On the Probability of Observing Misleading Evidence in Sequential Trials

by

Jeffrey D. Blume

A dissertation submitted to The Johns Hopkins University

in conformity with the requirements for the degree of

Doctor of Philosophy

Baltimore, Maryland

July, 1999

Abstract

Throughout the course of a clinical trial, study investigators are ethically obligated to monitor participant safety, as well as the accumulating evidence concerning the effectiveness of treatments (often known as a sequential trial). However, current statistical tools discourage continuous monitoring of study data. For example, it is well known that when conducting repeated significance tests on accumulating data the probability of a type one error approaches unity (Armitage's repeated significance testing paradox). Sequential, group sequential, and Bayesian methods have failed to fill this void in current practice for a variety of reasons.

An Evidential Paradigm, based on the Law of Likelihood, is examined in the context of continuous monitoring. This paradigm (1) uses likelihood ratios, not p-values, to measure the strength of statistical evidence and (2) provides a bound on, and control over, the frequency of observing both misleading and weak evidence. Instead of representing evidence against a null hypothesis, the likelihood function measures relative evidence supporting one simple hypothesis over another. Re-examination of accumulating evidence does not diminish its strength, because the likelihood function is unaffected by the number of examinations.

A procedure fashioned after the Law of Likelihood is proposed to accommodate composite hypotheses. Brownian Motion techniques are used to approximate and demonstrate control over the probability of observing misleading evidence. This procedure allows continuous monitoring for evidence of a clinically significant treatment effect over no treatment effect, while maintaining a low probability of observing evidence that might be construed as "misleading".

To my beloved Jewel

Acknowledgments

No amount of prose can adequately express my gratitude to the following individuals for their contributions to my graduate studies at Johns Hopkins:

My mentor, Richard Royall, for his unending intellectual generosity and guidance, and his personal friendship. Richard's dedication and perseverance in statistical research continues to set the standard to which I aspire in my own career. My growth and professional maturity is a reflection of his constant attention, caring, and brilliance.

Steven Piantadosi, for his expertise and insight on multiple issues, intellectual and professional guidance, and personal friendship during these past years. My regrets for not having found more time to spend with Steve in these last years.

Steven Goodman, for his personal friendship, the question motivating this work, and excellent roadmap of the foundations of statistical inference. Our many discussions on the foundations helped shape my passion for statistics.

Marie Diener-West, for her personal friendship and continual support during these past years. Extra time for my benefit was always available from Marie, as well as insightful comments.

Curt Meinert, his professional and keen expertise was invaluable during these past years. I have learned much from Curt, but most of all: "Pigs is pigs, data is data".

Scott Zeger and Charles Rohde for building such a splendid department in which to study and for their personal support over the years.

The faculty of the department: especially Ron Brookmeyer for many fascinating discussions and advise, Karen Bandeen-Roche for her support early in my academic career, Mei-Cheng Wang and Kung-Yee Liang for helpful professional guidance during this past year, and Subhash Lele for his training in the early years.

The staff of the department: Mary Joy Argo, for her headaches should have often been mine, and Mark Chiveral and Patty Hubbard for their friendship, kindness, and help from the beginning.

The senior students in the department: Paul Rathouz, Diana Miglioretti, Sterling Hilton, and Patrick Heagerty for their friendship and guidance; and my recent peers, especially Tom Travison, Mike Griswold, Elizabeth Garrett, Qian-li Xue, and Marc Starnes for the many discussions and moral support from which I have benefited both personally and professionally. A special thanks to my cohort and office-mates who found a way to deal with, and accept, my quirks for so many years.

And especially, my wife Jewel, who felt my pain and joy in this process almost as much as I did. I am indebted to her for her love, kindness, and understanding.

Table of Contents

List of Tables

List of Figures

Chapter 1

The Likelihood Paradigm

1.1 Introduction

Researchers continue to use Neyman-Pearsonian inferential procedures even in the presence of well documented shortcomings (e.g. the multiple looks and multiple comparison problems) [5],[9],[18]. One reason for this is that Neyman-Pearson type procedures enable the experimenter to measure and control the probability of making an erroneous decision [29]. To present a serious challenge to the Neyman-Pearson Paradigm, an alternative paradigm must also offer similar control over error probabilities or analogous quantities.

The Likelihood Paradigm is exactly such an alternative. It offers researchers an objective measure of the strength of evidence and control over the probabilities of observing misleading, weak, and strong evidence. The paradigm (1) formalizes the concepts of strong, weak, and misleading evidence and, (2) uses the probabilities of observing these types of evidence as tools for evaluating experimental designs. As a result, the probabilities of observing misleading or weak evidence provide quantities analogous to the Type I and Type II error probabilities of hypothesis testing, but do not represent the strength of the

evidence in the data.

The purpose of this chapter is to introduce the Likelihood Paradigm for interpreting observations as statistical evidence. The fundamental axiom of the Likelihood Paradigm is the Law of Likelihood. The Law of Likelihood provides a mechanism for interpreting observations, under a probability model, as statistical evidence about a parameter of interest. This requires the calculation of likelihood ratios and/or simple examination the likelihood function. Often a graph of the likelihood function is displayed to communicate the strength of the evidence over the entire parameter space. A detailed framework for conducting a likelihood-based statistical analysis is outlined in Edwards [13] and Royall [29].

1.2 The Law of Likelihood

Observations, in the context of a probability model, represent statistical evidence. The Law of Likelihood is an axiom for the interpretation of observations as statistical evidence.

> ***The Law of Likelihood*:** If the first hypothesis, H_1, implies that the probability for observing some random variable X is $f_1(X)$, while the second hypothesis, H_2, implies the probability is $f_2(X)$, then the observation $X = x$ is evidence supporting H_1 over H_2 if and only if $f_1(x) > f_2(x)$, and the likelihood ratio, $f_1(x)/f_2(x)$, measures the strength of that evidence.

Evidence for one simple hypothesis vis-à-vis another is measured by the likelihood ratio. The Law of Likelihood asserts that the hypothesis which is

better supported is the hypothesis which places a higher probability on the observed events. (A likelihood ratio equal to one indicates the observations do not support either hypothesis over the other.) This law represents a concept of statistical evidence that is essentially relative in nature; the data represent evidence for one hypothesis in relation to another. The data do not represent evidence for (or against) a single hypothesis.

A further implication of the Law of Likelihood is that statistical evidence and uncertainty have *different* mathematical forms. Uncertainty is measured through probabilities, which describe how often evidence of a particular type will be observed or how an experiment will perform over the so-called 'long run'. A separate mathematical quantity, the likelihood ratio, is required to measure the strength of the evidence after observations have been collected. This distinction, between the measure of the strength of evidence and the frequency with which such evidence occurs, is a critical implication of the Law of Likelihood. R.A. Fisher, who was often cryptic in his writings, made a clear distinction in this case: "In fact, as a matter of principle, the infrequency with which, in particular circumstances, decisive evidence is obtained, should not be confused with the force, or cogency, of such evidence" [17, p93].

1.3 Three Questions

After some data are collected in a scientific study, the statistician is often asked the following three questions:

1. What should I believe, now that I have collected these data?

2. What should I do, now that I have collected these data?

3. What do these data tell me about one hypothesis versus another? (How should I interpret these data as evidence regarding one hypothesis versus another?)

The answer to question one, namely Bayes theorem, requires not only the data, but knowledge of what was believed by the scientist prior to collecting the data [14]. Likewise the answer to question two, namely statistical decision theory, depends on not only the data, but on prior beliefs and expected losses and gains [34]. Furthermore, the answer to question three cannot be deduced from either question one's or questions two's answer. Therein lies a problem at the core of inferential statistics; the answers for questions one and two are often substituted for the answer to question three [29], [30].

Scientists turn to statistics for the answer to question three. The Law of Likelihood is well suited for this task because "the Law of Likelihood does not address question one or two, it is concerned with the objective evaluation of data as evidence, independently of prior beliefs or decision processes to which that evidence might be relevant" [31, p4]. In addition, the Law is consistent with Bayesian inference and decision theory [14], [29].

The distinction between the three questions is best illustrated through example. Consider a diagnostic test for some disease D where the properties of the test are listed in the following table (Table 1.1). This test for the presence of disease D has sensitivity 0.94 and specificity 0.98.

Suppose a doctor applies this test to individual X and a positive result is reported to the doctor. In this case, a simple calculation using Bayes theorem gives the doctor's probability that individual X has the disease given the positive

	Test Result	
Disease D	Positive	Negative
Present	0.94	0.06
Absent	0.02	0.98

Table 1.1: Probability Table for a Diagnostic Test.

test result, thus answering question number one. This probability will range from 0 to 1 depending on the prevalence of the disease. Whether or not the doctor should believe that individual X has the disease will depend on the rarity of the disease.

Likewise, the treatment which should be prescribed to individual X depends on the prevalence of the disease as well as the benefits and costs (monetary and otherwise) associated with the treatment. If available, the doctor may prescribe a simple, noninvasive, inexpensive treatment as a precautionary measure, even though the doctor may not believe individual X suffers from the disease. The answer to question number two depends not only on the answer to question number one, but on the interplay between costs and benefits, as well.

Neither what the doctor should believe nor what the doctor should do answers the third question. The positive test does not prove that the individual has the disease, but represents statistical evidence about the true disease state of individual X. Specifically, the third question asks, "What do the data say about the hypothesis that the disease is present verses the hypothesis that the disease is absent?" That is, what is the evidence for H_p (disease present) over H_a (disease absent)? Certainly a positive test result is evidence for H_p over H_a, but why? And how much better do the data support H_p over H_a? It is the Law of Likelihood that provides the answers to these fundamental questions.

By the Law of Likelihood, H_p is better supported over H_a because the likelihood of a positive test for a diseased individual is greater than the likelihood of a positive test for a nondiseased individual (e.g. $0.94 = P(T+|D+) > P(T+|D-) = 0.02$) and the strength of the evidence for H_p vis-à-vis H_a is the likelihood ratio of 47 (e.g. $P(T+|D+)/P(T+|D-) = 0.94/0.02 = 47$). A likelihood ratio of 47 indicates that the data strongly support H_p over H_a because the ratio is much larger than one. The following section will provide benchmarks for determining when a likelihood ratio is 'large' or 'small' i.e. when the evidence is 'strong' or 'weak'. For now, it is enough to understand why the data support H_p over H_a.

It is quite possible that individual X does not really have the disease. Regardless of the true disease status of individual X, the Law of Likelihood indicates that a positive test result is evidence to support H_p over H_a. That is, a positive test result is always evidence for H_p over H_a, even if the disease is really absent. When this happens, misleading evidence has been collected (in practice, it is unknown if evidence is misleading or not). The evidence has been interpreted correctly (via the Law of Likelihood) and no error has been made; it is the evidence itself that is misleading. Therefore, there are times when properly interpreted evidence can be misleading.

In the diagnostic test example, the probability that the positive test result for individual X constitutes misleading evidence is small ($P(T+|D-) = 0.02$). The probability of observing misleading evidence is an important property of the diagnostic test because the outcome of a single test is often used as a surrogate for the true disease status of an individual. We will see later (section 1.5) that, in general, it is not possible to frequently observe misleading evidence.

1.4 The Strength of Statistical Evidence

Suppose the likelihood ratio for H_1 versus H_2 is equal to some fixed number k. Now k measures the strength of evidence for the first hypothesis over the second hypothesis. If k is greater than one, then the Law of Likelihood indicates that the evidence favors H_1 over H_2. Likewise, if k is less than one, then the evidence supports H_2 over H_1. A likelihood ratio equal to one would indicate that the evidence does not support either hypothesis over the other; the evidence for H_1 vis-à-vis H_2 is neutral. The strength of the evidence for H_1 over H_2 can take any non-negative value, from zero (indicating overwhelming evidence for H_2 over H_1) to infinity (indicating the reverse). For the purpose of interpreting and communicating the strength of evidence, it is useful to divide the continuous scale of the likelihood ratio into descriptive categories, such as "weak", "moderate", and "strong" evidence. Such a crude categorization allows a quick and easily understood characterization of the evidence for one hypothesis over another.

Benchmark values of k = 8 and 32 are suggested to distinguish between weak, moderate, and strong evidence [29]. (Similar benchmarks have been proposed by others [20],[13],[23].) Observations resulting in likelihood ratio $1/8 \leq k \leq 8$ represent weak evidence; neither hypothesis is better supported over the other by a factor of 8 or more. For example, if $k = 1/7$ there is weak evidence to support H_2 over H_1. Those with a likelihood ratio of 8 or more (or less than 1/8) represent at least moderate (fairly strong) evidence. And those with a likelihood ratio greater than 32 (or less than 1/32) represent strong evidence. When interpreting evidence, these benchmarks are to be used as guidelines and

not cutoff points. Table 1.2 displays the suggested categories for interpreting the strength of evidence as given by likelihood ratios.

Range of Likelihood Ratio	Strength of Evidence
1-8	Weak
8-32	Moderate
over 32	Strong

Table 1.2: Benchmarks for the Strength of Statistical Evidence

1.5 Misleading Evidence

It is possible to observe strong evidence for H_2 over H_1 when, in actuality, H_1 is correct. Such evidence is misleading (refer to the previous diagnostic test example in section 1.3). *Misleading evidence is defined as strong evidence in favor of the incorrect hypothesis over the correct hypothesis.* After collecting data, the value of the likelihood ratio, say k, will determine whether strong, moderate, or weak evidence has been observed for one hypothesis versus another. The likelihood ratio will not indicate if the evidence is misleading. Because the correct hypothesis is unknown, it is not possible to determine if misleading evidence has been observed without collecting more observations.

The frequency with which misleading evidence is observed (called the probability of observing misleading evidence) is, in general, low. For any fixed sample size the probability of observing misleading evidence ($LR \geq k$) is always less than or equal to $1/k$ [39],[8],[29]. The bound on the probability of observing misleading evidence, $1/k$, applies to any pair of distributions with any structure. Thus, this bound has been named the Universal bound [29]. Mathematically, if

both $f(X)$ and $g(X)$ are probability density functions and $X \sim f(X)$ then

$$P_f\left(\frac{g(X)}{f(X)} \geq k\right) \leq \frac{1}{k} \tag{1.1}$$

A simple application of Markov's inequality will yield the result. The universal bound indicates that, for moderately large k, misleading evidence will not be observed very often. The probability of observing very strong misleading evidence (i.e. evidence supporting an incorrect hypothesis over the correct hypothesis by a factor of 32 or more) cannot exceed $1/32 = 0.031$. Notice that the Universal bound on the probability of observing misleading evidence requires that k is fixed *before* collecting observations.

In the case of multiple looks at the data, the likelihood function is unaffected. By contrast, the probability of observing misleading evidence increases with each look (from two to infinity) *but also remains bounded* by the Universal bound, see Robbins [27]. This result also holds for any pair of probability distributions. If both $f(X_n)$ and $g(X_n)$ are probability density functions and X_n is a vector of n observations with joint density $f(X_n)$ then

$$P_f\left(\frac{g(X_n)}{f(X_n)} \geq k \quad ; \text{for any } n = 1, 2, \ldots\right) \leq \frac{1}{k} \tag{1.2}$$

Thus, an experimenter who plans to examine the data with each new observation, stopping only when the data support H_g over H_f, will be eternally frustrated with probability at least $1 - 1/k$. Further discussion of Robbins' result and its application to clinical trials is continued in Chapter 2. The implication is that for *any* stopping rule the probability of observing misleading evidence cannot exceed the Universal bound.

Finally, the probability of observing misleading evidence is a key planning tool. This planning probability does not aid (or detract from) the evidence in the

data, i.e. it does not modify the likelihood ratio. However it does influence one's opinion or belief about the reliability of the study. But this issue is separate from that of objectively measuring the statistical evidence generated by data from a study (i.e. see §1.3). Even a poorly designed study generates statistical evidence that can be measured objectively.

1.6 Graphical Presentation of Statistical Evidence

After observations have been collected, the likelihood function can be graphed to provide a visual impression of the evidence about the parameter of interest. For presentation purposes, likelihood functions can be standardized by their maximum value (a constant). The scaling constant for the likelihood function can be chosen arbitrarily because the Law of Likelihood implies that likelihood functions are only meaningful up to a proportional constant. A brief 'statistical analysis' consists of presenting the likelihood function with a few simple summary measures to help describe the curve.

Suppose the parameter of interest is θ and denote the likelihood function as $L(\theta)$. The standardized likelihood function resides in the interval $[0, 1]$ and is defined as

$$\frac{L(\theta)}{\max_\theta L(\theta)} = \frac{L(\theta)}{L(\hat{\theta})} \tag{1.3}$$

The maximum of the likelihood function will always occur at the maximum likelihood estimator, $\hat{\theta}$ (MLE) for θ. Note that the maximum likelihood estimator is always the best supported value of the parameter in the sense that the likelihood ratio for the hypothesis $\theta = \hat{\theta}$ versus any other hypothesis $\theta = \theta_i$ is greater than or equal to one ($L(\hat{\theta})/L(\theta_i) \geq 1 \ \ \forall i$).

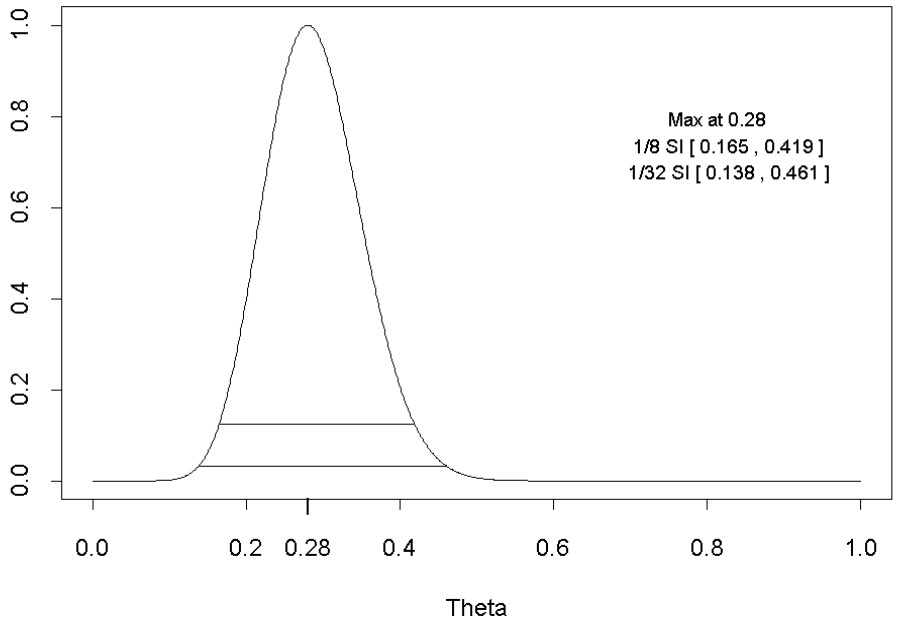

Figure 1.1: Likelihood Function for the Probability of a Head

To demonstrate, suppose I am interested in generating evidence about the probability of obtaining heads from the toss of a biased coin. Generating statistical evidence in this case requires tossing the biased coin. The coin is tossed 50 times and 14 heads are observed. The appropriate statistical model for this experiment is to consider each toss of the coin as an independent realization of a Bernoulli random variable with common probability of a head θ. Hence $X_1, X_2, \ldots, X_{50}$ are independent and identically distributed $Bernoulli(\theta)$ random variables. Thus the standardized likelihood function is:

$$L(\theta) = \frac{\theta^{14}(1-\theta)^{36}}{\hat{\theta}^{14}(1-\hat{\theta})^{36}} \quad \text{where} \quad \hat{\theta} = \frac{14}{50} = 0.28 \tag{1.4}$$

Figure 1.1 is the graphical presentation of the likelihood function (equation (1.4)). The y-axis is not labeled because only ratios of points on the likelihood function have evidential meaning. Notice that the best supported hypothesis is $\theta = \hat{\theta} = 0.28$. Also the hypothesis $\theta = 0.3$ is better supported over $\theta = 0.5$ by a factor of 143 ($L(0.3)/L(0.5) = 143$). The hypotheses most consistent with the data are those θ under the crest of the likelihood function. The evidence provided by the data indicate that the coin is biased.

To provide a summarization of the evidence about θ, a collection of all hypotheses that are 'consistent with the data' is presented. A set of hypotheses that are 'consistent with the data' at some level k is called a $1/k$ likelihood support interval. A $1/k$ likelihood support interval (SI) is defined as

$$\left\{\theta : \frac{L(\theta)}{\max_\theta L(\theta)} \geq \frac{1}{k}\right\} = \left\{\theta : \frac{L(\hat{\theta})}{L(\theta)} \leq k\right\} \tag{1.5}$$

Remaining consistent with the aforementioned benchmarks, k = 8 and 32 will be used to construct moderate and strong support intervals. The hypotheses (parameter values) in a 1/8 support interval are considered 'consistent with the data' because no alternative parameter value is better supported by a factor of more than eight over any other parameter value within the interval. In other words, there is no alternative hypothesis better supported by a factor of 8 or more over a hypothesis within the 1/8 support interval. At most, there is only weak evidence supporting some alternative parameter value over a parameter value within the interval. In our example, hypotheses suggesting the probability of a head is between 0.138 and 0.461 (the 1/32 SI drawn on the likelihood function) are consistent with the data at a strong level.

1.7 Hypothesis Testing and Likelihood Methods

There is a close mathematical connection between Neyman-Pearson hypothesis testing and the Likelihood Paradigm. Both paradigms use likelihood ratios as key quantities, but each for a different purpose. Hypothesis testing procedures *do not* place any interpretation on the numerical value of the likelihood ratio. The extremeness of an observation is measured, not by the magnitude of the likelihood ratio, but by the probability of observing a likelihood ratio that large or larger. It is the tail area, not the likelihood ratio, that is the meaningful quantity in hypothesis testing. By contrast, the Law of Likelihood shows that the likelihood ratio itself, measures the strength of statistical evidence.

These differences are subtle, but critically important. Consider the following common example: Observe $X_1, X_2, \ldots, X_n$ i.i.d. $f(X_i; \mu)$. Let $L_n(\mu) = \prod f(x_i; \mu)$ be the likelihood function. Suppose two hypotheses of interest are $H_0 : \mu = \mu_0$ and $H_1 : \mu = \mu_1$. The Type I error rate of hypothesis testing is defined as

$$P_0\left(\frac{L_n(\mu_1)}{L_n(\mu_0)} \geq k\right) = \alpha \quad \text{where } \alpha \text{ is } \textit{fixed.} \tag{1.6}$$

By contrast the probability of observing misleading evidence for μ_1 over μ_0 is

$$P_0\left(\frac{L_n(\mu_1)}{L_n(\mu_0)} \geq k\right) = M(n) \quad \text{where } k \text{ is } \textit{fixed.} \tag{1.7}$$

While the Type I error rate and the probability of observing misleading evidence appear to have the same mathematical form, they are actually very different. In hypothesis testing, the Type I error rate is fixed at α and the strength of evidence k, at which the test rejects, depends on α. In contrast, for the probability of observing misleading evidence, the strength of the evidence k is fixed and the

resulting probability, $M(n)$, depends on k.

Consider the case when the data are normally distributed with mean μ_0 and known variance σ^2. A simple calculation shows that the probability of observing misleading evidence for μ_1 is [26],[29].

$$M(n) = P_0\left(\frac{L_n(\mu_1)}{L_n(\mu_0)} \geq k\right) = \Phi\left[-\frac{\Delta\sqrt{n}}{2\sigma} - \frac{\sigma \ln k}{\Delta\sqrt{n}}\right]$$

where $\Delta = |\mu_1 - \mu_0|$ and $\Phi(\cdot)$ is the standard normal cumulative distribution function. The Type I error, on the other hand, forces the above probability to be fixed at α. Hence, we can calculate k such that

$$\Phi\left[-\frac{\Delta\sqrt{n}}{2\sigma} - \frac{\sigma \ln k}{\Delta\sqrt{n}}\right] = \Phi\left(-Z_\alpha\right) = \alpha$$

This implies for hypothesis testing, that the strength of the evidence for rejection by the test is

$$k = \exp\left\{Z_\alpha \frac{\Delta\sqrt{n}}{\sigma} - \frac{n\Delta^2}{2\sigma^2}\right\}$$

When conducting a hypothesis test it is possible for k to be less than one, indicating that the hypothesis test would reject H_0 when the evidence favored H_0 over H_1. For example, suppose $n = 30$ observations are collected to test $H_0 : \mu = \mu_0$ versus $H_1 : \mu = \mu_0 + \Delta$ with size $\alpha = 0.05$ when the two hypotheses of interest are one standard deviation apart, i.e. $\Delta = \sigma$. Now the hypothesis test rejects H_0 in favor of H_1 if the likelihood ratio, $L_n(\mu_1)/L_n(\mu_0)$, is greater than $k = 1/70$. Thus observations which support H_0 over H_1 by a factor of 70 or less (32 or greater is strong evidence for H_0 over H_1) would still lead to rejection of H_0.

A similar argument can be made concering the probability of observing weak evidence and the Type II error. While they have similar mathematical forms,

they represent different quantities. The probability of observing weak evidence can be expressed as

$$W(n) = P_1\left(\frac{1}{k} < \frac{L_n(\mu_1)}{L_n(\mu_0)} < k\right) = \Phi\left[\frac{\Delta\sqrt{n}}{2\sigma} + \frac{\sigma \ln k}{\Delta\sqrt{n}}\right] - \Phi\left[\frac{\Delta\sqrt{n}}{2\sigma} - \frac{\sigma \ln k}{\Delta\sqrt{n}}\right]$$

Figure 1.2 plots the probability of observing misleading and weak evidence ($M(n)$ and $W(n)$, respectively) with the Type I and II error probabilities, as a function of the sample size. All four curves are quite distinct. Hence the Type I error probability is not to be interpreted as the probability of observing misleading evidence and the Type II error probability is not to be interpreted as the probability of observing weak evidence.

As shown in Figure 1.2, the probabilities of observing misleading and weak evidence both converge to zero as the sample size increases [29]. (Note that the Type I error probability remains constant regardless of the total sample size.) Because the probability of observing misleading evidence naturally converges to zero as the sample size increases, sample size calculations are based on achieving some low level of the probability of observing weak evidence. Furthermore, if μ_0 is the true mean, then the law of large numbers implies that

$$\frac{L_n(\mu_0)}{L_n(\mu_i)} \xrightarrow{a.s.} \infty \quad \forall \ i \neq 0$$

ensuring that the likelihood ratio will eventually favor the true mean over any other alternative value by a large factor. This result applies very generally to a wide class of models.

The well known large sample result that minus twice the log-likelihood ratio has an approximate chi-square distribution can be used to show that 95% confidence intervals are approximately 1/6.67 likelihood support intervals. However

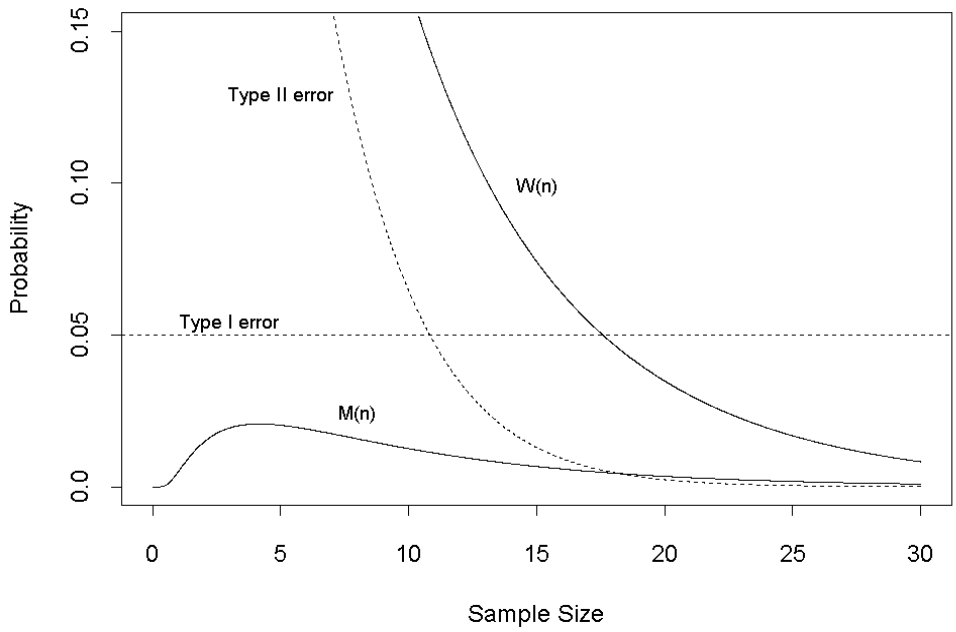

Figure 1.2: Probabilities of Weak and Misleading Evidence

in moderate to small samples, support intervals and confidence intervals are often dissimilar. There are three major advantages of support intervals over confidence intervals. The first is that the interpretation of a support interval does not depend only on long term frequencies of the repeated experiment. The second is that, unlike confidence intervals, support intervals do not depend on the sample space (e.g. binomial or negative binomial sampling resulting in the same outcome will yield the same support interval but different confidence intervals). The third is that likelihood support intervals summarize statistical evidence (unlike confidence intervals) and therefore do not suffer from multiple comparison or multiple look issues. More discussion of support intervals and confidence intervals can be found in [13],[29],[31],[30].

Finally, it is interesting to note that had Neyman and Pearson chosen to minimize a linear combination of the Type I and Type II error probabilities, (*e.g.* $\min(\alpha + k\beta)$), instead of fixing the Type I error rate and minimizing the Type II error rate, the resulting rejecting rule would coincide with the Law of Likelihood, i.e. reject when the likelihood ratio is greater than k [10].

1.8 Composite Hypotheses

The Law of Likelihood does not apply to composite hypotheses. However, in clinical trials, a comparison of a hypotheses such as 'no effect of the treatment' versus 'a treatment effect' is often of interest. The hypothesis of 'a treatment effect' is known as a composite hypothesis and suggests an often-broad range of possible values for the actual treatment effect. Composite hypotheses can lead to complications if the treatment effect is found to be 'statistically significant' but not 'clinically significant'. By comparing only simple versus simple hypotheses these complications can be avoided. This section is devoted to demonstrating the value of reporting all of the simple versus simple hypothesis comparisons rather than reporting a single summary measure representing the composite versus simple hypothesis comparison.

In order to obtain a single summary value that represents the evidence for a simple hypothesis versus a composite hypothesis (set of simple hypotheses), the relative evidence for the simple hypothesis versus each individual simple hypothesis from the composite set must be combined. However, evidence summarized over a composite space depends on the method of summarization. And, in most cases, the evidence for a simple hypothesis versus a composite hypothesis does

not accurately represent the evidence for any of the simple pairwise comparisons. For these reasons, the evidence provided by the data is best presented by the likelihood function. The following oversimplified example demonstrates this.

Suppose H_1 is supported over H_2 by a factor of $k = 1/10$, so that the data support H_2 over H_1. Suppose also that H_1 is supported over H_3 by a factor of $k = 100$. Consider comparing the support for H_1 versus the composite hypothesis $H_c = \{H_2 \text{ or } H_3\}$. We wish to obtain a single summary measure that represents the support for H_1 over H_c. Choosing the average as a summary measure would indicate that H_1 is supported over H_c by a factor of $k_c = 50.05$. However, the support for H_1 versus either hypothesis in the composite set is not 50.05. This composite support does not convey the correct directionality for the support of H_1 versus H_2, indicating inferences cannot be extended from the composite hypothesis to simple hypotheses within the composite space. Furthermore, the support for H_1 over H_c may change when summaries are performed in a different fashion, e.g. by taking the maximum ($k_c = 100$), minimum ($k_c = 1/10$), or weighted average ($k_c = 0.99(1/10) + 0.01(100) = 1.099$) (weights are chosen arbitrarily here).

In measuring the 'evidence' for a simple versus composite hypothesis, there is subjectivity in choosing the summarization technique for the composite space. By contrast, likelihood methods convey the relative pairwise support for each simple hypothesis but do not measure the 'evidence' for a simple versus composite hypotheses; thereby avoiding the specification of a subjective summarization technique. For this reason, reporting the likelihood function provides a more straightfoward and clear manner of presenting the results of a study. Simple

reporting of the likelihood function also allows for a composite summarization to be conducted at a later point in time, if desired.

