SCIENTIFIC TEST
& ANALYSIS
TECHNIQUES
CENTER OF EXCELLENCE

Bayesian Methods
in
Test and Evaluation:
A Decision Maker's Perspective

July 2022

Maj Victoria Sieck, PhD
Dr. Kyle Kolsti

The STAT COE provides independent STAT consultation to designated acquisition programs and special projects to improve Test & Evaluation (T&E) rigor, effectiveness, and efficiency.

About this Publication:
This work was conducted by the Scientific Test & Analysis Techniques Center of Excellence.

For more information:
Visit, www.AFIT.edu/STAT
Email, CoE@afit.edu
Call, 937-255-3636 x4736

Scientific Test & Analysis Techniques Center of Excellence
2950 Hobson Way
Wright-Patterson Air Force Base, Ohio

The views expressed are those of the author(s) and do not necessarily reflect the official policy or position of the Department of the Air Force, the Department of Defense, or the U.S. government.

Version: 1, FY22

Modernizing *the* Culture *of* Test *&* Evaluation

Table of Contents

The field of statistics can broadly be broken into two branches: frequentist and Bayesian. The frequentist approach is more common in general use, academia, and Department of Defense (DOD) Test & Evaluation (T&E); however, Bayesian methods can provide many benefits, to include: (1) potential for better estimation, (2) more natural interpretability, and (3) flexibility. This Best Practice provides a high-level summary of how Bayesian methods can help DOD decision makers make more effective and efficient decisions about whether a system under test would provide the needed capabilities to the warfighter, or other similar questions of interest. The goal of this Best Practice is to offer T&E leadership with a concise treatment of Bayesian methods that can be referred to if questions or concerns arise.

In addition to discussing key benefits to a Bayesian approach, a few starting questions are offered here for leadership to explore whether a Bayesian approach has been appropriately applied.

More detailed references concerning this topic can be found in the Appendix.

Overview of Bayesian Methods and Key Terms

Fundamentally, Bayesian methods combine previously-known information with new information to update beliefs about the event or characteristic in question. From a non-statistical viewpoint, Bayesian methods are a statistical manifestation of the observe-orient-decide-act Loop (OODA Loop). From a T&E perspective, Bayesian methods allow testers to quantify their *a priori* understanding of the system under test; this understanding is then combined with information gained from actual testing, resulting in an updated understanding of system performance.

While frequentist approaches to testing are constrained to only formally including data into an analysis (that is to say, only including information that can be directly observed), Bayesian methods readily consider additional sources of information in a formal, mathematical way. This incorporation of available information can be accomplished through the development and use of priors (D. Berry 1993; S. Berry, *et al.* 2011; R. Christensen, *et al.* 2011). It is worth noting that the term “information” refers to a variety of different sources. For example, information can consist of:

- Previous data (raw data or summary statistics) that comes from the same distribution as the current data (e.g., data that would be considered appropriate for direct pooling in the current test)
- Previous data that comes from a related, but different, distribution as the current data (e.g., developmental testing [DT] data compared to operational testing [OT] data)
- Data from different (perhaps legacy) systems that have already been fielded
- Subject matter expert (SME) opinion / institutional knowledge
- An understanding of the natural bounds the model parameters can take on (e.g., if we want to estimate the heights of fighter pilots prior to removing the height restriction, we know the mean height must be between 64” and 74”)
- Believing we know nothing about the system under test (perhaps because it is a novel system that has only just been developed)

Key terms in a Bayesian approach to testing include data model, prior distributions (also referred to as priors), posterior distributions, posterior probability, and credible intervals (also referred to as posterior intervals). These terms are briefly introduced here; more information about these terms can be found in the resources provided in the References and Appendix.

The structure and form of the population from which the data are sampled during testing is described through the data model, which is conditioned on model parameters. For example, normally distributed data is conditioned on the model parameters μ (the mean) and σ^2 (the variance). The data model is how data enters a Bayesian analysis, which is used to construct a likelihood function (which also could be represented as a joint density of the data). Within frequentist analysis, this is the same likelihood function that is used for inference. Within the Bayesian framework, every parameter in a data model is unknown and has a prior distribution associated with it that is developed independently of the data to be collected. This prior distribution is used to quantify the uncertainty surrounding a given model parameter and represents an individual's beliefs about the model parameter before seeing the data. Therefore, priors are a means of incorporating key information about the model parameter into a statistical analysis. More concretely, priors are a formal statistical method to incorporate past knowledge from similar tests (either data or expert understanding) into an analysis.

