SESUG Paper 289-2018

A Flexible Approach to Computing Bayes' Factors with PROC MCMC

Tyler Hicks, University of Kansas

ABSTRACT

In many instances, such as the one considered in this paper, the Bayes' factor enhances the meaningfulness of tests of hypotheses. The Bayes' factor measures the extent to which one hypothesis fits the data better than another. Although its analytic derivation is no small feat, Markov Chain Monte Carlo (MCMC) algorithms enables its successful empirical derivation. Using an example-based approach, this pedagogical paper explores an instance when the Bayes' Factor is clearly preferable to the *p*-value. The paper also describes Bayes' Factor calculation within the SAS/STAT® MCMC procedure. Although the calculation of Bayes' Factor is currently not a default option in PROC MCMC, this paper shows that its built-in functions allow its computation all the same.

Keywords: BAYES' FACTOR, PROC MCMC, BAYESIAN TESTS

INTRODUCTION

This paper draws upon a databased example with a long history as a pedagogical illustration in mainstream statistics. Suppose that a woman claims that she has an extrasensory ability to perceive thoughts. Suppose further that a contest was held to test her powers. In this contest, a man concentrated on a number between 1 and 10, and the woman intuited whether he was thinking of an odd or even number. In 26 trials, the woman made 18 correct guesses—i.e., an impressive accomplishment. It is clear that a person with telepathy is more likely to achieve a 70% hit rate than a random guesser, but it is unclear how much likelier? The Bayes' Factor (BF_{10}) is the measure of how much likelier.

In this example, a Bayes' Factor of 5.77 would indicate that telepathy fits the data 5.77 times better than chance. This finding is not equivalent to an endorsement of telepathy, though. Because the initial plausibility of telepathy is enormously small, a Bayes' factor of big stature is needed to make telepathy more probable than chance. For instance, if chance was held to be 1,000 times more initially plausible than telepathy, a Bayes' Factor of 1,000 is needed to make the credibility of telepathy even with chance. In this light, 5.77 looks small. This paper has three aims: it (a) shows how to calculate Bayes' Factor, (b) why it is worth calculating, and (c) example code to calculate it with the SAS/STAT® MCMC procedure.

HOW IS BAYES' FACTOR CALCULATED?

To calculate the BF_{10}**,** first select a meaningful set of hypotheses to compare and second calculate the likelihood of obtaining the data under each hypothesis. For step 1, this paper setups a binomial model for the number of correct guesses (shown below) and also uses this model to define the primary hypotheses of interest (telepathy vs. chance):

$$r \sim Bin(\theta, N) \qquad \text{[Binomial model for data]}$$

and

$$H_0\colon\ \theta = .50 \qquad \text{[.50 represents chance]}$$

$$H_1\colon\ \theta = .75 \qquad \text{[.75 represents telepathy]}$$

where r is the number of obtained successes (18), θ is the probability of success at each trial, and N is the total number of trials in the probability experiment (26). For now, .75 was chosen to represent telepathy because its value is equidistance between .50 (chance) and 1 (perfection).

For the second step in the calculation of BF_{10}, this paper computes model-based likelihoods of data at each of the two hypothesized values for θ (.50 vs. .75). To accomplish this masterstroke, the binomial likelihood function, $f(r|\theta, N)$, comes in handy:

$$f(r|\theta, N) = \binom{N}{r} \theta^r (1-\theta)^{n-r}$$

where $\binom{N}{r}$ is the binomial coefficient, r and N are both observable sample statistics, and the value of θ is set to be either .75 or .50. The binomial likelihood function calculates the likelihood of obtaining 18 successes in 26 trials given each hypothesis:

$$f(\theta = .75|r = 18, N = 26) = .13$$
$$f(\theta = .50|r = 18, N = 26) = .02$$

As the visualization in Figure 1 (below) indicates, the telepathy model will produce 18 successes in 26 trials at a much higher rate than the chance model. In fact, as a ratio of likelihoods, the BF_{10} compares the relative sizes of the two model-based likelihoods:

$$BF_{10} = \frac{f(\theta = .50|r = 18, \mathbf{N} = 26)}{f(\theta = .50|r = 18, \mathbf{N} = 26)} = \frac{.13}{.02} = 5.77 \quad (1)$$

To communicate this finding to a general audience, a report might say:

