursing home (center). Under this model

\[
y_{ij} = \mu_j + e_{ij},
\]

where \( \mu_j \) is the mean bereavement for staff members in the \( j \)th nursing home and \( e_{ij} \) is an error term with mean zero and variance \( \sigma^2 \).

### 2.4 Lecture: Model Without Center Effects

First, to set ideas, we consider an analysis of this data that assumes the mean bereavement for staff does not vary across nursing homes. This analysis shows how to do Bayesian analysis of a single random sample, like the one-sample analysis you have already learned. We will refer to this as a model without center effects. Under this model

\[
y_{ij} = \mu + e_{ij},
\]

where \( \mu \) is the mean bereavement for all staff regardless of what nursing home they work.

Consider the parameters that are unknown. The first is the average bereavement score \( (\mu) \) and the second is the variance of the error term \( (\sigma^2) \). We need to supply prior information to both of these parameters in order to perform a Bayesian analy-
sis. For this seminar, we will focus only on the prior for \( \mu \). We will consider three prior distributions which we will name the flat prior, the optimistic prior, and the pessimistic prior. These priors are actually just normal distributions and we can describe them by specifying the 95% credible intervals (CRI). The 95% CRI is the interval that contains 95% of the probability for this distribution. It sounds similar to the 95% confidence interval used in classical statistics, but we are not interpreting them as a long-term frequency.

Summarized in Table 2 are the 95% CRI for each of the prior distributions. The scale for the bereavement score runs from 1–5. Essentially, the flat prior’s CRI runs out of the bounds of possible values. It seems like an unrealistic prior but Bayesian analysts have found that an extremely wide interval reflects a situation with maximal uncertainty where there is hardly any prior information. Any prior distribution that extends well beyond the range of the actual data values will produce similar results. In contrast, the optimistic and pessimistic priors offer more information (see Table 2). As their names indicate, the optimist believes that the average bereavement is much closer to the highest score (5) whereas the pessimist believes that the average bereavement is much closer to the lowest score (1). In addition, their priors are informative as the length of their 95% CRI is very narrow (about 0.50).

Now consider a Bayesian analysis: take the experiment’s data (observed staff’s bereavement score) and combine it with the prior information to get total information (posterior distribution). It can be shown that the mean of this posterior distribution is

\[
\hat{\mu} = w \eta + (1 - w) \bar{y},
\]

which is a linear combination of \( \eta \), the prior mean, and \( \bar{y} \), the sample mean from the data. \( w = \left\{1/\text{var}(\eta)\right\}/\left\{n/\sigma^2 + 1/\text{var}(\eta)\right\} \) where var(\( \eta \)) is the prior variance of \( \eta \), and \( n \) is the sample size \( (n = 561) \). \( \sigma^2 \) can be replaced by the sample variance \( (\hat{s}^2 = 0.5818 \text{ in this example}) \). The value \( w \) is close to zero for a flat prior (i.e., a very wide prior 95% CRI). When there is little or no prior data, this estimate places all the weight on the data itself. In this example the 95% CRI for the flat prior goes from −196 to 196 which far exceeds the range of possible responses (1–5) which results in \( w = 0 \).

The overall mean of the observed data is \( \bar{y} = 3.31 \). With a sample size of 561, the value of \( w \) is approximately 0 for the flat prior and 0.06 for both the optimistic and pessimistic priors. The posterior mean is 3.31 (0.06 * 0 + 1 * 3.31) for the flat prior, 3.35 (0.06 * 4 + 0.94 * 3.31) for the optimistic prior, and 3.23 (0.06 * 2 + 0.94 * 3.31) for the pessimistic prior. Additionally, we have computed 95% CRI for the posterior distribution (Table 3). In general, such interval calculations are often done using the computer. The analytic result can be shown, but there is only so much we can do in an hour.

In summary, these intervals tell you that the average level of perceived bereavement score (posterior) is unlikely to be much different than 3.2 or 3.4 even in the presence of strong prior beliefs. There is strong evidence, in particular, that the average score is much higher than 3.0, which is the mid-point of the scale.

### 2.5 Lecture: Model with Center Effects

We now need to look at a more complex model that allows the mean perceived bereavement score to change depending on nursing home. Our classic ANOVA table (not shown) would certainly agree with a more complex model \( (F = 10.69, p < 0.001) \).

The Bayesian representation of a model with center effects is depicted graphically in Figure 1. From the Bayesian perspective, all of the parameters are considered random. In the graphical model, the parameters (circles) represent latent (unobserved) variables that are circled are latent (not observed) and need to be estimated. The variables in rectangles are observed.

![Figure 1. Model with center effects.](image)

Table 2. Prior credible intervals (CRI) for flat and informative cases.

<table>
<thead>
<tr>
<th>Prior</th>
<th>Mean, ( \eta )</th>
<th>( \text{sd}(\eta) )</th>
<th>( \eta \pm 1.96 \text{sd}(\eta) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Flat</td>
<td>0</td>
<td>100</td>
<td>−196–196</td>
</tr>
<tr>
<td>(2) Optimistic</td>
<td>4</td>
<td>1/8</td>
<td>3.76–4.25</td>
</tr>
<tr>
<td>(3) Pessimistic</td>
<td>2</td>
<td>1/8</td>
<td>1.76–2.25</td>
</tr>
</tbody>
</table>

Now consider a Bayesian analysis: take the experiment’s data (observed staff’s bereavement score) and combine it with the prior information to get total information (posterior distribution).

Table 3. Posterior credible intervals (CRI) for various priors.