Given the likelihood function, prior distribution(s), and newly-obtained data, Bayes' Theorem can be used to obtain the posterior distribution. The posterior distribution represents updated belief about all of the model parameters or corresponding system performance. Essentially, the posterior distribution is the combination of previous information through the prior with data through the likelihood function. Bayes' theorem when stated as a formula is actually quite simple; however, in practice, computational tools must commonly be used to numerically approximate the posterior distribution. This context is where the term "Markov chain Monte Carlo" (MCMC) will likely be encountered—a term that refers to a class of methods that could be used to obtain the posterior distribution.

Once the posterior distribution is obtained, various estimates of interests can be calculated. For example, point estimates or interval estimates for model parameters can be obtained to facilitate factor level analysis; the uncertainty (probability) in those point estimates can be captured in a credible interval. Alternatively, when the posterior distribution is a function of model parameters that defines performance metrics, posterior probability can be used to obtain the probability that a system will obtain the required threshold value for a measure or requirement, and similar questions of interest. For example, if the range of a mortar was modeled using a normal distribution, the posterior distribution could lead to inference about parameter values ("there is a 90% chance the standard deviation is between 45 and 64") or performance metrics ("there is a 92% chance the range exceeds the threshold of 250").

Benefits of a Bayesian Approach

While there are many benefits to a Bayesian approach to T&E, this Best Practice focuses on three main benefits: potential for better estimation, interpretability, and flexibility.

***Benefit 1:* Bayesian methods can obtain more precise estimates of system performance than classical methods.**

When all relevant information is not included in an analysis, it can leave testers spending limited and expensive resources to capture data that might be unnecessary, ultimately resulting in allocating resources in a sub-optimal manner or in having insufficient data at the end of the test. Bayesian methods are ideal for scenarios in which there is insufficient information available in a test, as all information thought to be relevant can be incorporated into the analysis. By incorporating additional information into the analysis, the standard deviation for model parameter estimates can be improved, making it possible for conclusions to be made with

greater certainty compared to the current approach. This can especially be true when testing has a limited sample size—a scenario that is common in DOD testing, due to cost and time constraints.

Frequentist methods only utilize observed data in accordance with the selected data model (through the likelihood function) to make inferences. This omission of any additional information, even that which may be obvious like "the gas mileage of my car must be greater than zero but less than 100 mpg," may not have much impact when the sample is very large. In contrast, when the sample size is not large, improbable (but possible due to random chance!) samples can potentially lead to inaccurate and even physically impossible conclusions. Furthermore, for a given test design, the precision of the final inference using frequentist methods can only be improved by using larger samples. These effects are amplified when samples are very small, a scenario often encountered when each test point is costly. In contrast, Bayesian methods can mitigate these problems with well-constructed priors.

<u>Benefit 2:</u> Bayesian methods result in easier interpretation for decision makers.

Because Bayesian analysis naturally leads to intuitive interpretations, results are easier to understand and communicate to non-statistician audiences. Specifically, the Bayesian approach enables testers to explicitly report the probability of a system obtaining the desired outcome (e.g., exceeding a requirement) by using posterior probability. This interpretability is in direct contrast to the frequentist view which results in indirect measures of system performance with more esoteric definitions, such as p-values or confidence intervals.

Consider testing a system where the interest is in determining if the time to send an email is less than 15 seconds. Let θ represent the time to send an email. From a frequentist approach, an analyst may consider the following hypothesis test: $H_0: \theta \geq 15$ versus $H_a: \theta < 15$, where the intent is to reject H_0 if the p-value is small enough (typically below 0.05). The interpretation of the p-value is the probability of obtaining a result as extreme or more extreme than the results obtained, assuming $\theta \geq 15$ is true—which is not only an indirect probability statement about the system's performance, but is also a non-intuitive statement for decision makers. Instead, decision makers are interested in whether that the system can send an email in under 15 seconds (a direct probability). Under a Bayesian framework, testers evaluate the direct probability that $\theta < 15$, a statement that decision makers can intuitively understand and is more information than a p-value. The probability can be evaluated in terms of risk that the decision maker is willing to take on, rather than evaluating how extreme a test result is.

Furthermore, Bayesian credible intervals (also known as posterior intervals) are the interpretation that decision makers are more intuitively able to understand. A 95% credible interval is interpreted as: the probability is 0.95 (i.e., 95%) that the parameter of interest exists within the interval. This is in contrast to the convoluted interpretation of a confidence interval: when we construct a confidence interval using the same procedures and methods, there is a 95% chance the true value of the parameter of interest will be contained in the interval. This has been interpreted as: if we were to run the test thousands of times and calculate a confidence interval for each time, approximately 95% of the confidence intervals would contain the true value (Meeker, 2007). However, once a confidence interval is calculated from a data set, the true value is either in the interval or it is not in the interval—which is unknown to the analyst. Moreover, since a confidence interval is not a based on a probability distribution, there is no way to know which values in the interval are most probable. In contrast, a credible interval will provide a wealth of information about the parameter of interest, including which values are probable, improbable, or most likely to be observed or even most likely.

