LA-UR-12-25008
Approved for public release; distribution is unlimited.

Title: Using a Binomial Model for Comparing Predictive Capabilities

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 Binomial Model for Comparing Predictive Capabilities

David E. Sigeti (XCP-8) and Robert A. Pelak (XTD-3)

LANL, LA-UR-12-XXXXX

September 12, 2012

Legacy and New Prediction Methodologies

- We assume that we have two methodologies (which we will call *legacy* and *new* methodologies) for predicting (or postdicting) the results of **the same** experiments.
  - Postdictions must be **honest**—no fitting to the result that is being postdicted
- We assume that we can determine which methodology does best at predicting a given experiment. We might do this by comparing:
  - The discrepancies between predicted and observed results for a single observable;
  - Some kind of weighted combination of discrepancies for several observables.

Postdiction

- Examples of honest postdiction:
  - Postdictions from a “common model” with either no knobs or an automatic procedure for setting them
    - The lack of knobs guarantees honest postdiction
    - We can compare postdictions from a common model to legacy predictions
    - We can compare postdictions from two different versions of a common model to track progress in the development of the model
  - A “clean room” postdiction in which the people doing the postdiction have no access to original documents and are fed what they need to know by others.
    - This is difficult to implement in a small organization.
- A subtle problem with postdiction: **What if the results we are trying to postdict have been part of a whole class of results to which we have effectively “tuned” our entire methodology?**

Comparing Legacy and New Prediction Methodologies: Two Questions

- **If the legacy methodology does better on some experiments and the new methodology on others, how do we determine which is best?**
  - Inference and hypothesis testing
- **How many comparisons between legacy and new methodologies do we need to do in order to answer the question above?**
  - Power analysis

A Statistical Model for Paired Predictions

The Binomial Model

- For each experiment, i, for which we have legacy and new predictions, we define a binary variable, x_i, by:

$$x_i \equiv \begin{cases} 1, & \text{if the new prediction for experiment } i \text{ is better} \\ & \text{than the legacy prediction} \\ 0, & \text{otherwise.} \end{cases}$$

- We model the x_i as independent draws from a single Bernoulli model with a well-defined probability, θ, that x_i takes the value 1,

$$\theta \equiv P(x_i = 1).$$

- If $\theta > 1/2$, then the new methodology is better than the legacy methodology.

The Binomial Model for Paired Predictions
Advantages and Disadvantages

- **The main advantages of the binomial model are simplicity and robustness.**
  - There is a single unknown quantity of interest, θ, which has an obvious meaning.
  - Power analysis is tractable.
  - We need to make almost no statistical modeling choices.
- **The main disadvantage is that we have discarded a lot of information.**
  - In particular, we take into account only **whether or not** a new prediction is better or worse than a legacy prediction, not **how much better or worse** it is.

Bayesian Inference on the Binomial Model

- We define $s \equiv s(\{x_i\}) \equiv \sum_i^n x_i$, the number of times 1 occurs in the set of x_i,
- Then, the likelihood of observing s in n trials is

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

- We assume a (conjugate) beta-distribution prior for θ with hyperparameters α_0 and β_0,

$$p(\theta|\alpha_0, \beta_0) = \frac{\Gamma(\alpha_0 + \beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} \theta^{\alpha_0 - 1} (1-\theta)^{\beta_0 - 1}.$$

- Then, the posterior distribution for θ is also a beta-distribution,

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

The Flat Prior and the Interpretation of α_0 and β_0

- If $\alpha_0 = \beta_0 = 1$, then the prior is flat.
- In that case, the posterior is

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

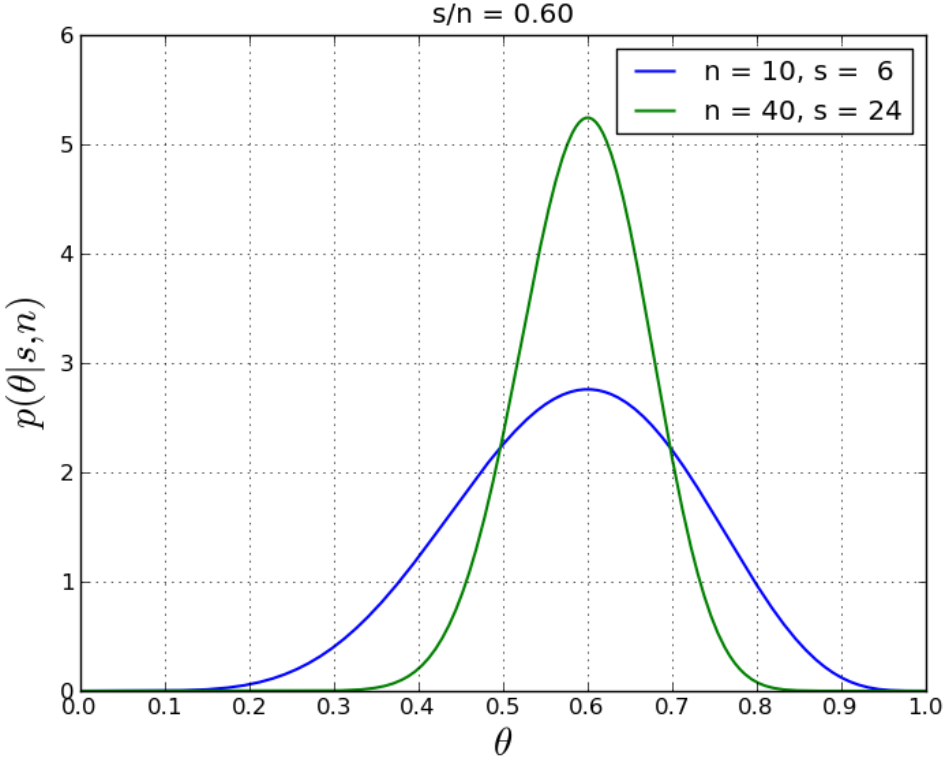
- **Note that choosing α_0 and β_0 to be integers greater than 1 is equivalent to using as our prior the posterior from an earlier analysis in which we used a flat prior and measured $\alpha_0 - 1$ ones and $\beta_0 - 1$ zeros.**
  - This is one of the pleasant consequences of using a conjugate prior.
- **If we are starting from scratch (no previously comparisons) it is hard to justify anything but a flat prior if we want our assessment of the legacy and new methodologies to be fair.**

