A Decision Theoretic Framework for Hypothesis and Significance Testing

Tyler Bonnett

A thesis submitted in partial fulfillment

of the requirements for the degree of

Master of Science

University of Washington

2018

Committee:

Ken Rice

Lurdes Inoue

Program Authorized to Offer Degree:

Biostatistics

University of Washington

Abstract

A Decision Theoretic Framework for Hypothesis and Significance Testing

Tyler Bonnett

Chair of Supervisory Committee:
Ken Rice, Professor
Department of Biostatistics

From its inception, statistical testing has been a controversial area. There are several philosophies of testing and inference, the most common among them being the so-called frequentist and Bayesian approaches. These approaches have often been viewed as at odds with one another. In this paper, we suggest that in many common testing scenarios this is not the case. We will approach testing from a decision theoretic standpoint, framing testing and inference as decisions to be made about a parameter. In doing so, we show that the commonly used methods of testing and inference answer different questions but can both provide valuable knowledge. We aim to help researchers move away from the viewpoint that one must be either a "frequentist" or a "Bayesian", as statisticians have often divided themselves in the past, and toward the recognition that both schools of thought can make relevant contributions to their research.

List of Figures

List of Tables

Contents

1 Introduction - History and Related Work

We begin with a brief review of the development of statistical testing from various perspectives as well as a review of some relevant literature which serves to motivate the methods we develop later. Readers who are already familiar with this body of literature should feel free to skip this section and move directly to Section 2. However, this section does contain a number of definitions which will be utilized throughout the work and which may be helpful to review for all readers before moving on to later sections.

1.1 Historical Foundations of Frequentist Statistical Testing

The problem of statistical testing and the controversies that have arisen as a result are typically said to date back to the 1920s, although we will see that the problems involved are much older. The setting initially seems straightforward: we are presented with a scientific hypothesis and wish to make a quantitative statement about what the available evidence tells us regarding the truth of that hypothesis. The proper way to go about developing and ultimately phrasing this quantitative statement, however, has been a hotly debated issue since its inception.

1.1.1 Significance Testing

Ronald Fisher originally developed what came to be known as the *significance testing* approach to this problem. In this approach one only considers testing a single hypothesis, deemed the null hypothesis. By convention, the null hypothesis is denoted H_0. Here, the term "null" refers to the fact that many tests are carried out to determine whether some effect (say, the effect of a new medical treatment) is different from zero (a "null" effect). The null hypothesis can, however, be framed differently. At the most basic level, the null hypothesis simply represents the current state of scientific knowledge. To begin to address the truth or falsehood of this hypothesis, Fisher is often credited with the development of the p-value—the probability, computed under the assumption that the null hypothesis is true, of obtaining an effect equal to or more extreme than the observed effect [1].

Consider a scenario in which we wish to test the null hypothesis that a new pharmaceutical treatment

intended to reduce blood pressure has some true, unknown effect which we will call θ. We could imagine that θ represents the average true reduction in systolic blood pressure in millimeters of mercury (mmHg) among patients who use the treatment. We will assume that, prior to experimentation, we do not know whether the treatment has any effect and therefore state the null hypothesis $H_0 : \theta = 0$. We then collect data, which we will call X, most likely by giving the treatment to a sample of patients and measuring their change in systolic blood pressure after some predetermined time period. We then compute a test statistic $T(X)$ (such as the sample mean or median) which succinctly summarizes the data and has a known sample distribution. For example, if we choose to summarize our data by computing the sample mean, then according to the central limit theorem the sample distribution of this test statistic is asymptotically normal (with the approximation to normality "kicking in" beyond a sample size of 30) [2]. Much work has been done to characterize the sample distribution of the many test statistics which could feasibly be used to summarize data in statistical problems. Once we have observed the test statistic, knowing these sample distributions allows us to define the probability (assuming the null hypothesis is true) of observing values of the test statistic as extreme or more extreme than what was actually observed. This probability is the p-value. In the example described above, the p-value addresses the question: If θ is actually 0 (if the null hypothesis is true) and the sample mean we have computed is $T(x) = \bar{X}$, what is the probability of having observed a value of $T(x)$ as extreme or more extreme than $\bar{X}$? Mathematically, the p-value is the probability $P_{\theta=0}(T(X) \geq T(x))$, where the subscript notation $\theta = 0$ reminds us that the computation assumes the truth of the null hypothesis. This approach to measuring evidence is intuitive to many. In our example, if we assume that the true effect of the new drug is no reduction in systolic blood pressure and yet we observe an average reduction in systolic blood pressure among those taking the drug of 10mmHg (with a relatively small sample standard deviation of, say, 3mmHg), then the evidence intuitively seems to contradict the null hypothesis. That is, we can intuit that if the null hypothesis is true then the probability of getting an average reduction in systolic blood pressure of 10mmHg or more is probably small. Another way of phrasing this statement is that the p-value (which we will not actually compute for this example) would probably be small. By this logic, small p-values are thus considered evidence which contradicts the assumption of the null hypothesis because they indicate that, if the null hypothesis *is* true, then we have observed an event which would be considered quite rare.

There is, however, a question of how small a p-value must be in order for the researcher to feel confident in declaring that the null hypothesis is not true (i.e. "rejecting the null"). This value is known as the *significance threshold*. In most cases in practice, the significance threshold is chosen to be 0.05. In the Fisherian view of statistics, results which correspond to p-values below the significance threshold are considered *statistically significant*. That is, if the probability (under the null) of observing a test statistic at least as extreme as $T(x)$ is less than or equal to 0.05, then we say the results are significant at the 0.05 level. The arbitrariness of this threshold for declaring results either significant or non-significant has contributed to much of the soon-to-be-discussed misuse of p-values in applied statistical literature.

Although Fisher is responsible for popularizing the canonical $\alpha = 0.05$ threshold for statistical significance, his credit for developing the p-value itself may not be entirely deserved. In fact, computations of p-values date back at least as far as the 1700s. In 1710, John Arbuthnot reportedly examined birth records in London and roughly computed the probability of his observation that the number of males born exceeded the number of females for every year from 1629 to 1710 under the assumption that the probabilities of giving birth to a male or female were equal. Finding that the probability of an outcome such as this was exceedingly small if the true probabilities were equal, Arbuthnot rejected that notion [3]. Later, in the 1770s, the highly renowned Pierre-Simon Laplace considered the same question. Laplace ultimately addressed the problem by modelling the number of male births as a binomial random variable [4] and also computed a p-value describing the probability of observing a more extreme birth ratio than the actual observation under the null hypothesis that both sexes were equally likely [5]. The p-value as a means of evaluating hypotheses was formally introduced by Karl Pearson around 1900 during the development of Pearson's chi-squared test [5].

1.1.2 Hypothesis Testing

The Fisherian approach to testing is distinct from the *hypothesis testing* approach developed by Jerzy Neyman and Egon Pearson [6]. According to the Neyman-Pearson hypothesis testing approach, one should consider evaluating both the null hypothesis, H_0, as well as an alternative hypothesis denoted H_1. In this view, according to the strength of the evidence, one could either reject the null hypothesis in favor of believing the alternative or, otherwise, fail to reject the null hypothesis. This framework

allows the statistician to compute two important probabilities: the probability of a *Type I error*, which occurs when we falsely reject a true null hypothesis, and the probability of a *Type II error*, which occurs when we fail to reject a null hypothesis that is in fact false. In the Fisherian approach, the probability of a Type I error is the rate of false rejections of a true null hypothesis. Unlike the hypothesis testing approach, however, Fisherian significance testing does not easily incorporate the concept of a Type II error. By convention, the probability of a Type I error, also called the Type I error rate, is denoted α. Also by convention, the Type II error rate is denoted β. Neyman and Pearson proposed that by limiting these error probabilities (which can be accomplished, for example, by decreasing the significance threshold or utilizing large samples) one could guarantee, in the long run, that only a small proportion of the decisions reached by the testing approach would be incorrect. The Neyman-Pearson framework also grants us the notion of the *power* of a test: the probability of appropriately rejecting a null hypothesis when it is in fact false. By definition, the power of a test is equal to $1 - \beta$.

From the beginning, the differences of opinion regarding the proper way of proceeding with statistical testing led to surprisingly bitter disputes between proponents of the two dominating schools of thought. More recently, there is often little distinction made between the significance testing and hypothesis testing approaches and the use of p-values as a measure of evidence against the null hypothesis is widespread in nearly all fields of research [7]. Later, we will develop a framework for *null hypothesis significance testing*—a hybrid approach which incorporates the concepts of Type I and Type II errors but also resembles significance testing in that it allows for making no decision as data dictates. As a whole, these approaches to testing are often labelled the "frequentist" approach, a reference to the fact that it defines probabilities in terms of the long run frequency of events.

1.2 Criticisms of the P-value Approach to Testing

The frequentist approach to testing makes decisions based on p-values. Despite their ubiquitous use in the scientific literature, there are a number of well-documented and common misunderstandings regarding the use and interpretation of p-values [8, 9, 10]. A primary concern among statisticians regarding the use of p-values is the common misinterpretation of the p-value as a probabilistic statement about the truth of the null hypothesis. This is incorrect. P-values do not measure the probability

that a hypothesis is true. To the contrary, they are computed under the assumption that the null hypothesis is true. Nor do p-values represent the probability that data could have been produced by random chance, as is sometimes misinterpreted. Instead, p-values are a statement about data and the likelihood of observing more extreme data under a specified hypothesis. These misunderstandings have been addressed extensively. Many authors have devoted time to explaining what p-values are and are not and the misconceptions surrounding their interpretation by researchers without statistical training (and sometimes, unfortunately, among statisticians as well) [11, 12, 13, 14]. The misuse of p-values has been suggested as one factor in the so-called "replication crisis" in scientific publishing, which has also been extensively addressed [15, 16, 17, 18].

Another concern with the use of p-values is that their use can contribute to a body of literature that is ripe with statistically significant results which, even if they are repeatable, may not be of practical importance. This criticism is often discussed in the context of medicine, where researchers should be encouraged to distinguish between *statistical* significance (indicated by a p-value) and *clinical* significance (indicated, usually, by meaningful effect sizes). It is possible, especially when sample sizes are large, to achieve a highly statistically significant result despite the estimated effect size of a treatment being quite small. In response to this, critics of p-values have pointed out that emphasis on publishing statistically significant results can obscure whether the results are of scientific importance [19, 20].

Furthermore, as we alluded to in the definition of the significance threshold, the frequentist dichotomization of results as either significant or non-significant requires deciding on an arbitrary cutoff for what we consider significant. This issue has prompted researchers to investigate the distribution of p-values reported in studies published in reputable journals. One such study by Masicampo and Lalande found an overabundance of barely significant p-values relative to what would be expected based on the observed distribution of p-values in other ranges, which they suggested might be attributed to publication bias [21]. The dichotomy also obscures the fact that small, statistically *non-significant* changes to outcomes (such as group means or regression coefficients) can be enough to push a result from above to below the threshold for significance. This fact has led some authors to criticize the frequentist philosophy because "the difference between ‘significant’ and ‘not significant’ is itself not statistically significant" [22].

Another notable problem with p-values which we will describe in some detail now which was originally

pointed out in work by Edwards, Lindman, and Savage [23] as well as Dickey [24] and later expanded upon by Berger and Sellke [25] is that, in a certain sense, they can overstate the evidence in favor of the alternative hypothesis. To show this, Berger and Sellke employed methodology from the *Bayesian* philosophy of statistics. We describe their methods here in order to highlight a criticism of p-values, but readers unfamiliar with Bayesian methods should refer to Section 1.3, below, before proceeding. Berger and Sellke derive lower bounds on the probability of the null hypothesis being true implied by p-values. They consider a variety of testing scenarios, the most simple of which involves testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ (a so-called "two-sided" test, since the alternative is the θ could be on either side of θ_0). In this setting, Berger and Sellke define $0 < \pi_0 < 1$ as the prior probability of H_0 and $\pi_1 = 1 - \pi_0$ as the prior probability of H_1. They further let $g(\theta)$ denote the *prior probability distribution* of θ given that H_1 is true. Using this notation, the *marginal density* of X is given by

$$m(x) = f(x|\theta_0)\pi_0 + (1 - \pi_0)m_g(x),$$

where

$$m_g(x) = \int f(x|\theta)g(\theta)d\theta.$$

Then the *posterior probability* of H_0 is

$$\begin{aligned} P(H_0|x) &= f(x|\theta_0) \times \frac{\pi_0}{m(x)} \\ &= \left[1 + \frac{1 - \pi_0}{\pi_0} \times \frac{m_g(x)}{f(x|\theta_0)}\right]^{-1}, \end{aligned}$$

and the *posterior odds ratio* of H_0 to H_1 is

$$\begin{aligned} \frac{P(H_0|x)}{P(H_1|x)} &= \frac{\pi_0}{1 - \pi_0} \times \frac{f(x|\theta_0)}{m_g(x)} \\ &= \text{Prior odds ratio} \times \text{Bayes factor for } H_0 \text{ versus } H_1. \end{aligned}$$

Berger and Sellke then derive lower bounds on $P(H_0|x)$ using the *Bayes factor* (a quantity we will discuss in more detail in Section 1.4) $B_g(x) = \frac{f(x|\theta_0)}{m_g(x)}$. Specifically, they show that for G a class of distribution

functions

$$\underline{B}(x, G) = \inf_{g \in G} B_g(x) = \frac{f(x|\theta_0)}{\sup_{g \in G} m_g(x)},$$

where "inf" and "sup" refer to the infimum and supremum of the indicated sets, respectively. Using this derived lower bound on the Bayes factor, Berger and Sellke then point out a lower bound on the posterior probability of the null hypothesis is given by

$$\underline{P}(H_0|x) = \left[1 + \frac{1 - \pi_0}{\pi_0} \times \frac{1}{\underline{B}(x, G)}\right]^{-1}.$$

Finally, Berger and Sellke show that $\underline{B}(x, G) = e^{-t^2/2}$ and

$$\underline{P}(H_0|x) = \left[1 + \frac{1 - \pi_0}{\pi_0} \times e^{t^2/2}\right]^{-1},$$

where t represents the theoretical value of a t distribution [4] corresponding to a particular p-value. Table 1.2.1 shows various two-sided p-values, their corresponding t-values, and the lower bound on the probability of the null hypothesis that the p-value implies when $\pi_0 = \frac{1}{2}$ (so that, initially, the null and alternative are considered equally likely).

Table 1.2.1: Comparison of p-values and $\underline{P}(H_0|x, G)$ as described by Berger and Sellke

p-value (p)	t	$\underline{P}(H_0\|x, G)$	$\underline{P}(H_0\|x, G)/pt$
0.10	1.645	0.205	1.25
0.05	1.960	0.128	1.30
0.01	2.576	0.035	1.36
0.0001	3.291	0.0044	1.35

We see, for example, that a p-value of 0.05 corresponds to a lower bound on the probability of the null hypothesis of 0.128. This demonstrates that improperly assigning a probabilistic interpretation to the p-value leads to exaggerated confidence in conclusions. If we make the common mistake of interpreting the p-value as the probability that the null hypothesis is true, then we would improperly assume that a p-value of 0.05 means that the null hypothesis has only a 5% chance of being true. Berger and Sellke demonstrate that in a very general setting the actual probability that the null is true in this case is at least 12.8%. In other settings the lower bound on this probability can be even higher. When we restrict

attention to the class of normal distributions [4], for instance, the lower bound implied by $p = 0.05$ is 32.1%. These results apply specifically to the case where we are interested in testing a point null hypothesis, but the implication can still be troubling.

There have been proposals that the scientific community should lower the commonly used p-value threshold of 0.05 [26]. While this may aid in reproducibility of significant results (because it would effectively require a higher burden of proof for deeming results significant), critics would point out that a new threshold would not be any less arbitrary than the 0.05 threshold and the approach would still be subject to the previous criticisms outlined in this section.

One final concern with the use of p-values which we will mention here, and which we consider less concerning since a number of solutions to this problem exist, is the issue of multiple testing. Although this topic is discussed thoroughly in introductory statistics courses, it bears repeating that when performing multiple statistical tests it is necessary to adjust the resulting p-values accordingly. For an example of why such a correction is needed, consider a scenario in which we wish to perform n independent tests, each one at significance level α. For an individual test, the probability of making a Type I error is α. But if we consider the overall Type I error rate (also referred to as the familywise error rate, FWER), then we find

$$\begin{aligned}
\mathbb{P}[\text{At least one Type I error in n tests}] &= 1 - \mathbb{P}[\text{No Type I errors in n tests}] \\
&= 1 - (\mathbb{P}[\text{No Type I error in an individual test}])^n \\
&= 1 - (1-\alpha)^n.
\end{aligned}$$

For $n = 20$ and a level of significance $\alpha = 0.05$ for each individual test, the FWER is inflated by the multiple comparisons to 0.642. In other settings, such as genetic association studies where the number of individual tests performed can easily be in the thousands, the multiple comparisons problem is a serious concern if left unaddressed.

One method of adjustment for the multiple comparisons problem is the Bonferroni correction [4], which corrects the overall Type I error rate for n tests by reducing the significance level for each test to $\frac{\alpha}{n}$, where α is the desired FWER [27]. Another popular method of handling this issue is the Benjamini-Hochberg procedure

, which seeks to control the false discovery rate (the expected proportion of errors committed by falsely rejecting null hypotheses, FDR) [28]. We will further elaborate on these procedures for multiple testing adjustments in Section 4, where we will show how the framework we introduce for testing can incorporate these procedures.

1.2.1 Addressing Misunderstandings: The ASA Statement on P-Values

The justified concerns that statisticians have professed about the use and misuse of p-values have prompted much debate about the direction that the field should go regarding testing and measuring the strength of evidence. Recently, one journal banned the use of p-values completely [29]. In 2016, the American Statistical Association (ASA) released a statement addressing the proper use of p-values [30] which outlined many of the same concerns described above and carefully defined best-practices for the use of p-values. The ASA statement concluded with advice for statisticians:

"Good statistical practice, as an essential component of good scientific practice, emphasizes principles of good study design and conduct, a variety of numerical and graphical summaries of data, understanding of the phenomenon under study, interpretation of results in context, complete reporting and proper logical and quantitative understanding of what data summaries mean. No single index should substitute for scientific reasoning." [30]

While it represents an important contribution to the field of scientific research in that it provides a succinct and authoritative guide for the use of p-values, the ASA statement did not seek to provide detailed suggestions for other approaches for addressing the truth of hypotheses. That is, while the statement correctly identifies that no single index should be used to evaluate evidence without context, it does not point us toward any new indices or new methods of contextualizing the weight of evidence.

Importantly, in agreement with the ASA, we do not consider the widespread misinterpretation and misuse of p-values as a reason to scrap this measure altogether. Instead, we will use these misunderstandings as motivation to develop a framework for testing which makes it more clear what p-values do and do not tell us. Our framework will use the ASA statement as a jumping-off point. We will show how, for some

scientific questions of interest, p-values are an optimal tool for decision-making (even from a Bayesian perspective—a philosophy of statistics which we expand on in the next section). At the same time, we will show that alternative questions naturally lead us to considering alternative measures.

1.3 Historical Foundations of Bayesian Inference

The namesake and foundations of Bayesian inference are due to Thomas Bayes (1701-1763), an English statistician and minister. Bayes developed a formal approach to making statistical inferences by combining previous understanding (called the "prior belief" or, more simply, the "prior") with observations from a current experiment, resulting in a data-updated version of the prior belief known as the "posterior belief" (usually shortened to "posterior"). In its simplest form, this process of updating prior beliefs with data is accomplished using *Bayes' Theorem*:

$$P[B|A] = \frac{P[A|B]P[B]}{P[A|B]P[B] + P[A|B^c]P[B^c]},$$

where $P[A]$ denotes the probability of some event A, $P[B|A]$ denotes the probability of event B, conditional on knowing event A, and B^c denotes the complement of event B. Bayes' primary work containing the theorem that bears his name was unpublished at the time of his death. Bayes' literary executor, Richard Price, is responsible for editing the work and seeing it published in 1763 [31]. However, as is the case with much of mathematics, the development of so-called Bayesian methodology could be attributed to many statisticians, among them Price and the aforementioned Pierre-Simon Laplace. Laplace, evidently, was unaware of Bayes' work when he reproduced and extended many of the results in 1774 [5].

A statement of Bayes' theorem which is more directly useful when we are interested in inference about a continuous, unknown parameter θ and have at our disposal data $x_1, \ldots, x_n$ is

$$h(\theta|x_1, \ldots, x_n) = \frac{f(x_1, \ldots, x_n|\theta)g(\theta)}{\int f(x_1, \ldots, x_n|\theta)g(\theta)d\theta},$$

where our prior beliefs about the parameter θ are encompassed in the *prior distribution* $g(\theta)$, the function

$f(x_1, \ldots, x_n|\theta)$ denotes the *likelihood function* of the data for a given parameter value θ, the integral in the denominator is over the entire support of θ, and $h(\theta|x_1, \ldots, x_n)$ denotes the *posterior distribution* of θ conditional on having observed the data.

In the Bayesian paradigm, rather than seeing probability as a quantification of the long run frequency of events, probability is interpreted either as a reasonable expectation representing the state of current knowledge (the "objective Bayesian" view) or as a quantification of personal belief (the "subjective Bayesian" view). Much has been written about the differences between these views of probability and we will not attempt to review those philosophical works thoroughly here. For our purposes, it will be sufficient to understand the basic premise of the Bayesian approach and the quantities that Bayesian statisticians use to make decisions.

1.4 Bayesian Statistical Testing

Harold Jeffreys was the first to develop a fundamental theory of Bayesian inference and proposed comparing hypotheses using Bayesian posterior probabilities in 1939 [32]. His approach can be generally outlined as follows.

We will consider a null hypothesis $H_0 : \theta \in \Theta_0$ as well as an alternative $H_1 : \theta \in \Theta_1$, for Θ_0 a set of values of θ such that the null hypothesis is true and Θ_1 a set corresponding to values of θ such that the alternative hypothesis is true. We denote the likelihood function of the data for a given parameter value by $f(x|\theta)$. Using the pre-experimental knowledge available to us, we can specify prior probabilities for the truth of each of the two hypotheses under consideration, $P(H_0)$ and $P(H_1)$, as well as prior probability densities $\pi_0(\theta)$ and $\pi_1(\theta)$ on Θ_0 and Θ_1, respectively.

Using this notation, the marginal likelihood under each respective hypothesis is given by

$$m(x|H_i) = \int_{\Theta_i} f(x|\theta_i)\pi_i(\theta)d\theta, \qquad i = 0, 1.$$

After computing each of the respective marginal likelihoods, we can compute a *Bayes factor*,

$$B_{01} = \frac{m(x|H_0)}{m(x|H_1)},$$

and posterior probabilities of each of the hypotheses,

$$P(H_0|x) = \frac{P(H_0)m(x|H_0)}{P(H_0)m(x|H_0) + P(H_1)m(x|H_1)} = 1 - P(H_1|x).$$

These can be related using

$$\frac{P(H_0|x)}{P(H_1|x)} = \frac{P(H_0)}{P(H_1)} \times B_{01}.$$

It is often the case that a researcher will consider each of the competing hypotheses equally likely prior to conducting an experiment, in which case $P(H_0) = P(H_1)$ and the Bayes factor is simply the ratio of the posterior probability of H_0 to H_1. In this way, the Bayes Factor measures how much the experiment influenced our thinking about the hypothesis.

Jeffreys suggested a scale for interpretation of the value of the Bayes factor B_{01} as it relates to making decisions about which hypothesis is most supported by the combination of prior beliefs and data (Table 1.4.1).

Table 1.4.1: Bayes Factor Interpretations proposed by Jeffreys

Bayes Factor B_{01}	**Strength of Evidence**
>30	Very strong evidence for H_0
10 - 30	Strong evidence for H_0
3 - 10	Moderate evidence for H_0
1-3	Anecdotal evidence for H_0
1	No evidence
$\frac{1}{3}$ - 1	Anecdotal evidence for H_1
$\frac{1}{10}$ - $\frac{1}{3}$	Moderate evidence for H_1
$\frac{1}{30}$ - $\frac{1}{10}$	Strong evidence for H_1
$<\frac{1}{30}$	Very strong evidence for H_1

A modern exposition of Bayes factors is provided by Kass and Raftery [33], who also developed an alternative scale for interpreting Bayes factor values (Table 1.4.2). Kass and Raftery emphasize a number of positive attributes regarding the use of Bayes factors. Among these attributes are that Bayes factors allow researchers to incorporate external information into their evaluation of evidence about a

hypothesis, that Bayes factors are very general and there are several available techniques for computing them, and that Bayes factors can be converted and used as weights for various models in order to build composite estimates that take into account model uncertainty.

