HOMELAND SECURITY
COMMUNITY OF BEST PRACTICES

Bayesian Model Checking

August 2022

Dr. James Theimer, CTR

To enhance T&E science through multidisciplinary collaboration and deliver it to the DHS workforce through independent consultation and tailored resources.

About this Publication:
This work was conducted by the Homeland Security Community of Best Practices under contract FA8075-18-D-0002, Task FA8075-21-F-0074.

For more information:
Visit, www.AFIT.edu/STAT
Email, HSCoBP@afit.edu

Homeland Security Community of Best Practices
2950 Hobson Way
Wright-Patterson Air Force Base, Ohio

Version: 1, FY22

THEORY into PRACTICE

Executive Summary

One potential risk in applying Bayesian methods is that it can lead an analyst to draw flawed conclusions if the prior knowledge of the system under test does not describe the present system well. This report offers a Box model-checking criterion that can detect when a Bayesian prior does not describe the current system well. This Box method computes how probable it is that we are to see the data observed, or something less probable, given our selected prior. If the probability computed is very low, then can we conclude that the prior does not appear to describe our system well. This report describes the computation of the probability we are using as criterion. The report gives an example of how an inaccurate prior could lead to flawed conclusions, and also how the method described could identify this condition.

Keywords: Bayesian statistics, Model checking, Model assessment

Table of Contents

Introduction

Program managers care about the risk of accepting a bad system and about the risk of rejecting a good one. Bayesian methods offer a natural way of discussing how to balance those risks. Bayesian methods also offer a way of updating prior knowledge by using collected data, which may offer a rigorous way of combining information, so that decision makers can make program decisions with all prior information.

A classic objection to using Bayesian methods is that one can create a prior distribution that makes acceptance of a system inevitable. Model-checking methods exist for Bayesian models, which can detect priors that may guarantee the acceptance of the system. This best practice aims to discuss a method proposed by Box (Box, 1980) to help programs identify prior distributions that are inconsistent with the data collected. The Box Criterion can check if the observed data appear very unlikely given the prior distribution. If the data are very unlikely, the prior appears inconsistent with the data and it is reasonable to question validity of the prior. Offering this Box method to T&E practitioners and program managers should encourage people to use Bayesian methods.

This Best Practice first reviews some basic aspects of Bayesian methods. Then, the Box Criterion is introduced and is applied to a sample problem to show that it can identify prior distributions that are inconsistent with the data collected. The next section shows examples of the probability of observing a given number of successes given different probabilities of success. Overall, the example provided demonstrates that the Box Criterion will more than likely reject a model if it does not fit the data collected. The conclusion offers comments on the implications of rejecting the model and summarizes key takeaways.

Background

Model checking is a common part of statistical practice, so a way is needed to determine if the prior and sampling distributions found in a Bayesian model appear consistent with the data collected. The sampling distribution describes the probability of certain outcomes of experiments based on model parameters, and the prior distribution is the Bayesian probability of those parameters. If the prior and sampling distributions are not consistent with the data then it can be inferred that the model may poorly describe the system being tested. In general, a model can have a problem with either the sampling distribution, or the prior distribution. This report will offer an example that assumes the sampling distribution fits the system well, so that the problems are only in the prior distribution.

For the purposes of this Best Practice, the word probability will be used in two senses: frequentist probability and Bayesian probability. Frequentist probability is defined in terms of the relative frequency. Relative frequency is the ratio of the number of times an event occurs in relation to the number of observations (Hogg, McKeen, & Craig, 2013, p. 2). Alternatively, Bayesian probability is the level of confidence that a statement is true. (Hogg, McKean, & Craig, 2013) on pg. 2 refer to this as "subjective probability" and illustrate it as the willingness to make a bet that a statement is true.

For example, a program might need to estimate the Bayesian probability that the probability of success is greater than some value. In this report the probability of success will be thought of in

the frequentist sense, and Bayesian uses will be identified as Bayesian probabilities, which are used to quantify confidence that a requirement has been met. An advantage of the Bayesian approach to probability is that it directly discusses our confidence in things like the probability of success being greater than some level. In effect, the program is making a bet that the system will work as required. As the trade-offs that the program need to make become more complicated, Bayesian analysis offers decision makers a more intuitive way to discuss confidence.

