National
Council on
Compensation
Insurance, Inc.

Facts and Fiction of Statistical Significance

Overview

- Issues Surrounding Reported Statistical Evidence
- Misinterpretation of *p*-Values
- The Fair Coin
- Calibrating *p*-Values as Measures of Evidence
- Digression: Bayesian Meta-Analysis
- Variable Selection
- References
- Appendix

The literature on issues surrounding classical significance and hypothesis testing is extensive. It is not the purpose of this presentation to provide a comprehensive overview of the literature. Instead, it is intended to point out the problem in a parsimonious way and then offer solutions for decision-makers and applied researchers

Public Discourse Surrounding Reported Statistical Evidence

- In recent years, a host of articles in newspapers and news magazines have drawn attention to issues related to reported statistical evidence
- Among these articles are Matthews [8] (*Financial Times,* 2004), *The Economist* [12] (2007), Freedman [4] (*The Atlantic*, 2010), Lehrer [7] (*The New Yorker,* 2010), and Carey [2] (*The New York Times*, 2011)
- Further, the U.S. Supreme Court in MATRIXX INITIATIVES, INC., ET AL. v. SIRACUSANO ET AL., on March 22, 2011, decided in a case on side effects of pharmaceuticals that the "premise that statistical significance is the only reliable indication of causation is flawed" (*)

* http://www.supremecourt.gov/opinions/10pdf/09-1156.pdf, p. 2. The Supreme Court decision raises the question of decision-making in the presence of outcomes for which the null hypothesis ("no effect") cannot be rejected. The error rate in rejecting the null hypothesis when it is true (type II error rate) is rarely available

Issues Surrounding Reported Statistical Evidence

- The *p*-value obtained in classical hypothesis testing (as developed by Neyman and Pearson) is not a measure of evidence against the null hypothesis
  - It can be shown that a *p-value* of about 0.05 (e.g., 0.05 after rounding) is about as likely to come from the null as from the alternative
- Multiple comparisons
  - For a pre-selected type I error rate α (of, for instance, 5 percent), the more tests are done, the more likely there is a chance outcome of $p \leq \alpha$, which is then reported as statistically "significant"
- Magnitudes of small effects tend to be overestimated
  - Unusually large effects are more likely to generate an outcome $p \leq \alpha$, particularly in small samples where the standard errors are high, whereas effects closer to what is ordinarily observed remain unreported (due to a lack of statistical significance)

The above list of issues surrounding reported statistical evidence is not meant to be exhaustive

Implications of Overstated Evidence

- The exaggeration of the evidence associated with common interpretations of *p*-values as evidence against the null contributes to false positives
  - In biomedical research, for instance, false positives may lead to recommendations for medical therapies that ultimately prove ineffective
- Results obtained in multiple comparisons are frequently not replicable due to being chance outcomes
  - The Bonferroni correction is one of several approaches to adjust classical hypothesis testing for multiple comparisons
- When studies are replicated, the magnitudes of the measured effects tend to be smaller, especially when the original study is based on a small sample

An Example from the Medical Sciences

- In 2005, John P.A. Ioannidis published an article in the *Journal of the American Medical Association (JAMA)* where he examined "all original clinical research studies published in 3 major general clinical journals or high-impact-factor specialty journals in 1990-2003 and cited more than 1000 times in the literature" (*)
- Of the resulting 49 highly cited original clinical research studies, 45 claimed that the intervention was effective; of these 45 studies, 11 (24 percent) remained "largely unchallenged"
- Of the 34 studies that were challenged by subsequent research, 7 (21 percent) were contradicted, another 7 (21 percent) had the magnitudes of their effects scaled back, and 20 (59 percent) were replicated

* Ioannidis, John P.A. "Contradicted and Initially Stronger Effects in Highly Cited Clinical Research," *Journal of the American Medical Association (JAMA),* Vol. 294, p. 218–228, 2005, http://www.givewell.org/files/methods/Ioannidis%202005-Contradicted%20and%20Initially%20Stronger%20Effects%20in%20Highly%20Cited%20Clinical%20Research.pdf

The Objective of the Presentation

- Show how to calibrate *p*-values as measures of evidence
- As a diversion, demonstrate how to deal with statistically "significant" and, alternatively, statistically "insignificant" results obtained from small samples
- Offer an introduction to variable selection as a Bayesian approach to statistical inference and decision-making under uncertainty (*)
  - Variable selection is evidential—it provides a posterior probability of the null hypothesis being true
  - In variable selection regression models, all covariates are retained—the models do not have to be re-estimated in the presence of covariates for which no (sufficiently strong) statistical evidence has been obtained

* In Bayesian statistics, inference and decision-making under uncertainty are conceptually identical

Classical Hypothesis Testing

Developed by Jerzy Neyman and Egon Pearson

- The researcher chooses a type I error rate α that defines a critical region for rejecting the null hypothesis; a commonly chosen error rate is α=0.05 (five percent)
- For instance, in a standard regression framework, the null hypothesis of "no effect" is rejected if (and only if) for a given regression coefficient the observed *t*-statistic is in the critical region of "$t \geq t_{(\alpha/2)}$ =1.960" (when making use of the asymptotic properties) (*)
- Because there is a one-to-one correspondence between an observed *t*-statistic and its tail probability, the critical region can alternatively be defined in terms of this tail probability, which is known as the *p*-value, thus substituting $p \leq \alpha$ for $t \geq t_{(\alpha/2)}$ (**)
- Thus, in repeated sampling, using the decision rule $p \leq \alpha$, the null hypothesis is falsely rejected 5 percent of the time

* Making use of the asymptotic properties of, for instance, Maximum Likelihood estimators, calls for the use of the standard normal distribution (or, equivalently, *t*-distribution with infinite degrees of freedom)
** The use of the tail probability as the critical region was not part of the original concept of classical hypothesis testing developed by Neyman and Pearson; see Hubbard and Bayarri [6]

Classical Hypothesis Testing

Error Rates are Not Measures of Evidence

- The decision rule $p \leq \alpha$ does not establish evidence against the null
- In the Neyman-Pearson framework, the decision to accept or reject a hypothesis is based on cost-benefits considerations that weigh the cost of committing a type I error (rejecting the null when it is true) against the cost of committing a type II error (not rejecting the null when it is false)
- Says Neyman: (*)

  "Thus, to accept a hypothesis *H* means only to decide to take action *A* rather than action *B*. This does not mean that we necessarily believe that the hypothesis *H* is true … [whereas rejecting a hypothesis *H*] … means only that the rule prescribes action *B* and does not imply that we believe that *H* is false."

* Neyman, Jerzy, *First Course in Probability and Statistics*, 1950, New York: Holt, p. 259-260. Quoted (net of bracket) from Hubbard and Bayarri [6]

Classical Hypothesis Testing

Misinterpretation of *p*-Values as Error Rates

- The *p*-value obtained in the context of Neyman-Pearson hypothesis testing is frequently misinterpreted as a frequency-based type I error rate
  - The error rate α is a pre-selected value, whereas the *p*-value is a data-dependent random variable
- According to Hubbard and Bayarri [6], the source of this misinterpretation arises from an amalgamation of R.A. Fisher's concept of *significance testing* (in the absence of an alternative hypothesis) and the Neyman-Pearson concept of *hypothesis testing*
  - In Fisher's approach, a small *p*-value indicates that "[e]ither an exceptionally rare chance has occurred or the theory is not true."(*) The smaller the *p*-value, the stronger the weight of the evidence (Hubbard and Bayarri [6])

* Fisher, Ronald A., *Statistical Methods and Scientific Inference*, 2nd ed., Edinburgh: Oliver and Boyd, 1959, p. 39. Quoted from Hubbard and Bayarri [6]. See Appendix 1 for a comparison of R.A. Fisher with Neyman-Pearson

The Case of the Fair Coin

Probability of Observing an Outcome Given a Model M

- Let us assume that, in an experiment where we fixed the number of tosses to 200, heads shows up 115 times
- What are the probabilities associated with this outcome if…

M_1 ("coin is fair") holds true, thus assuming q=0.5?

(*q* being the probability of heads in a given toss)

M_2 ("coin is not fair") holds true, assuming a uniform prior for *q*? (*)

$$P(X = 115|M_1) = \binom{200}{115}\left(\frac{1}{2}\right)^{200} = 0.005956$$

$$P(X = 115|M_2) = \int_0^1 \binom{200}{115} q^{115}(1-q)^{85}dq = \frac{1}{201} = 0.004975$$

See http://en.wikipedia.org/wiki/Bayes_factor for the example stated above. The R code for this problem is displayed in Appendix 2

* A uniform distribution is a rectangular continuous distribution with support on the interval (0,1)

The Case of the Fair Coin

Frequentist and Bayesian Approaches Deliver Opposing Results

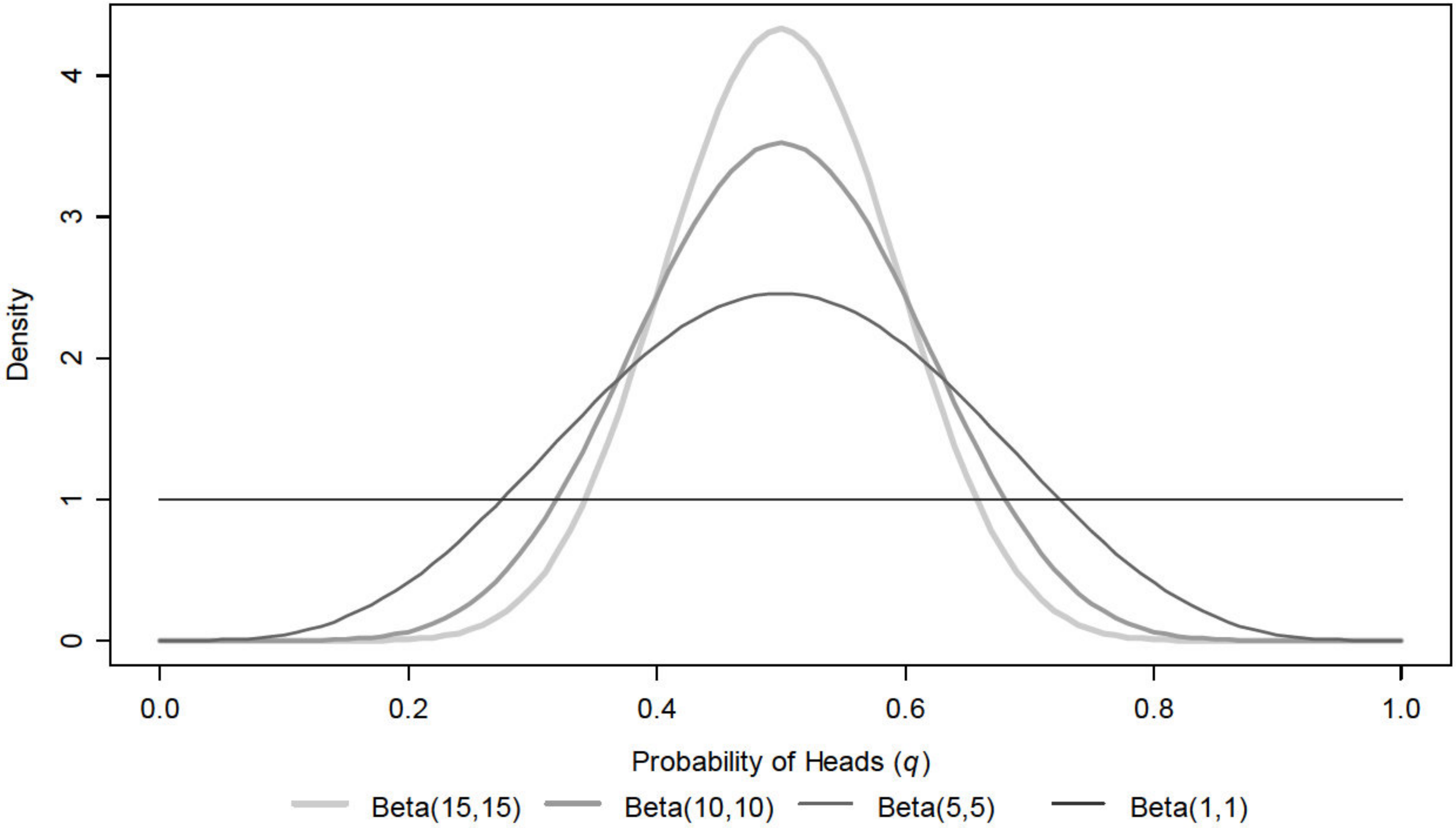
- The example shows that 115 heads in 200 tosses is *more likely* under the hypothesis of a fair coin (the null) than under the hypothesis of a biased coin
- Yet, in classical hypothesis testing, the null is rejected:

  The probability of 115 or more heads in a fixed number of 200 tosses equals 0.02, which results in a *p*-value of 0.04 in a two-tailed test
- Although the outcome is rare under the null, under the alternative of the coin not being fair, the outcome is even rarer
- The discrepancy that arises between the frequentist and Bayesian results under a uniform prior is known as Lindley's paradox

See http://en.wikipedia.org/wiki/Bayes_factor for the example stated above and http://en.wikipedia.org/wiki/Lindley's_paradox for a discussion of Lindley's paradox. The R code for this problem is displayed in Appendix 2