Table 1.4.2: Bayes Factor Interpretations proposed by Kass and Raftery

Bayes Factor B_{01}	**Strength of Evidence**
>150	Very strong evidence for H_0
20 - 150	Strong evidence for H_0
3 - 20	Positive evidence for H_0
1 - 3	Not worth more than a bare mention

The use of Bayes factors for making decision about competing hypotheses will prove to be an important component in many of the tests we develop later in Section 2.

1.5 Criticisms of the Bayesian Approach

There have been a number of criticisms of the Bayesian approach [34]. Typically, these criticisms point to the role of the prior distribution in the analysis. Usually, the concern among frequentists regarding the use of prior distributions in Bayesian analysis is that while prior information is often vague, the prior distribution is specified completely [35]. More succinctly, frequentists object to the subjectivity of the prior distribution. There are two reasonable responses to this from a Bayesian perspective. First, a Bayesian statistician would point out that while this is true, the same criticism could be equally applied to the frequentist approach, in which the statistician often assumes a specific parametric form of the likelihood function. Furthermore, a Bayesian statistician would argue, Bayesian statistics provides methods for evaluating the role of the prior in the analysis. Hierarchical analyses (in which one specifies prior distributions for the prior parameters themselves) and sensitivity analyses (in which one evaluates the robustness of inference to the prior distribution by performing the analysis with a variety of priors) can deal with ambiguity in the specification of the prior and are encouraged as good practices in the Bayesian paradigm [36, 37, 38]. Finally, it is also true that many Bayesian analyses are performed with so-called "noninformative", or "flat", priors which spread the density of the prior distribution evenly

among reasonable values of the parameter (although we will see in Section 3.2 that priors of this type may not be ideal in many cases, a point which has also been addressed previously [39]). Work has been done toward the selection of prior distributions using formal rules which would eliminate some of the subjectivity involved in Bayesian procedures [40].

1.6 Relationships between the Bayesian and Frequentist Approaches and Previous Suggestions for a Way Forward

Statisticians have been debating the relative merits of the Bayesian and frequentist paradigms for almost a century and will continue to do so. However, many of the issues which currently divide statisticians with respect to their preferred approach are philosophical or pedagogical and there is increasing recognition that both approaches can offer something of merit to statistical practice [41]. There is a growing body of literature offering connections between the Bayesian and frequentist point of view [42, 43]. An approachable and concise overview of the interplay between Bayesian and frequentist analysis is given by Bayarri and Berger [41]. We now give our own brief introduction to some previous work regarding these relationships with an eye toward the decision theoretic framework we develop later that incorporates ideas from both perspectives. Many authors have made suggestions for areas of compromise between the Bayesian and frequentist viewpoints or have suggested other ways of performing statistical tests. This body of literature is quite large and we will not attempt to list all of these works here, but we will point interested readers to a few selected works.

For an initial example, we turn to the first task in many statistical analyses: sample size determination. This task asks statisticians to strike a balance between the costs of experimentation and the strength of the statement we can ultimately make about the results. From the frequentist perspective, this often means specifying the desired power of the eventual statistical test to detect an effect of a certain size at a specific significance level and then selecting the smallest sample which will provide that power. A Bayesian perspective might instead seek to describe a prior distribution for the hypothetical effect and choose the sample size which provides the desired rate of correct classification of hypotheses as either true or false. The considerations at play for both the Bayesian and frequentist in exactly these scenarios have been written about by Inoue, Berry, and Parmigiani, who also provide a framework for identifying

mappings between the two approaches [44]. Other related work includes Spiegelhalter and Friedman [45]; Adcock [46]; Weiss [47]; and Pham-Gia and Turkkan [48].

Furthermore, in contrast to the aforementioned work of Berger and Sellke describing the sometimes large discrepancies between p-values and Bayesian probabilities that the null hypothesis is true in a two-sided, point null test (Section 1.2), Casella and Berger have pointed out that there are also hypothesis testing scenarios in which these two quantities can be reconciled. Specifically, in a one-sided testing setting it can be shown that for many classes of prior distributions the infimum of the Bayesian posterior probability that the null hypothesis is true is equal to the p-value [49]. That is, the discrepancy pointed out previously between low, highly significant p-values and the somewhat higher probability of a false null hypothesis (which we previously called a limitation of p-values since they are often misinterpreted as such a posterior probability, thus leading to overstating the evidence against the null) is problem-specific. In the two-sided, point-null case it is true that misinterpreted p-values can overstate the weight of evidence. But in the one-sided case, under reasonable priors, Bayesian statisticians can find some utility in frequentist approaches and vice versa. As pointed out by Casella and Berger, this leads us to question what important factors differentiate the two problems. According to Casella and Berger, differences between the Bayesian and frequentist measures will obtain when the Bayesian specifies that the prior mass is concentrated at a point (the null) and the remainder is allowed to vary over the alternative (as is the case in traditional two-sided testing) [49]. But in other scenarios the Bayesian and frequentist measures may well overlap and thus quantities from one perspective may have interpretations within the other. Pratt gives a thorough review of the ways in which a Bayesian statistician might make use of the standard inference measures developed in the frequentist paradigm [50]. This work discusses Bayesian uses for insufficient statistics, relationships between frequentist unbiased estimators and Bayesian posterior means, and approximate Bayesian properties of confidence intervals. Pratt also goes into detail on a Bayesian perspective on p-values and common uses of tests, with the general conclusion being similar to the work of Casella and Berger in that only certain one-tailed p-values are presented as having a Bayesian interpretation. In the framework we develop later we will expand on this idea by describing scenarios in which p-values are the only reasonable decision-making tool from a Bayesian perspective.

Poole promoted a Bayesian-leaning approach to rectifying the problems with p-values by suggesting that statisticians should report a graph of the entire *p-value function*—a plot of all possible p-values that would have been reported for a wide range of plausible test statistics [51]. Goodman appealed to reporting the minimum Bayes factor for the null hypothesis [52]. In situations where the prior probability distribution is symmetric and descending around the null value to be tested against, this minimum Bayes factor is $-ep \ln p$, where p is the fixed sample size p-value. This quantity is similar to and computed from the same information as a p-value but, Goodman argued, a conceptual step in the right direction because it encourages viewing data as one source of information among many. Greenland and Poole point out that a direct Bayesian interpretation of p-values can be made when two extreme types of prior distributions are considered: point-mass and uniform priors [53]. DeGroot provides examples where the tail areas in the distribution of test statistics used to compute p-values can be interpreted as either Bayesian posterior probabilities or likelihood ratios [54]. Finally, Wagenmakers suggested a switch from p-value methodology to model selection methodology using the Bayesian information criterion [55].

1.7 Two Additional Approaches to Measuring the Weight of Evidence

There are two areas of previous work which we will review in slightly more detail than those above because we will see them both again in Section 2.2.1's derivation of an expression for the overall risk of testing procedures we encounter in our decision-theoretic framework.

1.7.1 Expected P-Values

P-values are random variables [56]. They are functions of observed data and hence have sampling distributions. Since this is the case, we may be inclined to ask about the expectation of the p-value under various hypotheses, similar to the proposal by Poole [51].

When the null model consists of a single distribution, it is well known that p-values are uniformly distributed under the null hypothesis [57]. In 1966, Dempster and Schatzoff initially proposed the *expected significance level* (which we will call the expected p-value), defined as the expected value of the observed p-value under a simple alternative hypothesis, as a criterion for comparing the sensitivities of

competing test statistics [58, 59]. In 1999, Sackrowitz and Samuel-Cahn pointed out that despite the early work by Dempster and Schatzoff, the stochastic aspect of p-values had often been ignored in the scientific literature [60]. Sackrowitz and Samuel-Cahn then went on to highlight several uses of expected p-values in hypothesis testing.

Since we will encounter expected p-values later in this work, we will now formally define them (using the same notation introduced by Sackrowitz and Samuel-Cahn) and outline some of their potential usefulness.

Let T be a test statistic (a random variable) with a distribution under the null hypothesis H_0 given by F_0. Straightforwardly, the p-value is the random variable $X = 1 - F_0(T)$. For a specified alternative θ, denote the distribution of T under that hypothesis as F_θ. Using this notation, the power of a level α test based on T is then $\mathbb{P}_\theta[X \leq \alpha]$ and, as Sackrowitz and Samuel-Cahn point out, this can also be expressed as

$$\mathbb{P}_\theta[X \leq \alpha] = 1 - F_\theta(F_0^{-1}(1 - \alpha)).$$

Letting α take all values between 0 and 1, Sackrowitz and Samuel-Cahn then obtain the distribution function of the p-value under the alternative. The EPV closely resembles the power of a test, but it depends on the specified alternative rather than the significance level. As such, this quantity is useful for measuring the performance of a test even in situations when the power function is difficult to evaluate because one can measure the strength of a test for a given true value of θ by evaluating the EPV at that alternative. Smaller values of the EPV correspond to stronger tests. Finally, Sackrowitz and Samuel-Cahn have argued, the expected p-value can help guide the interpretation of p-values.

1.7.2 Rejection Odds and Rejection Ratios

Another alternative to traditional testing using p-values was proposed by Bayarri, Benjamin, Berger, and Sellke. Their approach focuses on evaluating the odds of correctly rejecting a false null hypothesis relative to the odds of incorrectly rejecting a true null hypothesis [61].

Bayarri et al. describe this quantity in a setting where we wish to test $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$. Respectively referring to the prior odds of the null and alternative hypotheses as π_0 and π_1, they define

the pre-experimental odds of correct to incorrect rejection of the null hypothesis as

$$O_{pre} = \frac{\pi_1}{\pi_0} \times \frac{(1-\bar{\beta})}{\alpha},$$

where $1-\bar{\beta}$ is the average power of the test: $1-\bar{\beta} = \int(1-\beta(\theta))\pi(\theta)d\theta$. The term $\frac{1-\bar{\beta}}{\alpha}$ is the *pre-experimental rejection ratio* of H_1 to H_0. Bayarri et al. argue that the rejection ratio "represents the evidentiary impact of statistical significance" by taking into account the role of power when rejecting the null hypothesis. They also point out that the present emphasis on statistical significance motivates a faulty line of reasoning in that it encourages the belief that if evidence is unlikely under the null hypothesis then it must automatically be likely under the alternative hypothesis. This is not necessarily the case. As Bayarri et al. detail, if the pre-experimental rejection ratio is too small (a threshold of 16:1 is suggested for common scenarios), then even a statistically significant finding is not particularly informative since the average power of the test was low relative to the significance threshold. When this occurs, either because the significance threshold was high or the power of the test was low (or a combination of the two), significant results can plausibly be due to random chance. Higher values of the rejection ratio indicate smaller chances that significant results are due to chance, which suggests that they could feasibly be used to measure, in some sense, the likelihood that significant results are reproducible.

Bayarri et al. suggest that researchers report the pre-experimental rejection ratio, which incorporates frequentist notions of error probabilities, when presenting experimental designs. After a study has been conducted, they suggest reporting Bayes factors (also referred to as post-experimental rejection ratios) when presenting results.

1.8 Relevant Concepts in Decision Theory

We now introduce the framework of statistical decision theory—a field which formalizes the decision making process. This introduction will be necessarily brief. We do not attempt to outline all of the possible ways of evaluating decisions, focusing instead on describing only the concepts which are directly relevant to the our own work. A broad overview is provided in Parmagiani and Inoue [62], whose notation we will use throughout this section. As Parmagiani and Inoue point out, concepts in decision

theory are rooted in ideas from Bernoulli [63], Laplace [64], and Gauss [65] and were later formalized and generalized by Wald [66].

1.8.1 Loss Functions and Bayes Decisions

In the language of decision theory, the decision maker is tasked with choosing among a set of possible actions (i.e. decisions). The consequences of these actions depend on the state of nature, which we consider unknown. We will therefore formalize the notion of losses incurred for taking each of these actions as a function of the state of nature.

We denote the set of possible actions with $\mathcal{A}$ and an individual action to be taken with a. To be consistent with the notation used previously to describe the effect to be tested, we will denote the true, unknown state of nature by θ. A *loss function* is a real-valued function $L(\theta, a)$ which describes the loss incurred by taking action a when the true state of nature is θ. The term "loss" is used by convention and refers to a common scenario in which we imagine that we incur a penalty for taking actions which are not optimal, in some sense. Loss functions can also be thought of as measuring the (negative) utility of certain actions.

An intuitive way of evaluating which action a is optimal among the set of possible actions $\mathcal{A}$ is to consider the expected loss, where the expectation is taken over all the possible states of nature θ. This requires that we specify how probable each of the various possible values of θ are. We will denote these probabilities $\pi(\theta)$. The expected loss for an action a is then

$$\mathbb{E}[\text{Loss}]_a = \int_\Theta L(\theta, a)\pi(\theta)d\theta,$$

where Θ represents the space of possible values of θ. An action a^* is called *Bayes* if it minimizes this expected loss:

$$a^* = argmin \int_\Theta L(\theta, a)\pi(\theta)d\theta.$$

1.8.2 Risk and Optimality

A primary motivation of decision theory is to describe optimal rules for deciding on which action in $\mathcal{A}$ should be taken. For data x taking values in the set $\mathcal{X}$, let $f(x|\theta)$ be a probability density function. We wish to find a function $\delta(x)$, called a *decision rule*, which takes elements of the space $\mathcal{X}$ (data) and returns actions from $\mathcal{A}$ (actions, which are hopefully optimal in some sense).

For a decision rule δ, we define a risk function

$$R(\theta, \delta) = \int_X L(\theta, \delta) f(x|\theta) dx$$

which describes the average loss when using rule δ to make decisions, averaged over the possible values of x and their respective likelihoods.

A related concept is the *Bayes risk*. For a prior distribution $\pi(\theta)$ describing the probability of the various possible states of θ, the Bayes risk is

$$r(\pi, \delta) = \int_\Theta R(\theta, \delta)\pi(\theta) d\theta.$$

The Bayes risk is the average risk associated with a rule δ for a specific true state of nature θ, where the average is computed over the possible values of θ. An important concept in the work we develop in Section 2 is that of the *Bayes rule*, the decision rule with Bayes risk equal to the infimum Bayes risk among all decisions δ. That is, a decision rule δ^* is called Bayes with respect to a prior distribution π if

$$r(\pi, \delta^*) = \inf_\delta r(\pi, \delta).$$

The Bayes rule minimizes Bayes risk with respect to a prior distribution. However, the motivation for Bayes rules can also be seen from a frequentist perspective: the Bayes rule also minimizes posterior expected loss for every realization of the data X.

In the sections that follow, we will develop a testing procedure in simple settings by first carefully stating loss functions corresponding to actions taken and the state of nature and then deriving the Bayes rule for the setting. Our goals will be to demonstrate how the two dominating philosophies of statistics can

find common ground in these testing procedures—we will see that the Bayes rule will direct us to make decisions based on either a p-value or a Bayes factor depending on the question of interest—and to describe the risk (in a frequentist and Bayesian sense) that we face when making various decisions.

2 A Decision-Theoretic Framework for Statistical Testing

In this section, we outline a decision-theoretic framework for statistical testing. We develop tests as decisions with carefully chosen loss functions. From this perspective, we show how a variety of testing approaches obtain. This framework is intentionally simple—the goal being to help readers understand at a basic level what each type of analysis covered tells them. The framework encompasses hypothesis testing as well as significance testing and will be shown to motivate the use of one- and two-sided p-values and Bayes factors.

2.1 One-Sided Hypothesis Tests

To illustrate how this framework motivates the use of p-values, we begin with a simple testing scenario in which we are concerned with testing a hypothesis about the sign of a parameter θ. We give no consideration to the case where $\theta = 0$ and instead focus on two possibilities: $\theta > 0$ or $\theta < 0$.

As we have mentioned, this framework is intentionally simplistic. One could object to ignoring the possibility that $\theta = 0$ on the ground that we have oversimplified the problem. Later, we will show that this decision theoretic framework can indeed encompass the point null that $\theta = 0$. However, we point out that testing using point null hypotheses have been frequently criticized as unrealistic [67, 68]. Most often, researchers point out in their objections that rarely does any researcher actually believe that the point null hypothesis could be true (that is, that the parameter θ could ever be exactly equal to 0). Furthermore, we acknowledge that testing regarding only the sign of θ, without making inference on the specific value of θ, may also seem overly simplistic. However, we note that there are a variety of real-world statistical problems in which testing for the sign of a parameter commonly occurs. Moreover, one goal of this work is to demonstrate that the testing scenario should guide the choice of which measure

of evidence we use to make decisions. In simple settings such as the one we consider first, we will obtain reasonably simple results. As the testing scenario becomes more complex, the decision rules we derive will do so as well.

Our initial example permits only two possible decisions: $\theta > 0$ ("d=Above") or $\theta < 0$ ("d=Below"). These decisions may be either correct or incorrect depending on which of the two possible states of nature we consider is true. With each of these decisions, depending on the true state of nature, we will associate a loss for making the decision. In this case, the framework leads to the four loss functions given in Table 2.1.1, where l_{TA}, l_{TB}, l_{FA}, and l_{FB} denote the losses incurred when we make a true (T) or false (F) decision that θ is above (A) or below (B) zero.

Table 2.1.1: Decision Framework - One-Sided Hypothesis Test

	Decision	
	d=Above	d=Below
Loss when $\theta > 0$	l_{TA}	l_{FB}
$\theta < 0$	l_{FA}	l_{TB}

In this example and those that follow, it will be necessary to state reasonable restrictions on the loss functions associated with each decision. In this case, since naturally it is preferable to make a correct decision, we will specify that $l_{TA} < l_{FA}$ and $l_{TB} < l_{FB}$. In fact, it is sensible to require that the framework penalizes incorrect decisions more than correct decisions regardless of whether those decisions are that $\theta > 0$ or that $\theta < 0$. Therefore, we require that $max(l_{TA}, l_{TB}) < min(l_{FA}, l_{FB})$. These losses are depicted in Figure 2.1.1.

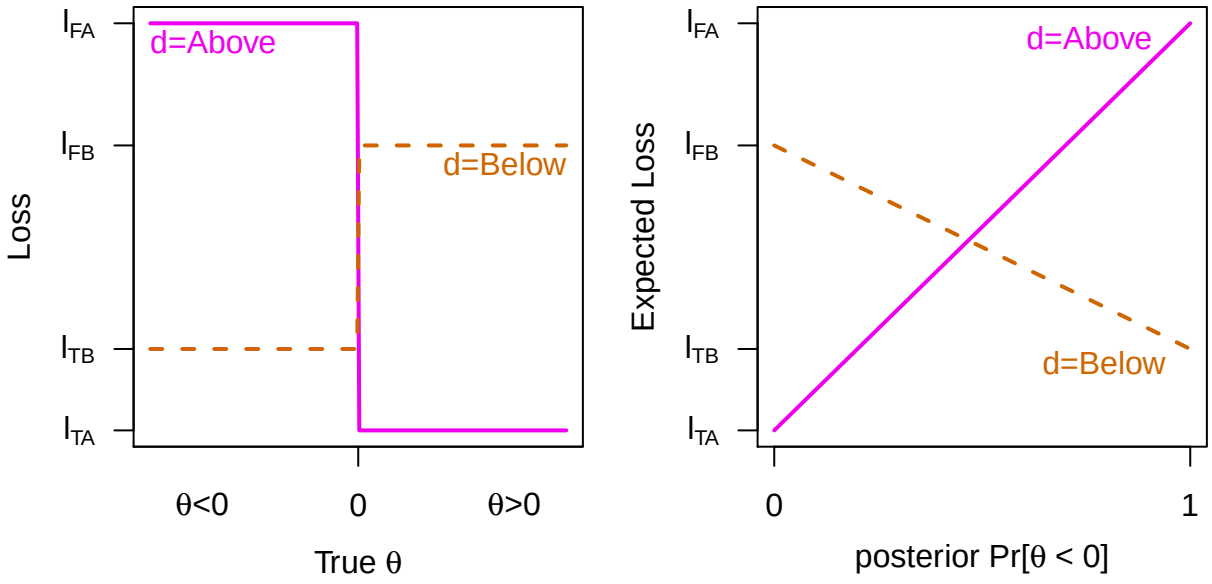


Figure 2.1.1: Loss functions and expected losses for a one-sided hypothesis test

Throughout the examples detailed in this section, we will use the same general approach. First, we will carefully construct the loss functions as we have above. Next, we will compute the expected loss when making each of the possible decisions. We will then relate these expected losses to derive a rule for when it is sensible to make each decision.

In this example, the loss incurred when making decision d=Above is

$$\text{Loss}_{d=Above} = l_{TA} I_{\theta>0} + l_{FA} I_{\theta<0},$$

where I_x is an indicator of event x. Taking the expectation, we find

$$\begin{aligned}\mathbb{E}[\text{Loss}]_{d=Above} &= l_{TA}\,\mathbb{P}[\theta > 0] + l_{FA}\,\mathbb{P}[\theta < 0] \\ &= l_{TA}(1 - \mathbb{P}[\theta < 0]) + l_{FA}\,\mathbb{P}[\theta < 0].\end{aligned}$$

Similarly, we find that

$$\begin{aligned}\text{Loss}_{d=Below} &= l_{TB} I_{\theta<0} + l_{FB} I_{\theta>0} \\ \implies \mathbb{E}[\text{Loss}]_{d=Below} &= l_{TB}\,\mathbb{P}[\theta < 0] + l_{FB}\,\mathbb{P}[\theta > 0]\end{aligned}$$

$$= l_{TB}\,\mathbb{P}[\theta < 0] + l_{FB}(1 - \mathbb{P}[\theta < 0]).$$

We see that both of the expected losses derived above are functions of $\mathbb{P}[\theta < 0]$ (Figure 2.1.1). From a Bayesian perspective, this quantity is the area of the posterior distribution below $\theta = 0$. By comparing the expected losses for each decision, we can then derive the Bayes rule for this test (the decision rule which minimizes the expected loss). For the decision d=Above, the Bayes rule states to make the decision if and only if

$$\begin{aligned}
&\mathbb{E}[\text{Loss}]_{d=Above} < \mathbb{E}[\text{Loss}]_{d=Below} \\
\implies\ & l_{TA}(1 - \mathbb{P}[\theta < 0]) + l_{FA}\,\mathbb{P}[\theta < 0] < l_{TB}\,\mathbb{P}[\theta < 0] + l_{FB}(1 - \mathbb{P}[\theta < 0]) \\
\implies\ & l_{TA}\,\mathbb{P}[\theta < 0] + l_{FA}\,\mathbb{P}[\theta < 0] < l_{FB} - l_{TA} + l_{TB}\,\mathbb{P}[\theta < 0] - l_{FB}\,\mathbb{P}[\theta < 0] \\
\implies\ & \mathbb{P}[\theta < 0](l_{FA} - l_{TA} + l_{FB} - l_{TB}) < l_{FB} - l_{TA} \\
\implies\ & \mathbb{P}[\theta < 0] < \frac{l_{FB} - l_{TA}}{l_{FA} - l_{TA} + l_{FB} - l_{TB}}.
\end{aligned}$$

The Bayes rule for decision d=Below can be derived similarly, resulting in the Bayes rule framework outlined in Table 2.1.2.