> *"BF analysis indicates that obtaining 18 successes in a rigorous test like this one is 5.77 times likelier to happen in a scenario were telepathy is real as opposed to fake."*

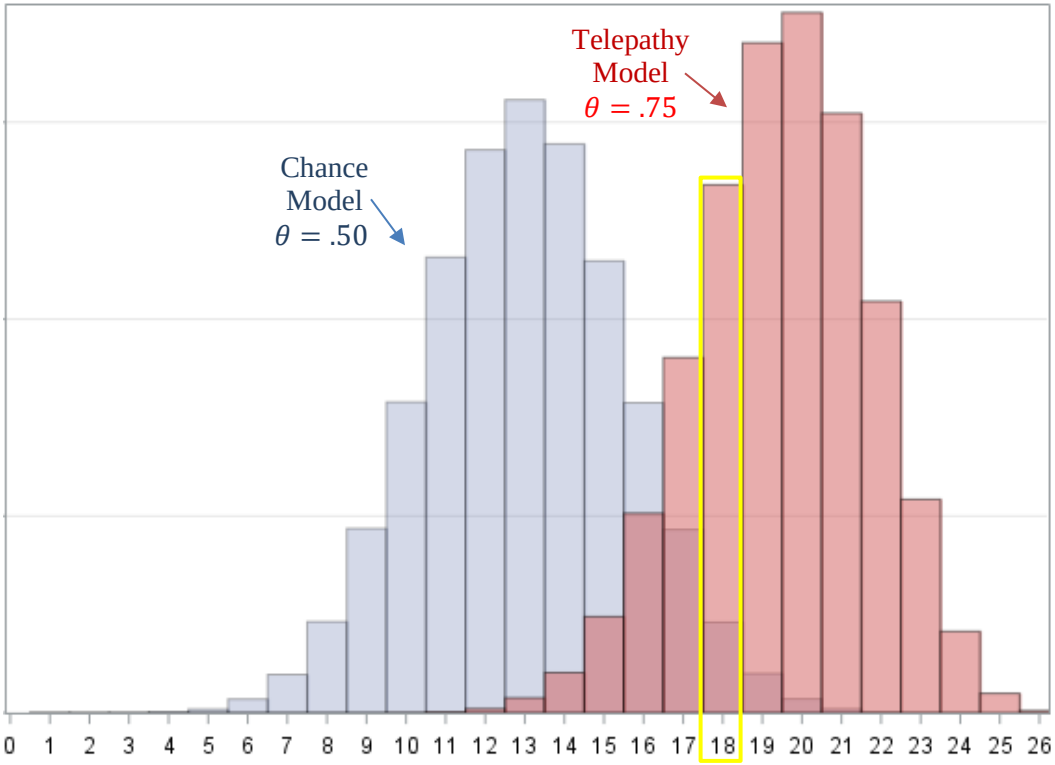

Figure 1. Sampling Distributions of the "Chance Model" (.50) and "Telepathy Model" (.75)

WHY CALCULATE BAYES' FACTOR?

Because criticisms of *p*-values are everywhere (e.g., Carver, 1978), this explanation of why Bayes' factor is preferable will be brief. Generally speaking, the *p*-value is the probability that a chance setup would generate a test statistic more extreme than the obtained test statistic if the null hypothesis was true. In this particular example, a *p*-value would be the probability that a random guesser would perform better than the woman under identical conditions. To calculate the correct *p*-value, a tester needs to recover the stopping rule implemented in the experiment—a proviso frequently forgotten (Howson & Urbach, 2006). For example, was the correct stopping rule to desist at the 26th trial or was it to keep on running trials until the 18th success was hit? Its statistical significance depends on the stopping rule.

Suppose that the stopping rule was to stop at 26 trials. The *p*-value is then statistically significant. This is because the *p*-value is now the probability that a random guesser exceeds 18 correct guesses in 26 trials (no more trials, no less trials). The number of successes is thus the test statistic. Asymptotic theory shows that the requisite sampling distribution is binomial. As depicted in Figure 2, the binomial leads to a categorial rejection of the null hypothesis. (Note also that the *p*-value procedure strangely rejects chance here even though the initial probability of telepathy is next to null for most scientists).