The Case of the Fair Coin

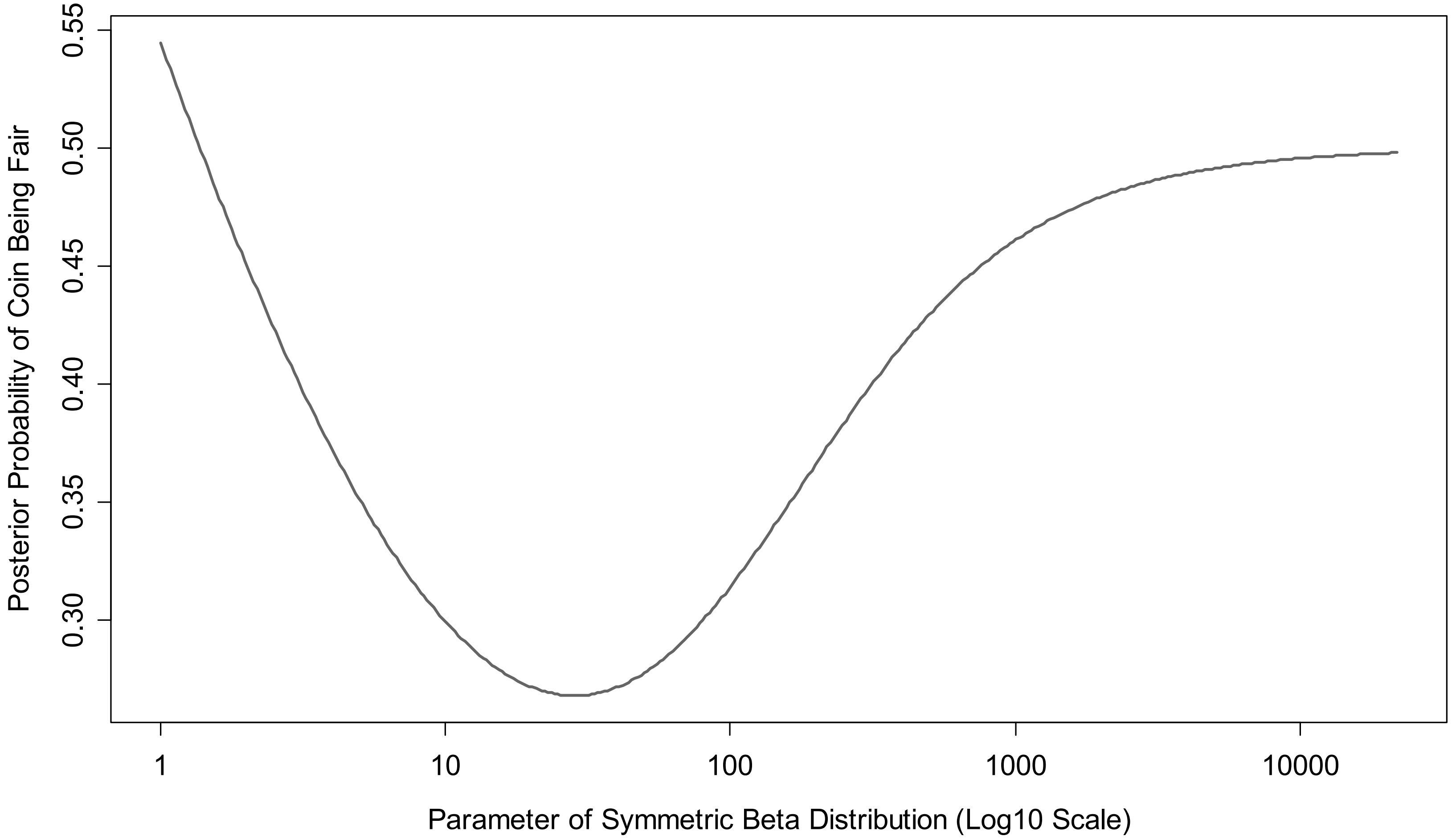
Symmetric Beta Priors for Probability *q*



The Beta(1,1) distribution is equivalent to the uniform distribution

The Case of the Fair Coin

Probability of the Coin Being Fair, Using Alternative Symmetric Beta Priors for *q*



The computations are based on Albert [1]. See Appendix 3 for the R code that generates this chart. The minimum posterior probability that the coin is fair equals 26.8 percent

The Case of the Fair Coin

Summary

- Using alternative symmetric beta distributions as priors for the fairness parameter of the coin, *q*, it has been shown that the posterior probability of the coin being fair when 115 heads come up in a fixed number of 200 tosses amounts to *at least* 26.8 percent
- On the other hand, the *p*-value equals 0.04, thus calling for a rejection of the null in classical hypothesis testing (when using an error rate of α=0.05)
- The *p*-value calculation is based on the event "115 heads or more," which has not been observed. What has been observed is exactly 115 heads, and it is this specific event that the posterior probability of fairness is based on
- This example illustrates the difference between a measure of evidence ("based on this sample, the probability that the coin is fair equals at least 26.8 percent") and an error rate ("in repeated sampling, rejecting the null for a *p*-value not greater than 5 percent is the wrong decision 5 percent of the time")

Calibrating p-Values as Measures of Evidence

What Is in a *p*-Value?

- Central to the work by Sellke, Bayarri, and Berger [11] is the following argument:
  - "The point, however, is that, if a study yields *p* = 0.046, this is the actual information, not the summary statement $0 < p < 0.05$. The two statements are very different from an evidentiary perspective, and replacing the former by the latter is simply an egregious mistake." (p. 5)
- The authors then demonstrate that a *p*-value of 0.05 is only weak evidence against the H_0—this is because "a *p-value* near 0.05 is essentially as likely to arise from H_1 as from H_0" (p. 5) (*)
- Finally, the authors show how to calibrate *p*-values as approximate lower bounds…

  (1) on the Bayes factor *B(p)* (**) of H_0 to H_1 and

  (2) on the frequentist conditional type I error probability $\alpha(p)$

* See http://www.stat.duke.edu/~berger/applet2/pvalue.html for a JAVA applet for *p*-value simulations
** The Bayes factor is an odds ratio, calculated as the ratio of marginal likelihoods of two competing models (hypotheses)

Calibrating p-Values as Measures of Evidence

Bayes Factor and Conditional Type I Error Probability

- The calibrations of *p*-values as Bayes factors, *B(p)*, and, alternatively, as frequentist conditional type I error probabilities, α*(p)*, hold for $p < 1/e$ (where *e* is Euler's constant)
  - The approximate lower bound for the Bayes factor of H_0 to H_1 (that is, stated as the reciprocal value of the odds against the null) reads:

    $$B(p) = -e\,p\,\log(p)$$

  - The approximate lower bound for the frequentist conditional type I error probability equals:

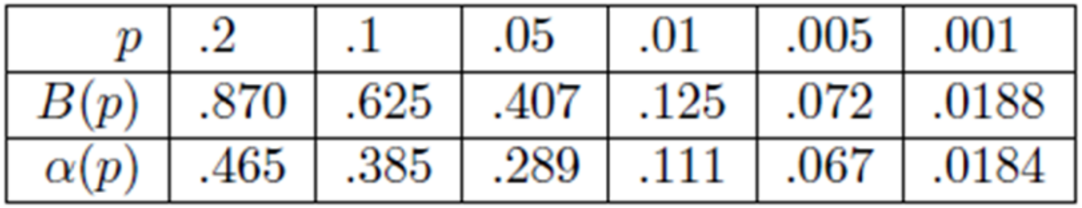
    $$\alpha(p) = (1 + [-e\,p\log(p)]^{-1})^{-1}$$

See Sellke, Bayarri, and Berger [11]

Calibrating p-Values as Measures of Evidence

Bayesian Interpretation and Conversion Table

- The expression for α*(p)* is the same as the posterior probability of H_0 that results from *B(p)* under the assumption of H_0 and H_1 having equal prior probabilities (of 0.5)
- Hence, α*(p)* can be interpreted as (the lower bound on) the posterior probability that the null hypothesis it true
- Below a conversion table for selected *p*-values:



p	.2	.1	.05	.01	.005	.001
$B(p)$	.870	.625	.407	.125	.072	.0188
$\alpha(p)$	.465	.385	.289	.111	.067	.0184

See Sellke, Bayarri, and Berger [11]

Calibrating p-Values as Measures of Evidence

Example

- As an example, let the *p*-value of a given regression coefficient be about 0.05 (*)
  - Then, using the approximation by Sellke, Bayarri, and Berger [11], the lower bound of the conditional type I error probability $\alpha(p)$ equals 0.289
  - This finding can be interpreted as saying that for H_0 and H_1 having equal *prior* probabilities of being true, the *posterior* probability of H_0 being true equals at least 28.9 percent
  - Further, for H_0 and H_1 having equal *prior* probabilities of being true, the Bayesian odds against the null are *at most* 2.5 to 1, as shown below:

    $1/B(p) = (1\text{-}0.289)/0.289 \approx 2.5$
  - In summary, the evidence against the null differs by a wide margin from the common interpretation that, for a *p*-value of 0.05, the evidence against the null is 20:1

* Because *p*-values are continuously distributed, the probability mass at 0.05 is nil; hence the concept of the *p*-value being in a small interval around 0.05. An example of such an interval are the *p*-values equal to 0.05 after rounding

Digression: Bayesian Meta-Analysis

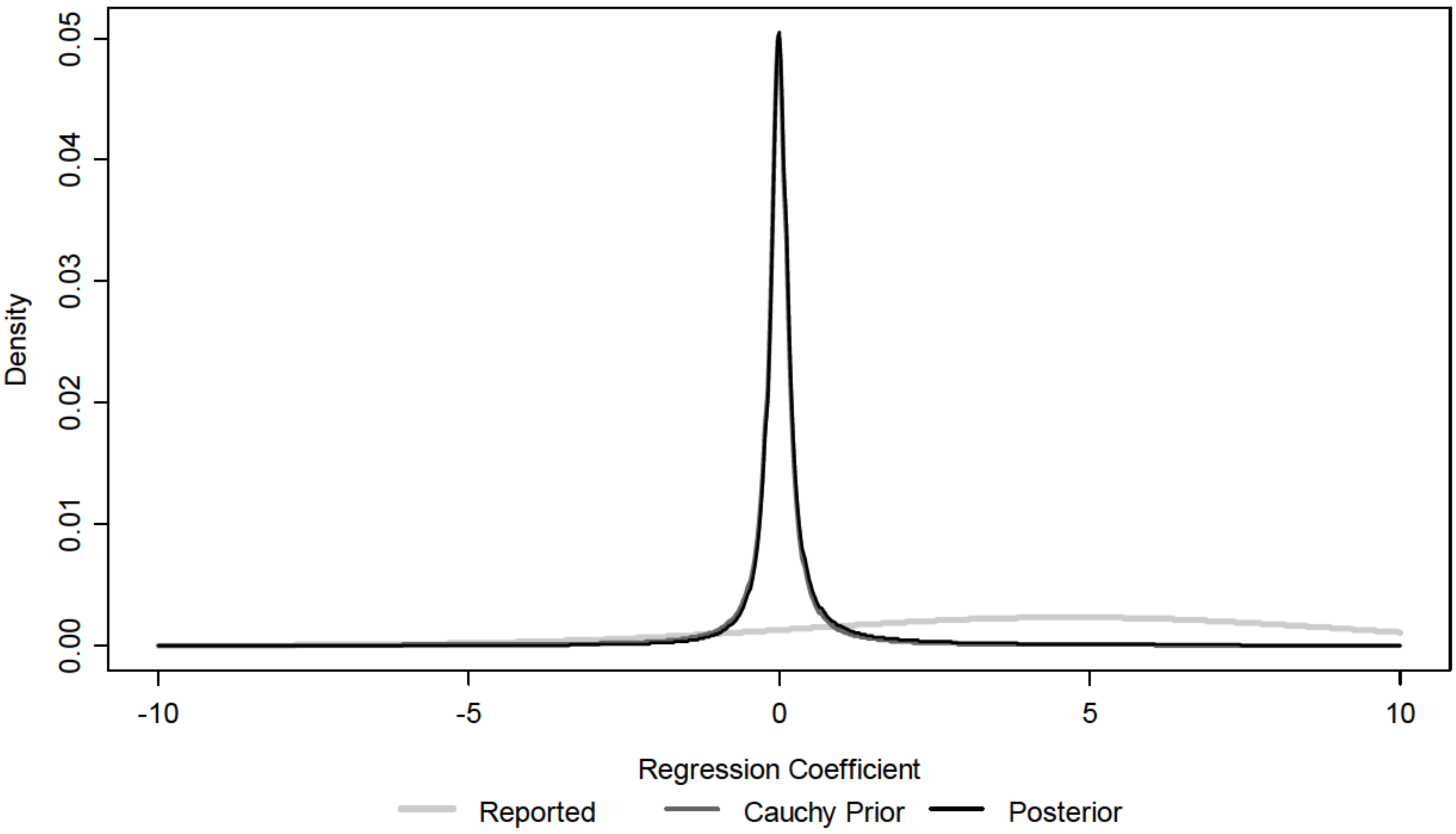
Gelman and Weakliem [5]