Table 2.1.2: Bayes Rule - One-Sided Hypothesis Test

	Decision	
	d=Above	d=Below
do d if and only if	$\mathbb{P}[\theta < 0] < \frac{l_{FB}-l_{TA}}{l_{FA}-l_{TA}+l_{FB}-l_{TB}}$	$\mathbb{P}[\theta < 0] > \frac{l_{FB}-l_{TA}}{l_{FA}-l_{TA}+l_{FB}-l_{TB}}$

From the Bayesian perspective, when the sample size is large and therefore the prior distribution for θ is dominated by the available data, the tail area $\mathbb{P}[\theta < 0]$ is approximately the frequentist p-value from a one-sided test under reasonable assumptions [54]. In other words, the Bayes rule for the above test (under reasonable regularity conditions) is a decision based only on the observed p-value and whether it is above or below some threshold $\frac{l_{FB}-l_{TA}}{l_{FA}-l_{TA}+l_{FB}-l_{TB}}$. Thus, in this simplistic scenario, we find a Bayesian analog of a classical one-sided hypothesis test. Importantly, the Bayes rule we have derived above (which is framed as a p-value) expresses the only reasonable choice a statistician approaching the problem from a Bayesian perspective could make. In this instance, both the Bayesian and frequentist statistician will

always agree on which decision should be made and that decision will always be made by a p-value.

To make this connection between approaches more clear, we can carefully reconstruct the loss functions to arrive at a more convenient expression for the threshold value used in the Bayes rule. Without loss of generality, we can scale the loss functions using addition or subtraction or multiplying by a constant. Here, it will be convenient to specify $l_{FB} - l_{TA} = \alpha$ and $l_{FA} - l_{TA} + l_{FB} - l_{TB} = 1$, yielding the final version of this testing framework presented in Table 2.1.3.

Table 2.1.3: Bayesian Analog of a One-Sided Hypothesis Test

	Decision	
	d=Above	d=Below
Loss when $\theta > 0$	l_{TA}	$l_{TA} + \alpha$
$\theta < 0$	$l_{TB} + 1 - \alpha$	l_{TB}
$\mathbb{E}$[Loss]	$l_{TA} + (l_{TB} - l_{TA} + 1 - \alpha)\,\mathbb{P}[\theta < 0]$	$l_{TA} + \alpha + (l_{TB} - l_{TA} - \alpha)\,\mathbb{P}[\theta < 0]$
do d if and only if	$\mathbb{P}[\theta < 0] < \alpha$	$\mathbb{P}[\theta < 0] > \alpha$

One advantage of approaching hypothesis testing from this decision-theoretic viewpoint is that it provides a direct, loss function-based interpretation of the p-value threshold α. Whereas the typical p-value threshold of 0.05 has been rightly criticized as arbitrary, this framework makes it clear that α corresponds to the excess penalty incurred by incorrect decisions, relative to correct decisions, when $\theta > 0$. Similarly, $1 - \alpha$ is the excess penalty incurred by incorrect decisions, relative to correct decisions, when $\theta < 0$. The ratio $\frac{\alpha}{1-\alpha}$ tells us the relative weight we put on incorrect decisions when $\theta > 0$ compared to when $\theta < 0$. Values of α above 0.5 indicate that we incur greater losses for incorrect “Above" responses. Similarly, values of α below 0.5 indicate that we incur greater losses for incorrect “Below" responses. To make a connection with traditional hypothesis testing, if we choose to test at level $\alpha = 0.05$ in this framework and assume zero loss for correct decisions then we are making an implicit statement that incorrectly declaring $\theta > 0$ is 19 times worse than incorrectly declaring $\theta < 0$.

2.2 One-Sided Significance Tests

We have seen that simple hypothesis testing, from both a Bayesian and frequentist point of view, can be easily motivated using this decision-theoretic framework. The framework can also be adapted to the significance testing philosophy originally proposed by Fisher. Like the hypothesis testing approach outlined above, we will first develop the one-sided significance testing approach in simple setting.

We will consider the null hypothesis that $\theta < 0$. Consistent with the philosophy of significance testing, we will not consider an alternative hypothesis. Since the null is that $\theta < 0$, the possible actions available to us in the decision framework are d=Above and d=No Decision. We will ignore the semantic issue of having one of our "decisions" being to make no decision. The losses associated with the decision d=Above will be identical to those specified in section 2.1, but when making d=No Decision we will specify the same loss, l_N, regardless of the true state of nature.

This last assumption—that we incur the same loss for all values of θ when making no decision—is arguably significant. In some cases it does seem that the true state of nature should impact the loss when making no decision. For example, if we consider the decision-making setting of a court trial, then doing nothing when a defendant is guilty is potentially much worse than doing nothing when a defendant is innocent. However, we can also come up with examples where the distinction is not so obvious. In medical settings, doing nothing when a drug has a very small positive effect is most likely neither much better nor much worse than doing nothing when the drug has a very small negative effect. And furthermore, if the drug in question does happen to have a large effect in either direction, then we will see that the chances we would opt to do nothing are slim. Even so, one should carefully consider the validity of the assumptions we make here before applying this significance testing framework.

Finally, we assume that making no decision is neither as good as making a correct decision nor as bad as making an incorrect decision, and therefore we will assume that $l_{TA} < l_N < l_{FA}$. These losses and the corresponding expected losses as a function of the posterior probability that θ is negative are depicted in Figure 2.2.1.

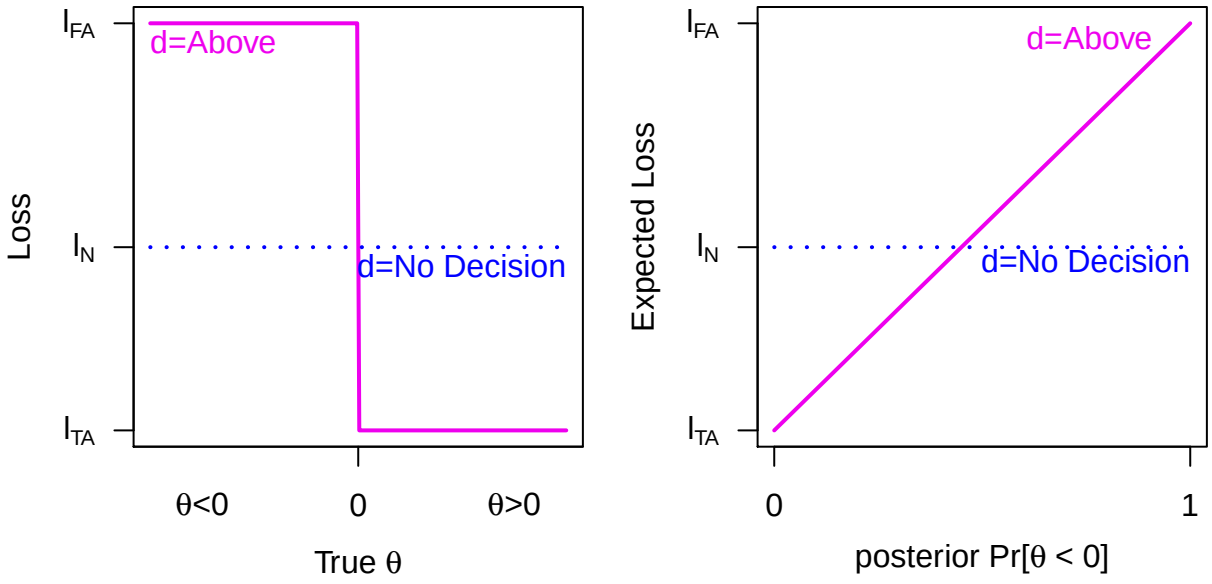


Figure 2.2.1: Loss functions and expected losses for a one-sided significance test

We then implement the same procedure we used for the one-sided hypothesis test, computing the expected loss for each decision and then deriving the Bayes rule for this test in terms of the tail area of the posterior distribution (from the Bayesian perspective), $\mathbb{P}[\theta < 0]$ (Table 2.2.1).

Table 2.2.1: Bayesian Analog of a One-Sided Significance Test

	Decision	
	d=Above	d=No Decision
Loss when $\theta > 0$	l_{TA}	l_N
$\theta < 0$	l_{FA}	l_N
$\mathbb{E}$[Loss]	$l_{TA}(1 - \mathbb{P}[\theta < 0]) + l_{FA}\,\mathbb{P}[\theta < 0]$	l_N
do d if and only if	$\mathbb{P}[\theta < 0] < \frac{l_N - l_{TA}}{l_{FA} - l_{TA}}$	$\mathbb{P}[\theta < 0] > \frac{l_N - l_{TA}}{l_{FA} - l_{TA}}$

Once again, we can conveniently rescale the loss functions in Table 2.2.1 by requiring $l_N - l_{TA} = \alpha$ and $l_{FA} - l_{TA} = 1$, resulting in the final form of this testing framework presented in Table 2.2.2.

Table 2.2.2: Rescaled Bayesian Analog of a One-Sided Significance Test

	Decision	
	d=Above	d=No Decision
Loss when $\theta > 0$	$l_N - \alpha$	l_N
$\theta < 0$	$l_N + 1 - \alpha$	l_N
$\mathbb{E}$[Loss]	$l_N - \alpha + \mathbb{P}[\theta < 0]$	l_N
do d if and only if	$\mathbb{P}[\theta < 0] < \alpha$	$\mathbb{P}[\theta < 0] > \alpha$

Once again, this view leads to a direct interpretation of the significance threshold α. In the significance testing framework, α is the extra utility (negative loss) gained when making a correct d=Above decision rather than no decision at all and $1-\alpha$ is the additional loss incurred when making an incorrect d=Above decision rather than making no decision at all.

2.2.1 Quantifying Loss

In Section 1.8.2, we introduced Bayes rules by describing their motivation: the Bayes rule with respect to a prior distribution π minimizes the Bayes risk and provides minimal frequentist loss by minimizing the loss for every realization of the data. In this section, we will further investigate frequentist connections to the Bayes rule by describing how much expected loss we face when making decisions according to the Bayes rule. The quantity which will be our focus is a frequentist measure in the sense that it is an expected value, but it does incorporate Bayesian methods in that it relies on prior information. We will refer to this quantity as the *frequentist expectation of the minimized posterior loss* (FEMPL).

For now, we remain in the one-sided significance testing setting considered in Section 2.2. Without loss of generality, we will assume $l_N = \alpha$ (Table 2.2.3).

Table 2.2.3: Rescaled Bayesian Analog of a One-Sided Significance Test with $l_N = \alpha$

	Decision	
	d=Above	d=No Decision
Loss when $\theta > 0$	0	α
$\theta < 0$	1	α
$\mathbb{E}$[Loss]	$\mathbb{P}[\theta < 0]$	α
do d if and only if	$\mathbb{P}[\theta < 0] < \alpha$	$\mathbb{P}[\theta < 0] > \alpha$

We find that the FEMPL is simply the minimum of $\mathbb{P}[\theta < 0]$, which we will call P for notational convenience, and α. This quantity, which we will denote $\mathcal{R}(\theta, \alpha)$ since it resembles the expected risk we face as a function of θ and α, can also be expressed as

$$\begin{aligned}
\mathcal{R}(\theta, \alpha) &= \mathbb{E}[\min(P, \alpha)] \\
&= \mathbb{P}[P < \alpha]\,\mathbb{E}[P|P < \alpha] + \mathbb{P}[P > \alpha]\,\mathbb{E}[\alpha|P > \alpha] \\
&= \mathbb{P}[P < \alpha]\,\mathbb{E}[P|P < \alpha] + \mathbb{P}[P > \alpha]\alpha \\
&= \alpha - \mathbb{P}[P < \alpha]\,\mathbb{E}[\alpha - P|P < \alpha] \\
&= \alpha\Big(1 - \frac{\mathbb{P}[P < \alpha]}{\alpha}\mathbb{E}[\alpha - P|P < \alpha]\Big).
\end{aligned}$$

There are a few terms in this expected minimized posterior loss which are familiar. The probability $\mathbb{P}[P < \alpha]$, which is a function of θ, is the power of the test when $\theta > 0$ and the Type I error rate when $\theta < 0$. We also see that the FEMPL is bound between 0 (when $\frac{\mathbb{P}[P<\alpha]}{\alpha}\mathbb{E}[\alpha - P|P < \alpha] = 1$) and 1 (when $\frac{\mathbb{P}[P<\alpha]}{\alpha}\mathbb{E}[\alpha - P|P < \alpha] = 0$).

Interestingly, the two quantities which appear to determine $\mathcal{R}(\theta, \alpha)$ are quantities we have seen previously. The conditional expectation $\mathbb{E}[\alpha - P|P < \alpha]$ is a scaled, conditional version of the expected p-value Sackrowitz and Samuel-Cahn have suggested can be used as a guide to interpreting p-values (Section 1.7.1). To see this, note that

$$\mathbb{E}[\alpha - P|P < \alpha] = \alpha(1 - \mathbb{E}[P/\alpha|P < \alpha]).$$

We will call the term $\mathbb{E}[P/\alpha|P < \alpha]$ a *scaled expected p-value.*

Furthermore, the ratio $\frac{\mathbb{P}[P<\alpha]}{\alpha}$ is the rejection ratio previously described by Bayarri et al (Section 1.7.2). The product of these two quantities (the rejection ratio and the scaled expected p-value) can then be said to capture the frequentist properties of Bayes rules. In Sections 3.1.2 and 3.2.1, we show that both of these two terms need to be included in the product and one cannot, for example, fully capture the frequentist properties of the Bayes rule by only describing the power of the test.

2.3 Two-Sided Significance Tests

Thus far we have demonstrated that in simple settings where our test is only concerned with the sign of θ and sample sizes are large, both Bayesian and frequentist statisticians will ultimately make decisions using p-values. This result is not limited to one-sided tests. We will now consider a somewhat more flexible version of the same question, in which we still concern ourselves with the sign of θ but combine ideas from Section 2.1 (where we made either the decision d=Above or d=Below) and Section 2.2 (where we introduced the possibility of making d=No Decision). Loss functions for each of the possible scenarios in this two-sided significance test are given in Table 2.3.1 and depicted in Figure 2.3.1.

Table 2.3.1: Decision Framework - Two-Sided Significance Test

	Decision		
	d=Above	d=No Decision	d=Below
Loss when $\theta > 0$	l_{TA}	l_N	l_{FB}
$\theta < 0$	l_{FA}	l_N	l_{TB}

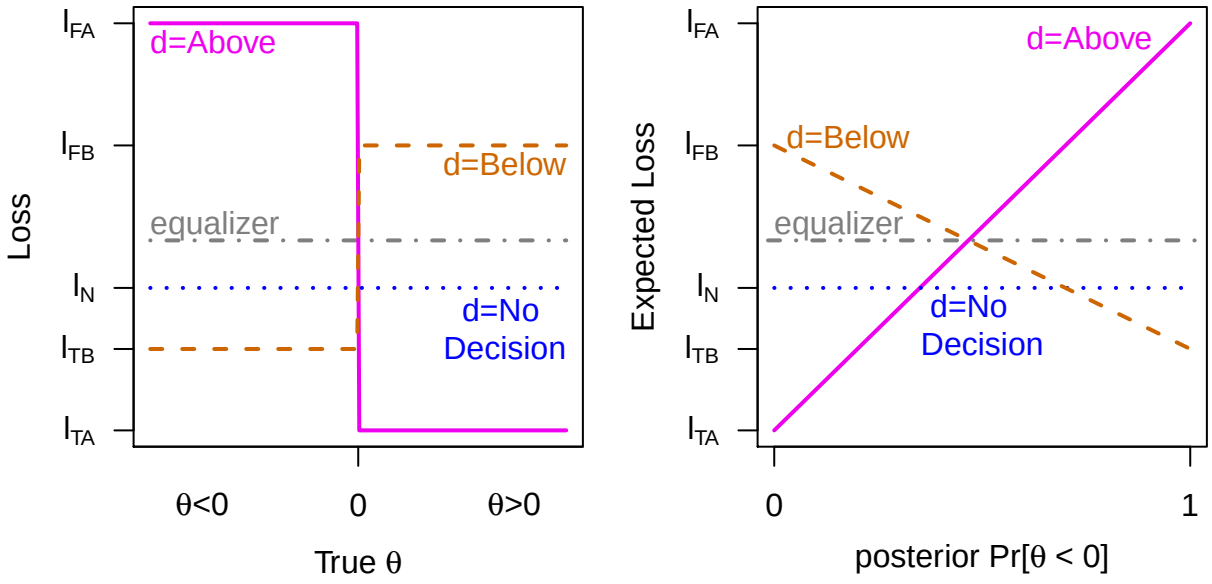


Figure 2.3.1: Loss functions and expected losses for a two-sided significance test

As in previous sections, we can impose some logical restrictions on the loss function in Table 2.3.1. We will repeat the assumptions that $l_{TA} < l_{FA}$ and $l_{TB} < l_{FB}$. When we did not give ourselves the option of making no decision, we outlined the logical assumption that $max(l_{TA}, l_{TB}) < min(l_{FA}, l_{FB})$ since the loss for making a correct decision should always be less than the loss for making incorrect decisions. Now, we will assume that making no decision is no worse than making an incorrect decision and no better than making a correct decision. Specifically, this assumption is that $max(l_{TA}, l_{TB}) < l_N < min(l_{FA}, l_{FB})$. We will also impose a slightly more strict condition on the loss function for making no decision. We imagine a decision rule which returns either d=Above or d=Below at random such that the average loss for the rule when $\theta < 0$ is equal to the average loss when $\theta > 0$. We call this rule the *equalizer rule*, and will assume that the loss incurred when making no decision, l_N, is no worse than the loss for this equalizer rule. This condition states that if we make no decision (an action we will take only when the data are not conclusively in favor of d=Above or d=Below), we are not worse off than if we had simply guessed randomly in the absence of conclusive evidence. Mathematically, this rule requires that

$$l_N < \frac{l_{FA}(l_{FB} - l_{TB}) + l_{TB}(l_{FA} - l_{TA})}{l_{FB} - l_{TB} + l_{FA} - l_{TA}} = \frac{l_{FB}(l_{FA} - l_{TA}) + l_{TA}(l_{FB} - l_{TB})}{l_{FB} - l_{TB} + l_{FA} - l_{TA}}.$$

Using the same procedures described above, we can quickly compute the expected loss for each of these

decision rules as well as the Bayes rule (Table 2.3.2).

Table 2.3.2: Bayesian Analog of a Two-Sided Significance Test

	Decision		
	d=Above	d=No Decision	d=Below
Loss when $\theta > 0$	l_{TA}	l_N	l_{FB}
$\theta < 0$	l_{FA}	l_N	l_{TB}
$\mathbb{E}$[Loss]	$l_{TA} + (l_{FA} - l_{TA})\,\mathbb{P}[\theta < 0]$	l_N	$l_{TB} + (l_{FB} - l_{TB})\,\mathbb{P}[\theta < 0]$
Bayes rule: do d if and only if	$\mathbb{P}[\theta < 0] < \frac{l_N - l_{TA}}{l_{FA} - l_{TA}}$	Otherwise	$\mathbb{P}[\theta < 0] > 1 - \frac{l_N - l_{TB}}{l_{FB} - l_{TB}}$

We see that this two-sided testing framework yields two thresholds for the left tail area $\mathbb{P}[\theta < 0]$. When the tail area (or, in large sample, the p-value) is small (specifically, less than $\frac{l_N - l_{TA}}{l_{FA} - l_{TA}}$), the Bayes rule tells us to make decision d=Above. On the other hand, if the tail area $\mathbb{P}[\theta < 0]$ is large (specifically, greater than $1 - \frac{l_N - l_{TB}}{l_{FB} - l_{TB}}$), then we have sufficient evidence to conclude using this rule that $\theta < 0$ and we will make decision d=Below. The equalizer rule described above guarantees that the lower threshold is strictly below the upper threshold. Values of $\mathbb{P}[\theta < 0]$ which fall between these two thresholds (i.e. when $\mathbb{P}[\theta < 0]$ is neither particularly small or particularly large) will lead to us making d=No Decision.

Traditionally, the conclusion of two-sided testing in the frequentist view is determined by comparing a two-sided p-value to two thresholds α_A and α_B. As we have done in previous sections, the comparison between the frequentist view and the Bayesian analog described above can be made direct by rescaling the loss functions carefully. We will introduce a scaling factor $\gamma \in [0, 1]$ and scale the loss such that $l_{FA} - l_{TA} = \gamma$ and $l_{FB} - l_{TB} = 1 - \gamma$ (Table 2.3.3).

Table 2.3.3: Bayesian Analog of a Two-Sided Significance Test

	Decision		
	d=Above	d=No Decision	d=Below
Loss when $\theta > 0$	$l_N - \alpha_A \gamma$	l_N	$l_N + (1 - \gamma)(1 - \alpha_B)$
$\theta < 0$	$l_N + \gamma(1 - \alpha_A)$	l_N	$l_N - \alpha_B(1 - \gamma)$
$\mathbb{E}$[Loss]	$l_N + \gamma(\mathbb{P}[\theta < 0] - \alpha_A)$	l_N	$l_N + (1 - \gamma)(1 - \mathbb{P}[\theta < 0] - \alpha_B)$
Bayes rule: do d if and only if	$\mathbb{P}[\theta < 0] < \alpha_A$	Otherwise	$\mathbb{P}[\theta < 0] > 1 - \alpha_B$

The scaling factor γ describes how strongly we wish to penalize different decisions when $\theta > 0$ compared

to when $\theta < 0$. For example, if we consider it worse to incorrectly decide d=Above than to incorrectly decide d=Below then we can impose larger losses to the scenario where we incorrectly choose d=Above by making γ close to one. This scenario might plausibly occur if, for example, we are deciding whether the effect of a new medical treatment is positive or negative. In that case, if we incorrectly decide d=Above, then we risk sending an ineffective treatment to larger, more expensive clinical trials (or worse, we make it available to the general public without knowing that it is actually harmful). In practice, we could increase γ accordingly to guard against such a false positive. Of course, this necessarily means that we will incur comparatively less loss for incorrectly deciding d=Below. In our example, this would mean that we were relatively more likely to improperly reject an effective drug than to approve an ineffective drug. Obviously, the trade-off between these considerations should be context-specific. In any case, the value of γ can be said to quantify this trade-off in some sense. Setting $\gamma = 0$ yields a one-sided significance test in which we will either make no decision or, if the evidence is particularly strong, decide d=Below.

In Section 2.2.1, we rescaled the loss functions (without loss of generality) so that the loss for making correct decisions was zero. Scaling the loss functions in Table 2.3.3 similarly is equivalent to specifying

$$l_N - \alpha_A\gamma = l_N - \alpha_B(1-\gamma),$$

which can be acomplished by letting $\alpha_A = \alpha_B$ and $\gamma = \frac{1}{2}$. By doing so, we eliminate the possibility of making no decision since for $\mathbb{P}[\theta < 0] < \alpha_A = \alpha_B$ we would decide d=Above and for $\mathbb{P}[\theta < 0] > \alpha_A = \alpha_B$ we would decide d=Below. Therefore, in this case, scaling the loss functions as we have before can result in reducing this significance test to a hypothesis test.

Finally, just as we have observed in previous sections, this framework lends itself to a loss-based interpretation of the thresholds α_A and α_B that may be lacking from the usual frequentist view: the ratio $\frac{\alpha_A}{1-\alpha_A}$ describes the ratio of extra utility to extra loss when we decide d=Above correctly instead of incorrectly and $\frac{\alpha_B}{1-\alpha_B}$ describes a similar ratio for scenarios in which we make decision d=Below.

2.4 Hypothesis Testing with Bayes Factors

Thus far, we have focused on deliberately simple situations in which we seek to answer a question about the sign of a parameter θ. In these scenarios we have derived Bayesian analogs to one- and two-sided hypothesis and significance tests, showing that Bayesians too can find some use in p-values and, conversely, that frequentists can find some use in Bayesian methods.

We will now turn our attention to a different question. Rather than concerning ourselves with the absolute sign of θ, we will instead ask how the data obtained have influenced our thinking about θ. Specifically, we will ask whether the data have provided evidence that θ is smaller or larger than previously thought—a question of relative support. To address this, we must have a way of comparing the prior for θ to the data-updated posterior estimate of θ.

Our approach to this problem will be to introduce a new parameter θ^* that we call a *clone parameter*. This independent parameter will have exactly the same prior distribution as θ but, importantly, will never be updated. We will then make decisions about the ordering of the updated estimate θ relative to θ^*. Since the question we seek to answer is now about relative support, the decisions available to us will be that $\theta < \theta^*$ ("Smaller") or $\theta > \theta^*$ ("Larger"). The loss functions for this scenario are outlined in Table 2.4.1, where we have introduced new subscript notion for the losses describing the true state of θ^* (either **P**ositive or **N**egative), the true state of θ (either **P**ositive or **N**egative), and the decision made (**S**maller or **L**arger).

