Bayesian Inferential Statistics Implemented in R

Jack McQuestion

University of Wisconsin-Eau Claire

Department of Mathematics

Faculty Mentor: Christopher Davis

The Power of AND

Abstract

Conventional frequentist statistics taught in undergraduate courses are obviously better than nothing, yet suffer from systematic failures that allow for easy p-hacking and consistent over-estimation of significance and effect sizes. Use of frequentist statistics is the strongest factor contributing to the replication crisis in the social sciences. Bayesian statistics, by contrast, have numerous advantages and can be tailored to any possible inferential problem. Many researchers and academics have advocated for Bayesianism, but only relatively recently have advances in computer technology allowed certain statistical methods to become practical. I wrote a script that allows one to carry out statistical inference using Bayesian statistics in R, An open-source programming language and software environment. The only assumption that the population parameters is an unknown, but fixed, Bernoulli distribution, though it can be generalized to categorical distributions and even multivariate numeric data. Methods are provided for constructing credibility intervals, establishing correlations, and comparing groups. Additional recommendations are given for data analysis and good scientific practice in general.

Background

In statistics there are two main schools of thought; frequentist statistics and Bayesian statistics. The main difference between frequentist statistics and Bayesian statistics is that Bayesian statistics support the idea of the prior: the degree of belief in something before any data is encountered. Frequentist statistics do not support this and instead rely entirely on support that is usually measured independently of prior and posterior probabilities: only the likelihood (usually for the null hypothesis) is considered. In addition, frequentist statistics treat probability as a fixed, but unknown parameter whereas bayesian statistics treat probability as a measure of uncertainty [1, 3, 4, 5].

Needless to say, this limits the kind of problems that frequentism can solve, though frequentist tests of probability have remained standard science due to social inertia, the early influence of Ronald Fisher, and the fact that many methods of Bayesian inference were impossiblle without digitial computers (in particular, MCMC methods) [1]

The foundation of Bayesian statistics is the equation:

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} \tag{1}$$

Which states that the likelihood of anything (such as an experimental hypothesis or statistical model) given the data, $P(A|B)$, is equal to the prior, P(A), multiplied by the odds given by the likelihood function, $P(B|A)$, which is then normalized against all other parameters under consideration(the denominator). The process of updating based produces a posterior distribution that describes what the probability of anything in the model is.

The foundations of Bayesian thought are very well established to the point where classical deductive logic is simply a special case of Bayesian probability with all truth values set to either 1 or 0 [1]. Unfortunately, Bayesian methods are almost never taught in introductory statistics courses (including at Eau Claire) and the problems with frequentist statistics have been well-documented elsewhere [1, 2, 3, 4]. The two main objections against Bayesian inferential statistics, among those who are aware of the distinction, are:

1. The choice of prior is too subjective.
2. Bayesian methods are fine in theory, but they are computationally intractable for any non-trivial problem.

Both of these are untrue using R, as will be demonstrated shortly.

Main Objective

To create a general system of Bayesian inference that can be applied to real-world problems, it is necessary to set some general constraints and goals. In accordance with the structural requirement put forth by Jaynes [1], for how an ideal reasoning agent would function, the system should have the following qualities:

1. If a conclusion can be reasoned out in more than one way, then every possible way must lead to the same result. Differing results from differing methods (as is common in frequentism) is not allowed. All roads must lead to Rome.
2. The method takes into account all available evidence. It cannot arbitrarily ignore some of the information, basing its conclusions only on what remains. It is non-ideological.
3. The method always represents equivalent states of knowledge by equivalent plausibility assignments. That is, if two quantities are the same in every way (except perhaps for the choice of labels) they should be described using identical probability statements.

Additionally, we will assume that all population distributions are some form of Bernoulli distribution.

The Method

A Bernoulli distribution is a type of a binary random variable that takes state 1 with probability p and state 2 with probability $1-p$. A useful metaphor is the idea of a coin that is biased towards either heads or tails to some extent. A Bernoulli distribution with p = 0.7 would denote a biased coin that comes up heads with probability 0.7 and tails with probability 0.3. A Bernoulli distribution with p = 0.5 would represent a fair coin. The likelihood function of a Bernoulli distribution is easy to calculate, which why the problem of determining whether a coin is fair or not (and to what extent it may be biased) has a very long basis in statistical literature [1, 3, 5]. The typical Bayesian approach is to model it using a beta function with varying options for the prior. A beta function is a probability density function (PDF) that is used to model various forms of binomial data. It is given by the equation:

$$Y = X^a * (1 - X)^b * NC \tag{2}$$

Where a and b are both non-negative integers, each representing the number of observations for some outcome category. The NC is a normalizing constant that ensures that the total area under the curve integrates to 1. This PDF represents our state of uncertainty as to which distribution (point along the x-axis) is the correct one. Each possible Bernoulli distribution has an infinitesimal chance of being correct given the data, but we can still make observations as to what sort of distributions are likely or unlikely. The prior is a flat, or non-informative prior where all possible Bernoulli distributions are considered equally likely before any data is observed. This is done by setting a and b equal to zero, making the PDF a constant function.

Let a cell be defined as a collection of observations (a and b) and the resulting posterior distribution from the beta function on the domain of [0, 1]. An example of this is seen in Figure 1. To determine whether or not a given cell supports a given hypothesis, integrate over the highest density interval (HDI) and see what values are included are excluded along the x-axis. If the highest density interval containing 95% of the probability mass excludes a Bernoulli distribution with p = 0.5, this is considered equivalent to rejecting the null hypothesis at the 0.05 level. The values on the edge of the HDI and the overall width of the HDI can also be considered a crude measure of effect size.