- In small samples, the magnitude of statistically "significant" effects are likely overestimated
  - Due to high standard errors associated with the estimation of regression coefficients, extreme outcomes are more likely to make it past the filter of statistical "significance"
- Conversely, statistically "insignificant" effects may be worth exploring, as statistical tests have little power in small samples (that is, the null has little chance of being rejected)
- Gelman and Weakliem [5] suggest using a Cauchy prior (and a normal likelihood) to compute the probability that an estimated regression coefficient is indeed positive (or negative, depending on the case) or within a given interval
  - The scale parameter of the Cauchy distribution is calibrated such that an interval between zero and a judgmentally chosen magnitude for the regression coefficient covers (for instance) 90 percent of the probability mass of the density
  - The Cauchy distribution has fairly flat tails—if the reported large effect has been measured with a sufficiently small standard error, then this effect is able to manifest itself in the posterior distribution even when the magnitude is outside the 90 percent interval defined by the Cauchy prior

See Appendix 4 for R code that applies the concept by Gelman and Weakliem [5] to an example discussed in these authors' paper

Digression: Bayesian Meta-Analysis

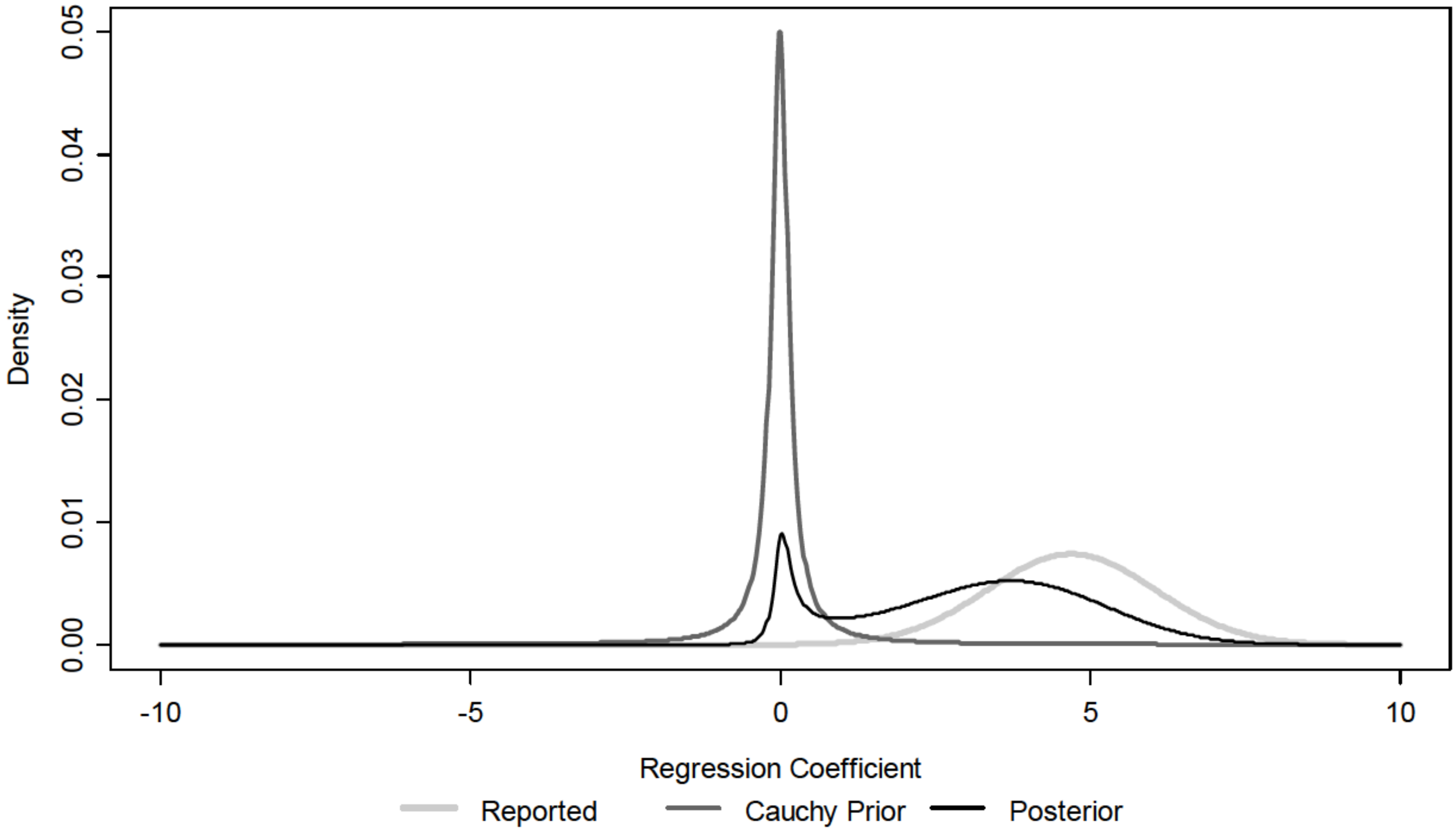
Example in Gelman and Weakliem [5]



The posterior is nearly identical with the prior due to a large standard error of the reported regression coefficient. The mode (highest point of density) or median may serve as the posterior estimate

Digression: Bayesian Meta-Analysis

Hypothetical Example of a Smaller Reported Standard Error



The flat tails of the Cauchy distribution accommodate the evidence associated with regression coefficients that come with small standard errors. The mode (highest point of density) close to the reported distribution may serve as the posterior estimate

Variable Selection

- Motivation
- Approaches
- Performance
- Implementation in R Using JAGS

Motivation

- Variable selection is evidential—through a dichotomous variable that governs the variable inclusion decision, the model delivers a posterior probability for the null hypothesis being true (or false)
- Variable selection does not automatically falsely reject the null hypothesis α × 100 percent of the time
  - Further, the researcher can control the sparseness of the model by adjusting the prior probability that regression coefficients are non-zero (or, for model space approaches, by adjusting the prior distribution for the dimension of the model space)
- Variable selection is not a remedy for the problem of overestimated effects in small samples

In Bayesian variable selection, the maximum number of covariates that can be safely included in the model may exceed the number of observations 10 to 15 times. See O'Hara and Sillanpää [9]

Approaches to Variable Selection

- Model space approach
  - Reversible Jump MCMC
- Indicator model selection
  - Kuo-Mallick Method
  - Gibbs Variable Selection
- Stochastic Search Variable Selection (SSVS)
  - Nonhierarchical spike and slab priors
  - Hierarchical spike and slab priors
    - Generalized linear spike and slab (GLSS)

MCMC: Markov Chain Monte Carlo simulation
The above list of variable selection approaches is not meant to be exhaustive. For an overview on variable selection models see O'Hara and Sillanpää [9]

Approaches to Variable Selection

Alternatives to SSVS

- Reversible Jump MCMC
  - Elegant and comparatively fast but unwieldy in high dimensions
  - High initial investment
- Kuo-Mallick Method
  - The approach is simple but suffers from poor mixing if the prior for the regression coefficient is too vague
- Gibbs Variable Selection
  - The approach addresses the poor mixing properties of the Kuo-Mallick method, but requires tuning (using pilot runs)

Stochastic Search Variable Selection

Nonhierarchical Spike and Slab Priors

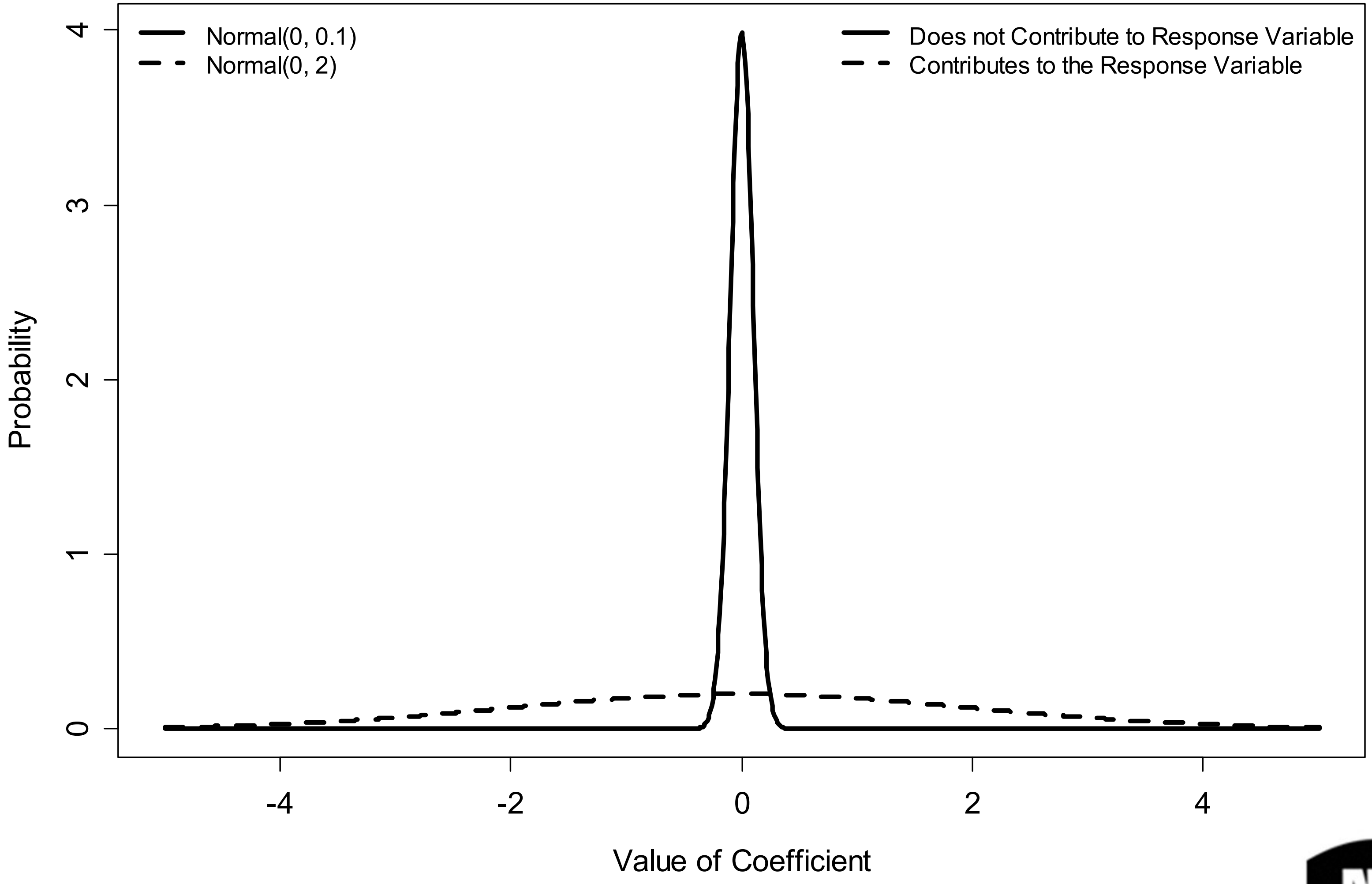
- The prior for the regression coefficient is a mixture of two normal distributions
  - The first distribution (the spike) is "sufficiently concentrated around zero" such that draws from this distribution can be "safely replaced by zero" (*)
  - The second distribution (the slab) is diffuse, thus "admitting the non-zero coefficients" (*)
- An auxiliary variable, γ, typically with a Bernoulli prior, facilitates the mixture
  - $\gamma = 0$ indicates the coefficient originates from the distribution concentrated around zero
  - $\gamma = 1$ indicates the coefficient originates from the diffuse distribution
- Tuning of the variances of the two normal distributions that make up the spike and the slab prior of the regression coefficient can be challenging (**)
  - On one hand, the variance of the spike has to be sufficiently small; on the other hand, if this variance is too restrictive, the Markov chain has difficulty moving between the states "coefficient is zero ($\gamma = 0$)" and "coefficient is non-zero ($\gamma = 1$)"

(*) See Pang and Gill [10]

(**) Typically, the covariates are standardized to improve mixing of the Markov chains, and the dependent variable is scaled to a "reasonable" order of magnitude. Although such scaling does not obviate tuning, it makes the choice of the precisions for the spike and slab prior less dependent on the analyzed data set

Stochastic Search Variable Selection

Nonhierarchical Spike and Slab Priors



Generalized Linear Spike and Slab (GLSS)