There are other fundamental situations in statistics where comparisons are made between only simple versus simple hypothesis. A power curve is such an example. By definition, the power of a test depends critically on two distinct simple hypotheses. A power curve displays the power of a hypothesis test with a fixed simple null hypothesis for every possible simple alternative hypothesis. Hence, the power for a fixed null hypothesis is different for each separate simple alternative. Therefore, like statistical evidence, the power of a test with a fixed null hypothesis depends on the simple alternative hypothesis. When examining the power of a test, no attempt is made to summarize the power over a composite space. (However, Bayesians often determine the 'pre-posterior risk' which is similar to summarizing power over a composite space [40]).

Accurately describing the power of a test for some fixed null hypothesis requires the presentation of the entire power curve. Analogously, accurately representing the evidence requires the presentation of the entire standardized likelihood function. Power and evidence both depend on relative comparisons of simple versus simple hypotheses.

1.9 Summary

Under a probability model, the Law of Likelihood indicates how to objectively measure the strength of evidence provided by the data. Its dependence on two simple hypotheses is not a flaw, but a strength. The probability of observing misleading evidence is naturally controlled, even with mulitple looks at

the data. And the probability of observing weak evidence can be controlled through sample size. In addition, the Law of Likelihood is consistent with statistical methodology such as Bayesian inference and decision analysis. Most importantly, the Likelihood paradigm separates the measure of the strength of evidence from the probability of observing such evidence, allowing each quantity to be dealt with in their distinct roles.

Chapter 2

Misleading Evidence: The Case of Two Simple Hypotheses

2.1 Introduction

An experiment or observational study produces observations which, under a probability model, represent statistical evidence. As discussed in section 1.2, the Law of Likelihood is an axiom for interperting such evidence. After the data have been collected, the strength of the evidence will be determined by the likelihood ratio. Whether the evidence is weak or strong will be clear from the numerical value of the likelihood ratio. However, it remains unknown if the evidence is misleading or not. Fortunately, strong misleading evidence is not often observed.

The probabilities of observing misleading, weak, and strong evidence are important characteristics of any planned experiment. These probabilities can provide assurance that the experiment will produce what is desired (in terms of statistical evidence) and they are key tools for evaluating the study design. In this chapter we examine in detail the probability of observing strong misleading evidence in the case of two fixed simple hypotheses under various experimental

designs. It is shown that even for designs incorporating continuous monitoring of the data, with termination of study recruitment as soon as strong evidence in one direction is observed, the probability of observing strong misleading evidence tends to remain well below the Universal bound.

Five classes of experimental designs, defined by different restrictions on sample size, will be studied. In the first class, Fixed Sample Size designs, observations are collected until a pre-determined sample size is reached. The second class, Open designs, are open ended, i.e. there are no limits or restrictions on the sample size. The third class of designs, Truncated designs, refers to designs where the sample size, while not fixed, may not exceed a pre-determined upper limit. Delayed designs (the fourth class) require at least a pre-determined minimum number of observations. Finally, Interval designs, constrain the sample size between a pre-determined minimum and maximum number of observations.

Within each class except the first, we study the stopping rule whose purpose is to generate strong evidence for H_1 over H_0, where both simple hypotheses are fixed. This stopping rule terminates the study at the smallest sample size allowed by the design where strong evidence supporting H_1 over H_0 is obtained. (Herein we take the phrases 'terminate, stop, or end the study' to mean the termination of participant accrual and cessation of the 'inferior treatment', but not necessarily designated follow up.) Experimental designs incorporating this stopping rule entail continuous monitoring of the evidence and are known as sequential trials [10], [2], [43]. The study stops, not after collecting a predetermined number of observations, but when the observations themselves represent sufficiently strong evidence for H_1 over H_0.

The above stopping rule is designed to maximize the probability of generating misleading evidence for H_1 in each class of designs and for that reason is an important case study. If the maximum probability is small (and we show that it remains well below the Universal bound), then *any* other stopping rule will have an even smaller chance of generating strong misleading evidence for H_1. In short, the above stopping rule provides the greatest chance of observing misleading evidence for H_1 in each class of designs.

While the likelihood function (and hence statistical evidence) does not depend upon the stopping rule [8], [4], the probability that a particular experimental design will generate misleading evidence does depend on the stopping rule. Nevertheless, for *any* study design with a stopping rule dependent on the evidence between two fixed simple hypotheses, the Universal bound applies and, with probability greater then $1 - 1/k$, strong misleading evidence will never be produced (see section 1.5, [27]). Practically, this implies that an investigator searching for evidence for his favorite hypothesis over the correct hypothesis is likely (with probability greater than $1 - 1/k$) *never* to find such evidence. This property of statistical evidence is an excellent scientific safeguard; it is difficult to deliberately, or otherwise, collect strong misleading evidence.

Throughout this chapter, the five classes of experimental designs are examined under the model where observations $X_1, X_2, \ldots, X_n, \ldots$ are i.i.d. $N(\mu, \sigma^2)$ where σ^2 known. Let $L_n(\mu) = \prod f(x_i; \mu)$ represent the corresponding likelihood function for n observations and let $S_n = X_1 + \cdots + X_n$. We begin by considering the probability of generating misleading evidence under the first class of designs, Fixed Sample Size experiments.

2.2 Class I: Fixed Sample Size Designs

We begin by examining the probability that a Fixed Sample Size experiment will produce weak or misleading evidence about the value of a normal mean. Suppose the Fixed Sample Size experiment will collect n observations. The planning probabilities depend on the sample size and the distance between the two fixed simple hypotheses H_1 and H_0. Note that these probabilities have been derived elsewhere [26], [29], [31].

Consider the evidence provided by observations $x_1, \ldots, x_n$ for $H_1 : \mu = \mu_1$ versus $H_0 : \mu = \mu_0$. By the Law of Likelihood, the strength of that evidence is given by the likelihood ratio

$$\frac{L_n(\mu_1)}{L_n(\mu_0)} = \exp\left\{\frac{n(\mu_1 - \mu_0)}{\sigma^2}\left(\frac{S_n}{n} - \frac{\mu_1 + \mu_0}{2}\right)\right\} \tag{2.1}$$

When the true mean is μ_0, the probability that the observations will support μ_1 over μ_0 by a factor of k or more (i.e. the probability that the Fixed Sample Size design will produce misleading evidence) is

$$P_0\left(\frac{L_n(\mu_1)}{L_n(\mu_0)} \geq k\right) = \Phi\left[-\frac{\Delta\sqrt{n}}{2\sigma} - \frac{\sigma \ln k}{\Delta\sqrt{n}}\right] \tag{2.2}$$

where $\Delta = |\mu_1 - \mu_0|$ and $\Phi(\cdot)$ is the standard normal CDF.

The probability that the Fixed Sample Size experiment will produce weak evidence favoring either hypothesis when μ_0 is the true mean is

$$P_1\left(\frac{1}{k} < \frac{L_n(\mu_1)}{L_n(\mu_0)} < k\right) = \Phi\left[\frac{\Delta\sqrt{n}}{2\sigma} + \frac{\sigma \ln k}{\Delta\sqrt{n}}\right] - \Phi\left[\frac{\Delta\sqrt{n}}{2\sigma} - \frac{\sigma \ln k}{\Delta\sqrt{n}}\right] \tag{2.3}$$

(Note that this is the same as the probability when μ_0 is the true mean.)

If the difference Δ is measured in units of standard errors, say $\Delta = c\sigma/\sqrt{n}$,

then equation (2.2) becomes

$$P_0\left(\frac{L_n(\mu_1)}{L_n(\mu_0)} \geq k\right) = \Phi\left[-\frac{c}{2} - \frac{\ln k}{c}\right] \tag{2.4}$$

where c is the absolute difference between μ_1 and μ_0 in standard error units. Figure 2.1 is a graph of the probability of observing misleading evidence for μ_1 (equation 2.2) as a function of the distance between the two hypotheses (c) when $k = 8, 32$. It has been named the 'Bump function' by Royall [31] because of its graphical appearance. Examination of the Bump function reveals that the maximum probability of observing misleading evidence (over all c) is $\Phi[-\sqrt{2\ln k}]$ or 0.021 when $k = 8$. This maximum occurs at $\sqrt{2\ln k}$ standard errors [31]. Note that the maximum probability of observing misleading evidence is much less than the Universal bound of $1/8 = 0.125$ [31].

For values of Δ near zero there is essentially no chance of finding misleading evidence for μ_1. This happens because μ_1 and μ_0 specify distributions so similar that only very extreme observations will produce strong evidence supporting μ_1 and those extreme obervations are improbable under μ_0. As the difference between the two hypotheses grows, the Bump function increases until it reaches its maximum value at $\Delta = \sqrt{2\ln k}$ standard errors. At this maximum, observations which would support $\mu_1 = \mu_0 + \sqrt{2\sigma \ln k/n}$ over μ_0 are more likely to occur, because those observations are not too extreme under H_0. The Bump function decreases after reaching its maximum until there is essentially no chance of observing strong misleading evidence for μ_1. This is because observations which would support μ_1 over μ_0 (for large values of Δ) are very improbable under μ_0.

The Bump function describes the probability of observing misleading evidence for H_1 over H_0 in a Fixed Sample Size study. Next, we examine the

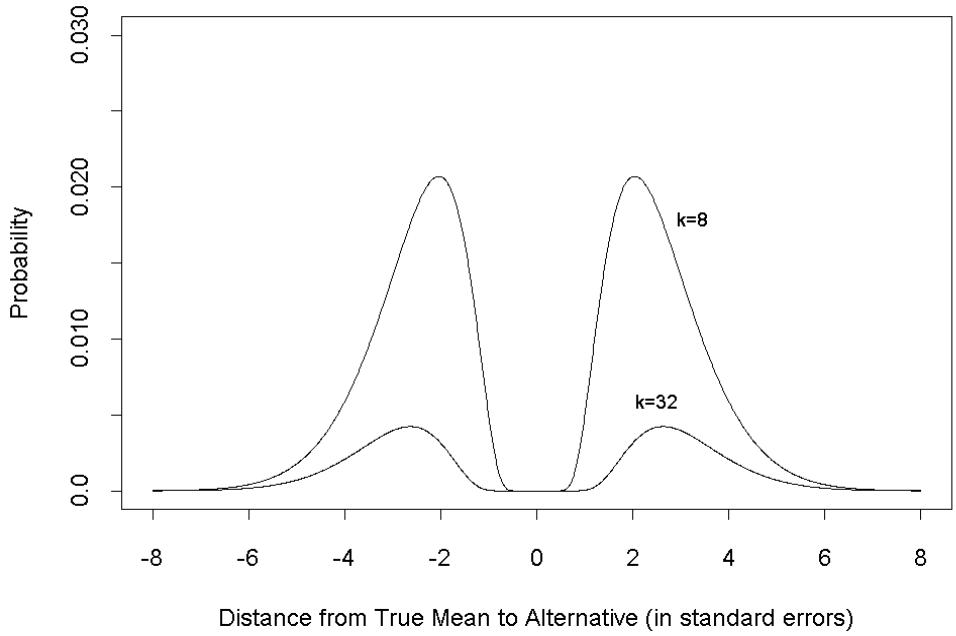

Figure 2.1: The Bump function

probability of observing misleading evidence for the class of Open designs.

2.3 Class II: Open Designs

The class of Open designs refers to experimental designs which are open-ended, i.e. there are no limits or restrictions on the sample size. The term 'Open Design' is commonly used in the sequential trial literature to refer to open-ended sequential hypothesis testing procedures [41], [43]. The class of Open designs considered here, while analogous in some respects, is not to be confused with its hypothesis-testing counterpart whose criteria to stop the study depends on the outcome of a repeated significance test. All subsequent references to 'Open designs' refer to designs whose criteria for stopping a study is based on the

evidential interpretation of likelihood ratios according to the Law of Likelihood, unless otherwise specified.

As discussed in section 2.1, we examine the stopping rule whose purpose is to generate strong evidence for H_1 over H_0, where both simple hypotheses are fixed. This stopping rule ends the study at the smallest sample size where strong evidence supporting H_1 over H_0 is obtained. The point at which the study stops is now a random variable called a stopping time, say N, and is defined as

$$N = \min\left\{n : n \geq 1, \frac{L_n(\mu_1)}{L_n(\mu_0)} \geq k\right\} \tag{2.5}$$

An Open design employing stopping time (2.5) continues sampling (possibly forever) until strong evidence for H_1 over H_0 is obtained. Thus, this design is *severely biased* in favor of H_1. Henceforth, an Open design employing this stopping rule will be called a Biased Open design. The Biased Open design maximizes the chance of generating misleading evidence for H_1 out of the entire class of Open designs.

The probability that the Biased Open design produces strong misleading evidence does not, under the normal model, achieve the Universal bound. Consider generating evidence about the value of a normal mean, say $H_1 : \mu = \mu_1$ versus $H_0 : \mu = \mu_0$. The probability that the Biased Open design produces strong misleading evidence for H_1 over H_0 can then be expressed in terms of the probability that the stopping time N is finite (see equation (2.5)).

Theorem 2.1 *(Biased Open Design) Suppose $X_1, X_2, \ldots$ are i.i.d. normal random variables with mean μ_0 and fixed known variance σ^2. Define a stopping time as $N = \inf\{n : n \geq 1, L_n(\mu_1)/L_n(\mu_0) \geq k\}$. Then*

1. $P_0\left(N < \infty\right) \cong \exp\left\{-\rho\Delta/\sigma\right\}/k$

2. $P_1\left(N < \infty\right) = 1$ *for any* μ_1

where $\Delta = |\mu_1 - \mu_0|, \rho \cong 0.583$ *and the subscript on the probability denotes the true mean.*

The proof of Theorem 2.1 is deferred to the end of this section. Therefore we have

$$P_0\left(\frac{L_n(\mu_1)}{L_n(\mu_0)} \geq k \text{ ; for any } n = 1, 2, \ldots\right) = P_0\left(N < \infty\right) \cong \frac{\exp\left\{-\rho\Delta/\sigma\right\}}{k} \quad (2.6)$$

where the subscript on the probability denotes the true mean, $\Delta = |\mu_1 - \mu_0|$, and $\rho \cong 0.583$ is a constant representing the expected overshoot of S_n over a linear boundary in the normal case (the appendix provides a more detailed explanation about ρ, see section A.3.3). If the difference Δ is measured in units of standard deviations, say $\Delta = c\sigma$, then equation (2.6) becomes

$$P_0\left(N < \infty\right) \cong \frac{\exp\left\{-\rho c\right\}}{k} \quad (2.7)$$

Figure 2.2 is a graphic representation of the probability that the Biased Open design produces strong misleading evidence for μ_1 over μ_0 when $k = 8$ (equation 2.7). The function on the RHS of equation (2.7) is called the 'Tepee' function because of its graphical appearance. The Bump function at $n = 1$ has been plotted for comparison purposes. Remember that, for the Bump function, the scale of Δ is expressed in standard errors, and hence the function would 'move' inwards for sample sizes larger than $n = 1$. By contrast, the scale of Δ for the Tepee function is expressed in standard deviations.

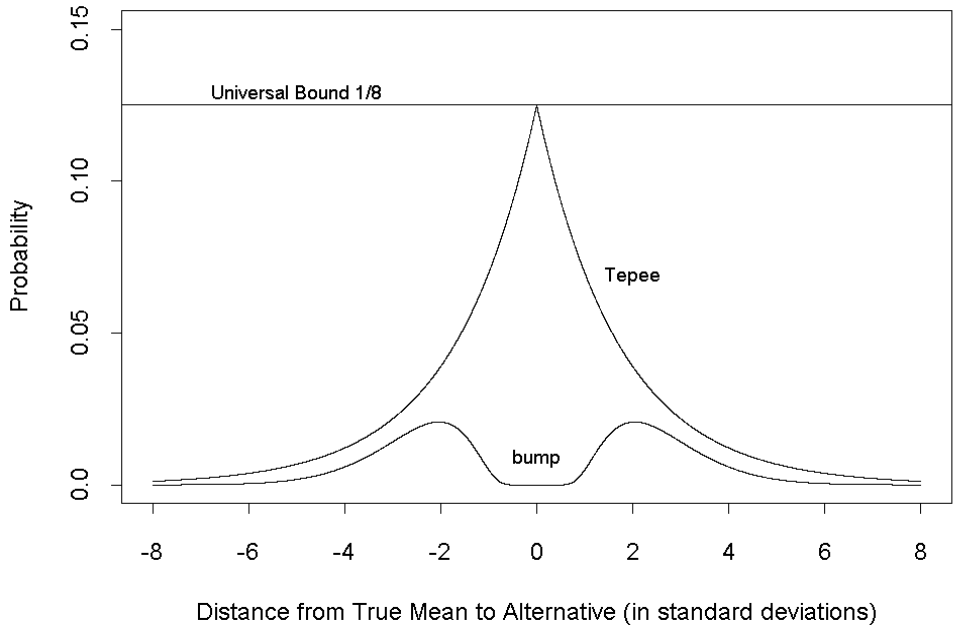

Figure 2.2: The Biased Open Design (Tepee function)

The Bump function represents the probabilty that a Fixed Sample Size design produces strong misleading evidence for μ_1 over μ_0 with $n = 1$ observation. The Tepee function is the analogous probability under the Biased Open design where the sample size is allowed to grow until strong evidence for H_1 is obtained. Notice that for large alternatives ($c > \sqrt{2 \ln k} = 2.04$) the Tepee function provides values that are only a little larger than the Bump function. For distant alternatives (greater than 3 or 4 standard deviations) the Bump function shows that there is little chance of a single observation representing strong misleading evidence for μ_1 over μ_0. Furthermore, the Tepee function shows that, this probability cannot be substantially increased, even if we continue to sample until such strong misleading evidence is obtained.

For $c < \sqrt{2 \ln k} = 2.04$ the probability increases steadily as c approaches zero. Interestingly enough, $\lim_{c \to 0} P_0(N < \infty) = \frac{1}{k}$ but $P_0(N < \infty) = 0$ at $c = 0$. Hence the Tepee function is really discontinuous at $c = 0$ and never achieves the Universal bound. This discontinuity is not reflected in Figure 2.2.

Lastly, when μ_1 is the true mean, the probability that the Biased Open design will produce strong evidence for μ_1 over μ_0 is unity because there is no limit on the sample size. The Biased Open design will continue to sample until it produces strong evidence supporting μ_1 over μ_0. Intuitively, the Law of Large Numbers suggests that this is correct because the likelihood ratio will converge to infinity under H_1. Mathematically we have

$$P_1\left(\frac{L_n(\mu_1)}{L_n(\mu_0)} \geq k \text{ ; for any } n = 1, 2, \ldots\right) = P_1(N < \infty) = 1 \tag{2.8}$$

The proof of this result is included at the end of this section.

Unfortunately, it is impossible to implement any Open design in practice because the design may require too large a sample size. In the following section the class of Truncated designs are considered.

Proof (Theorem 2.1):

Without loss of generality take $\mu_0 = 0$ and $\sigma^2 = 1$. It suffices to consider the case when $\mu_1 > 0$. Then we have

$$\begin{aligned} N &= \inf\left\{n : n \geq 1, \frac{L_n(\mu_1)}{L_n(0)} \geq k\right\} \\ &= \inf\left\{n : n \geq 1, \mu_1 S_n - n\frac{\mu_1^2}{2} \geq \ln k\right\} \\ &= \inf\left\{n : n \geq 1, S_n \geq \frac{\ln k}{\mu_1} + n\frac{\mu_1}{2}\right\} \end{aligned} \tag{2.9}$$

If $\tilde{\tau}$ represents the corrected Brownian motion process from section A.3.3, we have by equation (A.10)

$$\begin{aligned} P_0\left(N<\infty\right) &= P_0\left\{\tilde{\tau}<\infty\right\} \\ &= \lim_{m\to\infty} P_0\left\{\tilde{\tau}\leq m\right\} \\ &\cong \lim_{m\to\infty}\Phi\left[-\frac{\ln k}{\Delta}m^{-\frac{1}{2}}-\frac{\Delta}{2}m^{\frac{1}{2}}\right] \\ &\quad + \lim_{m\to\infty}\left\{\frac{\exp\left\{-\rho\Delta\right\}}{k}\Phi\left[-\left(\frac{\ln k}{\Delta}+2\rho\right)m^{-\frac{1}{2}}+\frac{\Delta}{2}m^{\frac{1}{2}}\right]\right\} \\ &= \frac{\exp\left\{-\rho\Delta\right\}}{k} \end{aligned} \tag{2.10}$$

Replace Δ with Δ/σ where $\Delta=|\mu_1-\mu_0|$ for complete generality. To prove number two of Theorem 2.1 consider an analogous argument when H_1 is true. By equation (A.11) we have

$$\begin{aligned} P_1\left(N<\infty\right) &= P_1\left\{\tilde{\tau}<\infty\right\} \\ &= \lim_{m\to\infty} P_1\left\{\tilde{\tau}\leq m\right\} \\ &\cong \lim_{m\to\infty}\Phi\left[-\frac{\ln k}{\Delta}m^{-\frac{1}{2}}+\frac{\Delta}{2}m^{\frac{1}{2}}\right] \\ &\quad + \lim_{m\to\infty} k\exp\left\{\rho\Delta\right\}\Phi\left[-\left(\frac{\ln k}{\Delta}+2\rho\right)m^{-\frac{1}{2}}-\frac{\Delta}{2}m^{\frac{1}{2}}\right] \\ &= 1 \end{aligned} \tag{2.11}$$

QED •

The Law of Large numbers also yields this result. The reader is encouraged

to consult the appendix for a detailed description and explanation of the tools used to complete the proof.

2.4 Class III: Truncated Designs

Variables beyond statistical control, i.e. time, money, and participant availability, often restrict the sample size of a study. In situations where the sample size is restricted, Open designs are not executable. For these situations Truncated designs are often used. The class of Truncated designs refers to experimental designs where the sample size may increase, but not exceed a pre-determined upper limit of m observations.

As discussed in section 2.1, we examine the stopping rule whose purpose is to generate strong evidence for H_1 over H_0, where both simple hypotheses are fixed. This stopping rule ends the study at the smallest sample size where strong evidence supporting H_1 over H_0 is obtained. The point at which the study stops is now a random variable called a stopping time, already defined in equation (2.5) as N. A Truncated design employing stopping time N continues sampling until strong evidence for H_1 over H_0 is obtained or m observations are collected, whichever occurs first. This design is also a biased design favoring H_1. Henceforth, a Truncated design employing this stopping rule will be called a Biased Truncated design.

In contrast with the Biased Open design, the Biased Truncated design may generate strong evidence supporting H_0 over H_1 or even weak evidence. However, weak evidence can only be collected if strong evidence for H_1 over H_0 is not, e.g. the study is forced to terminate on the m^{th} observation regardless of

the observed evidence.

The Biased Truncated design maximizes the chance of generating misleading evidence for H_1 out of the entire class of Truncated designs with an upper limit of m observations or less. The probability that the Biased Truncated design will generate strong misleading evidence is less than the corresponding probability for the Biased Open design (Tepee function, equation 2.7), and greater than the corresponding probability for the Fixed Sample Size design (Bump function, equation 2.2). In fact, the Tepee function is the limiting probability of generating misleading evidence for H_1 under the Biased Truncated design as m increases to infinity.

The probability that the Biased Truncated design produces strong misleading evidence for H_1 over H_0 can be expressed in terms of the probability that the stopping time N does not exceed m observations. Therefore we have

$$P_0\left(\frac{L_n(\mu_1)}{L_n(\mu_0)} \geq k \quad ; \text{for any } n = 1, 2, \ldots, m\right) = P_0\left(N \leq m\right)$$

Directly from the appendix, equation (A.10), with Δ/σ replacing Δ for complete generality, we have

$$\begin{aligned} P_0\left(N \leq m\right) &= P_0\left\{\tilde{\tau} \leq m\right\} \\ &\cong \Phi\left[-\frac{\sigma \ln k}{\Delta} m^{-\frac{1}{2}} - \frac{\Delta}{2\sigma} m^{\frac{1}{2}}\right] \\ &\quad + \frac{\exp\left\{-\rho\Delta/\sigma\right\}}{k} \Phi\left[-\left(\frac{\sigma \ln k}{\Delta} + 2\rho\right) m^{-\frac{1}{2}} + \frac{\Delta}{2\sigma} m^{\frac{1}{2}}\right] \end{aligned} \tag{2.12}$$

where $\Delta = |\mu_1 - \mu_0|$ and $\rho \cong 0.583$. If the difference Δ is measured in units of standard deviations, say $\Delta = c\sigma$, then equation (2.12) becomes

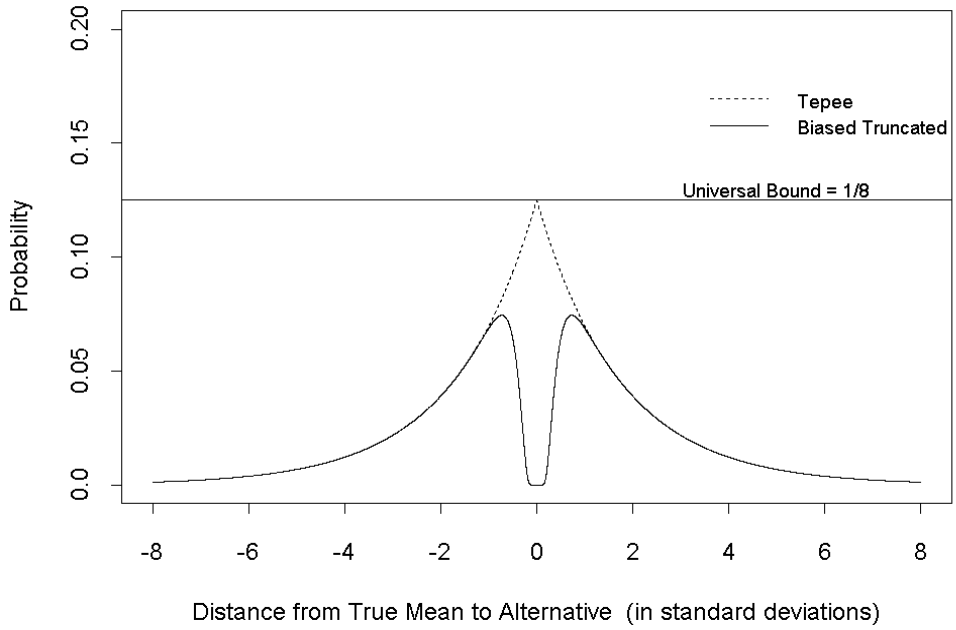

Figure 2.3: The Biased Truncated Design with $m = 20$ observations

$$P_0\left(N \leq m\right) \cong \Phi\left[-\frac{\ln k}{c}m^{-\frac{1}{2}} - \frac{c}{2}m^{\frac{1}{2}}\right] + \frac{\exp\left\{-\rho c\right\}}{k}\Phi\left[-\left(\frac{\ln k}{c} + 2\rho\right)m^{-\frac{1}{2}} + \frac{c}{2}m^{\frac{1}{2}}\right] \quad (2.13)$$

Figure 2.3 is a graph of the probability that the Biased Truncated design (solid line) generates misleading evidence when $m = 20$ and $k = 8$. The dotted line is the Tepee function plotted for comparison. For values of $\Delta > 1$ standard deviation, the probability of generating misleading evidence for the Biased Truncated design is almost that of the Biased Open design. This indicates that the probability of generating misleading evidence for alternative values one standard deviation or further from the true value accumulates early, at least before the 20^{th} observation. Also, notice also the Bump-like shape for

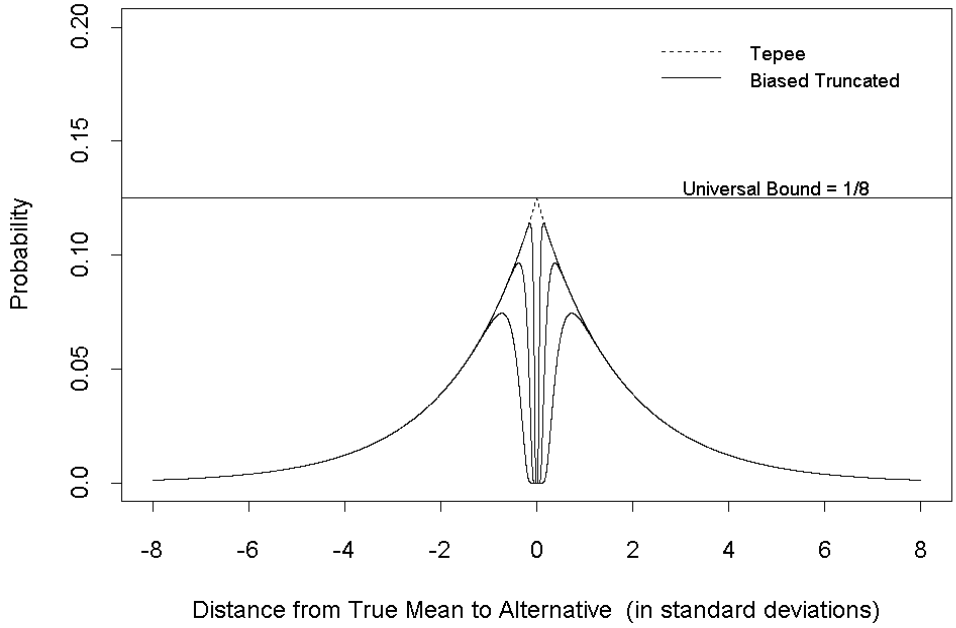

Figure 2.4: The Biased Truncated Design with m=20,100,1000 observations

this Biased Truncated design.

Figure 2.4 displays the probability of generating misleading evidence for H_1 from three different Biased Truncated designs. The Truncated designs have maximum sample sizes of 20, 100, and 1,000. Notice that as m increases, the shape of the curve for the Biased Truncated design approaches the limiting Tepee shape. Also notice how quickly the probabilities increase for alternatives in a small neighborhood of the true value (Δ close to zero).

The probabilty that the Biased Truncated design produces strong evidence for H_1 over H_0 when μ_1 is the true mean is

$$P_1\left(\frac{L_n(\mu_1)}{L_n(\mu_0)} \geq k \ \text{ ; for any } n = 1, 2, \ldots, m\right) = P_1\left(N \leq m\right) = P_1\left\{\tilde{\tau} \leq m\right\}$$

$$\cong \ \Phi\left[-\frac{\sigma \ln k}{\Delta} m^{-\frac{1}{2}} + \frac{\Delta}{2\sigma} m^{\frac{1}{2}}\right]$$

$$+k \exp\left\{\rho\Delta/\sigma\right\} \Phi\left[-\left(\frac{\sigma \ln k}{\Delta} + 2\rho\right) m^{-\frac{1}{2}} - \frac{\Delta}{2\sigma} m^{\frac{1}{2}}\right] \tag{2.14}$$

where $\Delta = |\mu_1 - \mu_0|$ and $\rho \cong 0.583$. Equation (2.14) is directly from the appendix, equation (A.11), with Δ/σ replacing Δ for complete generality.

The Biased Truncated design is effectively a Biased Open design with a limit on the sample size. Therefore, the probability that the Biased Truncated design generates misleading evidence depends on the pre-determined limit on the sample size. Interestingly, figure 2.3 shows that for the Biased Open design the chance of generating misleading evidence for alternatives 'far' from the true value accumulates early (before the m^{th} observation) and for alternatives 'near' the true value accumulates late (after the m^{th} observation). To explore this phenomona in detail we consider another class of designs called Delayed designs, which do not allow the study to stop before a certain number of observations have elapsed.

2.5 Class IV: Delayed Designs

The class of Delayed designs refers to experimental designs where the sample size must accumulate at least a pre-determined minimum number of observations, say m_0. A Delayed design first collects m_0 observations and then determines whether to continue sampling based on its stopping rule. In practice, a Delayed design might accomodate an investigator's unwillingness to stop a study with

only a small number of participants. This unwillingness may stem from a fear that the impact of the trial in the medical community will be small, due solely to the small sample size (even though the likelihood ratio in question may be large). This notion raises ethical issues which will not be addressed here. Statistically, the Delayed design allows the random sequence of log likelihood ratios to accumulate and stabilize for the first m_0 observations, effectively lowering the variability of the log likelihood ratio in the stopping rule.

