

Mixture Priors 101: Using SAS® to Obtain Powerful Frequentist Inferences with Bayesian Methods

Tyler Hicks, University of Kansas, Lawrence, KS
Jeffrey D. Kromrey, University of South Florida, Tampa, FL

ABSTRACT

Solely based on background information, scientists know which values for parameters are more probable than others. In Bayesian statistics, these judgements are called prior probabilities. Analysts can use Bayesian methods to obtain powerful frequentist inferences with prior probabilities. However, if prior probabilities are misspecified, they can wreak havoc on inferences. A mixture prior density provides a tool to make inferences more robust to specification errors in prior probabilities. The purpose of this paper is to demonstrate that setting mixture priors is easy using PROC MCMC. Researchers can use PROC MCMC to implement mixture priors to execute powerful statistical inferences. This paper provides a rationale for mixture priors, presents annotated PROC MCMC code, and then concludes with an example.

Keywords: PROC MCMC, MIXTURE PRIORS, ROBUSTNESS, POWER

INTRODUCTION

Researchers can produce frequentist or Bayesian modes of argumentation to justify important statistical inferences (Hacking, 2001). For example, in a frequentist mode, they may justify a formula for computing a 95% interval estimate because it yields intervals which enclose the true parameter 95% of the time. Alternatively, in a Bayesian mode, they may justify the same formula because of its conformity to a timeless logical rule of induction (i.e., Bayes' rule). Statisticians debate about which mode of justification is superior. Some frequentists argue that frequencies are more objective than logical intuitions. However, Bayesians, such as John Keynes (1923), have countered that "we are all dead in the long run" (pg. 80). A brief rehearsal then of how Bayesians justify estimates is in order.

Bayesian statisticians select among rival estimates for a parameter by balancing posterior probabilities, $P(\theta|Y)$. This notation is read as the probability of θ given Y , wherein θ denotes a parameter and Y sample data. Figure 1 depicts a normal-shaped posterior density for θ . Its center represents the most credible estimate for θ ; its spread, uncertainty. Studying Figure 1, one can see that the value 50 holds more posterior probability than 25 or 75, respectively. Summary statistics from a posterior density, such as the quartile values, can provide researchers with handy point and interval estimates for parameters.

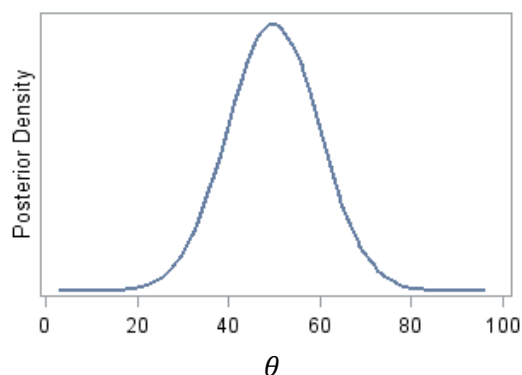


Figure 1. Example posterior density generated with PROC MCMC

The rule for obtaining posterior probabilities, Bayes' rule of conditionalization, is the eponymous formula of Bayesian statistics:

$$P(\theta|Y) \propto P(Y|\theta)P(\theta) \quad (1)$$

This rule states that the posterior probability is proportional to the product of the likelihood, $P(Y|\theta)$, and the prior probability, $P(\theta)$. The likelihood, which is the probability of the data given a parameter value, can be found using an algorithm (i.e., likelihood function). However, setting up the prior probability, which is the probability of the parameter value apart from sample data, requires some artistic judgement.

Special prior densities, called *power priors*, encode background knowledge about parameters (e.g., a variance cannot have a negative value) to facilitate the recovery of priors from limited sample information. A power prior, denoted below as $P_p(\theta)$, dominates the posterior when the sample size is sufficiently small. Accordingly, when power priors turn out to be misspecified (i.e., the expectations about parameters encoded in them are inaccurate), Bayesian methods can yield misleading posteriors: Garbage in, garbage out.