The situation in which there can be two outcomes (the system works, or it does not) is an example of binary data, and the probability of observing *y* success out of *n* trials is described by the binomial distribution. A convenient model is one where the Binomial distribution is used as the sampling distribution and Beta distribution is used as the prior. These are conjugate distributions, so the posterior also has a Beta distribution. If the prior distribution has parameters *a* and *b* the posterior distribution for the Bayesian probability of the estimated probability of success will have parameters $y + a$ and $n - y + b$. The estimated probability of success will be represented by θ, and is given by

$$p(\theta|y) = \frac{\Gamma(n+a+b)}{\Gamma(y+a)\Gamma(n-y+b)}\theta^{y+a-1}(1-\theta)^{n-y+b-1}\ 0 < \theta < 1. \quad (1)$$

The function $\Gamma(y)$ is the gamma function and it can be computed in many math packages, including Excel. In Equation 1 the hyper-parameter *a* acts as an additional number of successes, *b* acts as an additional number of failures, and $a + b$ acts as an additional number of observations. This is a way of thinking about the relative weighting of prior and experimental observations; $a + b$ is the weight given to the prior data and n the weight given to the experimental data.

The predictive prior distribution is the probability distribution of observations we would expect given our model and our prior, marginalizing over the model parameter(s). The prior predictive distribution of observations is related to the prior distribution of the parameter and the sampling distribution (Gelman, et al., 2013) p. 7 by,

$$p(y) = \int f(y|\theta)\, f(\theta)\, d\theta. \quad (2)$$

For our model, it can be shown that (Christensen, Johnson, Branscum, & Hanson, 2011) p. 58,

$$f(y) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\binom{n}{y}\frac{\Gamma(y+a)\Gamma(n-y+b)}{\Gamma(n+a+b)}. \quad (3)$$

Equation 3 is in the form of the Beta Binomial distribution. The prior predictive distribution is a Bayesian probability for the number of successes we will observe in a test.

Having reviewed some concepts that will be used, an example of the problem will be examined to motivate the need for model-checking.

Using Bayesian Methods to Assess if a Requirement has Been Met

The use of the Box Criterion will be shown with an example that mimics a typical program situation where it must be demonstrated that the system will work with high probability in a

given situation.

The use of model checking will be illustrated by computing the posterior distribution of θ, given in Equation (1), with three prior distributions of θ. The results are shown in Table 1 and Table 2. In Table 1 the ratio of successes to observations is held constant at 0.95 and in Table 2 at 0.85. It is assumed that it is required that the probability of success is greater than 0.9, and that a Bayesian probability greater than 0.9 is required.

The Bayesian probability is computed by integrating the posterior distribution over values greater than 0.9. This is the sort of analysis required to prove a system will work with high probability and with high confidence. The prior of $a = 1$ and $b = 1$ is a reference prior where it is assumed the prior is formulated based on no prior knowledge. The prior where $a = 99$ and $b = 1$ is a situation where there is a strong belief that the system will work. The prior where $a = 18$ and $b = 2$ is a model where the Beta distribution of the prior has a mean of 0.9 and it is weighted as if there were 20 prior observations, so this distribution falls between the others in many properties. The mean of the prior distribution also falls between the observed proportions of successes in the two tables. The intent was to have an example that was not very strong and not completely representative of either data set. Table 1 shows the effect of collecting 20, 40, and 80 trials, so the relative weight of the prior distribution and the data is changed as well.

Table 1

Estimated Bayesian Probability that the Probability of Success is Greater than 0.9 for Different Sample Sizes Where Success is Observed 95% of the Time and with Different Priors.

	a = 1 ,b = 1	a = 18 ,b = 2	a = 99 ,b = 1
y = 19 n = 20	0.63527	0.762168	0.999949
y = 38 n = 40	0.791425	0.853697	0.999941
y = 76 n = 80	0.917295	0.938848	0.99995

Table 1 shows that many data points have to be collected to achieve a Bayesian probability of 0.9, unless a very strong prior is used. The data indicate the probability of success is most likely to be about 95%. However, the reported Bayesian probability, that the probability of success exceeds 90%, changes based on the selected prior. This illustrates that if prior data is available it could help confirm that the requirement is met with fewer trials. It also shows why the answer to the problem of what prior to use may not simply be to use a reference prior. If one has prior knowledge, there are advantages to using it.