Providing both posterior probability and a credible interval answers two critical questions decision makers want answered, in a more intuitive and understandable way than current methods allow for: “how likely is it that the system can perform?” and “what is the variability (risk) in this assessment?” In addition to the intuitive interpretations, Bayesian methods are also more geared to decision-making problems than traditional methods. Traditional hypothesis testing revolves around “proving” a hypothesis (e.g., making the statement “the system meets requirements” or not), while Bayesian methods focus on making decisions (e.g., making statements about how well the system performing and how likely is it to obtain specifications).

<u>Benefit 3:</u> Flexibility in the Bayesian approach allows for the potential to terminate testing early, as understanding of a system is continually updated.

Under the Bayesian framework, every data point collected is an update to a prior belief. Therefore, at any point in test, interim results can become final results if enough information has been collected to adequately understand the system under test. This not only facilitates cost savings (terminating early and saving costly test resources), but also demonstrates the built-in sequential testing approach of Bayesian methods.

A test that is conducted under a Bayesian approach starts with a prior. The prior is then updated with incoming data (either a complete test or updated at some point during test), which results in a posterior distribution that reflects the updated understanding of system performance. This posterior distribution then can become a prior for the next set of data to be obtained (or modified, if appropriate), which will be updated by the data and result in a new posterior distribution. This approach to testing leverages not only current data but also past relevant data to provide the testers with continually updated information about the system under test. Furthermore, as alluded to, Bayesian methods provide the flexibility to change the type of prior being used across the continuum of testing. This provides testers with a flexible approach to testing that allows for the potential to end testing early, while also making the most use of previous relevant data in a flexible and appropriate way.

Discussion: The Importance of Careful Prior Construction

The posterior distribution depends on both the prior(s) and the data model through the likelihood function. Conceptually, the results of a Bayesian analysis may be thought of as a weighted combination of the prior information and the data information. This is a benefit of Bayesian analysis, as small sample sizes may be augmented by additional information, to ultimately make more precise conclusions about system performance. However, when the prior is too strong relative to the amount of data being collected (i.e., a narrow distribution indicating a strong understanding of the parameters), the choice of prior may have a large influence on the test conclusions. It is important to note that as the sample size increases, the data will overwhelm the prior, leading to results that are less sensitive to the choice of the prior. Practically, however, the cost of increasing the sample size must be weighed against the risk of making decisions that depend too heavily on prior information (S Berry, *et al.* 2011). Therefore, it is important that priors are carefully constructed to benefit the analysis by capturing the appropriate amount of information for the test at hand, not detract from it. Not only does this highlight the importance of sample sizes and understanding how informative the selected priors are, but it also highlights the importance of sensitivity analysis—evaluating the effect that changing priors has on conclusions. How priors are developed, a critical component of a Bayesian analysis, is a driver of many of the questions in the next section.

Key Questions to Ask Testers

In order for Bayesian methods to be effective, efficient, and appropriate (e.g., avoid inadvertently biasing results), it is imperative that priors are developed by a team that includes both an expert in Bayesian prior development and an expert in/on the system. To this end, a few starting questions are offered here for leadership to explore whether a Bayesian approach has been appropriately applied. These questions largely focus on prior development, because the prior is unique to Bayesian analysis while the data model and likelihood function is common to both frequentist and Bayesian methods.

Question 1: Did you select a team member, or seek outside consultation from someone, who is an expert in Bayesian analysis during the development of priors and the execution of analysis?

The answer to this question is the first line of defense for test leadership in ensuring appropriate statistical methods were applied. If the answer is "yes," it is recommended that the remaining questions be asked. If "no", it is recommended that no more questions are asked, and external consultation is brought in with expertise in the area—as would be the case for any analytical technique proposed, traditional or Bayesian.

Question 2: When were the priors developed?

As discussed previously, priors are an *a priori* representation of beliefs. Therefore, they must be created by experts before collecting (or at least seeing) the data from the experiment. It is recommended that test teams have documentation that all stakeholders represented on the test team agreed to the priors before exposure to the test data. If priors are developed after seeing the data, serious issues may arise ranging from inadvertent introduction of bias to the more damaging perception of "data snooping" (changing the prior to get the desired results).

Question 3: What information was used to build the prior, and how does the prior account for any (possible) differences between the information in the prior and the data (to be) collected?

Recall that there are different types of information that can be built into a prior. It is important to understand what data is being used in a prior. For instance, if it was SME opinion, was enough variability incorporated into the prior to account for any potential biases? Were multiple SMEs consulted, and information appropriately combined? If data was used, did it come from the same population (i.e., was it data that would have been used to directly answer a question of interest)? If the data did not come from the same population (e.g., perhaps the system under test has changed since the last time test was conducted), were differences appropriately accounted for? Of note, if the data comes from a different population (e.g., system changes occurred), there are priors that can be used to appropriately account for this, but a SME in Bayesian methods should be consulted to ensure the previous (potentially dissimilar) data is being appropriately accounted for.