Operated by Los Alamos National Security, LLC for the U.S. Department of Energy's NNSA

NNSA

The Posterior for θ, Two Examples

- The priors are flat.
- The two posteriors shown have the same $s/n = 0.6$ and different sample sizes, 10 and 40.
- Any reasonable estimates for θ from the two cases will be very close but the width of the posterior for the larger sample size is much narrower than for the smaller size.
- **Using the Bayesian posterior automatically takes into account the effects of sample size and "statistical significance".**



Operated by Los Alamos National Security, LLC for the U.S. Department of Energy's NNSA

NNSA

What Should We Report?

- **It may be perfectly satisfactory to summarize the posterior for θ with some kind of measure of central tendency and some kind of measure of width.**
  - Mean and standard deviation (analytic)
  - Median and 5% and 95% percentiles (easily obtained numerically)
- **However, we may need to address the basic question of "improvement or no improvement" more directly.**
  - The posterior probability that the new methodology is better than the old is the probability that $\theta > 1/2$, which is

$$P\left[\theta > 1/2 | s, \alpha_0, \beta_0\right] \equiv \int_{1/2}^{1} d\theta \, p(\theta | s, \alpha_0, \beta_0)$$
$$= 1 - I_{1/2}(s + \alpha_0, n - s + \beta_0)$$

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

Operated by Los Alamos National Security, LLC for the U.S. Department of Energy's NNSA

NNSA

Two Hypotheses and Their Posterior Probabilities

- **What, however, if we are required to answer "yes" or "no" to the question, "Is the new methodology better than the old?"**
- **Now, we are in the realm of "hypothesis testing".**
- We have, in effect, two mutually exclusive and exhaustive hypotheses:
  - $H_0 \iff \theta \leq 1/2$, corresponding to no improvement in predictive capability; and
  - $H_1 \iff \theta > 1/2$ corresponding to improvement.
- Of course, we have the posterior probability for H_1 from the previous slide:

$$P\left[H_1 | s, \alpha_0, \beta_0\right] = 1 - I_{1/2}(s + \alpha_0, n - s + \beta_0)$$

Operated by Los Alamos National Security, LLC for the U.S. Department of Energy's NNSA

NNSA

A Simple Binary Test

- **If we are forced to make a binary choice between H_0 and H_1, it is hard to see how we can, in good conscience, do anything but choose H_0 if its posterior probability is greater than the posterior probability for H_1 and vice versa.**
- In other words we choose between H_0 and H_1 according to,

$$\begin{aligned} \text{If } P\left[H_1|s,\alpha_0,\beta_0\right] &> P\left[H_0|s,\alpha_0,\beta_0\right], \text{ decide for } H_1, \\ P\left[H_1|s,\alpha_0,\beta_0\right] &\leq P\left[H_0|s,\alpha_0,\beta_0\right], \text{ decide for } H_0. \end{aligned}$$

- It is easy to show that $P\left[H_1|s,\alpha_0,\beta_0\right] > 1/2$ (and thus $P\left[H_1|s,\alpha_0,\beta_0\right] > P\left[H_0|s,\alpha_0,\beta_0\right]$) if and only if $s > \frac{1}{2}\left(n+\beta_0+\alpha_0\right)$. So, our test becomes:

$$\begin{aligned} \text{If } s &> \frac{1}{2}\left(n+\beta_0+\alpha_0\right), \text{ decide for } H_1; \\ \text{If } s &\leq \frac{1}{2}\left(n+\beta_0+\alpha_0\right), \text{ decide for } H_0. \end{aligned}$$

- Note that, when the prior is flat, the condition to decide for H_1 becomes $s > n/2$.

The Posterior Confidence

- Although our simple binary test is certainly straightforward, it is not entirely satisfactory in that it makes no distinction, for example, between a case where the posterior probability for H_1 is 0.6 and a case where it is 0.99.
- This consideration leads naturally, as we will show, to a three-way test where we choose between H_0, H_1, and "Do not know".
- The challenge here is to provide an operational definition for "Do not know".
- In order to do this, we first define the *posterior confidence, C*, as the posterior probability that the simple binary test gives the right answer. In other words,

$$C \equiv C(s) \equiv \begin{cases} P\left[H_1|s\right], & \text{when } P\left[H_1|s\right] > P\left[H_0|s\right], \\ P\left[H_0|s\right], & \text{when } P\left[H_1|s\right] \leq P\left[H_0|s\right]. \end{cases}$$