Neutral priors can provide a convenient substitute for a power prior when background knowledge is too limited to set a power prior. Jeffreys priors are perhaps the quintessential neutral prior (Jeffreys, 1946). A Jeffreys prior, denoted below as $P_J(\theta)$, is proportional to the square root of the determinant of Fisher's Information Matrix for the data:

$$P_J(\theta) \propto \sqrt{\det I(\theta)} \quad (2)$$

Jeffreys priors satisfy the *Invariance Principle*, which states that information yielded by posteriors should be constant across monotonic transformations of the parameter [e.g., Inferences should not depend upon whether the Jeffreys prior was set on θ or, for example, its natural log, $\ln(\theta)$].

Historically, methodologists justified settings priors with arguments consistent with the the Bayesian paradigm (Howson & Urbach, 2006). Yet, it is possible to integrate prior probabilities into a frequentist paradigm, if their implementation yields estimates with desirable frequency properties. These Bayesian-frequentist hybrid procedures can help statisticians capitalize on priors while staying solidly rooted in the frequentist paradigm. The mixture prior, which merges a power and neutral prior into one prior, can be an excellent tool for building Bayesian-frequentist hybrid procedures as it can improve upon – or, at least preserve – a statistical procedure's frequency properties.

MIXTURE PRIORS

To blend a power and Jeffreys prior into a mixture prior, denoted below as $P_m(\theta)$, a new parameter, say φ , can be introduced into the model. It takes on two possible values, 1 and 0. The conditional probability of θ given φ is:

$$P_m(\theta, \varphi) \begin{cases} P_J(\theta) & \text{if } \varphi = 1 \\ P_p(\theta) & \text{if } \varphi = 0 \end{cases} \quad (3)$$

The parameter φ also needs a prior density. If a binary prior is set on φ [i.e., $\varphi \sim B(p_0)$] then p_0 can be given a high value, such as 0.8, 0.9, or .95, to indicate that the researcher is sure the power prior is specified correctly, or p_0 can be given an impartial value, such as 0.5, in order not to privilege either prior option.

It can be exceedingly difficult to analytically derive a posterior density using a mixture prior, but it can be empirically derived using Markov Chain Monte Carlo (MCMC) simulation. The MCMC procedure expanded Bayesian modeling capabilities in the SAS programming environment. It uses a random walk Metropolis algorithm to simulate posterior samples in order to recover posterior parameters. The logic of this simulation technique has been amply described for the newcomer elsewhere (Kruschke, [2011, Chapter 7]). For present purposes, it suffices to recall that at each step in the random walk an observation is drawn from the posterior density. This creates a "chain" of simulated observations.

At each step in the random walk, either a 1 or 0 value is drawn for φ . If $\varphi=0$, then the θ value will be drawn from $P_J(\theta)$ and used to derive the posterior sample. However, if $\varphi=1$ then a θ value will instead be drawn from $P_p(\theta)$. If $\varphi = 1$ on every draw then the posterior sample would be equivalent to the one which would have been reached using only $P_p(\theta)$. if $\varphi = 0$ on every draw then the posterior sample would be equivalent to the one which would have been reached using only $P_J(\theta)$. If the value of φ fluctuates back and forth between 1 and 0 then the posterior sample will be a mixture of the posterior samples derivable from those two priors. Table 1 illustrates a posterior sample for a parameter (θ) produced by a mixture prior.

Records	MCMC Draws from Posterior Densities			
	$\theta \sim P_p(\theta)$	$\theta \sim P_J(\theta)$	$\varphi \sim B(0.5)$	$\theta \sim P_m(\theta)$
1	4.2178	9.5215	1.0000	4.2178
2	4.2178	5.6897	0.0000	5.6897
3	4.3703	7.7248	1.0000	4.3703

Table 1. Example of the first three records from a posterior sample drawn from a posterior density derived with a mixture prior generated using MCMC simulation methods

Building the Prior

As previously mentioned, to construct a mixture prior one must merge a power prior and Jeffreys prior together. To illustrate building a mixture prior with an example, below is a brief description of a quasi-experimental study where researchers sought to quantify the effect of an intervention using a mixture prior.