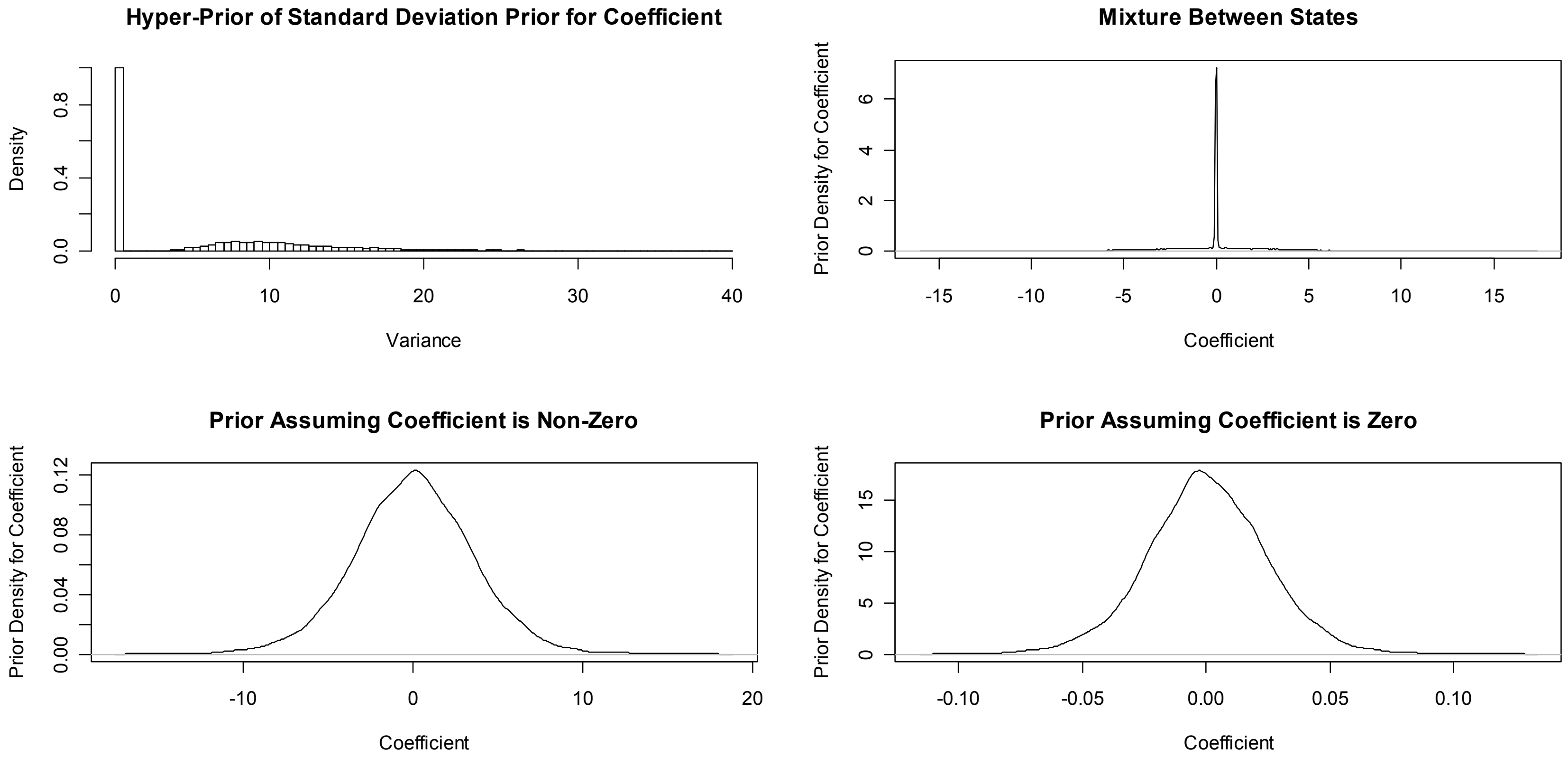
Hierarchical Spike and Slab Priors

- GLSS is an SSVS approach in spirit, with generalized spike and slab priors
- The prior for the regression coefficient is a mixture of two normal distributions
  - SSVS typically has fixed values for the scale parameters of the two normal distributions
  - GLSS places a prior on one of these scale parameters and specifies a fixed ratio between the two
- An auxiliary variable γ facilitates the mixture
  - SSVS typically specifies a Bernoulli hyper-prior for γ with p = 0.5
  - GLSS places a prior on p (instead of imposing a fixed value of, for instance, 0.5)
- The use of hierarchical spike and slab priors makes GLSS adaptive, thus providing a degree of self-tuning

See Pang and Gill [10]

Generalized Linear Spike and Slab (GLSS)

Hierarchical Spike and Slab Priors—Shrinkage Versus Confounding



Hyper-Prior of Standard Deviation Prior for Coefficient

Mixture Between States

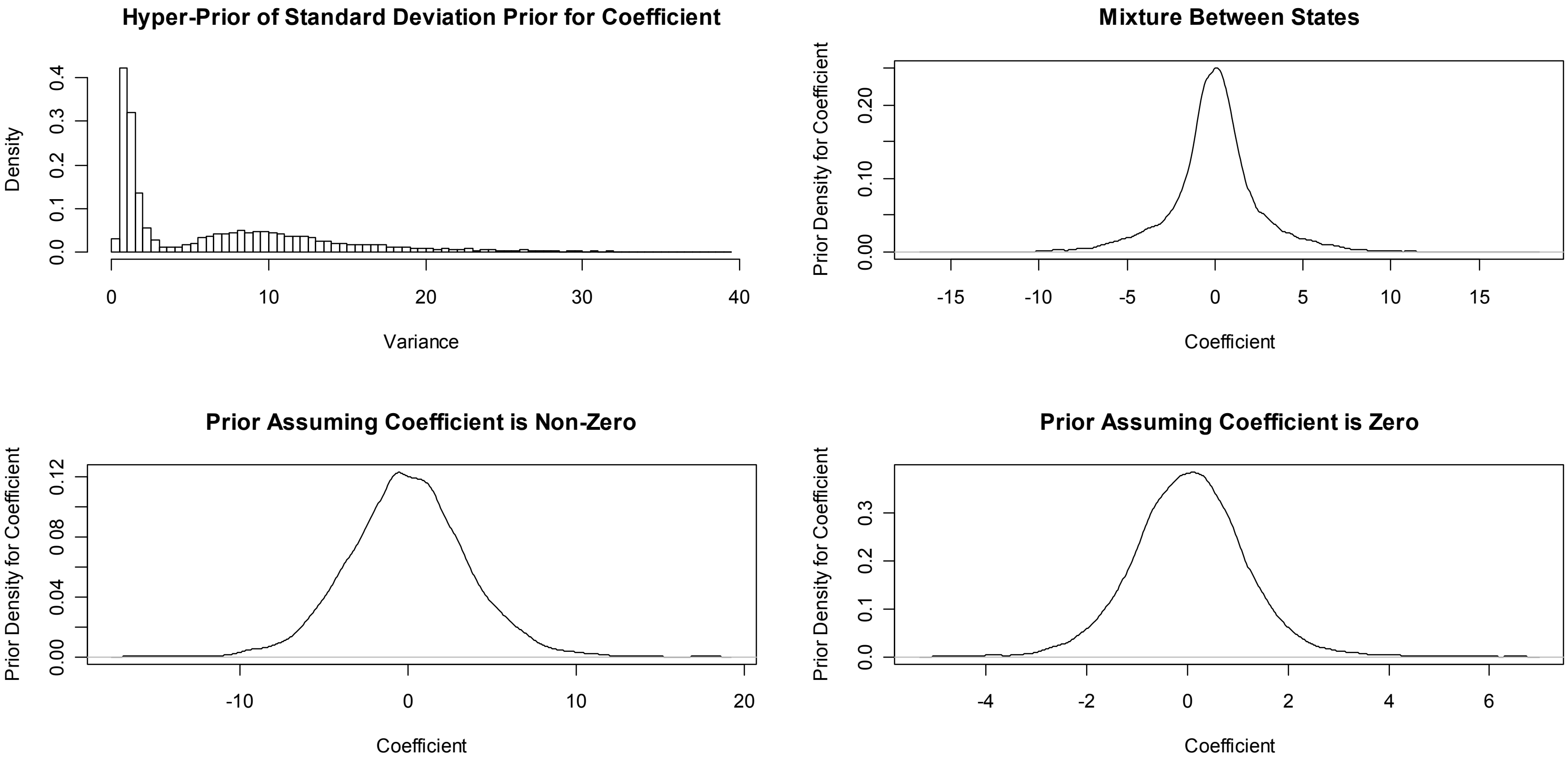
Prior Assuming Coefficient is Non-Zero

Prior Assuming Coefficient is Zero

Critical for the GLSS approach is the choice of the inverse gamma prior for the variance displayed in the upper left-hand chart. If the probability mass that connects the spike and the slab priors is too thin, then the Markov chain has difficulty moving between the states "coefficient is zero" and "coefficient is non-zero." On the other hand, too much confounding diminishes the ability of the model to discriminate between zero and non-zero coefficients. Further, confounding is data-dependent, which makes tuning dependent on the context. See Pang and Gill [10]

Generalized Linear Spike and Slab (GLSS)

Hierarchical Spike and Slab Priors—Shrinkage Versus Confounding


Hyper-Prior of Standard Deviation Prior for Coefficient

Mixture Between States

Prior Assuming Coefficient is Non-Zero

Prior Assuming Coefficient is Zero

Critical for the GLSS approach is the choice of the inverse gamma prior for the variance displayed in the upper left-hand chart. If the probability mass that connects the spike and the slab priors is too thin, then the Markov chain has difficulty moving between the states "coefficient is zero" and "coefficient is non-zero." On the other hand, too much confounding diminishes the ability of the model to discriminate between zero and non-zero coefficients. Further, confounding is data-dependent, which makes tuning dependent on the context. See Pang and Gill [10]

GLSS Performance Evaluation

Two Data Generating Processes

- GLSS is the preferred approach as it requires comparatively little tuning
- The performance is measured for data sets for which the DGP is known
  - Two DGP are considered (both of which require GLM techniques)
    - Poisson with overdispersion
    - Logistic with co-linearity
- The performance is evaluated for samples large (N=1000) and small (N=100)
- The performance is assessed in the presence of model misspecification
- Decisions for inclusion of covariates are compared to classical hypothesis testing

DGP: Data Generating Process
GLM: Generalized Linear Model

GLSS Performance Evaluation

Overdispersed Poisson Process

- Data Generating Process

$$Y_i \sim Poisson(\lambda_i) \quad i = 1, \ldots, N$$

$$\lambda_i = e^{\alpha + \boldsymbol{X}_i \boldsymbol{\beta} + \varepsilon_i}$$

$$X_{ij} \sim Normal(0,1) \;\; j = 1, \ldots, 10$$

$$\varepsilon_i \sim Normal(0, 0.25^2)$$

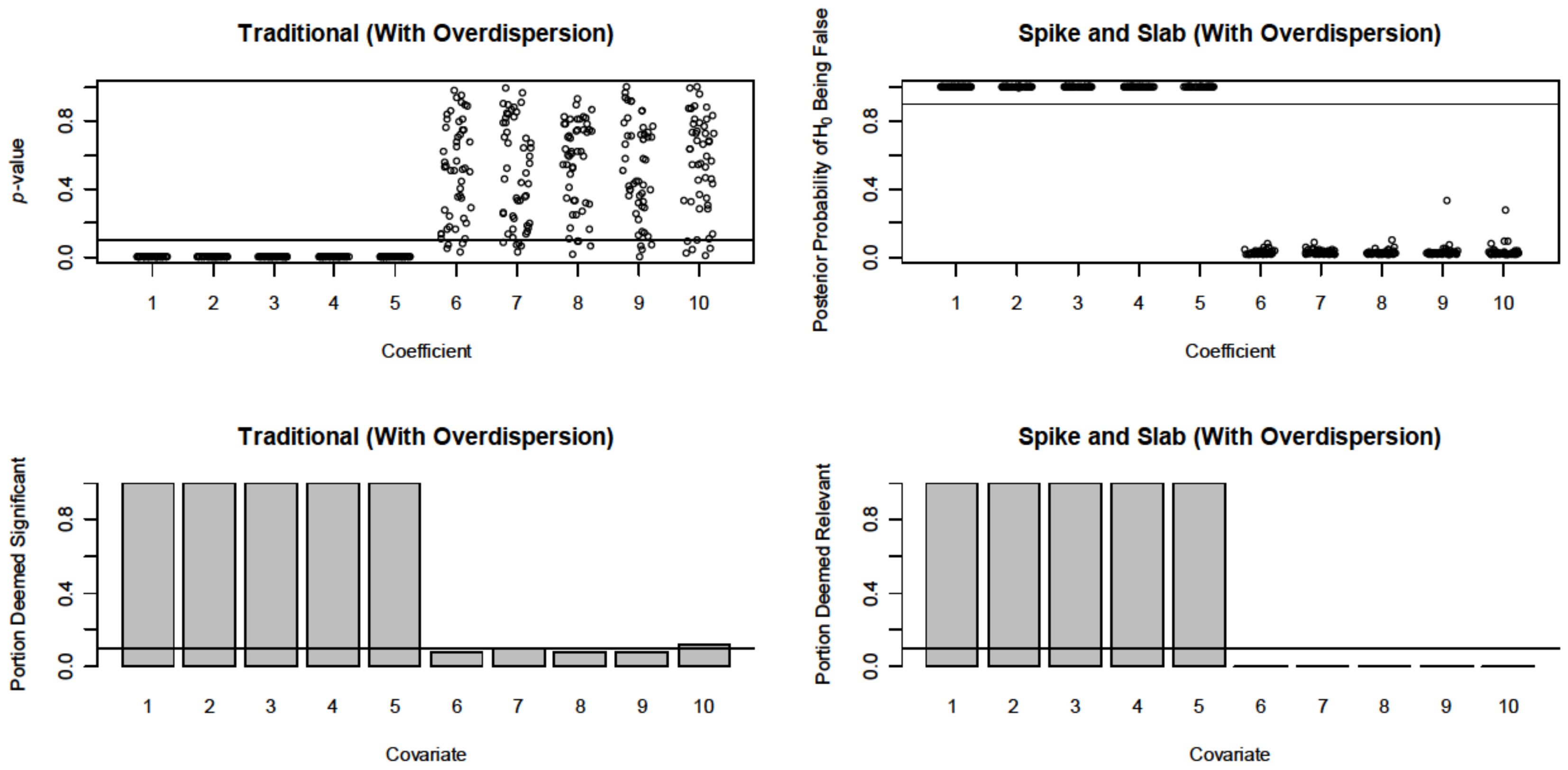
$$\alpha = 0.5$$

$$\boldsymbol{\beta} = [0.6, \quad -0.2, \quad 0.45, \quad -0.35, \quad 0.23, \quad 0, \quad 0, \quad 0, \quad 0, \quad 0]^T$$