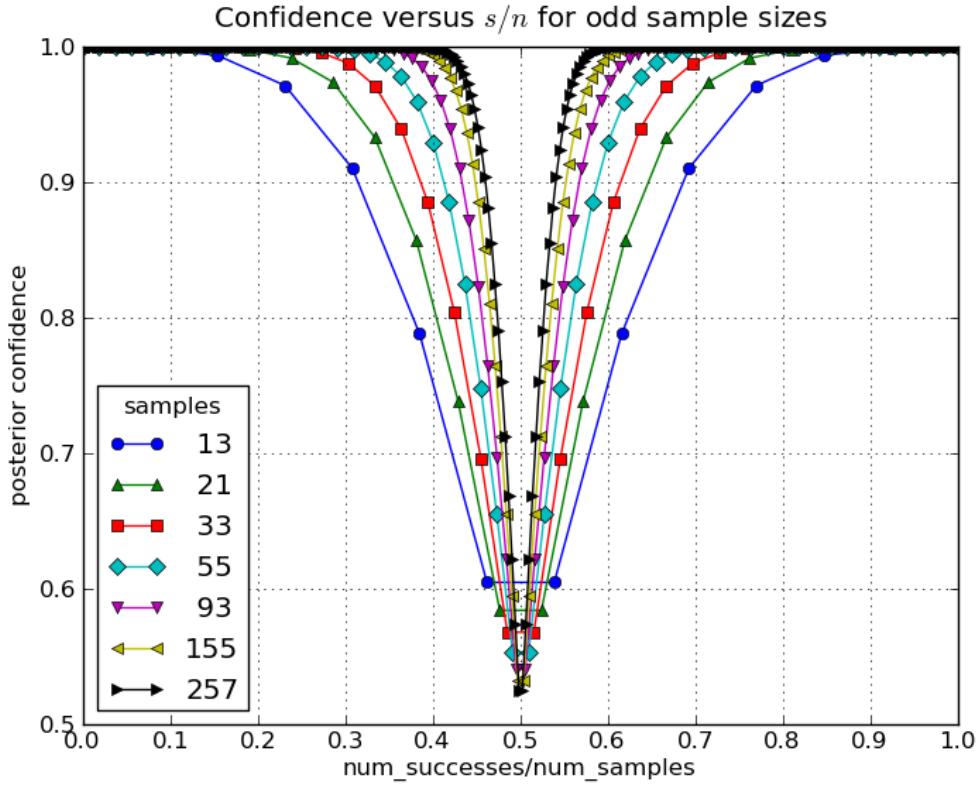
Operated by Los Alamos National Security, LLC for the U.S. Department of Energy's NNSA

NNSA

The Posterior Confidence

As a Function of the Data and Sample Size for a Flat Prior

- The confidence is never less than 0.5.
- For fixed sample size, the confidence increases as the measured s/n moves away from $1/2$.
- Except when $s/n = 1/2$, the confidence increases with sample size for a given s/n.
- 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.



A Three-Way Test

- We are now in a position to demand a given level of confidence, $\phi : 1/2 < \phi < 1$.
- Thus, we have the following three-way test:
  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 we do not have the required level of confidence in either hypothesis).
- In the following, we will take $\phi = 0.9$, although other choices are certainly possible.

Predicting the Results of the Tests

- We are using 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, $P[H_1|s]$, is greater than the posterior probability for no improvement, $P[H_0|s]$, and thereby provides us with our binary test; and
  2. The posterior confidence, C, which allowed us to define our three-way test.
- Once we have taken our data and computed the posterior density for θ, these two quantities will have perfectly definite values.
- However, now, 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.
- In order to determine how many simulations 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.

The Binary Test

Predictive Probabilities for Type 1 and 2 Errors

- **It is useful to examine the predictive probabilities for making Type 1 and 2 errors when we use the simple binary test.**
  - We have made a *Type 1* error if we have concluded that H_1 is true when, in fact, H_0 is true.
  - We have made a *Type 2* error if we have concluded that H_0 is true when, in fact, H_1 is true.
- **We can easily compute the probabilities for making these types of errors as a function of the true value of θ.**
  - Of course, we can only make a Type 1 error when $\theta \leq 1/2$ and we can only make a Type 2 error when $\theta > 1/2$.
- **We can also compute the overall probability of making either kind of error by averaging the probability as a function of θ over θ.**
  - This kind of averaging is heavily dependent on our choice of prior.

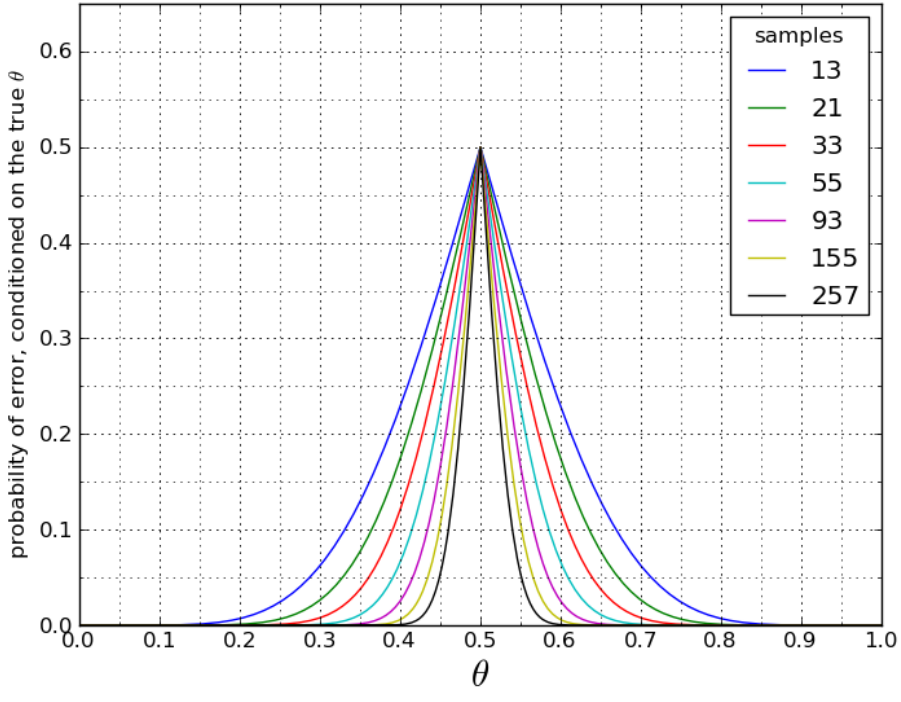
Operated by Los Alamos National Security, LLC for the U.S. Department of Energy's NNSA

