LA-UR-12-24643
Approved for public release; distribution is unlimited.

Title: Using a Simple Binomial Model to Assess Improvement in Predictive Capability: Sequential Bayesian Inference, Hypothesis Testing, and Power Analysis

Author(s): Sigeti, David E.
Pelak, Robert A.

Intended for: arXiv
Report

Disclaimer:
Los Alamos National Laboratory, an affirmative action/equal opportunity employer,is operated by the Los Alamos National Security, LLC for the National NuclearSecurity Administration of the U.S. Department of Energy under contract DE-AC52-06NA25396. By approving this article, the publisher recognizes that the U.S. Government retains nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Departmentof Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

Using a Simple Binomial Model to Assess Improvement in Predictive Capability: Sequential Bayesian Inference, Hypothesis Testing, and Power Analysis

David E. Sigeti and Robert A. Pelak

September 9, 2012

Abstract

We present a Bayesian statistical methodology for identifying improvement in predictive simulations, including an analysis of the number of (presumably expensive) simulations that will need to be made in order to establish with a given level of confidence that an improvement has been observed. Our analysis assumes the ability to predict (or postdict) the same experiments with legacy and new simulation codes and uses a simple binomial model for the probability, θ, that, in an experiment chosen at random, the new code will provide a better prediction than the old. This model makes it possible to do statistical analysis with an absolute minimum of assumptions about the statistics of the quantities involved, at the price of discarding some potentially important information in the data. In particular, the analysis depends only on whether or not the new code predicts better than the old in any given experiment, and not on the magnitude of the improvement. We show how the posterior distribution for θ may be used, in a kind of Bayesian hypothesis testing, both to decide if an improvement has been observed and to quantify our confidence in that decision. We quantify the predictive probability that should be assigned, prior to taking any data, to the possibility of achieving a given level of confidence, as a function of sample size. We show how this predictive probability depends on the true value of θ and, in particular, how there will always be a region around $\theta = 1/2$ where it is highly improbable that we will be able to identify an improvement in predictive capability, although the width of this region will shrink to zero as the sample size goes to infinity. We show how the posterior standard deviation may be used, as a kind of "plan B metric" in the case that the analysis shows that θ is close to 1/2 and argue that such a plan B should generally be part of hypothesis testing. All the analysis presented in the paper is done with a general beta-function prior for θ, enabling sequential analysis in which a small number of new simulations may be done and the resulting posterior for θ used as a prior to inform the next stage of power analysis.

Contents

1 Introduction

In this paper, we describe a Bayesian statistical methodology for identifying improvement in predictive simulations. In particular, we address the question of how many simulations will need to be run and compared to experimental results in order to have a given expected confidence in any detection of improvement or lack of same.[1]

We assume the existence of a "legacy" simulation methodology that has been used to predict experiments in the past and a "new" simulation methodology that may be used either to predict new experiments[2] or to "postdict" the experiments for which the legacy methodology originally produced predictions[3]. The most important constraint in our approach here is that we assume that it is possible to use both the legacy and new methodologies to predict (or postdict) *the same* experiments. In other words, we assume that we will have a set of *paired predictions,* one pair for each experiment.

In order to frame precisely the question of the number of experiments for which we will need paired predictions, we need to define the process that we

[1] The investigation of this question is traditionally referred to as *power analysis* in statistics.

[2] In this case the legacy methodology must still be available to be used on the new experiments.

[3] In this case, the essential requirement is that the postdictions be *honest*, that is, that they do not use the known experimental results to fit simulation inputs or otherwise "tune" the simulation.

will use to assess predictive capability based on a given set of paired predictions. We will assume in this paper that we have a method to determine how well the legacy and new methodologies do *on a given experiment.* We might make this determination, for example, by looking at the discrepancies from the two methodologies between predicted and measured values for a given experimental observable. Or, we could do the same comparison for some weighted combination of discrepancies for multiple observables. However we make the determination, the result will be a set of binary variables, x_i, one for each experiment, i, for which we have compared legacy and new predictions. We will adopt the convention that a value $x_i = 1$ implies that the new methodology does better at predicting the results of experiment i (a "success") and a value of $x_i = 0$ that the new methodology does not do better than the old (a "failure").

In general, we do not expect either the new or the legacy methodology to do better on all the experiments. Thus, in the context of our discussion so far, we may frame our previously vague questions on identifying improvement in predictive capability and the number of simulations required to do so more specifically:

1. If the legacy methodology does better on some experiments and the new methodology on others, how do we determine which is best?

2. How many comparisons between legacy and new methodologies do we need to do in order to answer the question above?

In order to answer these questions, we must next develop a statistical model for the binary data, x_i.

We do this in Section 2, where we propose a simple binomial model with a single unknown parameter, θ, which gives the probability that an experiment chosen at random will be better predicted by the new than the legacy methodology. Of course, $\theta > 1/2$ corresponds to improvement in predictive capability and $\theta \leq 1/2$ corresponds to no improvement (or, if $\theta < 1/2$, to a deterioration). In Section 3, we describe the properties of the binomial model and the result of applying Bayesian inference to it using the x_i as our data. This result will be, as is typical in Bayesian analysis, a posterior distribution for the unknown parameter θ.

Assuming that we will have a posterior for θ, we need to know how we will use it to assess whether or not we have detected an improvement in predictive capability. Given a posterior for θ, the obvious first step is to use it to compute the posterior probability that the new methodology makes better predictions than the old—which is just the posterior probability that θ is greater than $1/2$, $P[\theta > 1/2|d]$, where our probability, being a posterior result, is conditioned on our data, d (see Section 4). It may be that this probability is all that is needed in a particular context, in which case we will be done.

However, if some definite decision must be made as to whether or not we have detected improvement, then we have entered the realm of what conventional (frequentist) statistics calls "hypothesis testing". Because we are using a Bayesian approach we have a tool, the posterior distribution for θ, that is not available to frequentist statisticians and our use of this tool will cause our approach to diverge substantially from conventional hypothesis testing. In particular, for given data and a given decision, we can always compute the posterior probability that our decision is correct (a probability that we call the *posterior confidence*). The availability of the posterior confidence allows us to lay out a three-way choice for evaluating predictive capability: Decide that an improvement is present (at a given level of confidence); decide that an improvement is not present (also at the given level of confidence); or decide that we do not know (again, at the given level of confidence). We present our results on hypothesis testing in Section 4 with Section 4.2 dealing specifically with the three-way test.

In the (usually not very satisfactory) situation where the data dictate that we do not know whether or not there has been an improvement, it will always be the case that the data are also telling us that θ is close to $1/2$. In this case, it makes sense to do some further analysis on the posterior. In particular, it makes sense to examine the width of the posterior in order to establish just how far from $1/2$ θ might possibly be. Any reasonable measure of the width of the posterior (together with any reasonable measure of central tendency, which will be close to $1/2$) constitute additional pieces of information that should be presented to decision makers in the case that our hypothesis tests fail to produce a clear answer. Doing so makes it possible, for example, to decide if any possible improvement in predictive capability (quantified by θ) is simply too small (θ is too close to $1/2$) to make any practical difference. We believe that this kind of "plan B" should probably be part of any hypothesis testing, at least where the hypotheses correspond to an unknown continuous parameter being in particular regions. See Section 4.3 for details.

Having laid out the procedure that we will use, once we have data, to detect any improvement in predictive capability, we can assess the number of paired predictions that we are likely to need in order to be able to achieve a given level of confidence in our results. We do this by evaluating the predictive (sometimes called *preposterior*) probabilities for making certain kinds of errors and for achieving a particular level of confidence. These probabilities may be usefully evaluated either conditionally, as a function of the true value of θ, or unconditionally, by averaging over the possible true values of θ.[4] The results of this analysis are presented in Section 5.

In accord with the program just described, we now proceed to describe our model for the statistics of paired predictions.

2 A Model for the Statistics of Paired Predictions

We will assume that we have paired predictions for n experiments and that we can produce the corresponding set of x_i, where we set $x_i = 1$ if we have a "success", a situation where the new prediction is better than the old, and $x_i = 0$, otherwise.

We will model the x_i as independent, identically distributed random (binary) variables with a well-defined probability, θ ($0 \leq \theta \leq 1$), to take the value 1 and, of course, probability $1 - \theta$ to take the value 0. Naturally, determining a posterior distribution for θ will be the central task for our Bayesian analysis.

In the language of mathematical probability, we will model our samples x_i as a sequence, $\boldsymbol{x} \equiv \{x_i : 1 \leq i \leq n\}$, of *Bernoulli trials* (flips of a weighted coin). In statistics, this is known as the *binomial model.* Not surprisingly, given its simplicity, it is one of the most common and well-understood probabilistic models.

[4] It should be noted that this last operation, averaging over the true value of θ, is only available in a Bayesian analysis. Because the averaging is done with respect to the prior for θ, preposterior θ-averaged results will be heavily dependent on the choice of prior for θ in a way that does not become arbitrarily small as we compute results for larger and larger (potential) datasets. This behavior is quite different from posterior quantities, which become asymptotically independent of the prior as we take more data. This dependence presents no problem if our prior has a strong basis as, for example, in the case where the prior is simply the posterior from an earlier analysis of some substantial amount of data. However, in cases where the prior has no strong basis, as when we use a noninformative prior for lack of a better choice, the predictive probabilities averaged over θ should be taken with a substantial grain of salt.

It is worth noting that, in addition to being simple and well-understood, the binomial model has the advantage that we may carry through the entire program presented in the introduction to this paper without having to worry about the joint distribution of the discrepancies from the legacy and new methodologies, nor their correlation, nor even their respective marginal distributions. All that matters when we use the binomial model is the unknown θ, which we may infer from a sample of paired predictions, reduced to the data x_i.

This simplicity has been bought at the price of ignoring certain information in the paired predictions. In order to be able to work with the binomial model, we have had to restrict our attention to the binary quantities x_i. We are thus ignoring any information we have about *how much* better or worse the new methodology is at predicting a given experiment. If, for example, we are making our decision on whether or not the new methodology is doing better in a given experiment based on the discrepancies from the legacy and new methodologies for a single observable, we are effectively basing our analysis on the *sign* of the difference between the discrepancies and are ignoring the *magnitude* of the difference. There is certainly the possibility that working with the full difference would allow a better analysis, although any improvement would, in turn, need to be paid for with the work to model the joint distribution for the two discrepancies (or at least the distribution for their difference) and with any accompanying uncertainties or errors introduced by this modeling. An analysis, parallel to this one, that takes into account the magnitudes of the differences is currently in the works.

We next proceed to present basic results on the binomial model and on the process of Bayesian inference on it.

3 Bayesian Inference on the Binomial Model

As discussed above, the binomial model assumes that we have a set of n binary samples, x_i, and, further, that the x_i may be modeled as independent, identically distributed random binary variables with a well-defined probability, θ ($0 \leq \theta \leq 1$), to take the value 1 and probability $1 - \theta$ to take the value 0. For a more detailed discussion of the binomial model and references, see Appendix A.1.

If we define s as the number of times that 1 appears in $\boldsymbol{x}$, $s \equiv s(\boldsymbol{x}) \equiv$

$\sum_i^n x_i$, then the likelihood, given θ, of observing s in n trials is

$$P(s|\theta) \equiv \binom{n}{s} \theta^s (1-\theta)^{n-s} \equiv \frac{n!}{s!(n-s)!} \theta^s (1-\theta)^{n-s}. \tag{1}$$

Note that s is an integer so the likelihood is an actual probability, not a density.[5] Of course, s can take the values $0, 1, \ldots, n$.[6]

Given this likelihood and a prior probability distribution for θ, $p(\theta)$, we may apply Bayes law to compute a posterior probability distribution for θ, given an observed sequence $\boldsymbol{x}$,

$$p(\theta|s) = \frac{P(s|\theta)\, p(\theta)}{\int d\theta\, P(s|\theta)\, p(\theta)}. \tag{2}$$

We will assume a conjugate prior[7] for θ, which, in the case of the binomial model, is an instance of the beta distribution on $[0, 1]$,

$$p(\theta) \equiv \mathrm{Be}(\theta|\alpha_0, \beta_0) \equiv \frac{\Gamma(\alpha_0 + \beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)}\, \theta^{\alpha_0 - 1} (1-\theta)^{\beta_0 - 1}, \tag{3}$$

where our prior has hyperparameters $\alpha_0 > 0$ and $\beta_0 > 0$.[8] The case $\alpha_0 = 1$ and $\beta_0 = 1$ corresponds to a flat (noninformative) prior. Then, as shown in Equation 58 in Appendix A.3, the posterior is

$$\begin{aligned} p(\theta|s) &= \mathrm{Be}(\theta|\alpha_0 + s, \beta_0 + n - s) \\ &= \frac{\Gamma(n + \alpha_0 + \beta_0)}{\Gamma(s + \alpha_0)\Gamma(n - s + \beta_0)}\, \theta^{s + \alpha_0 - 1} (1-\theta)^{n - s + \beta_0 - 1}. \end{aligned} \tag{4}$$

[5] Throughout this paper we use uppercase P to represent probabilities and lowercase p to represent probability densities. Moreover, we place the argument of P in parentheses (as in $P(s|\theta)$) if it is an elementary quantity like s. If the argument of P is a composite predicate as in $P[s > n/2|\theta]$, then we place the argument in square brackets.

[6] The astute reader will note that the likelihood in Equation 1 is conditional on the number of samples, n, as well as θ. The same will be true of all the probabilities in this paper so, for the sake of simplicity of expression, we will leave this dependence implicit.

[7] For a brief discussion of the use of conjugate priors in Bayesian inference, see Appendix A.2.

[8] Of course, the prior probability is dependent on α_0 and β_0 and might just as well be written as $p(\theta|\alpha_0, \beta_0)$. The same is true of any other probabilities that depend on the prior. As with the case of dependence on the number of samples, n, we will leave this dependence implicit for the sake of simplicity of expression.

If we assume that α_0 and β_0 are integers then, since s and n are integers, the arguments of our Gamma functions are integers and we may use the fact that, for a positive integer argument n, $\Gamma(n) = (n-1)!$ to write

$$p(\theta|s) = \frac{(n+\alpha_0+\beta_0-1)!}{(s+\alpha_0-1)!(n-s+\beta_0-1)!}\,\theta^{s+\alpha_0-1}(1-\theta)^{n-s+\beta_0-1}. \tag{5}$$

It is worth noting that using a prior with α_0 and β_0 equal to integers > 1 is equivalent to using as a prior the posterior from an earlier application that used a flat prior to analyze data with $\alpha_0 - 1$ ones and $\beta_0 - 1$ zeros.

The posterior mean is given by Equation 60,

$$\mathrm{E}(\theta|s) = \frac{s+\alpha_0}{n+\alpha_0+\beta_0}, \tag{6}$$

the posterior variance by Equation 62,

$$\mathrm{Var}(\theta|\boldsymbol{x}) = \frac{(s+\alpha_0)(n-s+\beta_0)}{(n+\alpha_0+\beta_0)^2(n+\alpha_0+\beta_0+1)}, \tag{7}$$

and the posterior mode by Equation 64

$$\theta_{\max} = \frac{s+\alpha_0-1}{n+\beta_0+\alpha_0-2}. \tag{8}$$

In the next major section (4) we will explain how we will use the posterior for θ to test for improvement in predictive capability. Before we proceed, however, we want to discuss, in the next two subsections, two questions that may have occurred to the reader: The question of "statistical significance"; and the question of the role of experimental uncertainty (or error).

3.1 "Statistical Significance"

It is important to understand that the use of the Bayesian posterior for θ in our analysis automatically takes into account what is usually termed "statistical significance." The posterior distribution becomes more sharply peaked as more data are included in the analysis, guaranteeing that our confidence in our results increases appropriately as we include more data. This may be illustrated with a simple example.

Two typical posterior distributions for the binomial model are shown in Figure 1. The priors for θ are flat in both cases. The ratio of successes to samples, s/n, equals 0.6 in both cases, but the sample sizes differ by a factor of four. Any reasonable estimates for θ that one might make from the two posteriors would be very close: The posterior medians are very close (0.58811 for the 10 sample case and 0.59676 for the 40 sample case) as are the posterior means (0.58333 for the 10 sample case and 0.59524 for the 40 sample case), and the posterior modes are identical, $\theta_{\max} = s/n = 0.6$. On the other hand, the posterior for the larger sample size is clearly much sharper, which shows how the Bayesian posterior automatically reflects the effects of sample size—even though any reasonable estimates for θ that one might make from the two posteriors would be very close, any reasonable error bounds that one would put on the estimate will be much narrower in the case of the larger sample size.