This is well established in the literature [1, 2, 3, 5]. but there is more that can be done.

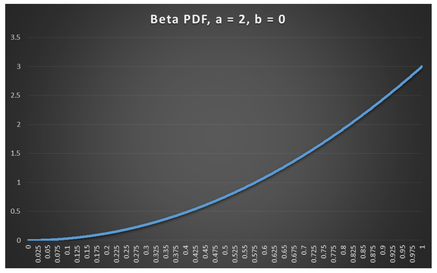

Figure 1: The PDF for a = 2, b = 0. The probability mass is concentrated on Bernoulli distributions where outcome "a" is more likely to occur.

Key Insight

All information can be represented in binary. Since the method makes inferences on binary data, it can be generalized to any data type, and most, if not all, forms of statistical inference (comparing groups. correlation, regression, model fitting, etc.)

New Applications

The method can be generalized to make inference about non-binary data via "forking" together different cells into a tree (see Figure 2). This is done by arranging the outcome types into arbitrarily determined supersets, assigning each superset to one side of a cell as a possible outcome type, and then subdividing the various intervals until we arrive at the probabilities for any individual outcome. If, for example, there are four possible outcomes, we would join together three cells in order to make inferences about the true population distribution, as depicted in Figure 2. In more formal terms, the method can make inferences about categorical distributions ("a coin with more than two sides") in addition to Bernoulli distributions, someting that is traditionally done in Bayesian statistics with dirchlet distributions. However, the mathematics behind dirchlet inference are complicated by things like multidimensional integration and Gibbs sampling that make it significantly harder to understand and implement, especially for novices.

Making inferences about categorical distributions is possible because the cells higher in the tree are systematically controlled by those lower in the tree. Observations "trickle up" until they reach the top. One interesting phenomenon is that observations lower in the tree count for less than those higher in the tree. Specifically, an observation in superset a does not increment a by one, but by 2^{-i} where i is the number of levels the cell is from the top of the tree. While unintuitive, this makes the method consistent with generalizing Laplace's rule of succession to an arbitrary number of outcomes as formulated by Jaynes [1]. Even more specifically, this allows (with psedocounts) allows the method to work when the number of outcome types is not a power of 2.

Since cells can be forked repeated and iteratively, there is no limit to the ultimate number of outcome types that can be inferenced using the method. For working with interval or ratio data, the sets become subranges of whatever range the data was found on, to arbitrary precision, with each additional cell representing a new subrange that is mutually exclusive from the ranges on the opposite side of the tree. This is convenient because it means that the method effectively does not discrimate based on data type, something that occurs with frequentist tests. a variable being "numeric" is a difference of degree rather than type.

Mathematically, The trend is that the number of cells needed is equal to the number of outcome types minus one. As well, the HDI width must be adjusted to account for the number of cells. If, for example, there are three cells, and we want confidence of 95% for our final values, we must find the $\sqrt[3]{.95}$ HDI for each individual cell. Correlation (for example, heght and inteligence) is done by comparing the probability mass distribution of trees while correcting for sample sizes, an area of research that is still ongoing.

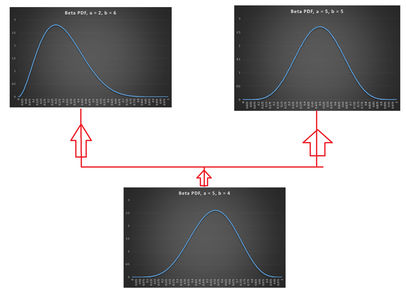

Figure 2: A sample tree for four outcome types. The tree uses three cells.

References

1. Jaynes, E. T. (2003). Probability Theory: The Logic of Science. Cambridge, NY: Cambridge University Press.
2. Berry, D. (1997). Teaching Elementary Bayesian Statistics with Real Applications in Science. The American Statistician, 51(3), 241-246. doi:10.2307/2684895
3. Justus, J. (2011). Evidentiary inference in evolutionary biology. Biology & Philosophy, 26(3), 419-437. doi:10.1007/s10539-010-9205-7
4. Wagenmakers, E. J., Lee, M., Lodewyckx, T., Iverson, G. (2010). Bayesian Versus Frequentist Inference. In Boelen, P. A. (Ed.), Bayesian Evaluation of Informative Hypotheses (pp. 181-207). New York, NY: Springer
5. Glickman, M. E., & Van Dyk, David A. (2007). Basic Bayesian Methods. In Ambrosius, W. T., (Ed.), Topics in Biostatistics: Methods in Molecular Biology (pp. 319-338). Totowa, NJ: Humana Press.

Pending Research

Numerous problems remain to be solved. A functioning implementation that compares groups on the basis of probability mass distribution (rather than intervals) is still in progress with eventual optimization towards look-up tables. As well, the implementation for categorical distributions is still buggy as identical cell counts do not always produce identical plausibility statements, meriting further work. Additionally, I intend to test the black raven problem using the script to determine which conclusions are actually supported in practice.

Acknowledgements

Special Thanks to Dr. Christopher Davis, Dr. Jessica Kraker, and Dr. Mohammad Aziz for their dedication and insight. Also thanks to Learning and Technology Services for printing the poster.