Within the class of Delayed designs, we examine the stopping rule whose purpose is to generate strong evidence for H_1 over H_0, where both simple hypotheses are fixed. This stopping rule ends the study at the smallest sample size for which there is strong evidence supporting H_1 over H_0. The point at which the study stops is now a random variable called a stopping time, already defined in equation (2.5) as N. A Delayed design employing stopping time N continues sampling after m_0 have been collected until strong evidence for H_1 over H_0 is obtained. This Delayed design is biased in favor of H_1 and, like the Biased Open design, has the possibility of sampling forever. Henceforth, a Delayed design employing this stopping rule will be called a Biased Delayed design.

The Biased Delayed design maximizes the chance of generating misleading evidence for H_1 out of the entire class of Delayed designs requiring at least m_0 observations. The probability that the Biased Delayed design generates misleading evidence for H_1 over H_0 is

$$P_0\left(\frac{L_n(\mu_1)}{L_n(\mu_0)} \geq k \ \ ; \text{for any } n = m_0, m_0+1, \ldots\right)$$

$$= \quad P_0\left(m_0 \leq N < \infty\right)$$

$$
\begin{aligned}
&= P_0\{\tilde{\tau} < \infty\} - P_0(\tilde{\tau} \leq m_0 - 1) \\
&\cong \frac{\exp\{-\rho\Delta/\sigma\}}{k}\Phi\left[\left(\frac{\sigma \ln k}{\Delta} + 2\rho\right)(m_0-1)^{-\frac{1}{2}} - \frac{\Delta}{2\sigma}(m_0-1)^{\frac{1}{2}}\right] \\
&\quad -\Phi\left[-\frac{\sigma \ln k}{\Delta}(m_0-1)^{-\frac{1}{2}} - \frac{\Delta}{2\sigma}(m_0-1)^{\frac{1}{2}}\right] \qquad (2.15)
\end{aligned}
$$

using equation (A.10) where $\Delta = |\mu_1 - \mu_0|$ and $\rho \cong 0.583$. If the difference Δ is measured in units of standard deviations, say $\Delta = c\sigma$, then equation (2.15) reduces to

$$
\begin{aligned}
P_0(m_0 \leq N < \infty) &\cong \frac{\exp\{-\rho c\}}{k}\Phi\left[\left(\frac{\ln k}{c} + 2\rho\right)(m_0-1)^{-\frac{1}{2}} - \frac{c}{2}(m_0-1)^{\frac{1}{2}}\right] \\
&\quad -\Phi\left[-\frac{\ln k}{c}(m_0-1)^{-\frac{1}{2}} - \frac{c}{2}(m_0-1)^{\frac{1}{2}}\right] \qquad (2.16)
\end{aligned}
$$

Figure 2.5 is a graph of the probability that the Biased Delayed design produces misleading evidence, equation (2.16) when $m_0 = 5$ and $k = 8$. The dotted line is the Tepee function plotted for reference. An interesting feature of this graph is that the probability, for the Biased Delayed design at large alternatives ($\Delta > 2$ standard deviations), is essentially zero. This indicates that almost the entire probability of generating misleading evidence for large alternatives accumulates extremely early in the sequence of observations, at least before the fifth observation. By contrast, the probability of generating misleading evidence for alternatives close to the true value ($\Delta < 1/2$ standard deviations) is hardly reduced from the Tepee function by the delaying. This indicates that for alternatives close to the true value, the chance of generating misleading evidence accumulates later in the sequence of observations.

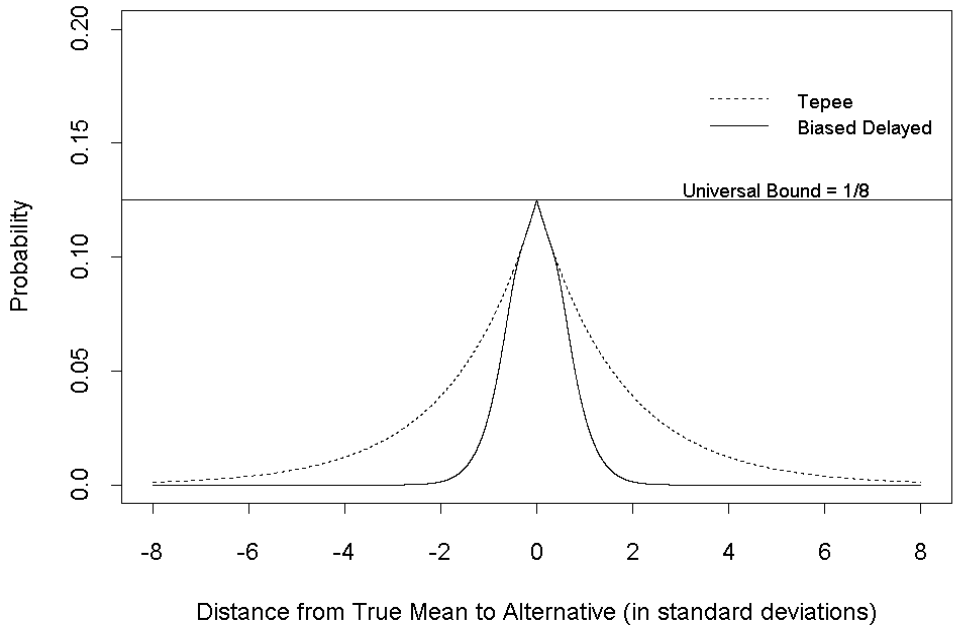

Figure 2.5: The Biased Delayed Design with $m_0 = 5$ observations

Figure 2.6 is a graph of the Biased Delayed design for various values of $m_0 = 5, 20, 70$. Notice that the peak of the curve slims out as m_0 increases. This has to be the case because the Tepee function is the limiting case of the Biased Delayed design when m_0 approaches to zero. The Biased Delayed design provides a way of reducing the probability of generating misleading evidence for alternatives not in a small neighborhood of the true value.

The probability that the Biased Delayed design will produce strong evidence for H_1 over H_0 when μ_1 is the true mean is

$$P_1\left(\frac{L_n(\mu_1)}{L_n(\mu_0)} \geq k \ \ ; \text{for any } n = m_0, m_0+1, \ldots\right)$$

$$= \quad P_1\left(m_0 \leq N < \infty\right)$$

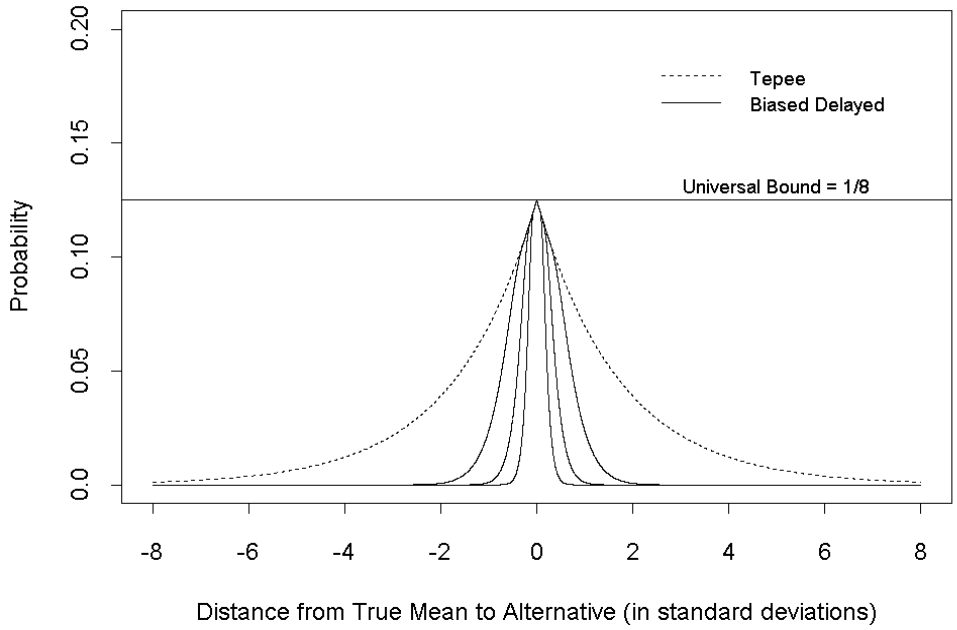

Figure 2.6: The Biased Delayed Design with $m_0 = 5, 20, 70$ observations

$$
\begin{aligned}
&= P_1\left\{\tilde{\tau} < \infty\right\} - P_1\left(\tilde{\tau} \leq m_0 - 1\right) \\
&\cong \Phi\left[\frac{\sigma \ln k}{\Delta}(m_0-1)^{-\frac{1}{2}} - \frac{\Delta}{2\sigma}(m_0-1)^{\frac{1}{2}}\right] \qquad (2.17) \\
&\quad -k\exp\left\{\rho\Delta/\sigma\right\}\Phi\left[-\left(\frac{\sigma \ln k}{\Delta} + 2\rho\right)(m_0-1)^{-\frac{1}{2}} - \frac{\Delta}{2\sigma}(m_0-1)^{\frac{1}{2}}\right]
\end{aligned}
$$

where $\Delta = |\mu_1 - \mu_0|$ and $\rho \cong 0.583$. Using equation (A.11) from the appendix with Δ/σ replacing Δ yields the result.

Finally, in situations where the sample size is restricted, Delayed designs are not executable. Combining the qualities of both a Delayed design and a Truncated design yields the Interval design, considered in the next section.

2.6 Class V: Interval Designs

The class of Interval designs refers to experimental designs where the sample size must accumulate at least m_0 observations, but may not exceed m observations. An Interval design first collects m_0 observations, and then determines whether to continue sampling based on its stopping rule until m observations are collected, at which point the study terminates regardless of the statistical evidence. An Interval design is effectively a combination of a Delayed design and a Truncated design, guaranteeing at least m_0 observations but no more than m observations. The class of Interval designs is the most versatile of all the classes of designs discussed in this chapter.

Within the class of Interval designs, we examine the stopping rule with the purpose of generating strong evidence for H_1 over H_0, where both simple hypotheses are fixed. This stopping rule ends the study at the smallest sample size where strong evidence supporting H_1 over H_0 is obtained. The point at which the study stops is now a random variable called a stopping time, already defined as N. An Interval design employing stopping time N continues sampling after m_0 have been collected until either strong evidence for H_1 over H_0 is obtained or m observations are collected, whichever occurs first. This design is also a biased design favoring H_1. Henceforth, an Interval design employing this stopping rule will be called a Biased Interval design.

It is possible for the Biased Interval design to generate strong evidence for H_0 over H_1 or even weak evidence. This would happen if strong evidence for H_1 over H_0 is not obtained between the m_0^{th} and m^{th} observations. The Biased Interval design maximizes the chance of generating misleading evidence for H_1 over H_0

out of the entire class of Intervals designs allowing at least m_0 observations and no more than m observations. The probability that the Interval design generates strong misleading evidence for H_1 is the probability that N is between m_0 and m. Hence we have

$$\begin{aligned}
& P_0\left(\frac{L_n(\mu_1)}{L_n(\mu_0)} \geq k \ \ ; \text{for any } n = m_0, m_0+1, ..., m\right) \\
= \ & P_0\left(m_0 \leq N \leq m\right) \\
= \ & P_0\left\{\tilde{\tau} \leq m\right\} - P_0\left(\tilde{\tau} \leq m_0 - 1\right) \\
\cong \ & \Phi\left[-\frac{\sigma \ln k}{\Delta} m^{-\frac{1}{2}} - \frac{\Delta}{2\sigma} m^{\frac{1}{2}}\right] - \Phi\left[-\frac{\sigma \ln k}{\Delta}(m_0-1)^{-\frac{1}{2}} - \frac{\Delta}{2\sigma}(m_0-1)^{\frac{1}{2}}\right] \\
& + \frac{\exp\left\{-\rho\Delta/\sigma\right\}}{k}\left\{\Phi\left[-\left(\frac{\sigma \ln k}{\Delta} + 2\rho\right) m^{-\frac{1}{2}} + \frac{\Delta}{2\sigma} m^{\frac{1}{2}}\right]\right. \\
& \qquad \left. - \Phi\left[-\left(\frac{\sigma \ln k}{\Delta} + 2\rho\right)(m_0-1)^{-\frac{1}{2}} + \frac{\Delta}{2\sigma}(m_0-1)^{\frac{1}{2}}\right]\right\} \qquad (2.18)
\end{aligned}$$

where $\Delta = |\mu_1 - \mu_0|$ and $\rho \cong 0.583$. Using equation (A.10) from the appendix with Δ replaced with Δ/σ yields the result. If the difference Δ is measured in units of standard deviations, say $\Delta = c\sigma$, then equation (2.18) becomes

$$\begin{aligned}
P_0\left(m_0 \leq N \leq m\right) \ \cong \ & \Phi\left[-\frac{\ln k}{c} m^{-\frac{1}{2}} - \frac{c}{2} m^{\frac{1}{2}}\right] - \Phi\left[-\frac{\ln k}{c}(m_0-1)^{-\frac{1}{2}} - \frac{c}{2}(m_0-1)^{\frac{1}{2}}\right] \\
& + \frac{\exp\left\{-\rho c\right\}}{k}\left\{\Phi\left[-\left(\frac{\ln k}{c} + 2\rho\right) m^{-\frac{1}{2}} + \frac{c}{2} m^{\frac{1}{2}}\right]\right. \\
& \qquad \left. - \Phi\left[-\left(\frac{\ln k}{c} + 2\rho\right)(m_0-1)^{-\frac{1}{2}} + \frac{c}{2}(m_0-1)^{\frac{1}{2}}\right]\right\} \qquad (2.19)
\end{aligned}$$

Figure 2.7 is a graph of the Biased Interval design when $m_0 = 6$ and $m = 100$. The dotted line is the Tepee function. Notice that the Biased Interval design

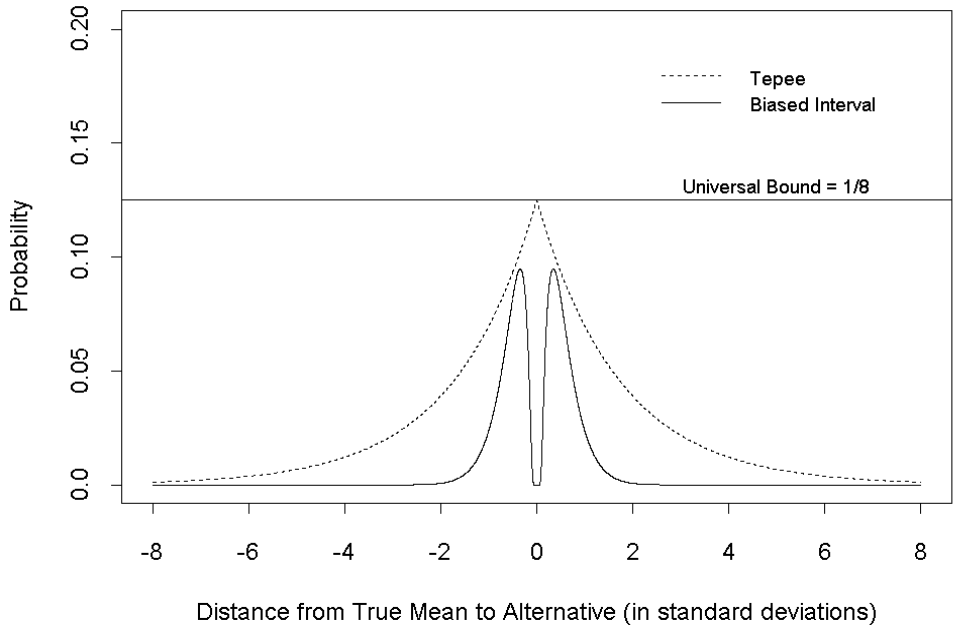

Figure 2.7: The Biased Interval Design with $m_0 = 6$ and $m = 100$ observations

produces a Bump function shape for the probability of generating misleading evidence. In fact, any 'Bump' like shape is possible by varying m_0 and m. The tails of the Bump shape are caused by the 'delaying' and the valley is caused by the 'truncating'.

Lastly, the probability that the Biased Interval design produces strong evidence for H_1 over H_0 when μ_1 is the true mean is

$$P_1\left(\frac{L_n(\mu_1)}{L_n(\mu_0)} \geq k \;\; ; \text{for any } n = m_0, m_0+1, ..., m\right)$$

$$= \quad P_1(m_0 \leq N \leq m)$$

$$= \quad P_1\{\tilde{\tau} \leq m\} - P_1(\tilde{\tau} \leq m_0 - 1)$$

$$\begin{aligned}
\cong\ & \Phi\left[-\frac{\sigma \ln k}{\Delta} m^{-\frac{1}{2}} + \frac{\Delta}{2\sigma} m^{\frac{1}{2}}\right] - \Phi\left[-\frac{\sigma \ln k}{\Delta}(m_0-1)^{-\frac{1}{2}} + \frac{\Delta}{2\sigma}(m_0-1)^{\frac{1}{2}}\right] \\
& + k \exp\{\rho\Delta/\sigma\} \left\{ \Phi\left[-\left(\frac{\sigma \ln k}{\Delta} + 2\rho\right) m^{-\frac{1}{2}} - \frac{\Delta}{2\sigma} m^{\frac{1}{2}}\right] \right. \\
& \left. - \Phi\left[-\left(\frac{\sigma \ln k}{\Delta} + 2\rho\right)(m_0-1)^{-\frac{1}{2}} - \frac{\Delta}{2\sigma}(m_0-1)^{\frac{1}{2}}\right] \right\} \qquad (2.20)
\end{aligned}$$

Using equation (A.11) from the appendix with Δ replaced with Δ/σ for complete generality and $\Delta = |\mu_1 - \mu_0|$ and $\rho \cong 0.583$.

2.7 Summary

According to the Law of Likelihood, the strength of the evidence in a given body of observations is not affected by the stopping rule. However, the planning probabilities (i.e. the probability that a particular design will generate misleading, weak, or strong evidence) depend upon the stopping rule. Study Designs that incorporate continuous monitoring of the data, with termination of the study as soon as strong evidence in one direction is observed, have a greater probability of generating misleading evidence than their Fixed Sample Size counterparts. However, the maximum probability, over all designs, that misleading evidence will be generated for a fixed simple alternative is, in general, small. This maximum probability is shown by the Tepee function to be well below the Universal bound for most alternatives (see section 2.3). Hence the practical and ethical advantages of employing study designs that incorporate continuous monitoring of the evidence with termination of the trial when strong evidence is first obtained, far outweigh the small and controllable increase in the probability of generating misleading evidence.

Chapter 3

Composite Hypotheses

3.1 Introduction

In Chapter 2, we examined a stopping rule whose purpose is to generate strong evidence for H_1 over H_0, where both H_1 and H_0 are fixed simple hypotheses. That stopping rule ends the study at the smallest sample size allowed by the design where strong evidence supporting H_1 over H_0 is obtained (equation (2.5)). The stopping rule favors H_1 and has the greatest chance of generating strong evidence for H_1 over H_0. This stopping rule is subject to criticism, however, because of its focus on the single simple alternative H_1. A more flexible stopping rule, based on likelihood ratios, is considered in this chapter. We now ask whether, for a given simple H_0 and a sequence of observations, some hypothesis (within a specified set of alternatives) can be found that is better supported than H_0.

More precisely, within each class of designs, we study the stopping rule whose purpose is to generate strong evidence for *any simple* alternative from a set (represented by the fixed composite hypothesis H_C) over the fixed simple hypothesis H_0. This stopping rule ends the study at the smallest sample size

allowed by the design whenever any simple alternative in H_C is better supported over H_0 by a factor of k or more. Experimental designs incorporating this stopping rule also entail continuous monitoring of the evidence. The study stops, not after collecting a predetermined number of observations, but when the observations themselves represent sufficiently strong evidence for some simple alternative in H_C over H_0. This stopping rule favors a range of hypotheses and, in that sense, is more biased than the stopping rule considered in Chapter 2.

For example, consider the problem of generating evidence about a real-valued parameter of interest $\theta \in \Theta$, where $L_n(\theta)$ is the likelihood function based on n observations. Now suppose we are interested in generating evidence for any simple alternative in $H_C : \theta \in \Theta_C \subseteq \Theta$ over $H_0 : \theta = \theta_0$. The stopping time associated with the above stopping rule is then defined as

$$N = \min \left\{ n : n \geq 1, \sup_{\theta \in \Theta_C} \frac{L_n(\theta)}{L_n(\theta_0)} \geq k \right\} \tag{3.1}$$

The random stopping time (3.1) identifies the smallest sample size at which the evidence supports some simple alternative in the set Θ_C over θ_0 by a factor of k or more.

It is tempting to interpret the quantity

$$\sup_{\theta \in \Theta_C} \frac{L_n(\theta)}{L_n(\theta_0)} \stackrel{\text{def}}{=} L^* \tag{3.2}$$

as a measure of evidence for H_C over H_0. However this interpretation is neither implied nor sanctioned by the Law of Likelihood. The Law states:

> If H_C and H_0 imply that the probability that a random variable X takes the value x is $P_C(x)$ and $P_0(x)$ respectively, then the observation $X = x$ is evidence supporting H_C over H_0 if and only if

> $P_C(x)/P_0(x) > 1$ and the likelihood ratio $P_C(x)/P_0(x)$ measures the strength of that evidence.

To measure the evidence for H_C over H_0, the probability model must provide for calculation of $P_C(x)$ [12], effectively reducing the composite hypothesis to a simple hypothesis. (This is, in essence, what a Bayesian analysis does; the prior distribution provides a mechanism for reducing the composite hypothesis to a simple hypothesis and then the Law of Likelihood is applied to the resulting Bayes factor.) But L^* fails to satisfy this condition because

$$\int P_C(X)dX = \int \sup_{\theta\in\Theta_C} \prod_{i=1}^{n} f(X_i;\theta)dX_n = \int \prod_{i=1}^{n} f(X_i;\theta_n(X))dX_n \neq 1$$

For example, one observation from a normal distribution yields $\theta_n(X) = \mu_1(X) = X$ and the resulting integral is infinite. Thus, L^* does not specify a vaild probability measure for $P_C(X)$ and cannot be interpreted as measuring the evidence for H_C over H_0.

More importantly, observations that produce $L^* = k$ only sometimes represent k-strength evidence in favor of H_C over H_0 and often support H_0 over H_C [11]. In fact, if the prior probability of H_0 is non-zero, then the posterior probability of H_0 can be increased by such observations. In light of this, we refrain from calling such observations misleading. Instead we refer to them as they are: observations supporting some simple alternative in a specified set over H_0 by a factor of k or more. These observations may be enticing or inveigling, but they are nothing more. An example, due to Armitage, which bears directly on this issue and has been important in discussions on the foundations of statistical inference, is examined in the next section.

Sections three through eight examine in detail the probability of generating strong evidence for *some* simple alternative (within a specified set) over a fixed simple hypothesis H_0. Throughout this chapter, the five classes of experimental designs studied in Chapter 2 are examined under the model where $X_1, X_2, \ldots, X_n, \ldots$ are i.i.d. $N(\mu, \sigma^2)$ where σ^2 is known. Let $L_n(\mu) = \prod f(x_i; \mu)$ represent the corresponding likelihood function for n observations and let $S_n = X_1 + \cdots + X_n$.

3.2 Armitage's Paradox

Consider the simple hypothesis $H_0 : \mu = 0$ and the composite alternative $H_C : \mu \neq 0$. Suppose an Open design is employed so that sampling continues until strong evidence for some $\mu \neq 0$ over $\mu = 0$ is obtained. Then stopping time (3.1) becomes

$$N = \min \left\{ n : n \geq 1, \sup_{\mu} \frac{L_n(\mu)}{L_n(0)} \geq k \right\} \tag{3.3}$$

Armitage's paradox states that $P_0(N < \infty) = 1$, but is most commomly cited in terms of repeated significance tests on a true null hypothesis [2],[1]. The result is actually a consequence of the Law of the Iterated Logarithm [22]. See also Feller [16] who is given credit for identifying the 'problem' and Armitage [3] for later developments. First we show that, indeed, $P_0(N < \infty) = 1$ for any $k > 1$.

Theorem 3.1 ***(The Law of the Iterated Logarithm)***. *Suppose observations $X_1, X_2, \ldots$ are i.i.d. with $E[X_1] = 0$ and $0 < \sigma^2 = Var(X_1) < \infty$. Then,*

$$\limsup_{n \to \infty} \frac{S_n}{\sqrt{2n \log \log n}} \overset{\text{a.s.}}{=} \sigma$$

The Law of the Iterated Logarithm states that the supremum of the random sequence $S_n; n = 1, 2, \ldots$ grows at the rate of $\sqrt{2n \log\log n}$. When the supremum of the random sum S_n is scaled by exactly $\sqrt{2n \log\log n}$, their ratio converges to σ. However, if the supremum of the random sequence $S_n; n = 1, 2, \ldots$ is scaled by something less than $\sqrt{2n \log\log n}$, their ratio will grow to infinity. For example, Theorem 3.1 indicates that the ratio $\sup_n S_n/\sqrt{n}$ converges to infinity as n increases to infinity.

If $\mu = 0$ is the true mean, then the probability that an Open design, using stopping time (3.3), generates strong evidence for some alternative $\mu \neq 0$ over $\mu = 0$ is

$$\begin{aligned} P_0(N < \infty) &= P_0\left(\sup_{\mu} \frac{L_n(\mu)}{L_n(0)} \geq k \ ; \text{for any } n = 1, 2, \ldots\right) \\ &= P_0\left(\frac{|S_n|}{\sqrt{n}} \geq \sqrt{2\sigma^2 \ln k} \ ; \text{for any } n = 1, 2, \ldots\right) \\ &= 1 \quad \text{for any fixed } k > 1 \end{aligned} \tag{3.4}$$

The last line follows from the Law of the Iterated Logarithm. Equation (3.4) says, under H_0, with probability one, sooner or later it will be found that *some* value $\mu^* \neq 0$ is better supported than $\mu = 0$ by a factor of at least k.

One misinterpretation of equation (3.4) is that, when observations are generated under $H_0 : \mu = 0$, the likelihood stopping rule (3.3) will always produce 'strong evidence against H_0' [2],[1]. But according to the Law of Likelihood, observations only represent evidence in a relative fashion, never 'for' or 'against' a single hypothesis. It is true that, with probability one, some value $\mu^* \neq 0$ will eventually be better supported over the true mean $\mu = 0$ by a factor of k or more. However, it does not follow that the data are 'evidence against H_0'

because some value of $\mu = \mu^*$ is better supported over zero (discussion of this point can be found in [29, §1.7.1, §1.10]). Remember that $\mu = 0$ will be better supported over an infinite number of simple alternatives. Furthermore, more often that not, μ^* will be in a small neighborhood of zero where there is no practical difference between μ^* and zero. (This happens because the MLE $\bar{X}_n$ converges to zero.) In this case the data still indicate, for all practical purposes, that the true mean is zero.

Another supposed implication of equation (3.4) is that, when H_0 is true, finding some alternative $\mu^* \neq 0$ that is better supported over H_0 is misleading. However such observations are frequently not misleading because the posterior probability of H_0 increases. The following Bayesian analysis, due to Cornfield, demonstrates this point [11].

Cornfield argued that the appropriate way to entertain the possibility of the truth of H_0 is to assign H_0 a non-zero prior probability. Denote the prior probability of $H_0 : \mu = 0$ by p and the posterior probability of H_0, given t_n the mean of n observations, as $P(H_0|t_n)$. Without loss of generality take $\sigma^2 = 1$. Consider the prior for $\mu \neq 0$ to be

$$\frac{(1-p)}{\tau}\phi[\mu/\tau] \quad \text{for } \mu \neq 0$$

where

$$\phi[x] = \frac{1}{\sqrt{2\pi}} \exp\{-x^2/2\}$$

Then by Bayes' Theorem and some algebra, Cornfield gives

$$P\left(H_0|t_n\right) = \frac{p}{p + (1-p)LR} \tag{3.5}$$

where

$$LR = \frac{P\left(t_n|H_C\right)}{P\left(t_n|H_0\right)} = \frac{1}{\sqrt{n\tau^2+1}} \exp\left\{\frac{n^2\tau^2 t_n^2}{2(n\tau^2+1)}\right\} \tag{3.6}$$

The likelihood ratio for H_C to H_0 will be less than 1 when

$$\left(\sqrt{n}t_n\right)^2 < \left(1+\frac{1}{n\tau^2}\right)\ln(n\tau^2+1)$$

and hence the posterior probability of H_0 increases when this condition is met. Furthermore, when some $\mu^* \neq 0$ is better supported over $\mu = 0$ by a factor of k or more we have

$$\frac{L_n(t_n)}{L_n(0)} \geq k$$

which happens when

$$\left(\sqrt{n}t_n\right)^2 \geq 2\ln k$$

Therefore, when $\mu = 0$, observations representing k strength evidence for some $\mu^* \neq 0$ over $\mu = 0$ that increase the posterior probability of H_0 satisfy

$$\sqrt{2\ln k} \leq \sqrt{n}|t_n| < \sqrt{\left(1+\frac{1}{n\tau^2}\right)\ln(n\tau^2+1)} \tag{3.7}$$

For large n, condition (3.7) implies that the posterior probability of H_0 will almost always increase when some $\mu^* \neq 0$ is found to be better supported over $\mu = 0$ by a factor of k or more. This happens because the quantity on the RHS of condition (3.7) approaches infinity as n increases.

In the upcoming sections we restrict attention to the 'one-sided' case; can some simple alternative in $\mu > 0$ be found that is better supported over $\mu = 0$ by a factor of k or more. More precisely, we examine the case when $H_C : \mu \geq \mu_1$ and $H_0 : \mu = \mu_0$ where $\mu_1 = \mu_0 + \Delta$ for some fixed $\Delta > 0$. This provides a way to distinguish clinically important effects from those effects which might

be technically non-zero, but clinically no different from zero. For example, we are not interested in whether *any* hypothesis $\mu > 0$ is better supported over $\mu = 0$, only if any clinically important (effective) hypothesis, $\mu \geq \mu_1$, is better supported over $\mu = 0$. The difference $\Delta = \mu_1 - \mu_0$ is typically called an indifference zone.