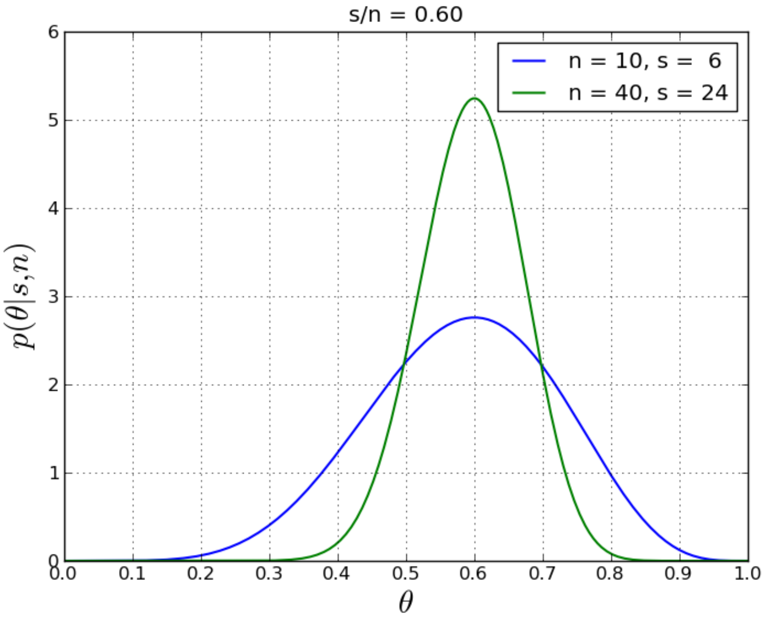

Figure 1: Typical posteriors from inference on the binomial model, assuming a flat prior for θ. The two cases correspond to different sample sizes but the same ratio of successes to samples, s/n. The posterior for the larger sample size is clearly much sharper, even though any reasonable estimates for θ that one might make from the two posteriors would be very close.

3.2 The Role of Experimental Uncertainty

Modeling experimental uncertainties for legacy experiments is often extremely difficult. Often, experimentalists have not provided any experimental uncertainties at all. When they have, it is often difficult to interpret the quantities they have provided. For example, it may be unclear whether the stated uncertainties were intended as one-sigma or as two-sigma uncertainties, or as some kind of strict bounds. If the experiments were performed many years before the statistical analysis, there may be no way to answer such questions as the experimentalists may be unavailable due to career changes, retirement, or death. Even when the experimentalists are available, memories may be faulty or may fail altogether.

The interpretation of experimental uncertainties becomes even more difficult when one is analyzing multiple experiments performed over a long period of time by different experimentalists, or even different generations of experimentalists. Uncertainties in outputs from different experiments may have been assessed with very different methodologies—everything from formal statistical analysis to expert judgment (which may itself derive from anything from long personal experience to folk knowledge) to a "scientific wild-assed guess" is likely to show up.

In light of these difficulties (which can show up in new as well as legacy experiments), it is good to have a statistical methodology available that is not dependent on detailed modeling of the experimental uncertainties. Our approach requires no such modeling.

One may reasonably ask, then, whether experimental uncertainty plays any role in our approach at all. It does, but only implicitly. In effect, the experimental uncertainty is aggregated with the modeling error in the codes to produce an overall assessment of predictions. If experimental error becomes large compared to the differences between legacy and new predictions, then the probability that a new prediction will be better than a legacy prediction will become more random, that is, θ will move toward $1/2$.

It would certainly be possible to incorporate knowledge about experimental uncertainties into our analysis. On the simplest level, it makes sense to take a quick look at predictions from legacy and new codes and to ask if they are generally within experimental uncertainty (assuming that one has at least a rough idea of what the experimental uncertainty should be). We think that this should be done in any case.

If predictions from both codes are generally within experimental uncertainties but one code does better than the other in a large percentage of the cases, then one should probably reexamine one's assumptions about experimental uncertainties, just as one would if a linear fit to some data produced a line that fit the data much better than the assumed experimental uncertainties should allow. If it is consistently possible to predict results to within a small fraction of experimental uncertainty, one's assumed uncertainty is probably too high.

More formally, it would be possible to separate the discrepancy between the prediction and the experimental value into a part due to experimental error and a part due to modeling error. Such an approach would require strong assumptions about the statistics of the predictions and of the experimental results and would lose the simplicity of the binomial model. Extending our analysis in this direction is not currently one of our priorities but we would be open to doing so if we can be convinced that it would be valuable.

4 Testing for Improvement in Predictive Capability Based on Binomial Inference

Our next task is to explain how we will use the posterior for θ to decide if we have observed improvement in predictive capability. As discussed in the Introduction, the critical question is whether or not θ is greater than 1/2: If it is, then predictive capability has improved; if it is not, there has been no improvement. In the language of conventional hypothesis testing, it makes sense to take the hypothesis that "θ is less than or equal to 1/2" as a null hypothesis (no improvement) with the complementary hypothesis, "θ is greater than 1/2", as the alternative hypothesis:

$$\begin{aligned} H_0 &\iff \theta \leq 1/2 \\ H_1 &\iff \theta > 1/2. \end{aligned} \tag{9}$$

These two hypotheses are exhaustive and mutually exclusive so, if we assign a probability ρ to the null hypothesis, we must assign probability $1 - \rho$ to the alternative hypothesis. (Note that, although we will use the language of null and alternative hypotheses, we will, unlike in frequentist analysis, treat the two hypotheses symmetrically.)

Since we have a posterior distribution for θ, we may use it to compute the posterior probability for H_0, which is just the posterior probability that

θ is less than or equal to 1/2,

$$P\big[H_0|s\big] \equiv P\big[\theta \leq 1/2|s\big] \equiv \int_0^{1/2} d\theta\, p(\theta|s). \tag{10}$$

Similarly, for H_1 we have,

$$P\big[H_1|s\big] \equiv P\big[\theta > 1/2|s\big] \equiv \int_{1/2}^{1} d\theta\, p(\theta|s). \tag{11}$$

(We also have the simple relation $P[H_1] = 1 - P[H_0]$.)

The integrals in Equations 10 and 11 can be evaluated in terms of a known function. Using Equation 4, we have

$$\begin{aligned} P\big[H_0|s\big] &= \int_0^{1/2} d\theta\, \mathrm{Be}(\theta|s+\alpha_0, n-s+\beta_0) \\ &= I_{1/2}(s+\alpha_0, n-s+\beta_0), \end{aligned} \tag{12}$$

where $I_x(\alpha, \beta)$ is the *regularized incomplete beta function*, and

$$\begin{aligned} P\big[H_1|s\big] &= \int_{1/2}^{1} d\theta\, \mathrm{Be}(\theta|s+\alpha_0, n-s+\beta_0) \\ &= 1 - I_{1/2}(s+\alpha_0, n-s+\beta_0). \end{aligned} \tag{13}$$

The capability to compute the regularized incomplete beta function is found in standard statistics packages such as the SciPy `stats` package.

Now, it is important to consider that, once we have a posterior for θ and the resulting posterior probabilities for H_0 and H_1, we may not need to specify anything further. In other words, it may be perfectly acceptable to report a simple statement of the form, “The probability that the new methodology is better than the old is x%”, where x is just the posterior probability for H_1. The sharpening of the posterior with sample size that was demonstrated in Figure 1 guarantees that such a statement already takes account of the question of “statistical significance”, so there is no essential factor that is being left out with such a statement.

On the other hand, a stronger statement of the form, “The new methodology is definitely better than the old,” may be required. In other words, we may be required to make a definite choice between the two hypotheses. In that case, it makes sense to consider “hypothesis tests” of H_0 and H_1. The outcomes of any such tests should, of course, be determined by the posterior probabilities for H_0 and H_1. We will investigate two tests for deciding if improvement in predictive capability has been observed:

1. A simple binary test, described in Section 4.1 immediately below; and

2. A more sophisticated three-way test, described in Section 4.2.

In addition to allowing us to make a definite statement about improvement of predictive capability, our analysis of these tests we will also provide an answer (actually more than one) to the question, "How many new predictions do we need to do in order to be able to decide if we have observed an improvement in predictive capability?"

In addition to the two tests mentioned above, we will also describe, in Section 4.3, posterior quantities that should prove useful to decision makers when the three-way test fails to produce a conclusive answer.

4.1 A Simple Binary Test

If we are forced to make a simple binary choice between H_0 and H_1, it is hard to see a justification for anything other than the following procedure: If the posterior probability for θ to be greater than 1/2 is greater than the probability for it to be less than or equal to 1/2, decide for H_1; if not, decide for H_0. More formally,

$$\begin{aligned} &\text{If } P\big[\theta > 1/2|s\big] > P\big[\theta \leq 1/2|s\big], \text{ decide for } H_1,\\ &\text{If } P\big[\theta > 1/2|s\big] \leq P\big[\theta \leq 1/2|s\big], \text{ decide for } H_0. \end{aligned} \tag{14}$$

(Note that the two conditions on the posterior distribution for θ corresponding to deciding for H_1 and H_0 are equivalent to the posterior median being greater than 1/2 on the one hand, or less than or equal to 1/2 on the other.)

One might expect that, in order to perform this test, we would need to integrate the posterior for θ (Equations 4 or 5) from 0 to 1/2 (or equivalently, from 1/2 to 1). However, we show in Appendix A.4 that we may simply examine the posterior expectation, given by Equation 6, because the posterior median is greater than 1/2 if and only if the posterior expectation is greater than 1/2. Moreover, both of these conditions are equivalent to the posterior mode being greater than 1/2 so that almost all simple binary tests that one might reasonably propose are equivalent.

Using our formula for the posterior expectation, Equation 6, we can represent the situation where the posterior expectation is greater than 1/2, and thus where the posterior median and mode are greater than 1/2 and the bulk of the posterior probability lies to the right of 1/2, thusly:

$$\begin{aligned} \mathrm{E}(\theta|s) > \frac{1}{2} &\iff 2\left(s+\alpha_0\right) > n+\alpha_0+\beta_0 \\ &\iff s > \frac{1}{2}\left(n+\beta_0-\alpha_0\right). \end{aligned} \tag{15}$$

For notational convenience, we will define here the set of values of s corresponding to H_0 as S_0 and the set of values of s corresponding to H_1 as S_1,

$$\begin{aligned} S_0 &\equiv \left\{ s \,\middle|\, s \leq \frac{1}{2}\left(n+\beta_0-\alpha_0\right) \right\} \\ S_1 &\equiv \left\{ s \,\middle|\, s > \frac{1}{2}\left(n+\beta_0-\alpha_0\right) \right\} \end{aligned} \tag{16}$$

remembering, of course, that s is restricted to the range $0 \leq s \leq n$. Then, our proposed test may be implemented as:

$$\begin{aligned} &\text{If } s \in S_1 \text{, decide for } H_1; \\ &\text{If } s \in S_0 \text{, decide for } H_0. \end{aligned} \tag{17}$$

We discuss this test and its implications in Section 5 below.

4.2 A Three-Way Test Based on the Posterior Confidence

Beyond deciding if we have or have not observed an improvement in predictive capability, we will want to know the (posterior) probability that our decision is correct. Obviously, a posterior probability of, say, 0.6 that θ is greater than 1/2 and a corresponding probability of 0.4 that it is not will not constitute strong evidence for an improvement in predictive capability. Thus, in order to state with confidence that we have observed an improvement, we will want to require some lower limit to this probability, such as, for example, 0.9. Similarly, in order to state with confidence that we have *not* observed an improvement, we will want to require some lower limit, say 0.9, to the probability that θ is less than or equal to 1/2.[9]

If the posterior probability for H_0 is greater than or equal to the posterior probability for H_1 (or, equivalently, if the posterior expectation is less than or equal to 1/2), then the posterior probability that our choice is correct will be given by the posterior probability for H_0, $P\big[H_0|s\big]$. Similarly, If the posterior probability for H_1 is greater than the posterior probability for H_0 (or, equivalently, if the posterior expectation is greater than 1/2), then the posterior probability that our choice is correct will be given by the posterior probability for H_1, $P\big[H_1|s\big] = 1 - P\big[H_0|s\big]$. We will call this probability the *posterior confidence* or just the *confidence*[10]. Specifically, we define the posterior confidence, C, to be

$$C \equiv C(s) \equiv P\big[H_0|s\big], \text{ when } P\big[H_1|s\big] \leq 1/2; \text{ and} \tag{18}$$

$$C \equiv C(s) \equiv P\big[H_1|s\big], \text{ when } P\big[H_1|s\big] > 1/2. \tag{19}$$

Using Equations 12 and 13, we have analytic expressions for the confidence,

$$C(s) = I_{1/2}(s + \alpha_0, n - s + \beta_0), \qquad \text{when } P\big[H_1|s\big] \leq 1/2; \text{ and} \tag{20}$$

$$C(s) = 1 - I_{1/2}(s + \alpha_0, n - s + \beta_0), \text{ when } P\big[H_1|s\big] > 1/2. \tag{21}$$

(Of course, we can, as discussed above in Section 4.1, determine easily whether or not $P\big[H_1|s\big] \leq 1/2$ by determining whether or not s is less than or equal to $\frac{1}{2}\left(n + \beta_0 - \alpha_0\right)$.) Note that the confidence will never be less than 0.5, because the posterior mean will always be on the side of 1/2 where the most posterior probability lies. (As proven in Appendix A.4.)

Figure 2 shows posterior confidence as function of s/n for various sample sizes, assuming a flat prior for θ. Note that, because our data, s, are discrete, the confidence is defined only at the markers and not in between—we have shown the lines connecting the markers only to aid in reading the figure. The figure shows clearly that:

1. The confidence is never less than 0.5;

[9] Note that our ability to characterize our confidence in our choice with a probability is dependent on our having a posterior distribution for θ and, thus, is a specifically Bayesian capability which is not available in non-Bayesian hypothesis testing.

[10] "Posterior confidence" and "confidence" are not standard terminology—they have been adopted here strictly for convenience.

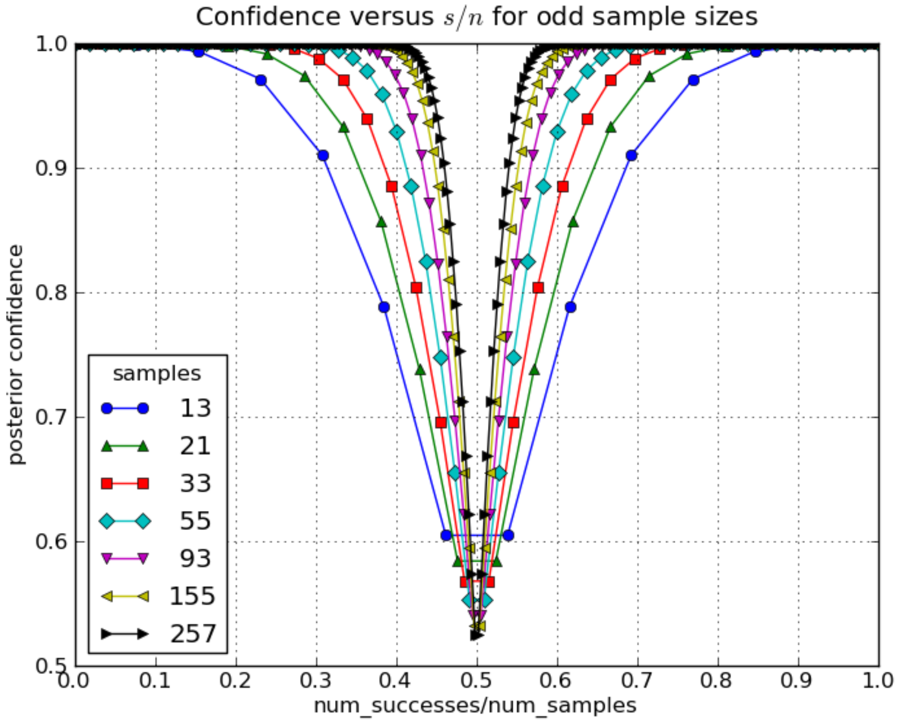

Figure 2: Posterior confidence as a function of s/n for various sample sizes, assuming a flat prior for θ. [Figure: figs/confidence_examples.png]

2. For fixed sample size, the confidence increases as the measured s/n moves away from $1/2$; and

3. Except when $s/n = 1/2$, the confidence increases with sample size for a given s/n.

In addition, all computations with even sample sizes (not shown on the figure) give a confidence of exactly 0.5 when $s = n/2$, as one would expect from the symmetry of the problem in that case.

If we go beyond a simple binary choice and demand a given level of confidence in our decision, $\phi : 0 < \phi < 1$, we are led naturally to a three-way choice between: 1) Decide definitely for H_1 (if the posterior probability of H_1 is greater than or equal to ϕ); 2) Decide definitely for H_0 (if the posterior probability for H_0 is greater than or equal to ϕ); 3) Decide that we do not know (because the posterior does not give the required level of confidence in either hypothesis).

