# TEACHING BAYESIAN STATISTICS TO UNDERGRADUATES: WHO, WHAT, WHERE, WHEN, WHY, AND HOW ®

W. M. Bolstad
University of Waikato
New Zealand

*At the present time, frequentist ideas dominate most statistics undergraduate programs, and the exposure to Bayesian ideas in undergraduate statistics is very limited. There are historical reasons for this frequentist dominance. Efron (1986) concluded that frequentists had captured the high ground of objectivity (p. 4). Bayesian methods have superior performance, often even outperforming frequentist procedures when evaluated under frequentist criteria. In the past, Bayesian methods were of limited practical use, since analytic solutions for the Bayesian posterior distributions were only possible in a few cases, and the numerical calculation of the posterior often was not feasible because of lack of computer power. Recent developments in computing power, and the development of Markov chain methods for sampling from the posterior have made Bayesian methods possible, even in very complicated models. It is clearly unsatisfactory for our profession that most of our students are not being introduced to the best methods available. In this paper I make a proposal for how our profession should deal with this challenge, by giving my answers to the journalistic "who, what, where, when, why, and how" questions about the place of Bayesian Statistics in undergraduate statistical education.*

## INTRODUCTION

We need a strategy for statistical education in the new millennium. David Moore (2001) noted that "although the discipline of statistics is healthy, its place in academe is not"(p.1). He contends that "our future depends strongly on achieving a more prominent place in undergraduate education beyond the first methods course".

Professional statisticians continue to be in high demand by employers, in industry, government, and education. The July 2001 issue of Amstat News has over 20 pages of advertisements for professional statisticians at all levels, in health and other industries, government, and educational institutions. Shettle and Gaddy (1998) found that in the US slightly over half (55%) of individuals with doctorate degrees in statistics were employed in the academic sector, with most of the remainder divided between the private sector (28%) and the government sector (10%). The median salary levels were comparable to other doctoral professionals. Demets, Woolson, Brooks and Qu (1998) conclude that while the supply of PhD statisticians seems to be increasing, a low unemployment rate suggests the profession is not saturated, in contrast to some other scientific fields. It is clear that there is a continuing demand for people with training in statistics.

Approximately 50% of total statistics degrees awarded in the USA during 1995 were at Masters level, and only about 33% at Bachelors level, (Iglewicz, 1998). Furthermore, from 1981 to 1995 the rate of increase in Bachelor degrees with statistics as a major (18%) was substantially less than that for Masters degrees (58%) and Doctoral degrees (60%) in statistics. Clearly statistics as a career option has low visibility to undergraduates. Yet, very large proportions of undergraduates are required to take a service course in statistics. We need to enthuse the best students in these service courses into taking further statistics courses. We aren't doing that now. This should not surprise us. As professional statisticians, we know Bayes methods have great theoretical advantages, and that these advantages can now be realized in practice with the advent of Markov chain Monte Carlo methods. I suggest that introducing Bayesian methods may be the key to revitalizing our undergraduate programmes in statistics. My answers to the journalistic questions on the place of Bayesian statistics posed in the title follow (but in a different order).

## WHERE

Introduction to Bayesian Statistics should be an alternative to the standard Introduction to Statistics first year service course. Most students only take one statistics course. This is our one opportunity to engage them.

WHAT

Introduction to Bayesian Statistics should to cover the same topics as our first year service course in statistics. These include techniques for gathering data such as methods for random sampling. The difference between observational studies and designed experiments should be emphasized. The data gathering method can justify the probability model we use in the analysis. Bayesian inferences should be made on the same parametric models as the service course, binomial proportions, normal means, and differences between normal means.

WHO

Mathematically well prepared students should be encouraged to take the Introduction to Bayesian Statistics course instead of the standard course. Bayesian statistics uses the rules of probability to make inferences, and that requires dealing with formulae. The actual calculus used is minimal. They only have to know that integrals represent areas under a curve. They don't have to evaluate them. Students who have previously passed the standard first year service course in statistics previously should also be allowed in.

WHEN

Now. We know that Bayesian estimates and intervals perform very well, often outperforming frequentist ones, even when evaluated using frequentist methods. What are we waiting for?

WHY

If we continue to put forward less effective frequentist methods of inference in our Introduction to Statistics courses, the decline of our discipline in the academic world will continue. Statistical methods are being re-invented in information technology, eg. machine learning, data mining, etc. These fields would be able to develop faster if there was more involvement with statistics. Statistics departments and statistics within mathematics departments are weak players in the contest for resources. We make heavy use of computing technology, but we are usually resourced at a lower level than Computer Science. We don't have a lock on Bayesian ideas. If there is a significant development of teaching Bayesian inference in other Departments, our profession will lose out. It is also true as statisticians, we can teach Bayesian statistical inference better than others.