Above, the scale parameter of the normal distribution is parameterized as the variance. Generally, in Bayesian statistical models (and in the JAGS code that comes with this presentation), the scale parameter of the normal distribution is parameterized as the precision (which equals the reciprocal of the variance)

GLSS Performance Evaluation

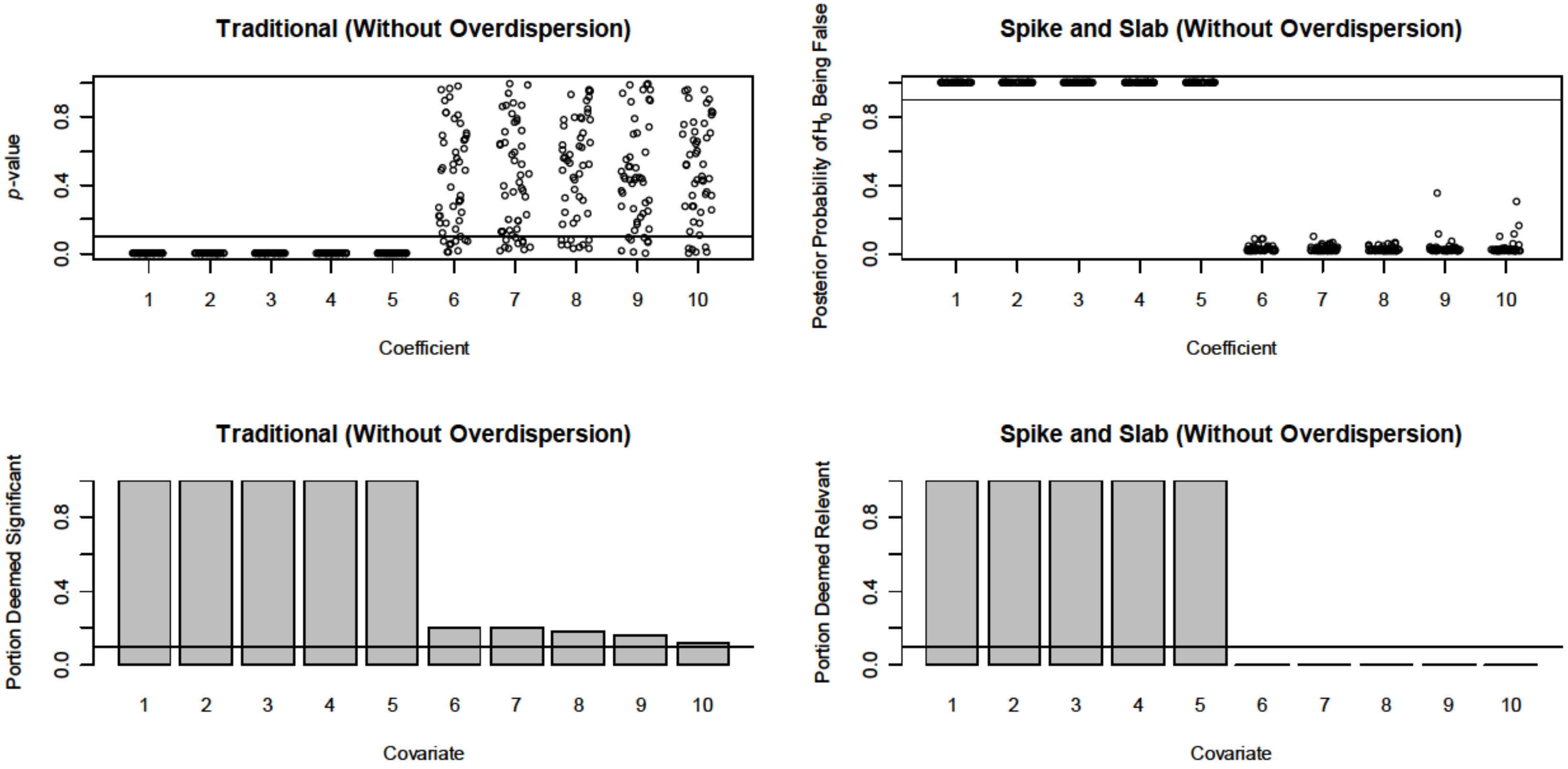
Poisson Process, N=1000, Estimated With Overdispersion



The data generating process is known and is a Poisson process with overdispersion. The first five regression coefficients are nonzero, whereas the latter five are zero. 50 random data sets are estimated. The charts on the left-hand side show the iteratively reweighted least squares approach with traditional hypothesis testing (α=0.1). The charts on the right-hand side illustrate the variable selection approach, implemented by means of MCMC

GLSS Performance Evaluation

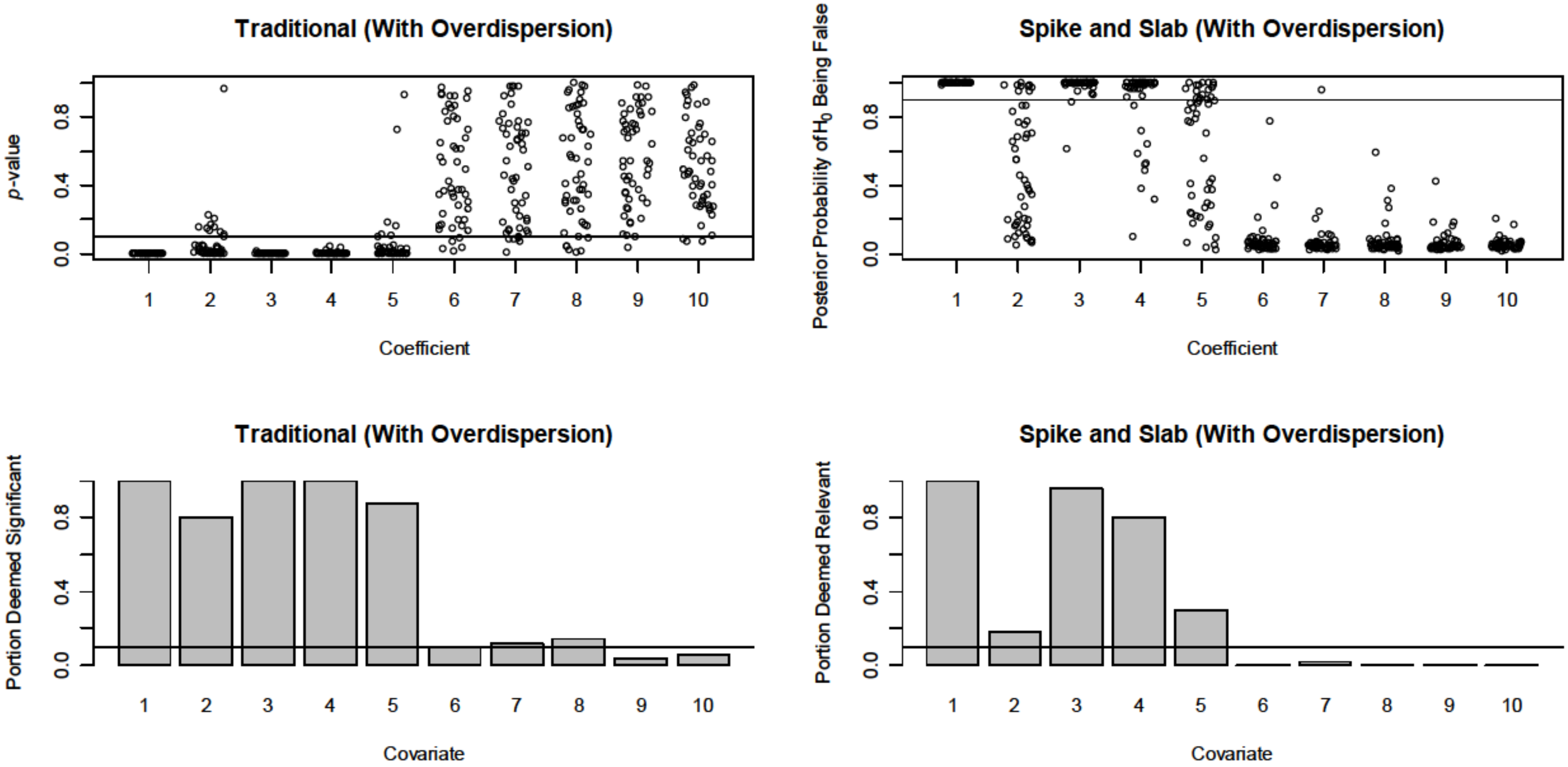
Poisson Process, N=1000, Misspecified (Estimated Without Overdispersion)



The data generating process is known and is a Poisson process with overdispersion. The first five regression coefficients are nonzero, whereas the latter five are zero. 50 random data sets are estimated. The charts on the left-hand side show the iteratively reweighted least squares approach with traditional hypothesis testing (α=0.1). The charts on the right-hand side illustrate the variable selection approach, implemented by means of MCMC

GLSS Performance Evaluation

Poisson Process, N=100, Estimated With Overdispersion



Traditional (With Overdispersion)

Spike and Slab (With Overdispersion)

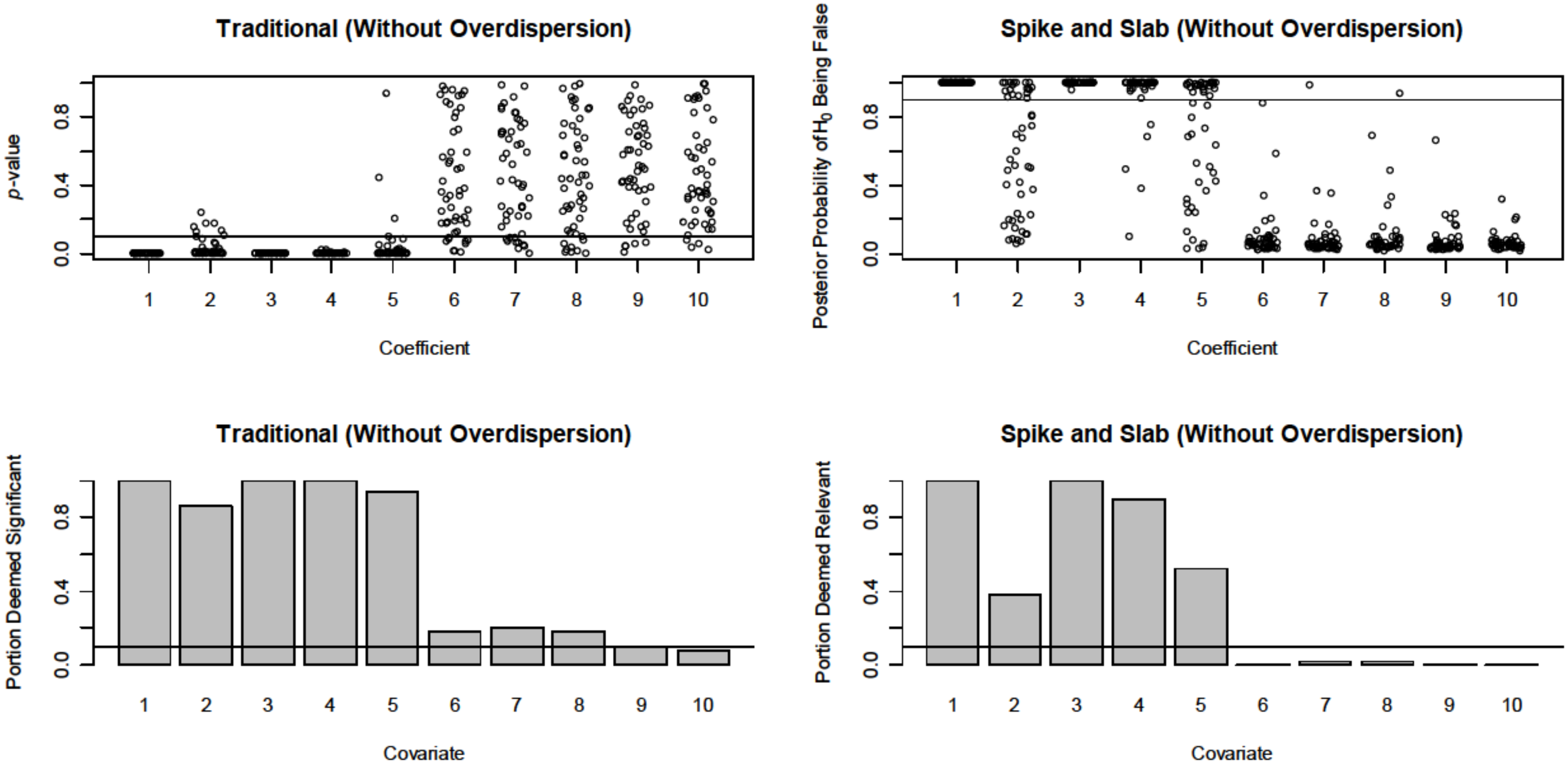
Traditional (With Overdispersion)