Formally speaking, we consider two mutually exclusive (but not exhaustive) possible outcomes of our Bayesian analysis:

$$\begin{aligned} O_0 &\iff P\big[H_1|s\big] \leq 1/2 \text{ and } C(s) \geq \phi\text{; and} \\ O_1 &\iff P\big[H_1|s\big] > 1/2 \text{ and } C(s) \geq \phi. \end{aligned} \tag{22}$$

Of course, we know how to determine when the posterior probability of H_1 is greater than 1/2 from our data and prior through Equation 17 so we may write our two outcomes as:

$$\begin{aligned} O_0 &\iff s \in S_0 \text{ and } C(s) \geq \phi\text{; and} \\ O_1 &\iff s \in S_1 \text{ and } C(s) \geq \phi. \end{aligned} \tag{23}$$

Then, our test is:

$$\begin{aligned} &\text{If } O_0\text{, decide definitely for } H_0\text{;} \\ &\text{If } O_1\text{, decide definitely for } H_1\text{;} \\ &\text{If neither } O_0 \text{ nor } O_1\text{, decide that we do not know.} \end{aligned} \tag{24}$$

We discuss this test further in Section 5.2.

4.3 Examining the Width of the Posterior When the Confidence is Too Low

It should be obvious that having to choose the third case in Equation 24 is not likely to be very satisfactory, even though it will be the data that dictate our choice. Fortunately, there is a sensible “plan B” that we can employ in this case. The data will not tell us to choose the third case except when it is also telling us that θ is close to 1/2. Now, in practice, there will certainly be some region around $\theta = 1/2$ where the improvement in predictive capability effectively quantified by $\theta - 1/2$ is simply too small to matter. Thus, when the three-way test is inconclusive, the width of the posterior distribution of θ, together with a point estimate for θ which will necessarily be close to 1/2, will be useful quantities for guiding any decisions that need to be made (like, for example, decisions about how many more simulations should be done).

Any reasonable measure of the width of the posterior and any reasonable measure of central tendency should be satisfactory in this context. The mean of the posterior is given by Equation 6 and the standard deviation of the posterior may be obtained immediately by taking the square root of the variance as given by Equation 7. We propose reporting these quantities, especially when the three-way test is inconclusive.

In general, whenever one works with discrete hypotheses that are defined by partitioning the range of a continuous variable (like θ) into disjoint regions, there will be a problem similar to the one we have just discussed when the variable is close to the boundary between the regions. Thus, it seems to us that the kind of plan B that we have just outlined should be a feature of hypothesis tests in many common situations.

We present results on the problem of predicting the posterior standard deviation in Section 5.3 below.

5 Evaluating and Predicting the Results of the Tests

In our hypothesis testing described in the preceding section, we proposed to use two quantities computed from the data to decide whether or not we have observed an improvement in predictive capability:

1. The number of ones in our data, s, which tells us whether or not the posterior probability for improvement, $P\big[H_1|s\big]$, is greater than the posterior probability for no improvement, $P\big[H_0|s\big]$ and thereby provides us with the binary test described in Section 4.1; and

2. The posterior confidence defined by Equations 18 and 19, which allowed us to define our three-way test in Section 4.2.

Once we have taken our data and computed the posterior density for θ, these two quantities will have perfectly definite values. However, before we have taken any data, these quantities are unknown and, hence, must be described as random variables characterized by probability distributions, which we will call *predictive* distributions. Bayesian statisticians sometimes call them, perhaps more descriptively, *preposterior* distributions [1, chapter 7].

In order to determine how many pre- or post-dictions we will need to do with our new simulation methodology, we need to characterize the predictive probability distributions for these two posterior quantities. In addition, examining the predictive distributions will give us insight into our binary and three-way hypothesis tests, showing us, for example, the situations in which they can and cannot be expected to give definite answers.

For each of the two posterior quantities, we will begin by characterizing the predictive distribution of the quantity, as a function of the true value of θ. There will not be anything particularly "Bayesian" about this characterization (other than the fact that the two quantities come from a posterior distribution). Then, we will characterize the predictive distribution, averaged over the true value of θ. This will be a Bayesian analysis that depends on the prior density that we assign to θ, before taking any data.

We present our analysis of the predictive distribution of s in Section 5.1 immediately below and our analysis of the predictive distribution of the confidence in the following Section 5.2.

In addition to the two quantities just discussed, we have also identified (in Section 4.3) a role for the posterior mean and standard deviation in informing decision making based on our analysis, especially when the three-way test does not produce a definite result. The primary predictive quantity associated with these posterior quantities that we examine is the predictive mean of the posterior standard deviation, which we do in Section 5.3 below.

5.1 The Binary Test: The Probability of Making Type 1 and Type 2 Errors

In evaluating a proposed binary test, frequentist hypothesis testing makes use of the concepts of Type 1 and Type 2 errors. A *Type 1* error occurs when one incorrectly rejects the null hypothesis (a false positive) and a *Type 2* error occurs when one incorrectly fails to reject the null hypothesis (a false negative). Frequentists rely almost exclusively on the predictive probabilities of making Type 1 and Type 2 errors for quantitative evaluation of a test.

Although our approach here is more symmetric than the frequentist one and we will have the posterior confidence to guide us in evaluating the results of any test, the concepts of Type 1 and Type 2 errors and their predictive probabilities are still useful. We have chosen as our null hypothesis, H_0, the proposition that $\theta \leq 1/2$, and as our alternative hypothesis, H_1, the proposition that $\theta > 1/2$. Then, we will have made a Type 1 error if we have inferred that $\theta > 1/2$ when it is actually the case that $\theta \leq 1/2$. Conversely, we will have made a Type 2 error, if we have inferred that $\theta \leq 1/2$ when it is actually the case that $\theta > 1/2$.

Given θ, we already know the distribution of s, as given by Equation 1. Based on this, we can compute the probability of making a Type 1 error, as a function of the true value of θ. A Type 1 error can only occur when the true value of θ is less than or equal to 1/2. The probability of making such an error is equal to the predictive probability (conditioned on the value of θ) that the posterior mean will be greater than 1/2, in other words, that s will be greater than $\frac{1}{2}(n + \beta_0 - \alpha_0)$. Writing in terms of the set S_1 defined in Equation 16, this probability is,

$$P\big[s \in S_1|\theta\big] = \sum_{\{s|s\in S_1\}} \binom{n}{s} \theta^s (1-\theta)^{n-s}, \qquad \theta \leq 1/2. \tag{25}$$

Similarly, we can compute the probability of making a Type 2 error, conditioned on the true value of θ. A type 2 error can only occur when the true value of θ is greater than 1/2 and the probability of making such an error is equal to the predictive probability that s will be less than or equal to $\frac{1}{2}(n + \beta_0 - \alpha_0)$, conditioned on the true value of θ. Again using Equation 16, this probability is,

$$P\big[s \in S_0|\theta\big] = \sum_{\{s|s\in S_0\}} \binom{n}{s} \theta^s (1-\theta)^{n-s}, \qquad \theta > 1/2. \tag{26}$$

It is worthwhile plotting these functions for the simple case of a flat prior ($\alpha = \beta = 1$). In this case, the set S_0 of values of s that correspond to H_0 is $\{s \mid s \leq n/2\}$ and S_1 is $\{s \mid s > n/2\}$ (from Equation 16).

Figure 3 shows the predictive probabilities for Type 1 and Type 2 error for the case of a flat prior when the number of samples is odd. (The reason for separating the cases of odd and even sample sizes will be explained below.) The different sample sizes have been chosen to be approximately equally spaced logarithmically. Since our H_0 lies to the left of $\theta = 1/2$ and our H_1 lies to the right, the curve to the left of 1/2 gives the probability of making a Type 1 error and the curve to the right of 1/2 gives the probability of making a Type 2 error. (Of course, the boundary case $\theta = 1/2$ is in H_0 but even if it were in H_1 the figure would not change.) The figure is symmetric around $\theta = 1/2$.

The figure shows that the probability of error if the true value of θ is 1/2 is 50%, regardless of the number of samples. This should really not be surprising. If $\theta = 1/2$, then s is just as likely to be greater than $n/2$ as to be

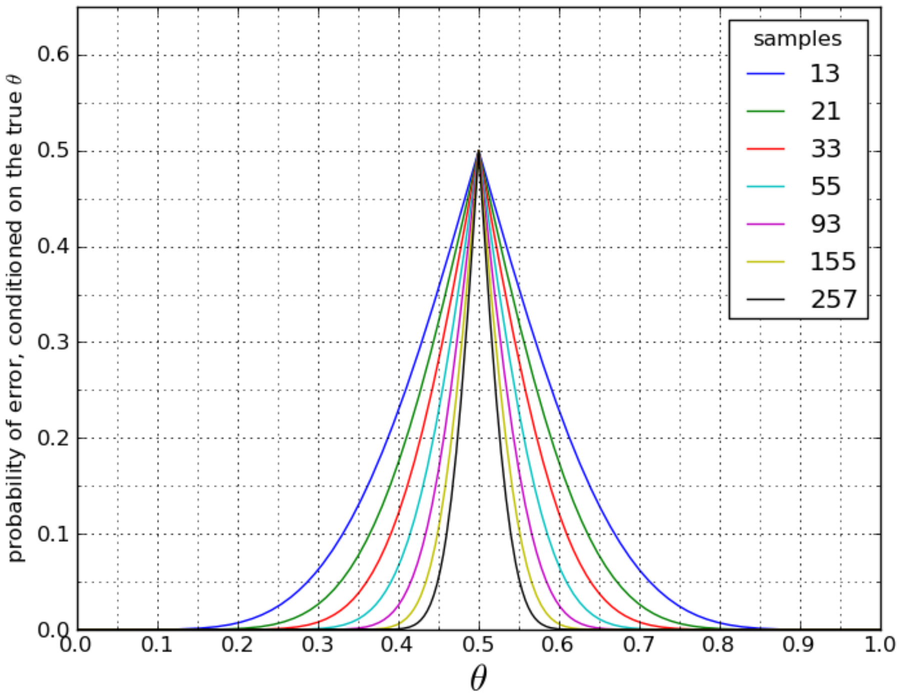

Figure 3: Probability of error, conditioned on the true value of θ, for the binary test, assuming a flat prior for θ. For $\theta \leq 1/2$ the curve gives the probability for a Type 1 error; for $\theta > 1/2$, the probability for a Type 2 error. The number of samples, n, is odd.

less and we are as likely to incorrectly decide for H_1 as to correctly decide for H_0. The root of this indeterminacy lies in the fact that the case $\theta = 1/2$ is in our H_0 but is arbitrarily close to points in H_1. Obviously, we could have chosen to put the case $\theta = 1/2$ is in our H_1 but we would still have the same issue. As long as we are partitioning the (continuous) range of θ into two disjoint regions, we are always going to see this behavior when the true value of θ lies on the boundary between the regions.

For all other values of the true θ, however, the figure shows clearly that the probability of error goes down monotonically with the number of samples. Not surprisingly, the probability of error is higher the closer the true θ is to 1/2. A few characteristic values are presented in Table 1. From the table

n	l-10%	u-10%	l-5%	u-5%
13	0.33	0.67	0.28	0.72
55	0.415	0.585	0.39	0.61
257	0.46	0.54	0.45	0.55

Table 1: Probabilities of error from Figure 3. n is the number of samples, l-10% is the lower value of θ at which the probability of error equals 10%, u-10% is the upper value at which the probability of error equals 10%, and l-5% and u-5% are the corresponding values for 5% probability of error.

we see that, for example, with 13 samples there is a less than 10% chance of error if θ is outside the range $[0.33, 0.67]$ and a less than 5% chance of error if θ is outside the range $[0.28, 0.72]$.[11] Roughly speaking, we can do reasonably well if θ is no closer than ± 0.2 to 1/2. If we go to 55 samples, we can do reasonably well if θ is no closer than ± 0.1 to 1/2. 257 samples allows us to do well outside of $1/2 \pm 0.05$.

Figure 4 shows the plot analogous to Figure 3 when the number of samples is even. The first thing one notices about this plot is that it is dramatically asymmetric. In particular, the probability for making a Type 2 error at a given distance from $\theta = 1/2$ is larger than the probability of making a Type 1 error at the same distance.

This asymmetry is a consequence of the discrete nature of our data, s, and the asymmetry between our null and alternative hypotheses that results from deciding to put the case $\theta = 1/2$ in H_0 rather than H_1. In the previous case of an odd number of samples, every possible data value leading to a decision for H_0 had a complementary value on the other side of 1/2 leading to a decision for H_1. Hence, the symmetry of Figure 3. With an even number of samples, one possible value for s is exactly $n/2$ and in this case we decide for the null hypothesis as specified by Equation 17. There is no complementary data value leading to a decision for H_1, so the figure is asymmetric. Of course, we could choose to put the case $\theta = 1/2$ in H_1 but this would just have the effect of flipping the asymmetry.

A table analogous to Table 1 for the case of even sample sizes is given in Table 2. As long as θ does not get too close to 1/2 and the sample size is not too small, the two figures give similar values for the probability of error

[11] The very similar ranges for 10% and 5% probability of error are due to the vary rapid increase in the probability of error as θ approaches 1/2, as is apparent in Figure 3.

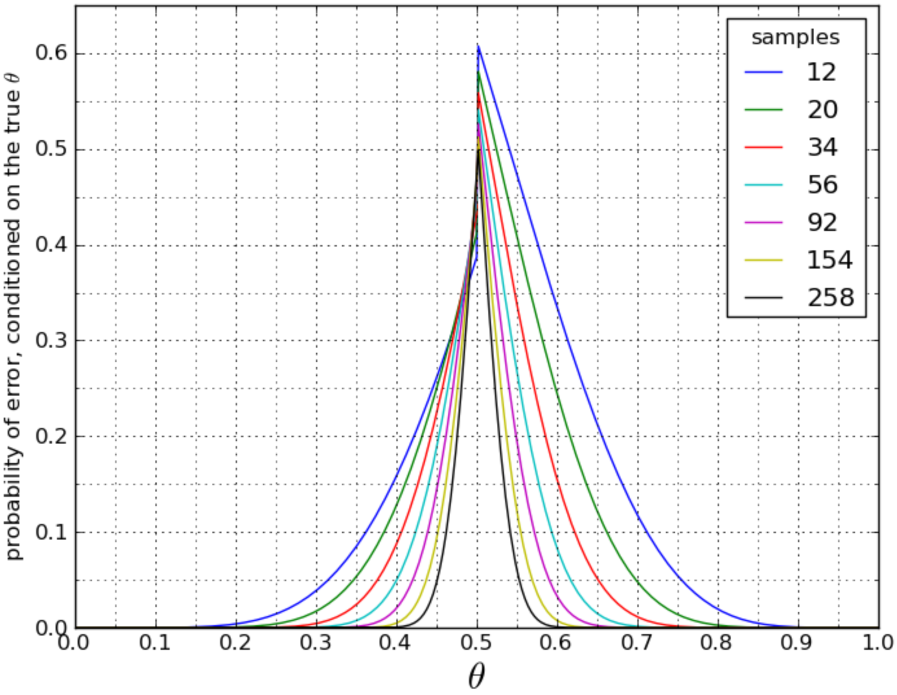

Figure 4: Probability of error, conditioned on the true value of θ, for the binary test, assuming a flat prior for θ. For $\theta \leq 1/2$ the curve gives the probability for a Type 1 error; for $\theta > 1/2$, the probability for a Type 2 error. The number of samples, n, is even.

for similar sample sizes.

The import of all this will become more clear when we examine the overall predictive probabilities for making Type 1 and Type 2 errors. The overall predictive probability of making a Type 1 error is just the joint predictive probability that $s \in S_1$ and $\theta \leq 1/2$,

$$P\bigl[s \in S_1, \theta \leq 1/2\bigr], \tag{27}$$

where it is may be helpful to the reader to note that this quantity is implicitly dependent on the number of samples, n, and on the prior hyperparameters α_0 and β_0. This quantity has the form $P\bigl[s \in U, \theta_0 < \theta \leq \theta_1\bigr]$ with $U = S_1$, $\theta_0 = 0$, and $\theta_1 = 1/2$. Quantities of this form are evaluated in Appendix A.5.

n	l-10%	u-10%	l-5%	u-5%
12	0.36	0.71	0.31	0.75
56	0.42	0.59	0.40	0.62
258	0.46	0.54	0.45	0.55

Table 2: Probabilities of error from Figure 4. n is the number of samples, l-10% is the lower value of θ at which the predictive probability of error equals 10%, u-10% is the upper value at which the probability of error equals 10%, and l-5% and u-5% are the corresponding values for 5% probability of error.

