SCIENTIFIC TEST & ANALYSIS TECHNIQUES
CENTER OF EXCELLENCE

Assessing if a System Meets a Requirement Using Bayesian Inference

May 2023

Dr. James Theimer, Ctr

Distribution Statement A. Approved for public release; distribution unlimited.
Case Number 88ABW-2023-0543; Cleared 22 May 2023.

To develop and apply independent, tailored Scientific Test & Analysis Techniques solutions to Test and Evaluation that deliver insight to inform better decisions.

About this Publication:
This work was conducted by the Scientific Test & Analysis Techniques Center of Excellence under contract FA8075-18-D-0002, Task FA8075-21-F-0074.

For more information:
Visit, www.AFIT.edu/STAT
Email, AFIT.ENS.STATCOE@us.af.mil
Call, 937-255-3636 x4736

Technical Reviewers:
Dr. Cory Natoli, Ctr
Maj Victoria Sieck, PhD
Mr. Corey Thrush, Ctr

Copyright Notice: No Rights Reserved
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, FY23

Table of Contents

Introduction

Programs conduct testing to assess whether requirements are met; since Bayesian statistics views probability as describing the belief that a statement is true, Bayesian methods offer natural ways of making decisions that programs must make. Frequentist statistics, which are typically used, set a significance level prior to data collection, and offer techniques that control the probability that a null hypothesis will be incorrectly rejected, to this level of significance, when data have been collected. As stated by Kruschke and Liddell (2018; p.192), "frequentist hypothesis testing can only reject or fail to reject a particular hypothesis. It can never show evidence in favor of a hypothesis." Programs frequently ask about the evidence in favor of a hypothesis, and Bayesian methods offer a way of doing this. This best practice will show how to use Bayesian methods to assess if requirements have been met in a common example situation, where the data acquired are only whether the system succeeded in a function or not.

Background

This document will introduce Bayesian methods to assess the plausibility of a hypothesis. It will discuss how to use the outputs of a Bayesian analysis tool produced by the Homeland Security Community of Best Practices (HS COBP) to make decisions about whether requirements have been shown to have been met in the light of test results. The document assumes that the reader has a basic knowledge of statistics, such as is covered in the HS COBP STAT workshop, which is training available to Department of Homeland security employees as a three-day short course. Many readers will also have taken a semester course in statistics as a part of their education. This Best Practice will use some of the general concepts that they learned in that course. It is also assumed that the reader has some basic knowledge of Bayesian statistics, such as the steps in fitting a Bayesian model to data.

A frequently encountered situation is one in which a program wishes to assess the probability of a system working. The data are records of either success (the system works) or failure (the system does not work). This is referred to as a binomial response. Because the HS COBP must help programs with this kind of analysis, it will be used to illustrate ideas about Bayesian hypothesis testing. This paper will assume that the team needs to show that a system has a probability of success greater than some required value, so methods will be presented for doing this. Prior to conducting the test, the program will need to decide what amount of risk can be accepted of failing to reject a system that does not meet the requirement. The risk can be treated as a threshold for subjective probability. Bayesian methods will be used to assess if the subjective probability of the system not meeting the requirement is acceptable. Occasionally, programs are concerned about the possibility of incorrectly rejecting a system that meets a requirement, so this risk will also be mentioned.

When hypotheses are tested with frequentist methods, they are evaluated by two types of possible errors, which are related to two hypotheses. As this paper will describe later, Bayesians can also quantify the risk, but they do so differently. The frequentist selects a null hypothesis, which is assumed true until it is disproved. The null hypothesis can be rejected when it is actually true, which is called Type I error, and tests are designed so that the probability of incorrectly rejecting the null is α. The alternative hypothesis is some statement that is only accepted if the null hypothesis is rejected, which is when the calculated p-value is less than α. We can also fail to reject the null hypothesis when the null hypothesis is false, which is called Type II error, which has a probability expressed as β. Hypothesis tests are discussed in Kensler & Freeman (2022), Hogg, McKean, & Craig (2013), and Montgomery & Runger (2014). The design of binomial tests is covered in Scientific Test and Analysis Techniques Center of

Excellence (STAT COE) Best Practices by Truett (2022), Ortiz (2022), and Burke, Key, & Wurscher (2022). It is also discussed in Montgomery & Runger (2014). After testing, a frequentist will say that the null hypothesis has been rejected, or not, at a given significance level, with a test designed with a given power based on an estimate of what will be observed.