Table 2

Estimated Bayesian Probability that the Probability of Success is Greater than 0.9 for Different Sample Sizes Where Success is Observed 85% of the Time and with Different Priors

	a = 1 ,b = 1	a = 18 ,b = 2	a = 99 ,b = 1
y = 17 n = 20	0.151965	0.350394	0.99829
y = 34 n = 40	0.110192	0.234205	0.988348
y = 68 n = 80	0.0584953	0.116694	0.916047

Table 2 shows a situation where the observed proportion of successes in the sample is 0.85, so it does not support that the system has a probability of success greater than 0.9. Even with 80 trials, the result does not indicate that the system should be rejected when $a = 99$ and $b = 1$.

This is an example of an overconfident prior, where the prior is such that the system will not be rejected unless a vast amount of data are collected. A way of flagging overconfident priors like this is needed. In order to offer proof from the data that a system meets a requirement, the validity of the model needs to be demonstrated and the model must show that the Bayesian probability that the probability of success meets the requirement and is sufficiently high.

The Box Criterion

As illustrated above, one could select a prior such that it basically guarantees that the requirement will be met. An approach to checking the model was proposed by Box (Box, 1980), who suggested a *p* test based on the prior predictive probability of the observed value. In our case, this is done by summing over the probability of the observed value, and values with lower probability than that observed. This can be expressed as

$$\boldsymbol{p} = \sum_{\boldsymbol{y}:\boldsymbol{f}(\boldsymbol{y})\leq\boldsymbol{f}(\boldsymbol{y}_{\boldsymbol{obs}})} \boldsymbol{f}(\boldsymbol{y}), \tag{4}$$

where f(y) is the prior predictive distribution computed with Equation 3. If the value of *p* is small, the value observed is surprising, and it suggests that either the sampling distribution, or the prior does not represent the system well. This can be computed by calculating f(y) for the number of successes observed, and calculating a set of all the values of f(y) for the integers from zero to *n*. One can then select all elements of this set with values less than or equal to the predictive prior probability of the observed number of observations, and take the total of all of them.

Application of the Box Criterion

Table 3 and Table 4 show the application of the Box Criterion to our sample problem. It is assumed that the hypothesis that the model is acceptable is rejected if the criterion falls below 0.05. We see that if the ratio of the number of success to the number of observations is 0.95 and we have collected 80 observations, then we reject the prior where $a = 99$ and $b = 1$. The prior is a Beta distribution with a mean of 0.99, so with many data points the test is sensitive enough to detect that the data does not fit. For the reference distribution, the prior is not rejected, but the distribution is such that all observations are equally likely, so none of the possible number of observations is surprising. For the observed frequency proportion of 0.85, the prior is rejected when $a = 99$ and $b = 1$ under all three notional samples, so it would have flagged this prior as not reflecting the data to a significant degree.

Table 3

Box Criterion for Different Sample Sizes Where Success is Observed 95% of the Time and with Different Priors.

	a = 1,b = 1	a = 18,b = 2	a = 99,b = 1
y = 19 n = 20	1.	1.	0.168067
y = 38 n = 40	1.	1.	0.0813262
y = 76 n = 80	1.	1.	0.0382423

Table 4

Box Criterion for Different Sample Sizes Where Success is Observed 85% of the Time and with Different Priors.

	a = 1 ,b = 1	a = 18 ,b = 2	a = 99 ,b = 1
y = 17 n = 20	1.	0.322245	0.00416333
y = 34 n = 40	1.	0.268424	0.000427453
y = 68 n = 80	1.	0.271232	0.0000388784