Substituting the specific values for U, θ_0, and θ_1 above into Equation 74 in that appendix, and recognizing that $I_0(s+\alpha_0, n-s+\beta_0) = 0$, we have,

$$\begin{aligned} P\big[s \in S_1, \theta \leq 1/2\big] \\ &= \frac{\Gamma(\alpha_0+\beta_0)\, n!}{\Gamma(\alpha_0)\Gamma(\beta_0)\Gamma(n+\beta_0+\alpha_0)} \\ &\quad \cdot \sum_{s \in S_1} \frac{\Gamma(s+\alpha_0)\Gamma(n-s+\beta_0)}{s!\,(n-s)!}\, I_{1/2}(s+\alpha_0, n-s+\beta_0), \end{aligned} \tag{28}$$

where, once again, $I_x(\alpha, \beta)$ is the regularized incomplete beta function. For the case of a flat prior, Equation 77 gives the dramatic simplification,

$$P\big[s \in S_1, \theta \leq 1/2\big] = \frac{1}{n+1} \sum_{s \in S_1} I_{1/2}(s+1, n-s+1), \qquad \text{for } \alpha_0 = \beta_0 = 1. \tag{29}$$

Similarly, the overall probability of making a Type 2 error is just the joint probability that $s \in S_0$ and $\theta > 1/2$,

$$P\big[s \in S_0, \theta > 1/2\big] \tag{30}$$

Now, we substitute $U = S_0$, $\theta_0 = 1/2$, and $\theta_1 = 1$ in $P\big[s \in U, \theta_0 < \theta \leq \theta_1\big]$

and, recognizing that $I_1(s+\alpha_0, n-s+\beta_0) = 1$, Equation 74 gives us

$$\begin{aligned} P\big[s \in S_0, \theta > 1/2\big] \\ &= \frac{\Gamma(\alpha_0+\beta_0)\, n!}{\Gamma(\alpha_0)\Gamma(\beta_0)\Gamma(n+\beta_0+\alpha_0)} \\ &\quad \cdot \sum_{s\in S_0} \frac{\Gamma(s+\alpha_0)\Gamma(n-s+\beta_0)}{s!\,(n-s)!}\, \big(1 - I_{1/2}(s+\alpha_0, n-s+\beta_0)\big)\,. \end{aligned} \tag{31}$$

For the case of a flat prior, Equation 77 gives the simplified form,

$$\begin{aligned} P\big[s \in S_0, \theta > 1/2\big] & \qquad\qquad \text{for } \alpha_0 = \beta_0 = 1, \\ &= \frac{1}{n+1} \sum_{s\in S_0} \big(1 - I_{1/2}(s+1, n-s+1)\big)\,. \end{aligned} \tag{32}$$

It is important to note that, because the overall predictive probabilities for making Type 1 and Type 2 errors are computed with respect to the prior distribution for θ, the results are critically dependent on the choice of prior. Posterior results become independent of our choice of prior as we take more data but preposterior quantities are computed without the benefit of data and so remain dependent on priors no matter how large the proposed sample size. In particular, if we were to take a prior that, unlike our current uniform prior, weights the values of θ near 1/2 more than values near 0 and 1, the overall probabilities of Type 1 and Type 2 errors would increase, because values of θ near 1/2 have a greater predictive probability of producing these errors. The lesson here is that our results for overall probabilities of producing Type 1 and Type 2 errors are only as good as our choice of prior. Where this prior has a solid basis, as when it is the posterior from an earlier analysis on a substantial amount of data, then we can have a high level of confidence in our overall probabilities. Where the prior represents only very vague prior knowledge, the overall probabilities must be taken as suggestive but not definitive.

Figures 5 and 6 show the overall probabilities of error for even and odd sample sizes, respectively, out to a sample size of 70, assuming a flat prior for θ. With an odd number of samples, the probabilities for Type 1 and Type 2 errors are equal. With an even number of samples, Type 1 errors are less probable than Type 2 errors but the total probability of error at a given

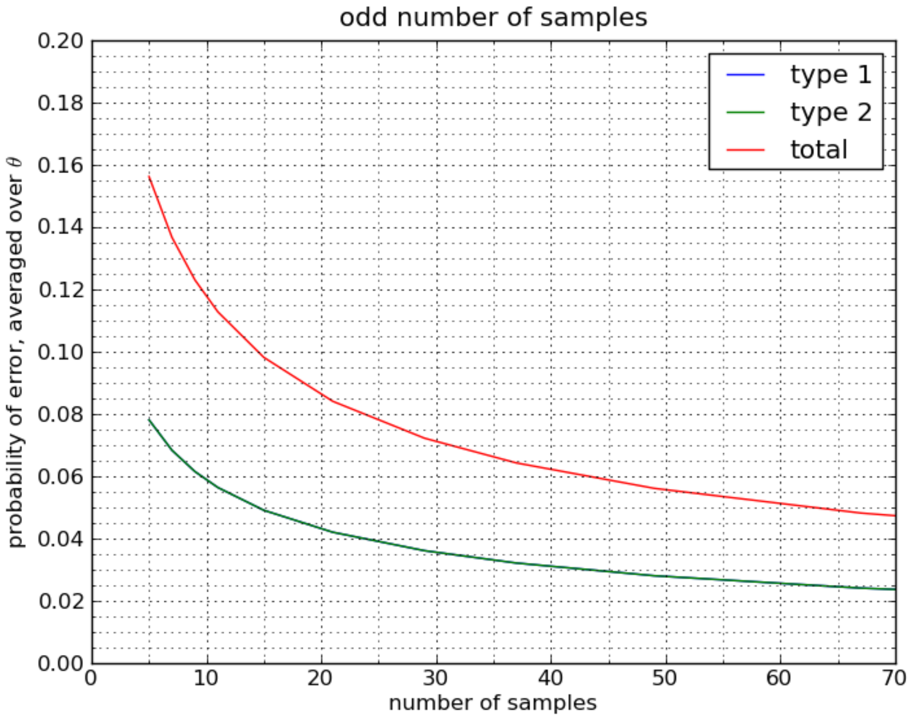

Figure 5: Probability of error, averaged over the true value of θ, for the binary test, assuming a flat prior for θ. The number of samples, n, is odd. The probabilities for Type 1 and Type 2 errors are equal and the green lies directly on top of (and overwrites) the blue curve.

sample size is almost identical to the case of odd sample size, once we get out to about 15 samples.

A couple of observation that can be made by inspecting these graphs are that the total probability of error is under 10% if the sample size is over about 15 and under 5% for sample sizes over about 65.

If we plot the total probabilities of error on log-log graphs, we see that the error is falling off as about $1/\sqrt{n}$ for large sample size, n, as is generally expected in statistics. Log-log plots are shown in Figures 7 and 8. These figures show that the total probability of error is about 2.54% at roughly 250 samples.

We conclude our discussion of the simple binary test with the following

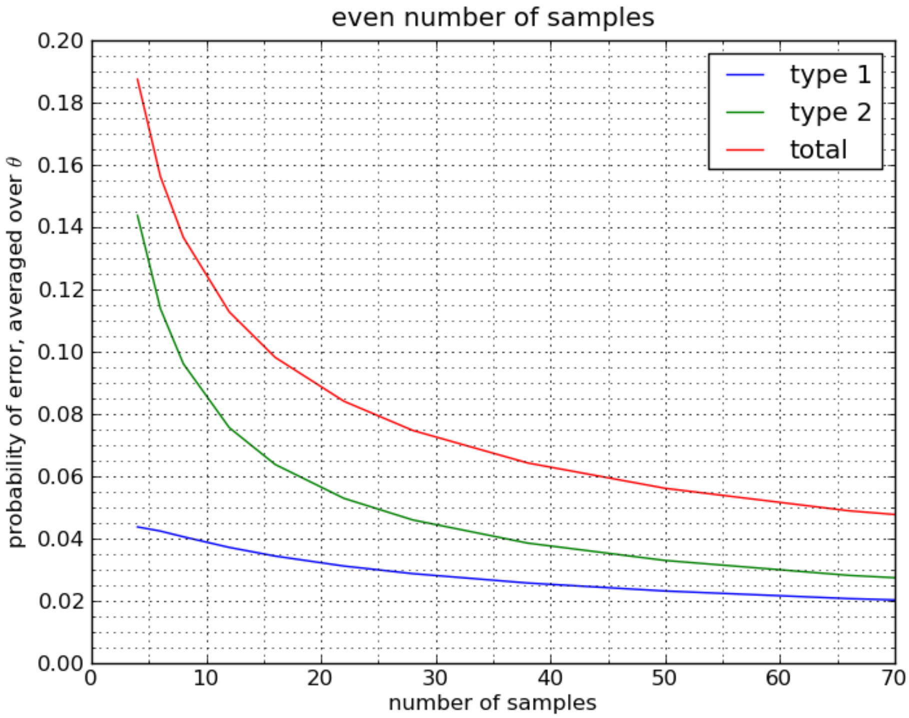

Figure 6: Probability of error, averaged over the true value of θ, for the binary test, assuming a flat prior for θ. The number of samples, n, is even. The probabilities for Type 1 and Type 2 errors are different, with Type 2 error always being more probable than Type 1.

observations about the case of a flat prior for θ:

1. The probability of error when the true value of θ is $1/2$ is always 50%, regardless of sample size.

2. The probability of error for all other values of θ goes to zero with sample size but, at any given sample size, there is always a region around $1/2$ where the probability of error is rapidly increasing toward 50% at $\theta = 1/2$.

3. If we define this region more precisely as the region over which the probability exceeds either 10% or 5%, we get similar results for the

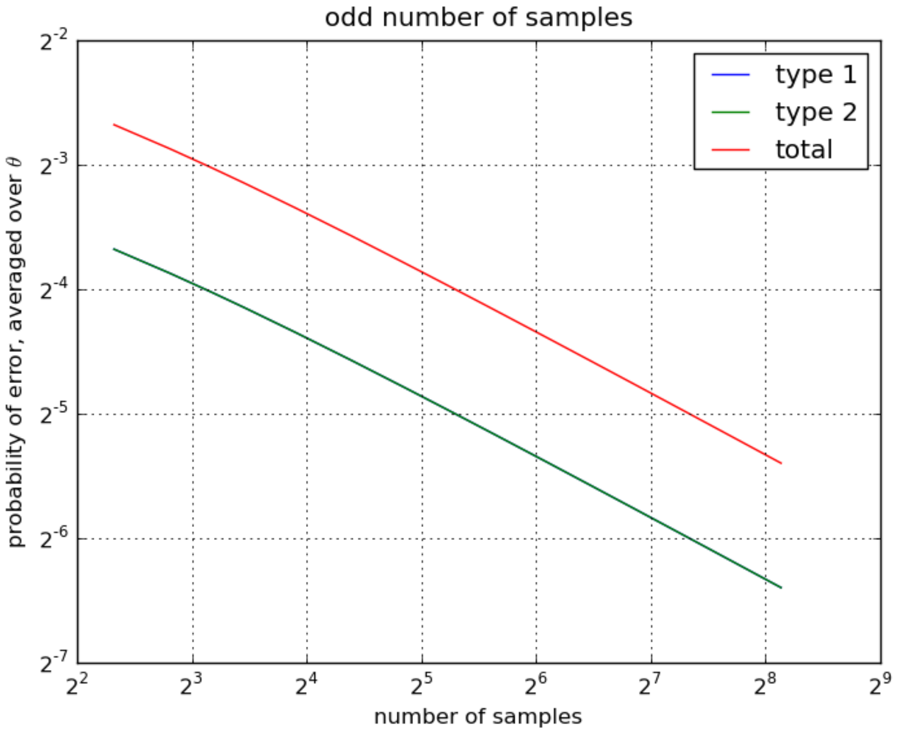

Figure 7: Probability of error, averaged over the true value of θ, for the binary test, assuming a flat prior for θ. The number of samples, n, is odd. Once again, the probabilities for Type 1 and Type 2 errors are equal and the green curve lies directly on top of and overwrites the blue curve.

width of the region. At 13 samples, the width of the region is roughly ±0.2, at 55 samples it is roughly ±0.1 and at 257 samples it is roughly ±0.05.

4. The total probability of error, averaged over θ, is under about 10% for sample sizes over about 15, under about 5% for sample sizes over about 65, and about 2.54% at about 250 samples.

Given the complexity of the plots with even sample sizes, we will restrict our attention hereafter to odd sample sizes.

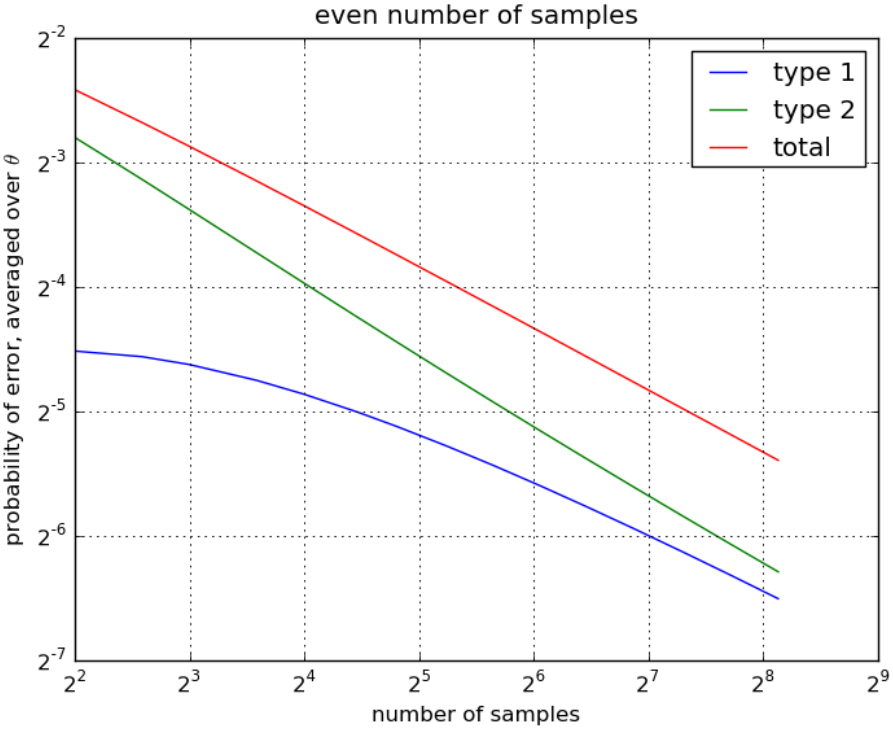

Figure 8: Probability of error, averaged over the true value of θ, for the binary test, assuming a flat prior for θ. The number of samples, n, is even. Once again, Type 2 error is always more probable than Type 1.

5.2 Predictive Probabilities for the Confidence

In the case of the predictive distribution for the confidence, the computation is a bit more complicated. We begin by reminding the reader that, for a fixed number of samples, n, there are only finitely many possible values for the integer-valued data, s: Specifically, $0 \leq s \leq n$. For each possible value of s and a given prior specified by (α_0, β_0) there is a particular posterior distribution for θ, $p(\theta|s)$, given by Equation 4 and, thus, a well-defined value for the posterior confidence defined by Equations 18 and 19 and given analytically by Equations 20 and 21. Moreover, we already know the predictive distribution for s given θ, $P(s|\theta)$, from Equation 1. Therefore, we can compute the predictive probabilities for the various values of the confidence—they are just the sums of the predictive probabilities of the corresponding values of s. Collectively, these probabilities are the predictive distribution for the confidence and we may compute any properties of this distribution that we want—the mean, the variance, quantiles, etc. However, it is important to be aware that, because the predictive distribution for the confidence is discrete rather than continuous, we cannot in general find a point corresponding exactly to a arbitrary quantile, such as 95%. Rather, if we wish to work with quantiles, we will need to compute quantities like "the smallest value of $s > \frac{1}{2}(n + \beta_0 - \alpha_0)$ for which the posterior confidence is at least 95%".

In general, the confidence that we achieve when we have made the correct decision (as with the simple binary test) has a very different significance than the confidence that we achieve when we have made the incorrect decision. We would hope that, when we have decided correctly, our confidence will be relatively high and when we have decided incorrectly, our confidence will be relatively low. Accordingly, in our analysis of the predictive distribution for the confidence, we will condition on both the true value of θ *and* the correctness of our simple binary test given by Equation 17. So, we will be interested in four cases:

1. $\theta \leq 1/2$ and $s \in S_0$,
2. $\theta > 1/2$ and $s \in S_1$,
3. $\theta \leq 1/2$ and $s \in S_1$, and
4. $\theta > 1/2$ and $s \in S_0$,

where S_0 or S_1 are defined in Equation 16 and membership of s in either S_0 or S_1 determines the result of the simple binary test as given by Equation 17. In the first two cases our simple binary test leads to the correct conclusion and in the last two cases, the incorrect conclusion. Case 3 corresponds to a Type 1 error in the simple binary test and case 4 to a Type 2 error.

In considering these cases, we will need to group together values of s that give us a particular answer to our binary test and which do so with the required level of confidence. In other words, we will be interested in the two sets of possible values of s, T_0 and T_1, given by,

$$T_0 \equiv \left\{ s \,\middle|\, s \leq \frac{1}{2}(n + \beta_0 - \alpha_0) \text{ and } C(s) \geq \phi \right\} \subset S_0 \tag{33}$$

and

$$T_1 \equiv \left\{ s \,\middle|\, s > \frac{1}{2}\left(n + \beta_0 - \alpha_0\right) \text{ and } C(s) \geq \phi \right\} \subset S_1, \tag{34}$$

where ϕ is our desired level of confidence and, of course, $0 \leq s \leq n$.

We will begin, as in our analysis of the simple binary test in Section 5.1 above, by computing the relevant predictive probabilities conditioned on the true value of θ. Then, we will compute overall predictive probabilities, averaged over the unknown value of θ.

