

## Invoking Survey Weights to Calculate Bayes' Factor with PROC MCMC to Generate More Generalizable Inferences

Tyler Hicks, University of Kansas; Graham G. Rifenbark, University of Connecticut

### ABSTRACT

In many instances, such as the one considered in this paper, SAS users rely on survey weights to enhance the representativeness of samples. As an example, pollsters use survey weights to better forecast elections with cheaper samples. Naturally, survey weights are also very relevant when generalizing statistical tests to populations. To engage in testing, Bayesian leaning statisticians turn to Bayes' Factors. The Bayes Factor (BF) quantifies the extent to which data support one model better than another. As an example,  $BF=5$  indicates data support the alternate model five times better than the null model. Although adjusting the BF with sampling weights is no small feat, Markov Chain Monte Carlo (MCMC) algorithms enables their successful empirical derivation. Using an example-based approach, this pedagogical paper shows how SAS users without extensive technical expertise in Bayesian methods and survey weights can fit a model to recover BFs from data that adjust for survey weights using in-built functions in the SAS/STAT® MCMC procedure. This paper thus helps SAS users take the extra step of marshaling the full power of survey weights in Bayesian testing with PROC MCMC.

### INTRODUCTION

Because of the centrality of survey weights in this paper, a brief conceptual introduction is in order. Broadly speaking, random data collection can be executed either by employing *simple random sampling* (SRS) or a *complex sampling design* (CSD). When estimating statistical models using data stemming from SRS, observations equally contribute to the likelihood function. Therefore, each observation is given a weight of 1 and the sum of weights equals the number of observations used in the analysis. On the other hand, when CSD is used to collect data, it is necessary either to increase or decrease the contribution (or influence) of the observations to the likelihood function so that meaningful inferences can be made about the population of interest. As such, the sum of weights corresponds to the population sought after (e.g., population of students with disabilities). That said, there are many instances in which influence on the likelihood function are not assumed uniform across observations. Take for example propensity score weighting in observational or quasi-experimental designs where it is not possible to randomly assign subject to treatment and control conditions.

When evaluating research hypotheses via statistical modeling, analysts can choose between frequentist methods and Bayesian methods, while the former has been the most popular approach for conducting analyses, Bayesian methods have become increasingly used for conducting data analysis. Bayesian methods consider the likelihood of the data, as do frequentist methods, however, prior knowledge (or beliefs) is also incorporated into their inferences. Ultimately, through simulation, it is possible to access posterior distributions and use them to make probabilistic inferences. Some advantages to using Bayesian methods are the following: (a) sample size does not have to be as large as is the case with frequentist statistical models, and (b) flexible hypotheses can be modeled and tested, unlike traditional null hypothesis testing where an "effect" is found when an estimate is found to be significantly different from zero and the decision is either to reject the null or fail to reject the null.

## THE BAYESIAN TOOLKIT

The Bayesian framework offers inferences that are straightforward and are probabilistic in nature. In fact, the p-values from the frequentist framework are often interpreted as if they stem from the posterior distribution, and the p-value is commonly misinterpreted as the probability of the null hypothesis when in fact, this that interpretation is emphatically not the case (Wagenmakers et al., 2018). Instead, Bayes Factors affords researchers the opportunity to more rigorously examine the extent to which the data fit the alternate better than the null. Bayes Factor does not aim to accept or reject the null hypothesis; rather it aims to gauge the relative fit of the hypotheses to the data. Effectively, evidence for each hypothesis (in this case, statistical model) is combined in a ratio and offers an estimate that is easy to interpret. For example, if the Bayes Factor is constructed as  $BF_{10}$ , then its ratio is in the direction of the alternative hypothesis (e.g., evidence for the alternative over evidence for the null):  $BF_{10}$  of 5.0 means the alternative model is five times as likely to have generated the data than the null model. Below, we discuss a method for constructing Bayes Factor with MCMC simulation.