3.3 Class I: Fixed Sample Size Designs

We begin by examining the probability that a fixed sample size experiment will produce some alternative in $\mu \geq \mu_1$ that is better supported over μ_0 where $\mu_1 = \mu_0 + \Delta$ ($\Delta > 0$ is fixed). The probability depends on the sample size and the distance between μ_0 and μ_1. Under a Fixed Sample Size design, the probability that some simple alternative in $H_C : \mu \geq \mu_1$ is better supported over $H_0 : \mu = \mu_0$ by a factor of k or more is

$$P_0\left(\sup_{\mu\geq\mu_1}\frac{L_n(\mu)}{L_n(\mu_0)}\geq k\right)=\begin{cases}\Phi\left[-\sqrt{2\ln k}\right] & 0<\Delta<\sqrt{2\sigma^2\ln k}/\sqrt{n}\\ \Phi\left[-\frac{\Delta\sqrt{n}}{2\sigma}-\frac{\sigma\ln k}{\Delta\sqrt{n}}\right] & \sqrt{2\sigma^2\ln k}/\sqrt{n}\leq\Delta\end{cases} \tag{3.8}$$

where $\Delta = \mu_1 - \mu_0$. A proof is included at the end of this section. If the difference Δ is measured in units of standard errors, say $\Delta = c\sigma/\sqrt{n}$, then equation (3.8) becomes

$$P_0\left(\sup_{\mu\geq\mu_1}\frac{L_n(\mu)}{L_n(\mu_0)}\geq k\right)=\begin{cases}\Phi\left[-\sqrt{2\ln k}\right] & 0<c<\sqrt{2\ln k}\\ \Phi\left[-\frac{c}{2}-\frac{\ln k}{c}\right] & \sqrt{2\ln k}\leq c\end{cases} \tag{3.9}$$

Formula (3.8) is called the C-bump function for Composite Bump function. Figure 3.1 is a graph of the probability that some simple alternative $\mu \geq \mu_1$ will be better supported over $\mu = \mu_0$, by a factor of $k = 8$ or more. On the x-axis is Δ, the distance between μ_1 and μ_0, measured in standard errors. Both the C-Bump function (solid line) and Bump function (dotted line) are plotted for

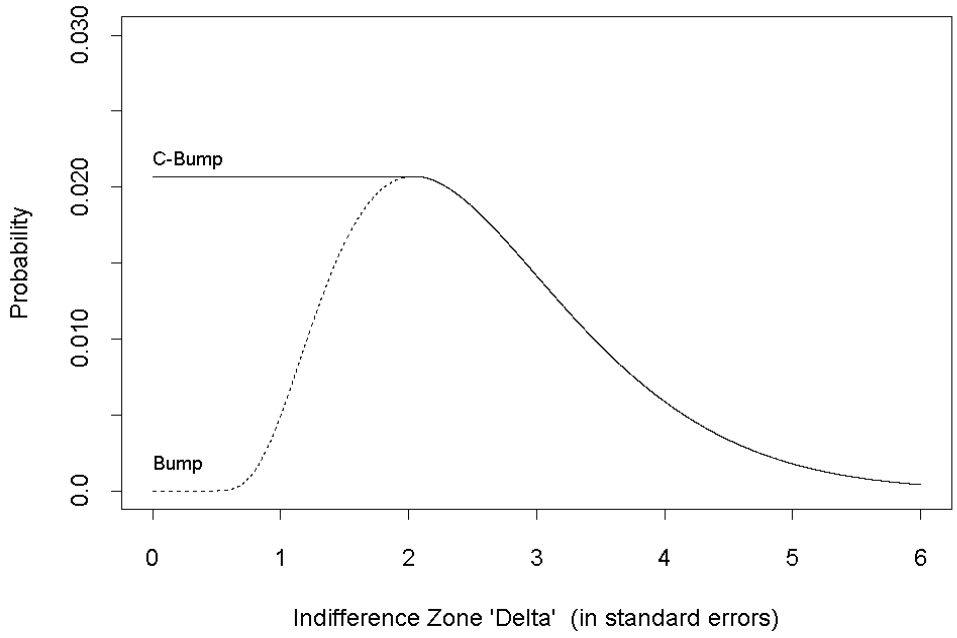

Figure 3.1: The C-Bump function

comparison purposes. Notice that the C-bump is exactly the same as the Bump function for $\Delta \geq \sqrt{2 \ln k}$ standard errors. For smaller Δ the probability remains constant at the maximum of the Bump function $\Phi\left[-\sqrt{2 \ln k}\right] = 0.021$ for $k = 8$. Remember, the Bump function is the probability, when $\mu = \mu_0$, of observing strong evidence for $\mu = \mu_1$ over $\mu = \mu_0$. By contrast, the C-Bump function is the probability of observing strong evidence that supports some alternative $\mu \geq \mu_1$ over $\mu = \mu_0$.

Theorem 3.1 *(C-Bump) Suppose observations $X_1, X_2, \ldots, X_n$ are i.i.d. Normal random variables with mean μ_0 and known variance σ^2. Then the probability of observing strong evidence for some $\mu \geq \mu_1 = \mu_0 + \Delta$ ($\Delta > 0$) over μ_0 is given by equation (3.8).*

Proof:

Without loss of generality take $\mu_0 = 0$. Then a simple calculation yields the following

$$\begin{aligned} P_0\left(\sup_{\mu\geq\Delta}\frac{L_n(\mu)}{L_n(0)}\geq k\right) &= P_0\left(\sup_{\mu\geq\Delta}\left\{\frac{n\mu}{\sigma^2}\left(\bar{X}_n-\frac{\mu}{2}\right)\right\}\geq\ln k\right) \\ &= P_0\left(\begin{array}{c}\left\{\bar{X}_n>\Delta \text{ and } \frac{n\bar{X}^2}{2\sigma^2}>\ln k\right\} \text{ or} \\ \left\{\bar{X}_n<\Delta \text{ and } \frac{n\Delta}{\sigma^2}\left(\bar{X}_n-\frac{\Delta}{2}\right)>\ln k\right\}\end{array}\right) \\ &= P_0\left(\begin{array}{c}\left\{\bar{X}_n>\max\left[\Delta,\frac{\sigma}{\sqrt{n}}\sqrt{2\ln k}\right]\right\} \text{ or} \\ \left\{\frac{\Delta}{2}+\frac{\sigma^2}{n\Delta}\ln k<\bar{X}_n<\Delta\right\}\end{array}\right)\end{aligned}$$

For $\left\{\frac{\Delta}{2}+\frac{\sigma^2}{n\Delta}\ln k<\bar{X}_n<\Delta\right\}$ to not be an empty set we need $\frac{\Delta}{2}+\frac{\sigma^2}{n\Delta}\ln k<\Delta$ which implies that $\Delta>\frac{\sigma}{\sqrt{n}}\sqrt{2\ln k}$, for all $\Delta>0$. Hence the two sets in the last line of the above equation are disjoint because $\max\left[\Delta,\frac{\sigma}{\sqrt{n}}\sqrt{2\ln k}\right]=\frac{\sigma}{\sqrt{n}}\sqrt{2\ln k}$, for all $\Delta<\frac{\sigma}{\sqrt{n}}\sqrt{2\ln k}$. Standardizing we have

$$P_0\left(\sup_{\mu\geq\Delta}\frac{L_n(\mu)}{L_n(0)}\geq k\right)=P_0\left(\frac{\sqrt{n}\bar{X}_n}{\sigma}\geq h_n(\Delta)\right) \tag{3.10}$$

where

$$h_n(\Delta)=\begin{cases}\sqrt{2\ln k} & 0<\Delta\sqrt{n}/\sigma<\sqrt{2\ln k} \\ \frac{\Delta\sqrt{n}}{2\sigma}+\frac{\sigma\ln k}{\Delta\sqrt{n}} & \sqrt{2\ln k}\leq\Delta\sqrt{n}/\sigma\end{cases} \tag{3.11}$$

Which yields equation (3.8) where $\Phi(\cdot)$ represents the standard normal CDF.

QED •

3.4 Class II: Open Designs

The class of Open designs refers to experimental designs which are open-ended, i.e. there are no restrictions on the sample size. As discussed in section (3.1), we examine the stopping rule whose purpose is to generate strong evidence for

some alternative in $H_C : \mu \geq \mu_1$ over $H_0 : \mu = \mu_0$ ($\mu_1 = \mu_0 + \Delta$) where both μ_0 and μ_1 are fixed. This stopping rule ends the study at the smallest sample size where some alternative $\mu \geq \mu_1$ is better supported over μ_0 by a factor of k or more. The point at which the study stops is now a random variable called a stopping time, defined as

$$N = \min \left\{ n : n \geq 1, \sup_{\mu \geq \mu_1} \frac{L_n(\mu)}{L_n(\mu_0)} \right\} \tag{3.12}$$

An Open design employing stopping time (3.12) continues sampling (possibly forever) until strong evidence for some alternative $\mu \geq \mu_1$ over μ_0 is obtained. Thus this design is a biased experimental design favoring hypotheses in H_C. Henceforth, an Open design employing this stopping rule will be called a Biased Composite Open Design or C-Biased Open design.

The probability that the C-Biased Open Design collects observations that support some alternative $\mu \geq \mu_1$ over μ_0 by a factor of k or more can be expressed as the probability that N is finite.

$$P_0 \left(\sup_{\mu \geq \mu_1} \frac{L_n(\mu)}{L_n(\mu_0)} \geq k \ \text{ ; for any } n = 1, 2, \ldots \right) = P_0(N < \infty) \tag{3.13}$$

1. For $\Delta < \sqrt{2\sigma^2 \ln k}$

$$\begin{aligned} P_0(N < \infty) \ &\leq \ \frac{\sqrt{\ln k}}{2k\sqrt{\pi}} \log \left(\frac{2\sigma^2 \ln k}{\Delta^2} \right) \\ &+ \ \frac{\exp\{-\rho\Delta/\sigma\}}{k} \Phi \left[\frac{\sigma \ln k + 2\rho\Delta}{A(\Delta)} - \frac{A(\Delta)}{2\sigma} \right] - \Phi \left[-\frac{\sigma \ln k}{A(\Delta)} - \frac{A(\Delta)}{2\sigma} \right] \end{aligned}$$

2. For $\Delta \geq \sqrt{2\sigma^2 \ln k}$

$$P_0(N < \infty) \cong \frac{\exp\{-\rho\Delta/\sigma\}}{k}$$

where $\Delta = \mu_1 - \mu_0, A(\Delta) = \sqrt{2\sigma^2 \ln k - \Delta^2}$, and $\rho \cong 0.583$. The proof is included at the end of this section. Now, if the difference Δ is measured in units of standard deviations, say $\Delta = c\sigma$, then equation (3.13) becomes

$$P_0\left(\sup_{\mu \geq \mu_1} \frac{L_n(\mu)}{L_n(\mu_0)} \geq k \ \text{; for any } n = 1, 2, \ldots\right) = P_0(N < \infty) \tag{3.14}$$

1. For $c < \sqrt{2 \ln k}$

$$\begin{aligned} P_0(N < \infty) &\leq \frac{\sqrt{\ln k}}{2k\sqrt{\pi}} \log\left(\frac{2 \ln k}{c^2}\right) \\ &+ \frac{\exp\{-\rho c\}}{k} \Phi\left[\frac{\ln k + 2\rho c}{A(c)} - \frac{A(c)}{2}\right] - \Phi\left[-\frac{\ln k}{A(c)} - \frac{A(c)}{2}\right] \end{aligned}$$

2. For $c \geq \sqrt{2 \ln k}$

$$P_0(N < \infty) \cong \frac{\exp\{-\rho c\}}{k}$$

where $A(c) = \sqrt{2 \ln k - c^2}$ and $\rho \cong 0.583$.

Another bound on $P_0(N < \infty)$ was derived by Gary Lorden in a 1973 paper [24]. Lorden's bound is

$$P_0\left(N < \infty\right) \leq \frac{1}{k}\left[1 + \frac{\sqrt{\log k}}{2\sqrt{\pi}} \frac{1}{2} \log\left(\frac{2 \ln k}{c^2}\right)\right] \tag{3.15}$$

where $c = \Delta/\sigma$ and $\Delta = \mu_1 - \mu_0$. The reader is referred to his paper for a proof of the result. Equation (3.14) is smaller than Lorden's bound (3.15). The reason for this is that equation (3.14) approximates the actual probability for $c \geq \sqrt{2 \ln k}$ standard deviations, while Lorden bounds this portion of the overall probability with the universal bound.

Figure 3.2 is a graph of the probability, under H_0, that the C-biased Open design (solid line) generates evidence for some simple alternative $\mu \geq \mu_1$ over μ_0.

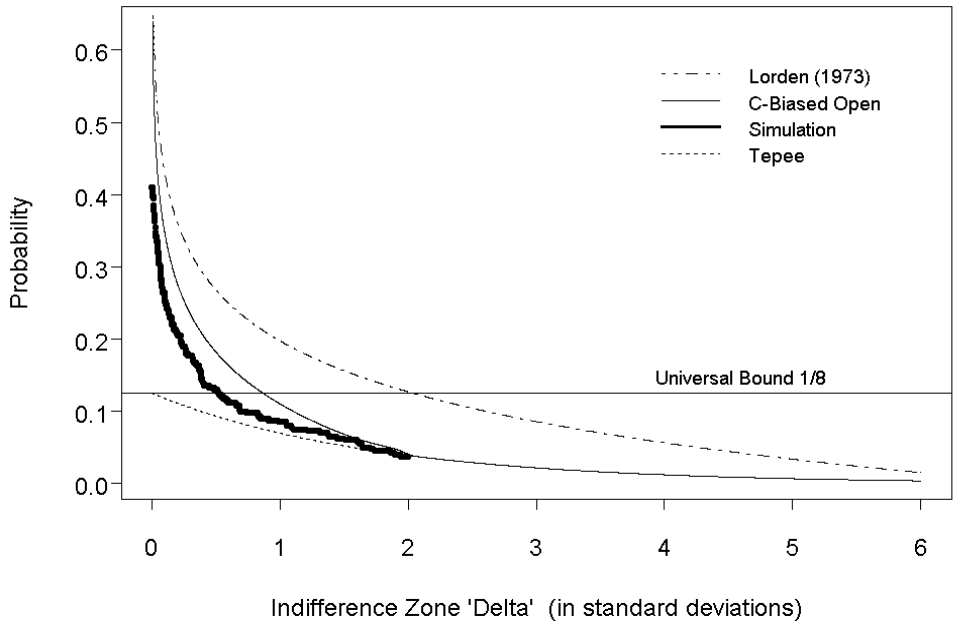

Figure 3.2: The C-Biased Open design

Lorden's bound (dashed/dotted line), the Tepee function (dotted line), and a S-plus simulation (large dots) of the probability are plotted for referance purposes. The S-plus simulation gives a visual impression of the tightness of the bound for the C-Biased Open design. However, the simulation has the limitation of finite sampling and thus represents a lower bound on the true probability. Notice that the estimated probability under the simulation, the Tepee function, and the C-Biased Open design all converge at $\sqrt{2 \ln k}$ standard deviations. This is a natural consequence of our stopping rule and the restricted alternative set as can been seen from the following proof.

Theorem 3.2 *(C-Biased Open Design). For $X_1, X_2, \ldots$ i.i.d. Normal random variables with mean μ_0 and known variance σ^2, the probability that the C-Biased*

Open design generates evidence for some alternative $\mu \geq \mu_1$ over $\mu = \mu_0$ is given by equation (3.14).

Proof:

Without loss of generality take $\mu_0 = 0$ and $\sigma^2 = 1$. Stopping time (3.12) reduces to

$$\begin{aligned} N &= \inf\left\{n : n \geq 1, \sup_{\mu \geq \Delta}\left[\frac{L_n(\mu)}{L_n(0)}\right] \geq k\right\} \\ &= \inf\left\{n : n \geq 1, \sup_{\mu \geq \Delta}\left[\mu S_n - n\frac{\mu^2}{2}\right] \geq \ln k\right\} \\ &= \inf\left\{n : n \geq 1, S_n \geq \inf_{\mu \geq \Delta}\left[\frac{\ln k}{\mu} + n\frac{\mu}{2}\right]\right\} \end{aligned} \tag{3.16}$$

Now consider that $\frac{d}{d\mu}\left[\frac{\ln k}{\mu} + n\frac{\mu}{2}\right] = 0$ when $n = \frac{2\ln k}{\mu^2}$. For $\mu > 0$ the second derivative is positive, indicating a minimum at $n = \frac{2\ln k}{\mu^2}$. Hence for $\mu > \sqrt{\frac{2\ln k}{n}}$ the function $\left[\frac{\ln k}{\mu} + n\frac{\mu}{2}\right]$ increases in μ and for $\mu < \sqrt{\frac{2\ln k}{n}}$ the function $\left[\frac{\ln k}{\mu} + n\frac{\mu}{2}\right]$ decreases in μ.

As a result, for $n \geq \frac{2\ln k}{\Delta^2}$, we have $\inf_{\mu \geq \Delta}\left[\frac{\ln k}{\mu} + n\frac{\mu}{2}\right] = \left[\frac{\ln k}{\Delta} + n\frac{\Delta}{2}\right]$. For $n < \frac{2\ln k}{\Delta^2}$, we have $\inf_{\mu \geq \Delta}\left[\frac{\ln k}{\mu} + n\frac{\mu}{2}\right]$ occurs at $\mu = \sqrt{\frac{2\ln k}{n}} > \Delta$ which implies $\inf_{\mu \geq \Delta}\left[\frac{\ln k}{\mu} + n\frac{\mu}{2}\right] = \sqrt{2\ln k}\sqrt{n}$. Stopping time (3.16) can then be rewritten as

$$\begin{aligned} N &= \inf\left\{n : \left[1 \leq n < \frac{2\ln k}{\Delta^2}, S_n \geq \sqrt{2\ln k}\sqrt{n}\right] \text{ or } \right. \\ &\qquad \left.\left[\frac{2\ln k}{\Delta^2} \leq n, S_n \geq \frac{\ln k}{\Delta} + n\frac{\Delta}{2}\right]\right\} \\ &= \min\{N_1, N_2\} \end{aligned} \tag{3.17}$$

where $N_1 = \inf\left\{n : 1 \leq n < \frac{2\ln k}{\Delta^2}, S_n \geq \sqrt{2\ln k}\sqrt{n}\right\}$ or ∞ if no such n exists and $N_2 = \inf\left\{n : \frac{2\ln k}{\Delta^2} \leq n, S_n \geq \left[\frac{\ln k}{\Delta} + n\frac{\Delta}{2}\right]\right\}$ or ∞ if no such n exists. This

method effectively splits the stopping boundary into a section that is linear in n and another section that is proportional to $\sqrt{n}$. Notice that N_2 is the stopping time studied in chapter 2 (equation (2.5)) for two simple hypotheses $\mu = \Delta$ and $\mu = 0$.

Now we have

$$\begin{aligned} P_0\left(N < \infty\right) &= P_0\left(1 \le N < \frac{2\ln k}{\Delta^2}\right) + P_0\left(\frac{2\ln k}{\Delta^2} \le N < \infty\right) \\ &= P_0\left(1 \le N_1 < \frac{2\ln k}{\Delta^2}\right) + P_0\left(\frac{2\ln k}{\Delta^2} \le N_2 < \infty\right) \quad (3.18) \end{aligned}$$

For $\Delta < \sqrt{2\ln k}$ use Theorem 2 of [24] to bound the first term of equation (3.18).

$$\begin{aligned} P_0\left(1 \le N_1 < \frac{2\ln k}{\Delta^2}\right) &\le \int_1^{\frac{2\ln k}{\Delta^2}} \frac{\sqrt{\ln k}}{2k\sqrt{\pi}} \frac{1}{t} dt \\ &= \frac{\sqrt{\ln k}}{2k\sqrt{\pi}} \log\left(\frac{2\ln k}{\Delta^2}\right) \quad (3.19) \end{aligned}$$

The second term is approximated in the following manner.

$$\begin{aligned} P_0\left(\frac{2\ln k}{\Delta^2} \le N_2 < \infty\right) &= P_0\left(N < \infty\right) - P_0\left(N \le \frac{2\ln k}{\Delta^2} - 1\right) \\ &\approx P_0\left(\tilde{\tau} < \infty\right) - P_0\left(\tilde{\tau} \le \frac{2\ln k}{\Delta^2} - 1\right) \\ &= \frac{\exp\{-\rho\Delta\}}{k} \Phi\left[\frac{\ln k + 2\rho\Delta}{\sqrt{2\ln k - \Delta^2}} - \frac{\sqrt{2\ln k - \Delta^2}}{2}\right] \\ &\quad -\Phi\left[-\frac{\ln k}{\sqrt{2\ln k - \Delta^2}} - \frac{\sqrt{2\ln k - \Delta^2}}{2}\right] \quad (3.20) \end{aligned}$$

where $\tilde{\tau}$ represents the adjusted Brownian motion process (see Appendix section A.3.3 for details) and $\rho \cong 0.583$. Adding the first term (equation 3.19) and the

second term (equation 3.20) yields equation (3.18). For $\Delta \geq \sqrt{2 \ln k}$ we have $P_0(N < \infty) = P_0(1 \leq N_2 < \infty)$. And $N_2 = \tilde{\tau}$ is the stopping time from section 2.3, so the probability is exactly the Tepee function.

$$P_0(N < \infty) = P_0(1 \leq N_2 < \infty) = \frac{\exp\{-\Delta\rho\}}{k} \tag{3.21}$$

For complete generality replace Δ with Δ/σ where $\Delta = \mu_1 - \mu_0$.

QED •

3.5 Class III: Truncated Designs

The class of Truncated Designs refers to experimental designs where the sample size may increase, but not exceed a pre-determined upper limit of m observations. These designs provide an alternative to the Open designs which are not executable in practice. We examine the stopping rule whose purpose is to generate strong evidence for some simple alternative in $H_C : \mu \geq \mu_1 = \mu_0 + \Delta$ ($\Delta > 0$) over $H_0 : \mu = \mu_0$ where both μ_0 and μ_1 are fixed.

This stopping rule ends the study at the smallest sample size where some alternative $\mu \geq \mu_1$ is better supported over μ_0 by a factor of k or more. The point at which the study stops is now a random variable called a stopping time, already defined in equation (3.12). A Truncated design employing stopping time (3.12) continues sampling until strong evidence for some simple alternative $\mu \geq \mu_1$ over μ_0 is obtained or m observations are collected, whichever occurs first. This design is also a biased design favoring hypotheses in H_C. Henceforth, an Truncated design employing this stopping rule will be called a Biased Composite Truncated Design or C-Biased Truncated design.

The probability that the C-Biased Truncated Design generates strong evidence for some $\mu \geq \mu_1$ over μ_0 can be expressed as the probability that $N \leq m$. For each fixed μ_1, the probability, $P_0(N \leq m)$, will depend on whether $m \geq 2\sigma^2 \ln k/\Delta^2$ or not.

Theorem 3.3 *(C-Biased Truncated Design). For $X_1, X_2, \ldots, X_m$ i.i.d. Normal random variables with mean μ_0 and known variance σ^2, the probability that the C-Biased Truncated design generates strong evidence for some $\mu \geq \mu_1$ over $\mu = \mu_0$ is the following.*

$$P_0\left(\sup_{\mu \geq \mu_1} \frac{L_n(\mu)}{L_n(\mu_0)} \geq k \ ; \text{ for any } n = 1, 2, \ldots, m\right) = P_0(N \leq m) \tag{3.22}$$

1. For $\Delta \geq \sqrt{2\sigma^2 \ln k}$

$$\begin{aligned} P_0(N \leq m) &\cong \Phi\left[-\frac{\sigma \ln k}{\Delta\sqrt{m}} - \frac{\Delta\sqrt{m}}{2\sigma}\right] \\ &\quad + \frac{\exp\{-\rho\Delta/\sigma\}}{k}\Phi\left[-\left(\frac{\sigma \ln k}{\Delta} + 2\rho\right)\frac{1}{\sqrt{m}} + \frac{\Delta\sqrt{m}}{2\sigma}\right] \end{aligned}$$

2. For $\sqrt{2\sigma^2 \ln k/m} \leq \Delta < \sqrt{2\sigma^2 \ln k}$

$$\begin{aligned} P_0(N \leq m) &\leq \frac{\sqrt{\ln k}}{2k\sqrt{\pi}} \log\left(\frac{2\sigma^2 \ln k}{\Delta^2}\right) \\ &\quad + \Phi\left[-\frac{\sigma \ln k}{\Delta}\frac{1}{\sqrt{m}} - \frac{\Delta\sqrt{m}}{2\sigma}\right] - \Phi\left[-\frac{\sigma \ln k}{A(\Delta)} - \frac{A(\Delta)}{2\sigma}\right] \\ &\quad + \frac{\exp\{-\rho\Delta/\sigma\}}{k}\left\{\Phi\left[-\left(\frac{\sigma \ln k}{\Delta} + 2\rho\right)\frac{1}{\sqrt{m}} + \frac{\Delta\sqrt{m}}{2\sigma}\right]\right. \\ &\quad \left. - \Phi\left[-\left(\frac{\sigma \ln k + 2\rho\Delta}{A(\Delta)}\right) + \frac{A(\Delta)}{2\sigma}\right]\right\} \end{aligned}$$

3. For $\Delta < \sqrt{2\sigma^2 \ln k/m}$

$$P_0\left(N \leq m\right) \leq \frac{\sqrt{\ln k}}{2k\sqrt{\pi}} \log\left(m\right)$$

where $\Delta = \mu_1 - \mu_0, A(\Delta) = \sqrt{2\sigma^2 \ln k - \Delta^2}$and $\rho \cong 0.583$.

A proof is inculded at the end of this section. Now, if the difference Δ is measured in units of standard deviations, say $\Delta = c\sigma$, then equation (3.22) becomes

$$P_0\left(\sup_{\mu \geq \mu_1} \frac{L_n(\mu)}{L_n(\mu_0)} \geq k \ \text{ ; for any } n = 1, 2, \ldots, m\right) = P_0\left(N \leq m\right) \tag{3.23}$$

1. For $c \geq \sqrt{2 \ln k}$

$$\begin{aligned} P_0\left(N \leq m\right) &\cong \Phi\left[-\frac{\ln k}{c\sqrt{m}} - \frac{c\sqrt{m}}{2}\right] \\ &\quad + \frac{\exp\left\{-\rho c\right\}}{k} \Phi\left[-\left(\frac{\ln k}{c} + 2\rho\right)\frac{1}{\sqrt{m}} + \frac{c\sqrt{m}}{2}\right] \end{aligned}$$

2. For $\sqrt{2 \ln k/m} \leq c < \sqrt{2 \ln k}$

$$\begin{aligned} P_0\left(N \leq m\right) &\leq \frac{\sqrt{\ln k}}{2k\sqrt{\pi}} \log\left(\frac{2 \ln k}{c^2}\right) \\ &\quad + \Phi\left[-\frac{\ln k}{c}\frac{1}{\sqrt{m}} - \frac{c\sqrt{m}}{2}\right] - \Phi\left[-\frac{\ln k}{A(c)} - \frac{A(c)}{2}\right] \\ &\quad + \frac{\exp\left\{-\rho c\right\}}{k}\left\{\Phi\left[-\left(\frac{\ln k}{c} + 2\rho\right)\frac{1}{\sqrt{m}} + \frac{c\sqrt{m}}{2}\right]\right. \\ &\quad \left. - \Phi\left[-\left(\frac{\ln k + 2\rho c}{A(c)}\right) + \frac{A(c)}{2}\right]\right\} \end{aligned}$$

3. For $c < \sqrt{2 \ln k/m}$

$$P_0\left(N \leq m\right) \leq \frac{\sqrt{\ln k}}{2k\sqrt{\pi}} \log\left(m\right)$$

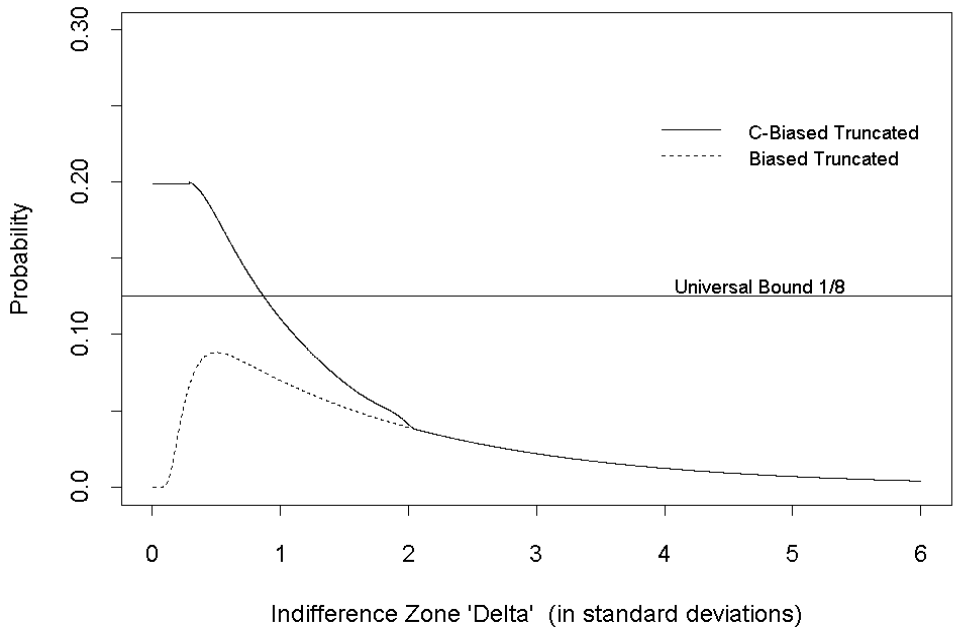

Figure 3.3: The C-Biased Truncated design when $m = 50$

where $A(c) = \sqrt{2 \ln k - c^2}$ and $\rho \cong 0.583$

Figure 3.3 is a graph of the probability that the C-Biased Truncated Design (solid line) generates evidence for some alternative $\mu \geq \mu_1$ over μ_0 by a factor of 8 or more when $m = 50$. The dotted line is the corresponding probability, under μ_0, that the Biased Truncated design generates strong evidence for $\mu = \mu_1$ over $\mu = \mu_0$. Notice for the C-Biased Truncated design, the probability of generating strong evidence levels off for all $\Delta < \sqrt{2 \ln k/m}$ standard deviations.

For $\Delta < \sqrt{2\sigma^2 \ln k/m}$, the probability of collecting observations that support some alternative $\mu \geq \mu_1$ over μ_0 by a factor of k or more cannot exceed $\frac{\sqrt{\ln k}}{2k\sqrt{\pi}} \log(m)$ which is independent of Δ and σ. If we represent this maximum probability by p, then calculate m so that the probability of generating such

evidence equals p, we have

$$m = \exp\left\{\frac{2k\sqrt{\pi}}{\sqrt{\ln k}}p\right\} \tag{3.24}$$

Figure 3.4 plots equation (3.24) for $k = 8, 32$. On the y-axis is the maximum probability (over all Δ) of generating strong evidence for some alternative $\mu > \mu_1 = \mu_0 + \Delta$ over μ_0 under the C-Biased Truncated design. And on the x-axis is the logarithm of the maximum sample size allowed by the design, $\log(m)$. The graph shows the large effect truncating has on the maximum probability (over all Δ) of generating strong evidence for some alternative $\mu > \mu_1 = \mu_0 + \Delta$ over μ_0. Of course as m increases, so does the corresponding probability, which converges to one. However, as we saw in section 3.2, it is important to keep the scale of μ in mind as the probability drops off quickly for increasing Δ. The purpose of Figure 3.4 is to show the degree of dependence of the maximum probability of generating strong evidence on the maximum possible sample size.

Proof: (C-Biased Truncated Design)

Without loss of generality take $\mu_0 = 0$ and $\sigma^2 = 1$. Define the usual stopping variable as:

$$\begin{aligned} N &= \inf\left\{n : n \geq 1, \sup_{\mu \geq \Delta}\left[\frac{L_n(\mu)}{L_n(0)}\right] \geq k\right\} \\ &= \inf\left\{n : n \geq 1, \sup_{\mu \geq \Delta}\left[\mu S_n - n\frac{\mu^2}{2}\right] \geq \ln k\right\} \\ &= \inf\left\{n : n \geq 1, S_n \geq \inf_{\mu \geq \Delta}\left[\frac{\ln k}{\mu} + n\frac{\mu}{2}\right]\right\} \end{aligned}$$

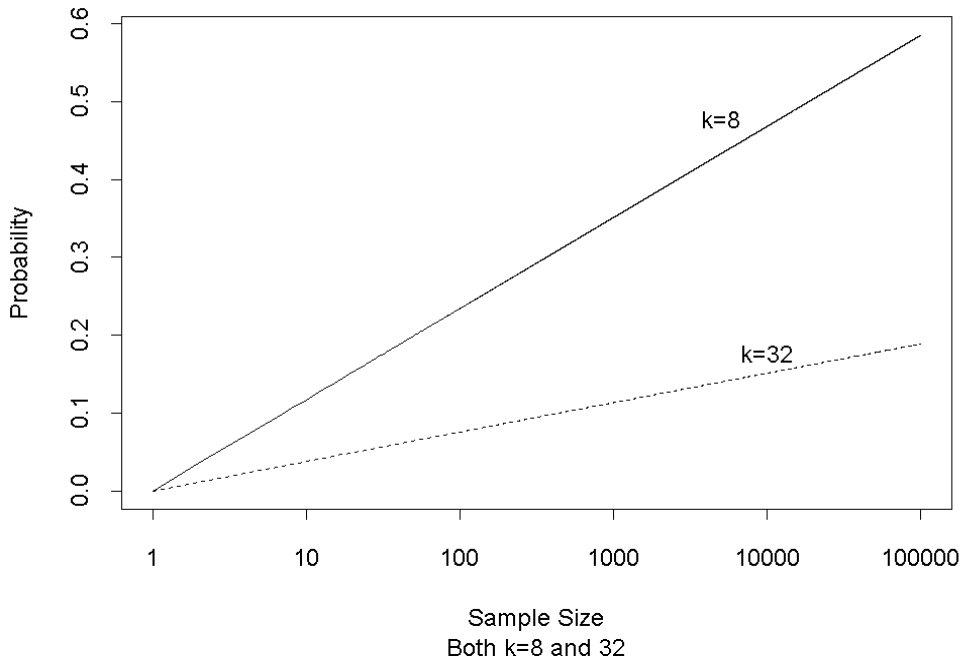

Figure 3.4: Class III: Maximum Probability versus Sample Size

Separate the probability of generating evidence for some $\mu \geq \Delta$ over $\mu = 0$ into:

$$P_0\left(N \leq m\right) = P_0\left(N_1 < \frac{2\ln k}{\Delta^2}\right) + P_0\left(\frac{2\ln k}{\Delta^2} \leq N_2 \leq m\right)$$

$$= \begin{cases} P_0\left(N_2 \leq m\right) & \Delta \geq \sqrt{2\ln k} \\ P_0\left(N_1 < \frac{2\ln k}{\Delta^2}\right) + P_0\left(\frac{2\ln k}{\Delta^2} \leq N_2 \leq m\right) & \sqrt{2\ln k} > \Delta \geq \sqrt{2\ln k/m} \\ P_0\left(N_1 < m\right) & \Delta < \sqrt{2\ln k/m} \end{cases}$$

where $N_1 = \inf\left\{n : 1 \leq n < \frac{2\ln k}{\Delta^2}, S_n \geq \sqrt{2\ln k}\sqrt{n}\right\}$ or ∞ if no such n exists and $N_2 = \inf\left\{n : \frac{2\ln k}{\Delta^2} \leq n, S_n \geq \left[\frac{\ln k}{\Delta} + n\frac{\Delta}{2}\right]\right\}$ or ∞ if no such n exists.