HOW

The ASA/Joint Curriculum Committee made three recommendations for any course whose goal is to introduce the nature of statistics to beginning students. These are:

1. Emphasize the elements of statistical thinking.
2. Incorporate more data and concepts, fewer recipes and derivations. Wherever possible automate computations and graphics.
3. Foster active learning.

INTRODUCTION TO BAYESIAN STATISTICS AT WAIKATO UNIVERSITY

I have designed a one semester Introduction to Bayesian Statistics course that is compatible with those recommendations. My main objectives are to cover the same topics that would be in a frequentist introduction to statistics course, but to do the inference from a Bayesian perspective. Any introductory statistics course must start with data gathering. We cover random sampling from a real population, the difference between observational studies and designed experiments, and basic experimental design concepts as randomisation, and pairing. Sound methods for displaying and summarizing data are also covered. These topics take about six lectures. Real data sets are used, including those gathered by the students themselves. Stigler (1977) is a good source of historically interesting scientific data sets.

I use an integrated approach involving lectures, tutorials, and computing, often involving small-scale Monte Carlo simulations. Each of the three strands approaches the topics from a different path, which caters for students with differing learning styles. Each strand reinforces the

others, and allows the students to learn to think statistically. The following examples illustrate the integrated approach.

EXAMPLE 1: SAMPLING FROM A REAL POPULATION

Students understand intuitively that to get good estimates of population parameters, the sample must be representative of the population. Random sampling gives a way to statistically control for an unknown characteristic of the population. If the sample size is large enough, a simple random sample drawn without replacement will be very close to being representative with respect to that characteristic.

If we know a characteristic of the population, for instance the proportions of each ethnic group, simple random sampling might give a sample that does not have each group represented in the proper proportions. Stratified random sampling allows us to control for the known characteristic by sampling each stratum in the proper proportion, while still controlling other unknown characteristics by the random sampling within each stratum. We also discuss cluster random sampling. Students understand that people in the same neighbourhood tend to be more similar than people in different neighbourhoods, so cluster random sampling will be less efficient for the same size sample.

These ideas are reinforced in a tutorial, where we have a population of 100 rods that are divided into three "ethnic groups" based on colour, and we want to estimate the mean "income" for the population of rods. The income for each rod is written on a flag attached to the rod. I have developed a sampling table, which has the sampling frame for each of the sampling methods. Students draw three random samples of size twenty, one using each type of random sampling. We look at the random samples drawn by the class, and see how well each method represents "ethnic group". Each student summarizes and graphically displays his/her sample of incomes.

We further reinforce this in a computing assignment. I have written macros that allow the students to draw 200 random samples of size 20 from the same population of 100 rods. They do a small scale Monte Carlo study to evaluate how effective is each method for getting samples representative with respect to "ethnic group". They also can look at the distribution of the mean incomes over the 200 samples. They compare these sampling distributions for the mean over the three sampling methods, and see that the stratified sampling distribution is more concentrated around the true population mean.

EXAMPLE 2: SEX, DRUGS AND ROCK & ROLL SURVEY

My Introduction to Bayesian Statistics class combines with the Introduction to Statistics course for the "Sex, Drugs, and Rock & Roll" survey we have developed at Waikato University, (Bolstad, Hunt, & McWhirter, 2001). This survey involves students using randomised response to gather sensitive information about the class (eg number of sex partners, marijuana usage) without anyone having to divulge their own specific information. We discuss other sex surveys such as the Hite report that are not based on any sort of randomisation. Students readily understand how there would be a temptation to either refuse to participate, or give false information if such a survey is done in a traceable way. This allows us to discuss non-sampling errors, such as interviewer bias, and non-responses due to refusal. The randomised response means there is no reason for anyone to put in false information, and nearly everyone participates, although participation is completely voluntary.

Because the incorrect answers are put in by a known random mechanism, we can use statistical methods to get estimates about the population. I think this is perhaps the most fundamental idea in statistics. Statistical methods for inference assume the data comes from a known random structure. It is risky to use to statistical methods of inference, frequentist or Bayesian, for data that arises any other way. In the tutorial we use the known random structure for the data to obtain unbiased (frequentist) estimates of the proportions of students who have had $i$ sex partners for $i$ =0, …,5+, and the posterior distribution for the proportion who have previously used marijuana.