### CALCULATING BAYES FACTORS VIA THE ROUTE OF HURDLE MODELING

In this section, we show how to setup and use a hurdle model as a fictional modeling device to compute the Bayes Factor with MCMC simulation. PROC MCMC is a general-purpose MCMC simulation procedure capable of fitting a wide range of models to data, including a two-stage hurdle model. In this case, to calculate the BF to compare an alternate ( $M_1$ ) vs null model ( $M_0$ ), pretend for the moment that, counterfactually speaking, the obtained sample data really came from a two-stage stochastic process in which stage-one outcomes impacted stage-two outcomes. In this imaginary scenario, stage-one outcomes are assumed controlled by a binary (Bernoulli) stochastic process; such that:

$$P(P = p) = \begin{cases} \Phi & p = 1 (M_1) \\ 1 - \phi & p = 0 (M_0) \end{cases}$$

Where  $\Phi$  denotes the prior probability (credibility) that the alternate model generated the sample data apart from seeing sample data. Incidentally, this hurdle model assumes then that the sample was generated from either the alternate or null models. It is common practice to set  $\Phi = 0.5$  (neutral; that is, assign equal prior probabilities between models). However, analysts can set the value of  $\Phi$  to any quantity that they judge to reflect the state of information. The stage-two outcomes, in turn, are controlled by whichever model is selected at stage one ( $M_1$  or  $M_0$ ). Take as an example, a simple logistic regression model such that we wish to test a null version (no effect) and an alternate version (effect).

Now, given a two-stage hurdle model, when we estimate the model using PROC MCMC and request this procedure monitor the parameter  $p$  during MCMC simulation. The MCMC output will then tell us the probability that  $p = 1$ , which means the probability that these data were generated by the alternate model instead of the null model. Once we obtain this probability, then we can calculate the Bayes Factor with this probability such that:

$$\frac{Prob(P = 1)}{1 - Prob(P = 1)} = BF$$

This equation holds when we set  $\Phi = .5$  because then only the relative likelihood of the data given models impacts  $Prob(P=1)$ . In other words,  $BF$  tells us, in this case, how much likelier the data is to be generated by the alternate model than the null model given both models had equal credibility apart from data.

## Handling Survey Weights

To take the extra step of deriving BF with sampling weights, a few words about sampling weights are in order. First, each observation had a sampling weight. These sampling weights reflect proportions in the population of interest. Second, when estimating Bayes Factor with sampling weights, it is important to rescale these sampling weights so that they sum to the actual sample size available for the analysis while still reflecting proportions in the population of interest. This is an important consideration, because if one forgets to rescale sampling weights to sum to the sampling size and weights sum to the population size then BF value (incorrectly) assume the accessible sample size really equals the population size. In fact, this type of rescaling occurs often behind the scenes in other procedures where users can request weights as a default option. To conduct this data pre-processing step, one can rescale sample weights by taking the ratio of sample size (i.e., number of observations) over the sum of original weights and multiply this ratio by the original weights:

$$Rescaled.w_i = Original.w_i * \frac{n}{\sum_{i=1}^n Original.w},$$

where,  $i$  corresponds to observation  $i$ ; and  $n$  corresponds to the number of observations available. By taking this approach it ensures that the sum of the rescaled weights comes to the number of observations available while still reflecting correct population proportions, which is the point of sampling weights. Afterwards, the likelihood function is multiplied by these rescaled weights so that output estimates reflect population proportions and standard errors reflect the actual sample size.

## APPLIED EXAMPLE WITH ANNOATED CODE

Our illustration of estimating Bayes Factor with sampling weights using PROC MCMC comes from the field of Special Education and concerns the transition of adolescents from secondary school to post-secondary education and/or careers. Beginning in 1985, the U.S. Department of Education started an initiative to investigate post-school outcomes, secondary school experiences, and characteristics of student with disabilities being served in the United States. To date, there have been three iterations of the National Longitudinal Transition Study with the first spanning from 1985 to 1990 (NLTS), the second spanning from 2001 to 2011 (NLTS2), and most recently from 2011 to 2014 (NLTS2012). For each iteration, a complex sampling scheme was used which considered locale (e.g., school districts), district enrollment size, and district socio-economic status. These nationally representative datasets are rich with information at the district level, parent perspective, and student perspective. Due to the sensitivity of the data sets, restricted-use licenses can be obtained from <https://nces.ed.gov/pubsearch/licenses.asp>. To illustrate the estimation and use of Bayes Factor using PROC MCMC, we will focus on the most recent iteration, the NLTS2012. In this iteration, participants were sampled from 432 school districts and in total, is comprised of 17,476 students with disabilities spanning all 12 IDEA disability categories stratified by age, as well as 4,483 students without disabilities – for more information about the NLTS2012 see Burghardt et al., (2017).