Program decision makers frequently wish to talk about the probability of rejecting a system that meets requirements or accepting one that does not, in the light of the data collected in testing. From a frequentist point of view, this is incorrect. For a frequentist, the hypothesis that the system does not meet a requirement either has been rejected, or we have failed to reject it. The probability of Type I error has been set and the Type II error has been assumed before the data collections, to set the criterion for rejecting the null hypothesis. The Bayesian, however, can talk about the probability that the system is acceptable or not, given the data observed. This is because the Bayesian thinks of probability as describing the belief that the estimated parameter takes on a given value. In the case discussed here, the parameter being estimated is the probability of success. The risk that the system is not acceptable is quantified by the posterior probability that the probability of success is less than the threshold. The belief that the system meets the requirement is measured by the subjective probability of the probability of success being greater than the threshold. The topic is discussed in much more detail, in Christensen et al. (2011), Gelman et al. (2013), and Reich & Ghosh (2019).

The data we will work with have a binomial response so we will need a Bayesian model for this type of data. The reader may recall that we will use Bayes' rule as a way of going from our prior knowledge of the probability distribution of the parameter and a likelihood function that describes a model of the observed variable given a fixed value of the parameter, to a posterior distribution function of the parameter given a data set. The posterior distribution can be computed based on the prior distribution and observed data by

$$p(\theta|y) = \frac{L(y|\theta)p(\theta)}{\int L(y|\theta)p(\theta)d\theta}, \tag{1}$$

where $p(\theta|y)$ is the posterior distribution, L is the likelihood function of the data collected, $p(\theta)$ is our prior belief or distribution, and the denominator is the marginal PDF of observing y successes. The probability of success is the parameter represented by θ. It turns out that a convenient prior is the Beta distribution,

$$p(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}, 0 < \theta < 1. \tag{2}$$

The parameters of the model are represented by a and b, and Γ is the Gamma function. The number of observations is given by n. The likelihood function is the Binomial distribution,

$$L(y|\theta) = \binom{n}{y}\theta^{y}(1-\theta)^{n-y}. \tag{3}$$

The posterior distribution is another Beta distribution, so the distributions are a conjugate family. The posterior is

$$p(\theta|y) = \frac{\Gamma(n+a+b)}{\Gamma(y+a)\Gamma(n-y+b)}\theta^{y+a-1}(1-\theta)^{n-y+b-1}, 0 < \theta < 1. \tag{4}$$

The posterior is a Beta distribution with parameters, a' and b', where

$$a' = a + y, \tag{5}$$

and

$$b' = n - y + b. \tag{6}$$

This is the model used in the spreadsheet discussed below, and was selected because, since it is a conjugate family of distributions, it has a closed form solution.

Bayesian hypothesis testing, when the hypothesis is simply testing that the parameter is above

or below a threshold level, can be done directly by computing the subjective probabilities of these events. This method of testing the hypothesis is mentioned by Christensen, Johnson, Branscum, & Hanson (2011) on pg. 56. Risks in Bayesian methods are quantified by the posterior probability above or below the required probability of success, and frequentists quantify risk with confidence levels. It seems plausible to use the value of $1 - \alpha$ as the value of required posterior probability, as they both quantify risk. Since the risks about which the team is concerned can be thought of as posterior probabilities, the Bayesian method offers a direct way of measuring the risk. This point is discussed in Shi & Yin (2021). The Excel spreadsheet that will be discussed later in this paper computes this posterior probability.

Some programs are concerned about both the possibility of rejecting a good system and of accepting a bad one, and this is why both have been thought of as risks. Figure 1 shows an example where the subjective probability of the variable falling below the threshold is very small. In this case, if the system was accepted there would be a very low risk of accepting a bad system. Alternatively, if the program rejects the system, there would be a very high risk of rejecting a good one. Figure 2 shows an example with a high subjective probability that the variable falls below the threshold. If the system is accepted, there is a high risk that it is bad. If it is rejected, there is a low risk it is good. The risk levels the program will accept need to be determined by program leadership. It is possible that the test is inconclusive. In this case only more testing can prove that the system, in its current state, passes.