We do the Bayesian analysis is in a computing assignment. We evaluate the posterior distributions for the proportion of males and females who have had $i$ sex partners for $i$ =0, …,5+, and the posterior distribution for the proportion who have previously used marijuana. This is done

numerically using a macro I have written to do the required integrations. We evaluate the posterior distributions of the difference between the corresponding male and female proportions and find Bayesian credible intervals for these differences. These intervals are used to see if there is any relationship between gender and the number of sex partners.

EXAMPLE 3: REACTION TIME EXPERIMENT.

This tutorial involves students in the collection, analysis, and interpretation of data on the reaction times for their two hands. The student first determines a normal prior for their mean reaction time, by matching moments with their prior belief. Each student performs an experiment where his/her partner holds a ruler between the subjects thumb and fingers, then drops it without warning. The distance the ruler drops before it is caught is measured, and converted to a reaction time. The experiment is performed ten times for each hand. The posterior distribution of the mean reaction time is found for each hand. The posterior distribution of the difference between means of the dominant hand and the non-dominant hand is calculated. This is used to test the hypothesis

$H_0 : \mu_d - \mu_n \geq 0 \quad vs. \quad H_1 : \mu_d - \mu_n < 0$

by calculating the posterior probability of the null hypothesis. The student uses this test to determine if their dominant hand has a shorter reaction time.

EXAMPLE 4: COMPARISON OF BAYESIAN AND FREQUENTIST ESTIMATORS

In this computing assignment, each student runs a small scale Monte Carlo simulation to compare the performance of the usual frequentist and Bayesian estimators of $\pi$ in terms of mean square. 5000 random samples of size 10 are taken from a binomial distribution with success probability $\pi$. Two estimators are calculated on each sample, the proportion of successes (frequentist), and Posterior mean using uniform prior (Bayesian). The mean and variance of the 5000 random samples are calculated. The difference between the mean and the true value gives the bias of each of the estimators. The bias and variance are combined into the mean square of the estimator, which measures the average squared distance from the true value. These biases and mean squares are calculated for $\pi$ = .1, .2, …, .9, and the points plotted and joined. Students then are asked about the bias of each of the two estimators, and to determine over which range the Bayesian estimator is closer, on average, than the frequentist estimator.

INFERENCE IN THE COURSE

Bayesian statistics uses probability to make inferences. We use the probability axioms to develop probability on events. The conditional probability of an event *B* given an event *A* is developed from the reduced universe given the event *A* has occurred. I like to use a Venn diagram (see Figure 1.) to illustrate Bayes theorem.

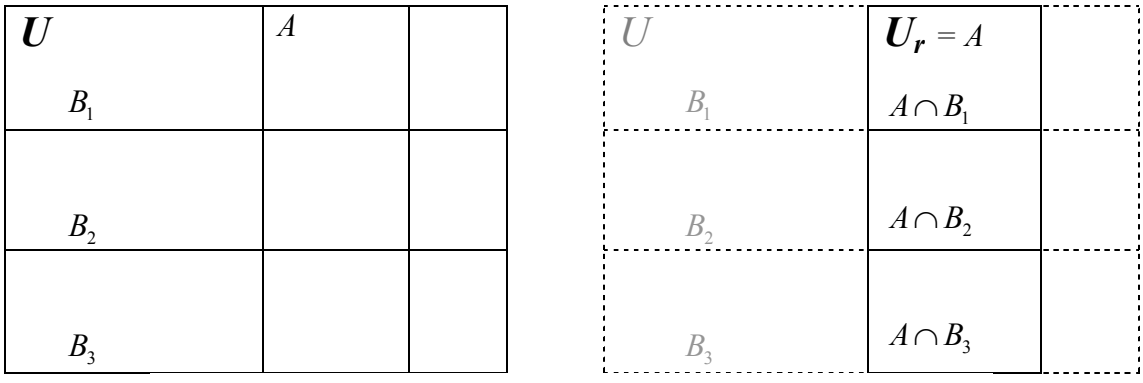


*Figure 1.* Venn Digram illustrating Baye's Theorem

We see that the posterior probability of $B_i|A$ equals their joint probability divided by the marginal probability of *A* which is just the sum of the joint probabilities summed over all *i.* This scales up the probabilities so the probability of the reduced universe equals one. Although the formula for conditional probability is symmetric for events *A* and *B*, we don't consider the events

symmetrically. We have a prior belief that event $B$ has occurred (its marginal probability). We use the analogous conditional probability formulae for $A$ given $B$ in the multiplication rule to find the joint probability, which is plugged into the formulae for the conditional probability of $B$ given $A$, which is called the posterior probability of $B$.