For $\Delta < \sqrt{2\ln k/m}$ use Theorem two of [24], yielding

$$\begin{aligned} P_0\left(N_1 < m\right) &\leq \int_1^m \frac{\sqrt{\ln k}}{2k\sqrt{\pi}}\frac{1}{t}dt \\ &= \frac{\sqrt{\ln k}}{2k\sqrt{\pi}}\log\left(m\right) \end{aligned}$$

Notice, in this case, $P_0(N \leq m)$ is independent of Δ indicating a constant probability for all $\Delta < \sqrt{2\ln k/m}$.

For the first piece, $\Delta \geq \sqrt{2\ln k}$, we use the standard adjusted Brownian motion approximation because $N_2 = \tilde{\tau}$ is the stopping time from Chapter 2, equation (2.5). Now we have

$$\begin{aligned} P_0(N_2 \leq m) &= P_0\{\tilde{\tau} \leq m\} \\ &\cong \Phi\left[-\frac{\ln k}{\Delta}m^{-\frac{1}{2}} - \frac{\Delta}{2}m^{\frac{1}{2}}\right] \\ &\quad + \frac{\exp\{-\rho\Delta\}}{k}\Phi\left[-\left(\frac{\ln k}{\Delta} + 2\rho\right)m^{-\frac{1}{2}} + \frac{\Delta}{2}m^{\frac{1}{2}}\right] \end{aligned}$$

For $\sqrt{2\ln k} > \Delta \geq \sqrt{2\ln k/m}$ the first half of the probability can be bounded using Theorem two of [24]

$$P_0\left(N_1 < \frac{2\ln k}{\Delta^2}\right) \leq \frac{\sqrt{\ln k}}{2k\sqrt{\pi}}\log\left(\frac{2\ln k}{\Delta^2}\right)$$

The second half of the quantity can be approximated using the adjusted Brownian motion probabilities. Let $A = \frac{2\ln k}{\Delta^2} - 1 = \frac{2\ln k - \Delta^2}{\Delta^2}$, so that

$$\begin{aligned} P_0\left(\frac{2\ln k}{\Delta^2} \leq N_2 \leq m\right) &= P_0(N_2 \leq m) - P_0\left(N_2 \leq \frac{2\ln k}{\Delta^2} - 1\right) \\ &= P_0(\tilde{\tau} \leq m) - P_0(\tilde{\tau} \leq A) \end{aligned}$$

From the appendix on Brownian motion section A.3.3 we have

$$\begin{aligned} P_0\left(\frac{2\ln k}{\Delta^2} \leq N_2 \leq m\right) &= \Phi\left[-\frac{\ln k}{\Delta}m^{-\frac{1}{2}} - \frac{\Delta}{2}m^{\frac{1}{2}}\right] \\ &\quad + \frac{\exp\{-\rho\Delta\}}{k}\Phi\left[-\left(\frac{\ln k}{\Delta} + 2\rho\right)m^{-\frac{1}{2}} + \frac{\Delta}{2}m^{\frac{1}{2}}\right] \end{aligned}$$

$$
\begin{aligned}
& -\Phi\left[-\frac{\ln k}{\Delta}A^{-\frac{1}{2}} - \frac{\Delta}{2}A^{\frac{1}{2}}\right] \\
& -\frac{\exp\{-\rho\Delta\}}{k}\Phi\left[-\left(\frac{\ln k}{\Delta}+2\rho\right)A^{-\frac{1}{2}} + \frac{\Delta}{2}A^{\frac{1}{2}}\right] \\
= \;& \Phi\left[-\frac{\ln k}{\Delta}m^{-\frac{1}{2}} - \frac{\Delta}{2}m^{\frac{1}{2}}\right] - \Phi\left[-\frac{\ln k}{\Delta}A^{-\frac{1}{2}} - \frac{\Delta}{2}A^{\frac{1}{2}}\right] \\
& +\frac{\exp\{-\rho\Delta\}}{k}\left\{\Phi\left[-\left(\frac{\ln k}{\Delta}+2\rho\right)m^{-\frac{1}{2}} + \frac{\Delta}{2}m^{\frac{1}{2}}\right]\right. \\
& \left.-\Phi\left[-\left(\frac{\ln k}{\Delta}+2\rho\right)A^{-\frac{1}{2}} + \frac{\Delta}{2}A^{\frac{1}{2}}\right]\right\}
\end{aligned}
$$

But $\sqrt{A}\Delta = \sqrt{\frac{2\ln k - \Delta^2}{\Delta^2}}\Delta = \sqrt{2\ln k - \Delta^2} = A(\Delta)$. So that:

$$
\begin{aligned}
P_0\left(\frac{2\ln k}{\Delta^2} \le N_2 \le m\right) \;=\; & \Phi\left[-\frac{\ln k}{\Delta}m^{-\frac{1}{2}} - \frac{\Delta}{2}m^{\frac{1}{2}}\right] - \Phi\left[-\frac{\ln k}{A(\Delta)} - \frac{A(\Delta)}{2}\right] \\
& +\frac{\exp\{-\rho\Delta\}}{k}\left\{\Phi\left[-\left(\frac{\ln k}{\Delta}+2\rho\right)m^{-\frac{1}{2}} + \frac{\Delta}{2}m^{\frac{1}{2}}\right]\right. \\
& \left.-\Phi\left[-\left(\frac{\ln k + 2\rho\Delta}{A(\Delta)}\right) + \frac{A(\Delta)}{2}\right]\right\}
\end{aligned}
$$

Combining the above equations and replacing Δ with Δ/σ where $\Delta = \mu_1 - \mu_0$ yields Theorem 3.3.

QED •

3.6 Class IV: Delayed Designs

The class of Delayed designs refers to experimental designs where the sample size must accumulate at least a pre-determined minimum number of observations, say m_0. A Delayed design first collects m_0 observations and then determines whether to continue sampling based on its stopping rule. Like the class of

C-Biased Open designs, these designs have no limit on the sample size. We examine the stopping rule whose purpose is to generate strong evidence for some alternative in $H_C : \mu \geq \mu_1$ over $H_0 : \mu = \mu_0$ ($\mu_1 = \mu_0 + \Delta$) where both μ_0 and μ_1 are fixed.

This stopping rule ends the study at the smallest sample size when some simple alternative $\mu \geq \mu_1$ is better supported over μ_0 by a factor of k or more. The point at which the study stops is now a random variable called a stopping time, already defined in equation (3.12). A Delayed design employing stopping time (3.12) continues sampling after m_0 observations have been collected until strong evidence for some simple alternative $\mu \geq \mu_1$ is better supported over μ_0. This design is also a biased design favoring hypotheses in H_C. Henceforth, a Delayed design employing this stopping rule will be called a Biased Composite Delayed Design or C-Biased Delayed design.

The probability that the C-Biased Delayed design generates strong evidence for some alternative $\mu \geq \mu_1$ over μ_0 can be expressed as the probability that $m_0 \leq N < \infty$. For each fixed μ_1, the probability $P_0(m_0 \leq N < \infty)$, depends on if $\Delta^2 < 2\sigma^2 \ln k/m_0$ and is given by Theorem (3.4).

Theorem 3.4 *(C-Biased Delayed Design). For observations $X_1, X_2, \ldots$ i.i.d. Normal random variables with mean μ_0 and known variance σ^2. The probability that the C-Biased Delayed design generates strong evidence for some simple alternative $\mu \geq \mu_1$ over $\mu = \mu_0$ is given by the following equation (3.25).*

$$P_0\left(\sup_{\mu \geq \mu_1} \frac{L_n(\mu)}{L_n(\mu_0)} \geq k \text{ for any } n = m_0, m_0+1, \ldots\right) = P_0\left(m_0 \leq N < \infty\right) \tag{3.25}$$

1. For $\Delta < \sqrt{2\sigma^2 \ln k/m_0}$

$$P_0(m_0 \leq N < \infty) \leq \frac{\sqrt{\ln k}}{2k\sqrt{\pi}} \log\left(\frac{2\sigma^2 \ln k}{m_0 \Delta^2}\right) - \Phi\left[-\frac{\sigma \ln k}{A(\Delta)} - \frac{A(\Delta)}{2\sigma}\right] + \frac{\exp\{-\rho\Delta/\sigma\}}{k} \Phi\left[\left(\frac{\sigma \ln k + 2\rho\Delta}{A(\Delta)}\right) - \frac{A(\Delta)}{2\sigma}\right]$$

2. For $\Delta \geq \sqrt{2\sigma^2 \ln k/m_0}$

$$P_0(m_0 \leq N < \infty) \leq -\Phi\left[-\frac{\sigma \ln k}{\Delta\sqrt{(m_0-1)}} - \frac{\Delta\sqrt{(m_0-1)}}{2\sigma}\right] + \frac{\exp\{-\rho\Delta/\sigma\}}{k} \Phi\left[\left(\frac{\sigma \ln k + 2\rho\Delta}{\Delta\sqrt{(m_0-1)}}\right) - \frac{\Delta\sqrt{(m_0-1)}}{2\sigma}\right]$$

where $\Delta = \mu_1 - \mu_0, A(\Delta) = \sqrt{2\sigma^2 \ln k - \Delta^2}$ and $\rho \cong 0.583$.

The proof is included at the end of this section. If the difference Δ is measured in units of standard deviations, say $\Delta = c\sigma$, then equation (3.25) becomes

$$P_0\left(\sup_{\mu \geq \mu_1} \frac{L_n(\mu)}{L_n(\mu_0)} \geq k \ ; \text{for any } n = m_0, m_0+1, \ldots\right) = P_0(m_0 \leq N < \infty) \tag{3.26}$$

1. For $c < \sqrt{2 \ln k/m_0}$

$$P_0(m_0 \leq N < \infty) \leq \frac{\sqrt{\ln k}}{2k\sqrt{\pi}} \log\left(\frac{2 \ln k}{m_0 c^2}\right) - \Phi\left[-\frac{\ln k}{A(c)} - \frac{A(c)}{2}\right] + \frac{\exp\{-\rho c\}}{k} \Phi\left[\left(\frac{\ln k + 2\rho c}{A(c)}\right) - \frac{A(c)}{2}\right]$$

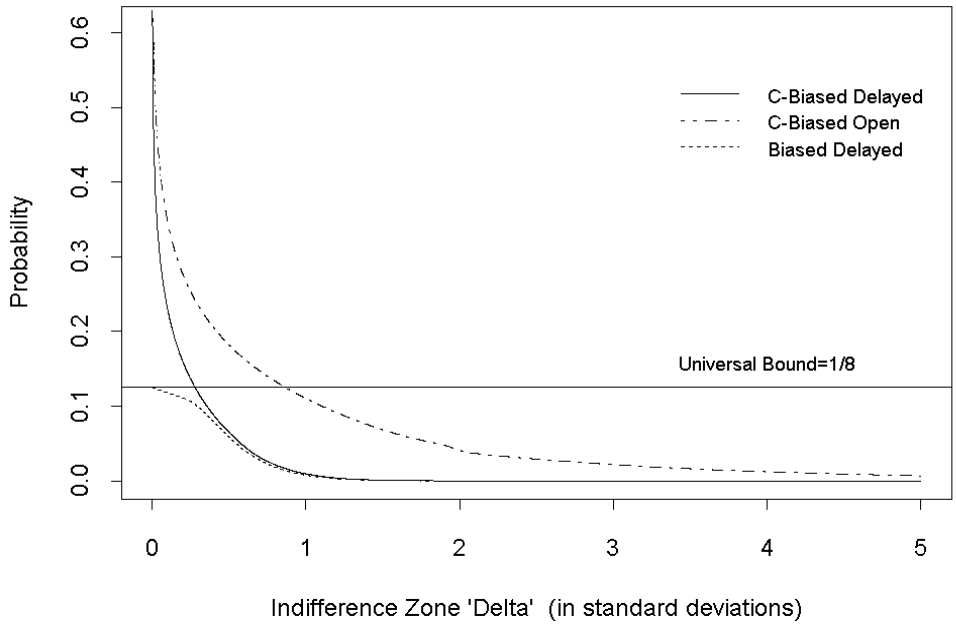

Figure 3.5: The C-Biased Delayed design when $m_0 = 10$

2. For $c \geq \sqrt{2 \ln k / m_0}$

$$P_0(m_0 \leq N < \infty) \leq -\Phi\left[-\frac{\ln k}{c\sqrt{(m_0-1)}} - \frac{c\sqrt{(m_0-1)}}{2}\right] + \frac{\exp\{-\rho c\}}{k}\Phi\left[\left(\frac{\ln k + 2\rho c}{c\sqrt{(m_0-1)}}\right) - \frac{c\sqrt{(m_0-1)}}{2}\right]$$

where $A(c) = \sqrt{2 \ln k - c^2}$ and $\rho \cong 0.583$.

Figure 3.5 is a graph of the probability of generating strong evidence for some alternative $\mu \geq \mu_1$ over μ_0 by a factor of 8 or more, under both the C-Biased Delayed design with $m_0 = 10$ (solid line) and under the C-Biased Open design (dashed and dotted). The Biased Delayed design with at least 10 observations (dotted line) is plotted as well, but represents the probability, when $\mu = \mu_0$, of

generating strong evidence for μ_1 over μ_0. Note that the Biased Delayed design provides a lower bound on the probability of generating strong evidence for some alternative in H_C over H_0. The C-Biased Open design is plotted to emphasize the reduction in the probability caused by delaying for 10 observations, which is large for moderate to larger values of Δ. Figure 3.5 demonstrates that a large majority of the probability, for moderate and larger values of Δ, occurs early (at least before the tenth observation). This suggests that the chance of generating evidence supporting some simple alternative $\mu \geq \mu_1$ over μ_0 can be reduced to almost zero in practice, by using a Delayed design.

Proof: (C-Biased Delayed Design)

Without loss of generality let $\mu_0 = 0$ and $\sigma^2 = 1$. The stopping variable N remains the same, see equation (3.12). For each fixed Δ the probability $P_0\left(m_0 \leq N < \infty\right)$ depends on whether $m_0 \geq 2\ln k/\Delta^2$. Now we have

$$
\begin{aligned}
& P_0\left(m_0 \leq N < \infty\right) \\
& = P_0\left(m_0 \leq N_1 < \frac{2\ln k}{\Delta^2}\right) + P_0\left(\frac{2\ln k}{\Delta^2} \leq N_2 < \infty\right) \\
& = \begin{cases} P_0\left(m_0 \leq N_2 < \infty\right) & \Delta \geq \sqrt{2\ln k/m_0} \\ P_0\left(m_0 \leq N_1 < \frac{2\ln k}{\Delta^2}\right) + P_0\left(\frac{2\ln k}{\Delta^2} \leq N_2 < \infty\right) & \Delta < \sqrt{2\ln k/m_0} \end{cases}
\end{aligned}
$$

For $\Delta < \sqrt{2\ln k/m_0}$ use Theorem two of Lorden [24] to bound the first term. Thus we have

$$
\begin{aligned}
P_0\left(m_0 \leq N_1 < \frac{2\ln k}{\Delta^2}\right) &\leq \int_{m_0}^{\frac{2\ln k}{\Delta^2}} \frac{\sqrt{\ln k}}{2k\sqrt{\pi}} \frac{1}{t} dt \\
&= \frac{\sqrt{\ln k}}{2k\sqrt{\pi}} \log\left(\frac{2\ln k}{\Delta^2 m_0}\right)
\end{aligned}
$$

The second term for $\Delta < \sqrt{2\ln k/m_0}$ is calculated using the adjusted Brownian motion results from the appendix, section A.3.3. Let $A = \frac{2\ln k}{\Delta^2} - 1 = \frac{2\ln k - \Delta^2}{\Delta^2}$, and we have:

$$\begin{aligned}
P_0\left(\frac{2\ln k}{\Delta^2} \le N_2 < \infty\right) &= P_0\left(N_2 < \infty\right) - P_0\left(N_2 \le \frac{2\ln k}{\Delta^2} - 1\right) \\
&= P_0\left(\tilde{\tau} < \infty\right) - P_0\left(\tilde{\tau} \le A\right) \\
&\cong \frac{\exp\{-\rho\Delta\}}{k} - \Phi\left[-\frac{\ln k}{\Delta}A^{-\frac{1}{2}} - \frac{\Delta}{2}A^{\frac{1}{2}}\right] \\
&\quad -\frac{\exp\{-\rho\Delta\}}{k}\Phi\left[-\left(\frac{\ln k}{\Delta} + 2\rho\right)A^{-\frac{1}{2}} + \frac{\Delta}{2}A^{\frac{1}{2}}\right] \\
&= \frac{\exp\{-\rho\Delta\}}{k}\Phi\left[\left(\frac{\ln k + 2\rho\Delta}{A(\Delta)}\right) - \frac{A(\Delta)}{2}\right] \\
&\quad -\Phi\left[-\frac{\ln k}{A(\Delta)} - \frac{A(\Delta)}{2}\right]
\end{aligned}$$

where $\sqrt{A}\Delta = \sqrt{2\ln k - \Delta^2} = A(\Delta)$

For $\Delta \ge \sqrt{2\ln km_0}$, the adjusted Brownian motion results from the appendix, section A.3.3 are again used.

$$\begin{aligned}
P_0\left(m_0 \le N_2 < \infty\right) &= P_0\left\{\tilde{\tau} < \infty\right\} - P_0\left\{\tilde{\tau} \le m_0 - 1\right\} \\
&\cong \frac{\exp\{-\rho\Delta\}}{k} - \Phi\left[-\frac{\ln k}{\Delta}\left(m_0 - 1\right)^{-\frac{1}{2}} - \frac{\Delta}{2}\left(m_0 - 1\right)^{\frac{1}{2}}\right] \\
&\quad -\frac{\exp\{-\rho\Delta\}}{k}\Phi\left[-\left(\frac{\ln k}{\Delta} + 2\rho\right)\left(m_0 - 1\right)m^{-\frac{1}{2}} + \frac{\Delta}{2}\left(m_0 - 1\right)^{\frac{1}{2}}\right] \\
&= \frac{\exp\{-\rho\Delta\}}{k}\Phi\left[\left(\frac{\ln k}{\Delta} + 2\rho\right)\left(m_0 - 1\right)^{-\frac{1}{2}} - \frac{\Delta}{2}\left(m_0 - 1\right)^{\frac{1}{2}}\right] \\
&\quad -\Phi\left[-\frac{\ln k}{\Delta}\left(m_0 - 1\right)^{-\frac{1}{2}} - \frac{\Delta}{2}\left(m_0 - 1\right)^{\frac{1}{2}}\right]
\end{aligned}$$

Combining the above equations and replacing Δ with Δ/σ where $\Delta = \mu_1 - \mu_0$ yields Theorem 3.4.

QED •

3.7 Class V: Interval Designs

The class of Interval designs refers to experimental designs where the sample size must accumulate at least m_0 observations, but may not exceed m observations. An Interval design first collects m_0 observations and then determines whether to continue sampling based on its stopping rule, until m observations are collected, at which point the study terminates regardless of the statistical evidence. An Interval design is effectively a combination of a Delayed and Truncated Design, guaranteeing at least m_0 observations, but no more than m observations.

Within the class of Interval designs, we study the stopping rule whose purpose is to generate strong evidence for some alternative $\mu \geq \mu_1$ over μ_0 by a factor of k or more, where both μ_0 and $\mu_1 = \mu_0 + \Delta$ ($\Delta > 0$) are fixed. This stopping rule ends the study at the smallest sample size when strong evidence for any single $\mu \geq \mu_1$ over μ_0 is obtained. The point at which the study stops is now a random variable called a stopping time, already defined in eqaution (3.12). An Interval design employing stopping time (3.12) continues sampling after m_0 observations have been collected until strong evidence for some simple alternative $\mu \geq \mu_1$ over μ_0 is obtained or m observations are collected, whichever come first. This design is also a biased design favoring alternative in $H_C : \mu \geq \mu_0$. Henceforth, an Interval design employing this stopping rule will be called a Biased Composite Interval Design or C-Biased Interval design.

The probability that the C-Biased Interval Design generates evidence supporting some alternative $\mu \geq \mu_1$ over μ_0 can be expressed as the probability that $m_0 \leq N \leq m$.

Theorem 3.5 *(C-Biased Interval Design) For observations $X_1, X_2, \ldots$ i.i.d. Normal random variables with mean μ_0 and known variance σ^2, the probability that the C-Biased Interval design generates evidence for some simple alternative $\mu \geq \mu_1$ over $\mu = 0$ is given as follows.*

$$P_0\left(\sup_{\mu \geq \mu_1} \frac{L_n(\mu)}{L_n(\mu_0)} \geq k \text{ ; for any } n = m_0, m_0+1, \ldots, m\right) = P_0(m_0 \leq N \leq m) \tag{3.27}$$

1. For $\Delta < \sqrt{2\sigma^2 \ln k/m}$

$$P_0(m_0 \leq N \leq m) \leq \frac{\sqrt{\ln k}}{2k\sqrt{\pi}} \log\left(\frac{m}{m_0}\right)$$

2. For $\sqrt{2\sigma^2 \ln k/m} \leq \Delta < \sqrt{2\sigma^2 \ln k/m_0}$

$$\begin{aligned} P_0(m_0 \leq N \leq m) \quad \leq \quad & \frac{\sqrt{\ln k}}{2k\sqrt{\pi}} \log\left(\frac{2\sigma^2 \ln k}{m_0 \Delta^2}\right) \\ & + \Phi\left[-\frac{\sigma \ln k}{\Delta\sqrt{m}} - \frac{\Delta\sqrt{m}}{2\sigma}\right] - \Phi\left[-\frac{\sigma \ln k}{A(\Delta)} - \frac{A(\Delta)}{2\sigma}\right] \\ & + \frac{\exp\{-\rho\Delta/\sigma\}}{k}\left\{\Phi\left[-\left(\frac{\sigma \ln k + 2\rho\Delta}{\Delta\sqrt{m}}\right) + \frac{\Delta\sqrt{m}}{2\sigma}\right]\right. \\ & \left. - \Phi\left[-\left(\frac{\sigma \ln k + 2\rho\Delta}{A(\Delta)}\right) - \frac{A(\Delta)}{2\sigma}\right]\right\} \end{aligned}$$

3. For $\Delta \geq \sqrt{2\sigma^2 \ln k/m_0}$

$$\begin{aligned} P_0\left(m_0 \leq N \leq m\right) &\cong \Phi\left[-\frac{\sigma \ln k}{\Delta\sqrt{m}} - \frac{\Delta\sqrt{m}}{2\sigma}\right] \\ &- \Phi\left[-\frac{\sigma \ln k}{\Delta\sqrt{(m_0-1)}} - \frac{\Delta\sqrt{(m_0-1)}}{2\sigma}\right] \\ &+ \frac{\exp\{-\rho\Delta/\sigma\}}{k}\left\{\Phi\left[-\left(\frac{\sigma \ln k + 2\rho\Delta}{\Delta\sqrt{m}}\right) + \frac{\Delta\sqrt{m}}{2\sigma}\right]\right. \\ &\left. -\Phi\left[-\left(\frac{\sigma \ln k + 2\rho\Delta}{\Delta\sqrt{(m_0-1)}}\right) - \frac{\Delta\sqrt{(m_0-1)}}{2\sigma}\right]\right\} \end{aligned}$$

where $\Delta = \mu_1 - \mu_0, A(\Delta) = \sqrt{2\sigma^2 \ln k - \Delta^2}$ and $\rho \cong 0.583$.

The proof is deferred until the end of this section. Now, if the difference Δ is measured in units of standard deviations, say $\Delta = c\sigma$, then equation (3.27) becomes

$$P_0\left(\sup_{\mu \geq \mu_1} \frac{L_n(\mu)}{L_n(\mu_0)} \geq k \text{ ; for any } n = m_0, m_0+1, \ldots, m\right) = P_0\left(m_0 \leq N \leq m\right) \tag{3.28}$$

1. For $c < \sqrt{2 \ln k/m}$

$$P_0\left(m_0 \leq N \leq m\right) \leq \frac{\sqrt{\ln k}}{2k\sqrt{\pi}} \log\left(\frac{m}{m_0}\right)$$

2. For $\sqrt{2 \ln k/m} \leq c < \sqrt{2 \ln k/m_0}$

$$\begin{aligned} P_0\left(m_0 \leq N \leq m\right) &\leq \frac{\sqrt{\ln k}}{2k\sqrt{\pi}} \log\left(\frac{2 \ln k}{m_0 c^2}\right) \\ &+ \Phi\left[-\frac{\ln k}{c\sqrt{m}} - \frac{c\sqrt{m}}{2}\right] - \Phi\left[-\frac{\ln k}{A(c)} - \frac{A(c)}{2}\right] \end{aligned}$$

$$+\frac{\exp\{-\rho c\}}{k}\left\{\Phi\left[-\left(\frac{\ln k+2\rho c}{c\sqrt{m}}\right)+\frac{c\sqrt{m}}{2}\right]\right.$$
$$\left.-\Phi\left[-\left(\frac{\ln k+2\rho c}{A(c)}\right)-\frac{A(c)}{2}\right]\right\}$$

3. For $c \geq \sqrt{2\ln k/m_0}$

$$\begin{aligned}P_0\left(m_0 \leq N \leq m\right) &\cong \Phi\left[-\frac{\ln k}{c\sqrt{m}}-\frac{c\sqrt{m}}{2}\right]\\ &- \Phi\left[-\frac{\ln k}{c\sqrt{(m_0-1)}}-\frac{c\sqrt{(m_0-1)}}{2}\right]\\ &+ \frac{\exp\{-\rho c\}}{k}\left\{\Phi\left[-\left(\frac{\ln k+2\rho c}{c\sqrt{m}}\right)+\frac{c\sqrt{m}}{2}\right]\right.\\ &\left.-\Phi\left[-\left(\frac{\ln k+2\rho c}{c\sqrt{(m_0-1)}}\right)-\frac{c\sqrt{(m_0-1)}}{2}\right]\right\}\end{aligned}$$

where $A(c)=\sqrt{2\ln k-c^2}$ and $\rho \cong 0.583$.

Figure 3.6 graphs the probability that the C-Biased Interval design generates strong evidence for some simple alternative $\mu \geq \mu_1$ over μ_0 when $m_0 = 5$, $m = 1,000,000$, and $k = 8$ (solid line). The corresponding probabilities for the C-Biased Open design (dotted/dashed line) and the Biased Interval design (dotted line) are drawn for comparison. The Biased Interval design is a lower bound on the probability when using the C-Biased Interval design, because it represents the probability of generating strong evidence for $\mu = \mu_1$ over μ_0, under μ_0. Notice that, for $\Delta < 1$ standard deviations, the slope of the C-Biased Interval design is steeper than that of the C-Biased Open design.

Both delaying and truncating combine to drastically lower the probability of generating strong evidence for some alternative $\mu \geq \mu_1$ over μ_0 at moderate to

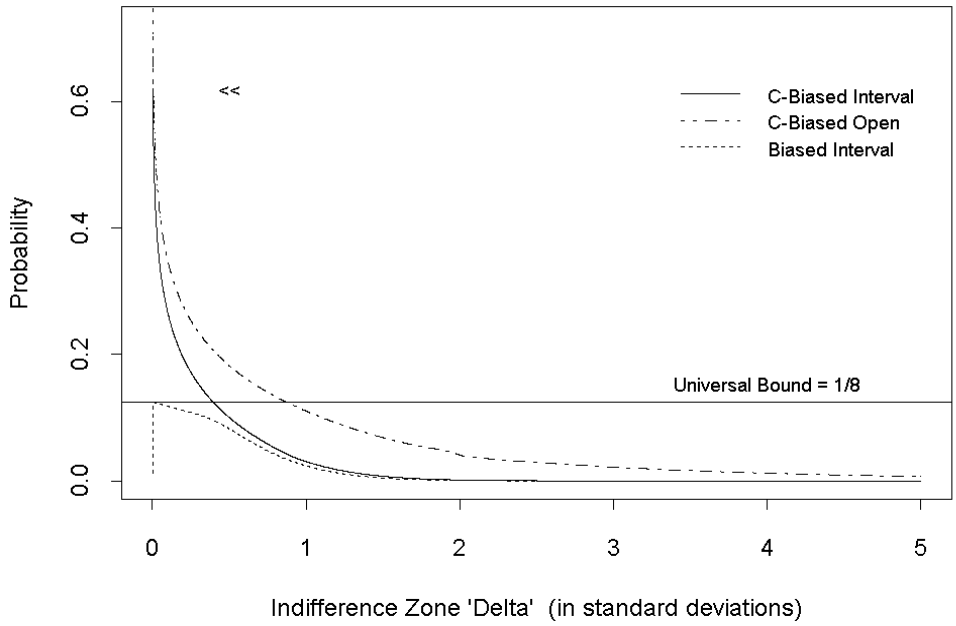

Figure 3.6: The C-Biased Interval design, when $m_0 = 5, m = 1,000,000$

large values of Δ. It is the delaying which plays the major role in lowering the probability. Truncating holds the maximum probability, below unity. The two arrows (<<) denote the maximum probability (about 0.62) under the C-Biased Interval design of collecting observations that support some simple alternative in H_C over H_0 by a factor of 8 or more.

The maximum probability of generating strong evidence for the C-Biased Interval design occurs at $\Delta = 0$. This probability cannot exceed $[\sqrt{\ln k}/(2k\sqrt{\pi})] \ln(m/m_0)$. If we represent the maximum probability by p, we can then calculate the sample size ratio $R = m/m_0$ which will achieve this probability at $\Delta = 0$. Figure 3.7 is a graph of the maximum probability that the C-Biased Interval design generates strong evidence for some alternative $\mu \geq 0$ over μ_0 versus the log sample size

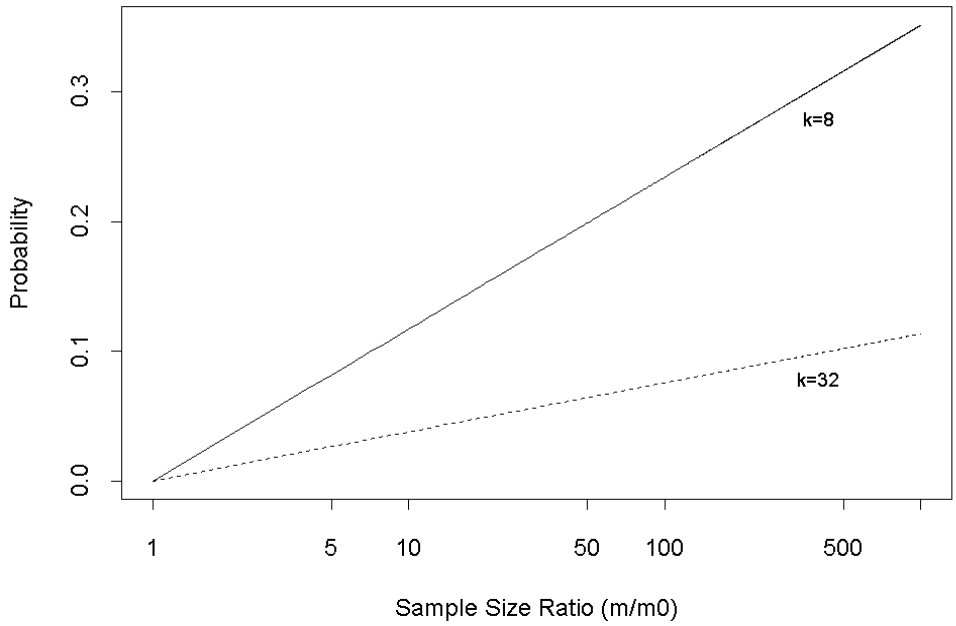

Figure 3.7: Class V: Maximum Probability versus Sample Size Ratio

ratio, for both $k = 8, 32$.