NNSA

Predictive Probabilities for Type 1 and 2 Errors

As a Function of the True θ, Flat Prior, Odd Sample Size

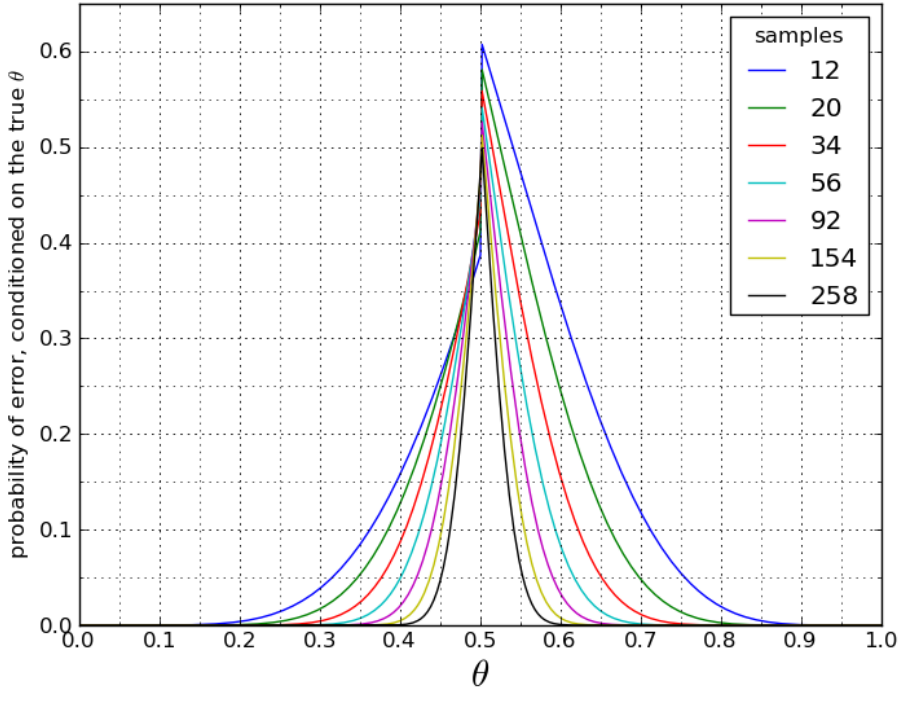
- The probability of error when the true value of θ is $1/2$ is always 50%, regardless of sample size.
- The probability of error for all other values of θ goes to zero with sample size but there is always a region around $1/2$ where the probability of error is rapidly increasing to 50% at $1/2$.
- If we define this region more precisely as the region over which the probability exceeds either 10% or 5%, we get similar results for the width of the region.
  - At 13 samples, the width of the region is roughly ± 0.2.
  - At 55 samples it is roughly ± 0.1.
  - At 257 samples it is roughly ± 0.05.



Predictive Probabilities for Type 1 and 2 Errors

As a Function of the True θ, Flat Prior, Even Sample Size

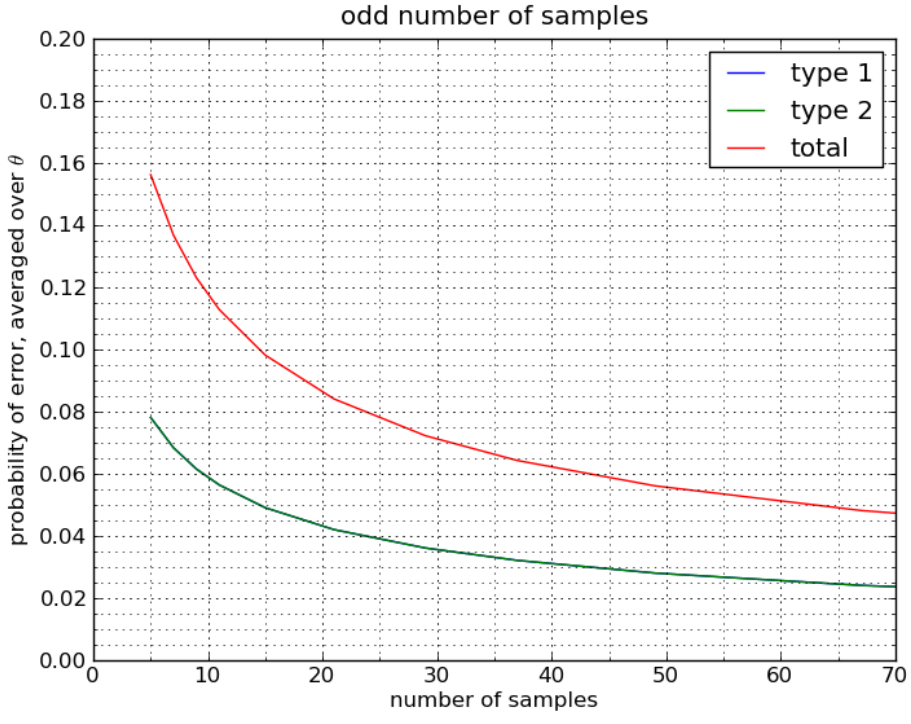
- The plot is dramatically asymmetric, especially near $\theta = 1/2$.
- The asymmetry is due to the discrete nature of the data and the slight asymmetry in the definitions of our hypotheses, where we chose to put the case $\theta = 1/2$ in H_0 rather than in H_1.
- Away from $\theta = 1/2$ and for sample size not too small, the asymmetry is small and the probabilities for Type 1 and Type 2 errors are close to the probabilities with odd sample size.
- For clarity, we will restrict ourselves to examining odd sample size hereafter.