5.2.1 Predictive Probabilities Conditioned on the True Value of θ

We will first consider the situation where our simple binary test produces the correct answer, cases 1 and 2 above. Given a value for θ, the predictive probability of achieving the desired confidence ϕ when the binary test gives the correct answer is $P\big[s \in T_0 \,|\, \theta\big]$ for $\theta \leq 1/2$ and $P\big[s \in T_1 \,|\, \theta\big]$ for $\theta > 1/2$. These quantities are easily computed from the likelihood, Equation 1,

$$\begin{aligned} P\big[s \in T_0 \mid \theta\big] & \qquad\qquad \theta \leq 1/2, \\ = \sum_{\{s|s\in T_0\}} & P(s|\theta) \\ = \sum_{\{s|s\in T_0\}} & \binom{n}{s}\theta^s(1-\theta)^{n-s} \end{aligned} \tag{35}$$

and,

$$\begin{aligned} P\big[s \in T_1 \mid \theta\big] & \qquad\qquad \theta > 1/2, \\ = \sum_{\{s|s\in T_1\}} & \binom{n}{s}\theta^s(1-\theta)^{n-s} \end{aligned} \tag{36}$$

A plot of these functions for the case of a flat prior and a variety of sample sizes is shown in Figure 9. The desired level of confidence is $\phi = 0.9$. For the most part, the figure behaves as we would expect—true values of θ that are farther from 1/2 are more likely to result in at least 90% confidence than values closer to 1/2 and larger sample size generally results in a higher probability of achieving at least 90% confidence than smaller sample size. However, there is a region around $\theta = 1/2$ where some of the curves for different sample sizes cross. This is due to the discrete nature of our data and

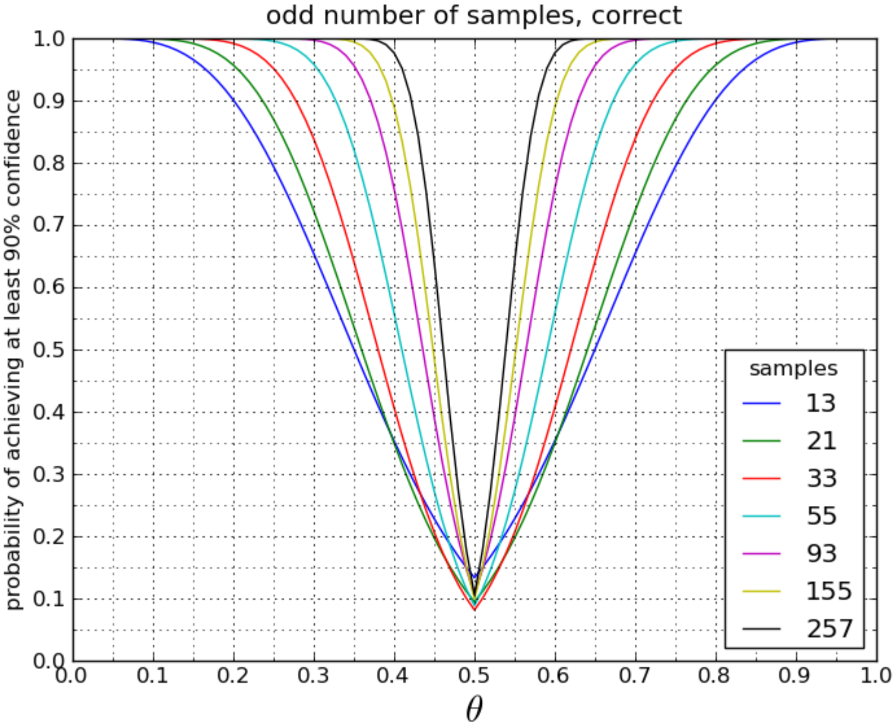

Figure 9: Predictive probability of achieving at least 90% confidence when the simple binary test gives the correct answer, assuming a flat prior for θ. The number of samples, n, is odd.

our resulting inability to precisely specify a quantile like “90% confidence”. As we increase our number of samples, the actual cutoff probability for including possible data values in the sums in Equations 35 and 36 fluctuates in the region just above 90% and we get the otherwise counterintuitive result of a larger number of samples sometimes having a slightly reduced predictive probability of achieving at least 90% confidence. The effect is largely restricted to the smaller sample sizes and the neighborhood of $\theta = 1/2$. This is apparently a well-known phenomenon when performing preposterior analysis of systems with discrete data [2].

It is also worth noting that the probability of (correctly) achieving at least 90% confidence given $\theta = 1/2$ jumps around 0.1 with no apparent pattern as we change the sample size. This behavior, too is a consequence of the

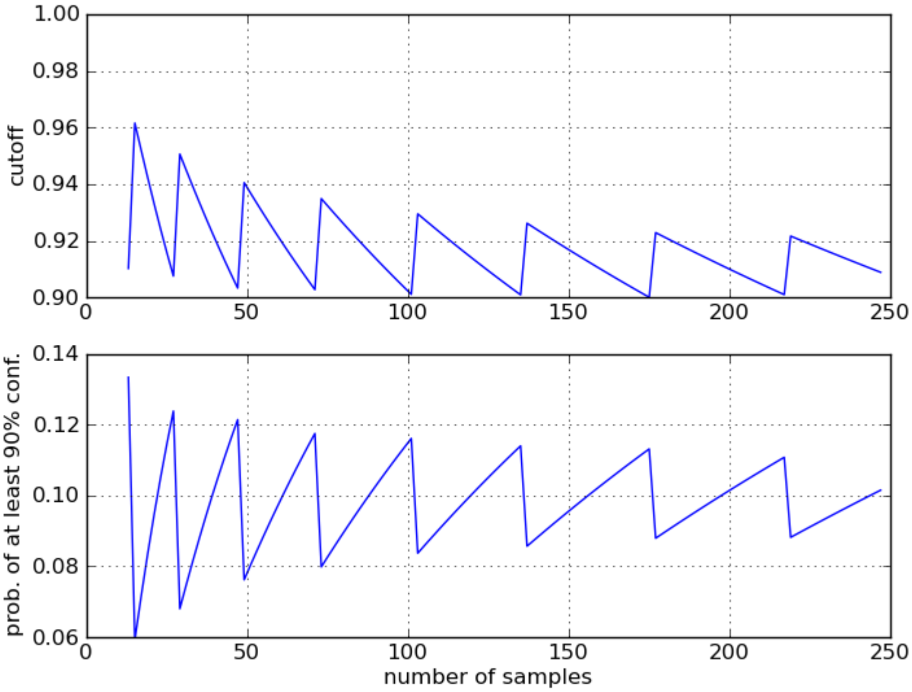

Figure 10: The horizontal axis for both plots is the sample size. The upper plot gives the actual cutoff for the confidence when we consider only data values s for which the associated posterior confidence is at least 90%. The lower plot shows the probability of achieving a confidence of at least 90% when the true value of θ is $1/2$.

changes in the effective cutoff probability as we change sample size. That this is so may be seen from Figure 10. The figure shows how the actual cutoff probability changes with sample size and the effect these changes have on the probability of achieving at least 90% confidence when $\theta = 1/2$. The vertical axis in the upper plot is the actual cutoff probability, that is, the smallest value of the posterior confidence, taken over all possible data values s, that is greater than or equal to 90%. The vertical axis in the lower plot is the probability of achieving at least 90% confidence when the true value of θ is $1/2$.

Note that, even with over 200 samples, the actual cutoff can be as high as

92% and it is as high as 96% when the number of samples is in the teens. It is clear that the fluctuations of the probability of achieving 90% confidence at $\theta = 1/2$ are driven by the cutoff—when the cutoff is well above 90%, the probability of achieving at least 90% confidence is relatively low and when the cutoff approaches 90%, the probability of achieving at least 90% confidence is relatively high. It appears that the probability of achieving at least 90% confidence when $\theta = 1/2$ is approaching 10% as the number of samples goes to infinity, but only very slowly.

n	l-90%	u-90%	l-80%	u-80%
13	0.20	0.80	0.25	0.75
55	0.33	0.67	0.35	0.65
257	0.42	0.58	0.44	0.56

Table 3: Probabilities of achieving at least 90% confidence from Figure 9. n is the number of samples, l-90% is the lower value of θ at which the predictive probability of achieving at least 90% confidence equals 90%, u-90% is the upper value at which the probability of achieving at least 90% confidence equals 90%, and l-80% and u-80% are the corresponding values for 80% probability of achieving 90% confidence.

It is instructive to tabulate some characteristic values from the plot in Figure 9 as we did for the simple binary test. The results are shown in Table 3. If we are willing to accept an 80% predictive probability of achieving 90% confidence, and are not concerned with a difference of θ from 1/2 of less than 0.15, then it appears that 55 samples will suffice. If, on the other hand, we are concerned with a difference of θ from 1/2 of more than 0.06 (again accepting 80% predictive probability for achieving our target confidence), then we will need many more samples, on the order of 250.

We next consider cases 3 and 4 from page 31 above, where the simple binary test produces the wrong answer. It is easy enough to derive formulae for the probability of achieving a given confidence and getting the wrong answer, analogous to Equations 35 and 36. The quantities we want are $P\big[s \in T_1 \,|\, \theta\big]$ for $\theta \leq 1/2$ and $P\big[s \in T_0 \,|\, \theta\big]$ for $\theta > 1/2$ and these evaluate to,

$$\begin{aligned} & P\big[s \in T_1 \mid \theta\big] \qquad\qquad \theta \leq 1/2, \\ & = \sum_{\{s|s \in T_1\}} \binom{n}{s} \theta^s (1-\theta)^{n-s} \end{aligned} \tag{37}$$

and,

$$\begin{aligned} P\big[s \in T_0 \mid \theta\big] & \qquad\qquad \theta > 1/2, \\ = \sum_{\{s \mid s \in T_0\}} \binom{n}{s} \theta^s (1-\theta)^{n-s} & \end{aligned} \tag{38}$$

The resulting plot, for the case $\phi = 90\%$ and a flat prior for θ, is shown in Figure 11. The probabilities are comfortingly low, even for the case of 13 samples. The probability of incorrectly achieving 90% confidence never exceeds 0.13 and is mostly much lower. We may have trouble achieving our desired level of confidence but, if we do, we can be reasonably certain that we have not become confident in the wrong conclusion.

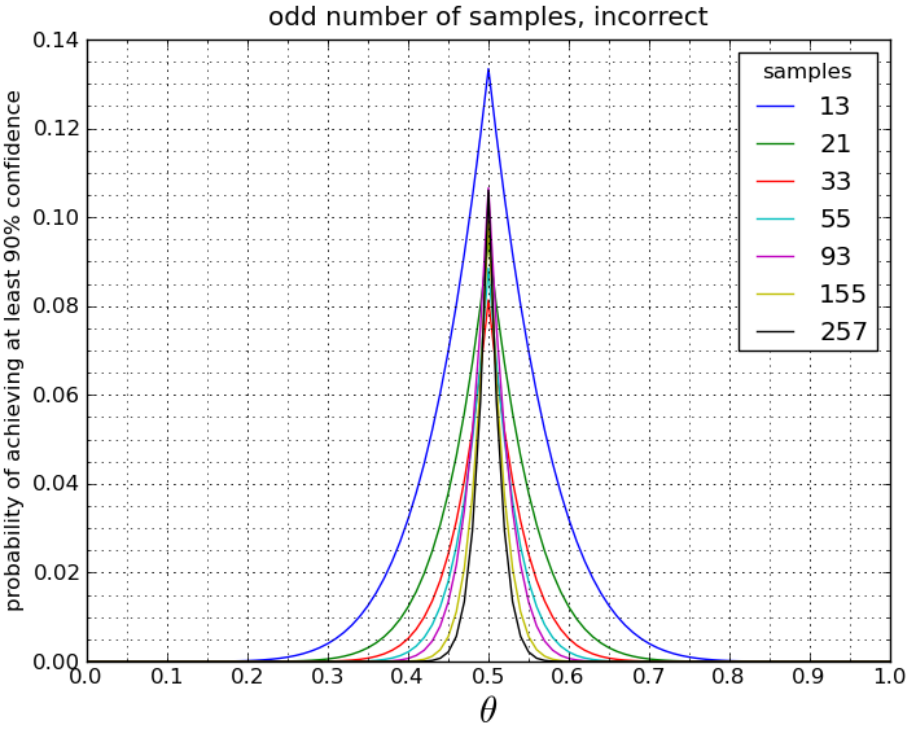

Figure 11: Predictive probability of achieving at least 90% confidence when the simple binary test gives the incorrect answer, assuming a flat prior for θ. The number of samples, n, is odd.

5.2.2 Predictive Probabilities Averaged over the True Value of θ

Now, we consider the overall predictive probability to achieve a given level of confidence, averaged over the true value of θ. As in the section above, we will first consider the situation where our simple binary test produces the correct answer, cases 1 and 2 on page 31 above.

The overall predictive probability that the simple binary test will give the correct answer and that we will achieve confidence ϕ is the sum of two probabilities corresponding to cases 1 and 2, respectively,

$$P\big[s \in T_0, \theta \leq 1/2\big] \tag{39}$$

and

$$P\big[s \in T_1, \theta > 1/2\big], \tag{40}$$

where T_0 and T_1 are defined by Equations 33 and 34, respectively. These quantities have the form $P\big[s \in U,\, \theta_0 < \theta \leq \theta_1\big]$ that is evaluated in Appendix A.5. Making the appropriate substitutions for U, θ_0, and θ_1, we have,

$$\begin{aligned} P\big[s \in T_0, \theta \leq 1/2\big] & \\ &= \frac{\Gamma(\alpha_0+\beta_0)\, n!}{\Gamma(\alpha_0)\Gamma(\beta_0)\Gamma(n+\beta_0+\alpha_0)} \\ &\quad \cdot \sum_{s \in T_0} \frac{\Gamma(s+\alpha_0)\Gamma(n-s+\beta_0)}{s!\,(n-s)!}\, I_{1/2}(s+\alpha_0, n-s+\beta_0) \end{aligned} \tag{41}$$

and

$$\begin{aligned} P\big[s \in T_1, \theta > 1/2\big] & \\ &= \frac{\Gamma(\alpha_0+\beta_0)\, n!}{\Gamma(\alpha_0)\Gamma(\beta_0)\Gamma(n+\beta_0+\alpha_0)} \\ &\quad \cdot \sum_{s \in T_1} \frac{\Gamma(s+\alpha_0)\Gamma(n-s+\beta_0)}{s!\,(n-s)!}\, \big(1 - I_{1/2}(s+\alpha_0, n-s+\beta_0)\big) \end{aligned} \tag{42}$$

For the situation where the binary test produces the wrong answer, the overall predictive probability that we will achieve confidence ϕ is the sum of two probabilities associated with cases 3 and 4,

$$P\big[s \in T_1, \theta \leq 1/2\big] \tag{43}$$

and

$$P\bigl[s \in T_0, \theta > 1/2\bigr], \tag{44}$$

which evaluate to

$$\begin{aligned} P\bigl[s \in T_1, \theta \leq 1/2\bigr] \\ &= \frac{\Gamma(\alpha_0 + \beta_0)\, n!}{\Gamma(\alpha_0)\Gamma(\beta_0)\Gamma(n + \beta_0 + \alpha_0)} \\ &\quad \cdot \sum_{s \in T_1} \frac{\Gamma(s + \alpha_0)\Gamma(n - s + \beta_0)}{s!\,(n-s)!}\, I_{1/2}(s + \alpha_0, n - s + \beta_0) \end{aligned} \tag{45}$$

and

$$\begin{aligned} P\bigl[s \in T_0, \theta > 1/2\bigr] \\ &= \frac{\Gamma(\alpha_0 + \beta_0)\, n!}{\Gamma(\alpha_0)\Gamma(\beta_0)\Gamma(n + \beta_0 + \alpha_0)} \\ &\quad \cdot \sum_{s \in T_0} \frac{\Gamma(s + \alpha_0)\Gamma(n - s + \beta_0)}{s!\,(n-s)!}\, \bigl(1 - I_{1/2}(s + \alpha_0, n - s + \beta_0)\bigr) \end{aligned} \tag{46}$$

For the case of a flat prior for θ, the overall probabilities of achieving at least 90% confidence for the correct and incorrect cases are given in Figures 12 and 13, respectively. In the case that the simple binary case produces the correct answer, we have an overall 80% predictive probability of achieving 90% confidence with 50 samples. 90% predictive probability occurs only with about 180 samples. In the case that the simple binary case produces the incorrect answer, the probability of achieving 90% confidence is already down to 1.8% at 13 samples and falls rapidly to about 0.6% at about 40 samples. Once again, we are very unlikely to incorrectly achieve 90% confidence.

It is appropriate to mention at this point that our choice of 90% as our target confidence is not entirely arbitrary. It is not difficult to compute the predictive mean of the confidence conditioned on the simple test giving the correct answer. The formulas are given in Appendix A.6 and the resulting plot in Figure 14. The salient point is that the predictive mean confidence is *never* less than 75%. It therefore seems to us that choosing a target confidence much less than 90% is likely to amount to no real target at all.