Table 2.4.1: Loss Functions for a Bayes Factor Hypothesis Test

	Decision	
	d=Smaller ($\theta < \theta^*$)	d=Larger ($\theta > \theta^*$)
Loss when $\theta^* > 0$, $\theta > 0$	l_{PPS}	l_{PPL}
$\theta < 0$	l_{PNS}	l_{PNL}
$\theta^* < 0$, $\theta > 0$	l_{NPS}	l_{NPL}
$\theta < 0$	l_{NNS}	l_{NNL}

We are operating in a simplified setting where the only information about θ^* and θ are their sign. Therefore, when both parameters have the same sign, we cannot determine an ordering between the two.

Because of this, two defensible conditions on the loss function in Table 2.4.1 are that $l_{PPS} = l_{PPL} = l_{PP}$ and $l_{NNS} = l_{NNL} = l_{NN}$ (i.e. that if we cannot determine a true ordering, both decisions are viewed equally). Furthermore, since correct decisions should be penalized less than incorrect decisions, we will also ensure that $l_{PNS} < l_{NPS}$ and $l_{NPL} < l_{PNS}$.

We will also introduce some simplifying notation in order to make the expression in the following section easier to understand. We will write P for $\mathbb{P}[\theta < 0]$ since this term is interpreted in the large-sample context as a p-value in each of these sections. Similarly, we will write P^* for $\mathbb{P}[\theta^* < 0]$. Using this notation, we obtain the expected losses and Bayes rule depicted in Table 2.4.2.

Table 2.4.2: Decision Framework - A Bayes Factor Hypothesis Test

	Decision	
	d=Smaller ($\theta < \theta^*$)	d=Larger ($\theta > \theta^*$)
Loss when $\theta^* > 0$, $\theta > 0$	l_{PP}	l_{PP}
$\theta < 0$	l_{PNS}	l_{PNL}
$\theta^* < 0$, $\theta > 0$	l_{NPS}	l_{NPL}
$\theta < 0$	l_{NN}	l_{NN}
$\mathbb{E}$[Loss]	$(1\text{-}P^*)(1\text{-}P)l_{PP}$	$(1\text{-}P^*)(1\text{-}P)l_{PP}$
	$+(1\text{-}P^*)Pl_{PNS}$	$+(1\text{-}P^*)Pl_{PNL}$
	$+P^*(1\text{-}P)l_{NPS}$	$+P^*(1\text{-}P)l_{NPL}$
	$+P^*Pl_{NN}$	$+P^*Pl_{NN}$
Bayes rule: do d if and only if	$\frac{(1-P^*)P}{P^*(1-P)} < \frac{l_{PNS}-l_{PNL}}{l_{NPL}-l_{NPS}}$	$\frac{(1-P^*)P}{P^*(1-P)} > \frac{l_{PNS}-l_{PNL}}{l_{NPL}-l_{NPS}}$

The Bayes rule in this scenario depends on the value of the fraction $\frac{(1-P^*)P}{P^*(1-P)}$. The quantity is a Bayes factor, originally introduced in section 1.4. In this case, $\frac{(1-P^*)P}{P^*(1-P)}$ is the Bayes factor for θ being positive.

As in the examples where the decision was made using a p-value and we were able to rescale the loss functions to make the comparison to frequentist testing more explicit, in this example we can rescale the losses to simplify the expression to which we compare the Bayes factor (Table 2.4.3).

Table 2.4.3: Scaled Bayes Factor Hypothesis Test

	Decision	
	d=Smaller ($\theta < \theta^*$)	d=Larger ($\theta > \theta^*$)
Loss when $\theta^* > 0$, $\theta > 0$	l_{PP}	l_{PP}
$\theta < 0$	0	1
$\theta^* < 0$, $\theta > 0$	B	0
$\theta < 0$	l_{NN}	l_{NN}
Bayes rule: do d if and only if	$\frac{(1-P^*)P}{P^*(1-P)} < B$	$\frac{(1-P^*)P}{P^*(1-P)} > B$

As we pointed out in section 1.4, previous authors have suggested values of the Bayes factor (B, above) corresponding to different levels of evidence. These thresholds have been critiqued as arbitrary. However, in this simple setting where our question is about the relative ordering of θ and its non-updated prior θ^*, the decision-theoretic framework makes it clear what different values of B are. Orderings can only be determined in two cases: when $\theta < \theta^*$ and when $\theta > \theta^*$. The value of the Bayes factor tells us how many times worse it is to make an incorrect decision when $\theta > \theta^*$ relative to when $\theta < \theta^*$. For example, if we use the threshold of B=20 proposed by Kass and Raftery to correspond to "Strong" evidence for θ being positive then we are making an implicit statement that incorrect decisions are 20 times worse when $\theta > \theta^*$ relative to when $\theta < \theta^*$. Different proposals for threshold values of the Bayes factor can still be defended, but this framework at least takes a step toward providing problem-specific context for those arguments.

Previous work has compared Bayes factors and p-values in the context of genome-wide association studies [69]. One important contribution of this work was the description of a specific prior yielding identical gene rankings between Bayes factors and p-values which was said to provide a link between the two approaches. However, rather than describing prior distributions which relate the Bayes factor to p-values, the results of this section indicate that a simpler approach might be to instead focus on the question of interest when deciding which summary measure we wish to interpret and report. As we have seen, there are scenarios in which getting the sign of the parameter correct is the primary concern and the only Bayesian approach to decision making is informed by a p-value (for example in statistical genetics, when we simultaneously evaluate the relative expression of thousands of genes and only wish

to categorize the signals as positive or negative). On the other hand, if we care more about the degree to which data has changed our prior beliefs, then decision-making using a Bayes factor is appropriate and the trade-off implied by the threshold value of the Bayes factor we choose is directly interpretable.

2.5 One-Sided Bayes Factor Significance Tests

We are also able to extend the Bayes factor testing framework described in Section 2.4 to significance testing. We will again specify θ^*, the clone parameter, as an independent parameter with exactly the same prior distribution as θ. We will again make decisions about the ordering of θ relative to θ^*. However, we will now either make the decision that $\theta < \theta^*$ ("Larger") or make "No Decision" (Table 2.5.1). Our previous rule for subscript notation remains mostly unchanged, with one addition being the "ND" subscript used for losses incurred when making no decision. Intuitively, these losses should be such that $l_{PNL} < l_{PNND} < l_{NPL}$ and $l_{PNL} < l_{NPND} < l_{NPL}$.

Table 2.5.1: Decision Framework - A One-Sided Bayes Factor Significance Test

	Decision	
	d=Larger ($\theta > \theta^*$)	d=No Decision
Loss when $\theta^* > 0$, $\theta > 0$	l_{N_1}	l_{N_1}
$\theta < 0$	l_{PNL}	$l_{PNND} = l_{N_1} + k$
$\theta^* < 0$, $\theta > 0$	l_{NPL}	$l_{NPND} = l_{N_2} + k$
$\theta < 0$	l_{N_2}	l_{N_2}

The losses displayed in Table 2.5.1 have been carefully chosen in such a way that the expected risk (which we will see to be a function of θ and θ^*) can be formulated similarly to and compared with the expected risk calculated previously for a one-sided significance test which did not incorporate the Bayes factor (Section 2.2.1). We will see shortly that the comparison is accomplished by setting $k = 0$, but we will leave the notation general for now. To simplify our derivation of the Bayes rule, we will adopt the following notation:

$$P_{PP} = P[\theta^* > 0]P[\theta > 0]$$

$$P_{PN} = P[\theta^* > 0]P[\theta < 0]$$

$$P_{NP} = P[\theta^* < 0]P[\theta > 0]$$

$$P_{NN} = P[\theta^* < 0]P[\theta < 0]$$

The expected losses for each of these decisions are then straightforward to derive algebraically:

$$\mathbb{E}[\text{Loss}]_{\text{d=Larger}} = P_{PP}l_{N_1} + P_{PN}l_{PNL} + P_{NP}l_{NPL} + P_{NN}l_{N_2}$$

$$\mathbb{E}[\text{Loss}]_{\text{d=No Decision}} = P_{PP}l_{N_1} + P_{PN}(l_{N_1} + k) + P_{NP}(l_{N_2} + k) + P_{NN}l_{N_2}.$$

Based on these expected losses, the Bayes rule is to do d=Larger if and only if

$$(1 - P^*)Pl_{PNL} + P^*(1 - P)l_{NPL} < (1 - P^*)P[l_{N_1} + k] + P^*(1 - P)[l_{N_2} + k]$$
$$\implies \frac{(1 - P^*)P}{P^*(1 - P)} < \frac{(l_{N_2} + k) - l_{NPL}}{l_{PNL} - (l_{N_1} + k)}.$$

where we have again used $\mathbb{P}[\theta^* < 0] = P^*$ and $\mathbb{P}[\theta < 0] = P$ to ease the notation somewhat. Once more, we find that when answering questions of relative support, the Bayes rule points us to the use of a Bayes factor (Table 2.5.2).

Table 2.5.2: Bayesian Analog of a One-Sided Significance Test

	Decision	
	d=Larger ($\theta > \theta^*$)	d=No Decision
Loss when $\theta^* > 0$, $\theta > 0$	l_{N_1}	l_{N_1}
$\theta < 0$	l_{PNL}	$l_{PNND} = l_{N_1} + k$
$\theta^* < 0$, $\theta > 0$	l_{NPL}	$l_{NPND} = l_{N_2} + k$
$\theta < 0$	l_{N_2}	l_{N_2}
$\mathbb{E}[\text{Loss}]$	$P_{PP}l_{N_1}$	$P_{PP}l_{N_1}$
	$+P_{PN}l_{PNL}$	$+P_{PN}(l_{N_1} + k)$
	$+P_{NP}l_{NPL}$	$+P_{NP}(l_{N_2} + k)$
	$+P_{NN}l_{N_2}$	$+P_{NN}l_{N_2}$
Bayes rule: do d if and only if	$\frac{(1-P^*)P}{P^*(1-P)} < \frac{(l_{N_2}+k)-l_{NPL}}{l_{PNL}-(l_{N_1}+k)}$	$\frac{(1-P^*)P}{P^*(1-P)} > \frac{(l_{N_2}+k)-l_{NPL}}{l_{PNL}-(l_{N_1}+k)}$

In this case, the Bayes factor threshold is $\frac{(l_{N_2}+k)-l_{NPL}}{l_{PNL}-(l_{N_1}+k)}$. Of course, we are again free to rescale these losses however we wish to arrive at a slightly more convenient expression for the Bayes factor. One

method of doing this is to let $(l_{N_2} + k) - l_{NPL} = B$ and $l_{PNL} - (l_{N_1} + k) = 1$, where the value B could then be compared to suggested Bayes factor interpretation guides available in the Bayes factor literature.

In Section 2.2.1, we derived an expression for the frequentist expected minimized posterior loss (FEMPL), which we denoted $\mathcal{R}(\theta, \alpha)$, for a one-sided significance test where our question of interest was about the sign of the parameter θ. In that case, the FEMPL was found to be

$$\mathcal{R}(\theta, \alpha) = \alpha\Big(1 - \frac{\mathbb{P}[P < \alpha]}{\alpha}\mathbb{E}[\alpha - P | P < \alpha]\Big),$$

where $\mathbb{P}[P < \alpha]$ gives the power of the test for $\theta > 0$ and the Type I error rate for $\theta < 0$. Since we are now considering another one sided significance test, this time with a question of interest regarding the relative support of $\theta < 0$ provided by the data, one question we might ask is whether the FEMPL is similar in both scenarios. To derive a similar expression in this new setting, we begin by examining the expected loss for the decision d=Larger:

$$\begin{aligned}\mathbb{E}[\text{Loss}] = \mathbb{E}\Big[min\Big(&P(l_{PNL}(1 - P^*) + l_{N_2}(P^*)) + (1 - P)(l_{N_1}(1 - P^*) + l_{NPL}(P^*)),\\ &P(l_{PNND}(1 - P^*) + l_{N_2}(P^*)) + (1 - P)(l_{N_1}(1 - P^*) + l_{NPND}(P^*))\Big)\Big].\end{aligned}$$

We see above that the expected loss for this test is expressed as the expected value of the loss associated with whichever decision yields minimum loss. We will now pay specific attention to the second term in this minimum: the loss associated with making no decision. Since we previously specified that $l_{PNND} = l_{N_1} + k$ and $l_{NPND} = l_{N_2} + k$, we can rewrite the second term in the expected loss to arrive at

$$\begin{aligned}\mathbb{E}[\text{Loss}] &= \mathbb{E}\Big[min\Big(P(l_{PNL}(1 - P^*) + l_{N_2}(P^*)) + (1 - P)(l_{N_1}(1 - P^*) + l_{NPL}(P^*)),\\ &\qquad k[P(1 - P^*) + P^*(1 - P)] + l_{N_1}(1 - P^*) + l_{N_2}P^*\Big)\Big].\\ &= \mathbb{E}\Big[min\Big(A, B\Big)\Big].\end{aligned}$$

Now, if we let $k = 0$, we obtain $B = l_{N_1}(1 - P^*) + l_{N_2}P^*$. The expected loss for making no decision is then constant in P.

Next we will shift and rescale the expected loss for making decision d=Larger so that we can compare the expected loss from this Bayes factor test directly to the expected loss from the one-sided significance test which did not incorporate Bayes factors. Note that when $P = 0$, $A = l_{N_1}(1 - P^*) + l_{N_2}P^*$. We subtract this quantity from A so that the loss for this decision is 0 when $P = 0$. We then have

$$A = P\Big([l_{PNL} - l_{N_1}](1 - P^*) + [l_{N_2} - l_{NPL}]P^*\Big).$$

Finally, if we then rescale this loss by dividing by the term in parentheses, we are left with $A = P$. After doing so, the FEMPL for this test can be written as

$$\mathbb{E}[\text{Loss}] = \mathbb{E}\left[min\Big(P, \alpha^*\Big)\right],$$

where $\alpha^* = l_{N_1}(1 - P^*) + l_{N_2}P^*$.

We can then express the FEMPL for this significance test in a comparable form to that from the earlier one-sided significance test (Section 2.2.1). We have

$$\begin{aligned}
\mathcal{R}(\theta, \theta^*, \alpha^*) &= \mathbb{P}(P < \alpha^*)E[P|P < \alpha^*] + \mathbb{P}(P > \alpha^*)\alpha^* \\
&= \alpha^* - \mathbb{P}(P < \alpha^*)E[\alpha^* - P|P < \alpha^*] \\
&= \alpha^*\Big(1 - \frac{\mathbb{P}(P < \alpha^*)}{\alpha^*}E[\alpha^* - P|P < \alpha^*]\Big).
\end{aligned}$$

Conveniently, in both the version of the test which focuses on absolute support for $\theta < 0$ (and thus results in a Bayes rule based on p-values) and in the version considered here which focuses on relative support for $\theta < 0$ (and thus results in a Bayes rule based on Bayes factors), we can express the minimized posterior loss in the same way. We comment more on this expression in the context of tests about a binomial probability in Section 3.1 and again in the context of test about the mean parameter of a normal distribution in Section 3.2.1.

2.6 Two-Sided Bayes Factor Significance Tests

In a two-sided Bayes factor significance test, we will again make use of the clone parameter setup and make a decision about the relative ordering of θ to θ^*. In this case, we also reserve the right to make

no decision. We will see that we can scale the losses such that middling values of the Bayes factor will result in making no decision.

In order to coerce the test into producing a decision based on the value of the Bayes factor, we first specify that the loss when making no decision in situations where a relative ordering cannot be determined by looking only at the sign of θ and θ^* (either when both are positive or when both are negative) must be equal to the loss when deciding d=Smaller or d=Larger in the same scenario. When both θ^* and θ are positive, we call this loss l_{PP}. When both are negative, we call the loss l_{NN}. Furthermore, we will set up the test so that we incur loss 0 for correct decisions about the relative ordering of θ and θ^* and loss 1 for incorrect decisions. Finally, when making no decision in cases where an ordering could be determined, we will incur constant loss equal to $\frac{B}{1+B}$.

Table 2.6.1: Decision Framework - A Two-Sided Bayes Factor Significance Test

	Decision		
	d=Larger ($\theta > \theta^*$)	d=No Decision	d=Smaller ($\theta < \theta^*$)
Loss when $\theta^* > 0$, $\theta > 0$	l_{PP}	l_{PP}	l_{PP}
$\theta < 0$	1	$\frac{1}{1+B}$	0
$\theta^* < 0$, $\theta > 0$	0	$\frac{1}{1+B}$	1
$\theta < 0$	l_{NN}	l_{NN}	l_{NN}
Bayes rule: do d if and only if	$\frac{\mathbb{P}[\theta^*<0]\,\mathbb{P}[\theta>0]}{\mathbb{P}[\theta^*>0]\,\mathbb{P}[\theta<0]} > B$	otherwise	$\frac{\mathbb{P}[\theta^*>0]\,\mathbb{P}[\theta<0]}{\mathbb{P}[\theta^*<0]\,\mathbb{P}[\theta>0]} > B$

In Table 2.6.1, we see that the Bayes rule tells us to make decision d=Larger if the Bayes factor $\frac{\mathbb{P}[\theta^*<0]\,\mathbb{P}[\theta>0]}{\mathbb{P}[\theta^*>0]\,\mathbb{P}[\theta<0]}$ is greater than B. This result is similar to other Bayes factor decisions we derived in Section 2.5: if the Bayes factor for θ being positive is larger than our chosen threshold B, then we will decide that $\theta > \theta^*$. However, unlike the one-sided test (which compared one Bayes factor to one threshold value), this two-sided version suggests that we should use a different Bayes factor to determine whether the decision d=Smaller is optimal. In this new case, we will decide d=Smaller ($\theta < \theta^*$) if $\frac{\mathbb{P}[\theta^*>0]\,\mathbb{P}[\theta<0]}{\mathbb{P}[\theta^*<0]\,\mathbb{P}[\theta>0]}$, the Bayes factor for θ being negative, is greater than B. Of course, it is a simple algebraic exercise to rearrange the Bayes factor and arrive at an equivalent rule for making decision d=Smaller:

$$\text{Do d=Smaller if and only if } \frac{\mathbb{P}[\theta^* > 0]\,\mathbb{P}[\theta < 0]}{\mathbb{P}[\theta^* < 0]\,\mathbb{P}[\theta > 0]} > B$$

$$\implies \mathbb{P}[\theta^* > 0]\,\mathbb{P}[\theta < 0] > B\,\mathbb{P}[\theta^* < 0]\,\mathbb{P}[\theta > 0]$$
$$\implies \frac{\mathbb{P}[\theta^* < 0]\,\mathbb{P}[\theta > 0]}{\mathbb{P}[\theta^* > 0]\,\mathbb{P}[\theta < 0]} < \frac{1}{B}.$$

Rearranging the rule for deciding d=Smaller yields a slightly simpler approach to the test. The Bayes rule then tells us to evaluate one Bayes factor, the Bayes factor for θ being positive. If that value exceeds B, we will decide $\theta > \theta^*$. Alternatively, if that value is smaller than $\frac{1}{B}$, then we will decide $\theta < \theta^*$. Otherwise, for $\frac{1}{B} < \frac{\mathbb{P}[\theta^* < 0]\,\mathbb{P}[\theta > 0]}{\mathbb{P}[\theta^* > 0]\,\mathbb{P}[\theta < 0]} < B$, the Bayes rule dictates making no decision.

To make clear the connection to previous significance tests we have described, we can also rearrange the form of this decision rule to arrive at an inequality for the p-value. As we have said, we will decide d=Larger if and only if

$$\frac{\mathbb{P}[\theta^* < 0]\,\mathbb{P}[\theta > 0]}{\mathbb{P}[\theta^* > 0]\,\mathbb{P}[\theta < 0]} > B$$
$$\implies \mathbb{P}[\theta^* < 0]\,\mathbb{P}[\theta > 0] > B\,\mathbb{P}[\theta^* > 0]\,\mathbb{P}[\theta < 0]$$
$$\implies \mathbb{P}[\theta^* < 0] - \mathbb{P}[\theta^* < 0]\,\mathbb{P}[\theta < 0] > B\,\mathbb{P}[\theta^* > 0]\,\mathbb{P}[\theta < 0]$$
$$\implies \mathbb{P}[\theta^* < 0] > \mathbb{P}[\theta < 0]\Big[B\,\mathbb{P}[\theta^* > 0] + \mathbb{P}[\theta < 0]\Big]$$
$$\implies \frac{\mathbb{P}[\theta^* < 0]}{B\,\mathbb{P}[\theta^* > 0] + \mathbb{P}[\theta^* < 0]} > \mathbb{P}[\theta < 0].$$

That is, assuming a sufficiently large sample, we can frame our decision rule as based on a p-value (i.e. the tail area of a posterior distribution for θ,$\mathbb{P}[\theta < 0]$), where significance is determined by

$$\mathbb{P}[\theta < 0] < \frac{\mathbb{P}[\theta^* < 0]}{B\,\mathbb{P}[\theta^* > 0] + \mathbb{P}[\theta^* < 0]}.$$

We can, of course, do similar operation with the part of the Bayes rule which deals with the decision d=Smaller to eventually arrive at another correspondence between the Bayes factor threshold and a p-value. In that case, it can be seen that the Bayes rule implies we would choose d=Smaller if and only if

$$\mathbb{P}[\theta < 0] < \frac{B\,\mathbb{P}[\theta^* < 0]}{B\,\mathbb{P}[\theta^* < 0] + \mathbb{P}[\theta^* > 0]}.$$

In both cases, we find a direct correspondence between this significance testing framework (where our question of interest is one of relative support for hypotheses) and the earlier scenarios we considered

(where our question was one of absolute support). In the latter case, the Bayes rule required the use of a p-value to make decisions. We now see that even the more nuanced question of interest can be formulated such that our decision ultimately comes down to the use of a p-value, albeit with a new value of the significance threshold which depends on some prior information through θ^*. In either case, it is apparent that both p-values and Bayes factors can play a valuable role in evaluating the strength of evidence.

3 Examples

In this section, we walk through two examples of how this decision-theoretic framework for testing can be implemented. In both examples we will approach the test first from a frequentist perspective and then from a Bayesian perspective. Our primary concerns will be to suggest reasonable loss functions, to show the expected loss we encounter as a function of the true parameter in question, and to demonstrate how the Bayesian and frequentist tests compare. First, in Section 3.1, we present a scenario in which we wish to test the value of a binomial probability parameter. Then, in Section 3.2, we outline a test for the mean of a normal distribution. In both examples, we will present results from one-sided as well as two-sided tests.

3.1 Testing a Binomial Parameter

We begin with an example in which we imagine we have observed n independent Bernoulli random variables X_i, $i = 1, \ldots, n$. We will consider tests about the binomial probability $\theta = \mathbb{P}[X_i = 1]$. For both one and two-sided tests, we will compare two approaches: a frequentist binomial exact test and a Bayesian beta-binomial model, with respect to their expected loss.

3.1.1 One-Sided Tests

For the one-sided case, we will test the null hypothesis that $\theta \leq \frac{1}{2}$ against the alternative that $\theta > \frac{1}{2}$. We frame this as a significance test, and therefore we plan to either make "No Decision" or decide "Above"

(that θ is above $\frac{1}{2}$). Without loss of generality, we will choose the loss functions for each of the possible decisions according to Table 3.1.1.