However, the above *p*-value is wrong if that was the wrong stopping rule. For example, if the stopping rule had really been to keep on running trials until the 18th success was hit then the correct *p*-value is not statistically significant. This is because the correct *p*-value is in fact the probability that a random guesser needs more than 26 trials to hit 18 successes. In other words, the number of trials is the test statistic as opposed to the number of successes. This means that the sampling distribution is negative binomial. As depicted in Figure 2, the negative binomial leads to a "fail to reject" decision for the null hypothesis. (The *p*-value procedure in this case learns nothing of substantive interests from this data).

It may seem strange that a stopping rule can determine whether a *p*-value is statistical significant or not, but when the stopping rule is known this is not a big deal. Unfortunately, recovery of the correct stopping rule may not be achievable. In fact, there may not even be a "correct" stopping rule to recover. Perhaps, the original experiment ended when time ran out. If so, there is no matter of fact about whether the binomial or negative binomial sampling distribution is correct. One consolation in this situation is that asymptotic theory guarantees sampling distributions of random test statistics become normalcy prevails as sample sizes rise so stopping rules are of no concern. However, this fact is no help with small samples. Moreover, test statistics from convenience samples may not even have random sampling distributions.

Fortunately, the Bayes' Factor is robust to the stopping rule problem for at least two reasons. First, whereas the classic *p*-value procedure is based on asymptotic theory, the Bayes' Factor procedure is based on likelihood theory. This means that it conforms to the likelihood principle (which states that the likelihood of the data is the only relevant factor about data for inference). The *p*-value procedure, in contrast, breaks the likelihood principle when it considers the sampling distribution of a test statistic in inference. Second, stopping rules are only reflected in the constants that differentiate the binomial, negative binomial, and Bernoulli likelihood functions from each and such constants get canceled out in Bayes' factors:

$$BF_{10} = \frac{\underline{\text{Binomial}}}{\cancel{\binom{N}{r}}\theta_1{}^r(1-\theta_1)^{n-r}} = \frac{\underline{\text{Negative Binomial}}}{\cancel{\binom{N-1}{r}}\theta_1{}^r(1-\theta_1)^{n-r}} = \frac{\underline{\text{Bernoulli}}}{\theta_1{}^r(1-\theta_1)^{n-r}} \qquad (2)$$

This computational property of likelihood ratios entails that binomial, negative binomial, and Bernoulli likelihood functions reproduce the same Bayes' Factors. These functions thus are members of an *equivalent class of likelihood function*. Thanks to this concept, analyst only needs to select the correct "class" of likelihood functions to calculate the BF_{10} (which is very achievable goal).