Spike and Slab (With Overdispersion)

The data generating process is known and is a Poisson process with overdispersion. The first five regression coefficients are nonzero, whereas the latter five are zero. 50 random data sets are estimated. The charts on the left-hand side show the iteratively reweighted least squares approach with traditional hypothesis testing (α=0.1). The charts on the right-hand side illustrate the variable selection approach, implemented by means of MCMC

GLSS Performance Evaluation

Poisson Process, N=100, Misspecified (Estimated Without Overdispersion)



The data generating process is known and is a Poisson process with overdispersion. The first five regression coefficients are nonzero, whereas the latter five are zero. 50 random data sets are estimated. The charts on the left-hand side show the iteratively reweighted least squares approach with traditional hypothesis testing (α=0.1). The charts on the right-hand side illustrate the variable selection approach, implemented by means of MCMC

GLSS Performance Evaluation

Bernoulli Process

- Data Generating Process

$$Y_i \sim Bernoulli(\pi_i) \quad i = 1, \dots, N$$

$$\pi_i = \frac{e^{\alpha + \boldsymbol{X}_i \boldsymbol{\beta}}}{1 + e^{\alpha + \boldsymbol{X}_i \boldsymbol{\beta}}}$$

$$X_{ij} \sim Normal(0,1) \; j = \; 1, \dots, 5$$

$$X_{i\{6:7\}} \sim MNormal\left([0, \quad 0], \begin{bmatrix} 1 & 0.85 \\ 0.85 & 1 \end{bmatrix}\right)$$

$$X_{i\{8:10\}} \sim MNormal\left([0, \quad 0, \quad 0], \begin{bmatrix} 1 & 0.6 & 0.36 \\ 0.6 & 1 & 0.6 \\ 0.36 & 0.6 & 1 \end{bmatrix}\right)$$

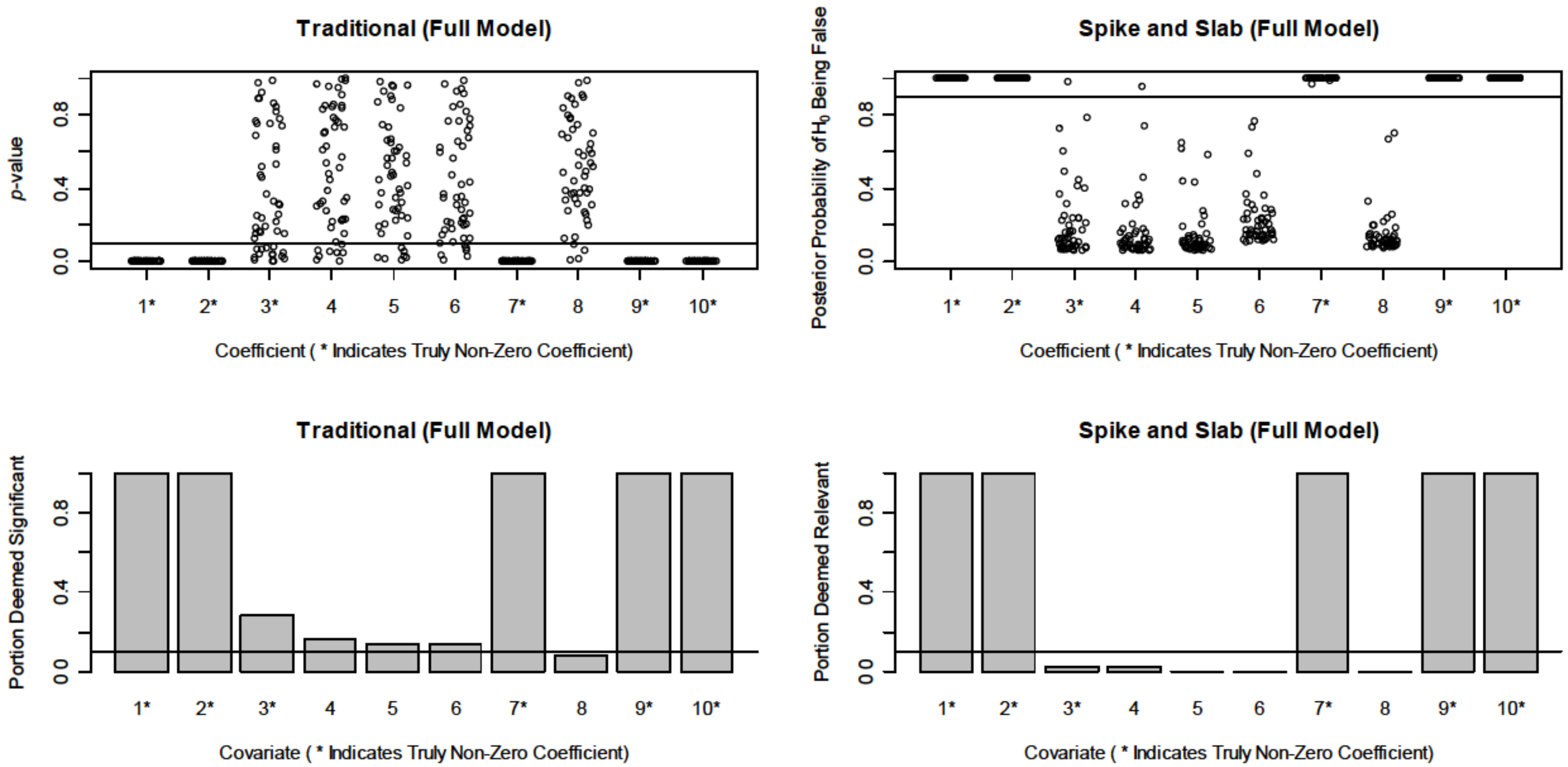
$$\alpha = 0.5$$

$$\boldsymbol{\beta} = [1.4, \quad -1.7, \quad -0.1, \quad 0, \quad 0, \quad 0, \quad -1, \quad 0, \quad -1.6, \quad 0.8]^T$$

Above, the scale parameter of the normal distribution is parameterized as the variance; similarly, for the multivariate normal distribution, the displayed matrix shows the variances and covariances. Generally, in Bayesian statistical models (and in the JAGS code that comes with this presentation), the scale parameter of the normal distribution is parameterized as the precision (which equals the reciprocal of the variance) or, in the multivariate case, the precision matrix (which equals the inverse of the covariance matrix)

GLSS Performance Evaluation

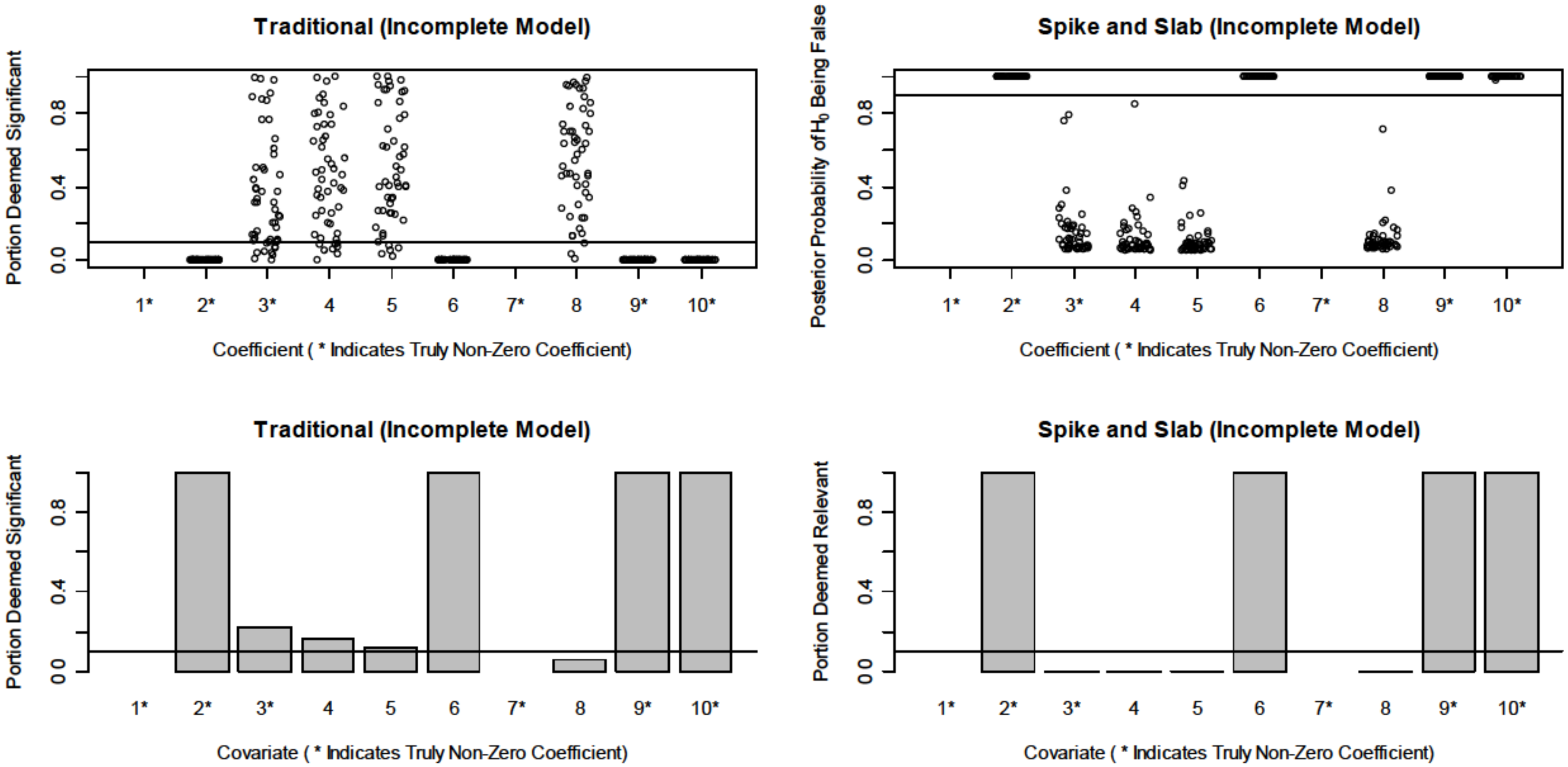
Bernoulli Process, N=1000, Full Model



The data generating process is known and is a Bernoulli process. Covariates 1 through 5 are independent. Covariates 6 and 7 are correlated and covariates 8 through 10 are correlated. 50 random data sets are estimated. The charts on the left-hand side show the iteratively reweighted least squares approach with traditional hypothesis testing (α=0.1). The charts on the right-hand side illustrate the variable selection approach, implemented by means of MCMC

GLSS Performance Evaluation

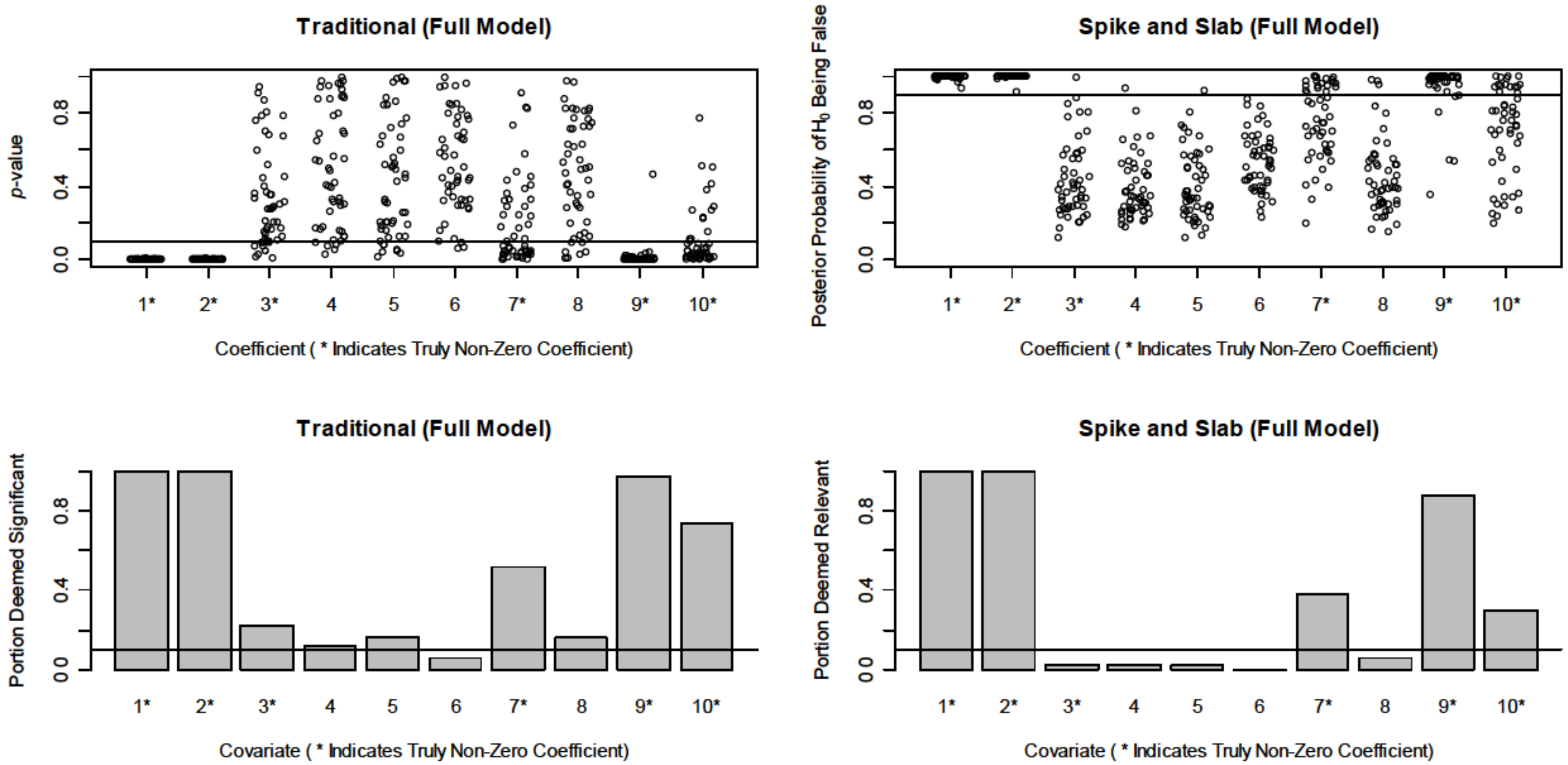
Bernoulli Process, N=1000, Misspecified (Incomplete) Model