Predictive Probabilities for Type 1 and 2 Errors

Averaged over θ, Flat Prior, Odd Sample Size

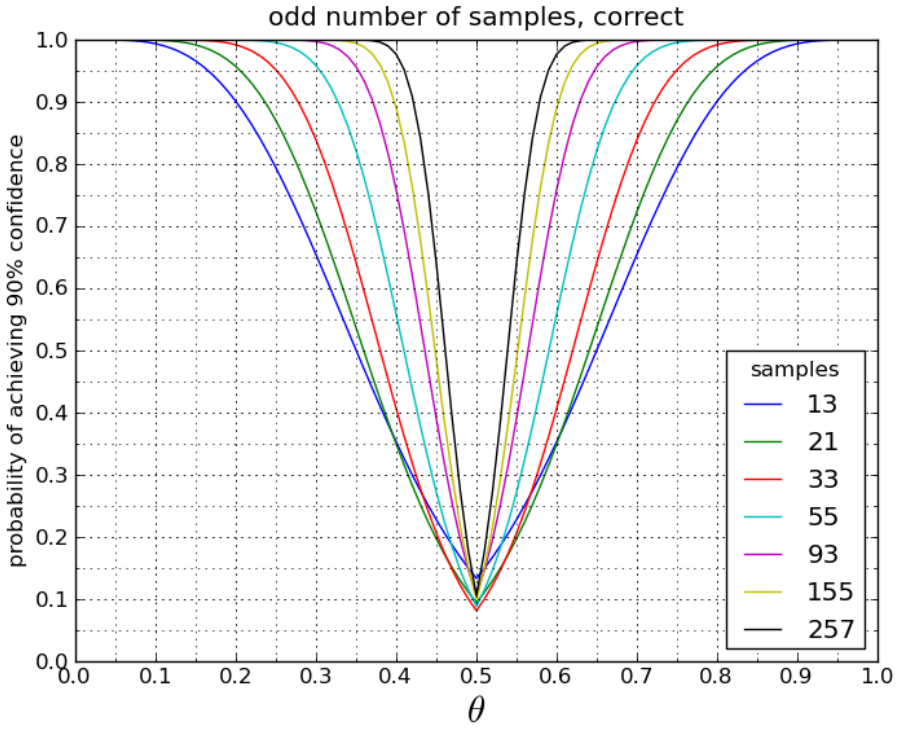
- For odd sample size, the probabilities for Type 1 and 2 errors are equal, so the curves lie on top of each other.
  - For even sample size, the total probability of error is similar to the case of odd sample size at similar sample sizes.
- The overall probability of error is under 10% for sample size over about 15 and under 5% for sample size over about 65.
- Remember, this result is strongly dependent on our choice of prior for θ!



Predictive Probabilities for The Three-Way Test

The Probability of Achieving 90% Confidence, as a Function of the True θ

- This shows the probability of achieving 90% confidence **and** being **correct** in the simple binary test.
- In general, 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.

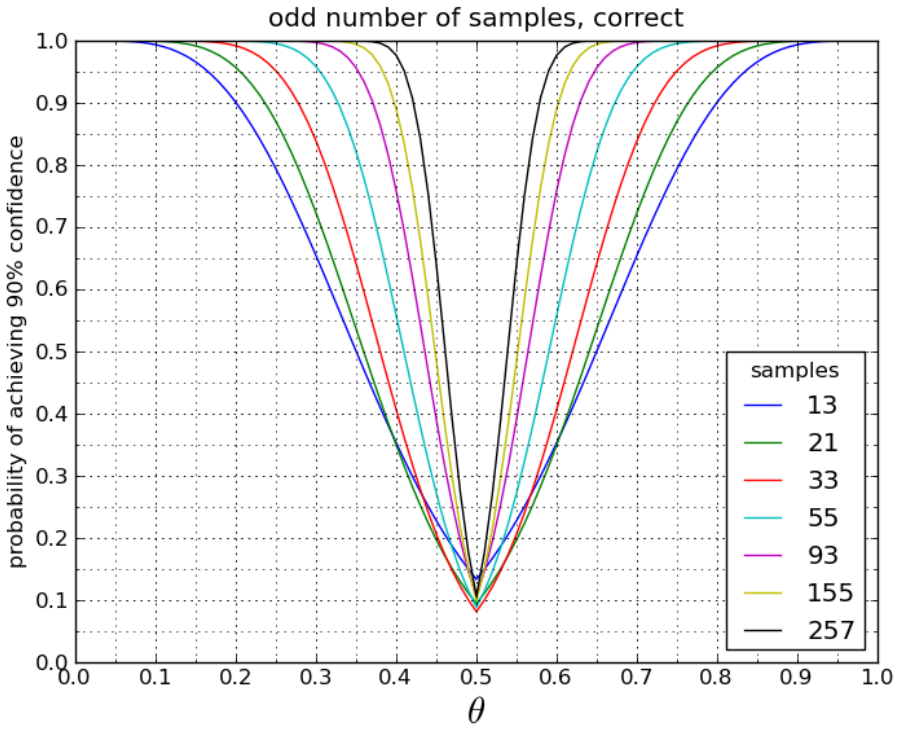


- However, there is a region around $\theta = 1/2$ where some of the curves for different sample sizes cross. This is due to the discrete nature of our data and our resulting inability to precisely specify a quantile like “90% confidence”.