We conclude our investigation of the three-way test with the following observations about the case of a flat prior for θ:

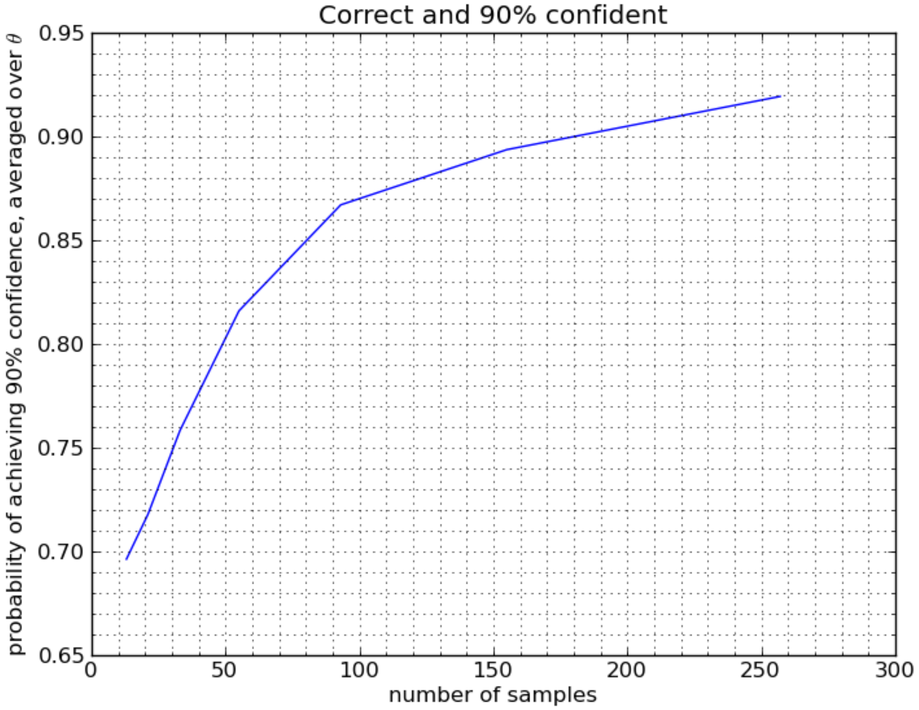

Figure 12: The overall predictive probability of achieving at least 90% confidence when the simple binary test gives the correct answer, assuming a flat prior for θ.

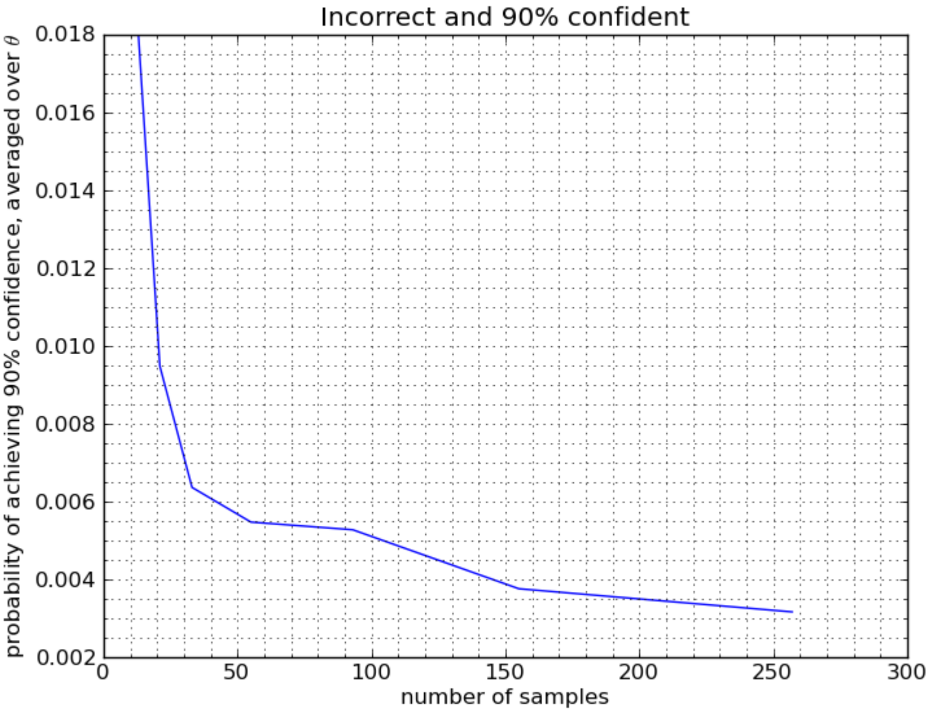

Figure 13: The overall predictive probability of achieving at least 90% confidence when the simple binary test gives the incorrect answer, assuming a flat prior for θ.

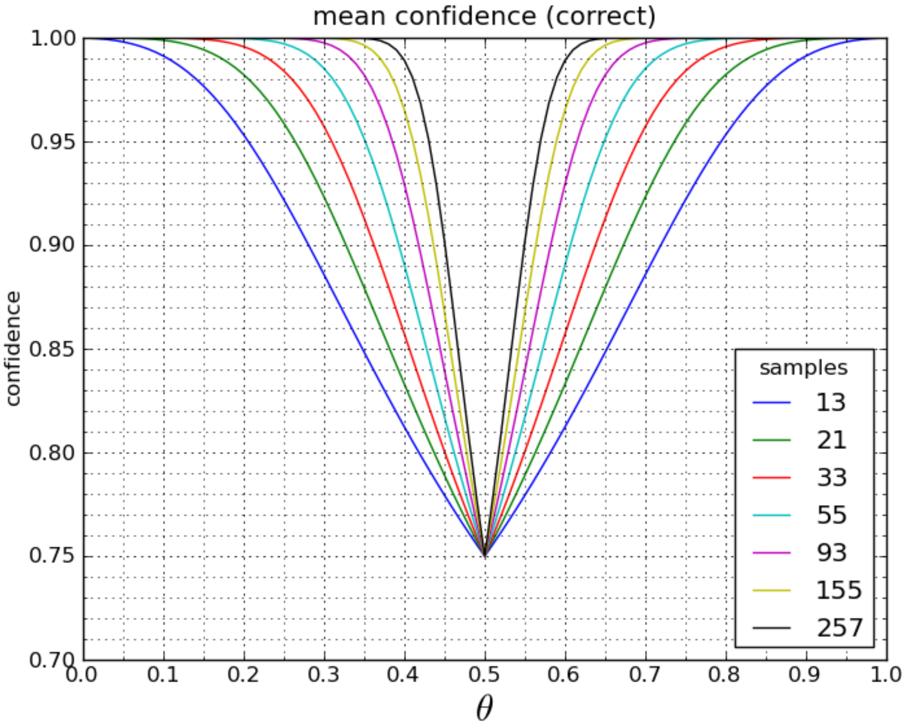

Figure 14: The predictive mean of the confidence when the simple binary test gives the correct answer, assuming a flat prior for θ.

1. At about 55 samples, we have an approximately 80% predictive probability of achieving at least 90% confidence, provided the true value of θ is no closer than 0.15 to 1/2.

2. At about 250 samples, we have an approximately 80% predictive probability of achieving 90% confidence, provided the true value of θ is no closer than 0.06 to 1/2.

3. We have an overall predictive probability of 80% of achieving 90% confidence (and being correct in our binary decision) with about 55 samples.

5.3 Predicting the Width of the Posterior When θ is Near 1/2

In Equation 88 in Appendix A.7, we give a formula for the predictive mean of the posterior standard deviation, conditioned on the true value of θ. A plot of the predictive mean as a function of θ for various sample sizes is shown in Figure 15. Note that our prediction for the posterior standard deviation goes down as the sample size increases (as expected) and that the prediction is relatively insensitive to the true value of θ, except in the vicinity of 0 and 1.

The same plot, but for the fairly extreme case of a prior with $\alpha_0 = 2$ and $\beta_0 = 10$, is shown in Figure 16. Note that the prior has not made a dramatic difference in the prediction for the posterior standard deviation, especially for the larger sample sizes, as long as we do not go very close to 0 or 1. This is desirable behavior because we are principally interested in the predictive mean of the posterior standard deviation when the data tell us that θ is close to 1/2. So the observed behavior guarantees that the predictive mean of the posterior standard deviation at a given sample size will be relatively insensitive to the data we have already analyzed in the cases in which we care about it.

This point is reinforced by Figures 17 and 18, which show the predictive mean of the posterior standard deviation averaged over the true value of θ, assuming a flat prior for θ in the former figure and a $\alpha_0 = 2$ and $\beta_0 = 10$ prior in the latter. Again, the differences between the two cases with very different priors are really quite moderate, maybe a factor of 1.5 at 30 samples, indicating that the predictive mean of the posterior standard deviation is a relatively stable with respect to prior knowledge.

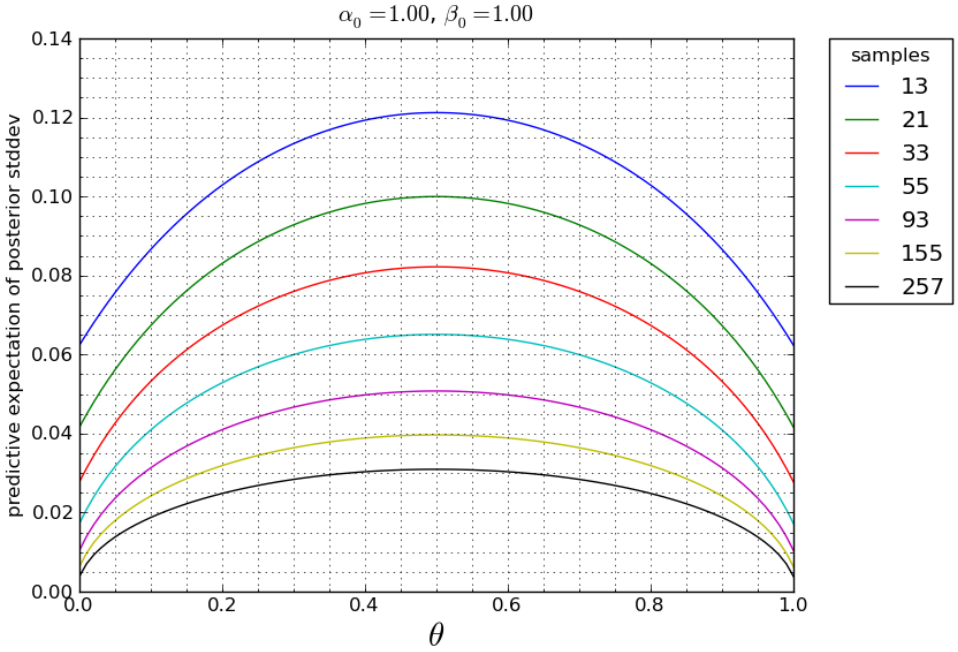

Figure 15: The predictive mean of the posterior standard deviation as a function of the true value of θ, assuming a flat prior for θ.

This point may be made even stronger if one takes into account the fact that a prior with $\alpha_0 = 2$ and $\beta_0 = 10$ is equivalent to having started with a flat prior and observed ten samples, with one sample being zero and the other nine samples being ones.[12] If we look at the value of the curve in Figure 18 at a given number of samples and match it with the value of the curve in Figure 17 at that number of samples plus ten, the values are very close. In other words, the posterior standard deviation depends fairly strongly on the sample size, but only quite weakly on the actual values of the data.

6 Sequential Power Analysis

The procedure that we have outlined for doing Bayesian inference on the binomial model, the tests that we described based on the posterior for θ, and

[12] See the comment on page 9 immediately below Equation 5.

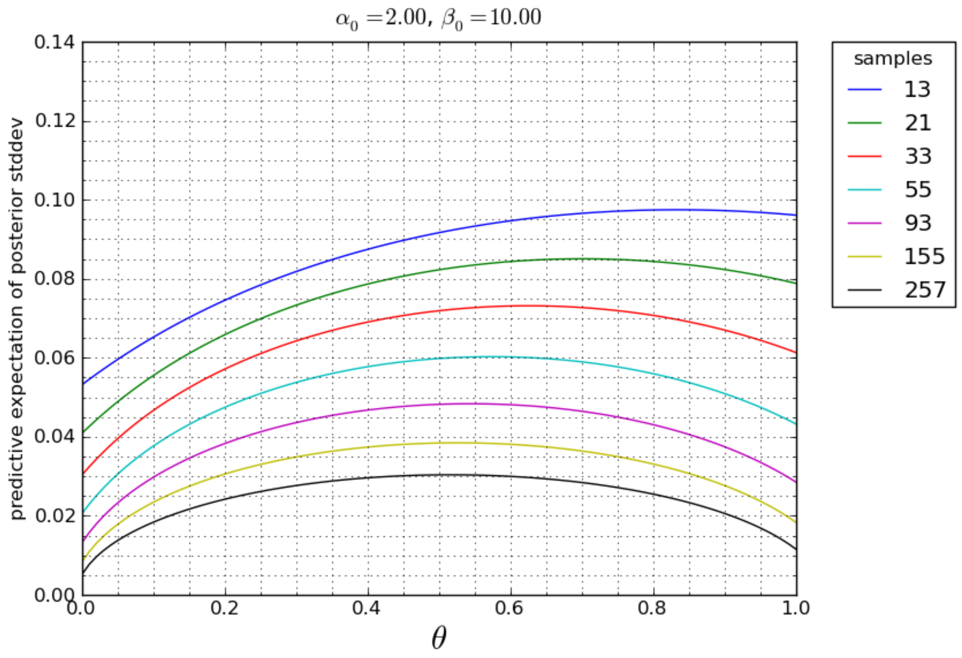

Figure 16: The predictive mean of the posterior standard deviation as a function of the true value of θ, assuming a prior for θ with $\alpha_0 = 2$ and $\beta_0 = 10$.

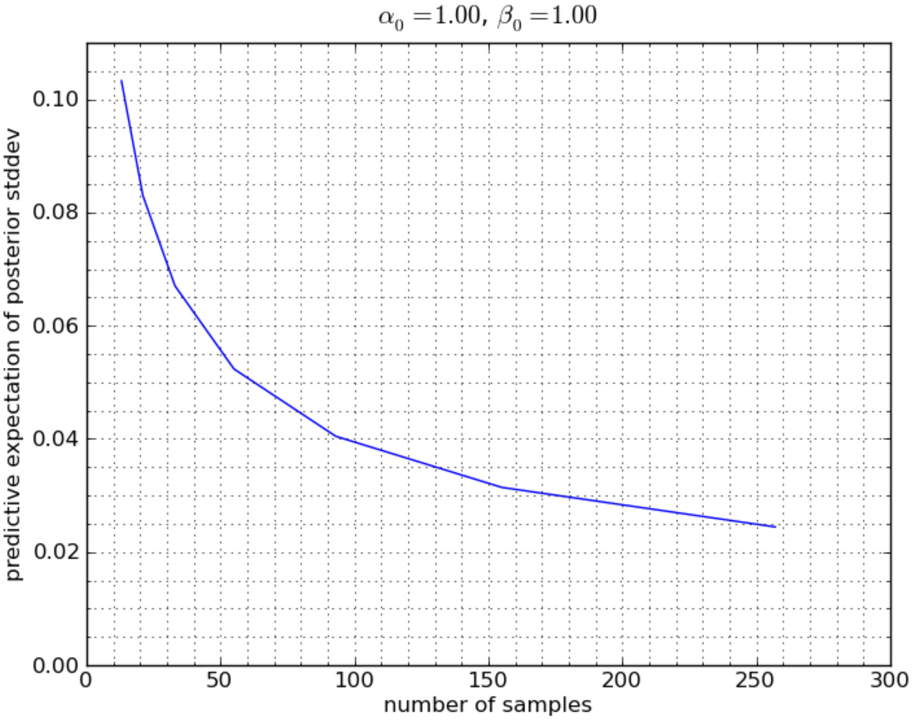

Figure 17: The predictive mean of the posterior standard deviation averaged over the true value of θ, assuming a flat prior for θ.