Dataset

The original education dataset used in this example had 106 records but summary statistics will suffice for analysis. The independent variable was a Multi-Tiered System of Supports (MTSS). SWIFT Center, the National Center on Inclusive School Reform at the University of Kansas, champions MTSS as a tool for building the capacity of the education system to support all students, with and without disabilities, in fully integrated schools. The dependent variable was literacy performance, as measured by the annual change in score on the state literacy test. Students from a school with MTSS and a school without MTSS were put into pairs using propensity score matching on key demographic and academic variables and their difference scores computed. The average difference score was the obtained test statistic.

```
DATA MTSS;
  INPUT N Mean StdDev StdErr;
  cards;
  106 5.6698 10.4030 1.0104
  ;
```

“N” was the sample size. “Mean” was the average difference score among observed matched pairs of students. “StdDev” was the observed standard deviation of difference scores. “StdErr” was the standard error for the mean.

A Bayesian-frequentist One-Sample t-test

Let $\bar{Y}_d$ denote the average observed difference in literacy performance among matched pairs of schoolchildren in the quasi-experimental study. Construed as a random variable, the distribution of $\bar{Y}_d$ follows a normal distribution. If, for expediency, the sample variance (S_d^2) substitutes for the population standard deviation, $\sigma_{\bar{Y}_d}$, then $\bar{Y}_d$ can be reasonably modeled as t -distributed:

$$\bar{Y}_d \sim t(\mu_{\bar{Y}_d}, \sigma_{\bar{Y}_d}, v) \quad (4)$$

where $\mu_{\bar{Y}_d}$ is the average effect, $\sigma_{\bar{Y}_d}$ is the standard error, and v is the degrees of freedom.

In this model, only one parameter needs to be recovered from sample data —namely, the average effect (μ_d). The standard error ($\sigma_{\bar{Y}_d}$) and degrees of freedom (v) are each a function of sample data:

$$\sigma_{\bar{Y}_d} = S_{\bar{Y}_d} / \sqrt{N} \quad v = N - 1 \quad (5)$$

where $S_{\bar{Y}_d}$ is the sample standard deviation and N the sample size.

To proceed with a One-Sample t -test, statisticians set a mixture prior on μ_d and compute its posterior density. Next, they abstract away a point and 95% interval estimate from the posterior density. The median of the posterior density can be the point estimate. The interval which excludes 5% of the posterior density with the lowest density can be the 95% interval estimate. However, statisticians must obtain a power prior and a Jeffreys prior to make a mixture prior.

A power prior for the effect can be elicited from an expert through a structured interview. Ask the expert to identify the lowest and highest plausible values for the effect in standard deviation units (i.e., Cohen's Δ). Let Δ_L and Δ_U denote the expert's chosen lowest and uppermost values for Δ . Next, compute an upper and lower boundary for the effect in the raw metric using the expert's judgements:

$$\mu_{d_L} = \Delta_L(S_{\bar{Y}_d}) \quad \mu_{d_U} = \Delta_U(S_{\bar{Y}_d}) \quad (6)$$

where $S_{\bar{Y}_d}$ is the sample standard deviation statistic, μ_{d_L} is the lower boundary for the effect in the original scale (μ_d), and μ_{d_U} is its upper boundary. Given obtained values for μ_{d_L} and μ_{d_U} , proceed to specify a normal-shaped power prior using formulas (6) and (7) below to derive its parameters, the mean (μ_0) and standard deviation (σ_0):

$$\mu_0 = (\mu_{d_U} + \mu_{d_L}) / 2 \quad (7)$$

$$\sigma_0 = \text{Round} \left[(\mu_{d_U} - \mu_0) / 3 \right] \quad (8)$$

where the standard deviation (σ_0) of the power prior is rounded to the nearest integer or half-integer (i.e., 0.5, 1, 1.5, 2, 2.5...) for convenience.

To illustrate setting up a power prior through an interview, suppose the expert expressed confidence the effect was moderate. Using standard conventions for interpreting Δ , a moderate effect falls between $\Delta_L = 0.5$ and $\Delta_U = 0.8$. So, one can translate the expert's judgements about the effect back into a raw metric using formula (6). If the obtained standard deviation was 10 units then $\mu_{d_L} = 5$ and $\mu_{d_U} = 8$. Next, using formula (7) and (8) one could obtain the following power prior $P_p(\theta) \sim N(6.5, 0.5^2)$. The mean is 6.5 ($= [(5 + 8)/2]$) and the standard deviation is 0.5 ($= \text{rounded } [([6.5 - 5]/3)]$).