Probability of Observing a Given Number of Successes

It is useful to think about the probability of observing a given number of success if the actual probability of success were known. The results above show results for two proportions of observed successes. Those proportions were chosen as being representative, which is to say likely to be observed, for different true probabilities of success (one above the requirement of 90% and one below it). This section expands the last section by investigating the probability of observing a given number of successes. The plot in Figure 1 shows the probability of observing a given number of successes given an actual probability of success of 0.95. It also shows the Bayesian confidence of the estimated probability of success, given here by θ, is greater than 0.9, and the Box criterion that would be obtained. The figure allows us to discuss the probability of observing that Bayesian probability and Box criteria given that probability of success. Several examples will be stepped through to demonstrate how the Box criterion works in several situations where the prior fits the observed data more or less well.

The example shown in Figure 1 is for a reference prior with a uniform PDF. No value is more likely than any other is, so the probability of observing that value, or something more extreme, is the same, so the Box Criterion is always one. The reference prior has worked in that it was not supposed to indicate prior knowledge, so any outcome is not surprising. In this case the mode of the PDF is near a value where the Bayesian probability of the probability of success being greater than 0.9 is approximately 0.9. This value is what is given in the entry in Table 1. The figure allows the estimation of the probability of obtaining other values and so of the risk of rejecting this system given that that many samples are collected.

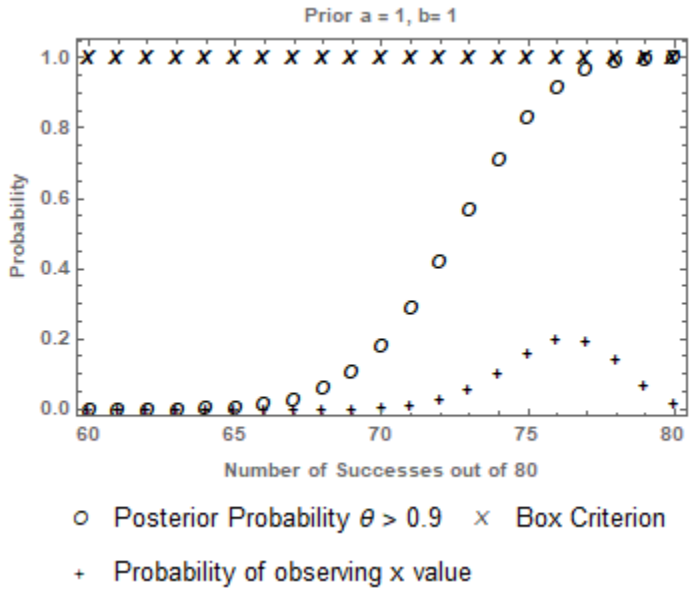

Figure 1
Reference prior probability of success is 0.95

Figure 2 shows an example where the prior distribution is $a = 99$ and $b = 1$. The mode of the prior is 1.0. The assumed actual probability of success in the figure is 0.95, and the difference is large enough that the most probable number of successes would produce Box criterion values such the prior would be rejected, as shown in Table 3. The figure also shows that the Bayes probability would create confidence that the probability of success was over 0.9 for any probable value of the number of observed successes. This is a situation where the Box criterion would have led to skepticism that the prior represents the system's performance well. The Bayesian probability would lead to acceptance of the system, but without this model checking procedure described here, the analyst might be unaware the system would pass the test no matter what the outcome of the test was.