The data generating process is known and is a Bernoulli process. Covariates 1 through 5 are independent. Covariates 6 and 7 are correlated and covariates 8 through 10 are correlated. 50 random data sets are estimated. The charts on the left-hand side show the iteratively reweighted least squares approach with traditional hypothesis testing (α=0.1). The charts on the right-hand side illustrate the variable selection approach, implemented by means of MCMC

GLSS Performance Evaluation

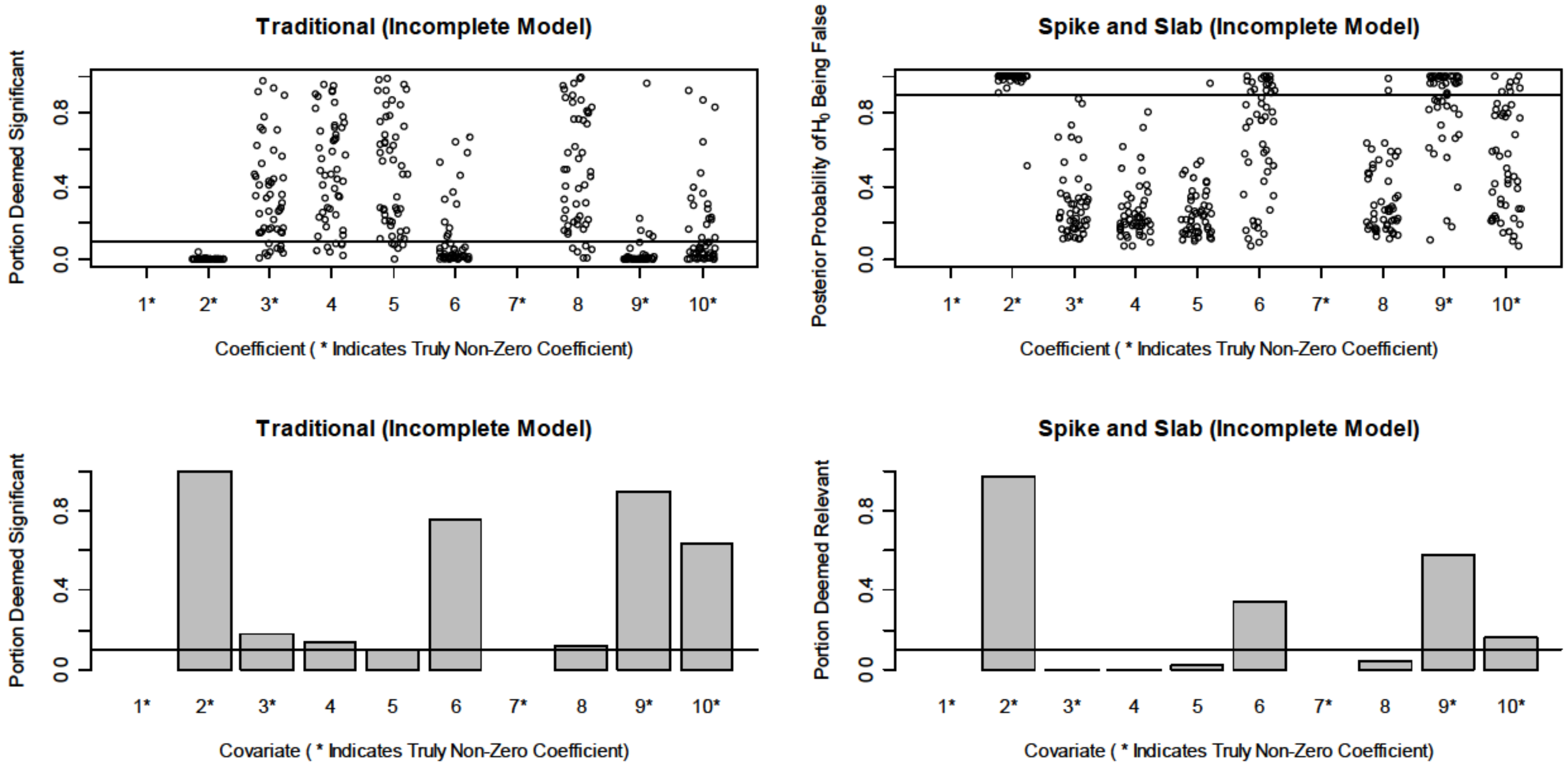
Bernoulli Process, N=100, Full Model



The data generating process is known and is a Bernoulli process. Covariates 1 through 5 are independent. Covariates 6 and 7 are correlated and covariates 8 through 10 are correlated. 50 random data sets are estimated. The charts on the left-hand side show the iteratively reweighted least squares approach with traditional hypothesis testing (α=0.1). The charts on the right-hand side illustrate the variable selection approach, implemented by means of MCMC

GLSS Performance Evaluation

Bernoulli Process, N=100, Misspecified (Incomplete) Model


Traditional (Incomplete Model)

Spike and Slab (Incomplete Model)

Traditional (Incomplete Model)

Spike and Slab (Incomplete Model)

The data generating process is known and is a Bernoulli process. Covariates 1 through 5 are independent. Covariates 6 and 7 are correlated and covariates 8 through 10 are correlated. 50 random data sets are estimated. The charts on the left-hand side show the iteratively reweighted least squares approach with traditional hypothesis testing (α=0.1). The charts on the right-hand side illustrate the variable selection approach, implemented by means of MCMC

SSVS and GLSS Implementation in R

Using JAGS as the Sampling Platform

- Poisson process with overdispersion
    - Known data generating process
        - Five coefficients that are zero
        - Five coefficients that are non-zero
- R, http://www.r-project.org/
    - Open source "software environment for statistical computing and graphics"
    - Implementation of the S language, which was developed at Bell Laboratories
- JAGS – Just Another Gibbs Sampler, http://sourceforge.net/projects/mcmc-jags/files/
    - "A program for the statistical analysis of Bayesian hierarchical models by Markov Chain Monte Carlo simulation"
    - Called from R using the package *rjags*, http://cran.r-project.org/web/packages/rjags/index.html
    - The R code and the JAGS model files are attached to this presentation

There are several R packages that perform variable selection, among which are the packages *spikeslab* and *spikeSlabGAM*

References

[1] Albert, Jim, *Bayesian Computation with R*, 2nd ed., 2009, New York: Springer.

[2] Carey, Benedict, "You Might Already Know This…," *The New York Times*, January 10, 2011, http://www.nytimes.com/2011/01/11/science/11esp.html.

[3] Carlin, Bradley P., and Thomas A. Louis, *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd ed., 2000, Boca Raton: Chapman & Hall/CRC.

[4] Freedman, David H., "Lies, Damned Lies, and Medical Science", *The Atlantic*, November 2010, http://www.theatlantic.com/magazine/archive/2010/11/lies-damned-lies-and-medical-science/8269/.

[5] Gelman, Andrew, and David Weakliem, "Of Beauty, Sex, and Power: Statistical Challenges in Estimating Small Effects," September 8, 2007, http://www.stat.columbia.edu/~gelman/research/unpublished/power.pdf.

[6] Hubbard, Raymond, and M.J. Bayarri, "Confusion Over Measures of Evidence (p's) Versus Errors (α's) in Classical Statistical Testing" (with discussion), *The American Statistician*, August 2003, Vol. 57, No. 3, 171-182.

[7] Lehrer, Jonah, "The Truth Wears Off — Is There Something Wrong With the Scientific Method?", *The New Yorker*, December 13, 2010, http://www.newyorker.com/reporting/2010/12/13/101213fa_fact_lehrer#ixzz1HTmXk2ni.

[8] Matthews, Robert, "Errors Behind Fluke Results," *Financial Times*, July 9, 2004.

[9] O'Hara, Robert B., and Mikko J. Sillanpää, "A Review of Bayesian Variable Selection Methods: What, How and Which," *Bayesian Analysis*, 2009, Vol. 4, No. 1, 85-118.

[10] Pang, Xun, and Jeff Gill, "Spike and Slab Prior Distributions for Simultaneous Bayesian Hypothesis Testing, Model Selection, and Prediction, of Nonlinear Outcomes," Working Paper, Washington University in St. Louis, http://polmeth.wustl.edu/mediaDetail.php?docId=914, July 13, 2009.

[11] Sellke, Thomas, M.J. Bayarri, and James O. Berger, "Calibration of *P*-values for Testing Precise Null Hypotheses," Duke University, Institute of Statistics and Decision Sciences, 1999, http://www.stat.duke.edu/~berger/papers/99-13.html.

[12] The Economist, "Signs of the Time: Why So Much Medical Research is Rot," February 22, 2007, http://www.economist.com/node/8733754?story_id=8733754.

A revised version of Sellke, Bayarri, and Berger [10] was published in *The American Statistician*, February 2001, Vol. 55, No. 1, 62-71

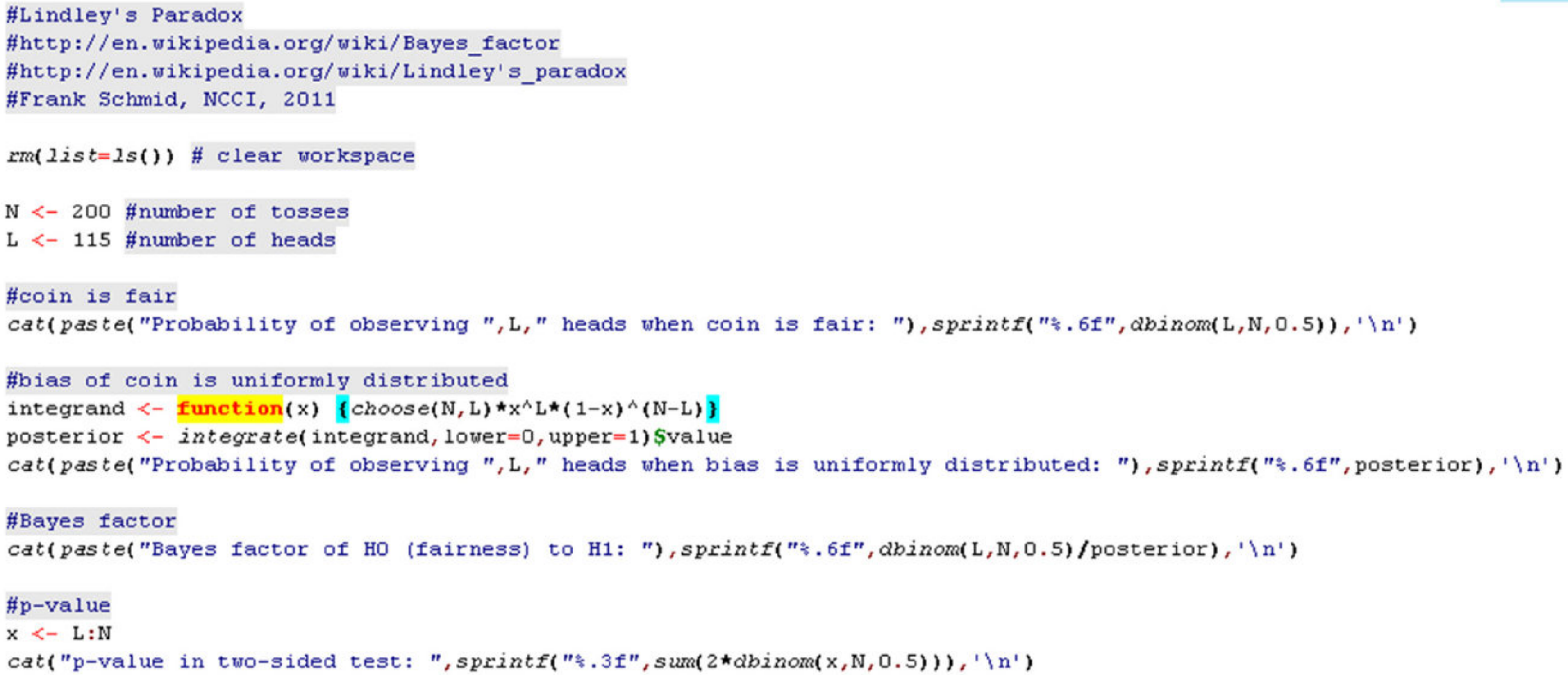
Appendix 1: Fisher versus Neyman-Pearson