## CONTEXT

Our specific example stems from a recently published registered report in *Exceptional Children* by Lombardi et al., (2022) which received Open Science Badges for open data, open materials, and for pre-registration of the study and hypotheses – all relevant materials can be found at <https://osf.io/nk3w5/>. In this registered report we were interested in examining whether the probability of receipt of college and career readiness (CCR) supports

differed based on disability type, race/ethnicity, household income, and the intersection of these characteristics. CCR supports were student-reported and were rated (yes/no; or numerically, 1/0): *help completing college applications* (variable: K9c1); *help with course selection* (variable: K9d1), *help reviewing college entrance exam results* (variable: K9e1), and *help arranging college visits* (variable: K9f1). We consulted the district variable (d\_y\_disability) to form nine groups of disability types: no disability, 504 plan, high incidence disabilities (e.g., specific learning disability), multiple disabilities, autism spectrum disorder, traumatic brain injury, sensory disabilities, orthopedic disabilities, and intellectual disability. For race/ethnicity, we used variables G2 (Hispanic, yes/no) and G3\_01 through G3\_05 to create the following six groups: White Hispanic, Black Hispanic, White non-Hispanic, Black non-Hispanic, Hispanic multiracial and other, non-Hispanic multiracial and other. For household income, we used the parent reported variable (p\_h\_income) to form low-, middle-, and high-income household groups. Per the Stage-1 registered report, when evaluating the intersection of all three student characteristics, we were able to use 19 groups.

## ANALYTIC APPROACH

To answer these questions, we first used PROC SURVEYLOGISTIC to model the log-odds of success (i.e., EVENT = '1') which allowed us to account for the CSD of the NLTS2012. We had hypothesized that White students without disabilities from high income households would, on average, have a higher probability of receiving CCR supports than others, therefore, we modeled this group as the reference and inserted indicators representing other groups (e.g., Black students with intellectual disability from low-income households). See below for this code:

```
proc surveylogistic data = dat;
cluster c_apsu;                                /*Specifies Cluster Variable*/
stratum c_astatum;                             /*Specify Stratum Variable*/
model apps(event='1') = asd hiInc id mult ortho plan504 sensory tbi;
weight y_weight_enrolledyouth;                 /*Specify Weight Variable*/
Run;
```

Following estimation of the logistic regression under the null hypothesis significance testing framework, we: (a) examined the significance of the partial log-odds estimates, (b) exponentiated these effects to retrieve the odds ratio, and (c) computed the model-implied probability of receipt for each group. Ultimately, if the odds ratio was not significantly different from unity, then we as researchers are forced to make a binary decision about whether an effect is there or not. As previously stated, leveraging Bayesian estimation and inference avoids the need to make such a zero/sum decision as we are able to report posterior probabilities and accumulate evidence for or against our hypotheses. Specifically, by estimating Bayes Factor we are able to evaluate the predictive merit of the two models (null and alternative), without necessarily rejecting either one. To accomplish this, we used PROC MCMC. Recall that Bayesian inference requires a prior distribution, data (i.e., likelihood), and results in the posterior distribution. To replicate the original model recovered in logistic regression from the NHST framework, we applied an uninformative prior [ $\sim$ NORMAL(0, VAR = 10, LOWER = -10, UPPER = 10)] which forces the Bayesian estimates to be influenced mostly from the data rather than the prior distribution. The following statement fit this model to sample data:

```
PROC MCMC DATA=dat NTU=1000 NMC=500000 THIN=10 PROPCOV=QuaNew SEED=10
MISSING=CC;
by testno;
```

```

PARAM B0 B1;
PRIOR B: ~NORMAL(0, VAR = 10, LOWER = -10, UPPER = 10);
MODEL GENERAL(weight*LOG(PDF("Bernoulli",apps,LOGISTIC(B0 + B1*focal))));
*User specified model statement modified for weights analysis;
ODS TABLE PostSumInt=AppsResult;
RUN;

```

The PROC MCMC statement invokes the procedure and species the input dataset *dat*. The NTU= option specifies the number of tuning iterations, NMC= option specifies the number of posterior simulation iterations. The THIN= option controls the thinning of the Markov chain and specifies that one of every 2 samples is kept. Thinning is often used to reduce the correlations among posterior draws. The PROPCOV=QuaNew controls the construction of the initial proposal covariance matrix (in this case, we requested quasi-Newton optimization). The SEED option specifies a seed for the MCMC simulation for reproducibility.