Predictive Probabilities for The Three-Way Test

The Probability of Achieving 90% Confidence: Some Characteristic Values

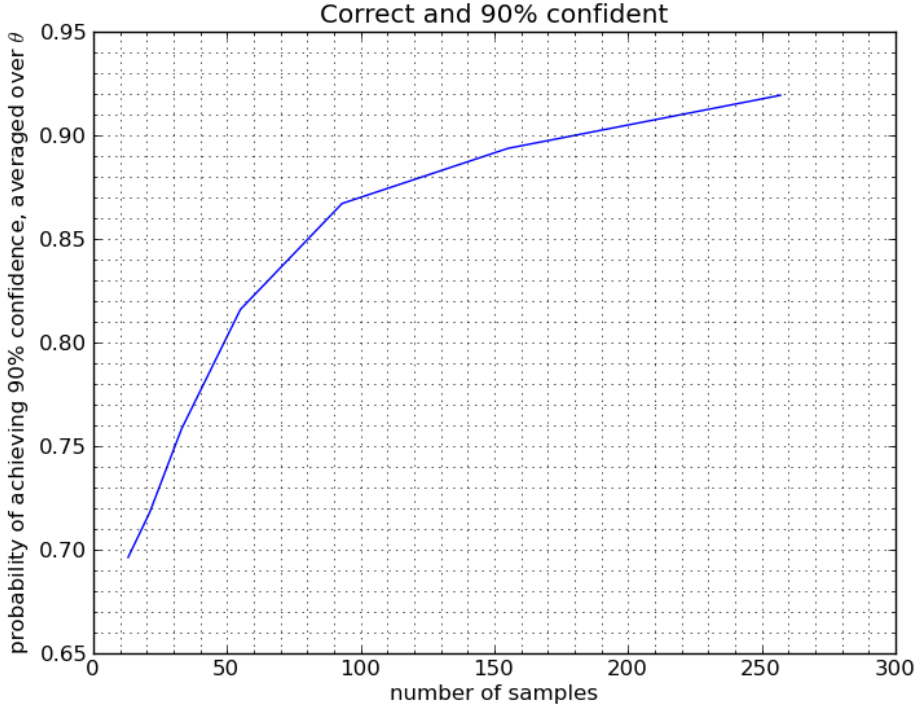
- If we are willing to accept an 80% predictive probability of achieving 90% confidence, and are not concerned with a difference of θ from $1/2$ of less than 0.15, then 55 samples will suffice.
- If, on the other hand, we are concerned with a difference of θ from $1/2$ of more than 0.06 (again accepting 80% predictive probability for achieving our target confidence), then we will need many more samples, on the order of 250.



Predictive Probabilities for The Three-Way Test

The Probability of Achieving 90% Confidence, Averaged over θ, Flat Prior

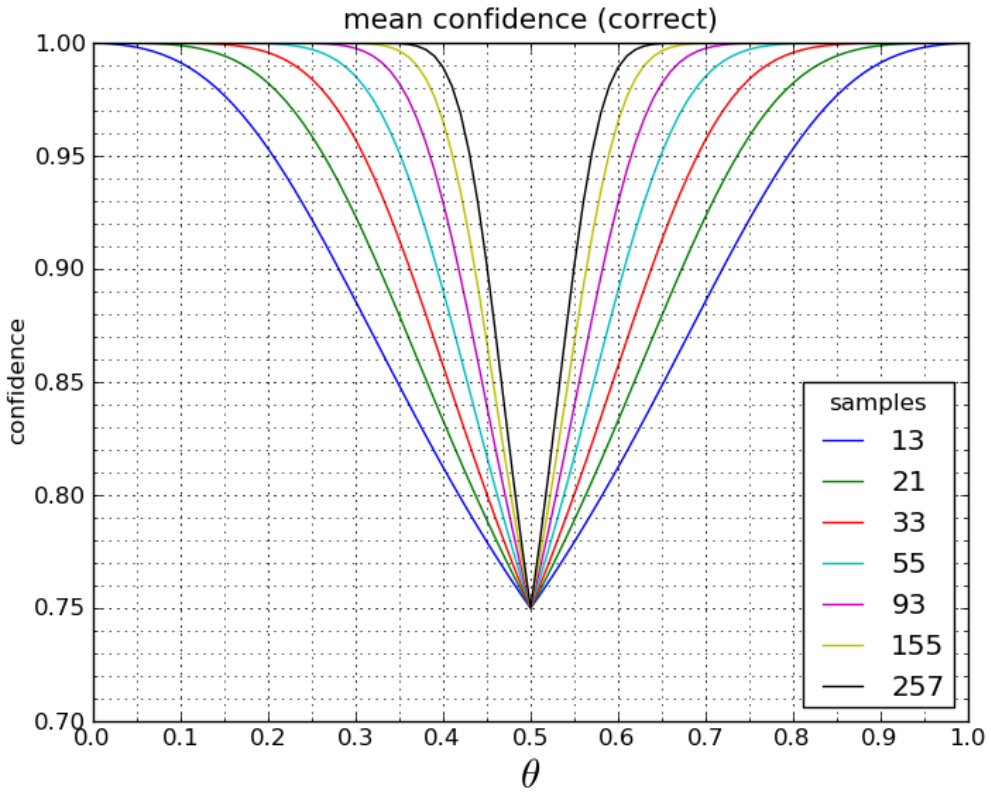
- The figure shows the overall predictive probability of achieving 90% confidence **and** being **correct** in the simple binary test.
- We have an overall 80% predictive probability of achieving 90% confidence with 50 samples.
- 90% predictive probability occurs only with about 180 samples.
- Remember, this result is strongly dependent on our choice of prior for θ!



The Predictive Mean of the Posterior Confidence

As a Function of the True Value of θ, Flat Prior

- Even when the true value of $\theta = 1/2$, the predictive mean of the posterior confidence is still 75%.
- **It really does not make sense to pick a target confidence much less than 90%.**



Sequential Analysis

- We can perform sequential analysis using the posterior for θ based on the comparisons we have already done as the prior for the next stage in our analysis.
  - Remember, if we have already analyzed (using a flat prior) n_0 data values with s_0 ones and $n_0 - s_0$ zeros, then our posterior (and the prior for the next stage) is a beta-distribution with $\alpha = s_0 + 1$ and $\beta = n_0 - s_0 + 1$
- This means that we can do comparisons for some experiments, then **reevaluate our estimates** of how many more comparisons will be necessary.
- Once we begin doing sequential analysis, **our priors will have a much more solid basis** (the comparisons done so far) so **our estimates of the probabilities averaged over θ will also have a much more solid basis.**
- Conversely, **our results conditioned on the true value of θ will be less meaningful because:**
  - Much of the range of θ will be eliminated by the previous analysis; and
  - Even within the range for θ that is left, different values will have very different prior weights.