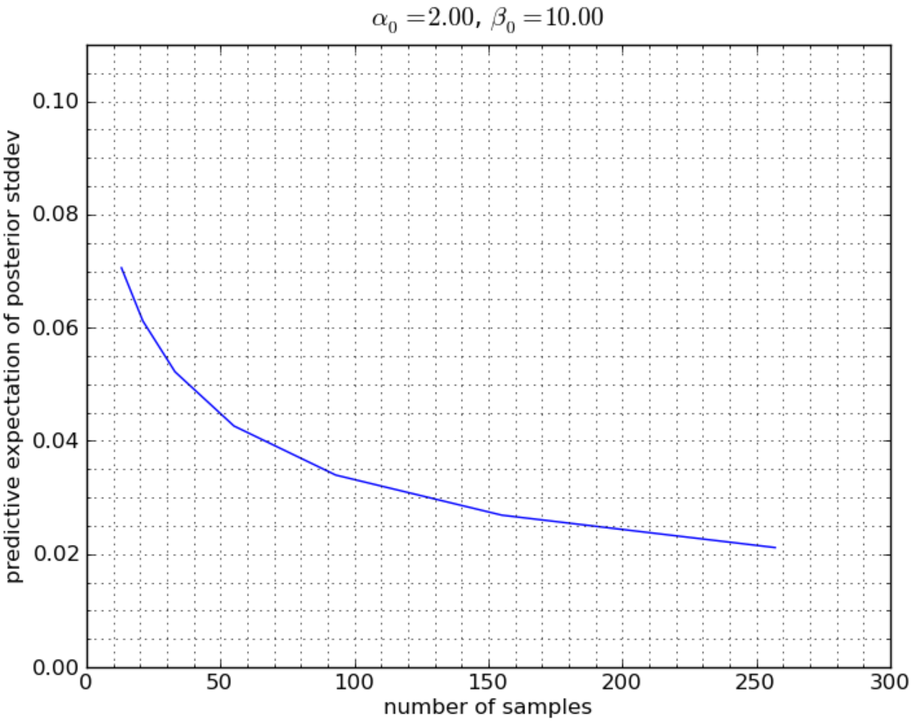

Figure 18: The predictive mean of the posterior standard deviation averaged over the true value of θ, assuming a prior for θ with $\alpha_0 = 2$ and $\beta_0 = 10$.

the power analysis that we did on the tests can all be performed sequentially. In other words, we can analyze some number of cases, x_i, produce a posterior for θ, and then use that posterior as a prior for further analysis, including power analysis.[13] This allows us to incrementally update our predictions about how much new data will be necessary to determine whether or not an improvement in predictive probability has been observed.

If we are using a prior for θ that comes from an earlier formal analysis, then the relative meaningfulness of the predictive probabilities conditioned on the true value of θ and averaged over θ will change from the examples that we have presented so far, which have used a flat prior for θ. On the one hand, predictive probabilities conditioned on the true θ will become less meaningful because our prior will eliminate more and more of the range of θ the more data used to derive it and, even within the range for θ that is left, different values will have very different prior weights. Conversely, predictive probabilities averaged over θ will become more meaningful because the prior that plays a critical role in their calculation becomes more well-founded. In other words, when we are starting from scratch and use a simple flat prior largely for convenience, the conditional predictive probabilities should play the dominant role. When we have data on which to base our prior, the averaged predictive probabilities should probably come to the fore.

7 Summary

We concluded Sections 5.1 and 5.2 with brief summaries of the numerical results of our analyses of the predictive probabilities associated with our two tests, in the case of a flat prior for θ. (See pages 28 and 42.) We will not repeat those results here but, rather, will present a high-level summary of the results that we have presented in the body of paper above.

In brief, in this paper we have:

1. Proposed a simple binomial model for comparison of paired predictions from legacy and new simulations methodologies, with a single unknown parameter, θ, which quantifies any improvement in predictive capability. (Section 2)

[13] Our use of a conjugate prior for θ makes this sequential analysis particularly simple because the beta-distribution form for all priors and posteriors is preserved with each iteration of the analysis.

2. Showed how to do Bayesian inference to obtain a posterior distribution for θ. (Section 3)

3. Showed, in Section 4, how to use the posterior distribution for θ to do a kind of Bayesian hypothesis testing that answers two questions:

   (a) Have we observed an improvement in predictive capability?

   (b) Have we achieved a given probability of improvement, a given probability of no improvement, or neither?

4. Stressed that, no matter how many data are taken, there will always be a region around $\theta = 1/2$ where question 3b above is virtually certain to receive an indefinite answer and discussed the ubiquity of this problem in hypothesis testing. (In Section 4.3 and throughout Section 5)

5. Showed how to use the posterior mean and standard deviation of θ as a kind of plan B to provide additional useful information to decision makers in the event that question 3b above receives an indefinite answer, and discussed the relevance of such an approach whenever one does hypothesis testing. (Section 4.3)

6. Performed a power analysis on the tests we have proposed and showed how the resulting predictive probabilities may be used to project, before simulations are run, how many simulations are likely to be necessary to answer the questions above. (Section 5)

7. Discussed the use of our results for sequential power analysis, where one performs a small number of simulations and uses the results to make a (more) informed projection of how many additional simulations will need to be run in order to answer the questions in item 3 above. (Section 6)

A Appendix: Some Standard Results and Detailed Calculations

This appendix provides some standard definitions and results from statistics and documents the details of several calculations in the main text. The calculations are extremely detailed with almost every step made explicit. This detail has been presented in order to make it easy to check the calculations.

A.1 Series of Bernoulli Trials and the Binomial Likelihood

In statistics, a *Bernoulli trial* is an observation that can produce one of only two possible outcomes (conventionally called 0 and 1, or “failure” and “success” respectively) with a well-defined probability, θ (where of course $0 \leq \theta \leq 1$), to take the value 1 and, of course, probability $1 - \theta$ to take the value 0. In simpler terms, a Bernoulli trial is equivalent to the flip of a weighted coin.

Assume that we have a sequence of n Bernoulli trials, $\boldsymbol{x} \equiv \{x_i\} : 1 \leq i \leq n$, that can be presumed to be identically distributed with parameter θ and to be independent, given θ. Then, the probability of obtaining a given sequence $\boldsymbol{x}$ is just

$$P(\boldsymbol{x}) = \theta^s (1 - \theta)^{n-s}, \tag{47}$$

where $s \equiv s\,(\boldsymbol{x}) \equiv \sum_i x_i$ is the number of times 1 appears in the sequence. Note that s can take the values $0, 1, \ldots, n$. Now, elementary combinatorics tells us that there are $\binom{n}{s}$ distinct binary sequences of length n that contain 1 exactly s times, where $\binom{n}{s} \equiv \frac{n!}{s!(n-s)!}$ is the binomial coefficient. Thus, we may write the likelihood, given θ, of observing s in n trials as

$$P(s|\theta) \equiv \binom{n}{s} \theta^s (1 - \theta)^{n-s} = \frac{n!}{s!(n-s)!}\, \theta^s (1 - \theta)^{n-s}. \tag{48}$$

This distribution for s is called the *binomial distribution*[14] and, when it represents the likelihood, the *binomial model.*[15]

A.2 Conjugate Priors in Bayesian Parameter Estimation

Appendix A.1 immediately above presents our unknown parameter, θ, and the likelihood that connects θ to our data, $\boldsymbol{x}$. The remaining essential ingredient for a Bayesian analysis is the prior for θ, $p(\theta)$. We choose to use a conjugate prior for θ, a concept that we explain in this appendix.

[14] See Gelman et al [3, Sections 2.1–2.5] or Bernardo and Smith [4, page 115 and Section 4.3.1] or almost any elementary statistics text for basic information on sequences of Bernoulli trials and the binomial distribution.

[15] Because the likelihood depends on the data only through the scalar function of the data, s, s is referred to as a *sufficient statistic* for the unknown parameter θ.

Conjugate priors are basic to the practice of Bayesian statistics, particularly when working with the kind of simple model that we are using in this paper. Simply put, given a likelihood and a particular parameter in the likelihood on which we are doing inference, a *conjugate prior* for the parameter is a prior that produces a posterior of the same functional form as the prior. A bit more specifically, a conjugate prior has a particular functional form with parameters of its own (usually called *hyperparameters*) that determine the specific prior. When this form for the prior is used with the corresponding likelihood, the result is a posterior with the same functional form but modified values for the hyperparameters.

Almost any basic text on Bayesian statistics will provide an extensive discussion of conjugate priors. See for example, Gelman et al [3] or Bernardo and Smith [4, Section 5.2]. Here, we will just outline what is probably the most common example.

Assume we have a Gaussian likelihood for some data, $\boldsymbol{d} \equiv \{d_i\}$,

$$p\left(\boldsymbol{d}|\theta\right) = \frac{1}{\left(2\pi\sigma^2\right)^{n/2}} \exp\left(\frac{1}{2\sigma^2}\sum_i \left(d_i - \theta\right)^2\right), \tag{49}$$

with an unknown mean, θ, and a known variance, σ^2. If we further assume a Gaussian prior for θ,

$$p(\theta) = \frac{1}{\sqrt{2\pi\tau_0^2}} \exp\left(\frac{1}{2\tau_0^2}\left(\theta - \mu_0\right)^2\right), \tag{50}$$

with mean μ_0 and variance τ_0^2, then the posterior for θ will also be Gaussian,

$$p(\theta|\boldsymbol{d}) = \frac{1}{\sqrt{2\pi\tau_1^2}} \exp\left(\frac{1}{2\tau_1^2}\left(\theta - \mu_1\right)^2\right), \tag{51}$$

with mean, μ_1, and variance, τ_1, that are functions of the data, the variance of the likelihood, and the mean and variance of the prior. (See either of the references above for details.)

Conjugate priors, where they exist, greatly simplify Bayesian parameter estimation. In particular, they make any kind of sequential analysis, where one repeatedly updates one's distribution for an unknown parameter using new data, much easier and more transparent.

Although the class of likelihoods and resulting inference problems that have conjugate priors is limited, it includes most of the common simple likelihoods and, fortunately, the binomial inference problem with which we are dealing here. We will now present our method for doing inference on the binomial model, using the appropriate conjugate prior.

A.3 The Posterior for the Binomial Model

Bernardo and Smith (B&S) [4] define the inferential process for the binomial model in the first box in their Appendix A.2 on page 436. They give a parametric form for the conjugate prior

$$p(x) = \mathrm{Be}(x|\alpha_0, \beta_0), \tag{52}$$

where Be is the beta distribution, defined by them in their Section A.1 in the first box on page 430 as

$$\mathrm{Be}(x|\alpha, \beta) \equiv \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, x^{\alpha-1}(1-x)^{\beta-1}. \tag{53}$$

Note that we have used the notation α_0 and β_0 to represent the values of α and β in our prior and have used bare α and β for the generic beta distribution. Γ is the well-known gamma function, given by

$$\Gamma(x) \equiv \int_0^\infty dt\, x^{n-1}\, e^{-x}. \tag{54}$$

For x an integer, n, we have,

$$\Gamma(n) = (n-1)!. \tag{55}$$

On their page 436, they give the posterior corresponding to the prior in Equation 52 as

$$p(\theta|z) = \mathrm{Be}(\theta|\alpha_0 + r, \beta_0 + n - r), \tag{56}$$

where their z equals our $\boldsymbol{x}$ and their r equals our $s \equiv s\,(\boldsymbol{x}) \equiv \sum_i x_i$. Writing in terms of our variables, we have

$$p(\theta|\boldsymbol{x}) = \mathrm{Be}(\theta|\alpha_0 + s, \beta_0 + n - s). \tag{57}$$

Plugging into the definition of the beta distribution, Equation 53, with the dummy $\alpha = s + \alpha_0$ and the dummy $\beta = n - s + \beta_0$, we have

$$\begin{aligned} p(\theta|\boldsymbol{x}) &= \frac{\Gamma(n+\alpha_0+\beta_0)}{\Gamma(s+\alpha_0)\Gamma(n-s+\beta_0)}\,\theta^{s+\alpha_0-1}(1-\theta)^{n-s+\beta_0-1} \\ &= \frac{(n+\alpha_0+\beta_0-1)!}{(s+\alpha_0-1)!(n-s+\beta_0-1)!}\,\theta^{s+\alpha_0-1}(1-\theta)^{n-s+\beta_0-1}, \end{aligned} \tag{58}$$

where the latter formula pertains when α_0 and β_0 are integers.

For $\mathrm{Be}(\theta|\alpha,\beta)$, B&S give the mean

$$\mathrm{E}(\theta|\alpha,\beta) = \frac{\alpha}{\alpha+\beta}. \tag{59}$$

For our posterior, Equation 58, this gives the posterior mean

$$\mathrm{E}(\theta|\boldsymbol{x}) = \frac{s+\alpha_0}{n+\alpha_0+\beta_0}. \tag{60}$$

B&S also give the variance for the beta distribution, $\mathrm{Be}(\theta|\alpha,\beta)$, as

$$\mathrm{Var}(\theta|\alpha,\beta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \tag{61}$$

For our posterior, this gives the posterior variance

$$\mathrm{Var}(\theta|\boldsymbol{x}) = \frac{(s+\alpha_0)(n-s+\beta_0)}{(n+\alpha_0+\beta_0)^2(n+\alpha_0+\beta_0+1)}. \tag{62}$$

The mode of the beta distribution is given by Gelman, et al [3, page 576] as,

$$\theta_{\max} \equiv \operatorname{argmax} \mathrm{Be}(\theta|\alpha,\beta) = \frac{\alpha-1}{\alpha+\beta-2}. \tag{63}$$

In the case of our posterior, this becomes,

$$\theta_{\max} = \frac{s+\alpha_0-1}{n+\beta_0+\alpha_0-2}. \tag{64}$$

A.4 Some Properties of the Beta Distribution Posterior

In this section, we show that testing the mean of a beta distribution to see if it is greater than 1/2 is equivalent to doing the same test on the mode and on the median.

We begin by showing that the mean and mode are always on the same side of $\theta = 1/2$. For a beta distribution with parameters α and β, $\mathrm{Be}(\theta|\alpha,\beta)$, the mean is given by Equation 59,

$$\mathrm{E}(\theta|\alpha,\beta) = \frac{\alpha}{\alpha+\beta}. \tag{65}$$

Simple algebra shows that,

$$\mathrm{E}(\theta|\alpha,\beta) > 1/2 \Longleftrightarrow \alpha > \beta. \tag{66}$$

The mode for the beta distribution is given in Equation 63,

$$\theta_{\max} = \frac{\alpha-1}{\alpha+\beta-2}. \tag{67}$$

Again simple algebra shows that,

$$\theta_{\max} > 1/2 \Longleftrightarrow \alpha > \beta, \tag{68}$$

So the condition for the mean to be greater than 1/2 is the same as the condition for the mode to be greater than 1/2.

Next, we will show that the mean being greater than 1/2 is also equivalent to the median being greater than 1/2. Although there is no analytical formula for the median of the beta distribution, we can show that the median is greater than to 1/2 if and only if $\alpha > \beta$ by showing that the condition on α and β is equivalent to the probability that θ is less than or equal to 1/2 being less than the probability that θ is greater than 1/2. Thus, we are interested in the relative magnitudes of the two quantities $P\big[\theta \leq 1/2|\alpha,\beta\big]$ and $P\big[\theta > 1/2|\alpha,\beta\big]$. These quantities are:

$$\begin{aligned} P\big[\theta \leq 1/2|\alpha,\beta\big] &\equiv \int_0^{1/2} d\theta\, \mathrm{Be}(\theta|\alpha,\beta) \\ &= \int_0^{1/2} d\theta\, \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, \theta^{\alpha-1}(1-\theta)^{\beta-1} \end{aligned}$$

and

$$\begin{aligned} P\big[\theta > 1/2|\alpha,\beta\big] &\equiv \int_{1/2}^{1} d\theta\, \mathrm{Be}(\theta|\alpha,\beta) \\ &= \int_{1/2}^{1} d\theta\, \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, \theta^{\alpha-1}(1-\theta)^{\beta-1}. \end{aligned}$$

Since we are only interested in the relative magnitudes of these quantities, we may ignore the normalization factor, which is identical in the two cases, and work with the two quantities,

$$J_0(\alpha,\beta) \equiv \int_0^{1/2} d\theta\, \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

and

$$J_1(\alpha, \beta) \equiv \int_{1/2}^{1} d\theta \, \theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

We wish to show that $J_0(\alpha, \beta)$ is less than $J_1(\alpha, \beta)$ if and only if $\alpha > \beta$.

We may show that the integral J_0 is less than the integral J_1 if we can rewrite both as integrals over the same domain with the integrand for J_1 being, at every point except possibly on a set of measure zero, greater than the integrand for J_0. Thus, we make the substitutions

$$\phi \equiv \frac{1}{2} - \theta, \quad \text{in the integral for } J_0,$$

and

$$\phi \equiv \theta - \frac{1}{2}, \quad \text{in the integral for } J_1.$$

Then, we have

$$\begin{aligned} J_0(\alpha, \beta) &= -\int_{1/2}^{0} d\phi \left(\frac{1}{2} - \phi\right)^{\alpha-1} \left(\frac{1}{2} + \phi\right)^{\beta-1} \\ &= \int_{0}^{1/2} d\phi \left(\frac{1}{2} - \phi\right)^{\alpha-1} \left(\frac{1}{2} + \phi\right)^{\beta-1} \end{aligned}$$

and

$$J_1(\alpha, \beta) = \int_{0}^{1/2} d\phi \left(\frac{1}{2} + \phi\right)^{\alpha-1} \left(\frac{1}{2} - \phi\right)^{\beta-1}.$$

These are integrals over the same domain so now we want to find the conditions under which the integrand for J_1 is always greater than the integrand for J_0. In other words, we want to find the conditions under which

$$\left(\frac{1}{2} + \phi\right)^{\alpha-1} \left(\frac{1}{2} - \phi\right)^{\beta-1} > \left(\frac{1}{2} - \phi\right)^{\alpha-1} \left(\frac{1}{2} + \phi\right)^{\beta-1}$$

Since the factor $\left(\frac{1}{2} + \phi\right)$ in this inequality is always greater than 0, the inequality is equivalent to

$$\left(\frac{1}{2} + \phi\right)^{\alpha-\beta} \left(\frac{1}{2} - \phi\right)^{\beta-1} > \left(\frac{1}{2} - \phi\right)^{\alpha-1}.$$

The factor $\left(\frac{1}{2} - \phi\right)$ is less than or equal to 0 only at the point $\phi = 1/2$, which we may safely ignore, since behavior of the integrand on a set of measure zero has no effect on the integral. Thus, our inequality is equivalent to

$$\left(\frac{1}{2} + \phi\right)^{\alpha-\beta} > \left(\frac{1}{2} - \phi\right)^{\alpha-\beta}.$$

This is equivalent to

$$\frac{1}{2} + \phi > \frac{1}{2} - \phi, \tag{69}$$

provided the exponent $\alpha - \beta$ is positive or, equivalently, provided

$$\alpha > \beta.$$

The relation in Equation 69 is equivalent to

$$\phi > 0.$$

Since this condition is satisfied over the domain of integration, except at the single point $\phi = 0$, we see that the condition that $\alpha > \beta$ imposed above is equivalent to $J_0(\alpha, \beta)$ being less than $J_1(\alpha, \beta)$ and thus, to the median of θ being greater than 1/2.