Table 3.1.1: Decision Framework - One-Sided Binomial Significance Test

	Decision	
	d=No Decision	d=Above
Loss when $\theta \leq \frac{1}{2}$	α	1
$\theta > \frac{1}{2}$	α	0

Our motivation for examining the properties of this test come partly from its simplicity. As the name of the test implies, an exact formula for the p-value is readily calculated. Adopting the notation $\sum_{i=1}^{n} X_i = T$, we have that $T \sim \text{Bin}(n, \theta)$ (the sum of n independent and identically distributed Bernoulli random variables is a binomial random variable). As such, the probability mass function for T is

$$\mathbb{P}[T = t] = \binom{n}{t} \theta^t (1-\theta)^{n-t}.$$

For an observed value of the test statistic, t_{obs}, the p-value is an upper bound on the probability (computed under the null hypothesis that $\theta \leq \frac{1}{2}$) of observing a test statistic at least as extreme or more extreme than t_{obs}, where in this case "more extreme" indicates greater values of t:

$$\begin{aligned} p &= \sum_{t=t_{obs}}^{n} \binom{n}{t} \theta^t (1-\theta)^{n-t} \\ &= \sum_{t=t_{obs}}^{n} \binom{n}{t} 0.5^n. \end{aligned}$$

Before we define the expected loss for this test, we must first use the above formula to define the probability that we make each of the respective decisions. For making "No Decision", an action which we denote d_0, this probability is

$$\mathbb{P}[d_0] = \sum_{t=0}^{n} \binom{n}{t} \theta^t (1-\theta)^{n-t} * I_{P>\alpha},$$

where the indicator function $I_{P>\alpha}$ is 1 if $P > \alpha$ and 0 otherwise. Similarly, the probability of deciding

"Above", an action we denote d_1, is

$$\mathbb{P}[d_1] = \sum_{t=0}^{n} \binom{n}{t} \theta^t (1-\theta)^{n-t} * I_{P \leq \alpha}.$$

Finally, the expected loss for a decision is given by

$$\mathbb{E}\,\text{Loss} = I_{\theta > \frac{1}{2}} \Big(\mathbb{P}[d_0]\alpha \Big) + I_{\theta \leq \frac{1}{2}} \Big(\mathbb{P}[d_0]\alpha + \mathbb{P}[d_1] \Big).$$

We can now examine the form of this expected loss, which is a function of the sample size n, the true parameter value θ, and the desired significance level α. In Figure 3.1.1, we show this expected loss for the familiar significance level $\alpha = 0.05$ and $n = 20$.

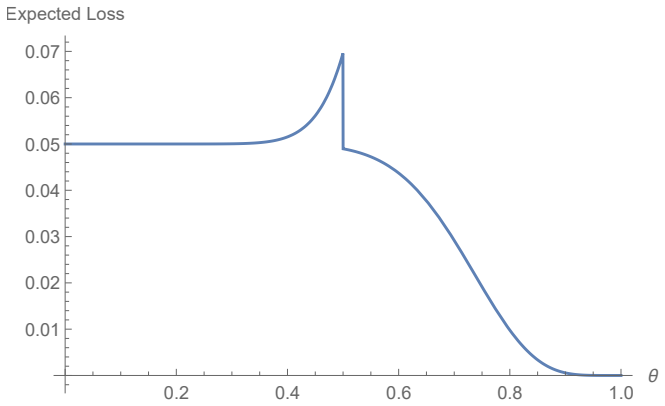
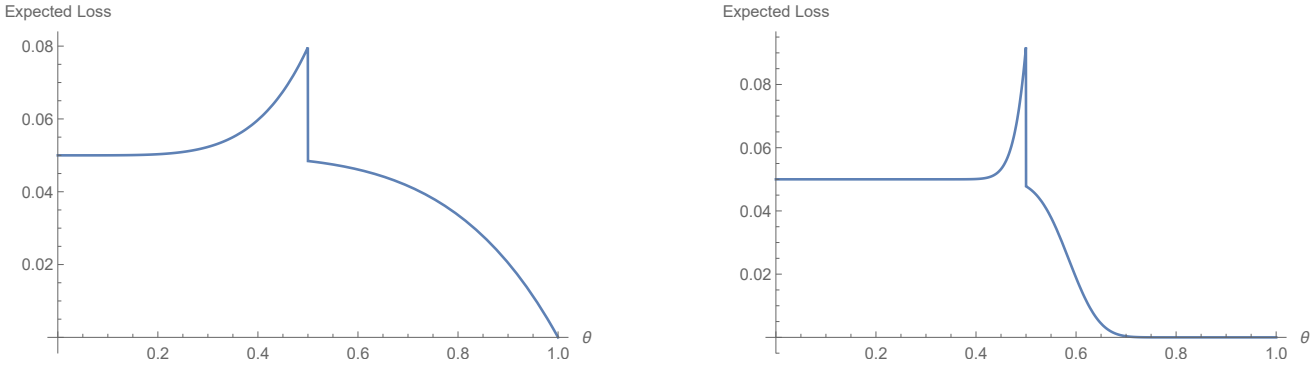


Figure 3.1.1: Expected loss for a binomial exact significance test with $n = 20$

In Figure 3.1.1, we see that for very small values of θ (i.e. cases where we will be quite sure that $\theta < \frac{1}{2}$), the expected loss is maintained at the significance level α. However, as θ approaches the null value of $\frac{1}{2}$, we observe a spike in the expected loss. The spike appears on the "No Decision" side of the null value in the one-sided significance test we consider here. In Section 3.1.3 we will see that in two-sided testing scenarios the spike is evenly distributed on both sides of the null value. Figure 3.1.1 also shows that for values of θ greater than the null, the expected loss tails off to zero. When θ is quite large (nearly 1), we will almost always correctly identify that $\theta > \frac{1}{2}$ and incur no loss.

As we alter the hypothetical sample size employed by this test, the shape of the expected loss function remains generally the same. For smaller values of n, the spike where the expected loss begins to exceed

the significance level of the test occurs for smaller values of θ and the expected loss tails off to zero beyond $\theta = \frac{1}{2}$ at a slower rate (Figure 3.1.2a). On the other hand, large values of n decrease the area of the spike and cause the expected loss to go to zero very quickly for values of θ larger than the null (Figure 3.1.2b).



(a) n=5 **(b) n=100**

Figure 3.1.2: Impact of changing sample size on the expected loss for a binomial exact significance test

Next, we address this problem from a Bayesian perspective. To do so, we will model the binomial probability using a *conjugate prior*—a prior distribution which, after being updated via the likelihood function, results in a posterior distribution belonging to the same family of densities as the prior. A common approach used for binary data is a model in which we specify a beta distribution for the prior for θ. By convention, we will call the parameters of this beta distribution α and β. To avoid confusion with the beta distribution parameter and the significance threshold, we will switch to calling the significance threshold α^* for the remainder of this example. Using

$$\theta \sim \text{Beta}(\alpha, \beta)$$

as a prior distribution and having binary data distributed according to

$$\sum_{i=1}^{n} X_i = T \sim \text{Binomial}(n, \theta),$$

the posterior distribution can be seen to be

$$(\theta | X_i) \sim \text{Beta}(\alpha + T, \beta + n - T),$$

and therefore the posterior has a density function given by

$$f(\theta|X_i) = \frac{1}{B(\alpha+T, \beta+n-T)}\theta^{\alpha+T-1}(1-\theta)^{\beta+n-T-1},$$

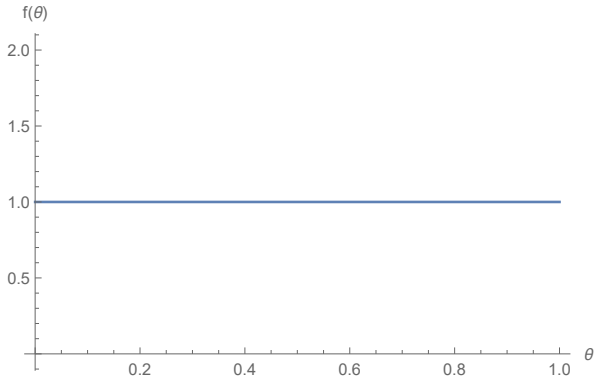
where B is the Beta function [4].

Since we have framed this as an upper-tailed one-sided significance test, our approach will be to decide "Above" for large values of T. More specifically, we will decide $\theta > \frac{1}{2}$ for values of T such that $F_{(\theta|X)}(\frac{1}{2}) < \alpha^*$, where F represents the cumulative distribution function of the posterior. That is, we will only decide that $\theta > \frac{1}{2}$ if and only if the value of the test statistic T is such that the posterior probability that θ is less than or equal to $\frac{1}{2}$ is less than the significance threshold α^*.

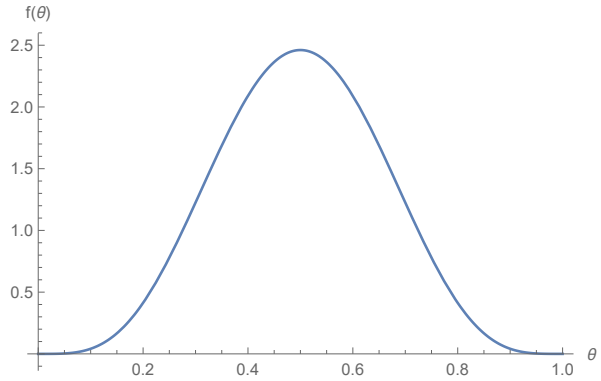
As in the frequentist version of this test, we will associate a constant loss α with making no decision. When we correctly decide $\theta > \frac{1}{2}$, we incur no loss. Otherwise, if we incorrectly decide $\theta > \frac{1}{2}$, we incur a loss of 1. Of course, as was the case in the frequentist setting, these losses can be scaled without loss of generality.

Since this is a Bayesian approach, the form of the prior (dictated by our chosen values for the beta distribution parameters α and β) plays a role in these decisions. Figure 3.1.3 shows the four types of priors we will consider. Figure 3.1.3a describes a flat prior in which both α and β are 1. This type of prior might be used in scenarios where we imagine that we have little to no prior information about θ and therefore consider all possible values of θ equally likely a priori. Figure 3.1.3b shows a symmetric prior, accomplished by letting $\alpha = \beta = 5$. The symmetric prior illustrates a scenario in which our prior information favors values of θ near $\frac{1}{2}$ but also gives some weight to other values of θ. Figure 3.1.3c describes an opposite scenario where $\alpha = \beta = \frac{1}{5}$, indicating that our prior belief points us toward an extreme value of θ but we are unsure whether the binomial probability is very small (near zero) or very large (near one). Finally, Figure 3.1.3d shows the situation most suited to the one-sided significance test setting we consider here in which our prior distribution gives much higher likelihood to values of θ below $\frac{1}{2}$.

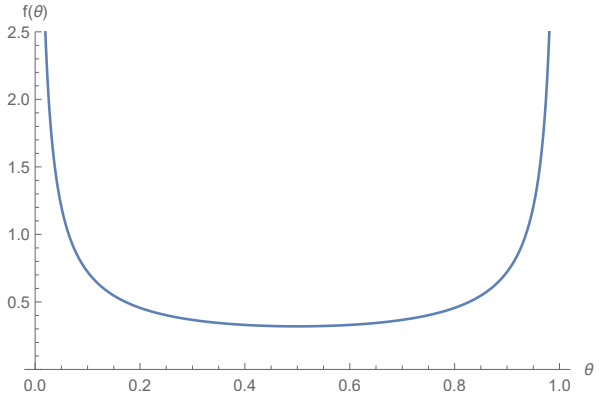
For each of these prior distributions, we will plot the expected loss (as a function of θ) for three sample sizes: $n = 5$, 20, and 100 (Figure 3.1.4).



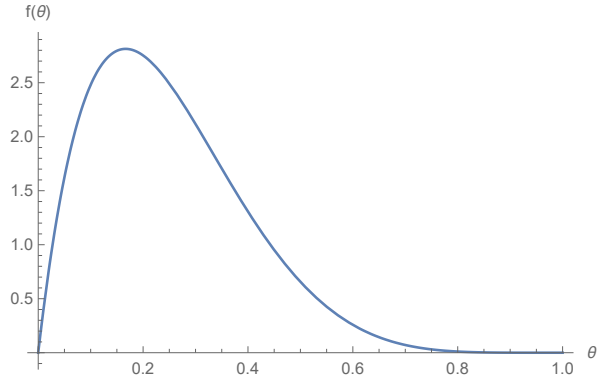
(a) $\alpha = \beta = 1$



(b) $\alpha = \beta = 5$

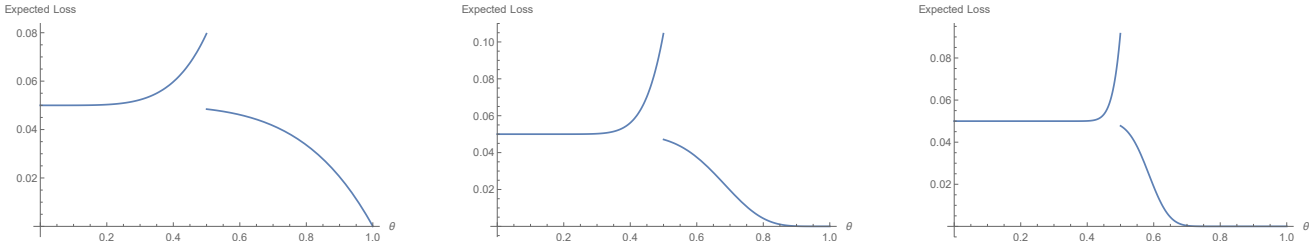


(c) $\alpha = \beta = \frac{1}{5}$

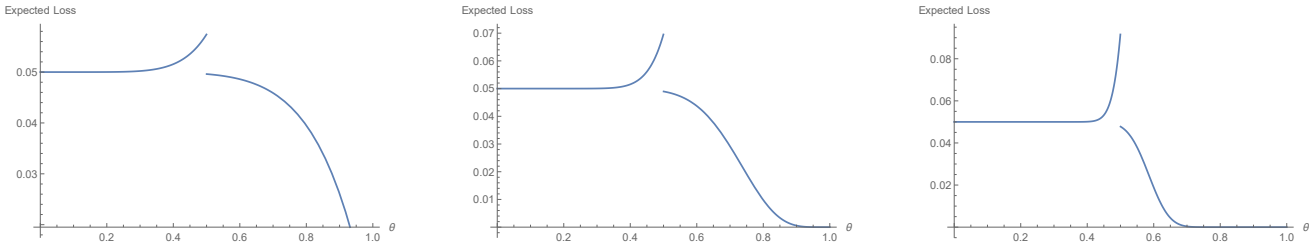


(d) $\alpha = 2$, $\beta = 6$

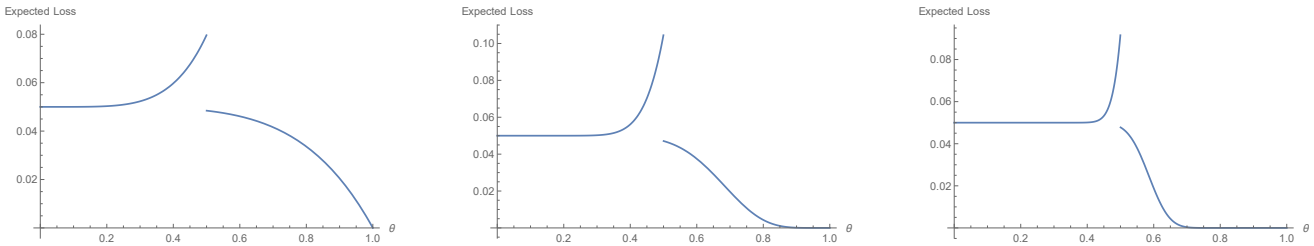
Figure 3.1.3: Four types of beta prior distributions



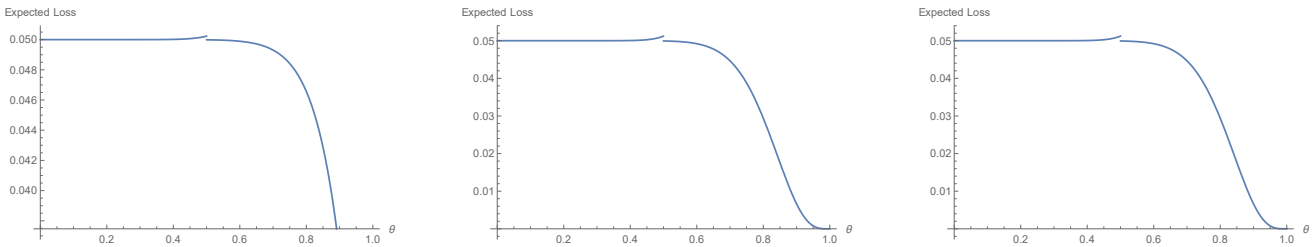
(a) $\alpha = \beta = 1$, $n = 5$ **(b)** $\alpha = \beta = 1$, $n = 20$ **(c)** $\alpha = \beta = 1$, $n = 100$



(d) $\alpha = \beta = 5$, $n = 5$ **(e)** $\alpha = \beta = 5$, $n = 20$ **(f)** $\alpha = \beta = 5$, $n = 100$



(g) $\alpha = \beta = \frac{1}{5}$, $n = 5$ **(h)** $\alpha = \beta = \frac{1}{5}$, $n = 20$ **(i)** $\alpha = \beta = \frac{1}{5}$, $n = 100$



(j) $\alpha = 2$, $\beta = 6$, $n = 5$ **(k)** $\alpha = 2$, $\beta = 6$, $n = 20$ **(l)** $\alpha = 2$, $\beta = 6$, $n = 100$

Figure 3.1.4: Expected loss for a Bayesian beta-binomial decision model using various priors and sample sizes

Figure 3.1.4 makes a few things clear. First, we find a direct comparison between the expected loss function in this Bayesian analysis and the expected loss function from the frequentist counterpart to this test (Figure 3.1.2). That is to say, it does not appear to make a difference in terms of the expected loss whether we choose to approach this problem from a frequentist or Bayesian perspective. In both cases, the expected loss spikes on the left hand side of the null value for θ and then drops off to zero as θ gets closer to one. The Bayesian approach does allow us to control the shape of the posterior somewhat by choosing the values of the prior parameters, but even for a wide variety of prior parameter values and sample sizes we see that the posterior is generally unchanged (Figures 3.1.4a-3.1.4i). However, we also observe that the prior distribution which most strongly reflects the prior knowledge which would lead us to consider this significance framework in the first place leads to an expected loss with a less prominent spike to the left of the null value. In this case, the right-skewed density of the prior distribution pulls down the spike considerably, resulting in an expected loss which is more consistently near the nominal level of the test.

3.1.2 Quantifying Loss for One-Sided Tests of a Binomial Parameter

Since we have derived a formula for the FEMPL in one-sided significance tests (Section 2.2.1), we might also be interested in describing this quantity for this example. For a significance threshold α^*, the FEMPL is

$$\begin{aligned}\mathcal{R}(\theta, \alpha^*) &= \mathbb{P}(P < \alpha^*)E[P|P < \alpha^*] + \mathbb{P}(P > \alpha^*)\alpha^* \\ &= \alpha^* - \mathbb{P}(P < \alpha^*)E[\alpha^* - P|P < \alpha^*] \\ &= \alpha^*\Big(1 - \frac{\mathbb{P}(P < \alpha^*)}{\alpha^*}E[\alpha^* - P|P < \alpha^*]\Big),\end{aligned}$$

where, since we will approach this from the Bayesian perspective and assume a sample large enough to enable interpretation of posterior distribution tails as p-values, P is the left tail area of the posterior distribution:

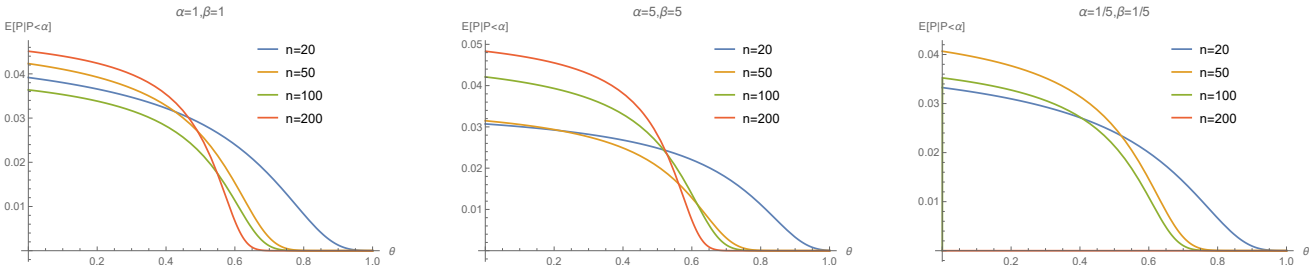
$$P = \int_0^{1/2} f(\theta|X)d\theta$$

$$= \int_0^{1/2} \frac{1}{B(\alpha + T, \beta + n - T)} \theta^{\alpha+T-1}(1-\theta)^{\beta+n-T-1} d\theta,$$

where B is the Beta function.

In Section 2.2.1, we were motivated to ask about the respective contributions of both of the dominating terms in this expression: the rejection ratio $\frac{\mathbb{P}(P<\alpha^*)}{\alpha^*}$ and the scaled expected p-value $E[\alpha^* - P|P < \alpha^*])$. We will now answer this question by examining each term in detail in our binomial test setting.

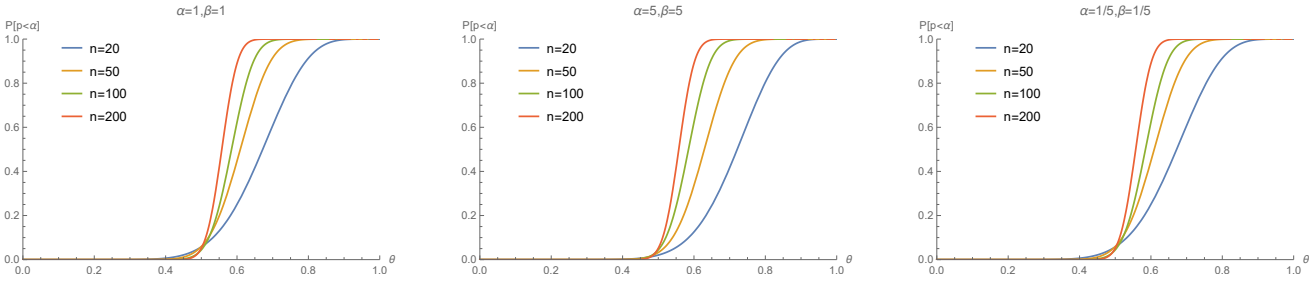
Figure 3.1.3 shows the prior distributions for θ implied by the three combinations of α and β we will consider. In Figure 3.1.5, we show the conditional expectation $E[P|P < \alpha^*]$ for each of the three combinations at a variety of sample sizes. We can see the general behavior that the conditional expected p-value is near the significance threshold when θ is very small but then decreases as θ grows larger, with the decline occurring more quickly for values of θ greater than the null.



(a) $\alpha = \beta = 1$ **(b)** $\alpha = \beta = 5$ **(c)** $\alpha = \beta = \frac{1}{5}$

Figure 3.1.5: Conditional expected p-values for a beta-binomial model test

The other major component of the risk as we derived it is what Bayarri et al. have called a *rejection ratio*: the ratio of the power of the test to the significance threshold. Figure 3.1.6 depicts the numerator of this quantity, $\mathbb{P}[p < \alpha^*]$, as a function of θ and n for the same three prior distributions considered above.

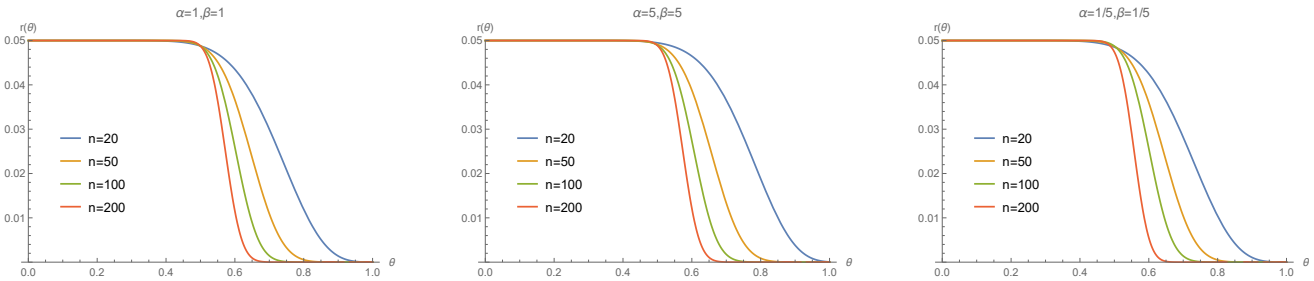


(a) $\alpha = \beta = 1$ **(b)** $\alpha = \beta = 5$ **(c)** $\alpha = \beta = \frac{1}{5}$

Figure 3.1.6: $\mathbb{P}[P < \alpha]$ **for a beta-binomial model test**

In comparing Figures 3.1.5 and 3.1.6, we can see that the conditional expected p-value and the rejection ratio are indeed providing unique information and both are relevant to our computation of the overall risk for this scenario. While the conditional expected p-value generally tails off from the significance level α toward zero as θ increases, the rejection ratio increases from zero to $\frac{1}{\alpha}$. As we see above, sample size does play a role in determining the rate of change of these quantities as θ increases.