Question 4: How informative is the prior relative to the data?

It is important to understand how informative priors are, relative to the amount of information to be obtained from the test. If the prior is too informative, it will be difficult for the data to overwhelm the prior, and inferences will be made mainly based on the prior. Therefore, a balance should be struck between a prior that is not too informative (i.e., can still be updated by the data), while still being informative enough to provide the benefit of Bayesian analysis. The

relative influence of the prior and the data is particularly important in tests with small sample sizes, where it should be investigated during the test planning/sizing phase

Question 5: What priors were used in sensitivity analysis, and how did that affect conclusions and recommendations?

As mentioned at the beginning of this section, sensitivity analysis is an important part of the Bayesian analysis. By using different priors to determine the impact of the prior on the conclusions made, decision makers can obtain an understanding of the risk in an assessment. To conduct sensitivity analysis, it is recommended that an informative prior and a non-informative prior are used at a minimum (note: the terms informative and non-informative are often discussed in Bayesian statistics, which is beyond the scope of this paper). Should using these two priors result in the same inference, decision makers can be assured in their decision to use an informative prior to obtain better estimates of system performance. However, if using different priors results in different decisions, decision makers can conclude that not enough data has been collected (i.e., the decision is being overly influenced by the prior). The decision maker can then decide if the question is important enough that more data is needed; or, if restricted by cost or time, the decision maker will gain an understanding in the risk associated with the decision.

Conclusion

This Best Practice has provided a high-level summary for leadership discussing how Bayesian methods can be used to support their testing strategies. Benefits include the potential for better estimation, interpretability, and flexibility in testing. In addition to the benefits it has also been acknowledged that there are areas that require careful considerations, such as prior development. To mitigate potential risks associated with incorrectly applying a Bayesian approach, it is recommended that a SME be identified, either within the test team or an external consultant such as one from the STAT COE—just as would be recommended for any statistical approach, whether traditional or Bayesian. When understood and properly applied, Bayesian methods add a powerful tool to the T&E professional's toolbox.

References

Berry, Donald A. “A Case for Bayesianism in Clinical Trials”. In: Statistics in Medicine 12.15-16 (1993), pp. 1377-1393. ISSN: 10970258. doi: 10.1002/sim.4780121504.

Berry, Scott M., Bradley P. Carlin, J. Jack Lee, and Peter Muller. *Bayesian adaptive methods for clinical trials*. Boca Raton, FL: CRC Press, 2011, p. 305. ISBN: 9781439825488.

Christensen, Ronald, Wesley Johnson, Adam Branscum, and Timothy Hanson. *Bayesian Ideas and Data Analysis*. Boca Raton, FL: CRC Press, 2011. ISBN: 978-1-4398-0354-7.

Meeker, William Q., Gerald J. Hahn, and Luis A. Escobar. *Statistical Intervals: a Guide for Practitioners and Researchers (2nd ed).* Wiley, 2007.

Appendix

Additional Resources for Bayesian Methods in DoD Testing

Dewald, Lee, Robert Holcomb, Sam Parry, and Alyson G. Wilson. “A Bayesian Approach to Evaluation of Operational Testing of Land Warfare Systems”. In: Military Operations Research 21.4 (2016), pp. 23-32. doi: 10.5711/1082598321423.

Dickinson, Rebecca M., Laura J. Freeman, Bruce A. Simpson, and Alyson G. Wilson. “Statistical Methods for Combining Information: Stryker Family of Vehicles Reliability Case Study”. In: Journal of Quality Technology 47.4 (2015), pp. 400-415.

National Research Council. Improved Operational Test and Evaluation Methods of Combining Test Information for the Stryker Family of Vehicles and Related Army Systems: Phase II Report. Washington, D.C.: National Academy Press, 2004.

National Research Council. Statistical Issues in Defense Analysis and Testing : Summary of a Workshop. Ed. by John E Rolph and Duane L Ste ey. Washington, D.C.: National Academy Press, 1994. doi: 10.17226/9686.

National Research Council. Statistics, Testing, and Defense Acquisition. : New Approaches and Methodological Improvements. Ed. by Michael L Cohen, John E Rolph, and Duane L Steffey. Washington, D.C.: National Academy Press, 1998.

Sieck, Victoria R.C. and Fletcher G.W. Christensen. “A framework for improving the efficiency of operational testing through Bayesian adaptive design”. In: Quality and Reliability Engineering International (2021). ISSN: 10991638. doi: 10.1002/qre.2802.