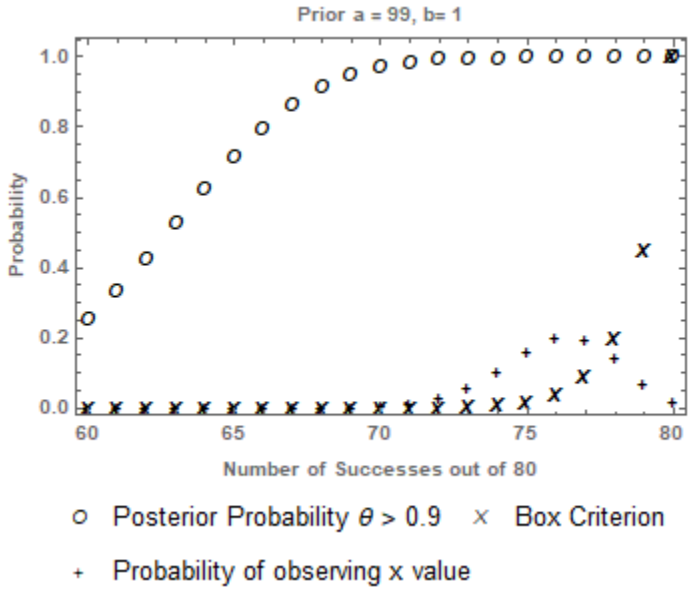

Figure 2
Over-optimistic Prior Probability of Success is 0.95

Figure 3 shows the same situation as Figure 2, except that it shows the PDF where the probability of success is 0.85. This is a probability of success below that required for the system. The prior is such that the Box Criterion would cause us to reject the prior for any reasonably likely observed number of successes. As pointed out in Table 2, for the mean number of observations, a Bayesian probability greater than 0.9 would still be estimated for meeting the probability of success requirement, and so the system would still be accepted if the decision were based only on evidence for meeting the requirement. This is an example of a situation where the Box criterion would have caused rejection of the prior, and the incorrect model would have caused overestimation of the Bayesian probability in a way that would have cause an incorrect judgement.

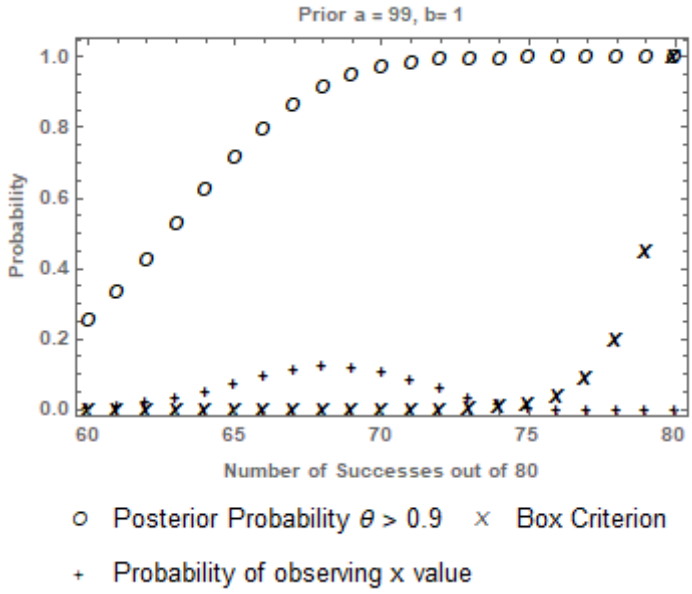

Figure 3
Over-optimistic Prior Probability of Success is 0.85

Figure 4 shows an example with a less informative prior with a mode number observation of 76. The figure shows a PDF for a probability of success of 0.95. If the probability of success were 0.85, it would have a PDF shown in Figure 3. In the example shown, the most likely number of success does not give Box Criterion scores that would cause us to question the model, so the difference between the model and the observed data is not deemed significant. The mode number of success estimated from Figure 3 is 68 if the true probability of success is 0.85. If the probability of success were 0.85, the most likely Box scores would be much lower, but still over 0.05, so the model would not be rejected. The estimated Bayesian probabilities are such that the probability of success being greater than 0.9 with Bayesian probability greater than 0.9 would be accepted, for the situation where the actual probability of success is 0.95. If the probability of success were 0.85, the model would be accepted, but the Bayesian probability would be low enough that the system would have been rejected for this reason as well. These are examples where the Box Criterion and the estimates from the model would have led to correct decisions.