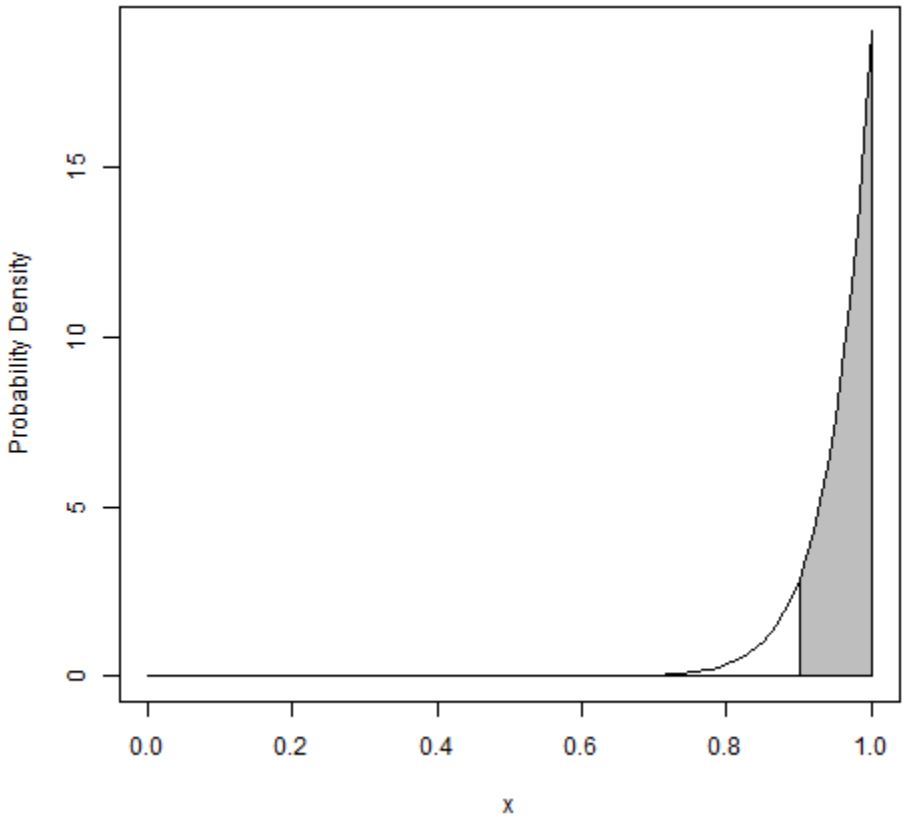

Figure 1
PDF of a Beta distribution with a = 19 and b = 1

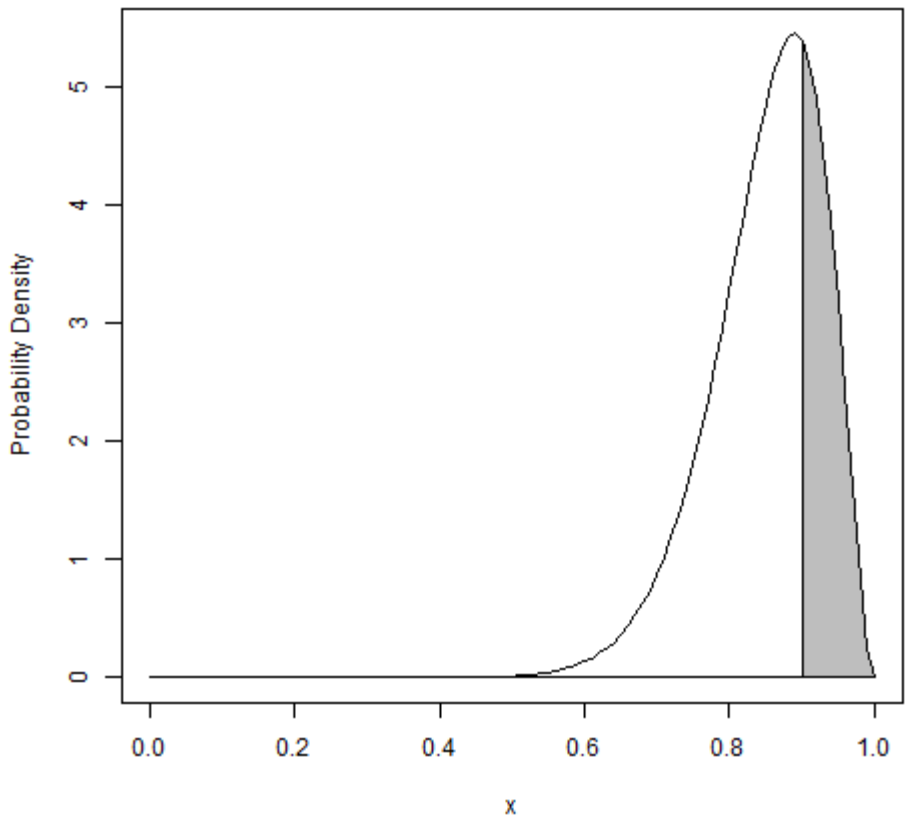

Figure 2
PDF of a Beta distribution a = 17 and b =3.

A second way to determine the plausibility of the hypothesis that the system meets the requirement, compared to the hypothesis that it does not, is to compute the Bayes Factor. Specifically, the Bayes Factor measures how much beliefs about the relative plausibility of two hypotheses are changed by data. In the example being given here, the team might test to see if the hypothesis that the probability of success is above some level is more plausible than the alternative. It will be used for this purpose in the spreadsheet described. Bayes Factor (BF) is given by Reich & Ghosh (2019),

$$BF = \frac{f(M_2|\boldsymbol{y})/f(M_1|\boldsymbol{y})}{f(M_2)/f(M_1)}. \tag{7}$$

The variables M_1, and M_2, are the two hypotheses. The data collected are represented by **y.** The variable y is shown in boldface to indicate that it may be a vector, though it is not in the case examined in this document. If p_0 is the requirement threshold for the subjective probability of the probability of success, M_1 is $\theta \geq p_0$ and M_2 is $\theta \leq p_0$. The numerator is the ratio of the probabilities of the two hypotheses given that we have collected data, which is to say as estimated by the posterior, and the denominator is the ratio of the probabilities of the models under the prior distribution. A large value would indicate that our belief in M_2 relative to M_1, should change significantly due to the observed data. If a Bayes Factor value of 10 or greater is observed, one may say that there is strong evidence to prefer M_2 (Kass & Raftery, 1995). Further evidence thresholds are given in Table 1.

Table 1

Description of the strength of evidence against M_1 taken from Kass & Raftery (1995)

Bayes Factor	Evidence against M_1