The PARMS statements identify two user-named parameters, B0 and B1. The PRIOR statement assigns a diffuse normal prior to these parameters. For better computation, we bounded the prior to be between -10 and 10 to rule out absurd values. The MODEL statement defines the response model for the data. In this case, we used the GENERAL option, which allows users to define their own response model with weights. By multiplying the loglikelihood of each observation by their sample weight parameter estimates will reflect population proportions while the certainty of inferences will still reflect sample sizes.

We then grouped estimates into either a positive effect or negative effect group to compute the Bayes Factor. This was necessary to give focus to the Bayes Factor, allowing us to examine evidence for or against the estimated effect. We determined that the smallest effect on the logistic scale we would be interested in finding was  $|.18|$ . Our rationale was that this effect is mathematically equivalent to  $d=|.1|$ , which is commonly considered to be the border line between a value being practically equivalent to the null and a value being practically different from the null. Therefore, in this case, when the frequentist estimated effect was negative, we set the alternative hypothesis (H1) to -0.18; and when the frequentist estimated effect was positive, H1 was 0.18; meanwhile H0 was fixed at zero, regardless of the direction of the effect. The parameter that corresponds to the probability of the alternative model is Phi and this parameter is given the following prior distribution of  $[\sim \text{BINARY}(0.5)]$ . The mean of posterior distribution of Phi corresponds is the estimate of Bayes Factor.

```

%LET N=50000; *NMC;
%LET H0=0; *Null Hypothesis for BF Analysis (Note: Single Point Hypothesis);
%LET H1=-0.18; *Alternate Hypothesis for BF Analysis (Note: Single Point Hypothesis);

PROC MCMC DATA=appsneg NTU=1000 NMC=500000 THIN=10 PROPCOV=QuaNew SEED=10
MISSING=CC; *Binary-Logistic Model;
by testNo;
PARAM Phi 0 B0;
PRIOR Phi ~ BINARY(0.5);
PRIOR B: ~NORMAL(0,VAR = 10, LOWER = -10, UPPER = 10);
/* MODEL
GENERAL(y_weight_enrolledyouth*LOG(PDF("Bernoulli",apps,LOGISTIC(B0 + ((1-Phi)*&H0 + Phi*&H1)*sensory_white_hi)))); *BF for Alternate;*/
MODEL GENERAL(weight*LOG(PDF("Bernoulli",apps,LOGISTIC(B0 + ((1-Phi)*&H0 + Phi*&H1)*focal)))); *BF for Alternate;

ODS TABLE PostSumInt=BF_neg;

```

```
RUN; *Approximate BF value;
```

The PARM statement identifies two parameters, Phi and B0. Phi is optionally giving an initial value of 0 and is used to create the artificial two-stage hurdle model needed to compute the Bayes Factor with sampling weights. The PRIOR statement assigns Phi a Binary prior distribution with probability of .5 (which means there is no prior preference for alternate over the null. Both are equally likely to generate the sample). The other parameter, b1, is the intercept. It is given a diffuse prior. The Model Statement uses the GENERAL option to allow the user to specify the response model. In this case, we modify the response model with sample weights to include the parameter Phi, which by randomly taking values of 0 and 1 in MCMC simulation, will, in this case, shift the response model between the alternate and null versions in proportion to the likelihood the data was generated by those models. Hence, the posterior of Phi is the probability the data came from the alternate rather than the null (assuming each prior probability for each model). As shown next, some postprocessing of the posterior sample can convert this information into the Bayes Factor:

```
DATA BF_neg; SET BF_neg; *Calculate BF value;
IF Parameter='Phi';
IF Mean=0 THEN BF=0; *Zero BF indicates inexact BF value but can infer value
is extremely favorable to null;
IF 0 < Mean < 1 THEN BF=Mean/(1-Mean);
IF MEAN=1 THEN BF=&N; *BF = N indicates inexact BF value but can infer value
is extremely favorable to alternate;
RUN;
PROC PRINT DATA=BF_neg NOOBS;
TITLE "Bayes Factor Analysis";
VAR BF;
FOOTNOTE "Model";
RUN; TITLE; FOOTNOTE;
```

In this DATA STEP, the posterior sample for parameter Phi is used to create the Bayes Factor with sample weights. One important thing to note is that IF-THEN statements in the DATA STEP are used to safeguard against instances when the alternate and null fit the data exceptionally better than the other and so MCMC simulation never selects the opposite model. In these extreme cases, PHI becomes either 0 or 1, which makes it impossible to estimate the exact BF with the formula because of a possible division by zero problem. To compensate, we automatically set the BF to be 0 if the Mean of Phi = 0 and to an arbitrarily high value &N (say 100) if Mean=1. However, if Mean is between 0 and 1 (as it normally will be the probability of one model does not overwhelm the probability of the other) then we use the formula to compute a precise BF.