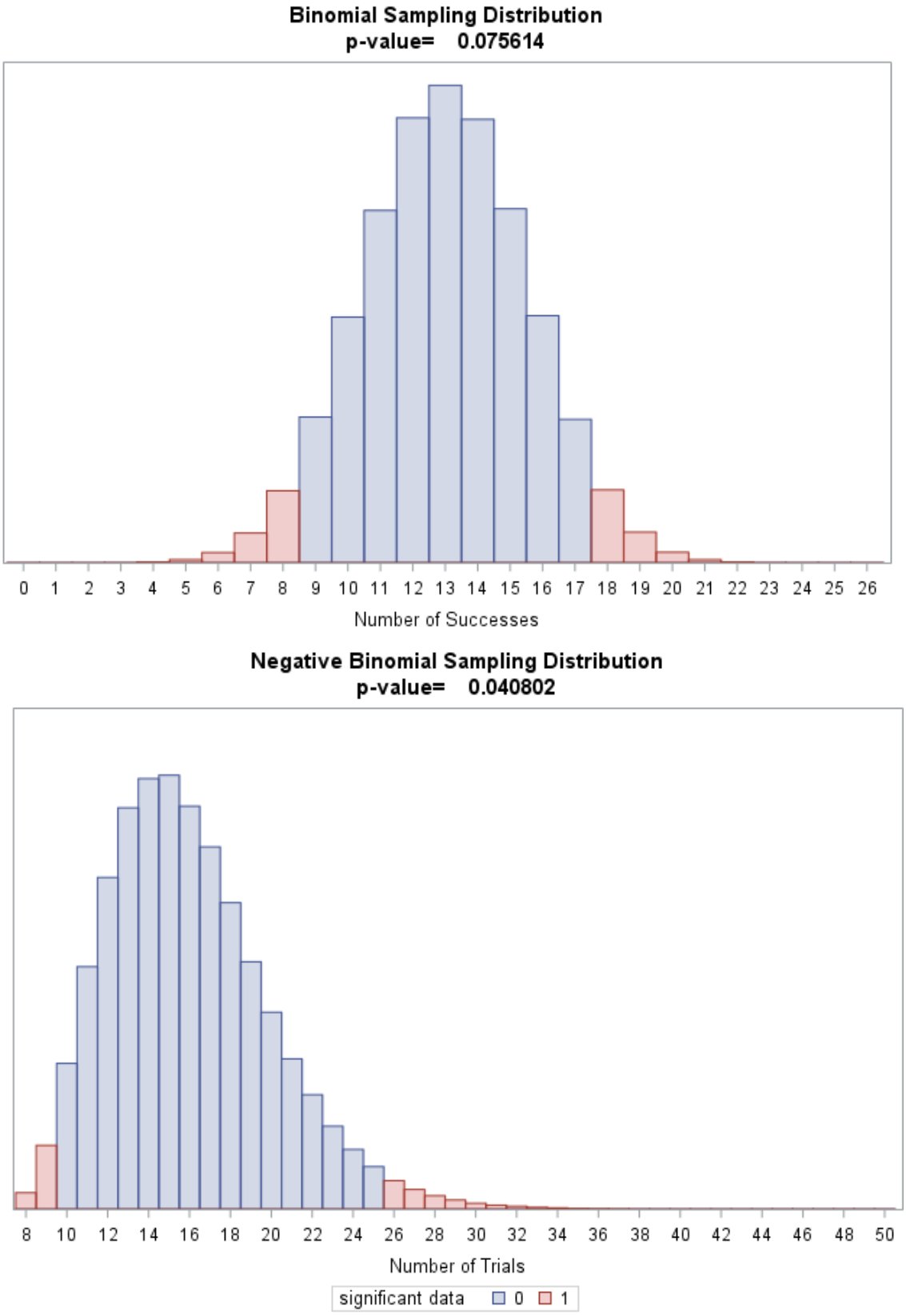

Figure 2. Sampling Distributions for the Same Null Model that Reflect Different Sampling Plans

CAN PROC MCMC CALCULATE A BAYES' FACTOR?

Bayes factor is not a default option in PROC MCMC, but it has built-in functions that allow its calculation. Chen (2008) provides an excellent refresher to consult on all the built-in functions of PROC MCMC and other syntax details. Table 1 (below) provides workable code to calculate the Bayes' Factor with PROC MCMC for the telepathy example. With `SEED=23`, PROC MCMC outputs $BF_{10} = 5.77$ (an estimate which matches the exact value arrived at earlier in the paper through analytic derivation). As Table 1 shows, this program has four basic parts. Note that the Bayesian hierarchical model fit to data in Part 2 is described in detail in the next section. For other technical details about this model, this paper refers interested readers to Kruschke (2011) presentation.

Part	Example Syntax	Notes

1	`DATA dataset;`
	`  r=18; n=26;`
	`RUN;`
		Records 18 successes and 26 trials into a SAS datafile.

2	`PROC MCMC DATA=dataset NMC=500000;`
	`  *Specify Parameters;`
	`  PARM M;`
	`  *Specify Prior;`
	`  PRIOR M ~ BINARY(0.5);`
	`  IF M=0 THEN THETA=0.50;`
	`  IF M=1 THEN THETA=0.75;`
	`  *Specify Liklihood Function;`
	`  MODEL R ~ BINOMIAL(N,THETA);`
	`  *Save Output;`
	`  ODS OUTPUT PostSumInt=BF;`
	`  RUN;`
		Implement PROC MCMC and fit a Bayesian hierarchical model to the data that contains the two hypotheses (.75 vs. .50) as subcomponents and save MCMC output to a SAS datafile (See next section for further details on why fitting this model work)

3	`Data BF; SET BF;`
	`  if Parameter='M';`
	`  BF=Mean/(1-Mean); RUN;`
		Postprocess MCMC output to calculate the Bayes' Factor

4	`PROC PRINT DATA=BF;` `  Var BF; run;`	Print Bayes' Factor.