Proof: (C-Biased Interval Design)

Without loss of generality let $\mu_0 = 0$ and the variance $\sigma^2 = 1$. Then for each fixed Δ the probability, $P_0\left(m_0 \leq N \leq m\right)$, will depend on if $\sqrt{2\ln k/m} < \Delta < \sqrt{2\ln k/m_0}$. Again, let $N_1 = \inf\left\{n : 1 \leq n < \frac{2\ln k}{\Delta^2}, S_n \geq \sqrt{2\ln k}\sqrt{n}\right\}$ or ∞ if no such n exists. And $N_2 = \inf\left\{n : \frac{2\ln k}{\Delta^2} \leq n, S_n \geq \left[\frac{\ln k}{\Delta} + n\frac{\Delta}{2}\right]\right\}$ or ∞ if no such n exists.

Now we have

$$P_0\left(m_0 \leq N \leq m\right) = P_0\left(m_0 \leq N_1 < \frac{2\ln k}{\Delta^2}\right) + P_0\left(\frac{2\ln k}{\Delta^2} \leq N_2 \leq m\right)$$

$$= \begin{cases} P_0\left(m_0 \le N_1 < m\right) & \Delta < \sqrt{2\ln k/m} \\ P_0\left(m_0 \le N_1 < \frac{2\ln k}{\Delta^2}\right) + P_0\left(\frac{2\ln k}{\Delta^2} \le N_2 \le m\right) & \sqrt{2\ln k/m} \le \Delta < \sqrt{2\ln k/m_0} \\ P_0\left(m_0 \le N_2 \le m\right) & \Delta \ge \sqrt{2\ln k/m_0} \end{cases}$$

For $\Delta < \sqrt{2\ln k/m}$ use Theorem two of Lorden [24]. Then the probability is bounded by

$$\begin{aligned} P_0\left(m_0 \le N_1 < m\right) &\le \int_{m_0}^{m} \frac{\sqrt{\ln k}}{2k\sqrt{\pi}} \frac{1}{t} dt \\ &= \frac{\sqrt{\ln k}}{2k\sqrt{\pi}} \log\left(\frac{m}{m_0}\right) \end{aligned}$$

For $\Delta \ge \sqrt{\frac{2\ln k}{m_0}}$, use the adjusted Brownian motion approximations from the appendix, section A.3.3.

$$\begin{aligned} P_0\left(m_0 \le N_2 \le m\right) &= P_0\left\{\tilde{\tau} \le m\right\} - P_0\left\{\tilde{\tau} \le m_0 - 1\right\} \\ &\cong \Phi\left[-\frac{\ln k}{\Delta} m^{-\frac{1}{2}} - \frac{\Delta}{2} m^{\frac{1}{2}}\right] - \Phi\left[-\frac{\ln k}{\Delta}\left(m_0 - 1\right)^{-\frac{1}{2}} - \frac{\Delta}{2}\left(m_0 - 1\right)^{\frac{1}{2}}\right] \\ &\quad + \frac{\exp\left\{-\rho\Delta\right\}}{k}\left\{\Phi\left[-\left(\frac{\ln k}{\Delta} + 2\rho\right) m^{-\frac{1}{2}} + \frac{\Delta}{2} m^{\frac{1}{2}}\right]\right. \\ &\quad \left. - \Phi\left[-\left(\frac{\ln k}{\Delta} + 2\rho\right)\left(m_0 - 1\right)^{-\frac{1}{2}} - \frac{\Delta}{2}\left(m_0 - 1\right)^{\frac{1}{2}}\right]\right\} \end{aligned}$$

For $\sqrt{2\ln k/m} \le \Delta < \sqrt{2\ln k/m_0}$ the first term can be bounded using Theorem 2 of Lorden [24]

$$\begin{aligned} P_0\left(m_0 \le N_1 < \frac{2\ln k}{\Delta^2}\right) &\le \int_{m_0}^{\frac{2\ln k}{\Delta^2}} \frac{\sqrt{\ln k}}{2k\sqrt{\pi}} \frac{1}{t} dt \\ &= \frac{\sqrt{\ln k}}{2k\sqrt{\pi}} \log\left(\frac{2\ln k}{\Delta^2 m_0}\right) \end{aligned}$$

For $\sqrt{2\ln k/m} \leq \Delta < \sqrt{2\ln k/m_0}$ the second term can be approximated using adjusted Brownian motion results from the appendix. Let $A = 2\ln k/\Delta^2 - 1 = (2\ln k - \Delta^2)/\Delta^2$ and $\sqrt{A}\Delta = \sqrt{2\ln k - \Delta^2} = A(\Delta)$ then we have:

$$\begin{aligned} P_0\left(\frac{2\ln k}{\Delta^2} \leq N_2 \leq m\right) &= P_0\left(\tilde{\tau} \leq m\right) - P_0\left(\tilde{\tau} \leq A\right) \\ &\cong \Phi\left[-\frac{\ln k}{\Delta}m^{-\frac{1}{2}} - \frac{\Delta}{2}m^{\frac{1}{2}}\right] - \Phi\left[-\frac{\ln k}{A(\Delta)} - \frac{A(\Delta)}{2}\right] \\ &\quad + \frac{\exp\{-\rho\Delta\}}{k}\left\{\Phi\left[-\left(\frac{\ln k}{\Delta} + 2\rho\right)m^{-\frac{1}{2}} + \frac{\Delta}{2}m^{\frac{1}{2}}\right]\right. \\ &\quad \left. -\Phi\left[-\left(\frac{\ln k + 2\rho\Delta}{A(\Delta)}\right) - \frac{A(\Delta)}{2}\right]\right\} \end{aligned}$$

Combining the above equations and replacing Δ with Δ/σ where $\Delta = \mu_1 - \mu_0$ yields Theorem 3.5.

QED •

3.8 Interval Design Examples

The purpose of this section is to provide a few numerical examples of the C-Biased Interval design. The C-Biased Interval design is the most practical design and provides an excellent way to emphasize the effect of the minimum sample size m_0 and the maximum sample size m on the probability of generating strong evidence for some alternative $\mu \geq \mu_1$ over μ_0. We fix $k = 8$ throughout these examples and use equation (3.28), with $\Delta = \mu_1 - \mu_0 = c\sigma$.

Figure 3.8 presents the probability, under μ_0, of generating evidence for some simple alternative $\mu \geq \mu_1$ over μ_0 under the C-Biased Interval design with minimum sample size of $m_0 = 5$ observations and varying maximum sample size

m. We see that the effect of truncating is to reduce the probability for small Δ, but for $\Delta > 0.5$ the probability is basically unaffected by the truncation point. This suggests that for values of $\Delta > 0.5$ the probability accumulates early, at least before the 100^{th} observation. After about 100 observations, for $\Delta > 0.5$ the probability of generating strong evidence for some alternative in H_C over H_0 increases very little.

In contrast, figure 3.9 shows that delaying provides a large reduction in the probability that the C-Biased design generates strong evidence for some alternative $\mu \geq \mu_1$ over μ_0 at all values of Δ. Figure 3.9 presents the probability under the C-Biased Interval design with a maximum sample size of $m = 1,000$ observations and varying minimum sample size m_0. The case with $m_0 = 1$ is technically the C-Biased Truncated Design at $m = 1,000$, because it is possible for the Interval design to stop on the m_0^{th} observation. We see that the delaying, even for a small number of observations, reduces the probability substantially. The reason for this is that the first few observations have the greatest chance of taking the random sum S_n over the square root boundary.

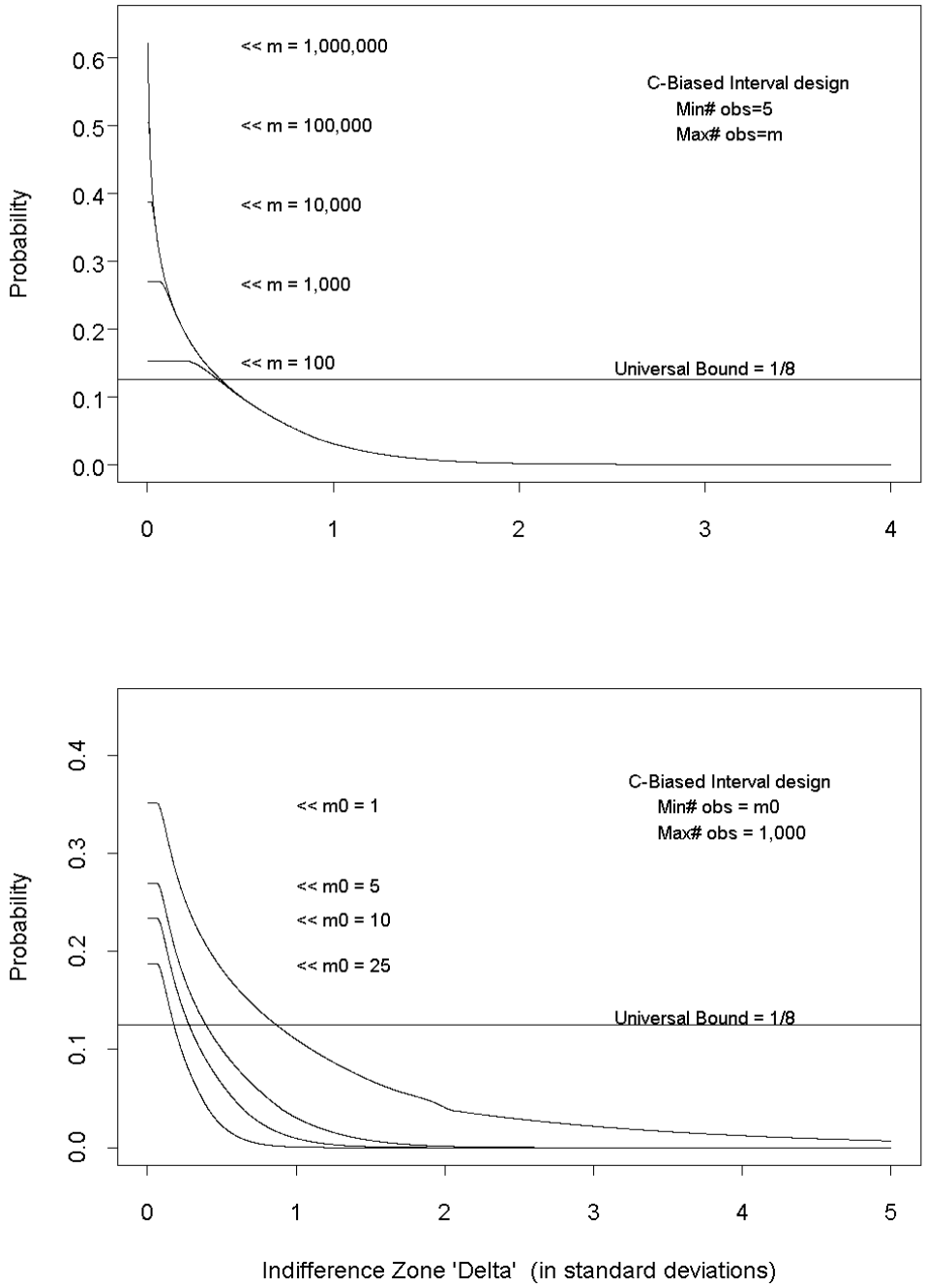

Figure 3.9: The C-Biased Interval design varying m_0

3.9 Summary

Within each class of designs, we examined the probability, under μ_0, of generating some alternative $\mu \geq \mu_1$ that is better supported over μ_0. The stopping rule we studied is designed to maximize this probability over all designs in the class. However, Armitage's paradox implies that when there are no restrictions on the sample size, the probability of generating evidence for some $\mu \geq \mu_1$ over μ_0 approaches unity as $\mu_1 \downarrow \mu_0$ (i.e. as the indifference zone becomes small). This is illustrated by figure 3.2 from the section on Open designs.

The probability of generating strong evidence for some alternative $\mu \geq \mu_1$ over μ_0 depends on the size of the indifference zone $\Delta = \mu_1 - \mu_0$, the upper limit on the sample size m, and the smallest possible sample size m_0 at which it is possible to stop the study. Delayed designs provide a way to reduce the probability for moderate to large indifference zones and Truncated designs provide a way to reduce the probability for small indifference zones. Thus the Interval designs seem the most practical choice, as they restrict the sample size by both delaying and truncating. In many cases, the probability of generating strong evidence for some alternative $\mu \geq \mu_1$ over μ_0 is at an acceptably low level.

Chapter 4

Beyond Normality

4.1 Introduction

The purpose of this chapter is to extend the results of Chapters 2 and 3 to a wider class of models. Initially we consider models from the single parameter exponential family and demonstrate that results in Chapters 2 and 3 can indeed be extended to this case. For models not in the single parameter exponential family class we consider a local analysis. More precisely, we prove that for alternatives close to the true value, the results of Chapter 2 and 3 apply to parametric models indexed by a scalar parameter of interest. In situations where fixed-dimensional nuisance parameters are present, the local analysis shows that these results apply for nearby alternatives when profile likelihood ratios are used to measure the strength of statistical evidence about the scalar parameter of interest, but do not apply when estimated likelihood ratios are used.

In the second section we prove that the curves in Chapter 2 which represent the probability that an experimental design generates strong or misleading evidence about the mean of a normal distribution provide a good, but local, approximation for parametric models in the single parameter exponential family

with their third moment equal to zero. For single parameter exponential family models with non-zero third moment, the curves in Chapter 2 apply in a similar fashion, incorporating an adjustment for skewness. For example, when using a Biased Open design under a single parameter exponential family model, the Tepee function represents the limiting probability of generating strong misleading evidence about the value of that parameter, regardless of the value of the third moment. But now the expected excess over the boundary, ρ, needs to be calculated for each distribution. The results in Chapter 3 depend on the same Brownian motion approximations as the curves in Chapter 2, plus a bound on random normal sums from Lorden [24]. Thus extension of the results in Chapter 3 to exponential family models is straightforward with the addition of a generalized bound, also from Lorden [24].

In the third section the normal model is examined in detail while in the fourth, fifth, and sixth sections a local analysis is considered for models not of the single parameter exponential family type. In a local analysis, the results of Chapters 2 and 3 apply under a wider class of models because, in part, Brownian motion is the natural limiting process associated with sums of independent log likelihood ratios. Standard Taylor expansion techniques are used to show that the sum of independent log likelihood ratios for nearby alternatives is approximately normally distributed in large samples. As we will see, the driving force behind these large sample results is the asymptotic normality of the score function. That is, the score function in large samples behaves like a normally distributed random walk, to which all of the Brownian motion approximations will apply. One interesting result is that, when nuisance parameters are present,

log profile likelihood ratios for nearby alternatives behave like 'true' log likelihood ratios in large samples, while estimated log likelihood ratios do not.

In Chapter 2, the Bump function was shown to represent the probability of generating misleading evidence for H_2 over H_1 under a fixed sample size design, where H_2 and H_1 were both fixed simple hypotheses about the mean of a normal distribution. Royall [31] shows that the Bump function is also the limiting curve representing the probability that a fixed sample size design generates misleading evidence about a scalar parameter of interest, say θ, under any smooth parametric model indexed by θ. However the variance, σ^2, is now replaced with the inverse of the expected Fisher Information $1/I(\theta)$. In situations where a fixed-dimensional nuisance parameter γ is present and a profile likelihood ratio is used in place of the 'true' likelihood ratio, he shows that the Bump function represents the limiting probability of generating misleading evidence in large samples, except now the variance σ^2 is replaced with the first element of the inverted information matrix, $1/(I_{\theta\theta}(1-\rho^2_{\theta\gamma}))$ (both $I_{\theta\theta}$ and $\rho^2_{\theta\gamma}$ are defined in section 4.5). In fact, for alternatives in a neighborhood of the true value, these results apply beyond the Bump function to any of the probability curves derived in Chapters 2 and 3, such as the Tepee function, C-Biased Open design, and the C-Biased Interval design with σ^2 replaced by the appropriate information quantity.

4.2 Distributions In The Exponential Family

Consider the model where observations $X_1, X_2, ...$ are i.i.d. single parameter exponential family with probability density function $f_\mu(X) = \exp\{\theta X - \psi(\theta)\}f(X)$.

Here $\psi(\theta)$ is convex, θ is the canonical parameter, and $\mu = \psi'(\theta)$ is a 1-1 strictly increasing function of θ, so one can indifferently consider μ to be a function of θ or θ a function of μ, say $\theta(\mu)$. General exponential family theory demonstrates $E_\theta[X_i] = \mu$ and $Var_\theta[X_i] = \psi''(\theta)$, where the prime denotes differentiation with respect to θ. Represent the likelihood function based on n observations as $L_n(\mu)$ and the sum of n observations as $S_n = X_1 + X_2 + \cdots + X_n$.

It will be convenient to assume that f is standardized in such as way that $\int xf(x)dx = 0$ and $\int x^2 f(x)dx = 1$, as a simple transformation on the X's will provide this case. This standardization implies that $\theta(0) = \psi'(0) = 0 (= \psi(0))$. In addition to the standardization, many of the references for this section consider the case where $\mu_1 < 0 < \mu_2$ and $\psi(\theta_1) = \psi(\theta_2)$ for the purpose of reducing algebraic computations. This case can always be achieved by redefining X_i to be $X_i - (\psi(\theta_2) - \psi(\theta_1))/(\theta_2 - \theta_1)$ and changing the corresponding value of μ.

4.2.1 Extending Results in Chapter 2

Recall the Biased Open design examined in Chapter 2, where the study terminated at the smallest sample size where strong evidence for H_2 over H_1 is obtained. Here both $H_2 : \mu = \mu_2$ and $H_1 : \mu = \mu_1$ are fixed simple hypotheses. After some algebra the stopping time can be expressed as

$$\begin{aligned} N &= \inf\left\{n : n \geq 1, \frac{L_n(\mu_2)}{L_n(\mu_1)} \geq k\right\} \\ &= \inf\left\{n : n \geq 1, S_n \geq \frac{\ln k}{\Delta} + n\frac{\psi(\theta_2) - \psi(\theta_1)}{\Delta}\right\} \\ &= \inf\left\{n : n \geq 1, S_n \geq b + n\eta\right\} \end{aligned} \tag{4.1}$$

where $\Delta = |\theta_2 - \theta_1|$.

Under the exponential family conditions outlined eariler, Siegmund uses Wald's likelihood ratio identity to demonstrate that ([37, (8.9),(10.6)], [35, 38])

$$P_{\mu_1}(N < \infty) = \frac{\exp\{-\rho_+ \Delta\}}{k} + o(e^{-1/\Delta}) \tag{4.2}$$

where ρ_+ is the expected excess over the boundary under the 'standardized' exponential family model. For unstandardized distributions and/or $\mu_1 > \mu_2$, we use $\Delta' = \left|\theta_2\sqrt{\psi''(\theta_2)} - \theta_1\sqrt{\psi''(\theta_1)}\right|$ in the first term, but not in the error. (Here, multiplying by $\sqrt{\psi''(\theta)}$ effectively unstandardizes the variance.) Remark (10.6) of [37] actually gives the error term for equation (4.2) as $o(\Delta)$. This is the correct error term if the random variables are only assumed to have mean zero and variance one, highlighting the fact that formula (4.2) can be a highly local result. However, the exponential family model implies stronger conditions on the tails of the distribution and the error term turns out to be exponentially small (see appendix 4 [37, p256]). Presumably, this implies that in exponential families, formula (4.2) provides a reasonably accurate approximation to the probability for moderately large Δ. Siegmund remarks that the exponential rate of convergence of the error helps to explain the extremely good numerical accuracy of these approximations [37, p266].

Formula (4.2) is the general Tepee function for a single parameter exponential family model (compare this with equation (2.6) for the normal case). While mathematically similar, Tepee functions for differing distributions often appear very different graphically, see section 4.2.3. Evaluation of equation (4.2) is straightforward since only the numerical calculation of ρ_+ is required. Methods for numerically evaluating ρ_+ can be found in [37, §10.4] and [44, p74]. Thus,

when using a Biased Open design, the general Tepee function describes the probability of generating misleading evidence for μ_2 over μ_1 locally, for models of the exponential family type. It is interesting to note that the probability of generating misleading evidence for μ_2 over μ_1 depends on the distance between their respective canonical parameters.

For designs like the Biased Truncated, Biased Delayed or Biased Interval, evaluation of the quantity $P_\mu(N \leq m)$, for some fixed m, is required to extend the entirety of results in Chapter 2 to exponential family models. It is convenient to express $P_\mu(N \leq m)$ in terms of the joint probability $P_\mu(N < m, S_m < c)$ where $c = b + \eta m$ as follows.

$$P_\mu\left(N \leq m\right) = P_\mu\left(S_m \geq b + \eta m\right) + P_\mu\left(N < m, S_m < c\right) \tag{4.3}$$

See appendix section A.3.2 for more details. Now suppose $\mu_1 < 0 < \mu_2$ in such a way that $\psi(\theta_1) = \psi(\theta_2)$. Theorem 10.45 of Siegmund [37, p220] gives for $j = 0$ (for μ_1) or 1 (for μ_2)

$$P_{\mu_j}\left(N < m, S_m < c\right) = \exp\left[-(-1)^j \Delta'(b + \rho_+)\right]$$

$$\times \Phi\left[\frac{c + \kappa/3 - 2(b + \rho_+)}{\sqrt{m + c\kappa/3}} + \frac{1}{2}(-1)^j \Delta' \sqrt{m + c\kappa/3}\right] + o(m^{-1/2}) \tag{4.4}$$

where $\Delta' = \left|\theta_2\sqrt{\psi''(\theta_2)} - \theta_1\sqrt{\psi''(\theta_1)}\right|$, $\kappa = E_0[X_1^3]$ where the distribution of X_1 is of the standardized exponential family form introduced earlier, and ρ_+ is the expected excess over the boundary. Equation (4.4) has the same form of the corresponding Brownian motion formula (A.4) with b replaced by $b+\rho_+$, c replaced by $c + \kappa/3$, and m replaced by $m + c\kappa/3$. Note that for normally distributed random variables, $\kappa = 0$ yields equation (A.8) (because of our standardization $\mu = \Delta/2$ and $\eta = 0$ in equation (A.8)).

When $\kappa \neq 0$, the exponential family distribution is skewed in some fashion. Now, the probabilities need to be adjusted to account for the non-normality of the exponential family distribution. This amounts to using (4.4) as the last term on the RHS of (4.3) and the first term on the RHS can be evaluated exactly. (If $\kappa = 0$ the first term on the RHS of (4.3) can be accurately approximated in moderately small sample sizes with a standard normal approximation, so that the results of Chapter 2 give a good approximation in this case.) For the Biased Truncated, Biased Delayed, and Biased Interval designs the derivations from Chapter 2 of the probabilities of generating strong and misleading evidence for H_2 over H_1 can be easily followed once a corrected version of $P_\mu(N \leq m)$ is obtained using equation (4.3). Finally, taking the limit of $P_\mu(N \leq m)$ as $m \to \infty$ yields the general Tepee fuction in equation (4.2). Note also that the Tepee function does not depend on κ.

4.2.2 Extending Results in Chapter 3

Recall the more flexible stopping rule examined in Chapter 3, which stopped the study at the smallest sample size when strong evidence for some alternative in $H_2 : \mu \geq \mu_2$ over $H_1 : \mu = \mu_1$ is obtained. Here both $H_2 : \mu = \mu_2$ and $H_1 : \mu = \mu_1$ are fixed simple hypotheses. Note that $\mu_2 > \mu_1$ implies $\theta_2 > \theta_1$. After some algebra the stopping time can be expressed as

$$\begin{aligned} N &= \inf\left\{n : n \geq 1, \sup_{\mu \geq \mu_2} \frac{L_n(\mu)}{L_n(\mu_1)} \geq k\right\} \\ &= \inf\left\{n : n \geq 1, \sup_{\theta \geq \theta_2} [\Delta S_n - n(\psi(\theta) - \psi(\theta_1))] \geq k\right\} \\ &= \inf\left\{n : n \geq 1, S_n \geq \inf_{\theta \geq \theta_2} \left[\frac{\ln k}{\Delta} + n\frac{\psi(\theta) - \psi(\theta_1)}{\Delta}\right]\right\} \end{aligned}$$

where $\Delta = \theta - \theta_1$. For simplicity and without loss of generality let $\theta_1 = 0$, so that stopping time N reduces to

$$N = \inf\left\{n : n \geq 1, S_n \geq \inf_{\theta \geq \theta_2}\left[\frac{\ln k}{\theta} + n\frac{\psi(\theta)}{\theta}\right]\right\} \tag{4.5}$$

Define $I(\theta) = \theta\psi'(\theta) - \psi(\theta)$. This quantity is called the information number by Lorden [24]. Now consider that

$$\frac{d}{d\theta}\left[\frac{\ln k}{\theta} + n\frac{\psi(\theta)}{\theta}\right] = \frac{1}{\theta^2}(nI(\theta) - \ln k) = 0$$

when $nI(\theta) = \ln k$. For $nI(\theta_2) \geq \ln k$ the derivative is non-negative on $[\theta_2, \infty)$, indicating a minimum at θ_2. For $nI(\theta_2) < \ln k$ the minimum is achieved at the solution, say θ_n, of $nI(\theta) = \ln k$, since the sign of the derivative changes from negative to positive at θ_n.

Now the stopping time (4.5) can be rewritten as

$$\begin{aligned} N &= \inf\left\{n : \left[1 \leq n < \frac{\ln k}{I(\theta_2)}, S_n \geq \frac{\ln k}{\theta_n} + n\frac{\psi(\theta_n)}{\theta_n}\right] \text{ or } \right. \\ &\qquad \left.\left[n \geq \frac{\ln k}{I(\theta_2)}, S_n \geq \frac{\ln k}{\theta_2} + n\frac{\psi(\theta_2)}{\theta_2}\right]\right\} \\ &= \min\{N_1, N_2\} \end{aligned}$$

where $N_1 = \inf\left\{n : 1 \leq n < \frac{\ln k}{I(\theta_2)}, S_n \geq \frac{\ln k}{\theta_n} + n\frac{\psi(\theta_n)}{\theta_n}\right\}$ or ∞ if no such n exists and $N_2 = \inf\left\{n : n \geq \frac{\ln k}{I(\theta_2)}, S_n \geq \left[\frac{\ln k}{\theta_2} + n\frac{\psi(\theta_2)}{\theta_2}\right]\right\}$ or ∞ if no such n exists.

Regarding n as a continuous variable on the interval from $[0, N^*]$ where $\ln k/I(\theta_*) = 1$, and $\ln k/I(\theta_2) \approx N^*$, with $nI(\theta_n) = \ln k$, Lorden [24, p638] demonstrates

$$P_0(N \leq N^*) = P_0(1 \leq N_1 \leq N^*) \leq \left(\frac{\ln k}{2\pi k^2}\right)^{1/2} \int_{\theta_2}^{\theta_*} \left(\frac{\psi''(\theta)}{I(\theta)}\right)^{1/2} d\theta \tag{4.6}$$

Often θ_* and θ_2 are replaced by their respective functions of sample size. For example, $\ln k/I(\theta_2) \approx N^*$ implies that θ_2 can be expressed as a function of N^*. Thus the probability can often be bounded by a function of the sample size, free of θ. See Chapter 3, section 3.7 for an example. Note that the generalized bound does not require the exponential distribution to be standardized.

We know from the previous section that $P(N_2 \leq m)$ is approximated by equation (4.3). For fixed sample sizes $m_0 < m$, we have

$$\begin{aligned} P_0\left(m_0 \leq N \leq m\right) &= P_0\left(m_0 \leq N < \frac{\ln k}{I(\theta_2)}\right) + P_0\left(\frac{\ln k}{I(\theta_2)} \leq N \leq m\right) \quad (4.7) \\ &= P_0\left(m_0 \leq N_1 < \frac{\ln k}{I(\theta_2)}\right) + P_0\left(\frac{\ln k}{I(\theta_2)} \leq N_2 \leq m\right) \end{aligned}$$

where the first term on the RHS is bounded as in equation (4.6) and the second term is approximated using equation (4.3).

The results of Chapter 3 are extended to exponential family models by using the extened Brownian motion tools derived for that purpose in the previous section and replacing the bound from Lorden's Theorem 2 with the generalized bound in equation (4.6). The derivations in Chapter 3 of the probability of generating strong evidence for some $\mu \geq \mu_2$ over μ_1 when μ_1 is the truth can be easily followed once a corrected version of $P_1(m_0 \leq N \leq m)$ is obtained using equation (4.7).

4.2.3 Example: Binomial Distribution

In this section the Tepee function for the Binomial distribution is developed. For simplicity, we begin with observations $X_1, X_2, \ldots, X_n$ i.i.d. Bernoulli(p). It is well known that the Bernoulli(p) distribution is of the single parameter

exponential family form with canonical parameter $\theta = \log[p/(1-p)]$ and $\psi(\theta) = \log[1 + \exp(\theta)]$. We examine the problem of using a Biased Open design to generate evidence for $H_2 : p = p_2$ over $H_1 : p = p_1 = 1/2$.

The stopping time for the Biased Open design stops the study at the smallest sample size where strong evidence for H_2 over H_1 is obtained. Now

$$N = \min\left\{n : n \geq 1, \frac{L_n(p_2)}{L_n(p_1)} \geq k\right\}$$

and by equation (4.2) we have

$$P_1\left(N < \infty\right) = \frac{\exp\{-\rho_+\Delta'\}}{k} + o(e^{-1/\Delta}) \tag{4.8}$$

where $\Delta' = \left|\log\left[\frac{p_2}{1-p_2}\right]\sqrt{4p_2(1-p_2)} - \log\left[\frac{p_1}{1-p_1}\right]\sqrt{4p_1(1-p_1)}\right|$, and $\rho_+ \cong 0.32$ (see next paragraph for details). For small to moderately large Δ, formula (4.8) gives a good approximation to the probability that a Biased Open design generates misleading evidence for p_2 over $p_1 = 1/2$.

The expected excess over the boundary, ρ_+, is evaluated under the 'standardized' Bernoulli distribution. To achieve the desired standardization define $Z_i = [(2X_i - 1) - (2p - 1)]/\sqrt{4p(1-p)}$, so that $E_p[Z_i] = 0$ and $Var_p[Z_i] = 1$. Notice that X_i is effectively transformed to a Bernoulli type random variable with $Var_p[2X_i - 1] = 4p(1-p)$. Now ρ_+ is actually a function of p, although the details are omitted here. This is the approach taken in Example 10.63 [37, p227], but Seigmund implicity uses $X_i^* = 2X_i - 1$ instead of the orignial random variable. Note that Siegmund's p^* is our p. Other methods are available [37, p107,p138], but seem to be less accurate [37, Table 10.1]. Using numerical integration techniques, Siegmund evaluates ρ_+ under the above standardized Bernoulli model as approximately 0.32.