<table>
<thead>
<tr>
<th>Prior</th>
<th>Mean</th>
<th>95% CRI</th>
<th>( w )</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Flat</td>
<td>3.31</td>
<td>3.25–3.37</td>
<td>0</td>
</tr>
<tr>
<td>(2) Optimistic</td>
<td>3.35</td>
<td>3.29–3.42</td>
<td>0.06</td>
</tr>
<tr>
<td>(3) Pessimistic</td>
<td>3.23</td>
<td>3.17–3.29</td>
<td>0.06</td>
</tr>
</tbody>
</table>
random variables and the data (rectangles) are observed values. The latent variables (parameters) are assumed to produce the observed data. In this model, rather than specifying a particular prior distribution for each of the values μ1, μ2, ..., μ24, we assume instead that the means for each center are sampled from a common prior distribution with mean η. This extra layer produces what is commonly called a hierarchical model.

This model allows you to consider changes to the experiment. Consider the dashed line to the extra mean parameter μ25 in the second model of Figure 1. This is a new group that we have not yet observed. The new group might represent staff members in a nursing home in Kansas outside of the nursing homes we observed. Using the hierarchical model, we can estimate this parameter via its posterior distribution. This would not be possible in a classical frequentist approach.

Now consider elaborating on this model by first defining a reasonable prior for the unknown parameters and then discussing the merits of the hierarchical model through the mean value and the CrI for the means of different nursing homes.

We specify a prior that is vaguely informative, η also has prior mean of 2.5 and 95% CrI of 0.54–4.46.

The mean for the parameter posterior distribution is

\[ \hat{\mu}_j = (1 - w_j) \bar{y} + w_j \bar{y}_j, \]

where \( w_j \) is in between 0 and 1. \( w_j = \{n_j/\sigma^2\}/\{n_j/\sigma^2 + 1/\text{var}(\eta)\} \), \( \text{var}(\eta) \) is the prior variance of η, and \( n_j \) is the sample size for nursing home \( j \). \( \sigma^2 \) can be replaced by the MSE from the ANOVA table (MSE = 0.4161 in this example). One can calculate an approximation to \( \text{var}(\eta) \); which is 0.2208. But this calculation requires more complex software commonly used in the Bayesian research articles because the formulas get more complex. The posterior mean of \( \mu_j \) is a linear combination of \( \bar{y} \), the population mean, and \( \bar{y}_j \), the sample mean from the data for the \( j \)th group. The value of \( w_j \) increases as the sample size for the individual nursing home increases. This is called “shrinkage.” If there is very little data for a particular nursing home, this model will produce an estimate for that nursing home that is close to the overall average. If the nursing home has a large sample size, but the mean for that nursing home is not very different from the overall average, then this model will still produce an average that is close to the overall mean. Only when a nursing home has a large sample size and the data from that nursing home is quite discrepant from the overall mean does this model produce markedly different estimates for that nursing home. This is actually quite sensible because you use something close to the overall mean unless there is strong evidence from the data from a particular nursing home that would make you believe otherwise.

Let’s look into the summary statistics of the nursing homes in Table 4 where the homes were rank ordered by their sample mean.

Notice that the posterior means for the nursing homes at the top of the table (nursing homes with the largest sample means) are still large, but are closer to the overall mean (3.31). The posterior means at the bottom of the table are also closer to the overall mean. The amount of shrinkage is dependent on the sample size. Nursing home 19, with a sample size of 10 shrinks by 0.14 units while nursing home 22, with a sample size of 22, shrinks by only 0.05 units. Nursing home 4, at the bottom of the list, shows a most dramatic shrinkage (0.28 units) because it has one of the smallest sample sizes.

We conclude the seminar by stating that the Bayesian credible intervals presented in this talk have a direct probability interpretation and are different than classical confidence intervals. Recall the interpretation of the classical confidence interval, where we have to state that if we collect 100 samples, we would expect approximately 95 confidence intervals produced from those samples to contain the true value. If you take the trouble to formulate a prior belief and apply a Bayesian analysis, then you are rewarded with a simpler and more direct interpretation of your results.

### Table 4. Nursing homes: ranked by sample mean.

<table>
<thead>
<tr>
<th>Home#</th>
<th>Sample mean</th>
<th>n</th>
<th>( w_j )</th>
<th>Posterior mean</th>
<th>95% CrI</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>4.08</td>
<td>10</td>
<td>0.83</td>
<td>3.94</td>
<td>3.57–4.32</td>
</tr>
<tr>
<td>3</td>
<td>3.98</td>
<td>16</td>
<td>0.89</td>
<td>3.90</td>
<td>3.61–4.21</td>
</tr>
<tr>
<td>22</td>
<td>3.95</td>
<td>22</td>
<td>0.91</td>
<td>3.90</td>
<td>3.63–4.16</td>
</tr>
<tr>
<td></td>
<td>. . . . .</td>
<td>. .</td>
<td>. . . . .</td>
<td>. . . . .</td>
<td>. . . . .</td>
</tr>
<tr>
<td>2</td>
<td>2.46</td>
<td>14</td>
<td>0.87</td>
<td>2.57</td>
<td>2.24–2.89</td>
</tr>
<tr>
<td>4</td>
<td>2.08</td>
<td>7</td>
<td>0.77</td>
<td>2.36</td>
<td>1.91–2.79</td>
</tr>
</tbody>
</table>

3. FOR TEACHERS: DISCUSSION AND FUTURE WORK

Will the students “get it”? To answer this question, we have to understand what we want the students to “get” in the context of a one-hour seminar. There are more and more statistical studies that use Bayesian data analysis. Therefore, we feel compelled to give the students a flavor of Bayesian ideas. We hypothesize that students who take this seminar (with proper prerequisites) will have a good taste of the main concepts of the Bayesian paradigm. The concepts involve taking prior information and updating it with data to gain updated insight via posterior information. After this one-hour seminar the students will, we hope, grasp these basic ideas.

[Received May 2007. Revised January 2008.]

REFERENCES