The Jeffreys prior for the effect can be computed with equation 2. The ability to specify Jeffreys priors with formulas makes setting them a more objective process than power priors which must be elicited from experts. Berger (1985) showed that a Jeffreys prior for an effect is equivalent to a non-normalized uniform prior:

$$P_j(\mu_d) \propto 1 \quad (9)$$

The Jeffreys prior depicted in equation 9, although perhaps not a realistic portrait of any real expert's opinion, can be said to exemplify the *Principle of Insufficient Reason*. This principle states when in doubt assign equal probabilities to every effect possibility.

The PROC MCMC Code

For computational efficacy, data analysts can model the sample mean rather than all 106 records because the mean perfectly summarizes all relevant sample information for a One-Sample *t*-test. It is the "sufficient statistic." The code for requesting the required MCMC procedure is as follows:

```
ODS GRAPHICS ON;
PROC MCMC DATA=MTSS NBI=50000 NMC=100000 MONITOR=(PHI MU); *Controls MCMC chain;
  PARM MU1 0 MU2 0 PHI 1;
  PRIOR PHI ~ BINARY(0.5); *P(φ)~Bin(0.5);
  PRIOR MU1 ~NORMAL(6.5, SD=0.5); * Pp(μ)~N(6.5, 0.52);
  PRIOR MU2 ~GENERAL(0); * Pj(μ)∝1;
  MU=(PHI*MU1)+(1-PHI)*MU2; * Programing code to generate PM(μ,φ);
  MODEL mean ~T(MU,SD=StdErr, N-1); *One-Sample t-test with mixture prior;
RUN;
ODS GRAPHICS OFF;
```

The lines in this code can be understood as follows:

```
PROC MCMC DATA=MTSS NBI=50000 NMC=100000 MONITOR=(PHI MU) (A)
```

Line A requests a posterior sample with 100,000 records using MCMC simulation (**NMC=100000**). The chain runs for 150,000 iterations in total but to reduce autocorrelation in the sample the first 50,000 records are discarded (**NBI=50000**). "NBI" denotes the "number of burn-ins" (e.g., discarded cases). Line A also requested only output for the parameters ϕ and μ (**MONITOR=(PHI MU)**).

```
  PARM MU1 0 MU2 0 PHI 1; (B)
```

Line B declares which parameters need to be estimated and sets a starting value in the MCMC chain for them.

```
  PRIOR PHI ~ BINARY(0.5); *P(φ)~Bin(0.5);
  PRIOR MU1 ~NORMAL(6.5, SD=0.5); * Pp(μ)~N(6.5, 0.52);
  PRIOR MU2 ~GENERAL(0); * Pj(μ)∝1;
  MU=(PHI*MU1)+(1-PHI)*MU2; * Programing code to generate PM(μ,φ); (C)
```

Line C specifies priors for each declared parameter. The prior for PHI is set to be neutral as 0.5 favors neither the Jeffreys nor power prior. The prior for MU1 is a power prior. It is the power prior elicited from an expert in the previous section. The prior for MU2 is a neutral Jeffreys prior. The last line of code (i.e., **MU=(PHI*MU1)+(1-PHI)*MU2**;) is the key to obtaining the requisite mixture prior for the focal parameter, MU.

```
  MODEL mean ~T(MU,SD=StdErr, N-1); (D)
```

Line D is the model statement. It species the likelihood function for the sample mean. As expected, it models the sample mean as *t*-distributed with a mean of MU and standard deviation equal to the obtained standard error (**SD=StdErr**) and degrees of freedom (**N-1**). In sum, the above PROC MCMC code constitutes a One-Sample *t*-test with a mixture prior appropriated for paired data.

Results

The PROC MCMC output for the example is reported below.

Posterior Summaries and Intervals					
Parameter	N	Mean	Standard Deviation	95% HPD Interval	
PHI	100000	0.9998	0.0141	1.0000	1.0000
MU	100000	6.3361	0.4437	5.4873	7.2254

The posterior parameters for φ (M=0.9998; SD=0.0141) indicate that the posterior sample generated for μ through the mixture prior favored the power prior over the neutral prior.