Ronald A. Fischer	Jerzy Neyman and Egon Pearson
p—measure of evidence	α—measure of error (error rate)
Significance testing	Hypothesis testing
There is no alternative hypothesis	Null (H_0) versus alternative (H_A)
Inductive inference	Inductive behavior
Evidential	Non-evidential
p value is stochastic	α is chosen by the researcher
p value represents a five$^{(*)}$ percent tail probability	α presents a five$^{(*)}$ percent tail probability

* Both Fisher and Neyman and Pearson considered other tail probabilities, for instance, 1 percent.
See Hubbard and Bayarri [6]

Appendix 2: Fair Coin, Uniform Prior

- R code for Lindley's Paradox

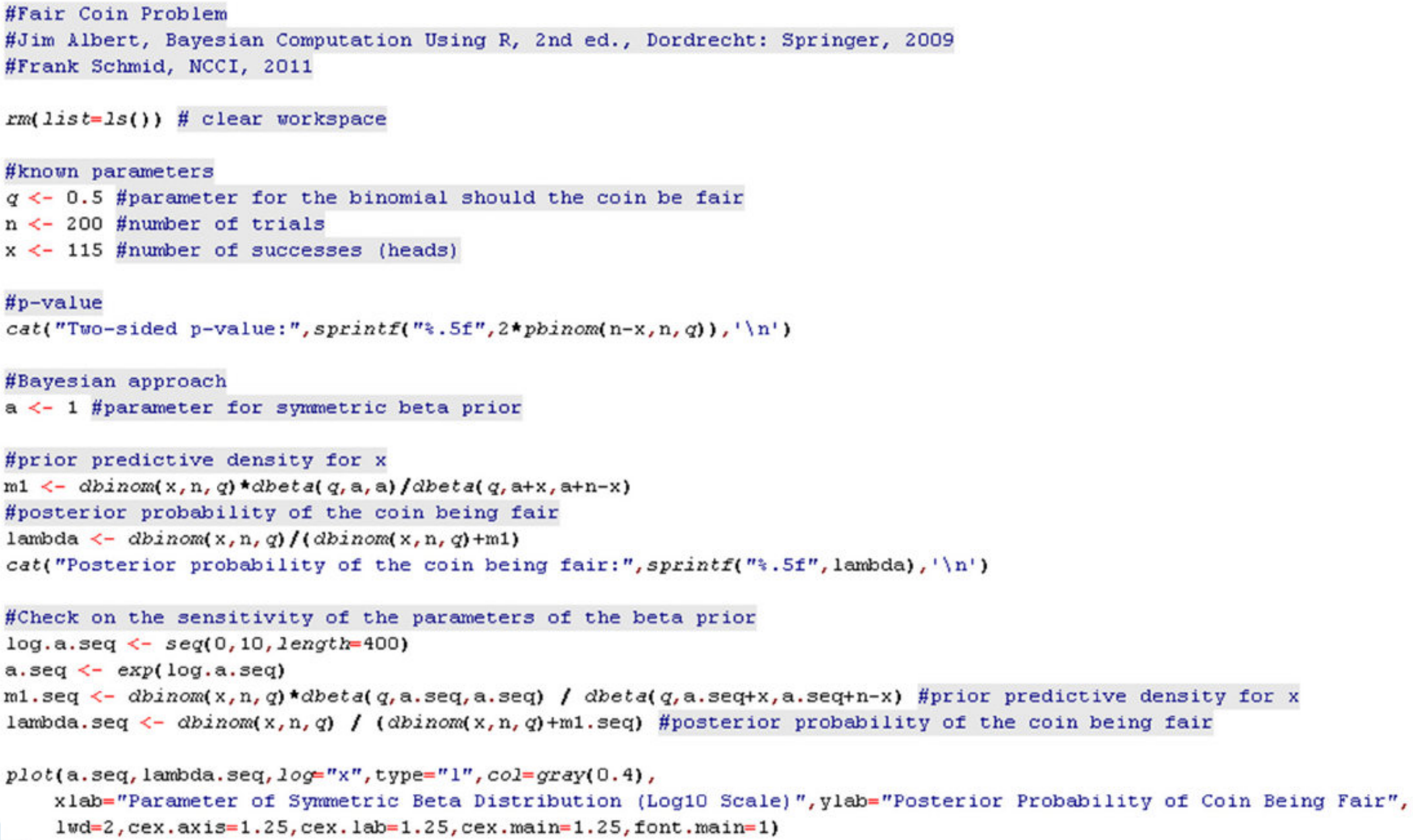


```
#Lindley's Paradox
#http://en.wikipedia.org/wiki/Bayes_factor
#http://en.wikipedia.org/wiki/Lindley's_paradox
#Frank Schmid, NCCI, 2011

rm(list=ls()) # clear workspace

N <- 200 #number of tosses
L <- 115 #number of heads

#coin is fair
cat(paste("Probability of observing ",L," heads when coin is fair: "),sprintf("%.6f",dbinom(L,N,0.5)),'\n')

#bias of coin is uniformly distributed
integrand <- function(x) {choose(N,L)*x^L*(1-x)^(N-L)}
posterior <- integrate(integrand,lower=0,upper=1)$value
cat(paste("Probability of observing ",L," heads when bias is uniformly distributed: "),sprintf("%.6f",posterior),'\n')

#Bayes factor
cat(paste("Bayes factor of H0 (fairness) to H1: "),sprintf("%.6f",dbinom(L,N,0.5)/posterior),'\n')

#p-value
x <- L:N
cat("p-value in two-sided test: ",sprintf("%.3f",sum(2*dbinom(x,N,0.5))),'\n')
```

Source: NCCI

Appendix 3: Fair Coin, Beta Priors

- R code for computing the minimum probability of the coin being fair

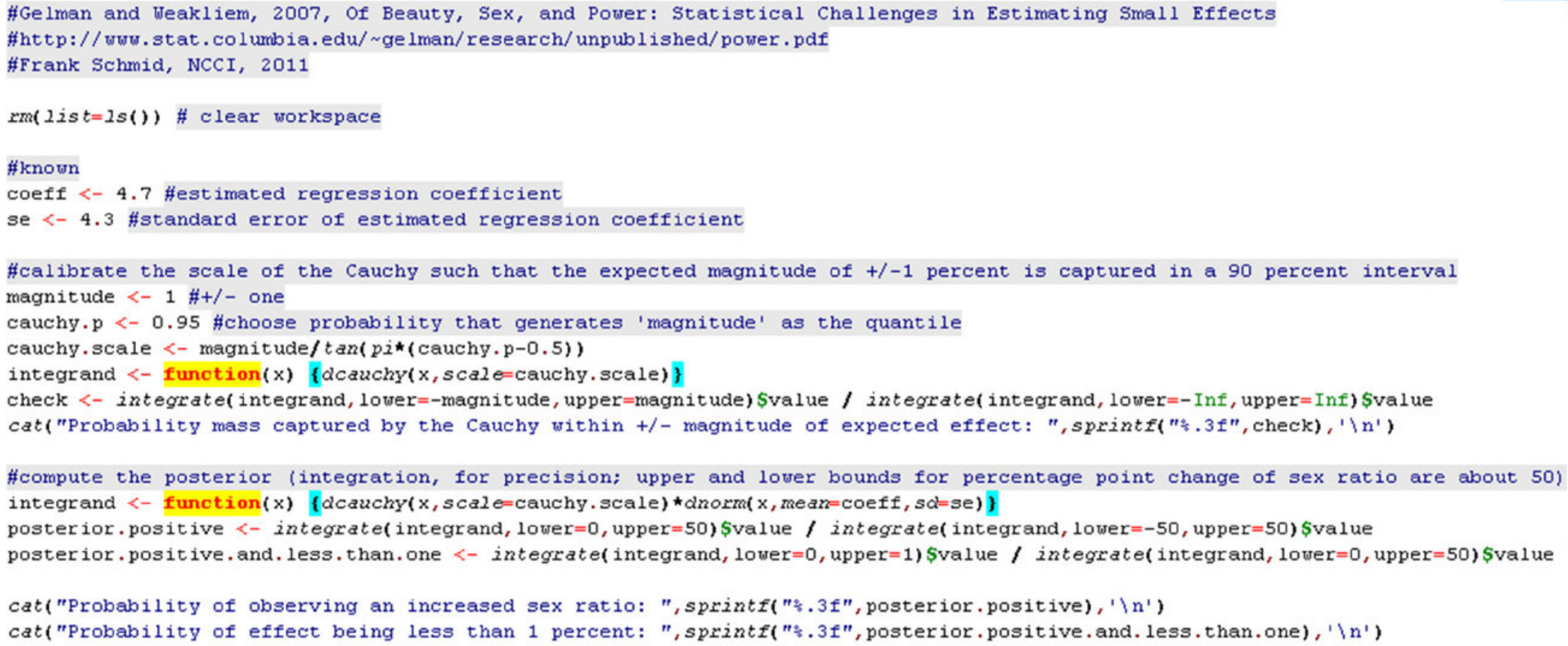


```
#Fair Coin Problem
#Jim Albert, Bayesian Computation Using R, 2nd ed., Dordrecht: Springer, 2009
#Frank Schmid, NCCI, 2011

rm(list=ls()) # clear workspace

#known parameters
q <- 0.5 #parameter for the binomial should the coin be fair
n <- 200 #number of trials
x <- 115 #number of successes (heads)

#p-value
cat("Two-sided p-value:",sprintf("%.5f",2*pbinom(n-x,n,q)),'\n')

#Bayesian approach
a <- 1 #parameter for symmetric beta prior

#prior predictive density for x
m1 <- dbinom(x,n,q)*dbeta(q,a,a)/dbeta(q,a+x,a+n-x)
#posterior probability of the coin being fair
lambda <- dbinom(x,n,q)/(dbinom(x,n,q)+m1)
cat("Posterior probability of the coin being fair:",sprintf("%.5f",lambda),'\n')

#Check on the sensitivity of the parameters of the beta prior
log.a.seq <- seq(0,10,length=400)
a.seq <- exp(log.a.seq)
m1.seq <- dbinom(x,n,q)*dbeta(q,a.seq,a.seq) / dbeta(q,a.seq+x,a.seq+n-x) #prior predictive density for x
lambda.seq <- dbinom(x,n,q) / (dbinom(x,n,q)+m1.seq) #posterior probability of the coin being fair

plot(a.seq,lambda.seq,log="x",type="l",col=gray(0.4),
    xlab="Parameter of Symmetric Beta Distribution (Log10 Scale)",ylab="Posterior Probability of Coin Being Fair",
    lwd=2,cex.axis=1.25,cex.lab=1.25,cex.main=1.25,font.main=1)
```

Source: NCCI

Appendix 4: Meta-Analysis

- R code for evaluating (statistically "significant" or statistically "insignificant") frequentist regression coefficients obtained from small samples



```
#Gelman and Weakliem, 2007, Of Beauty, Sex, and Power: Statistical Challenges in Estimating Small Effects
#http://www.stat.columbia.edu/~gelman/research/unpublished/power.pdf
#Frank Schmid, NCCI, 2011

rm(list=ls()) # clear workspace

#known
coeff <- 4.7 #estimated regression coefficient
se <- 4.3 #standard error of estimated regression coefficient

#calibrate the scale of the Cauchy such that the expected magnitude of +/-1 percent is captured in a 90 percent interval
magnitude <- 1 #+/- one
cauchy.p <- 0.95 #choose probability that generates 'magnitude' as the quantile
cauchy.scale <- magnitude/tan(pi*(cauchy.p-0.5))
integrand <- function(x) {dcauchy(x,scale=cauchy.scale)}
check <- integrate(integrand,lower=-magnitude,upper=magnitude)$value / integrate(integrand,lower=-Inf,upper=Inf)$value
cat("Probability mass captured by the Cauchy within +/- magnitude of expected effect: ",sprintf("%.3f",check),'\n')

#compute the posterior (integration, for precision; upper and lower bounds for percentage point change of sex ratio are about 50)
integrand <- function(x) {dcauchy(x,scale=cauchy.scale)*dnorm(x,mean=coeff,sd=se)}
posterior.positive <- integrate(integrand,lower=0,upper=50)$value / integrate(integrand,lower=-50,upper=50)$value
posterior.positive.and.less.than.one <- integrate(integrand,lower=0,upper=1)$value / integrate(integrand,lower=0,upper=50)$value

cat("Probability of observing an increased sex ratio: ",sprintf("%.3f",posterior.positive),'\n')
cat("Probability of effect being less than 1 percent: ",sprintf("%.3f",posterior.positive.and.less.than.one),'\n')
```

Source: NCCI