1 to 3.2	Not more than a bare mention
3.2 to 10	Substantial
10 to 100	Strong
>100	Decisive

Use of the Tool

Inputs Related to Prior Distribution

The HS COBP has developed an Excel spreadsheet for carrying out the calculations for the Bayesian posterior probability of meeting the requirement and the Bayes Factor. This may be obtained by U.S. Government agencies and their contractors by emailing afit.ens.hscobp@us.af.mil. This document will describe the information that needs to be entered and then will look at the outputs. The inputs will include the parameters of the prior distribution, and those that describe the data collected. The section of the spreadsheet which describes the prior and posterior distributions is shown in Figure 3. The entries shown on the column labeled "Prior" are the parameters of a Beta distribution where both parameters are 1. This is an example of a reference, or noninformative, prior and gives a uniform prior probability. The probability is equally likely that the parameter is any value between 0 and 1. This prior is easily overwhelmed by data. If prior knowledge suggested the system would most likely work in a test, the team might choose a prior distribution with a mean value near 1 instead. The sum of the a and b parameters act as if one has a prior number of tests equal to that value. For example, in Figure 1, the reference prior can be interpreted as having seen two tests prior to seeing the current test—one that was a success and one that was a failure. Selecting the sum of a and b sets how much the prior knowledge will be weighted relative to the data. Details of selecting priors are given in Christensen, Johnson, Branscum, & Hanson (2011) Section 5.1. The number of successes and the number of trials, entered as shown in Figure 4, are used in Equations 5 and 6 to compute the updated values of the Beta distribution parameters. These appear in the "Posterior" column of Figure 3.