Table 1. Example code for implementing a Bayes' Factor Procedure

Commentary on each part of the syntax:

Part 1. This syntax creates a SAS dataset with the needed sample data. Because the number of successes and number of heads are sufficient statistics, they "suffice" for the analysis.

Part 2. This syntax fits a Bayesian hierarchical model to the data in PROC MCMC (see the next section for details). The option `NMC=500000` requests that PROC MCMC generate a posterior sample for the model parameters that contains 500,000 iterations. This example syntax has one parameter, called, '`M`' (which is explained in the next section).

Part 3 and **Part 4.** This syntax manipulates the mean of the posterior probability for `M` in order to empirically derive an approximate BF_{10}. In this configuration, the likelihood of obtained data on the alternate hypothesis is in the denominator spot.

BAYESIAN HIERARCHINCAL MODEL

This section is at the heart of the paper. Although this section is short, the content is difficult to fully grasp in a single reading. Please read carefully.

In the telepathy example, the SAS program requested that PROC MCMC fit a "*binary-binomial*" model to the data. An example of a such model is:

$$\theta \sim Binary(\phi = .50) \quad \theta = \{.50, .75\} \qquad \text{[level 2: Binary Component]}$$

$$r \sim Binomial(\theta, N) \qquad \text{[level 1: Binomial Component]}$$

where r is the number of successes in an experiment out of *N* trials, θ is the probability of success at each trial, and ϕ is the probability that θ takes on a value of .75 as opposed to .50. As is shown below, the value of ϕ has to be .50 in order for this approach to calculating the correct Bayes' Factor. Importantly, this same model can be construed as representing a two-stage lottery. In the first stage of this lottery, a random value for θ must be drawn. The values .50 and .75 are equally probable selections for θ. In the second stage of this lottery, another random number (r) is drawn that could be any integer between 0 to *N* (inclusive). The probability of selecting a value of r is conditional on the selected value of θ .

To specify this binary-binomial model in PROC MCMC, a programmer must curve up its two components into a "prior distribution" and "likelihood function." Whereas the level-2 binary component is the "prior distribution" set on the θ parameter, the level-1 binomial component is the "likelihood function." PROC MCMC expects this format for model components even though this conventional description of Bayesian models become blurry in advanced statistics. That is, a principled distinction between prior distribution and likelihood function is a carryover from frequentist statistics and it breaks down in complex modeling situations. However, to make PROC MCMC happy it is best to accept the conventional format and define something as a prior distribution and something as likelihood function even though the model is really a two-stage lottery.

When PROC MCMC learns that 18 successes were obtained in the experiment out of 26 trials, the only job it wants to perform is the task of estimating the conditional probability that a .75 had been drawn in the first stage of the lottery for θ instead of .5. The conditional probabilities of .75 and .50 are different then the likelihoods of data under each hypothesis. However, because PROC MCMC knows that .75 and .50 have the same initial plausibility, the ratio of the conditional probabilities of .75 given data over .50 given data will be equivalent to the ratio of the likelihoods of data given .75 over .50 such that:

$$\frac{P(\theta = .75|r = 18)}{P(\theta = .50|r = 18)} = \frac{L(r = 18|\theta = .75)}{L(r = 18|\theta = .50)}$$

This mathematical equivalency undergirds the formula in Part 3 of the above SAS program that actually generates the approximate BF_{10}:

$$\frac{E(M)}{1 - E(M)} = BF_{10}$$

where *E(M)* is the posterior mean of parameter *M* found in the generated MCMC output, which in this case equals the conditional probability that .75 was drawn in the lottery given the obtained data.

Importantly, a normal-binary model configuration is also possible for more complex databased examples. At its first level, a normal process outputs a continuous random outcome. However, the mean parameter of this normal process is itself construed to also be a random value. Perhaps, the level-2 binary process randomly outputs 0 and 0.3 (in standard deviation units) for the level-1 mean parameter. This model is easily fit to data in PROC MCMC. Analysts could use it to obtain the approximate Bayes' Factors in t-test, regression, and ANOVA contexts (see Hicks, Rodriguez-Campos, & Choi, 2017 for further elucidation of this approach). In fact, with creative ingenuity for building hierarchical models, the sky is the limit to computing Bayes' Factor with PROC MCMC. Yet, troubleshooting PROC MCMC when calculating Bayes' Factors in advanced modeling situations is a topic that will need development in another paper.