Finally, we will conclude this example by showing the full FEMPL for this example, which is roughly a product of these two terms (Figure 3.1.7).



(a) $\alpha = \beta = 1$ **(b)** $\alpha = \beta = 5$ **(c)** $\alpha = \beta = \frac{1}{5}$

Figure 3.1.7: Minimized posterior expected loss for a beta-binomial model test

We notice that the form of the prior (that is, the values of α and β, make little difference to the general shape of this function. In all cases, the expected loss is maintained at the nominal level for values of θ in the space of the null hypothesis ($\theta \leq \frac{1}{2}$), and then drops to zero for large values of θ. For larger samples, this decrease is sharper. In other words, larger samples result in a test which can reject the null

for small effect sizes. This is, of course, consistent with our intuition about how tests should perform.

It is interesting that the risk can be shown to be a function of two quantities, a rejection ratio and a conditional version of the expected p-value, which have both been previously proposed as alternative measures of the strength of a testing procedure. This suggests that these quantities could be important measures of a test both in isolation as well as in tandem. In settings where one can derive an expression for the minimized posterior expected loss, it might be advisable that researchers present this quantity as well. In particular, the behavior of this function for values of θ corresponding to the alternative hypothesis could be valuable for helping researchers understand how much loss they can expect to incur for various effect sizes.

3.1.3 Two-Sided Tests

We can also consider a two-sided test of the binomial parameter θ and arrive at a similar comparison between a frequentist, exact testing approach and a Bayesian model approach. We again frame this as a significance test, and therefore we plan to either decide "Below" (that θ is below $\frac{1}{2}$), make "No Decision", or decide "Above" (that θ is above $\frac{1}{2}$). Without loss of generality, we will choose the loss functions for each of the possible decisions according to Table 3.1.2.

Table 3.1.2: Decision Framework - Two-Sided Binomial Significance Test

	Decision		
	d=Below	d=No Decision	d=Above
Loss when $\theta \leq \frac{1}{2}$	0	α	2
$\theta > \frac{1}{2}$	2	α	0

Again we will use the notation $\sum_{i=1}^{n} X_i = T$ and note that $T \sim \text{Bin}(n, \theta)$ (the sum of n independent and identically distributed Bernoulli random variables is a binomial random variable).

For this two-sided test, the two-sided p-value is

$$P = 2[\min(P_{\text{lower}}, P_{\text{upper}})],$$

where

$$P_{\text{lower}} = \sum_{t=0}^{t_{obs}} \binom{n}{t} \theta^t (1-\theta)^{n-t}$$

and

$$P_{\text{upper}} = \sum_{t=t_{obs}}^{n} \binom{n}{t} \theta^t (1-\theta)^{n-t}.$$

Assuming a significance level α for the test, the probability that our test statistic T takes on some observed value t such that we would make "No Decision" (an action which we denote d_0) is

$$\mathbb{P}[d_0] = \sum_{t:P>\frac{\alpha}{2}} \binom{n}{t} \theta^t (1-\theta)^{n-t}.$$

Similarly, the probability of deciding "Above", an action we denote d_{1A}, is

$$\mathbb{P}[d_{1A}] = \sum_{t:P<\frac{\alpha}{2} \text{ and } t>n/2} \binom{n}{t} \theta^t (1-\theta)^{n-t},$$

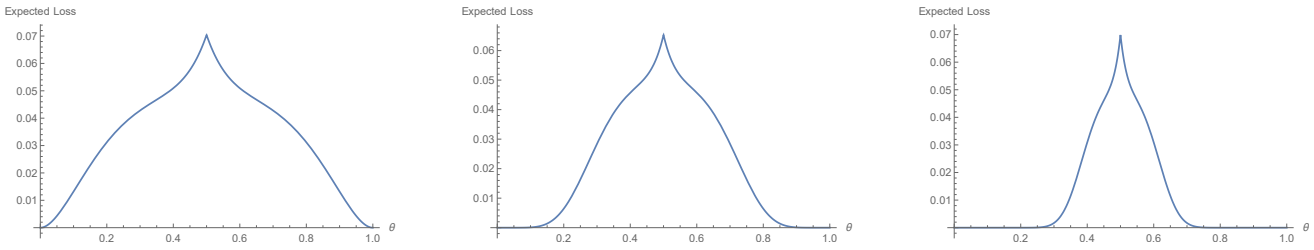
and the probability of deciding "Below", denoted d_{1B}, is

$$\mathbb{P}[d_{1B}] = \sum_{t:P<\frac{\alpha}{2} \text{ and } t<n/2} \binom{n}{t} \theta^t (1-\theta)^{n-t} * I_{P\leq\alpha}.$$

Finally, the expected loss for a decision is given by

$$\mathbb{E}[\text{Loss}] = I_{\theta>\frac{1}{2}}\Big(2\,\mathbb{P}[d_{1B}] + \mathbb{P}[d_0]\alpha\Big) + I_{\theta\leq\frac{1}{2}}\Big(2\,\mathbb{P}[d_{1A}] + \mathbb{P}[d_0]\alpha\Big).$$

We are now ready to examine the form of this expected loss function. In Figure 3.1.8, we show this expected loss for a level α test using three sample sizes.



(a) $n = 10$ **(b)** $n = 30$ **(c)** $n = 100$

Figure 3.1.8: Expected loss for a two-sided binomial exact significance test

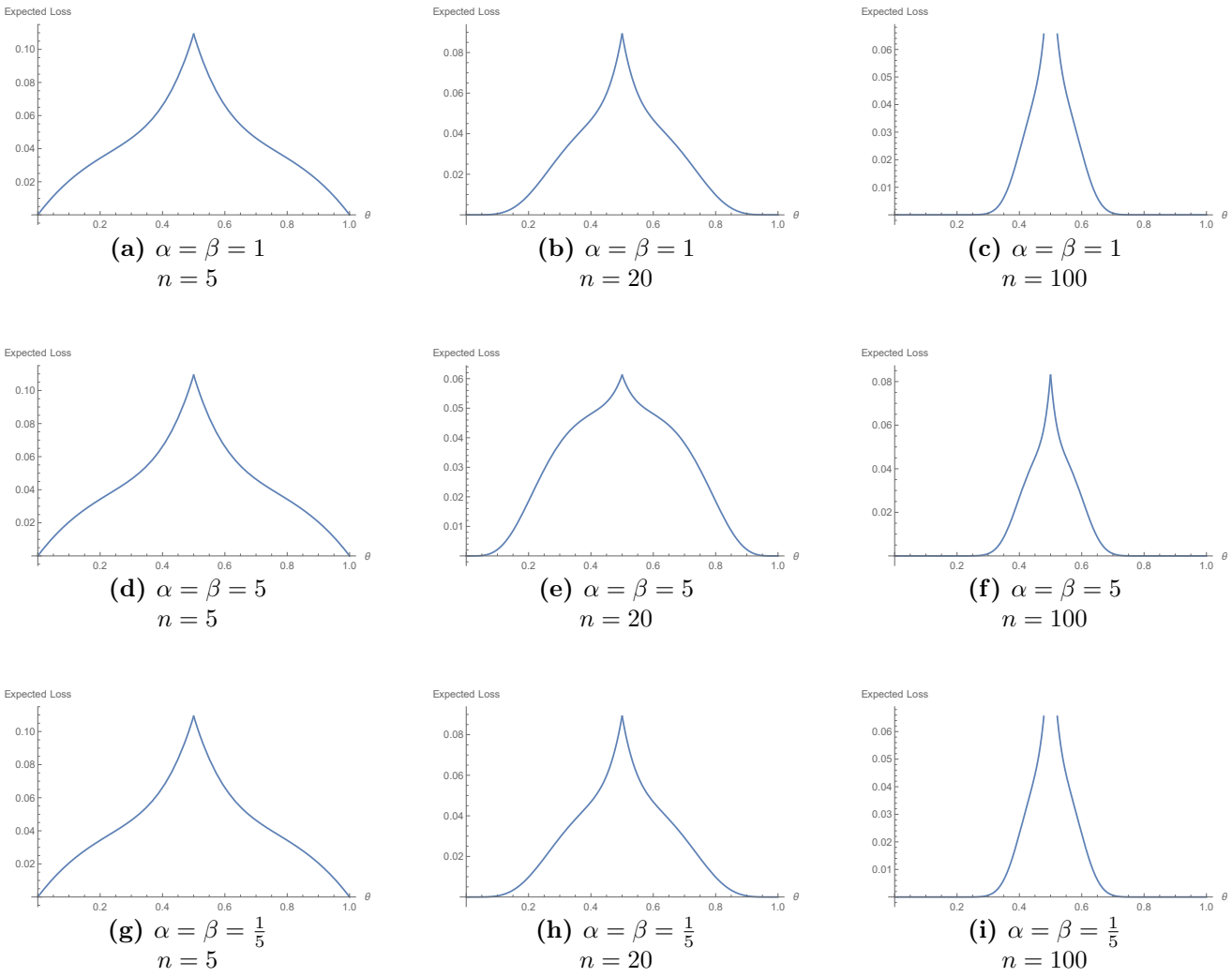
We find that the expected loss is low for extreme values of θ (that is, values near 0 and 1). This is sensible—at these values we should easily get the decision correct. For sample sizes near $n = 30$, we have almost zero expected loss for values of θ less than 0.10 of greater than 0.90. As before, in the one-sided example, we observe a spike in the expected loss near the null value. Previously this spike was only on the lower (null hypothesis) side of $\theta = \frac{1}{2}$, but in the two-sided case we see a noticeable spike in expected loss on both sides.

Again following the example set in the prior section, we will now turn to addressing this problem from a Bayesian perspective. To do so, we will again model the binomial probability using the beta distribution conjugate prior. Once again, we will call the parameters of this beta distribution α and β and we will switch to calling the significance threshold α^* for the remainder of this example.

In this two-sided approach, one possible outcome is that we will observe a large value of T (greater than $n/2$) and be motivated to decide "Above". More specifically, we will decide that $\theta > \frac{1}{2}$ for values of T such that $F_{(\theta|X)}(\frac{1}{2}) > 1-\alpha^*/2$, where F represents the cumulative distribution function of the posterior. Alternatively, we might observe a small value of T (less than $n/2$), in which case we will decide that $\theta < \frac{1}{2}$ if T is small enough that $F_{(\theta|X)}(\frac{1}{2}) < \alpha^*/2$. For middling values of T, we will reserve the right to make no decision.

Our setup of the loss functions for the Bayesian version of this test will mimic the frequentist setup. We will associate a constant loss α with making no decision. When $\theta > \frac{1}{2}$, we will either incur a loss of 2 if we incorrectly decide $\theta < \frac{1}{2}$ or incur no loss if we make the correct decision. Otherwise, when $\theta < \frac{1}{2}$, we will either incur a loss of 2 if we incorrectly decide $\theta > \frac{1}{2}$ or incur no loss if we make the correct decision.

Once again, the form of the prior (dictated by our chosen values for the beta distribution parameters α and β) plays a role in these decisions. Figure 3.1.3 shows the three types of priors we will consider for this two-sided test. For each of these prior distributions, we will plot the expected loss (as a function of θ) for three sample sizes: $n = 5$, 20, and 100 (Figure 3.1.9).



(a) $\alpha = \beta = 1$, $n = 5$ **(b)** $\alpha = \beta = 1$, $n = 20$ **(c)** $\alpha = \beta = 1$, $n = 100$

(d) $\alpha = \beta = 5$, $n = 5$ **(e)** $\alpha = \beta = 5$, $n = 20$ **(f)** $\alpha = \beta = 5$, $n = 100$

(g) $\alpha = \beta = \frac{1}{5}$, $n = 5$ **(h)** $\alpha = \beta = \frac{1}{5}$, $n = 20$ **(i)** $\alpha = \beta = \frac{1}{5}$, $n = 100$

Figure 3.1.9: Expected loss for a two-sided Bayesian beta-binomial decision model using various priors and sample sizes

Unsurprisingly, due to our specification of the loss functions as the same in both the exact test described previously and this beta-binomial model, the loss functions for the Bayesian version of this two-sided test are comparable to those from the exact test. Just as we observed in the frequentist exact test version (Figure 3.1.8), we see that increasing sample sizes results in a sharper decline toward zero expected loss on both sides of $\theta = \frac{1}{2}$ (Figure 3.1.9). We also see that in this example that smaller samples correspond to a larger spike in the expected loss near $\theta = \frac{1}{2}$.

The lesson of this example is partly that we can use our decision-theoretic framework in such a way that

the expected loss encountered in the frequentist version of the test has the same form as that encountered in the Bayesian version of the test. We note that the Bayesian approach is predictably somewhat more flexible, as is enables us to achieve different shapes of the expected loss function depending on our choices of the prior parameters α and β.

Thus far, we have pointed out the spike in the expected risk which occurs in the significance testing paradigm in both one-sided and two-sided tests but we have not commented on its implications for testing. Interestingly, the spike in the expected loss indicates that for values of θ such that the expected loss is greater than the significance threshold α, which occur when θ is near the null value (i.e. for small effect sizes), we would have been "better off" (in terms of having the lowest possible expected loss) not doing any testing at all. In other words, testing is futile for values of θ close to the null hypothesized value—if we had not done any testing whatsoever then we could have made no decision and incurred a loss α (equal to the significance threshold) which is less than the expected loss when actually running the test for the same true value of θ. An interesting question we might ask is: how close does the true value of θ have to be to the null hypothesized value in order for this to be the case? To make precise to the definition of closeness, we will turn to a new example of significance testing in which the test statistic is continuous rather than discrete. In this new setting not only will we see the same spike in expected value near the null value, but we will precisely calculate the "break-even" point (a value of θ) at which significance testing becomes futile. We will then go on to comment on the power of the test at these break-even points.

3.2 A Normal Location Testing Setting

In our first example, the test statistic of interest was discrete. We will now turn to a familiar testing scenario where the test statistic is continuous. We intend to perform a two-sided significance test of a population mean, denoted θ, with the null hypothesis being $H_0 : \theta = 0$ and the alternative $H_1 : \theta \neq 0$. Once again, we plan to approach this problem from two perspectives: frequentist and Bayesian. In the frequentist paradigm, we will test these hypotheses using a significance test version of the Wald test. For the Bayesian approach, we will again work in a conjugate setting. Just as in the previous example where we tested a binomial probability, our eventual goals will be to describe the expected loss function

for both the frequentist and Bayesian procedures and to make comparisons between the risk we face from both approaches.

We will begin with the frequentist testing approach. Given data X_i; $i = 1, \ldots, n$, our test statistic will be the sample mean $\bar{X}$. The central limit theorem provides the convenient results that the sample mean is normally distributed:

$$\bar{X} \sim N(\theta, \frac{\sigma^2}{n}),$$

where θ is the population mean and σ^2 the population variance. This allows us to define a threshold value for the sample mean beyond which (in absolute value) we will reject the null hypothesis that the population mean is zero. For $Z_{\alpha/2}$ the $\alpha/2^{th}$ quantile of the standard normal distribution, the threshold value of the sample mean is $\bar{X}_{thresh} = Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$. The probability that the sample mean is between $-\bar{X}_{thresh}$ and $\bar{X}_{thresh}$ is given by

$$\mathbb{P}[\text{No Decision}] = \Phi_{(\theta, \frac{\sigma}{\sqrt{n}})}(\bar{X}_{thresh}) - \Phi_{(\theta, \frac{\sigma}{\sqrt{n}})}(-\bar{X}_{thresh}),$$

where $\Phi_{(\theta, \frac{\sigma}{\sqrt{n}})}(x)$ is the cumulative distribution function of a $N(\theta, \frac{\sigma}{\sqrt{n}})$ random variable evaluated at x. As we have done in all of the two-sided significance tests encountered thus far we will associate loss α with making no decision and therefore the expected loss when making no decision is

$$\mathbb{E}[\text{Loss}]_{\text{No Decision}} = \alpha\Big[\Phi_{(\theta, \frac{\sigma}{\sqrt{n}})}(\bar{X}_{thresh}) - \Phi_{(\theta, \frac{\sigma}{\sqrt{n}})}(-\bar{X}_{thresh})\Big].$$

The expected loss when deciding d=Above and d=Below are derived in the same way. When deciding d=Above, we will either be correct and incur zero loss or be incorrect and incur loss 2. Similarly, when deciding d=Below we will either be correct and incur zero loss or be incorrect and incur loss 2. This corresponds to an expected loss for making a decision equal to

$$\mathbb{E}[\text{Loss}]_{\text{Decision}} = I_{\theta \leq 0} \times 2\Big[\sigma - \Phi_{(\theta, \frac{\sigma}{\sqrt{n}})}(\bar{X}_{thresh})\Big] + I_{\theta > 0} \times 2\Phi_{(\theta, \frac{\sigma}{\sqrt{n}})}(-\bar{X}_{thresh})\Big].$$

Finally, the expected loss for the testing procedure overall is the sum of the two expected losses above:

$$\mathbb{E}[\text{Loss}] = \alpha\Big[\Phi_{(\theta, \frac{\sigma}{\sqrt{n}})}(\bar{X}_{thresh}) - \Phi_{(\theta, \frac{\sigma}{\sqrt{n}})}(-\bar{X}_{thresh})\Big] + I_{\theta \leq 0} \times 2\Big[1 - \Phi_{(\theta, \frac{\sigma}{\sqrt{n}})}(\bar{X}_{thresh})\Big] + I_{\theta > 0} \times 2\Phi_{(\theta, \frac{\sigma}{\sqrt{n}})}(-\bar{X}_{thresh})\Big].$$

As we can see, this expected loss is a function of n and θ. We plot this function for $n = 1, 100,$ and 10000

in Figure 3.2.1.

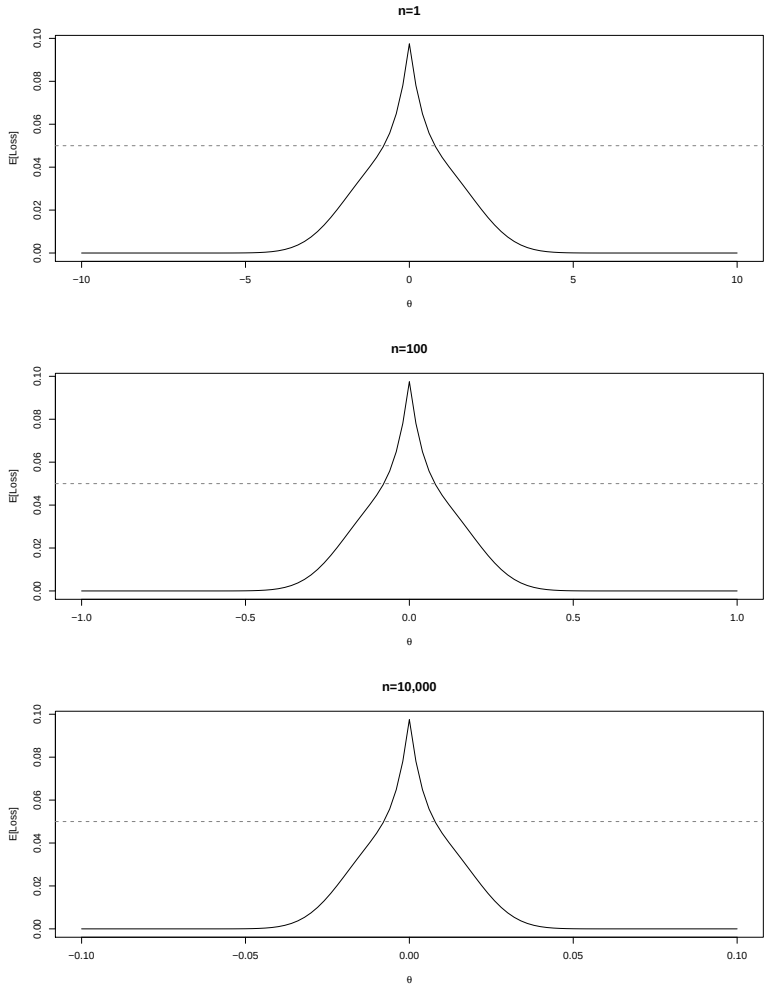


Figure 3.2.1: Expected loss for a classical Wald significance test. Dotted line drawn at level α.

We see that the expected loss in this significance test displays the same spike around the null value which we observed previously in the context of a test about a binomial probability. Moreover, noting the differing scales of the Y-axis in Figure 3.2.1, we can see that the width of this spike depends on the sample size: larger samples correspond to a thinner spike which indicates the comparatively smaller risk we face when we test the parameter in a much larger sample. We also note that for some values of θ, regardless of the sample size, the expected loss exceeds the nominal rate of the test (in this case, $\alpha = 0.05$). This is the same behavior we noted in the binomial tests of Section 3.1, where we commented

that the result indicates that test is futile for certain combinations of sample size and true effect size. That is, in certain cases we would be better off not doing any test, making no decision, and incurring the constant loss α.

Numerical root-finding software, available in R [70], Mathematica [71], and other statistics or mathematics programming software, can be used to determine the precise value of θ, for a given sample size, at which the expected loss drops below α. When n=1, for example (Figure 3.2.1, top), the break-even point is $\theta = 0.781$. When $|\theta|$ is greater than this value, we can enjoy lower expected loss by performing the test. However, when $|\theta| < 0.781$, significance testing is no longer worthwhile. Continuing our examination of this phenomenon, we note that when $n = 1$ the power of the test to detect a true difference of 0.781 is 12.2%. That is, significance testing at level 0.05 is worthwhile only so long as we can guarantee that the power of the test is 12.2%. In practice, 80% power is often viewed as a standard goal. This is fortunate—it indicates that in reasonable settings we would, in fact, be better off basing our eventual decision on the result of this significance test than opting to make no decision a priori.

Our next goal for this example will be to describe the expected loss for a Bayesian analog to the classical Wald test. The Bayesian approach to this test which we will consider is that of a normal-normal conjugate model. That is, we will specify a normally distributed prior for θ, the parameter of interest:

$$\theta \sim N(\theta_0, \tau^2),$$

where for now we will assume τ^2 is known. The clone parameter θ^* will be independently and identically distributed (i.e. $\theta^* \sim N(\theta_0, \tau^2)$). Assume that the data are also normally distributed so that

$$X_i \sim_{iid} N(\theta, \sigma^2) \qquad \forall \;\; i = 1, \ldots, n,$$

where σ^2 is considered known. This conjugate setting has been studied in great detail. The posterior distribution can be seen to have the following form:

$$(\theta|\bar{X}) \sim N\Big(\frac{\tau^2}{\tau^2 + \frac{\sigma^2}{n}}\bar{X} + \frac{\frac{\sigma^2}{n}}{\tau^2 + \frac{\sigma^2}{n}}\theta_0, \;\; \Big(\frac{1}{\tau^2} + \frac{n}{\sigma^2}\Big)^{-1}\Big).$$

For notational convenience later, we will also write the mean of the above distribution as a weighted

average of the sample mean and the prior distribution mean by writing

$$W = \frac{\tau^2}{\tau^2 + \frac{\sigma^2}{n}},$$

which in turn implies

$$1 - W = \frac{\frac{\sigma^2}{n}}{\tau^2 + \frac{\sigma^2}{n}},$$

and results in the following, slightly more convenient expression for the posterior distribution:

$$(\theta|X) \sim N\Big(W\bar{X} + (1-W)\theta_0, \quad \Big(\frac{1}{\tau^2} + \frac{n}{\sigma^2}\Big)^{-1}\Big).$$

We can use this form of the posterior distribution to determine the value of $P = \mathbb{P}(\theta < 0)$. By first standardizing the above distribution, this is straightforwardly seen to be

$$P = \Phi\Big(- \Big(W\bar{X} + (1-W)\theta_0\Big) \Big/ \Big(\frac{1}{\tau^2} + \frac{n}{\sigma^2}\Big)^{-1/2}\Big),$$

where Φ represents the cumulative distribution function of the standard normal distribution.