Each of the joint probabilities are found by multiplying the prior probability of $B_i$ times the conditional probability of $A|B_i$

Then we go to the joint random variable case, where the discrete random variable $X$ corresponds to an unobservable parameter and the $Y=y_j$ random variable is observed. The posterior distribution of $X$ given $Y=y_j$ is found by dividing each joint probability of $X=x_i$

and $Y=y_j$ by the sum of the joint probabilities summed over all $i$. The joint probabilities are found by multiplying the prior distribution by the corresponding probability of observing that particular value $y_j$ given each of the possible values $x_i$ , $P(Y=y_j \mid X=x_i)$. We see this follows the same pattern as Bayes theorem for events given above. The probability of observing that particular value $y_j$ as a function of the possible values of X is called the likelihood function. Posterior is proportional to prior times likelihood.

When the unobservable parameter random variable $X$ is continuous, the posterior distribution of $X$ given $Y=y_{obs}$ is observed is found by dividing each joint density value $f(x,y_{obs})$ by the integral of the joint density values integrated over all values $x$. Each value of the joint density is found by multiplying the prior density at $x$ by the likelihood function, the corresponding conditional probability density of observing $y_{obs}$ given that value $x$.

Bayesian inference methods are developed for the same models as in a standard introductory course: (binomial) proportions, normal mean, difference between normal means, difference between proportions, and simple linear regression. Subjective priors are chosen from the conjugate family by matching moments. The importance of graphing the prior to make sure that it reasonably reflects your belief is emphasized. The equivalent sample size is used to prevent students using a prior that is too precise relative to the sample size. Flat priors are used to represent prior ignorance. We demonstrate that using any reasonable prior doesn't change the posterior very much.

We use the posterior mean as an estimator. We show it has excellent characteristics when averaged over the sample space and performs better than the corresponding frequentist estimator in terms of mean square over the most of the possible parameter values. We introduce credible intervals for the parameter and contrast their useful interpretation with the backwards interpretation of the corresponding frequentist confidence interval. Hypothesis testing is well entrenched into science, so we feel we have to introduce it, but in a Bayesian manner. One-sided tests are performed by rejecting the null hypothesis whenever the posterior probability of the null hypothesis is below the level of significance. We test the credibility of a point null hypothesis versus a two-sided alternative by looking at whether the null value lies inside the corresponding credible interval.

We introduce the Student's t distribution as an approximation to be used when the variance is estimated form the sample. The idea of marginalization is introduced, but the calculations are beyond the level that I wish to go in an introductory class. The only place where we do marginalize is when we find the predictive distribution of a new observation by marginalizing out the parameter.

SUMMARY

I have taught this course several times at the University of Waikato over the past few years, and have always got good results. The students do understand the key ideas of Bayesian inference, and appreciate that it can give better performance than the corresponding frequentist methods. They get a good understanding the importance of randomisation in the gathering of data, and elementary data analysis. My experience with this course, along with others (Berry, 1997) and (Albert, 1997) show that an Introduction to Bayesian Statistics Course can be constructed that conforms to the ASA/Joint Curriculum recommendations, and it is a feasible alternative to standard Introduction to Statistics courses taught from the frequentist perspective. Good students not only cope with the direct use of probability inherent in Bayesian inference, they can

understand that frequentist inference is backwards by comparison. They realize that Bayesian inferences are often better than frequentist ones, even when evaluated by frequentist ideas.

I believe that having an Introduction to Bayesian Statistics available to well prepared students as an alternative to the standard Introduction to Statistics course is essential to the future of our field. We can no longer afford not to teach the best prepared students the best methods available

REFERENCES

Albert, J. (1997). Teaching Bayes' rule: A data oriented approach. *The American Statistician 51(3)*, 247-253.

Berry, D. (1997). Teaching elementary Bayesian statistics with real applications in science. *The American Statistician 51(3)*, 241-246.

Bolstad, W.M., Hunt, L.A., & McWhirter, J.L. (2001). Sex, drugs, and rock & roll survey in a first-year service course in statistics. *The American Statistician, 55(2),* 145-149.

Demets, D. L., Woolson, R., Brooks, C., & Qu, R. (1998). Where the jobs are: A study of amstat news job advertisements. *The American Statistician 52(4),* 303-307.

Efron, B. (1986). Why isn't everyone a Bayesian. *The American Statistician 40( 1),* 1-11.

Iglewicz, B. (1998). Selected information on the statistics profession. *The American Statistician*, *52(4),* 289-294.

Moore, D.S. (2001). Undergraduate programs and the future of academic statistics. *The American Statistician, 55(1),* 1-7.

Shettle, C., & Gaddy, C. (1998). The labour market for statisticians and other scientists. *The American Statistician 52(4),* 295-302.

Stigler, S. M. (1977). Do robust estimators work with real data? *The Annals of Statistics, 5(6)*, 1055-1098.