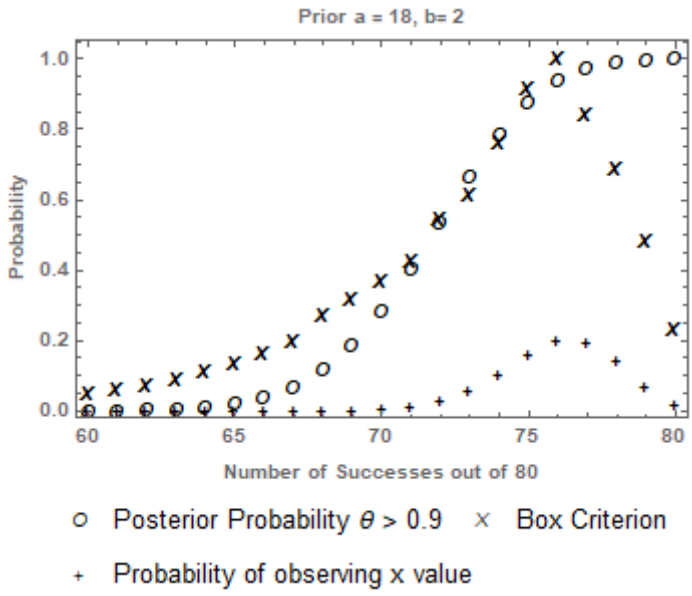

Figure 4

Illustrates an Example Where the Prior Roughly Describes the Probability of Success is 0.95

What to Do if the Model is Rejected

Inevitably, programs will want to know what to do if a model is rejected. Unfortunately, there is no easy answer. There is a general principle that the same data should not be used to build and to test the model, as this will lead to overconfidence in the results (Gareth, Witten, Hastie, & Tibshirani, 2013). The question of what to do if a model is rejected becomes one of model selection which is covered by Christensen et al. in Section 4.9, and by Gelman et al. in chapter 7, which offers proper ways of choosing a model. Model checking, by itself, could lead to the rejection of a model; however, it does not suggest a solution. Rather, it highlights that more data are needed to draw good conclusions. In practice, if an analyst finds that the model is rejected, they should consult an expert in Bayesian methods (to include but not limited to, consultants from the STAT COE) to determine the best way-forward.

Conclusion

The Box Criterion points out situations where the prior distribution is inconsistent with the observed data. More specifically, the Box criterion points out where the observed data are inconsistent with the model, including the prior distribution. A frequent concern to Bayesian methods is whether the prior does not correctly describe the system being tested. Examples illustrate how an analyst can apply this procedure after a test to provide assurance that the prior is credible.

For more information regarding Bayesian methods (Sieck, Kolsti, 2022), model checking (Burke, 2022), or probability (Sigler, 2022) please reference the STAT COE website: www.AFIT.edu/STAT.

References

Box, G. E. (1980). Sampling and Bayes' Inference in Scientific Modeling and Robustness. *Journal of the Royal Statistical Society A*, 383 - 430.

Burke, S. (2022, July 19). *Model Building Process Part 1: Checking Model Assumptions V 1.1.* Retrieved from AFIT/STAT Center of Excellence: https://www.afit.edu/STAT/statdocs.cfm?page=1126

Christensen, R., Johnson, W., Branscum, A., & Hanson, T. E. (2011). *Bayesian Ideas and Data Analysis.* Boca Raton, FL: CRC Press.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis* (3rd ed.). Boca Raton, FL: CRC Press.

Hogg, R. V., McKean, J. W., & Craig, A. T. (2013). *Introduction to Mathematical Statistics* (7th ed.). Boston: Pearson.

Kensler, J. (2022, July 19). *Interpreting Confidence Intervals.* Retrieved from AFIT/STAT Center of Excellence: https://www.afit.edu/STAT/statdocs.cfm?page=1126

Kensler, J., & Freeman, L. (2022, July 19). *Statistical Hypothesis Testing.* Retrieved from AFIT/STAT Center of Excellence: https://www.afit.edu/STAT/statdocs.cfm?page=1126

Montgomery, D. C. (2013). *Design and Analysis of Experiments* (8th ed.). Hoboken, NJ: John Wiley & Sons.

Reich, B. J., & Ghosh, S. K. (2019). *Bayesian Statistical Methods.* Boca Raton, FL: CRC Press.

Sigler, G. (2022 , July 19). *Statistics Reference Series Part 2: Probability.* Retrieved from AFIT/STAT Center of Excellence: https://www.afit.edu/STAT/statdocs.cfm?page=1126