Similarly to our approach for the classical Wald test, we will describe the losses for this test by first noting the cutoff values of sample mean which dictate which decision we will make. Since this is a two-sided test, we will decide d=Below if the left tail area of the posterior distribution for θ (left of $\theta = 0$) is less than $\frac{\alpha}{2}$. That is, we will make this decision if and only if

$$\begin{aligned}
&\Phi_{(W\bar{X}+(1-W)\theta_0, (\frac{1}{\tau^2}+\frac{n}{\sigma^2})^{-1})}(0) < \frac{\alpha}{2} \\
\implies &\Phi_{(0,1)}\Big(\frac{-(W\bar{X} + (1-W)\theta_0)}{(\frac{1}{\tau^2}+\frac{n}{\sigma^2})^{-1}}\Big) < \frac{\alpha}{2} \\
\implies &\frac{-(W\bar{X} + (1-W)\theta_0)}{(\frac{1}{\tau^2}+\frac{n}{\sigma^2})^{-1}} < Z_{\alpha/2} \\
\implies &W\bar{X} + (1-W)\theta_0 > (\frac{1}{\tau^2} + \frac{n}{\sigma^2})^{-1} Z_{1-\alpha/2} \\
\implies &\bar{X} > \frac{\Big((\frac{1}{\tau^2}+\frac{n}{\sigma^2})^{-1} Z_{1-\alpha/2}\Big) - (1-W)\theta_0}{W} = \bar{X}_{\text{high threshold}}.
\end{aligned}$$

Using a nearly identical derivation, the sample mean threshold for making the decision d=Above is

$$\bar{X} < \frac{\Big((\frac{1}{\tau^2}+\frac{n}{\sigma^2})^{-1} Z_{\alpha/2}\Big) - (1-W)\theta_0}{W} = \bar{X}_{\text{low threshold}}.$$

Exactly as we did when setting up the classical Wald significance test, we will imagine that we incur loss α when making no decision, an event which occurs with probability

$$\Phi_{(\theta,\frac{1}{\sqrt{n}})}(\bar{X}_{\text{high threshold}}) - \Phi_{(\theta,\frac{1}{\sqrt{n}})}(\bar{X}_{\text{low threshold}}).$$

Next, if $\theta \leq 0$, we will either incur no loss (if we get the decision correct) or a loss of 2 (if $\bar{X} > \bar{X}_{\text{high threshold}}$). Finally, $\theta > 0$, we will either incur no loss (if we get the decision correct) or a loss of 2 (if $\bar{X} < \bar{X}_{\text{low threshold}}$). Altogether, we find that an expression for the expected loss in this setting is

$$\begin{aligned}\mathbb{E}[\text{Loss}] =& \alpha\left[\Phi_{(\theta,\frac{1}{\sqrt{n}})}(\bar{X}_{\text{high threshold}}) - \Phi_{(\theta,\frac{1}{\sqrt{n}})}(\bar{X}_{\text{low threshold}})\right] \\ &+ I_{\theta\leq 0} \times 2\left[1 - \Phi_{(\theta,\frac{1}{\sqrt{n}})}(\bar{X}_{\text{high threshold}})\right] \\ &+ I_{\theta>0} \times 2\left[\Phi_{(\theta,\frac{1}{\sqrt{n}})}(\bar{X}_{\text{low threshold}})\right].\end{aligned}$$

The expected loss for this Bayesian approach to the classical Wald test is shown in Figure 3.2.2 for sample sizes of $n = 10$, 25, and 200. As we saw previously, the spike in risk that we encounter near the null value of θ becomes sharper as the sample sizes increase. As we expected, the expected loss for this test can be directly compared to the loss incurred in the Wald test. In Figure 3.2.2, we also mark a line show the maximum expected loss for each test. This maximum occurs at $\theta = \theta_0$, the null value, and is equal to $\alpha(2 - \alpha)$.

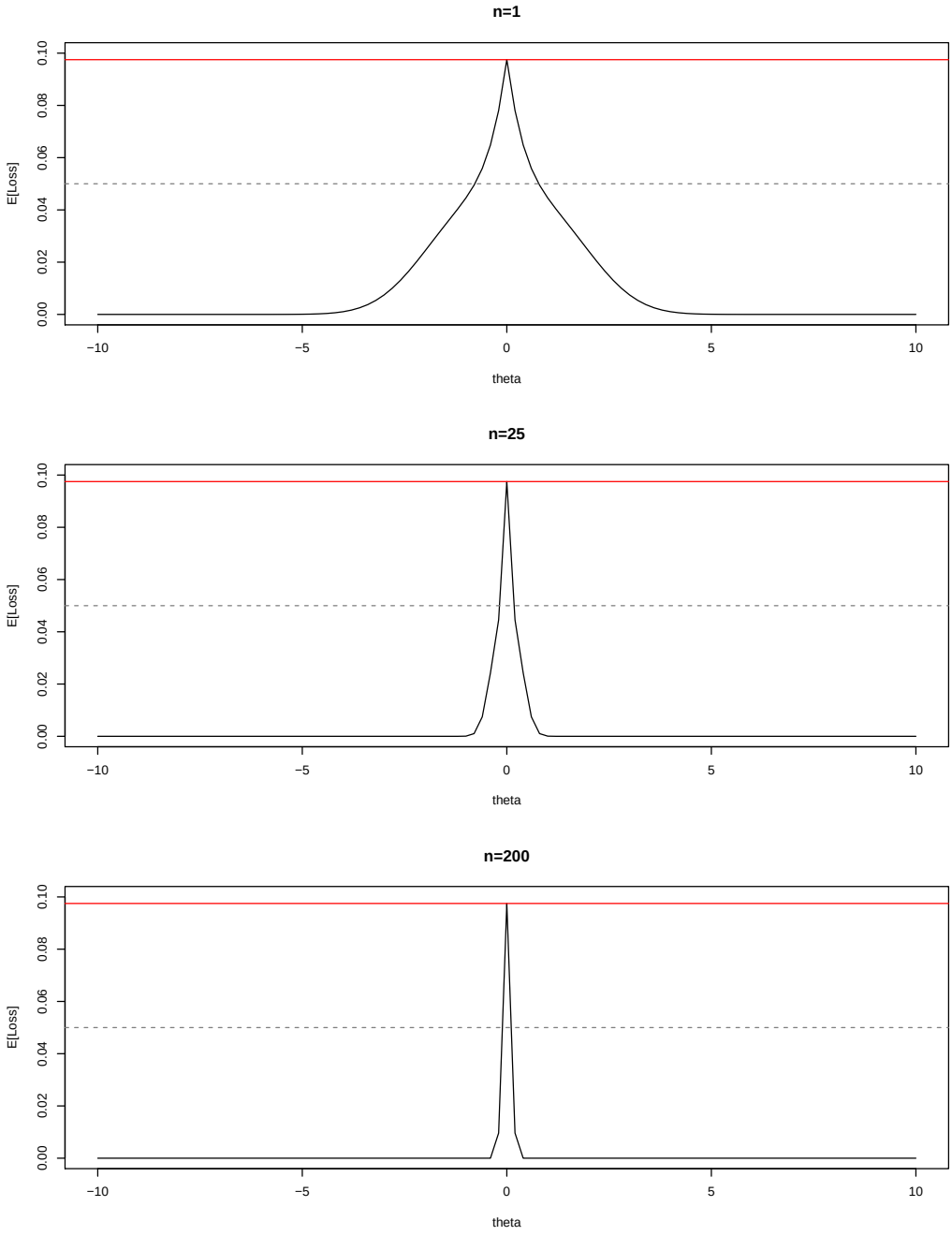


Figure 3.2.2: Expected loss for a Bayesian analog to the classical Wald test. Dotted line drawn at level α and red line drawn at $\alpha(2-\alpha)$.

We will also comment on the role of the prior distribution in determining the shape of the expected loss seen in Figure 3.2.2. In Figure 3.2.3, we show the expected loss for four types of normal prior

distributions by varying the standard deviation, τ, of those distributions. In all cases, we use $n = 1$ for convenience and fix the standard deviation of the likelihood, σ, at $\sigma = 1$. We will use $\theta_0 = 0$ as the mean of the prior distribution for θ in each case and imagine that we intend to test at level $\alpha = 0.05$. The expected loss for each scenario is shown in green. The density of the prior distribution for a given θ is displayed in blue.

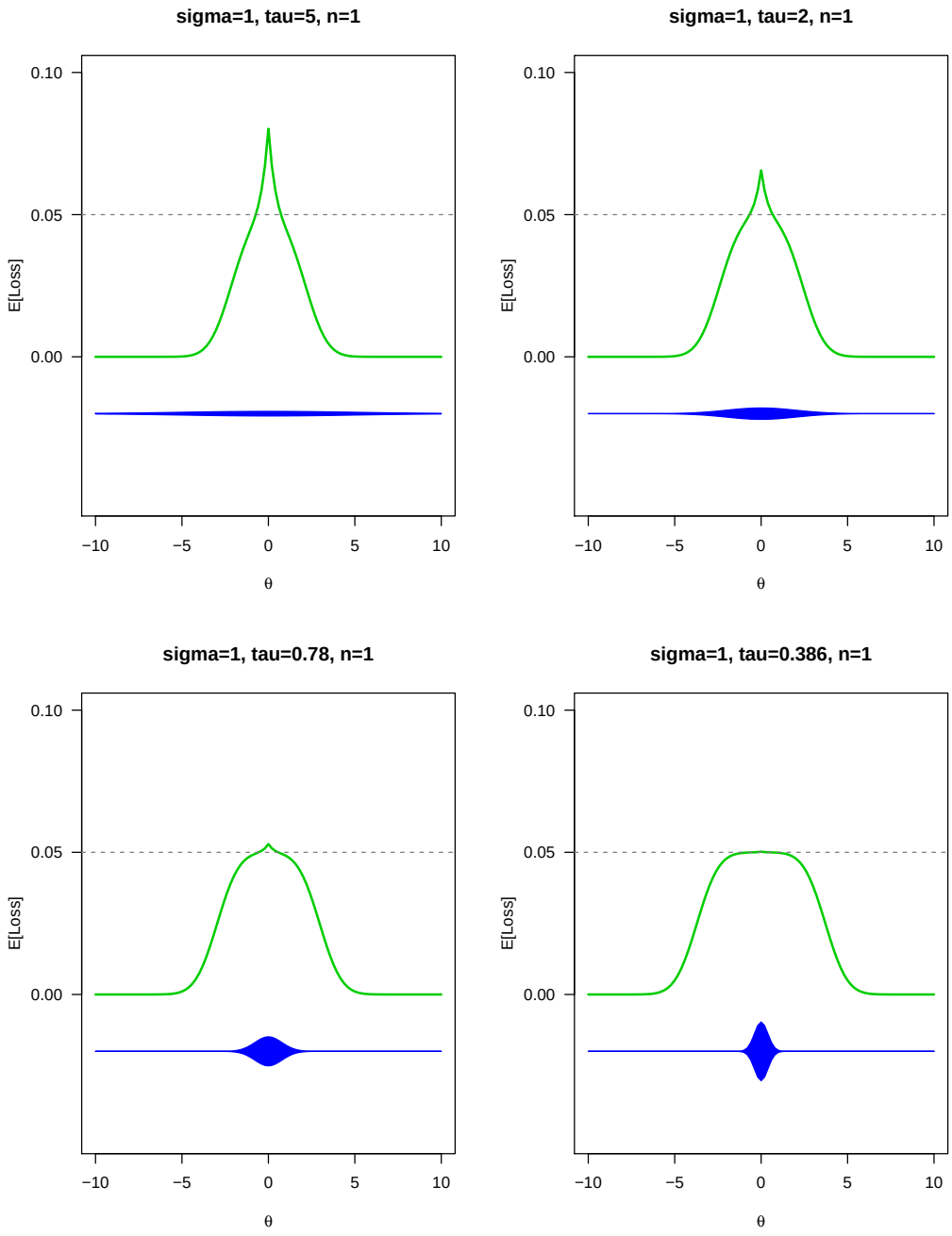


Figure 3.2.3: Expected losses (green) and prior distributions (blue) as a function of τ

Figure 3.2.3 motivates a discussion of how the shape of the prior distribution dictates the risk we face in this example. Researchers frequently assume that flat priors which spread the density more evenly on reasonable values of θ are somewhat less risky since they seem to assume less specific knowledge

about θ. However, we see in Figure 3.2.3 that flatter prior distributions (those with higher values of τ), actually produce more prominent spikes in the expected loss we face. On the other hand, when using prior distributions that place more density immediately surrounding θ_0, we see that the expected loss flattens out. In other words, the significance testing framework we have described goes against the intuition that flatter priors are a less risky specification. Although the Bayes risk in these scenarios can be seen to be less than the significance level α, the frequentist expected loss can be higher than α in a neighborhood near θ.

3.2.1 Quantifying Loss for Tests of a Normal Location Parameter

As we did in the binomial probability example, we also wish to define the frequentist expected minimized posterior loss (FEMPL) for this setting and examine how it changes as a function of θ and n. In this section, since there is no other prior parameter traditionally called α, we will switch back to the usual notation of letting α denote the significance threshold. Since our binomial probability example was in the context of a one-sided test, we will examine the FEMPL for this normal location setting in a one-sided testing context as well. The expression we previously defined for a one-sided significance test was

$$\begin{aligned}\mathcal{R}(\theta,\alpha) &= \mathbb{P}(P<\alpha)E[P|P<\alpha] + \mathbb{P}(P>\alpha)\alpha \\ &= \alpha - \mathbb{P}(P<\alpha)E[\alpha - P|P<\alpha] \\ &= \alpha\Big(1 - \frac{\mathbb{P}(P<\alpha)}{\alpha}E[\alpha - P|P<\alpha]\Big).\end{aligned}$$

We will break this expression into pieces and evaluate each piece separately in much the same way as we did in Section 3.1.2. We begin by deriving an expression for the probability of a significant p-value, $\mathbb{P}(P<\alpha^*)$, where P is the left tail area of the posterior distribution $(\theta|\bar{X})$:

$$P = \Phi\Big(-\Big(\frac{\tau^2}{\tau^2+\frac{\sigma^2}{n}}\bar{X} + \frac{\frac{\sigma^2}{n}}{\tau^2+\frac{\sigma^2}{n}}\theta_0\Big)\Big/\Big(\frac{1}{\tau^2}+\frac{n}{\sigma^2}\Big)^{-1/2}\Big).$$

Since P is a function of $\bar{X}$, we will plan to arrive at an expression for the desired probability (which is, up to a scaling factor of $\frac{1}{\alpha}$, the rejection ratio discussed by Bayarri et al.) by carefully rearranging

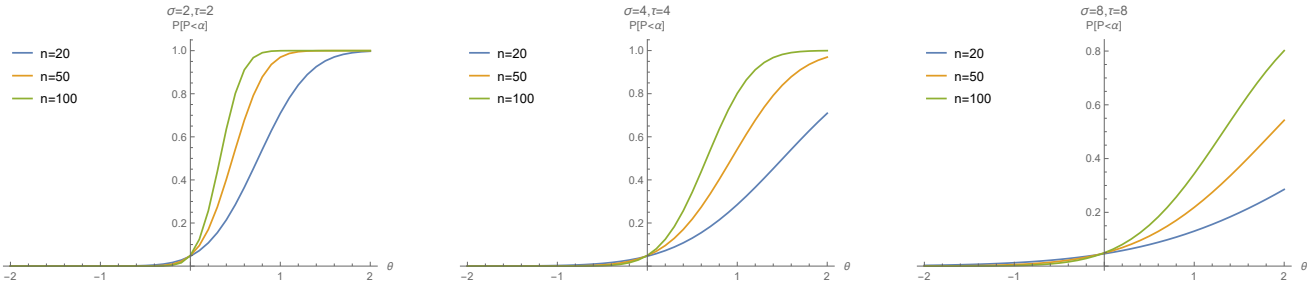
$\mathbb{P}(P < \alpha^*)$ into a probability statement about $\bar{X}$ and then invoke the central limit theorem. We have

$$\begin{aligned}
\mathbb{P}[P < \alpha^*] &= \mathbb{P}[\Phi\Big(-\Big(\frac{\tau^2}{\tau^2 + \frac{\sigma^2}{n}}\bar{X} + \frac{\sigma^2}{\tau^2 + \frac{\sigma^2}{n}}\theta_0\Big)\Big/\Big(\frac{1}{\tau^2} + \frac{n}{\sigma^2}\Big)^{-1/2}\Big) < \alpha^*] \\
&= \mathbb{P}[-\Big(\frac{\tau^2}{\tau^2 + \frac{\sigma^2}{n}}\bar{X} + \frac{\sigma^2}{\tau^2 + \frac{\sigma^2}{n}}\theta_0\Big)\Big/\Big(\frac{1}{\tau^2} + \frac{n}{\sigma^2}\Big)^{-1/2} < \Phi^{-1}(\alpha^*)] \\
&= \mathbb{P}[-\Big(\frac{\tau^2}{\tau^2 + \frac{\sigma^2}{n}}\bar{X} + \frac{\sigma^2}{\tau^2 + \frac{\sigma^2}{n}}\theta_0\Big) < \Phi^{-1}(\alpha^*)\Big(\frac{1}{\tau^2} + \frac{n}{\sigma^2}\Big)^{-1/2}] \\
&= \mathbb{P}[-\tau^2\bar{X} < (\tau^2 + \frac{\sigma^2}{n})\Big[\Phi^{-1}(\alpha^*)\Big(\frac{1}{\tau^2} + \frac{n}{\sigma^2}\Big)^{-1/2} + \frac{\sigma^2}{\tau^2 + \frac{\sigma^2}{n}}\theta_0\Big]] \\
&= \mathbb{P}[\bar{X} > -\frac{\tau^2 + \frac{\sigma^2}{n}}{\tau^2}\Big[\Phi^{-1}(\alpha^*)\Big(\frac{1}{\tau^2} + \frac{n}{\sigma^2}\Big)^{-1/2} + \frac{\sigma^2}{\tau^2 + \frac{\sigma^2}{n}}\theta_0\Big]].
\end{aligned}$$

Now, since the central limit theorem gives us the result that $\bar{X} \sim N(\theta, \frac{\sigma^2}{n})$, our expression for the power of the test (when the null is false) is

$$\begin{aligned}
\mathbb{P}[P < \alpha^*] &= \mathbb{P}[\bar{X} > -\frac{\tau^2 + \frac{\sigma^2}{n}}{\tau^2}\Big[\Phi^{-1}(\alpha^*)\Big(\frac{1}{\tau^2} + \frac{n}{\sigma^2}\Big)^{-1/2} + \frac{\sigma^2}{\tau^2 + \frac{\sigma^2}{n}}\theta_0\Big]] \\
&= 1 - \Phi\left(\frac{-\frac{\tau^2 + \frac{\sigma^2}{n}}{\tau^2}\Big[\Phi^{-1}(\alpha^*)\Big(\frac{1}{\tau^2} + \frac{n}{\sigma^2}\Big)^{-1/2} + \frac{\sigma^2}{\tau^2 + \frac{\sigma^2}{n}}\theta_0\Big] - \theta}{\sigma/\sqrt{n}}\right)
\end{aligned}$$

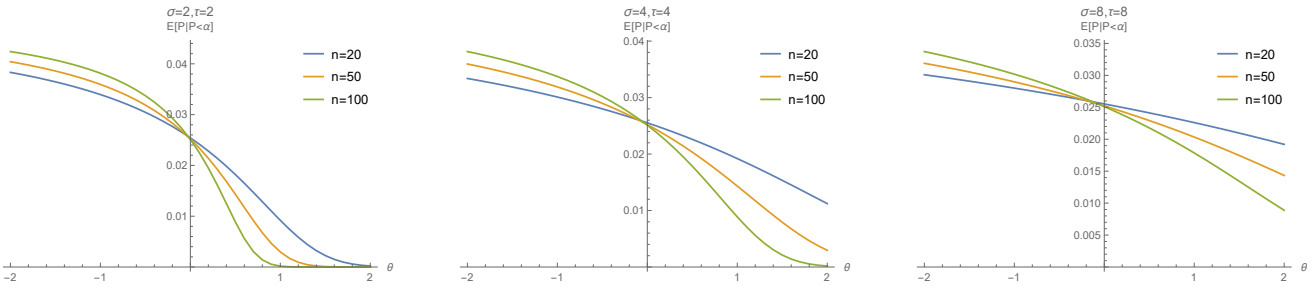
Figure 3.2.4 shows the above function for increasing values of the variances σ and τ. We see that in all cases, the probability of a significant p-value when the true mean θ is equal to the null (0, in this case) is 0.05. This behavior was, of course, determined by our choice of testing at level $\alpha = 0.05$. We can also observe the obvious benefit of larger sample sizes, which give us a higher probability of rejecting false null hypotheses. More importantly, we can see that the behavior of this function (the probability of a significant p-value, a function of θ) mimics the behavior we observed in our earlier tests of a binomial probability parameter (Figure 3.1.6).



(a) $\sigma = \tau = 2$ **(b)** $\sigma = \tau = 4$ **(c)** $\sigma = \tau = 8$

Figure 3.2.4: Probability of a significant p-value: one-sided significance test of the mean for increasing prior and likelihood variances

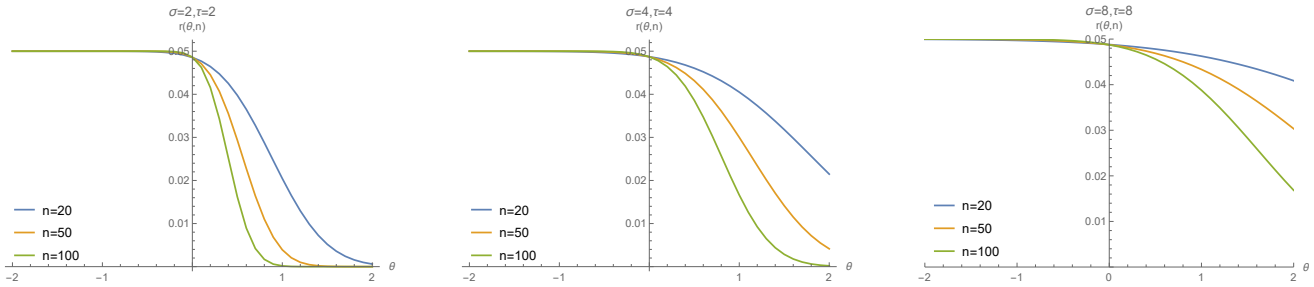
The other main component of the minimized posterior expected loss is the conditional expected p-value, $\mathbb{E}[P|P < \alpha]$ (Figure 3.2.5). We see that this quantity does indeed provide different information from the rejection ratio, which again suggests that both measures of evidence may be useful for describing the performance of a testing procedure.



(a) $\sigma = \tau = 2$ **(b)** $\sigma = \tau = 4$ **(c)** $\sigma = \tau = 8$

Figure 3.2.5: Conditional expected p-value: one-sided significance test of the mean for increasing prior and likelihood variance

Furthermore, we can observe that the posterior expected loss for this example (Figure 3.2.6) is maintained at level α for values of the true mean θ less than the null ($\theta = 0$, in this case) and then, beginning at θ near the null, steadily decreases toward zero. As would be expected, the risk decreases more sharply for larger sample sizes and less sharply when the variances of the prior distribution and likelihood increase.



(a) $\sigma = \tau = 2$ **(b)** $\sigma = \tau = 4$ **(c)** $\sigma = \tau = 8$

Figure 3.2.6: Bayes risk for a one-sided significance test of the mean for increasing prior and likelihood variance

4 Discussion

As we have pointed out, we expect that researchers will continue to have disagreements regarding the proper direction of the field of statistical testing. For many, the common misinterpretations and misuse of p-values makes them unpalatable. For others, the arguable subjectivity of Bayesian analyses is equally suspect. In this work, we have suggested that both approaches have something of use to offer statisticians. Rather than viewing the frequentist and Bayesian approaches as competing to be the most widely accepted philosophy, researchers should see them as tools—both of which should be included in the statistician's toolbox. Our work joins a growing body of literature suggesting that different tools are best suited to different jobs.