Prior Distribution		
Beta dist parameters	**Prior**	**Posterior**
beta param 1	1	98
beta param 2	1	4

Figure 3

Input of prior distribution parameters and output of posterior distribution parameters

Input Related to Data Collected

A description of the data collected is shown in Figure 4. The "Requirement, p_req" box is the entry of the threshold for the probability of success that must be exceeded to meet the requirement. In this case the probability of success must be at least 95%. "# Trials, N" shows that there were 100 trials, and "# Successes, X" shows that 97 were successes.

INPUT DATA		
Requirement, p_req	0.95	*For example, required probability of detection*
# Trials, N	100	
# Successes, X	97	

Figure 4
Input requirement to be satisfied and data collected

Plot of Distributions

The spreadsheet plots the Probability Density Functions (PDF) of the prior and posterior distributions, as shown in Figure 5. The requirement is shown by a red line. The posterior probability that the requirement is met is the area under the posterior PDF curve for probability of success values greater than the required value.

Calculations giving Posterior Probability of the Requirement being Met

The estimated posterior probability that the requirement is met is given in Figure 6. In this case, the posterior probability that the requirement is met is about 75%. The posterior probability is the updated subjective probability of meeting the requirement when the prior probability distribution has been updated due to the observed data using Bayes' rule. This means that there is a 25% posterior probability that the probability of success is less than the requirement. As argued above, frequentist confidence and Bayesian subjective probability both quantify risk and so it would seem reasonable to use the same numerical value for both. For instance, if the specification states that the requirement must be met with 80% confidence one would suggest using that threshold for the posterior probability requirement. Therefore, for the situation in Figure 6, we would reject the system as the subjective probability is less than 80%.

Bayes Factor

The computed value of the Bayes Factor is shown in Figure 7. The spreadsheet inserts a text description based on values in Table 1. In this case we conclude that the data would strongly lead us to update our belief to accept that the system meets the requirement.

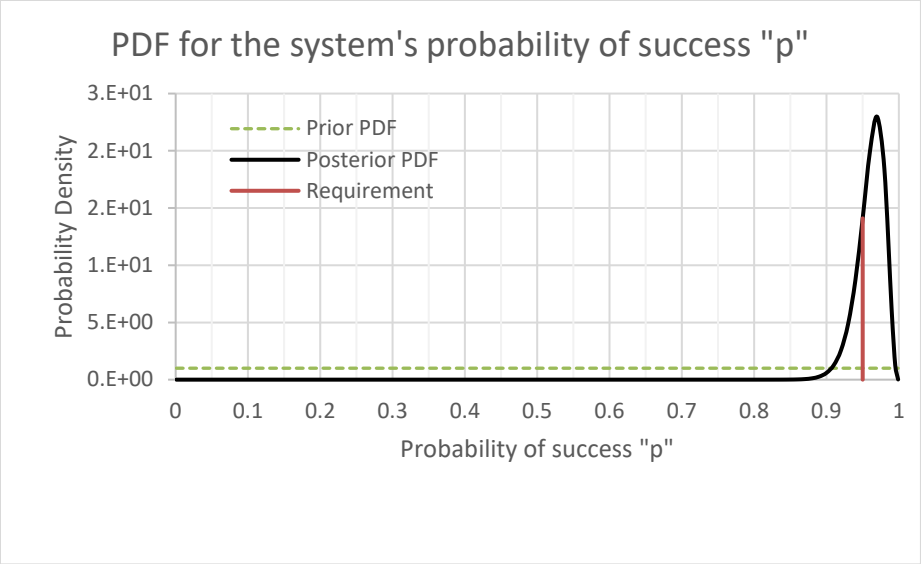

Figure 5
PDF of prior and posterior distributions

	CALCULATIONS		
	Expected p (mean)	0.961	
	prob(p > p_req)	74.9%	*"chance that the prob. of success exceeds the requirement"*

Figure 6
Computed estimated probability of success and posterior probability that probability of success is greater than the requirement

Bayes Factor Results	
Strength of evidence for p > p_req:	
56.73949394	***Strong***