THE MAGIC OF PRIOR DISTRIBUTIONS

Recall, that up until now $\theta = .75$ in the telepathy model. Representing "telepathy" with a single value (say .75) is actually suspect practice. Why not .74 or .76? A prior distribution over a range of plausible values solves this problem. For example, Hicks et al. (2017) suggested a uniform distribution for telepathy, $\theta \sim U(.65,1)$. This distribution is centered on .825 but it's standard deviation reflects an appropriate degree of uncertainty. It states that values between .65 and 1 equally count as telepathy, if true. Other shapes of prior distributions are also possible. The modest aim is just to adequately represent the hypothesis with a prior distribution that a skeptical audience will find to be acceptable practice for purposes of analysis.

For telepathy, table 2 (below) provides a modification of the SAS program that represents (not measures) the concept of telepathy with a uniform prior distribution. Comparison of the MCMC outputs from the two SAS programs, shows that the BF_{10} moved from 5.77 to 2.93 with this modification. This reduction in values is desirable. The new BF_{10} compares the expected likelihood for values in the entire range of possible telepathy values (i.e., values ranging from .65 to 1). Because this new range of telepathy values includes many values that will poorly fit 18 successes in 26 trials, such as the value of 1, the telepathy model fares much worse against the chance model even though the data is still more likely under it. The point of this section is that wise testers should only compare meaningful formulations of each hypothesis (Christensen, Johnson, Branscum, & Hanson, 2011).

Part	Example Syntax	Notes

2a	`PROC MCMC Data=dataset nmc=500000`
	`*Specify Parameters;`
	`Parm M MU;`
	`*Specify Prior;`
	`PRIOR MU ~ UNIFORM(0.65, 1.00);`
	`PRIOR M ~ BINARY(0.5);`
	`IF M=0 THEN THETA=0.50;`
	`IF M=1 THEN THETA=MU;`
	`*Specify Liklihood Function;`
	`MODEL R ~ BINOMIAL(N,THETA);`
	`*Save Output;`
	`ODS OUTPUT PostSumInt=BF;`
	`RUN;`
		Implement PROC MCMC and fit a revised Bayesian hierarchical model to the data that contains the two hypotheses (.75 vs. U(.65,.85)) as subcomponents and save MCMC output to a SAS datafile.

Table 2. Example code for implementing a Bayes' Factor Procedure

OVERCOMING SMALL SAMPLES

Bayesian estimation has been advertised as a solution to small samples. Yet, a misconception of Bayesian estimation is that all its small sample benefits derive from subjective prior distributions that make untestable assumptions. There is a grain of truth in this misconception but only a grain. Bayesian estimation has many resources for dealing with small samples. In fact, the prior distribution may not even be its most important such resource. For example, MCMC algorithms are much more stable in small sample situations than the optimization algorithms that typically drive maximum likelihood estimation. Another advantage, which this paper wants to highlight, is that Bayesian estimation is based on likelihood theory rather than asymptotic theory. As such, the Bayes' Factor procedure is optimum for small samples because it is robust to the stopping rule problem. Although it is true that central limit theorems guarantee normalcy in sampling distributions regardless of stopping rules, these central limit theorems will not start kick-in when samples are kept small. Thus, stopping rules will affect *p*-values in small samples. However, Bayes' factor analysis avoids this mess. Table 2 provides the PROC MCMC code that allows SAS users to empirically demonstrate the concept of an equivalent class of likelihood functions. For example, with the same SEED is set, the program using the negative binomial model will return the same Bayes' Factor as the equivalent program with the binomial model (e.g., try `SEED=23`).

Part	Original Example Syntax	Modified Example Syntax

1a	`DATA dataset;`
	`  r=18; n=26;`
	`RUN;`
		`DATA dataset;`
		`  r=18; n=26; k=n-r;`
		`RUN;`