Our focus has been primarily on scenarios in which we sought to assess and summarize evidence strictly about the sign of an underlying parameter. This narrow focus was deliberate—working in this setting allowed us to easily describe all possible courses of action and simplified the motivations for testing. For example, instead of broadly describing the trade-off between Type I error rates and the power of a test, we were able to interpret these quantities in what they imply for the loss we incur when making incorrect decisions. We argue that this is a more accessible approach to testing than the traditional pedagogy and may help researchers think critically about the relative value of correct and incorrect decisions when addressing questions of interest that are specific to their areas of expertise.

We have also seen that this framework provides a convenient connection to quantities which others have

suggested may be valuable new ways of measuring evidence and evaluating the strength of tests. In fact, a notable strength of this decision-theoretic approach is its generalizability to incorporate or motivate the use of many common statistical quantities. Here, we have shown its use for describing the trade-offs implied by one- and two-sided p-values, Bayes factors, Bayarri et al's rejection ratio, and Sackrowitz and Samuel-Cahn's expected p-value. Although we have not shown it here, the framework can also be extended to motivate the use of confidence intervals, credible intervals, and adjustments for multiple testing including the Bonferroni correction and the Benjamini-Hochberg procedure we addressed briefly in Section 1.2.

Our goal has not been simply to provide an alternative to p-values. Indeed, in many decision problems we have demonstrated that p-values are the only reasonable course of action. The true goal of this work has been to develop an intentionally simple framework that can be generalized to suit the needs of individual researchers. We hope that this framework may provide some illumination of common statistical procedures and measures of evidence for experts and non-experts alike.

4.1 Extension: Expected Loss Bands for Handling Nuisance Parameters

We will conclude by highlighting a possible extension of the concepts discussed here which we feel may be of interest and could be a valuable avenue for future research.

In Section 3.2 we outlined a test for the mean of a normally distributed random variable. We assumed that the data were normally distributed:

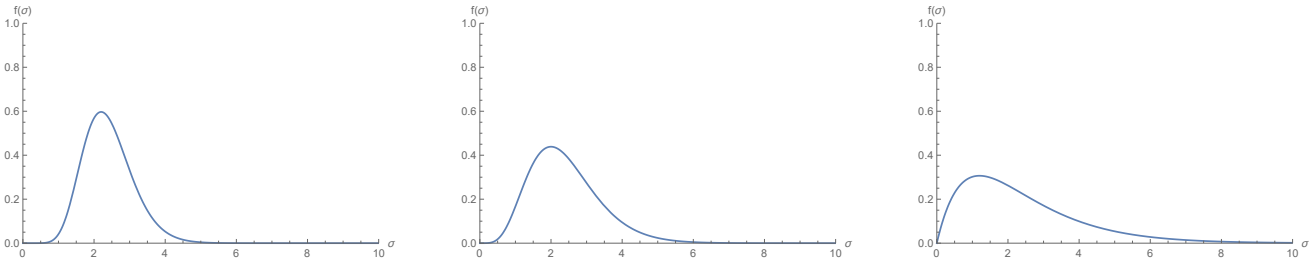
$$X_i \sim_{iid} N(\theta, \sigma^2) \qquad \forall \;\; i = 1, \ldots, n,$$

where the likelihood variance σ^2 was also considered known. In this setting, where our primary interest is in the mean θ but we must account for the variance σ^2 in our analysis, we refer to σ^2 as a *nuisance parameter*. When we encountered this nuisance parameter in Section 3.2, we partially handled the issue by simply assuming $\sigma = 1$ (Figure 3.2.3). Of course, it was unrealistic to imagine that we did not know the population mean, hence motivating our test of θ, but simultaneously maintain that we knew the population variance exactly. We now propose a more sophisticated method of handling this nuisance

parameter, which we will call an *expected loss band*, that could be implemented when we are unwilling to assume strong knowledge of the population variance.

Rather than assume a specific value for the variance, we will take a Bayesian approach and provide a prior distribution for σ, the standard deviation. When we move on to describing the expected loss as a function of θ, we will randomly draw n_σ values of σ from this prior distribution and then plot the expected loss function for each σ on a single plot. The resulting graph depicts a region of possible expected loss functions which depends on both the parameter of interest θ and the nuisance parameter σ, allowing us to get an idea of the role that both parameters play in our testing procedure at the same time.

Since the nuisance parameter we consider here is a standard deviation, which is bounded to be positive, we will specify priors with positive support. One reasonable prior which fits this criteria is the gamma distribution [4]. There are multiple parameterizations of the gamma distribution. We will choose to use the parameterization which incorporates a shape parameter, denoted k, and a scale parameter, denoted ξ. In this parameterization, the mean of the distribution is given by $k\xi$. Figure 4.1.1 shows the density of the distribution for three combinations of k and ξ which we will consider, all of which have mean 2.4 but which have varying amounts of spread.



(a) $k = 12$, $\xi = 0.2$ **(b)** $k = 6$, $\xi = 0.4$ **(c)** $k = 2$, $\xi = 1.2$

Figure 4.1.1: Three types of gamma prior distributions for σ

We will choose $n_\sigma = 100$, so that we will plot the expected loss in our normal conjugate model for 100 different random draws from each of the distributions in Figure 4.1.1.

In Figure 4.1.2, we plot the expected loss for the prior on σ described in Figure 4.1.1a in which $k = 12$ and $\xi = 0.2$. This prior represents the scenario in which we are most sure of the value of σ. For each of

the two plots in the figure, each line represents the expected loss function given a randomly generated value of σ from this prior. The result, which we will call an *expected loss band* for a given prior on σ, gives us an idea of the loss we can expect to incur and is more flexible than the previous depiction in which we simply assumed $\sigma = 1$. By randomly drawing values of σ from the prior, we can see a region of expected losses which could be considered reasonable. We also see the predictable impact of increasing the overall sample size n: when $n = 10$ we find that the width of the expected loss band decreases significantly.

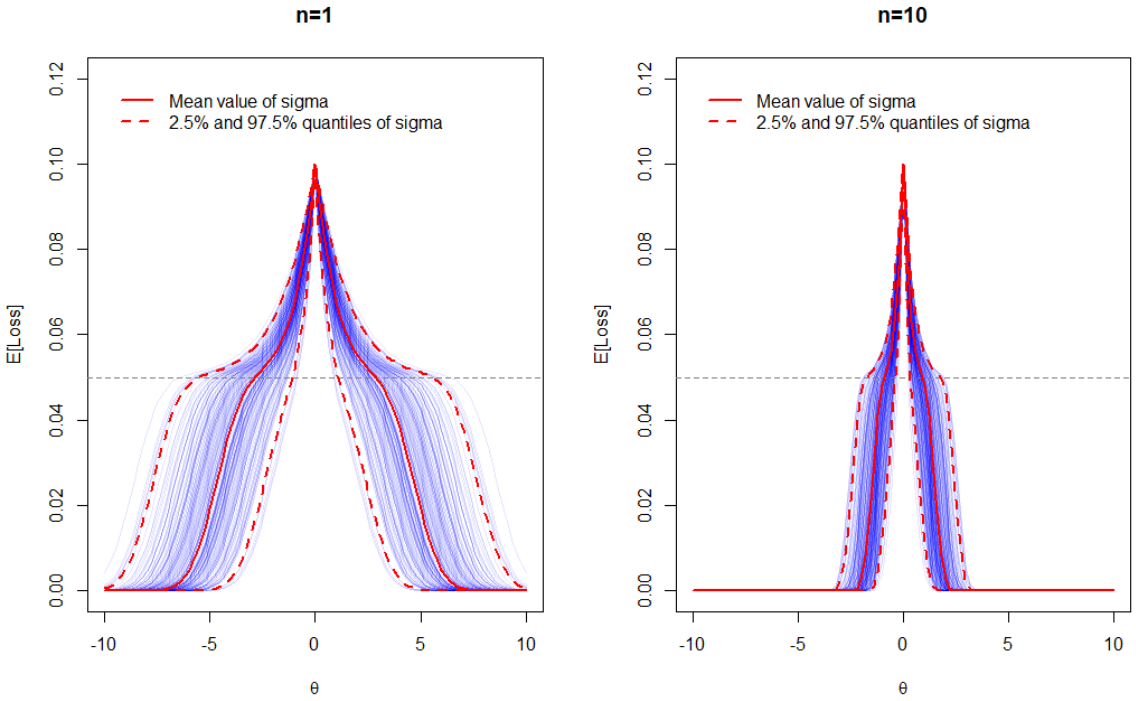


Figure 4.1.2: Expected loss band for two overall sample sizes given a prior on σ with $k = 12$ and $\xi = 0.2$

Next, in Figure 4.1.3, we show the expected loss band for the prior on σ described in Figure 4.1.1b. This prior describes a scenario in which we are slightly less knowledgeable about the true value of σ and thus is slightly more variable than the prior from Figure 4.1.1a. Once again, we can see a distinctive expected loss band. In this case, relative to the depiction in Figure 4.1.2, we see that the width of the expected loss band increases in accordance with the variability of the prior for σ.

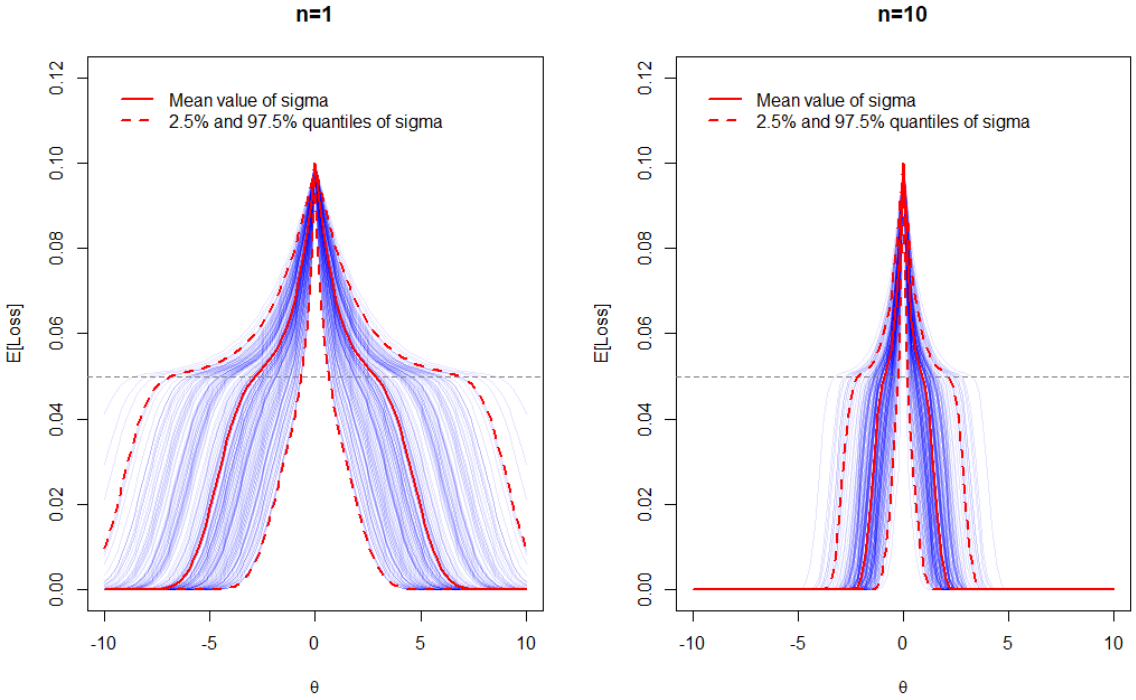


Figure 4.1.3: Expected loss band for two overall sample sizes given a prior on σ with $k = 6$ and $\xi = 0.4$

Finally, in Figure 4.1.4, we show the expected loss for the prior on σ described in Figure 4.1.1c, which describes the scenario where we are least knowledgeable about the true value of σ.

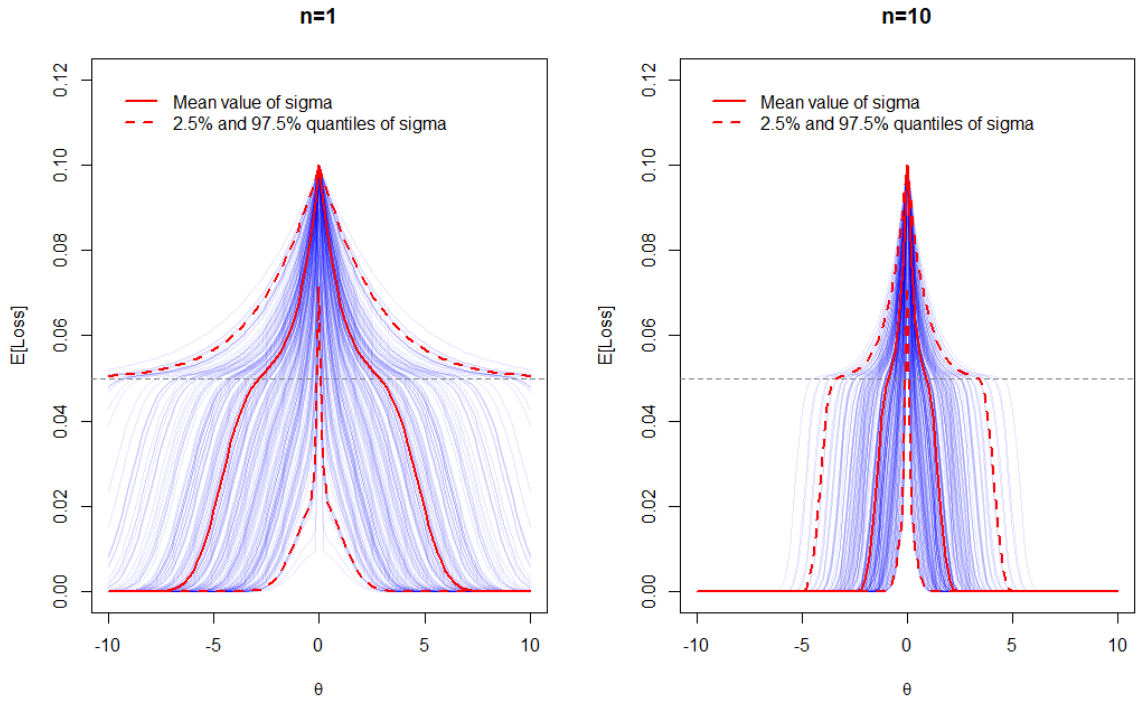


Figure 4.1.4: Expected loss band for two overall sample sizes given a prior on σ with $k = 2$ and $\xi = 1.2$

As we expected, we once again see that the width of the expected loss band increases as the variability of the prior increases. In this case, the expected loss band indicates that we have little reason to do significance testing at all: the expected loss is greater than the significance level for a wide range of true θ. That is, the more unsure we are about the true value of the nuisance parameter σ, the more likely it is that testing is futile and the more prudent it becomes to simply make no decision.

This Bayesian approach to evaluating the expected loss we expect to encounter when making decisions could be further generalized, and future research could be done to evaluate the behavior of the expected loss band in other settings. For instance, researchers might eventually be interested in determining the mean or median break-even point (the point on either side of θ_0 where the expected loss crosses the $\mathbb{E}[\text{Loss}] = \alpha$ threshold) suggested by the expected loss band as a means of determining whether testing is likely to be worthwhile. In general, it seems that this procedure could be a valuable way of evaluating when we stand to benefit from testing and when resources might be better spent elsewhere.

Finally, it also represents another way in which frequentists can stand to gain from adopting Bayesian approaches—although expected loss is a frequentist notion, a Bayesian take on expected loss may more accurately reflect researchers' usually-vague knowledge of nuisance parameters.

References

[1] R. Fisher, "Statistical methods for research workers," 1925.

[2] I. D. Dinov, N. Christou, and J. Sanchez, "Chapter 9: Central Limit Theorem," in *Journal of statistics education*, 2008.

[3] J. Arbuthnott, "An argument for divine providence, taken from the constant regularity observed in the births of both sexes," *Philosophical Transaction of the Royal Society of London*, 1710.

[4] G. Casella and G. L. Berger, *Statistical inference*. 1993.

[5] S. M. Stigler, "The History of Statistics: The Measurement of Uncertainty Before 1900.," 1987.

[6] J. Neyman and E. S. Pearson, "On the Problem of the Most Efficient Tests of Statistical Hypotheses," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 1933.

[7] D. J. Biau, B. M. Jolles, and R. Porcher, "P value and the theory of hypothesis testing: An explanation for new researchers," 2010.

[8] S. N. Goodman, "Toward evidence-based medical statistics. 1: The P value fallacy," 1999.

[9] S. Goodman, "A Dirty Dozen: Twelve P-Value Misconceptions," *Seminars in Hematology*, 2008.

[10] M. J. Schervish, "P Values: What They Are and What They Are Not," *The American Statistician*, 1996.

[11] H. W. Cohen, "P values: Use and misuse in medical literature," *American Journal of Hypertension*, 2011.

[12] D. Colquhoun, "An investigation of the false discovery rate and the misinterpretation of p-values," *Royal Society Open Science*, 2014.

[13] S. Greenland, S. J. Senn, K. J. Rothman, J. B. Carlin, C. Poole, S. N. Goodman, and D. G. Altman, "Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations," *European Journal of Epidemiology*, 2016.

[14] D. Chavalarias, J. D. Wallach, A. H. T. Li, and J. P. Ioannidis, "Evolution of reporting P values in the biomedical literature, 1990-2015," *JAMA - Journal of the American Medical Association*, 2016.

[15] A. Gelman and E. Loken, "The statistical Crisis in science," *American Scientist*, 2014.

[16] R. Peng, "The reproducibility crisis in science: A statistical counterattack," *Significance*, 2015.

[17] M. M. Levy, F. Albuquerque, J. D. Pfeifer, and J. P. A. Ioannidis, "Why most published research findings are false," *PLoS Medicine*, 2005.

[18] J. T. Leek and L. R. Jager, "Is Most Published Research Really False?," *Annual Review of Statistics and Its Application*, 2017.

[19] N. S. Jacobson and P. Truax, "Clinical Significance: A Statistical Approach to Defining Meaningful Change in Psychotherapy Research," *Journal of Consulting and Clinical Psychology*, 1991.

[20] K. L. Sainani, "Clinical Versus Statistical Significance," *PM and R*, 2012.

[21] E. J. Masicampo and D. R. Lalande, "A peculiar prevalence of p values just below .05," *Quarterly Journal of Experimental Psychology*, 2012.

[22] A. Gelman and H. Stern, "The difference between "significant" and "not significant" is not itself statistically significant," *American Statistician*, 2006.

[23] W. Edwards, H. Lindman, and L. J. Savage, "Bayesian statistical inference for psychological research," *Psychological Review*, 1963.

[24] J. M. Dickey and B. P. Lientz, "The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain," *The Annals of Mathematical Statistics*, 1970.

[25] J. O. Berger and T. Sellke, "Testing a point null hypothesis: The irreconcilability of P values and evidence," *Journal of the American Statistical Association*, 1987.

[26] J. P. Ioannidis, "The proposal to lower P value thresholds to .005," 2018.

[27] R. A. Armstrong, "When to use the Bonferroni correction," 2014.

[28] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, 1995.

[29] C. Woolston, "Psychology journal bans P values," *Nature*, 2015.

[30] R. L. Wasserstein and N. A. Lazar, "The ASA ' s statement on p-values : context , process , and purpose," *The American Statistician*, 2016.

[31] T. Bayes and R. Price, "An Essay towards Solving a Problem in the Doctrine of Chances By the Late Rev. Mr. Bayes," *Philosophical Transactions (1683-1775)*, 1763.

[32] H. Jeffreys, *Theory of Probability*. 1961.

[33] R. E. Kass and A. E. Raftery, "Bayes factors," *Journal of the American Statistical Association*, 1995.

[34] A. Gelman, "Objections to Bayesian statistics," *Bayesian Analysis*, 2008.

[35] C. P. Robert, *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. 2007.

[36] S. Van Dongen, "Prior specification in Bayesian statistics: Three cautionary tales," *Journal of Theoretical Biology*, 2006.

[37] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*. 1985.

[38] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*. 2004.

[39] J. W. Seaman, J. W. Seaman, and J. D. Stamey, "Hidden dangers of specifying noninformative priors," *American Statistician*, 2012.

[40] R. E. Kass and L. Wasserman, "The selection of prior distributions by formal rules," *Journal of the American Statistical Association*, 1996.

[41] J. O. Berger and M. J. Bayarri, "The Interplay of Bayesian and Frequentist Analysis," *Statistical Science*, 2004.

[42] A. J. Bonis, "Comparative Statistical Inference," *Technometrics*, 1976.

[43] D. B. Rubin, "Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician," *The Annals of Statistics*, 1984.

[44] L. Y. T. Inoue, D. A. Berry, and G. Parmigiani, "Relationship between bayesian and frequentist sample size determination," 2005.

[45] D. J. Spiegelhalter and L. S. Freedman, "A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion," *Statistics in Medicine*, 1986.

[46] C. J. Adcock, "Sample size determination: A review," 1997.

[47] R. Weiss, "Bayesian sample size calculations for hypothesis testing," *Journal of the Royal Statistical Society Series D: The Statistician*, 1997.

[48] T. Pham-Gia and N. Turkkan, "Determination of exact sample sizes in the Bayesian estimation of the difference of two proportions," *Journal of the Royal Statistical Society Series D: The Statistician*, 2003.

[49] G. Casella and R. L. Berger, "Reconciling bayesian and frequentist evidence in the one-sided testing problem," *Journal of the American Statistical Association*, 1987.

[50] J. W. Pratt, "Bayesian interpretation of standard inference statements," *Journal of the Royal Statistical Society: Series B (Methodological)*, 1965.

[51] C. Poole, "Beyond the confidence interval," *American Journal of Public Health*, 1987.

[52] S. N. Goodman, "Of p-values and bayes: A modest proposal," 2001.

[53] S. Greenland and C. Poole, "Living with P values: Resurrecting a bayesian perspective on frequentist statistics," 2013.

[54] M. H. DeGroot, "Doing what comes naturally: Interpreting a tail area as a posterior probability or as a likelihood ratio," *Journal of the American Statistical Association*, 1973.

[55] E. J. Wagenmakers, "A practical solution to the pervasive problems of p values," 2007.

[56] D. J. Murdoch, Y.-L. Tsai, and J. Adcock, "P -Values are Random Variables," *The American Statistician*, 2008.

[57] J. M. Robins, A. van der Vaart, and V. Ventura, "Asymptotic Distribution of P Values in Composite Null Models," *Journal of the American Statistical Association*, 2000.

[58] A. P. Dempster and M. Schatzoff, "Expected Significance Level as a Sensitivity Index for Test Statistics," *Journal of the American Statistical Association*, 1965.

[59] M. Schatzoff, "Sensitivity Comparisons Among Tests of the General Linear Hypothesis," *Journal of the American Statistical Association*, vol. 61, no. 314, pp. 415–435, 1966.

[60] H. Sackrowitz and E. Samuel-Cahn, "P values as random variables-expected P values," *American Statistician*, 1999.

[61] M. J. Bayarri, D. J. Benjamin, J. O. Berger, and T. M. Sellke, "Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses," *Journal of Mathematical Psychology*, 2016.

[62] G. Parmigiani and L. Inoue, *Decision Theory Principles and Approaches.* 2009.

[63] D. Bernoulli, "Specimen theoriae novae de mensura sortis," *Commentarii academiae scientiarum imperialis Petropolitanae*, 1738.

[64] P. S. LaPlace, "Théorie analytique des probabilités,"

[65] K. F. Gauss, *Theory of the combination of observations which leads to the smallest errors.* 1821.

[66] A. Wald, "Sequential Tests of Statistical Hypotheses," *The Annals of Mathematical Statistics*, 1945.

[67] J. Berkson, "Some Difficulties of Interpretation Encountered in the Application of the Chi-Square Test," *Journal of the American Statistical Association*, 1938.

[68] D. R. Anderson, K. P. Burnham, and W. L. Thompson, "Null Hypothesis Testing: Problems, Prevalence, and an Alternative," *The Journal of Wildlife Management*, 2000.

[69] J. Wakefield, "Bayes factors for Genome-wide association studies: Comparison with P-values," *Genetic Epidemiology*, 2009.

[70] M. J. Crawley, *The R Book*. 2007.

[71] P. Wellin, *Programming with Mathematica®: An introduction*. 2011.