Based on these results, one can say that their best guess given the evidence is that MTSS will improve the annual state literacy scores for schoolchildren is approximately 6 (i.e., $\mu_d = 6.3361$). One can be 95% sure that the actual improvement is approximately between 5 to 7 points, 95% CI [5.4873, 7.2254]. As expected, these findings agree with those of the traditional One-Sample *t*-test (except the traditional test yielded a slightly wider 95% CI, [3.6953, 7.7083]).

In the above example, a One-Sample *t*-test performed with a mixture prior yielded results comparable to a traditional One-Sample *t*-test. In the next section, it is demonstrated through simulation studies that mixture priors can improve the procedure's frequency properties given certain conditions. Statisticians thus have good reason to supplement a traditional One-Sample *t*-test with mixture priors.

SIMULATION STUDY

A simulation study was conducted to empirically verify that the One-Sample *t*-test with a mixture prior is powerful and controls Type I error. This section first describes the design of the simulation and then discusses the results.

Design

This simulation study utilized a 3 X 4 X 2 factorial design. The factors considered were (a) the prior (i.e., mixture, Jeffreys, and power prior), (b) sample size (i.e., 50, 100, 150, 200), and (c) Effect size (i.e., null, large). For each of the 24 conditions, 5,000 samples were simulated using the rand functions in SAS (SAS Institute Inc, 2008). Regardless of estimation method, the model shown in Equation 4 was fitted to the generated data and estimated using PROC MCMC. A posterior sample with 100,000 records was always requested for a procedure. The chain ran 150,000 iterations but to reduce autocorrelation the first 50,000 records were discarded. On the basis of a liberal criterion for Type I error control suggested by Bradley (1978), control was defined as a procedure's Type I error rate fell between .025 (= .05-.025) and .075 (= .05+.025).

For the sake of expediency, the three testing procedures are denoted by the relevant priors (i.e., "Mixture *t*-test," "Jeffreys *t*-test," and "Power *t*-test"). The Jeffreys *t*-test behaves identically to a standard Student's One-Sample *t*-test.

Point estimates

Figure 2 (pg. 6) provides a summary of the frequencies properties of the point estimates yielded by the three priors under the two effect conditions (i.e., condition 1: large population effect; Condition 2: null population effect). Bias assesses the degree to which the average point estimate fails to match the true parameter. The Root Mean Squared Error (RMSE) assesses the inconsistency of repeated point estimates. Ideally, point estimates should have minimal bias and RMSE.

Based on the information in Figure 2, one can observe that when the power prior was specified correctly (i.e., when the researchers expected a large effect and there really was a large effect) all three priors yield unbiased estimates but the mixed and power priors yielded more consistent estimates in repeated applications (i.e., lower RMSE). When the power prior was specified incorrectly (i.e., when the researchers expected a large effect but there was really no effect in the population) then the power prior yielded biased estimates. However, as sample size rose, the mixture prior began to, on average, deliver unbiased estimates. Given a sample size of 50 or greater, the mixture prior performed like the Jeffreys prior when the power prior was misspecified (otherwise, it outperformed the Jeffreys prior when the power prior was correctly specified).