A.5 A Useful Predictive Probability

There are several places in this work where we need to compute predictive probabilities of the form,

$$P\big[s \in U, \theta_0 < \theta \leq \theta_1\big] \tag{70}$$

for U some subset of the possible values of the data, s, and θ_0 and θ_1 some limits on θ such that $0 \leq \theta_0 \leq \theta_1 \leq 1$.[16] We may use the law of total probability and the definition of conditional probability to write,

$$\begin{aligned} P\big[s \in U, \theta_0 < \theta \leq \theta_1\big] &= \int_{\theta_0}^{\theta_1} d\theta\, P\big[s \in U, \theta\big] \\ &= \int_{\theta_0}^{\theta_1} d\theta\, P\big[s \in U | \theta\big]\, p(\theta), \end{aligned} \tag{71}$$

[16] In practice, we always have either $(\theta_0, \theta_1) = (0, 1/2)$ or $(\theta_0, \theta_1) = (1/2, 1)$, but the general problem is no more difficult to solve than these special cases.

where $p(\theta)$ is our prior for θ,

$$p(\theta) \equiv \mathrm{Be}(\theta|\alpha_0, \beta_0) \equiv \frac{\Gamma(\alpha_0+\beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)}\,\theta^{\alpha_0-1}(1-\theta)^{\beta_0-1}. \tag{72}$$

The quantity $P\big[s \in U|\theta\big]$ is just the predictive probability that s will be in U, given the true value of θ, which is given by the likelihood (Equation 48), summed over the values of s in U,

$$P\big[s \in U|\theta\big] = \sum_{s\in U} \binom{n}{s} \theta^s (1-\theta)^{n-s}. \tag{73}$$

Then, we have,

$$\begin{aligned} &P\big[s \in U, \theta_0 < \theta \le \theta_1\big] \\ &\quad = \int_{\theta_0}^{\theta_1} d\theta \sum_{s\in U} \binom{n}{s} \theta^s (1-\theta)^{n-s}\, \frac{\Gamma(\alpha_0+\beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)}\,\theta^{\alpha_0-1}(1-\theta)^{\beta_0-1} \\ &\quad = \frac{\Gamma(\alpha_0+\beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} \sum_{s\in U} \binom{n}{s} \int_{\theta_0}^{\theta_1} d\theta\, \theta^{s+\alpha_0-1}(1-\theta)^{n-s+\beta_0-1}. \end{aligned} \tag{74}$$

Writing the integrand in terms of the beta distribution through its definition in Equation 53, we have,

$$\begin{aligned} &P\big[s \in U, \theta_0 < \theta \le \theta_1\big] \\ &\quad = \frac{\Gamma(\alpha_0+\beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} \\ &\qquad \cdot \sum_{s\in U} \binom{n}{s} \int_{\theta_0}^{\theta_1} d\theta\, \frac{\Gamma(s+\alpha_0)\Gamma(n-s+\beta_0)}{\Gamma(n+\alpha_0+\beta_0)}\,\mathrm{Be}(\theta|s+\alpha_0, n-s+\beta_0). \end{aligned} \tag{75}$$

Moving the fraction out of the integral and writing the integral in terms of the regularized incomplete beta function, $I_x(\alpha, \beta)$, we have,

$$\begin{aligned} P\big[s \in U, \theta_0 < \theta \leq \theta_1\big] \\ &= \frac{\Gamma(\alpha_0 + \beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} \\ &\quad \cdot \sum_{s \in U} \binom{n}{s} \frac{\Gamma(s + \alpha_0)\Gamma(n - s + \beta_0)}{\Gamma(n + \beta_0 + \alpha_0)} \; \Big(I_{\theta_1}(s + \alpha_0, n - s + \beta_0) \\ &\qquad\qquad - I_{\theta_0}(s + \alpha_0, n - s + \beta_0)\Big) \\ &= \frac{\Gamma(\alpha_0 + \beta_0)\, n!}{\Gamma(\alpha_0)\Gamma(\beta_0)\Gamma(n + \beta_0 + \alpha_0)} \\ &\quad \cdot \sum_{s \in U} \frac{\Gamma(s + \alpha_0)\Gamma(n - s + \beta_0)}{s!\,(n - s)!} \; \Big(I_{\theta_1}(s + \alpha_0, n - s + \beta_0) \\ &\qquad\qquad - I_{\theta_0}(s + \alpha_0, n - s + \beta_0)\Big). \end{aligned} \tag{76}$$

When we have a flat prior ($\alpha_0 = \beta_0 = 1$), this expression simplifies to

$$\begin{aligned} P\big[s \in U, \theta_0 < \theta \leq \theta_1\big] & \qquad\qquad \text{for } \alpha_0 = \beta_0 = 1, \\ = \frac{1}{n + 1} \sum_{s \in U} \big(I_{\theta_1}(s + 1, n - s + 1) - I_{\theta_0}(s + 1, n - s + 1)\big). \end{aligned} \tag{77}$$

Unfortunately, there is no equivalent simplification for the general case.

A.6 The Predictive Mean and Variance of the Confidence

We seek to compute the predictive mean value of the confidence when the true value of θ is less than or equal to $1/2$ *and* when our test has correctly confirmed this, in other words, when s is less than or equal to $\frac{1}{2}\,(n + \beta_0 - \alpha_0)$. In order to do this, we need to compute the confidence for each possible outcome (value of s) and then add these quantities for all the outcomes on the "right" side, weighted by the probability of each outcome. Then, we need to normalize by dividing by the total probability to be on the right side. Thus, we seek to calculate:

$$\begin{aligned}\mathrm{E}\big(C(s) \mid s \in S_0,\, \theta\big) & \qquad\qquad\qquad \text{for } \theta \le 1/2,\ s \text{ correct} \\ &= \frac{1}{\sum_{s\in S_0} P\big(s|\theta\big)} \sum_{s\in S_0} C(s)\, P\big(s|\theta\big) \\ &= \frac{1}{\sum_{s\in S_0} P\big(s|\theta\big)} \sum_{s\in S_0} P\big(s|\theta\big)\, I_{1/2}(s+\alpha_0, n-s+\beta_0),\end{aligned} \tag{78}$$

where we have substituted the explicit value for $C(s)$ from Equation 20 and S_0 is defined by Equation 16.

Similarly, we seek to compute the mean value of the confidence when the true value of θ is less than or equal to $1/2$ *and* when our test has *incorrectly failed* to confirm this, in other words, when s is greater than $\frac{1}{2}\,(n+\beta_0-\alpha_0)$. In order to do this, we need to compute the confidence for each possible outcome (value of s) and then add these quantities for all the outcomes on the “wrong” side, weighted by the probability of each outcome. Then, we need to normalize by dividing by the total probability to be on the wrong side. Thus, we seek to calculate:

$$\begin{aligned}\mathrm{E}\big(C(s) \mid s \in S_1,\, \theta\big) & \qquad\qquad\qquad \text{for } \theta \le 1/2,\ s \text{ incorrect} \\ &= \frac{1}{\sum_{s\in S_1} P\big(s|\theta\big)} \sum_{s\in S_1} C(s)\, P\big(s|\theta\big) \\ &= \frac{1}{\sum_{s\in S_1} P\big(s|\theta\big)} \sum_{s\in S_1} P\big(s|\theta\big)\, \big(1 - I_{1/2}(s+\alpha_0, n-s+\beta_0)\big),\end{aligned} \tag{79}$$

where we have substituted the explicit value for $C(s)$ from Equation 21 and S_1 is defined by Equation 16.

Of course, we want to compute the same quantities when the true value of θ is greater than $1/2$. The mean value of the confidence when the true value of θ is greater than $1/2$ *and* when our test has correctly confirmed this, in other words, when s is greater than $\frac{1}{2}\,(n+\beta_0-\alpha_0)$, is:

$$\mathrm{E}\big(C(s) \mid s \in S_1,\, \theta\big) \qquad \text{for } \theta > 1/2,\ \ s \text{ correct}, \tag{80}$$

but notice that this is functionally the same as the quantity in Equation 79 above, although we will be evaluating it over a different region of θ.

Finally, we seek to compute the mean value of the confidence when the true value of θ is greater than $1/2$ *and* when our test has *incorrectly failed* to confirm this, in other words, when s is less than or equal to $\frac{1}{2}\left(n+\beta_0-\alpha_0\right)$. This quantity is:

$$\mathrm{E}\big(C(s) \mid s \in S_0,\, \theta\big) \qquad \text{for } \theta > 1/2,\ s \text{ incorrect}, \tag{81}$$

but notice, again, that this is functionally the same as the quantity in Equation 78, although we will be evaluating it over a different region of θ.

We also want to compute the corresponding variances. They are:

$$\begin{aligned}&\mathrm{Var}\big(C(s) \mid s \in S_0,\, \theta\big) \qquad\qquad\qquad \text{for } \theta \leq 1/2,\ s \text{ correct}\\ &\qquad = \frac{1}{\sum_{s\in S_0} P\big(s|\theta\big)} \sum_{s\in S_0} P\big(s|\theta\big)\left(C(s) - \mathrm{E}\big(C(s)|s\in S_0,\theta\big)\right)^2,\end{aligned} \tag{82}$$

and

$$\begin{aligned}&\mathrm{Var}\big(C(s) \mid s \in S_1,\, \theta\big) \qquad\qquad\qquad \text{for } \theta \leq 1/2,\ s \text{ incorrect}\\ &\qquad = \frac{1}{\sum_{s\in S_1} P\big(s|\theta\big)} \sum_{s\in S_1} P\big(s|\theta\big)\left(C(s) - \mathrm{E}\big(C(s)|s\in S_1,\theta\big)\right)^2.\end{aligned} \tag{83}$$

As was with the case with the mean values, the quantity

$$\mathrm{Var}\big(C(s) \mid s \in S_1,\, \theta\big) \qquad \text{for } \theta > 1/2,\ s \text{ correct}, \tag{84}$$

is functionally the same as the quantity in Equation 83. Similarly, the quantity

$$\mathrm{Var}\big(C(s) \mid s \in S_0,\, \theta\big) \qquad \text{for } \theta > 1/2,\ s \text{ incorrect}, \tag{85}$$

is functionally the same as the quantity in Equation 82. In all these computations, we get the values for the expectations from Equations 78 through 81 above, and we get the value for $C(s)$ from Equations 20 or 21 depending on whether s is less than or equal to $\frac{1}{2}\left(n+\beta_0-\alpha_0\right)$ or s is greater than $\frac{1}{2}\left(n+\beta_0-\alpha_0\right)$, respectively.

A.7 The Predictive Mean of the Posterior Standard Deviation

We seek to compute the predictive mean of the posterior standard deviation. We begin, as should now be familiar, by computing this predictive quantity conditioned on the true value of θ. Given the data, s, the posterior standard standard deviation, $\Sigma_\theta(s)$, is obtained by taking the square root of the posterior variance given by Equation 62,

$$\Sigma_\theta(s) = \sqrt{\frac{(s+\alpha_0)(n-s+\beta_0)}{(n+\alpha_0+\beta_0)^2(n+\alpha_0+\beta_0+1)}}. \tag{86}$$

Then, the predictive mean of this quantity is,

$$\mathrm{E}\left(\Sigma_\theta(s)|\theta\right) = \sum_s \Sigma_\theta(s)\, P\big(s|\theta\big), \tag{87}$$

which gives, using Equation 1 for $P\big(s|\theta\big)$,

$$\begin{aligned}
&\mathrm{E}\left(\Sigma_\theta(s)|\theta\right) \\
&\qquad = \sum_s \Sigma_\theta(s) \binom{n}{s} \theta^s (1-\theta)^{n-s} \\
&\qquad = \sum_s \sqrt{\frac{(s+\alpha_0)(n-s+\beta_0)}{(n+\alpha_0+\beta_0)^2(n+\alpha_0+\beta_0+1)}} \binom{n}{s} \theta^s (1-\theta)^{n-s},
\end{aligned} \tag{88}$$

which is straightforward to compute numerically.

We may next compute the overall version of the predictive mean of the posterior standard deviation, averaged over θ, $\mathrm{E}\left(\Sigma_\theta(s)\right)$. We need only multiply the quantity above by the prior for θ, given by Equation 3, and integrate over all θ, thusly,

$$\begin{aligned}
&\mathrm{E}\big(\Sigma_\theta(s)\big) \\
&\quad = \int_0^1 d\theta \sum_s \Sigma_\theta(s) \binom{n}{s} \theta^s (1-\theta)^{n-s} \mathrm{Be}(\theta|\alpha_0,\beta_0) \\
&\quad = \int_0^1 d\theta \sum_s \Sigma_\theta(s) \binom{n}{s} \theta^s (1-\theta)^{n-s} \frac{\Gamma(\alpha_0+\beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)}\, \theta^{\alpha_0-1}(1-\theta)^{\beta_0-1} \\
&\quad = \frac{\Gamma(\alpha_0+\beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} \sum_s \Sigma_\theta(s) \binom{n}{s} \int_0^1 d\theta\, \theta^{s+\alpha_0-1}(1-\theta)^{n-s+\beta_0-1}.
\end{aligned} \tag{89}$$

The integral is a beta function that evaluates to,

$$\mathrm{B}\left(s+\alpha_0, n-s+\beta_0\right) = \frac{\Gamma\left(s+\alpha_0\right)\Gamma\left(n-s+\beta_0\right)}{\Gamma\left(n+\beta_0+\alpha_0\right)}. \tag{90}$$

Thus, we have,

$$\begin{aligned}
&\mathrm{E}\big(\Sigma_\theta(s)\big)\\
&\quad= \frac{\Gamma(\alpha_0+\beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} \sum_s \Sigma_\theta(s) \binom{n}{s} \frac{\Gamma\left(s+\alpha_0\right)\Gamma\left(n-s+\beta_0\right)}{\Gamma\left(n+\beta_0+\alpha_0\right)}.\\
&\quad= \frac{\Gamma(\alpha_0+\beta_0)\, n!}{\Gamma(\alpha_0)\Gamma(\beta_0)\Gamma\left(n+\beta_0+\alpha_0\right)} \sum_s \Sigma_\theta(s) \frac{\Gamma\left(s+\alpha_0\right)\Gamma\left(n-s+\beta_0\right)}{s!(n-s)!}.\\
&\quad= \frac{\Gamma(\alpha_0+\beta_0)\, n!}{\Gamma(\alpha_0)\Gamma(\beta_0)\Gamma\left(n+\beta_0+\alpha_0\right)}\\
&\qquad\cdot \sum_s \sqrt{\frac{(s+\alpha_0)(n-s+\beta_0)}{(n+\alpha_0+\beta_0)^2(n+\alpha_0+\beta_0+1)}} \frac{\Gamma\left(s+\alpha_0\right)\Gamma\left(n-s+\beta_0\right)}{s!(n-s)!}\\
&\quad= \frac{\Gamma(\alpha_0+\beta_0)\, n!}{\Gamma(\alpha_0)\Gamma(\beta_0)\Gamma\left(n+\beta_0+\alpha_0\right)} \frac{1}{\sqrt{(n+\alpha_0+\beta_0)^2(n+\alpha_0+\beta_0+1)}}\\
&\qquad\cdot \sum_s \sqrt{(s+\alpha_0)(n-s+\beta_0)}\, \frac{\Gamma\left(s+\alpha_0\right)\Gamma\left(n-s+\beta_0\right)}{s!(n-s)!}
\end{aligned} \tag{91}$$

which is straightforward, if tedious, to compute numerically.

References

[1] James O. Berger. *Statistical Decision Theory and Bayesian Analysis.* Springer Verlag, 1985.

[2] David Higdon, personal communication, 2011.

[3] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis.* Chapman & Hall/CRC, second edition, 2003.

[4] José M. Bernardo and Adrian F. M. Smith. *Bayesian Theory.* John Wiley and Sons Ltd., 2000.