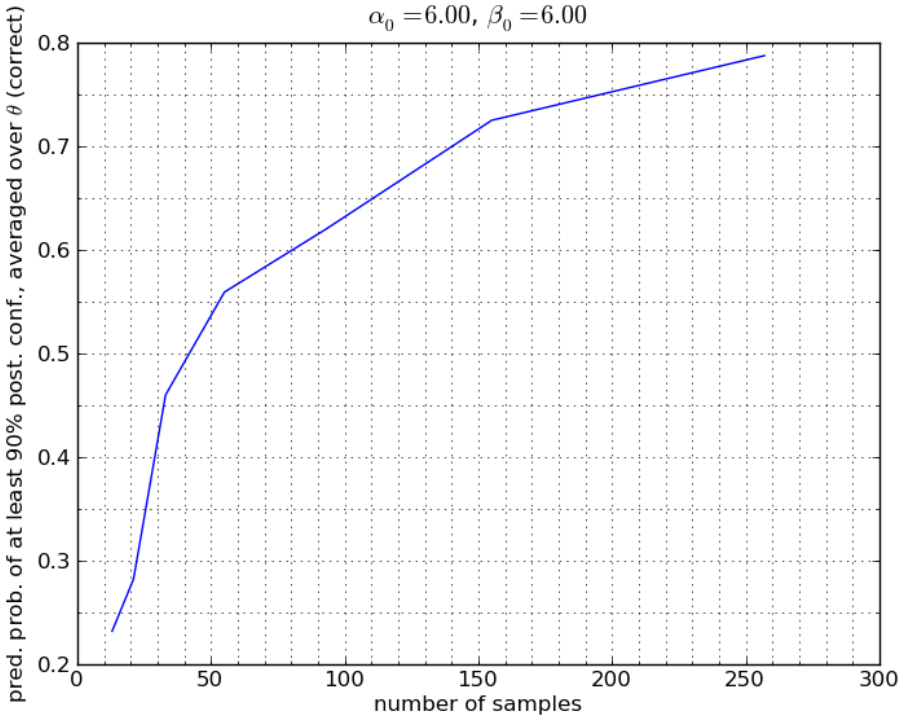
Sequential Analysis: The Effect on Power Analysis

- If we use the posterior from analysis of some data as a prior in the power analysis the results can change **dramatically**, even for a relatively small amount of data, because we now have some real information about where θ lies.
- If the data show that θ is far from 1/2, then we will already have considerable confidence and the power analysis will predict even greater confidence if we analyze more data.
- If the data show that θ is close to 1/2, then we will not have high confidence and the power analysis will predict that we will need **a lot** of new data to achieve even reasonable confidence.

Predictive Probabilities for The Three-Way Test

The Probability of Achieving 90% Confidence, Averaged over θ, $\alpha_0 = 6$ and $\beta_0 = 6$

- Assume we have observed ten samples with $s = 5$ ones and $n - s = 5$ zeros. Then $\alpha_0 = 6$ and $\beta_0 = 6$.
- Our confidence is 50% (the lowest possible value)
- If we take 50 future samples, we have only about a 55% predictive probability of achieving 90% confidence.
- Even at 250 future samples, we only have about a 79% probability of achieving 90% confidence.
- **We are in a situation where we have the lowest possible confidence and do not expect to be able to achieve 90% confidence in the future.**



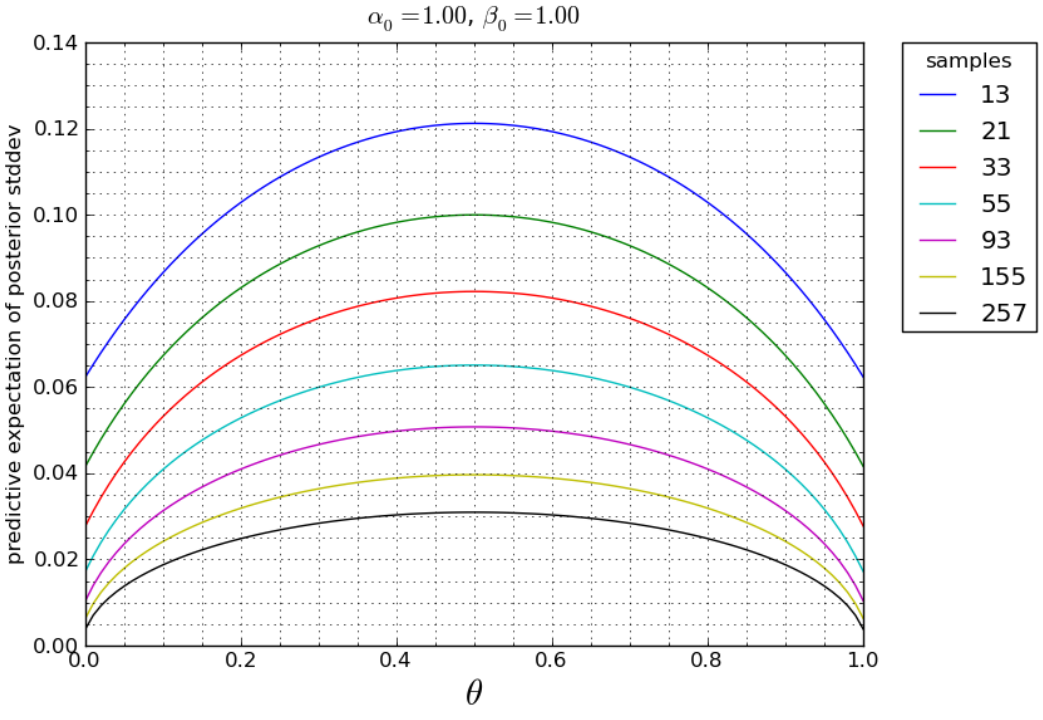
Plan B

- **The Problem:** The data may tell us that we have little confidence in either H_0 or H_1 and that we do not expect to attain much better confidence without a lot more data.
  - This is not a very satisfactory situation.
  - **But note that this will only happen when the data is also telling us that θ is close to 1/2.**
- **The Solution:** Characterize how far θ can be from 1/2 (given the current data) and how accurately we expect to be able to estimate it with more data.
  - **Justification:** There must be some region around $\theta = 1/2$ where the difference (of θ from 1/2) is not enough to make a difference.
  - **Metrics for characterizing how far θ can be from 1/2:** The posterior mean and standard deviation.
  - **Metric for the power analysis:** The predictive mean of the posterior standard deviation.