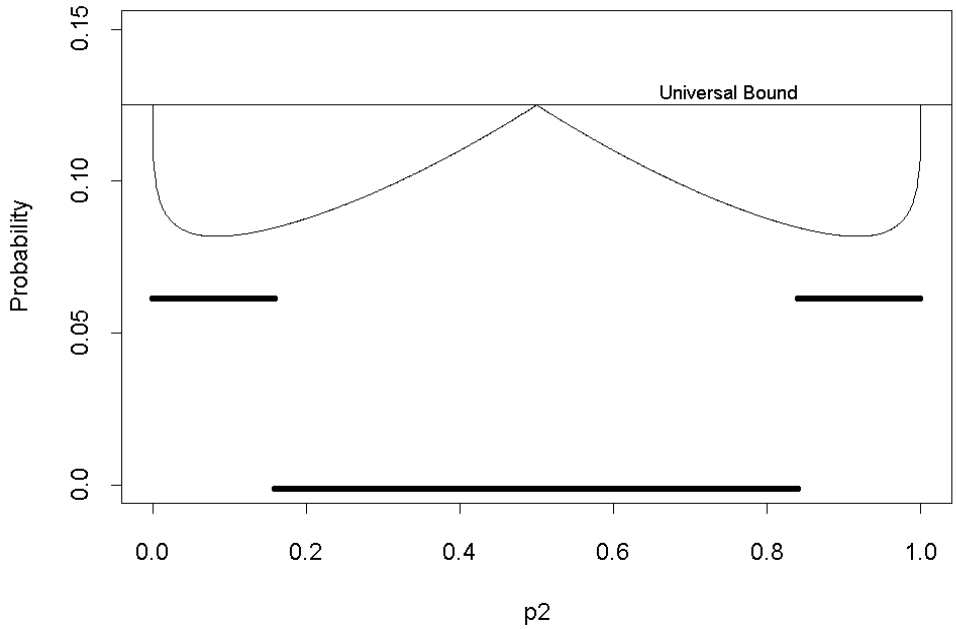

Figure 4.1: The Tepee Function for Bernoulli($p = 1/2$)

Figure 4.1 is a plot of the Tepee function (4.8) which represents the probability that the Biased Open design generates strong ($k = 8$) misleading evidence for p_2 over $p_1 = 1/2$. The horizontal lines represent the probability of observing misleading evidence for p_2 over $p_1 = 1/2$ on the 4^{th} observation. When $p_2 = 0$ or 1 the Tepee function equals the Universal bound $1/k$. This is a necessary component of the Tepee function because the probability of observing misleading evidence can achieve the Universal bound for such extreme p_2. For example, when $p_2 = 0$, the observations $X_1 + X_2 + X_3 = 0 = S_3$ give a likelihood ratio of $(1/0.5)^3 = 8$ in favor of $p_2 = 0$ over $p_1 = 1/2$. Under H_1, this evidence is misleading, and the probability of observing it is $P_1(S_3 = 0) = (0.5)^3 = 1/8$, the Universal bound.

Now suppose the observations $X_1, X_2, \ldots, X_n$ are i.i.d. Binomial(m, p) with m fixed. The Binomial distribution is also of the single parameter exponential

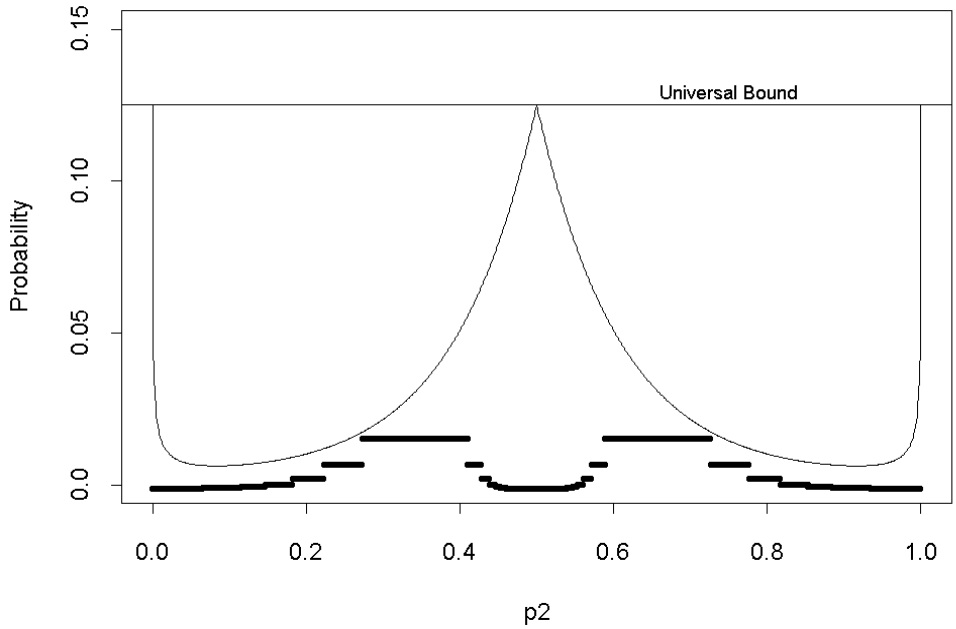

Figure 4.2: The Tepee Function for Binomial($m = 50, p = 1/2$)

family form with canonical parameter $\theta = \log[p/(1-p)]$ and $\psi(\theta) = m \log[1 + \exp(\theta)]$. Using an argument similar to that used to standardize the Bernoulli random variables (with $X^* = 2X_i - m$), formula (4.8) can be used for the Binomial distribution with $\Delta = \left|\log\left[\frac{p_2}{1-p_2}\right]\sqrt{4mp_2(1-p_2)} - \log\left[\frac{p_1}{1-p_1}\right]\sqrt{4mp_1(1-p_1)}\right|$. Although not explicity sanctioned by Siegmund, we use the same $\rho_+ \cong 0.32$ for the Binomial distribution. It seems natural that expected excess over the boundary will remain the same because both the slope of the boundary and the average drift per observation increase by a factor of m.

Figure 4.2 is a plot of the Tepee function (4.8) for the Binomial($m = 50, p_1 = 1/2$) distribution and it represents the probability that the Biased Open design generates strong ($k = 8$) misleading evidence for p_2 over $p_1 = 1/2$. The horizontal lines represent the probability of observing misleading evidence for p_2 over $p_1 = 1/2$ on the 1^{st} Binomial observation. Here the Tepee function is beginning

to look more like the corresponding Tepee function for the normal case, except it increases to the universal bound for extreme p_2. But under this Binomial model, there is essentially no chance of generating misleading evidence for extreme p_2 and the Tepee function incorrectly approximates this probability as quite large. This indicates how poor the local approximation in equation (4.8) can be if applied universally to all $\Delta = |\theta_2 - \theta_1|$.

4.3 The Normal Model

Before considering the local analysis in the following sections, it is instructive to examine the normal model in detail. Consider the model where observations $X_1, X_2, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$ with known variance σ^2, density function $f(X_i; \mu)$, and log likelihood function based on n observations $l_n(\mu)$.

4.3.1 Log Likelihood Ratios

The log likelihood ratio is exactly normally distributed, as may be expected under this model.

Remark 4.1 *(Normal Case). If $X_1, X_2, \ldots, X_n$ are i.i.d. $N(\mu_1, \sigma^2)$ with known variance σ^2, then for $\mu_2 = \mu_1 + c\sigma/\sqrt{n}$ ($c > 0$ is a constant), the log likelihood ratio is normally distributed with mean $-c/2$ and variance c^2.*

$$l_n(\mu_2) - l_n(\mu_1) = cZ - \frac{c^2}{2} \stackrel{d}{=} N\left(-\frac{c^2}{2}, c^2\right) \tag{4.9}$$

where $Z \sim N(0, 1)$.

Proof:

Direct calculation demonstrates

$$\begin{aligned} l_n(\mu_2) - l_n(\mu_1) &= \frac{(\mu_2 - \mu_1)S_n}{\sigma^2} - n\frac{\mu_2^2 - \mu_1^2}{2\sigma^2} \\ &= c\frac{S_n - n\mu_1}{\sigma\sqrt{n}} - \frac{c^2}{2} \\ &= cZ - \frac{c^2}{2} \end{aligned}$$

where Z is a standard normal random variable.

QED •

4.3.2 Approximating The Discrete Stopping Time

The stopping rule examined in Chapter 2 stopped the study at the smallest sample size where strong evidence for H_2 over H_1 is obtained. Here both $H_1 : \mu = \mu_1$ and $H_2 : \mu = \mu_2$ are fixed simple hypotheses. After some algebra the stopping time can be re-expressed as the following.

$$\begin{aligned} N &= \inf\left\{n : n \geq 1, \frac{L_n(\mu_2)}{L_n(\mu_1)} \geq k\right\} \\ &= \inf\left\{n : n \geq 1, \frac{S_n - n\mu_1}{\sigma^2} \geq \frac{\ln k}{\Delta} + n\frac{\Delta}{2\sigma^2}\right\} \\ &= \inf\left\{n : n \geq 1, \sigma S_n^* \geq \frac{\sigma \ln k}{\Delta} + n\frac{\Delta}{2\sigma}\right\} \end{aligned}$$

where $\Delta = |\mu_2 - \mu_1|$, and

$$\frac{S_n - n\mu_1}{\sigma^2} = \frac{\partial l_n(\mu)}{\partial \mu}|_{\mu_1} = S_n^*$$

is the score function at μ_1. Under μ_1, direct calculation shows the score function is normally distributed with mean zero and variance n/σ^2. In this case, stopping time N represents the first time misleading evidence for μ_2 over μ_1 is obtained.

The notation S_n^* is used to emphasize the fact that the score function is itself a random walk. The behavior of S_n^* can be approximated with the Brownian motion process $W(n/\sigma^2)$ with drift zero and variance per unit time $1/\sigma^2$. In this case the joint distributions of $S_n^*, n = 1, 2, \ldots$ and $W(n/\sigma^2), n = 1, 2, \ldots$ are exactly the same. Now $W(n/\sigma^2)\sigma$ is Brownian motion with drift zero and variance per unit time one. For continuous $t > 0$, define the stopping time τ to be

$$\tau \quad = \quad \inf \left\{ t : t > 0, W(t) \geq \frac{\sigma \ln k}{\Delta} + t\frac{\Delta}{2\sigma} \right\}$$

where $\Delta = |\mu_2 - \mu_1|$, and $W(t)$ is Brownian motion with drift zero and variance per unit time one. Now τ is the continuous time version of N and the Brownian motion approximation $P_{\mu_1}(N \leq m) \approx P_{\mu_1}(\tau \leq m)$ for some fixed m holds as described in the appendix, as do the adjusted Brownian motion probabilities $P_{\mu_1}(N \leq m) \cong P_{\mu_1}(\tilde{\tau} \leq m)$ in section A.3.3.

4.3.3 The Fixed Sample Size Design Revisited

Here we derive the probability that some alternative is better supported over the true mean μ_1 under a Fixed Sample Size design.

Theorem 4.1 *(Maximum Probability under a Fixed Sample Size design) For $X_1, X_2, \ldots, X_n$ are i.i.d. $N(\mu_1, \sigma^2)$ with known variance σ^2*

$$P_1 \left(\max_{\mu} \frac{L_n(\mu)}{L_n(\mu_1)} \geq k \right) = 2\Phi \left[-\sqrt{2 \ln k} \right]$$

Proof:

The result is verified by direct calculation using

$$\max_{\mu} L_n(\mu) = L_n(\overline{X}_n)$$

QED •

The probability that some alternative is better supported over the true mean μ_1 under a fixed sample size design is $2\Phi[-\sqrt{2\ln k}]$. This is twice the maximum of the C-Bump function of section 3.3, because only the 'one sided' case was considered in Chapter 3.

4.4 Distributions Indexed By A Single Parameter

In this section we consider the model where observations $X_1, X_2, \ldots, X_n$ are i.i.d. $f(X_i;\theta)$, where f is a smooth function of the real-valued parameter θ. Represent the log likelihood function for n observations as $l_n(\theta)$.

4.4.1 Log Likelihood Ratios

In large samples the log likelihood ratio at nearby alternatives is approximately normally distributed.

Theorem 4.2 *(Scalar Parameter Case). If $X_1, X_2, \ldots, X_n$ are i.i.d. $f(X_i;\theta_1)$ where f is a smooth function of the real-valued parameter θ, then in large samples for $\theta_2 = \theta_1 + c/\sqrt{nI(\theta_1)}$ ($c > 0$ is a constant), the log likelihood ratio is approximately normally distributed with mean $-c/2$ and variance c^2.*

$$l_n(\theta_2) - l_n(\theta_1) \approx \frac{c}{\sqrt{I(\theta_1)}} A_n + \frac{c^2}{2I(\theta_1)} B_n \xrightarrow{d} N\left(-\frac{c^2}{2}, c^2\right) \text{ as } n \to \infty \quad (4.10)$$

where

$$A_n = \frac{1}{\sqrt{n}} \frac{\partial l_n(\theta)}{\partial \theta}|_{\theta_1} \text{ and } B_n = \frac{1}{n} \frac{\partial^2 l_n(\theta)}{\partial \theta^2}|_{\theta_1} \text{ and } I(\theta_1) = E\left[\frac{\partial^2 l_n(\theta)}{\partial \theta^2}|_{\theta_1}\right]$$

is the expected Fisher information.

Proof:

Because θ_2 is in a $O(n^{-1/2})$ neighborhood of θ_1, direct Taylor expansion of the log likelihood ratio around θ_1 gives

$$l_n(\theta_2) - l_n(\theta_1) \approx \frac{c}{\sqrt{nI(\theta_1)}} \frac{\partial}{\partial\theta} l_n(\theta)|_{\theta_1} + \frac{c^2}{2nI(\theta_1)} \frac{\partial^2}{\partial\theta^2} l_n(\theta)|_{\theta_1} + R_n$$

The remainder is effectively a sum of n terms times a factor of $n^{-3/2}$. Thus, under mild conditions, $R_n = O_p(n^{-1/2})$. Now

$$A_n = \frac{1}{\sqrt{n}} \frac{\partial l_n(\theta)}{\partial\theta}|_{\theta_1} \xrightarrow{d} N(0, I(\theta_1))$$

and

$$B_n = \frac{1}{n} \frac{\partial^2 l_n(\theta)}{\partial\theta^2}|_{\theta_1} \xrightarrow{p} -I(\theta_1)$$

are standard results. Let $Z \sim N(0,1)$, then we have

$$l_n(\theta_2) - l_n(\theta_1) \xrightarrow{d} cZ - \frac{c^2}{2} \stackrel{d}{=} N\left(-\frac{c^2}{2}, c^2\right)$$

QED •

4.4.2 Approximating The Discrete Stopping Time

Now consider the stopping time N examined in Chapter 2, which equaled the smallest sample size where strong evidence for H_2 over H_1 is obtained. We consider the case of fixed simple hypothesis $H_1 : \theta = \theta_1$ and $H_2 : \theta = \theta_2$ where θ_2 is in a $O(n^{-1/2})$ neighborhood of θ_1. Using Taylor expansion methods, the stopping time N can be re-expressed as

$$N = \inf\left\{n : n \geq 1, \frac{L_n(\theta_2)}{L_n(\theta_1)} \geq k\right\}$$

$$= \inf\left\{n : n \geq 1, \frac{\partial l_n(\theta)}{\partial\theta}|_{\theta_1} \geq \frac{\ln k}{\Delta} + n\frac{-B_n\Delta}{2} - O_p(n^{-1/2})\right\}$$

$$= \inf\left\{n : n \geq 1, \frac{S_n^*}{\sqrt{I(\theta_1)}} \geq \frac{\ln k}{\Delta\sqrt{I(\theta_1)}} + n\frac{-B_n\Delta}{2\sqrt{I(\theta_1)}} - O_p(n^{-1/2})\right\}$$

where $\Delta = |\theta_2 - \theta_1|$, and

$$\frac{\partial l_n(\theta)}{\partial\theta}|_{\theta_1} = S_n^*$$

is the score function at θ_1. Under H_1, the stopping time N represents the first time misleading evidence for θ_2 is obtained.

It is well known that, under θ_1, the score function behaves asymptotically like a random walk with mean zero and variance $nI(\theta_1)$. Furthermore, under θ_1, $B_n \xrightarrow{p} -I(\theta_1)$ as $n \to \infty$. Because of this, the behavior of S_n^* can be approximated asymptotically with the Brownian motion process $W(nI(\theta_1))$ with drift zero and variance per unit time $I(\theta_1)$. In large samples the joint distributions of $S_n^*, n = 1, 2, \ldots$ and $W(nI(\theta_1)), n = 1, 2, \ldots$ are approximately the same. It can be verified that $W(nI(\theta_1))/\sqrt{I(\theta_1)}$ is Brownian motion with drift zero and variance per unit time one. For continuous $t > 0$, define the stopping time τ as

$$\tau = \inf\left\{t : t > 0, W(t) \geq \frac{\ln k}{\Delta\sqrt{I(\theta_1)}} + t\frac{\Delta\sqrt{I(\theta_1)}}{2}\right\}$$

where $\Delta = |\theta_2 - \theta_1|$ and $W(t)$ is Brownian motion with drift zero and variance per unit time one. Now, under θ_1, τ is the limiting continuous time version of N and the Brownian motion approximations $P_{\theta_1}(N \leq m) \approx P_{\theta_1}(\tau \leq m)$ for some fixed m hold for large samples at nearby alternatives as described in the appendix, as do the adjusted Brownian motion probabilities $P_{\theta_1}(N \leq m) \cong P_{\theta_1}(\tilde{\tau} \leq m)$ in section A.3.3.

An identical argument is used when θ_2 is fixed and θ_1 is approaching θ_2. Thus results of Chapter 2 apply for nearby alternatives in large samples, replacing σ^2 with $1/I(\theta_1)$. The results in Chapter 3 apply for nearby alternatives in large samples as well, because the replacement of σ^2 by $1/I(\theta_1)$ does not effect the usage of Theorem two of Lorden [24].

4.4.3 The Fixed Sample Size Design Revisited

Here we derive the probability that some alternative is better supported over the true value θ_1 under a Fixed Sample Size design with a large sample size.

Theorem 4.3 *(Maximum Probability Under a Fixed Sample Size Design) If $X_1, X_2, \ldots, X_n$ are i.i.d. $f(X_i;\theta_1)$ where f is a smooth function of the real-valued parameter θ, then*

$$\lim_{n\to\infty} P_1\left(\max_{\theta}\frac{L_n(\theta)}{L_n(\theta_1)} \geq k\right) = 2\Phi\left[-\sqrt{2\ln k}\right]$$

Proof:

Let $\hat{\theta}$ be the MLE for θ. Then we have

$$\max_{\theta}\frac{L_n(\theta)}{L_n(\theta_1)} = \frac{L_n(\hat{\theta})}{L_n(\theta_1)}$$

By Taylor series expansion of $\hat{\theta}$ around θ_1 in the log likelihood ratio we have

$$l_n(\hat{\theta}) - l_n(\theta_1) = (\hat{\theta}-\theta_1)\frac{\partial l_n(\theta)}{\partial\theta}|_{\theta_1} + \frac{(\hat{\theta}-\theta_1)^2}{2}\frac{\partial^2 l_n(\theta)}{\partial\theta^2}|_{\theta_1} + R \tag{4.11}$$

The remainder R is effectively a sum of n terms times a factor of $(\hat{\theta}-\theta_1)^3$. From the likelihood equation we have

$$0 = \frac{\partial l_n(\theta)}{\partial\theta}|_{\hat{\theta}} \approx \frac{\partial l_n(\theta)}{\partial\theta}|_{\theta_1} + (\hat{\theta}-\theta_1)\frac{\partial^2 l_n(\theta)}{\partial\theta^2}|_{\theta_1} \tag{4.12}$$

Solving equation (4.12) for $(\hat{\theta} - \theta_1)$ and substituting into equation (4.11) yields

$$l_n(\hat{\theta}) - l_n(\theta_1) = -\left[\frac{\partial l_n(\theta)}{\partial \theta}|_{\theta_1}\right]^2 / 2\left[\frac{\partial^2 l_n(\theta)}{\partial \theta^2}|_{\theta_1}\right] + R$$

But $(\hat{\theta} - \theta_1)^3 = O_p(n^{-3/2})$ so that $R = O_p(n^{-1/2})$ under mild conditions. Now both

$$\frac{1}{\sqrt{n}}\frac{\partial l_n(\theta)}{\partial \theta}|_{\theta_1} \xrightarrow{d} N\left(0, I(\theta_1)\right)$$

$$\frac{1}{n}\frac{\partial^2 l_n(\theta)}{\partial \theta^2}|_{\theta_1} \xrightarrow{p} -I(\theta_1)$$

are standard results. Thus we have

$$\left|\frac{\partial l_n(\theta)}{\partial \theta}|_{\theta_1}\right| / \sqrt{-\frac{\partial^2 l_n(\theta)}{\partial \theta^2}|_{\theta_1}} + R \xrightarrow{d} |Z| \quad \text{where } Z \sim N(0,1) \tag{4.13}$$

and the result follows.

QED •

4.5 Eliminating Nuisance Parameters: The Profile Likelihood

In this section we consider observations $X_1, X_2, \ldots, X_n$ as i.i.d. $f(X_i; \theta, \gamma)$ where f is a smooth function and both θ and γ are fixed dimensional parameters. Let $l_n(\theta, \gamma)$ be the log likelihood function. Suppose θ represents the parameter of interest, and γ the nuisance parameter. One approach to eliminating the nuisance parameter is to use the profile likelihood function for θ as if it were a true likelihood function. The profile likelihood function maximizes with respect to the nuisance parameter for each value of the parameter of interest. For fixed θ, define the profile likelihood function to be [21]

$$\max_{\gamma} L_n(\theta, \gamma) = L_n(\theta, \widehat{\gamma}(\theta)) = L_{pn}(\theta) \tag{4.14}$$

We only consider the case where θ and γ are real-valued. The extension to the vector case is straightforward.

4.5.1 Log Profile Likelihood Ratios

In large samples, the log profile likelihood ratio at nearby alternatives is approximately normally distributed and has the same limiting distribution as a 'true' likelihood ratio. This result has been proved elsewhere [31].

Theorem 4.4 *(Profile Likelihood Case) For $X_1, X_2, \ldots, X_n$ i.i.d. $f(X_i; \theta_1, \gamma_1)$ where f is a smooth function, both θ_1 and γ_1 are real-valued, and $\theta_2 = \theta_1 + c/\sqrt{nI_{\theta\theta}(1-\rho^2_{\theta\gamma})}$ ($c > 0$ is a constant), then in large samples the log profile likelihood ratio for θ_2 over θ_1 is approximately normally distributed with mean $-c/2$ and variance c^2.*

$$\begin{aligned} l_{pn}(\theta_2) - l_{pn}(\theta_1) \quad &\approx \quad \frac{c}{\sqrt{I_{\theta\theta}(1-\rho^2_{\theta\gamma})}} A_{pn} + \frac{c^2}{2I_{\theta\theta}(1-\rho^2_{\theta\gamma})} B_{pn} + R_n \qquad (4.15) \\ &\xrightarrow{d} \quad N\left(-\frac{c^2}{2}, c^2\right) \quad \text{as } n \to \infty \end{aligned}$$

where

$$A_{pn} = \frac{1}{\sqrt{n}} \frac{\partial l_n(\theta,\gamma)}{\partial \theta}|_{(\theta_1, \widehat{\gamma}(\theta_1))}$$

$$B_{pn} = \frac{1}{n} \frac{\partial^2 l_n(\theta,\gamma)}{\partial \theta^2}|_{(\theta_1, \widehat{\gamma}(\theta_1))} + \frac{1}{n} \frac{\partial^2 l_n(\theta,\gamma)}{\partial \theta \partial \gamma}|_{(\theta_1, \widehat{\gamma}(\theta_1))} \frac{d\widehat{\gamma}(\theta)}{d\theta}|_{\theta_1}$$

$$I_{\theta\theta} = E\left[\frac{\partial^2 l_n(\theta,\gamma)}{\partial \theta^2}|_{(\theta_1, \widehat{\gamma}(\theta_1))}\right]$$

$$\rho^2_{\theta\gamma} = \frac{I^2_{\theta\gamma}}{I_{\theta\theta} I_{\gamma\gamma}}$$

Proof:

Because θ_2 is in a $O(n^{-1/2})$ neighborhood of θ_1, direct Taylor expansion of the log likelihood ratio around θ_1 gives

$$\begin{aligned} l_{pn}(\theta_2) - l_{pn}(\theta_1) &\approx \frac{c}{\sqrt{nI_{\theta\theta}(1-\rho^2_{\theta\gamma})}} \frac{\partial}{\partial\theta} l_n(\theta,\gamma)|_{(\theta_1,\widehat{\gamma}(\theta_1))} \\ &\quad + \frac{c^2}{2nI_{\theta\theta}(1-\rho^2_{\theta\gamma})} \frac{\partial^2}{\partial\theta^2} l_n(\theta,\gamma)|_{(\theta_1,\widehat{\gamma}(\theta_1))} + R_n \end{aligned}$$

The remainder is effectively a sum of n terms times a factor of $n^{-3/2}$. Thus, under mild conditions, $R_n = O_p(n^{-1/2})$. Lemma 1 will show that $A_{pn} \xrightarrow{d} N(0, I_{\theta\theta}(1-\rho^2_{\theta\gamma}))$, while lemmas 2 and 3 will show that $B_{pn} \xrightarrow{p} -I_{\theta\theta}(1-\rho^2_{\theta\gamma})$. Thus $l_{pn}(\theta_2) - l_{pn}(\theta_1) \xrightarrow{d} cZ - c^2/2$, where $Z \sim N(0,1)$ and the result follows.

Lemma 1. $\frac{1}{\sqrt{n}} \frac{\partial}{\partial\theta} l_n(\theta,\gamma)|_{(\theta_1,\widehat{\gamma}(\theta_1))} \xrightarrow{d} N(0, I_{\theta\theta}(1-\rho^2_{\theta\gamma}))$

Proof: (Lemma 1)

$$\frac{\partial}{\partial\theta} l_n(\theta,\gamma)|_{(\theta_1,\widehat{\gamma}(\theta_1))} = \frac{\partial l_n(\theta,\gamma)}{\partial\theta}|_{(\theta_1,\widehat{\gamma}(\theta_1))} + \frac{\partial l_n(\theta,\gamma)}{\partial\gamma}|_{(\theta_1,\widehat{\gamma}(\theta_1))}$$

by the chain rule. Note that the second term is zero because $\widehat{\gamma}(\theta_1)$ is the MLE at fixed θ_1. Now by a Taylor expansion around γ_1 we have both

$$\frac{\partial l_n(\theta,\gamma)}{\partial\theta}|_{(\theta_1,\widehat{\gamma}(\theta_1))} \approx \frac{\partial l_n(\theta,\gamma)}{\partial\theta}|_{(\theta_1,\gamma_1)} + \frac{\partial^2 l_n(\theta,\gamma)}{\partial\theta\partial\gamma}|_{(\theta_1,\gamma_1)}(\widehat{\gamma}(\theta_1) - \gamma_1)$$

and

$$\begin{aligned} 0 = \frac{\partial l_n(\theta,\gamma)}{\partial\gamma}|_{(\theta_1,\widehat{\gamma}(\theta_1))} &\approx \frac{\partial l_n(\theta,\gamma)}{\partial\gamma}|_{(\theta_1,\gamma_1)} + \frac{\partial^2 l_n(\theta,\gamma)}{\partial\gamma^2}|_{(\theta_1,\gamma_1)}(\widehat{\gamma}(\theta_1) - \gamma_1) \\ \Longrightarrow \quad (\widehat{\gamma}(\theta_1) - \gamma_1) &= \frac{\frac{\partial l_n(\theta,\gamma)}{\partial\gamma}|_{(\theta_1,\gamma_1)}}{-\frac{\partial^2 l_n(\theta,\gamma)}{\partial\gamma^2}|_{(\theta_1,\gamma_1)}} \end{aligned}$$

Substituting into the first expression,

$$\frac{\partial l_n(\theta,\gamma)}{\partial\theta}|_{(\theta_1,\widehat{\gamma}(\theta_1))} \approx \frac{\partial l_n(\theta,\gamma)}{\partial\theta}|_{(\theta_1,\gamma_1)} + \frac{\frac{\partial^2 l_n(\theta,\gamma)}{\partial\theta\partial\gamma}|_{(\theta_1,\gamma_1)}}{-\frac{\partial^2 l_n(\theta,\gamma)}{\partial\gamma^2}|_{(\theta_1,\gamma_1)}} \frac{\partial l_n(\theta,\gamma)}{\partial\gamma}|_{(\theta_1,\gamma_1)} \qquad (4.16)$$

The ratio of second partials converges to $-I_{\theta\gamma}/I_{\gamma\gamma}$ (see Lemmas 2 and 3), and the conclusion follows from the asymptotic bivariate normality of the two score functions.

Lemma 2. $\frac{1}{n}\frac{\partial^2 l_n(\theta,\gamma)}{\partial\theta\partial\gamma}|_{(\theta_1,\widehat{\gamma}(\theta_1))} \xrightarrow{p} -I_{\theta\gamma}$ and $\frac{1}{n}\frac{\partial^2 l_n(\theta,\gamma)}{\partial\gamma\partial\gamma}|_{(\theta_1,\widehat{\gamma}(\theta_1))} \xrightarrow{p} -I_{\gamma\gamma}$

Proof: (Lemma 2)

These results require only standard Taylor expansion arguments. Hence only one result is demonstrated, the others following by a similar argument. By Taylor series expansion around γ_1 we have

$$\begin{aligned}\frac{1}{n}\frac{\partial^2 l_n(\theta,\gamma)}{\partial\theta\partial\gamma}|_{(\theta_1,\widehat{\gamma}(\theta_1))} &\approx \frac{1}{n}\frac{\partial^2 l_n(\theta,\gamma)}{\partial\theta\partial\gamma}|_{(\theta_1,\gamma_1)} + \frac{1}{n}\frac{\partial^3 l_n(\theta,\gamma)}{\partial\theta\partial\gamma^2}|_{(\theta_1,\gamma_1)}(\widehat{\gamma}(\theta_1)-\gamma_1) \\ &\xrightarrow{p} -I_{\theta\gamma}\end{aligned}$$

because $\widehat{\gamma}(\theta_1) \xrightarrow{p} \gamma_1$

Lemma 3. $\frac{d\widehat{\gamma}(\theta)}{d\theta}|_\theta \xrightarrow{p} -I_{\theta\gamma}/I_{\gamma\gamma}$

Proof: (Lemma 3)

$\frac{\partial l_n(\theta,\gamma)}{\partial\gamma}|_{\widehat{\gamma}(\theta)} = 0$ for all θ implies that the function on the left-hand side, say $g(\theta,\widehat{\gamma}(\theta))$, is a constant. Thus $\frac{dg}{d\theta} = 0$. But using the chain rule we have

$$\begin{aligned}\frac{dg}{d\theta} &= \frac{\partial g(\theta,\gamma)}{\partial\theta}|_{\widehat{\gamma}(\theta)} + \frac{\partial g(\theta,\gamma)}{\partial\gamma}|_{\widehat{\gamma}(\theta)}\frac{d\widehat{\gamma}(\theta)}{d\theta} \\ &= \frac{\partial^2 l_n(\theta,\gamma)}{\partial\theta\partial\gamma}|_{\widehat{\gamma}(\theta)} + \frac{\partial^2 l_n(\theta,\gamma)}{\partial\gamma^2}|_{\widehat{\gamma}(\theta)}\frac{d\widehat{\gamma}(\theta)}{d\theta}\end{aligned}$$

and the result follows from Lemma 2.