Source: See Best Practice

Figure 7
Bayes Factor with interpretation

Application of Spreadsheet to Examples

Table 2 shows what happens when the prior or requirement is changed. The results are discussed to show how one would reason based on the results. One prior is the reference prior used above, and the other is a weak prior that shows some confidence that the system will work, but not perfectly. In all cases it is assumed that there needs to be at least a posterior probability of 0.8 that the requirement is met for the system to pass. The data describing the experiment are those shown in Figure 4. If the required probability of success is at least 0.95,

we would reject the system if we used a reference prior. The system would pass with the weak prior. If the required probability of success is at least 0.9, the system passes for either prior. This required posterior probability should be established before testing to ensure objective consideration. The Bayes Factor scores are lower when using the weak prior. The Bayes Factor indicates how much the data changes our mind about the relative probabilities of the priors. In the case of the weakly informative prior we have some belief that the system will work, so the data changes our belief less than if we used the reference prior. The Bayes Factor is less for the higher required probability of success. If it takes more data to demonstrate a high probability of success, the data observed would change our mind less than in the case where we require less data to be convinced, due to the lower required probability of success.

Table 2
Examples of Bayes Factor for different situations

Required Probability of Success	0.9	0.95	0.9	0.95
a in prior	1	1	9	9
b in prior	1	1	1	1
Probability that Requirement Met	0.993	0.749	0.996	0.8
Bayes Factor	1232.84	56.74	163.81	6.81

Conclusion

This Best Practice has demonstrated the use of Bayesian methods to assess if a system has met requirements. The situation considered is the requirement that the probability of success is greater than some level, which is a commonly encountered sort of requirement. The posterior probability directly gives a measure of the combined evidence that we should believe that the requirement has been met. The Bayes Factor provides a measure of the degree to which the evidence has caused us to update our belief that the requirement has been met. The Best Practice has also demonstrated the inputs and outputs to an Excel spreadsheet for performing these calculations, available by request from the HS COBP at afit.ens.hscobp@us.af.mil.

References

Burke, S., Key, M., & Wurscher, K. (2022). *Categorical Data in a Designed Experiment Part 4: Estimating Power of Test Designs for a Binary Response.* Retrieved from AFIT/STAT Center of Excellence: https://www.afit.edu/stat/statcoe_files/Categorical%20Data%20in%20a%20Designed%20Experiment%20Part%204%20-%20Estimating%20Power%20of%20Test%20Designs%20for%20a%20Binary%20Response_Best%20Practice.pdf

Christensen, R., Branscum, A., Hanson, T. E., & Johnson, W. (2010). *Bayesian ideas and data analysis: an introduction for scientists and statisticians*. CRC Press, an imprint of Taylor and Francis. https://doi.org/10.1201/9781439894798

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.

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association, 90*(430), 773-795. https://doi.org/10.1080/01621459.1995.10476572

Kensler, J., & Freeman, L. (2022). *Statistical Hypothesis Testing*. Retrieved from AFIT/STAT Center of Excellence: https://www.afit.edu/stat/statcoe_files/Statistical%20Hypothesis%20Testing.pdf

Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin and Review*, **25**, 178–206. https://doi.org/10.3758/s13423-016-1221-4

Montgomery, D. C., & Runger, G. C. (2014). *Applied Statistics and Probability for Engineers* (6th ed.). Hoboken, NJ: Wiley.

Ortiz, F. (2022). *Categorical Data in a Designed Experiment Part 2: Sizing with a Binary Response.* Retrieved from AFIT/STAT Center of Excellence: https://www.afit.edu/stat/statcoe_files/Categorical%20Data%20in%20a%20Designed%20Experiment%20Part%202%20Sizing%20with%20a%20Binary%20Response.pdf

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

Shi, H., & Yin, G. (2021). Reconnecting p-Value and Posterior Probability Under One- and Two-Sided Tests. *The American Statistician*, 75(3), 265-275. https://doi.org/10.1080/00031305.2020.1717621

Truett, L. (2022). *Using Operating Characteristic (OC) Curves to Balance Cost and Risk.* Retrieved from AFIT/STAT Center of Excellence: https://www.afit.edu/stat/statcoe_files/Using%20Operating%20Characteristic%20Curves%20to%20Balance%20Cost%20and%20Risk%20V2.pdf