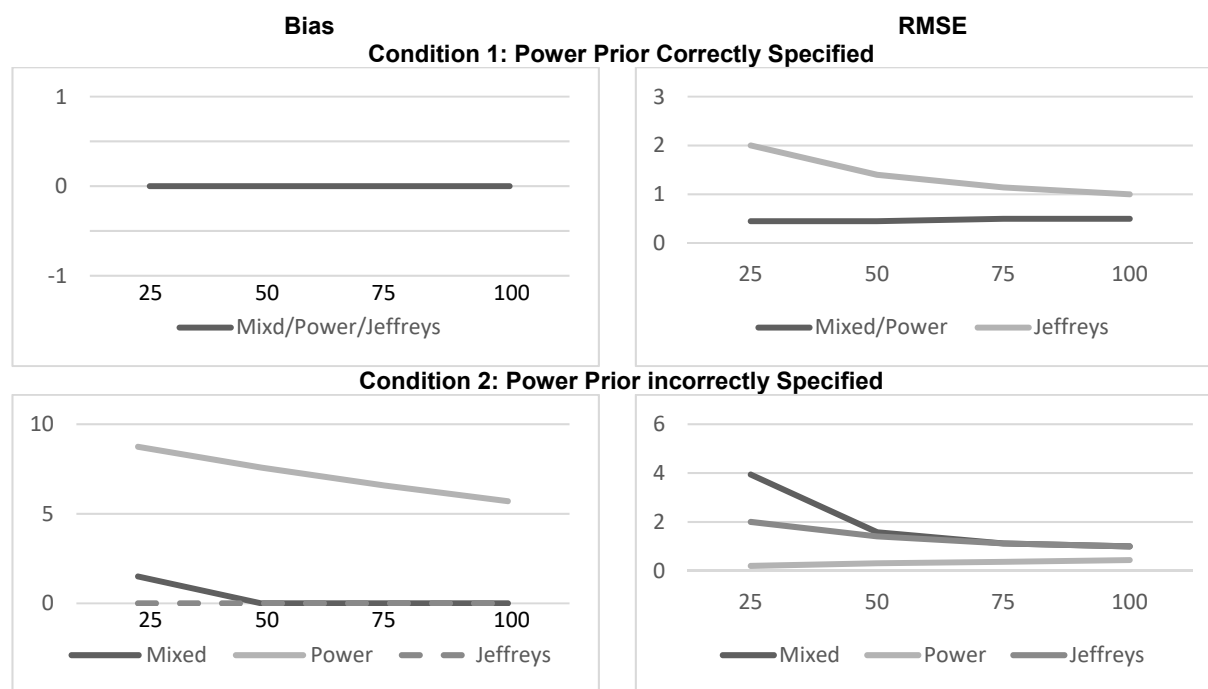


Figure 2. Frequency properties of the three priors' point estimates

95% Interval estimates

Figure 3 provides a summary of the frequencies properties of 95% interval estimates yielded by the three priors under the two effect conditions. Coverage assesses the accuracy of the interval estimates. Precision assesses how narrow the intervals were on average. Ideally, 95% interval estimates should be narrow and accurate.

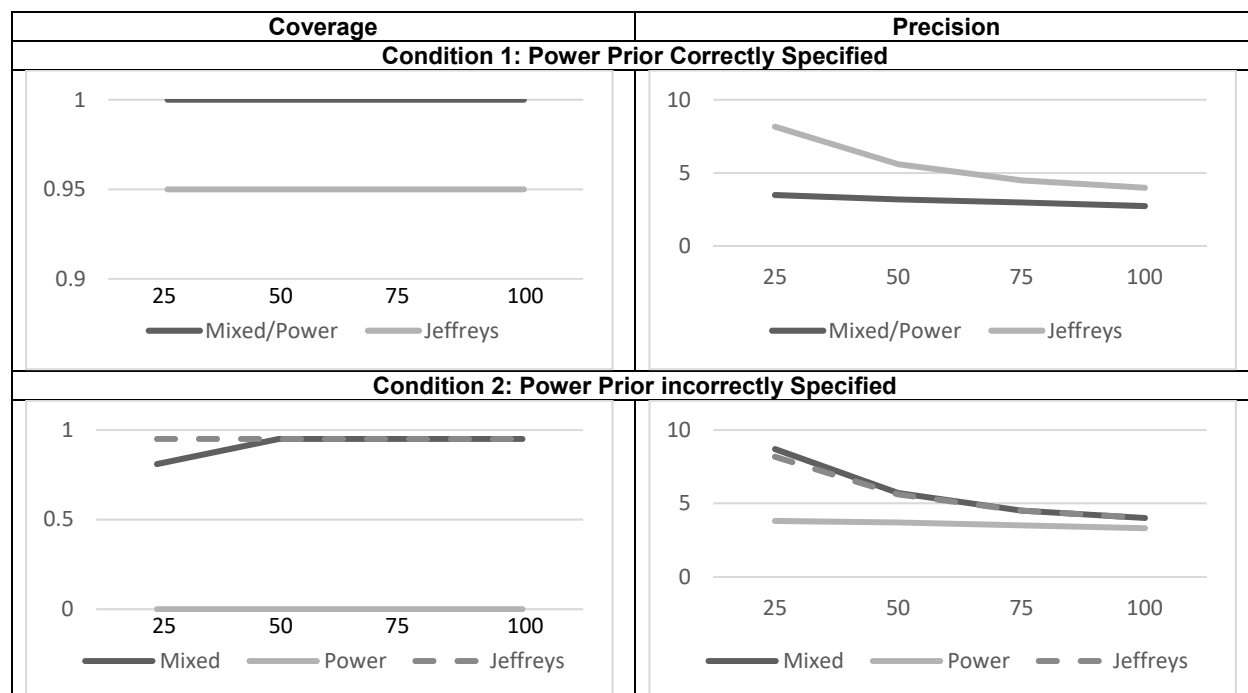


Figure 3. The frequency properties of point estimates of the different priors

Based on these results, when the power prior was specified correctly the mixture and power priors yielded intervals which were more precise and accurate than those of the Jeffreys prior. When the power prior was specified