2b	`PROC MCMC DATA=dataset NMC=500000;`
	`  *Specify Parameters;`
	`  PARM M;`
	`  *Specify Prior;`
	`  PRIOR M ~ BINARY(0.5);`
	`  IF M=0 THEN THETA=0.50;`
	`  IF M=1 THEN THETA=0.75;`
	`  *Specify Liklihood Function;`
	`  MODEL R ~ BINOMIAL(N,THETA);`
	`  *Save Output;`
	`  ODS OUTPUT PostSumInt=BF;`
	`  RUN;`
		`PROC MCMC DATA=dataset NMC=500000;`
		`  *Specify Parameters;`
		`  PARM M;`
		`  *Specify Prior;`
		`  PRIOR M ~ BINARY(0.5);`
		`  IF M=0 THEN THETA=0.50;`
		`  IF M=1 THEN THETA=0.75;`
		`  *Specify Liklihood Function;`
		`  MODEL K ~ NEGBIN(R,THETA);`
		`  *Save Output;`
		`  ODS OUTPUT PostSumInt=NBF;`
		`  RUN;`

	Output with binomial: BF_{10}=5.7755	Output with Negative Binomial: BF_{10}=5.7755

Table 3. Example code for demonstrating the concept of equivalent classes of likelihood functions

CONCLUSION

This paper focuses on the computation of Bayes' Factor in PROC MCMC. PROC MCMC is built to estimate posterior probabilities rather than the Bayes' Factor. The secret to obtaining Bayes' Factors with this MCMC procedure is to work with the estimates of posterior probabilities for the different values of θ in a binary-binomial model. Then, assuming that both values of θ had equal prior probablity, set the ratio of the estimated posterior probabilities to be equal to the ratio of the likelihoods of data under each hypothesis. This ratio of likelihoods is the Bayes' Factor. Thus, PROC MCMC is "tricked" into outputting the Bayes' Factor even though the MCMC procedure only wants to estimate posterior probabilities.

In wrapping up this paper, a few talking points about the complex relationship between *p*-values and Bayes' Factors are in order. First, it is best to see the Bayes Factor and the *p*-value as different tools in the toolbox as opposed to members of opposing statistical systems (Efron, 1986). In fact, the idea of a Bayesian *p*-value is becoming popular in the Bayesian system (especially in posterior predictive checking and Bayesian computation). These two types of test serve complementary functions in the Bayesian system. Whereas the Bayesian *p*-value assesses the "absolute" fit of a model in a posterior predictive check, the Bayes' Factor assesses the "relative fit" of alternate models in the modeling competition (Morey & Rouder, 2011). Because of this key difference in function, all *p*-value procedures require much stronger assumptions to be valid than Bayes' Factor procedures—which may mean that a valid *p*-value procedure is not always achievable. Thus, researchers should always check whether the Bayes' Factor may in fact be the more meaningful test before testing hypotheses (Kruschke, 2013).

REFERENCES

Bolstad, W. M. (2010). *Understanding computational Bayesian statistics.* Hoboken, NJ.: John Wiley & Sons, Inc.

Carver, R. (1978). The case against statistical significance testing. *Harvard Educational Review, 48*(3), 378-399.

Chen, F. (2009). *Bayesian modeling using the MCMC procedure (SAS Global Forum Paper 257-2009).* Cary, NC: SAS Institute Inc.

Christensen, R., Johnson, W., Branscum, A., & Hanson, T. E. (2011). *Bayesian ideas and data analysis* New York, NY: CRC Press.

Efron, B. (1986). Why Isn't Everyone a Bayesian. *The American Statistician, 40*(1), 1-5. Fisher, R. A. (1925). *Statistical Methods for Research Workers.* Edinburgh, KU: Oliver & Boyd.

Hacking, I. (2001). *An introduction to probability and inductive logic.* New York: Cambridge University Press.

Hicks, T., Rodriguiz-Campos, L., & Choi, J. H. (2017). Bayesian posterior odds ratios: Statistical tools for collaborative evaluations. *American Journal of Evaluation*, 1-12. doi:10.1177/109821407704302

Howson, C., & Urbach, P. (2006). *Scientific reasoning: The Bayesian approach* (3rd ed.). Chicago: Open Court.

Kruschke., J. (2011). *Doing Bayesian data analysis: A tutorial with R and BUGS.* New York, NY: Elsevier.

Kruschke. J. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology, 142*(2), 573-603.

Morey, R. D., & Rouder, J.N. (2011). Bayes' Factor approaches for testing null hypotheses. *Psychological Methods, 16,* 406-419.

CONTACT INFORMATION <HEADING 1>

Your comments and questions are valued and encouraged. Contact the author at:

Tyler Hicks
University of Kansas
1450 Jayhwak Blvd
Lawrence, KS 66045
tahicks@ku.edu