The Predictive Mean of the Posterior Standard Deviation

As a Function of the True θ, Flat Prior

- The predictive mean of the posterior standard deviation goes down with sample size for a given θ (duh).
- The predictive mean does not vary a lot as a function of θ, as long as θ is not too close to 0 or 1 (where we do not care about it).



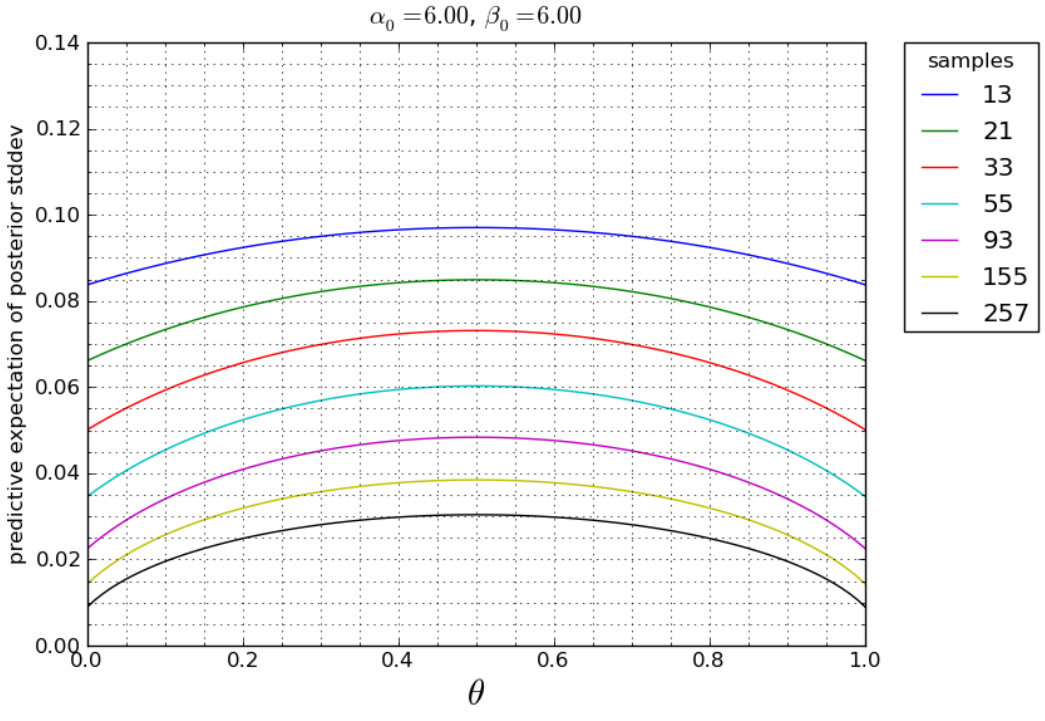
Operated by Los Alamos National Security, LLC for the U.S. Department of Energy's NNSA

NNSA

The Predictive Mean of the Posterior Standard Deviation

As a Function of the True θ, $\alpha_0 = 6$ and $\beta_0 = 6$

- Assume we have observed ten samples with $s = 5$ ones and $n - s = 5$ zeros. Then $\alpha_0 = 6$ and $\beta_0 = 6$.
- The predictive mean varies as a function of θ even less than it did with a flat prior.
- Except for the smallest sample sizes, the curves are very close to those on the previous slide.
- **Conclusion: The predictive mean of the posterior standard deviation is not very sensitive to prior information.**



Recommendations

- Start with a flat prior for θ (in order to be fair to both the methodologies).
- Decide for H_0 or H_1 based on the simple binary test, but be prepared to say, "I don't know," if the posterior confidence is not sufficiently high.
- Compute the posterior confidence for your choice and quote it to your audience.
- Compute the predictive probability for achieving 90% confidence for various sizes of (future) samples.
- If the confidence is low and the predictive probability of achieving 90% confidence with a reasonable amount of additional data is not sufficiently high, compute:
  - The mean and standard deviation of the posterior, so you will know how far θ can be from 1/2.
  - The predictive mean of the posterior standard deviation for various future sample sizes, so you will know how much additional data will be necessary to nail θ down further.

Some Final Thoughts

- In order to decide how many simulations we will need to do, we need to have some sense of how far θ needs to differ from 1/2 for us to regard the difference as making a difference.
  - The comments on the figures provide some guidance by showing how many simulations we will need as a function of the true value of θ.
  - Ultimately, the decision as to how far θ must be to 1/2 in order for us to care is a value judgment.
- The issues we have dealt with here—especially the existence of intractable cases of low confidence—are likely to show up whenever one tests hypotheses that correspond to regions of the real line (or any continuous space).
- The solutions we have proposed—the three-way test and our plan B—are also likely to be useful quite generally.

Why I Did Not Do Traditional Hypothesis Testing

- A point null hypothesis with $H_0 \Leftrightarrow \theta = 1/2$
  - Does anyone really think that θ is exactly 1/2 in this case?
  - Does anyone care?
- A region null hypothesis with $H_0 \Leftrightarrow \theta \leq 1/2$
  - The analysis must be based on the case that produces the worst type 1 error, which is $\theta = 1/2$.
  - To get a reasonable probability for type 1 error, we must put the division point for our choice far to the right of $\theta = 1/2$. This seems very wrong if we want to do a fair comparison of legacy and new methodologies.
  - The root of the problem is that the traditional approach lumps together a positive rejection of the alternative hypothesis and an inability to decide between the two hypotheses under the rubric "failure to reject the null hypothesis".

Why the Curves for the Predictive Probability of Attaining 90% Confidence Cross

- The priors are flat
- 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.
- It is clear that the odd behavior of the lower plot is due to the variation in the cutoff.