incorrectly then the behavior of the mixture and power prior deviated. The power prior was always inaccurate in this case. The coverage of the mixture prior, in contrast, was comparable to that of the Jeffreys prior.

A 95% interval estimate with a 100% coverage rate, although desirable if yielded by a Bayesian-frequentist hybrid procedure, can be very confusing for a statistician trained in the frequentist paradigm. But procedures which use prior probabilities can produce 95% intervals with a coverage rate different than 95% without conceptual difficulties. This is possible because the 95% in this case is defined by the proportion of the posterior density the interval encloses rather than its coverage rate. This definition of 95%, of course, only applies if the interval was produced by a procedure implementing prior probabilities. A procedure without prior probabilities, in contrast, must produce 95% intervals with a 95% coverage rate (Otherwise, there is no way to meaningfully define 95% in the 95% interval).

CONCLUSIONS AND IMPLICATIONS

This Monte Carlo simulation study demonstrates that mixture priors can improve the frequency properties of a procedure's estimates. Statisticians who are frequentist in their statistical philosophy select procedures by balancing frequency properties. A mixture prior should copy the behavior of the analogous power prior when the power prior is specified correctly and thereby outperform a standard neutral prior in regards to its point and interval estimates. However, when the power prior is incorrectly specified, a mixture prior should copy the behavior of the standard prior and thereby outperformed the power prior in regards to its point and interval estimates.

Bayesian statisticians championed the prior. But even researchers who are more partial to the frequentist paradigm can assimilate priors in their work by mixing neutral and power priors. Because the mixture prior blends a neutral and power prior together its long run performance is stable and predictable. One therefore does not need to borrow from a Bayesian paradigm to justify it (e.g., appeal to logical intuitions). Instead, one can show that the method works in the long run in order to justify it. Researchers can use mixture priors to integrate Bayesian methods into the frequentist paradigm.

The key advantage of a mixture prior is that it allows researchers to evaluate its frequency properties. For instance, if researchers carrying out an intervention study anticipated a large effect then they could set up a mixture prior on the parameter to conduct a One-Sample *t*-test and evaluate its frequency properties through simulation methods. They can evaluate its performance relative to standard procedures given small, medium, and large effects. This is possible because mixture priors have stable frequency properties.

REFERENCES

- Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis* (2nd ed.). New York, NY: Springer-Verlag.
- Hacking, I. (2001). *An introduction to probability and inductive logic*. New York: Cambridge University Press.
- Howson, C., & Urbach, P. (2006). *Scientific reasoning: The Bayesian approach* (3rd ed.). Chicago: Open Court.
- Jeffreys, H. (1946). An invariant form of the prior probability in estimation problems. *Proceedings from the Royal Society of London*, 186(1007), 453-461. doi:10.1098/rspa.1946.0056
- Keynes, J. (1923). *A track on monetary reform*. Toronto, ON: MacMillan.
- Kruschke. (2011). *Doing Bayesian data analysis: A tutorial with R and BUGS*. New York, NY: Elsevier.

Contact Information

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

Tyler Hicks, Ph.D.
University of Kansas
122 West Campus Rd,
Joseph R. Pearson Hall, Rm 521.
Lawrence, Kansas 66045
Email: tahicks@ku.edu

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.

SWIFT Center produced the data for this presentation under U.S. Department of Education, Office of Special Education Programs Grant No. H326Y120005. OSEP Project Officers Grace Zamora Durán and Tina Diamond served as the project officers. The views expressed herein do not necessarily represent the positions or policies of the Department of Education. No official endorsement by the U.S. Department of Education of any product, commodity, service or enterprise mentioned in this publication is intended or should be inferred.