## RESULTS

We now present a subset of results from both the frequentist NHST framework and the Bayesian framework. Additionally, we expose the differences in inferences from the two methodological approaches. We focus on the outcome variable regarding help with college applications.

### Frequentist Inference

The model-implied probability for White students from high income household without disabilities was .29 (Est = -0.88, SE = 0.14,  $p < 0.001$ ). Of the 18 effects, two were statistically significant: White students with sensory disabilities from high income

households ( $Pr = .16$ ,  $Est = -0.75$ ,  $SE = 0.31$ ,  $p = 0.017$ ;  $OR = 0.47$ ) and Black students without disabilities from low-income households ( $Pr = .45$ ,  $Est = 0.66$ ,  $SE = 0.31$ ,  $p = 0.033$ ,  $OR = 1.93$ ). Additionally, White students with multiple disabilities from high income households were found to have a model-implied probability of .16 (versus .29); however, their partial log-odds were not significantly different from the reference group ( $p = 0.08$ ).

## Bayesian Inference

Turning to Bayesian methods, we can gauge the amount of evidence for each of these effects. For instance, the difference in probabilities found between the reference group and White students with sensory disabilities from high income households (.29 versus .16) was found to be 5.39 times more likely to have generated the data than the null model. Similarly, the difference in probabilities between the reference group and Black students without disabilities from low-income households (.29 versus .45) was identified as the most likely source of the data and this model was 4.72 times more likely. In both of these cases, frequentist and Bayesian inference arrive at the same decision, with the latter affording a richer interpretation of the effect. On the other hand, the alternative model in which a difference in probabilities between the reference group and White students with multiple disabilities from high-income households exists was found to be 4.48 times more likely than the null model in which there was no difference in probabilities. In this example, frequentist and Bayesian inference come to different conclusions where the former tells us there is no difference based on the reject/fail to reject options of the NHST framework; whereas, the Bayes Factor suggests that true difference in probabilities is likely illustrating that it is more sensitive to differences when group sample sizes are small.

## CONCLUSION

This paper set out to accomplish two main objectives. First, we wanted to elucidate the difference between sampling designs (simple versus complex) and to advocate for use of sampling weights when available to produce the most meaningful results as possible given the population at hand (assuming inferences from sample to population is the aim). Second, we wanted to compare and contrast Bayesian and frequentist methods to illuminate how well they converge when using flat (or un-informative) priors so that Bayesian estimation can closely match results produced by frequentist estimation. This convergence, in turn, , affords researchers the opportunity to estimates Bayes Factors and bring Bayesian testing to supplement p-value tests. Third, we sought to highlight the use of Bayes Factors to gain a deeper understanding of p-value test results, relative to 0/1 decisions inherent in NHST from the frequentist perspective. Finally, we set out to demonstrate that it is feasible to estimate Bayes Factors in the context of complex sampling designs using PROC MCMC. Although our applied data example considered logistic regression, Bayes Factors can be estimated for other statistical models such as linear regression, mixed effect models, and latent variable models; this is made possible by invoking the GENERAL statement in the MODEL command where it is possible to call the needed functions to correspond to the model at hand.

## REFERENCES

1. Burghardt, J., Haimson, J., Lipscomb, S., Liu, A.Y., Potter, F., Waits, T., & Wang, S. (2017). National Longitudinal Transition Study 2012 design documentation (NCEE 2017-4021). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.
2. Lombardi, A. R., Rifenbark, G. G., Hicks, T., Taconet, A., & Challenger, C. (2022). College and Career Readiness Support for Youth With and Without Disabilities Based

on the National Longitudinal Transition Study 2012. *Exceptional Children*, 00144029221088940.

3. Wagenmakers, E. J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., ... & Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic bulletin & review*, 25(1), 35-57.

## ACKNOWLEDGMENTS

This is the text for the acknowledgments. This paragraph uses the PaperBody style, which uses the Verdana font, not the Arial font.

## CONTACT INFORMATION

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

Tyler A. Hicks, Ph.D.  
University of Kansas  
[tahicks@ku.edu](mailto:tahicks@ku.edu)

Graham G. Rifenbark, Ph.D.  
University of Connecticut  
[graham.rifenbark@uconn.edu](mailto:graham.rifenbark@uconn.edu)