Using Lemmas 2 and 3 while applying the Chain rule twice on the second derivative from the inital Taylor expansion of the log profile likelihood ratio

yields

$$\begin{aligned}
B_{pn} &= \frac{1}{n}\frac{\partial^2}{\partial\theta^2} l_n(\theta,\gamma)|_{(\theta_1,\widehat{\gamma}(\theta_1))} \\
&= \frac{1}{n}\frac{\partial}{\partial\theta}\left[\frac{\partial l_n(\theta,\gamma)}{\partial\theta}|_{(\theta_1,\widehat{\gamma}(\theta_1))} + \frac{\partial l_n(\theta,\gamma)}{\partial\gamma}|_{(\theta_1,\widehat{\gamma}(\theta_1))}\frac{d\widehat{\gamma}(\theta)}{d\theta}|_{\theta_1}\right] \\
&= \frac{1}{n}\frac{\partial}{\partial\theta}\left[\frac{\partial l_n(\theta,\gamma)}{\partial\theta}|_{(\theta_1,\widehat{\gamma}(\theta_1))}\right] \\
&= \frac{1}{n}\frac{\partial^2 l_n(\theta,\gamma)}{\partial\theta^2}|_{(\theta_1,\widehat{\gamma}(\theta_1))} + \frac{1}{n}\frac{\partial^2 l_n(\theta,\gamma)}{\partial\gamma\partial\theta}|_{(\theta_1,\widehat{\gamma}(\theta_1))}\frac{d\widehat{\gamma}(\theta)}{d\theta}|_{\theta_1} \qquad (4.17) \\
&\xrightarrow{p} -I_{\theta\theta}(1-\rho^2_{\theta\gamma}) \text{ as } n\to\infty
\end{aligned}$$

QED •

4.5.2 Approximating The Discrete Stopping Time

Recall the stopping time N examined in Chapter 2, which equaled the smallest sample size where strong evidence for H_2 over H_1 is obtained. Here consider the case of fixed simple hypothesis $H_1 : \theta = \theta_1$ and $H_2 : \theta = \theta_2$ where θ_2 is in a $O(n^{-1/2})$ neighborhood of θ_1. Using Taylor expansion methods, the stopping time N can be re-expressed as

$$\begin{aligned}
N &= \inf\left\{n : n\geq 1, \frac{L_{pn}(\theta_2)}{L_{pn}(\theta_1)} \geq k\right\} \\
&= \inf\left\{n : n\geq 1, \frac{\partial l_n(\theta,\gamma)}{\partial\theta}|_{(\theta_1,\widehat{\gamma}(\theta_1))} \geq \frac{\ln k}{\Delta} + n\frac{-B_{pn}\Delta}{2} - O(n^{-1/2})\right\} \\
&= \inf\left\{n : n\geq 1, \frac{S^*_n}{\sqrt{I_{\theta\theta}(1-\rho^2_{\theta\gamma})}} \geq \frac{\ln k}{\Delta\sqrt{I_{\theta\theta}(1-\rho^2_{\theta\gamma})}} + n\frac{-B_{pn}\Delta}{2\sqrt{I_{\theta\theta}(1-\rho^2_{\theta\gamma})}} - O(n^{-1/2})\right\}
\end{aligned}$$

where $\Delta = |\theta_2 - \theta_1|$, and

$$\frac{\partial l_n(\theta,\gamma)}{\partial\theta}|_{(\theta_1,\widehat{\gamma}(\theta_1))} = S^*_n$$

is the profile score function at θ_1. Under θ_1, the stopping time N represents the first time misleading evidence for θ_2 is obtained.

Arguments similar to those used to prove Theorem 4.4 show that under θ_1, the profile score function behaves asymptotically like a random walk with mean zero and variance $nI_{\theta\theta}(1-\rho^2_{\theta\gamma})$, where $I_{\theta\theta}$ and $(1-\rho^2_{\theta\gamma})$ are specified by Theorem 4.4. Likewise, similar Taylor expansion arguments demonstrate that under θ_1, $B_{pn} \xrightarrow{p} I_{\theta\theta}(1-\rho^2_{\theta\gamma})$ as $n \to \infty$.

Thus, under H_1, the behavior of S^*_n can be approximated asymptotically with a Brownian motion process $W(nI_{\theta\theta}(1-\rho^2_{\theta\gamma}))$ with drift zero and variance per unit time $I_{\theta\theta}(1-\rho^2_{\theta\gamma})$ [37, p64]. In large samples the joint distributions of $S^*_n, n = 1, 2, \ldots$ and $W(nI_{\theta\theta}(1-\rho^2_{\theta\gamma})), n = 1, 2, \ldots$ are approximately equal. It can be verified that $W(nI_{\theta\theta}(1-\rho^2_{\theta\gamma}))/\sqrt{I_{\theta\theta}(1-\rho^2_{\theta\gamma})}$ is Brownian motion with drift zero and variance per unit time one. For continuous $t > 0$, define the stopping time τ as

$$\tau \;=\; \inf\left\{ t : t > 0, W(t) \geq \frac{\ln k}{\Delta\sqrt{I_{\theta\theta}(1-\rho^2_{\theta\gamma})}} + t\frac{\Delta\sqrt{I_{\theta\theta}(1-\rho^2_{\theta\gamma})}}{2} \right\}$$

where $\Delta = |\theta_2 - \theta_1|$ and $W(t)$ is Brownian motion with drift zero and variance per unit time one. Now, under θ_1, τ is the limiting continuous time version of N and the Brownian motion approximations $P_{\theta_1}(N \leq m) \approx P_{\theta_1}(\tau \leq m)$ for some fixed m hold for large samples at nearby alternatives as described in the appendix, as do the adjusted Brownian motion probabilities $P_{\theta_1}(N \leq m) \cong P_{\theta_1}(\widetilde{\tau} \leq m)$ in section A.3.3.

An identical argument is used when θ_2 is fixed and θ_1 is approaching θ_2. Thus results of Chapter 2 apply for nearby alternatives in large samples, replacing σ^2 with $1/I_{\theta\theta}(1-\rho^2_{\theta\gamma})$. The results in Chapter 3 apply for nearby alternatives in

large samples as well, because the replacement of σ^2 by $1/I_{\theta\theta}(1-\rho^2_{\theta\gamma})$ does not affect the usage of Theorem two of Lorden [24].

4.5.3 The Fixed Sample Size Design Revisited

Here we derive the probability that some alternative is better supported over the true value θ_1, when profile likelihood functions are used under a Fixed Sample Size design with a large sample size.

Theorem 4.5 *(Maximum Probability Under a Fixed Sample Size Design) If $X_1, X_2, \ldots, X_n$ are i.i.d. $f(X_i;\theta,\gamma)$ where f is a smooth function and both θ and γ are real-valued, then*

$$\lim_{n\to\infty} P_1\left(\max_{\theta}\frac{L_{pn}(\theta)}{L_{pn}(\theta_1)} \geq k\right) = 2\Phi\left[-\sqrt{2\ln k}\right]$$

where

$$\max_{\gamma} L_n(\theta,\gamma) = L_n(\theta,\widehat{\gamma}(\theta)) = L_{pn}(\theta)$$

Proof:

Let $\hat{\theta}$ be the MLE for θ. Then we have

$$\max_{\theta}\frac{L_{pn}(\theta)}{L_{pn}(\theta_1)} = \frac{L_{pn}(\hat{\theta})}{L_{pn}(\theta_1)}$$

By Taylor series expansion of $\hat{\theta}$ around θ_1 in the log likelihood ratio we have

$$l_{pn}(\hat{\theta}) - l_{pn}(\theta_1) = (\hat{\theta}-\theta_1)A_p + \frac{(\hat{\theta}-\theta_1)^2}{2}B_p + R_p \tag{4.18}$$

where

$$A_p = \frac{\partial l_n(\theta,\gamma)}{\partial\theta}|_{(\theta_1,\widehat{\gamma}(\theta_1))}$$

$$B_p = \frac{\partial^2 l_n(\theta,\gamma)}{\partial\theta^2}|_{(\theta_1,\widehat{\gamma}(\theta_1))} + \frac{\partial^2 l_n(\theta,\gamma)}{\partial\theta\partial\gamma}|_{(\theta_1,\widehat{\gamma}(\theta_1))}\frac{d\widehat{\gamma}(\theta)}{d\theta}|_{\theta_1}$$

$$I_{\theta\theta} = E\left[\frac{\partial^2 l_n(\theta,\gamma)}{\partial\theta^2}|_{(\theta_1,\widehat{\gamma}(\theta_1))}\right]$$

$$\rho^2_{\theta\gamma} = \frac{I^2_{\theta\gamma}}{I_{\theta\theta}I_{\gamma\gamma}}$$

see section 4.5 for details. The remainder R_p is effectively a sum of n terms times a factor of $(\hat{\theta}-\theta_1)^3$. From the likelihood equation we have

$$0 = \frac{\partial l_n(\theta,\gamma)}{\partial\theta}|_{(\hat{\theta},\widehat{\gamma}(\hat{\theta}))} \approx A_p + (\hat{\theta}-\theta_1)B_p \tag{4.19}$$

from the note to Lemma 3, equation (4.17). Solving equation (4.19) for $(\hat{\theta}-\theta_1)$ and substituting into equation (4.18) yields

$$l_{pn}(\hat{\theta}) - l_{pn}(\theta_1) = -\frac{A_p^2}{2B_p} + R_p$$

But $(\hat{\theta}-\theta_1)^3 = O_p(n^{-3/2})$ so that $R_p = O_p(n^{-1/2})$ under mild conditions. Now both

$$\frac{1}{\sqrt{n}}A_p = A_{pn} \xrightarrow{d} N\left(0, I_{\theta\theta}(1-\rho^2_{\theta\gamma})\right)$$

$$\frac{1}{n}B_p = B_{pn} \xrightarrow{p} -I_{\theta\theta}(1-\rho^2_{\theta\gamma})$$

are proved in section 4.5 (see Lemmas 1, 2, and 3). Thus we have

$$\frac{|A_{pn}|}{\sqrt{-B_{pn}}} \xrightarrow{d} |Z| \quad \text{where } Z \sim N(0,1) \tag{4.20}$$

and the result follows.

QED •

4.6 Eliminating Nuisance Parameters: The Estimated Likelihood

In this section we again consider observations $X_1, X_2, \ldots, X_n$ as i.i.d. $f(X_i; \theta, \gamma)$ where f is a smooth function and both θ and γ are fixed dimensional parameters. Let $l_n(\theta, \gamma)$ be the log likelihood function. Suppose θ represents the parameter of interest, and γ the nuisance parameter. Another approach to eliminating the nuisance parameter is to use the estimated likelihood function for θ as if it were a true likelihood function. For example, the overall MLE might be used in place of the nuisance parameter as an estimated likelihood function. For fixed θ, define an estimated likelihood function to be

$$L_n(\theta, \hat{\gamma_n}) = L_{en}(\theta) \tag{4.21}$$

where $\hat{\gamma_n}$ is any consistent estimator of γ. We only consider the case where θ and γ are real-valued. The extension to the vector case is straightforward.

4.6.1 Log Estimated Likelihood Ratios

In large samples the log estimated likelihood ratio at nearby alternatives is approximately normally distributed. We notice that the log estimated likelihood ratio at nearby alternatives has a greater limiting variance than the 'true' likelihood ratio and thus does not behave like a 'true' likelihood ratio in large samples. This result has been proved elsewhere [31].

Theorem 4.6 *(Estimated Likelihood Case) For observations $X_1, X_2, \ldots, X_n$ i.i.d. $f(X_i; \theta_1, \gamma_1)$ where f is a smooth function, both θ_1 and γ_1 are real-valued, and $\theta_2 = \theta_1 + c/\sqrt{nI_{\theta\theta}}$ ($c > 0$ is a constant), the log estimated likelihood ratio*

for θ_2 over θ_1 is approximately normally distributed in large samples with mean $-c/2$ and variance $c^2/(1-\rho^2_{\theta\gamma})$.

$$\begin{aligned} l_{en}(\theta_2) - l_{en}(\theta_1) \quad &\approx \quad \frac{c}{\sqrt{I_{\theta\theta}}} A_{en} + \frac{c^2}{2I_{\theta\theta}} B_{en} + R_n \\ &\xrightarrow{d} \quad N\left(-\frac{c}{2}, \frac{c^2}{(1-\rho^2_{\theta\gamma})}\right) \text{ as } n \to \infty \end{aligned} \tag{4.22}$$

where

$$A_{en} = \frac{1}{\sqrt{n}} \frac{\partial l_n(\theta,\gamma)}{\partial \theta}|_{(\theta_1,\hat{\gamma})}$$

$$B_{en} = \frac{1}{n} \frac{\partial^2 l_n(\theta,\gamma)}{\partial \theta^2}|_{(\theta_1,\hat{\gamma})}$$

$$\rho^2_{\theta\gamma} = \frac{I^2_{\theta\gamma}}{I_{\theta\theta} I_{\gamma\gamma}}$$

Notice that the log profile likelihood ratio does not have the same limiting distribution as a 'true' likelihood ratio because its variance is inflated by the factor $1/(1-\rho^2_{\theta\gamma})$.

Proof:

Because θ_2 is in a $O(n^{-1/2})$ neighborhood of θ_1, direct Taylor expansion of the log likelihood ratio around θ_1 yields

$$l_{en}(\theta_2) - l_{en}(\theta_1) \approx \frac{c}{\sqrt{nI_{\theta\theta}}} \frac{\partial}{\partial\theta} l_n(\theta,\gamma)|_{(\theta_1,\hat{\gamma})} + \frac{c^2}{2nI_{\theta\theta}} \frac{\partial^2}{\partial\theta^2} l_n(\theta,\gamma)|_{(\theta_1,\hat{\gamma})} + R_n \tag{4.23}$$

The remainder is effectively a sum of n terms times a factor of $n^{-3/2}$. Thus, under mild conditions, $R_n = O_p(n^{-1/2})$.

Lemma 4 will show that $A_{en} \xrightarrow{d} N(0, I_{\theta\theta}/(1-\rho^2_{\theta\gamma}))$,and $B_{en} \xrightarrow{p} -I_{\theta\theta}$ is a standard result. Thus $l_{en}(\theta_2) - l_{en}(\theta_1) \xrightarrow{d} cZ^* - c^2/2$, where $Z^* \sim N(0, 1/(1-\rho^2_{\theta\gamma}))$ and the result follows.

Lemma 4. $\frac{1}{\sqrt{n}}\frac{\partial}{\partial\theta}l_n(\theta,\gamma)|_{(\theta_1,\hat{\gamma})} \xrightarrow{d} N(0, I_{\theta\theta}/(1-\rho^2_{\theta\gamma}))$

Proof: (Lemma 4)

By Taylor series expansion around γ_1

$$\frac{\partial l_n(\theta,\gamma)}{\partial\theta}|_{(\theta_1,\hat{\gamma})} \approx \frac{\partial l_n(\theta,\gamma)}{\partial\theta}|_{(\theta_1,\gamma_1)} + \frac{\partial l_n(\theta,\gamma)}{\partial\gamma\partial\theta}|_{(\theta_1,\gamma_1)}(\hat{\gamma}-\gamma_1)$$

From the likelihood equation and by Taylor series expansion,

$$\begin{aligned} 0 &= \frac{\partial l_n(\theta,\gamma)}{\partial\theta}|_{(\hat{\theta},\hat{\gamma})} \\ &\approx \frac{\partial l_n(\theta,\gamma)}{\partial\theta}|_{(\theta_1,\gamma_1)} + \frac{\partial^2 l_n(\theta,\gamma)}{\partial\theta^2}|_{(\theta_1,\gamma_1)}(\hat{\theta}-\theta_1) + \frac{\partial^2 l_n(\theta,\gamma)}{\partial\gamma\partial\theta}|_{(\theta_1,\gamma_1)}(\hat{\gamma}-\gamma_1) \end{aligned}$$

and

$$\begin{aligned} 0 &= \frac{\partial l_n(\theta,\gamma)}{\partial\gamma}|_{(\hat{\theta},\hat{\gamma})} \\ &\approx \frac{\partial l_n(\theta,\gamma)}{\partial\gamma}|_{(\theta_1,\gamma_1)} + \frac{\partial^2 l_n(\theta,\gamma)}{\partial\theta\partial\gamma}|_{(\theta_1,\gamma_1)}(\hat{\theta}-\theta_1) + \frac{\partial^2 l_n(\theta,\gamma)}{\partial\gamma^2}|_{(\theta_1,\gamma_1)}(\hat{\gamma}-\gamma_1) \end{aligned}$$

Together the likelihood equations imply that $(\hat{\gamma}-\gamma_1)$ is asymptotically equivalent to

$$\frac{\frac{\partial l_n(\theta,\gamma)}{\partial\gamma}|_{(\theta_1,\gamma_1)}\frac{\partial^2 l_n(\theta,\gamma)}{\partial\theta^2}|_{(\theta_1,\gamma_1)} - \frac{\partial l_n(\theta,\gamma)}{\partial\theta}|_{(\theta_1,\gamma_1)}\frac{\partial^2 l_n(\theta,\gamma)}{\partial\theta\partial\gamma}|_{(\theta_1,\gamma_1)}}{\left(\frac{\partial^2 l_n(\theta,\gamma)}{\partial\theta\partial\gamma}|_{(\theta_1,\gamma_1)}\right)^2 - \frac{\partial^2 l_n(\theta,\gamma)}{\partial\theta^2}|_{(\theta_1,\gamma_1)}\frac{\partial^2 l_n(\theta,\gamma)}{\partial\gamma^2}|_{(\theta_1,\gamma_1)}}$$

Replacing $(\hat{\gamma}-\gamma_1)$ in the first expression with the above ratio, the conclusion follows from the asymptotic bivariate normality of the two score functions.

QED •

4.6.2 Approximating The Discrete Stopping Time

Recall the stopping time N examined in Chapter 2, which equaled the smallest sample size where strong evidence for H_2 over H_1 is obtained. Here consider

the case of fixed simple hypothesis $H_1 : \theta = \theta_1$ and $H_2 : \theta = \theta_2$ where θ_2 is in a $O(n^{-1/2})$ neighborhood of θ_1. Using Taylor expansion methods, the stopping time N can be re-expressed as

$$\begin{aligned} N &= \inf\left\{n : n \geq 1, \frac{L_{en}(\theta_2)}{L_{en}(\theta_1)} \geq k\right\} \\ &= \inf\left\{n : n \geq 1, \frac{\partial l_n(\theta,\gamma)}{\partial\theta}|_{(\theta_1,\hat{\gamma})} \geq \frac{\ln k}{\Delta} + n\frac{-B_{en}\Delta}{2} - O_p(n^{-1/2})\right\} \\ &= \inf\left\{n : n \geq 1, \frac{S_n^*}{\sqrt{I_{\theta\theta}/(1-\rho_{\theta\gamma}^2)}} \geq \frac{\ln k}{\Delta\sqrt{I_{\theta\theta}/(1-\rho_{\theta\gamma}^2)}} + n\frac{-B_{en}\Delta}{2\sqrt{I_{\theta\theta}/(1-\rho_{\theta\gamma}^2)}} - O_p(n^{-1/2})\right\} \end{aligned}$$

where $\Delta = |\theta_2 - \theta_1|$, and

$$\frac{\partial l_n(\theta,\gamma)}{\partial\theta}|_{(\theta_1,\hat{\gamma})} = S_n^*$$

is the estimated score function at θ_1. Under θ_1, the stopping time N represents the first time misleading evidence for θ_2 is obtained.

Arguments similar to those used to prove Theorem 4.6 show that, under θ_1, the estimated score function evaluated at θ_1 behaves asymptotically like a random walk with mean zero and variance $nI_{\theta\theta}/(1-\rho_{\theta\gamma}^2)$, where $I_{\theta\theta}$ and $(1-\rho_{\theta\gamma}^2)$ are evaluated as specified by Theorem 4.6. Likewise Taylor expansion arguments demonstrate that under θ_1, $B_{en} \xrightarrow{p} I_{\theta\theta}$ as $n \to \infty$.

By the preceding remarks, the behavior of S_n^* can be asymptotically approximated with a Brownian motion process $W(nI_{\theta\theta}/(1-\rho_{\theta\gamma}^2))$ with drift zero and variance per unit time $I_{\theta\theta}/(1-\rho_{\theta\gamma}^2)$. In large samples the joint distributions of $S_n^*, n = 1, 2, \ldots$ and $W(nI_{\theta\theta}/(1-\rho_{\theta\gamma}^2)), n = 1, 2, \ldots$ are approximately the same. It can be verified that $W(nI_{\theta\theta}/(1-\rho_{\theta\gamma}^2))/\sqrt{I_{\theta\theta}/(1-\rho_{\theta\gamma}^2)}$ is Brownian motion with drift zero and variance per unit time one. For continuous $t > 0$, define the

stopping time τ as

$$\tau^* \quad = \quad \inf \left\{ t : t > 0, W(t) \geq \frac{\ln k}{\Delta} \frac{\sqrt{(1-\rho^2_{\theta\gamma})}}{\sqrt{I_{\theta\theta}}} + t\frac{\Delta}{2} \frac{\sqrt{I_{\theta\theta}}}{\sqrt{(1-\rho^2_{\theta\gamma})}} (1-\rho^2_{\theta\gamma}) \right\}$$

where $\Delta = |\theta_2 - \theta_1|$ and $W(t)$ is Brownian motion with drift zero and variance per unit time one. Under θ_1, τ^* is the limiting continuous time version of N. The Brownian motion approximations $P_{\theta_1}(N \leq m) \approx P_{\theta_1}(\tau^* \leq m)$ for some fixed m will hold in large samples for nearby alternatives as described in the appendix, as do the adjusted Brownian motion probabilities $P_{\theta_1}(N \leq m) \cong P_{\theta_1}(\widetilde{\tau}^* \leq m)$ in section A.3.3.

However, τ^* does not have the form of the stopping times considered in Chapters 2 and 3, because the slope of the boundary is reduced by $(1-\rho^2_{\theta\gamma})$. Thus, with stopping time τ^* it is actually easier to cross the boundary. Now the results of Chapters 2 and 3 are lower bounds on the corresponding probabilities of generating certain types of evidence, for nearby alternatives, in large samples, when estimated likelihood ratios are used in place of 'true' likelihood ratios. Remember that σ^2 is replaced with $(1-\rho^2_{\theta\gamma})/I_{\theta\theta}$ in this case.

An identical argument is used when θ_2 is fixed and θ_1 is approaching θ_2. Thus results of Chapter 2 will provide a lower bound on the probability for nearby alternatives in large samples, replacing σ^2 with $(1-\rho^2_{\theta\gamma})/I_{\theta\theta}$. The results in Chapter 3 will also provide a lower bound on the probability for nearby alternatives in large samples, because the replacement of σ^2 by $(1-\rho^2_{\theta\gamma})/I_{\theta\theta}$ does not effect the usage of Theorem two of Lorden [24].

4.6.3 The Fixed Sample Size Design Revisited

Here we derive the probability that some simple alternative is better supported over the true value θ_1 when estimated likelihood functions are used under a Fixed Sample Size design with a large sample. We note that the limiting probability is greater than the corresponding profile likelihood result.

Theorem 4.7 *(Maximum Probability Under a Fixed Sample Size Design) If $X_1, X_2, \ldots, X_n$ are i.i.d. $f(X_i; \theta, \gamma)$ where f is a smooth function and both θ and γ are real-valued, then*

$$\lim_{n\to\infty} P_1\left(\max_{\theta} \frac{L_{en}(\theta)}{L_{en}(\theta_1)} \geq k\right) = 2\Phi\left[-\sqrt{1-\rho^2_{\theta\gamma}}\sqrt{2\ln k}\right]$$

where $\hat{\gamma_n}$ is any consistent estimator of γ_1 and $L_n(\theta, \hat{\gamma_n}) = L_{en}(\theta)$

Proof:

Let $\hat{\theta}$ be the MLE for θ. Then we have

$$\max_{\theta} \frac{L_{en}(\theta)}{L_{en}(\theta_1)} = \frac{L_{en}(\hat{\theta})}{L_{en}(\theta_1)}$$

By Taylor series expansion of $\hat{\theta}$ around θ_1 in the log likelihood ratio we have

$$l_{en}(\hat{\theta}) - l_{en}(\theta_1) = (\hat{\theta} - \theta_1)A_e + \frac{(\hat{\theta} - \theta_1)^2}{2}B_e + R_e \tag{4.24}$$

where

$$A_e = \frac{\partial l_n(\theta, \gamma)}{\partial \theta}|_{(\theta_1, \hat{\gamma})}$$

$$B_e = \frac{\partial^2 l_n(\theta, \gamma)}{\partial \theta^2}|_{(\theta_1, \hat{\gamma})}$$

$$\rho^2_{\theta\gamma} = \frac{I^2_{\theta\gamma}}{I_{\theta\theta} I_{\gamma\gamma}}$$

see section 4.6 for details. The remainder R_e is effectively a sum of n terms times a factor of $(\hat{\theta} - \theta_1)^3$. From the likelihood equation we have

$$0 = \frac{\partial l_n(\theta, \gamma)}{\partial \theta}|_{(\hat{\theta}, \hat{\gamma})} \approx A_e + (\hat{\theta} - \theta_1) B_e \tag{4.25}$$

Solving equation (4.25) for $(\hat{\theta} - \theta_1)$ and substituting into equation (4.24) yields

$$l_{en}(\hat{\theta}) - l_{en}(\theta_1) = -\frac{A_e^2}{2B_e} + R_e$$

But $(\hat{\theta} - \theta_1)^3 = O_p(n^{-3/2})$ so that $R_e = O_p(n^{-1/2})$ under mild conditions. Now

$$\frac{1}{\sqrt{n}} A_e = A_{en} \xrightarrow{d} N\left(0, I_{\theta\theta}/(1 - \rho_{\theta\gamma}^2)\right)$$

$$\frac{1}{n} B_e = B_{en} \xrightarrow{p} -I_{\theta\theta}$$

where the first result is proven in section 4.6 (see Lemma 4) and the second is a standard result. Thus we have

$$\frac{|A_{en}|}{\sqrt{-B_{en}}} \xrightarrow{d} |Z^*| \quad \text{where } Z^* \sim N\left(0, 1/(1 - \rho_{\theta\gamma}^2)\right) \tag{4.26}$$

and the result follows.

QED •

4.7 Summary

The curves in Chapters 2 and three which represent the probability that an experimental design generates misleading or strong evidence about the mean of a normal distribution extend to non-skewed distributions in a one-parameter exponential family almost directly. For distributions in a one-parameter exponential family with third moment not equal to zero, only a simple adjustment to the results in Chapters 2 and 3 is required to account for the non-normality. In

general, for nearby alternatives, the Tepee function represents the probability that a Biased Open design generates misleading evidence about the mean of an exponential family distribution.

For distributions not in the exponential family model, a local analysis reveals that the curves representing the probability that an experimental design generates evidence of a certain type about the mean of a normal distribution apply in large samples. One needs only replace σ^2 by the appropriate information quantity. In situations where fixed-dimensional nuisance parameters are present, these results apply in large samples for nearby alternatives when profile likelihood ratios are used to measure the strength of statistical evidence about the parameter of interest, but do not apply when estimated liklelihood ratios are used.

When using a Fixed Sample Size design with a large sample size, the maximum probability that some simple alternative is better supported over the true value is approximately $2\Phi[-\sqrt{2\ln k}]$. In situations where nuisance parameters are present, the maximum probability that some simple alternative is better supported over the true value is also approximately $2\Phi[-\sqrt{2\ln k}]$ when Profile likelihood ratios are used, but greater when Estimated likelihood ratios are used.

Appendix A

Brownian Motion

A.1 Introduction

A particle suspended in a liquid or gas exhibits a ceaseless irregular random movement. This motion was first observed in 1827 by the English botanist Robert Brown [25]. In 1905, Einstein showed that Brownian motion was due to the continual bombardment or impacts of surrounding molecules (For a history of the theory of Brownian Motion see [42],[15]). N. Wiener was among the first to consider a mathematical process to describe Brownian motion [25, p28]. The Wiener process, also called a Wiener-Lévy process or Brownian motion process, originated as a model for Brownian motion and for price movements in stock and commodity markets [25].

Brownian motion is the natural limiting process for sums of independent random variables. Considerable simplification comes from approximating sums of independent random variables $X_1 + \cdots + X_n, n = 1, 2, \ldots$ in discrete time by a Brownian motion process $\{W(t), 0 \leq t < \infty\}$ in continuous time. For example, a series of log likelihood ratios is a random walk, and Brownian motion can be used as its asymptotic approximation (see Chapter 4). A Brownian motion

process can be thought of as a continuous time interpolation of the discrete time random walk and even though this approximation is often very crude, it can be used as a starting point for further improvement.

This appendix reviews some basic Brownian motion concepts required for calculating probabilities about a stopping time. As in [35], [36], [37], [38], [44] we derive $P(\tau \leq m)$ where τ is a Brownian motion stopping time and m is some fixed number, for the purposes of approximating $P(N \leq m)$ where N is the stopping time from Chapter 2, equation (2.5). We then seek to adjust $P(\tau \leq m)$ to attain better numerical accuracy in the discrete time case. The results presented here are well known. Both presentation and examples of this material are borrowed from [37], where Brownian motion processes are used to determine the significance levels of repeated significance tests.

A.2 What is Brownian Motion?

A Brownian Motion Process with drift μ and variance per unit time σ^2 is a family of random variables $\{W(t), 0 \leq t < \infty\}$ with the following properties (t is a continuous index often thought of as time):

1. $W(0) = 0$;

2. $P_{\mu,\sigma}\{W(t) - W(s) \leq x\} = \Phi[(x - \mu(t-s))/\sigma(t-s)^{\frac{1}{2}}], 0 \leq s < t < \infty$;

3. for all $0 \leq s_1 < t_1 \leq s_2 < t_2 \leq \ldots \leq s_n < t_n < \infty, n = 2, 3, \ldots$ the random variables $W(t_i) - W(s_i), i = 1, 2, \ldots, n$ are stochastically independent;

4. $W(t), 0 \leq t < \infty$, is a continuous function of t.

Properties 2 and 3 above are equivalent to the conditions that (1) the joint distribution of $W(t_1), \ldots, W(t_n)$ is Gaussian, and (2) $E_\mu[W(t)] = \mu t$, and (3) $Cov(W(t_i), W(t_j)) = \sigma^2 \min(t_i, t_j)$ for all n and $t_1, \ldots, t_n$. Thus at any fixed time t, the distribution of $W(t)$ is $N(t\mu, t\sigma^2)$.

To motivate why a Brownian motion process has these properties, we consider the case where $X_1, X_2, \ldots$ are independent and identically distributed random variables with mean zero and variance one. Represent the sum of the observations by $S_n = X_1 + \cdots + X_n$, where n is the total number of observations ($n = 0, 1, 2, \ldots$). The central limit theorem implies that $S_m/\sqrt{m}$ converges in law to a normally distributed random variable with mean zero and variance one. In fact, there is a more general result for fixed but possibly continuous t, $S_{[mt]}/\sqrt{m}$ converges in law to a normally distributed random variable with mean zero and variance t, where $[x]$ denotes the largest integer $\leq x$.

Let $W(t) = \lim_{m\to\infty} m^{-\frac{1}{2}} S_{[mt]}$, and consider whether $W(t)$ converges in law uniformly for all t ($0 \leq t < \infty$). Under mild conditions this is the case and a complete discussion of this point can be found in [7]. For our purposes it is enough to assume the limit exists. Therefore, by the above remark, $W(t)$ is normally distributed with mean zero and variance t.

Suppose we have fixed time points, for each $k = 1, 2, \ldots$ such that $0 \leq s_1 < t_1 \leq s_2 < t_2 \leq \ldots \leq s_k < t_k$. The random process will also converge at any fixed finite number of time points, so the random variables

$$(S_{[mt_j]} - S_{[ms_j]})/\sqrt{m}, \quad j = 1, 2, \ldots, k$$

must converge to $W(t_j) - W(s_j)$, $j = 1, 2, \ldots, k$, as independent normally distributed random variables with mean zero and variance $t_j - s_j$. Furthermore

the covariance of $W(t)$ and $W(s)$ for $s < t$ is given by

$$\begin{aligned} E[W(t)W(s)] &= \lim_{m\to\infty} m^{-1}E[S_{[mt]}S_{[ms]}] \\ &= \lim_{m\to\infty} m^{-1}E[S^2_{[ms]} + S_{[ms]}(S_{[mt]} - S_{[ms]})] \\ &= \lim_{m\to\infty} m^{-1}[ms] \\ &= s \\ &= \min(s,t) \end{aligned}$$

Now the process $\{W(t), 0 \le t < \infty\}$, is a Brownian motion process with the properties outlined at the beginning of this section. Note that $W(t)$ has not been shown to be a continuous function of t. This property is assumed to hold.

In the case where $X_1, X_2, \ldots$ are i.i.d. $N(\mu, \sigma^2)$ and $\{W(t), 0 \le t < \infty\}$ is a Brownian motion process with drift μ and variance per unit time σ^2, S_n and $W(n), n = 0, 1, \ldots$ have the same joint distribution, providing another way of determining the joint distribution of S_n. Seigmund states "Brownian motion is an interpolation of the discrete time random walk S_n, $n = 0, 1, \ldots$ which preserves the Gaussian distribution and which makes the paths of the process continuous in time. To the extent that many random walks are approximately normally distributed for large n, the Brownian motion process may be used as an asymptotic approximation to a large class of random walks and hence of log likelihood ratios" [37, p35].

As a final example, consider the likelihood ratio supporting $H_1 : \mu = \mu_1 > \mu_0$ over $H_0 : \mu = \mu_0$, in both the discrete and continuous time case. The familiar

likelihood ratio in discrete time is

$$L(n, S_n; \mu_1, \mu_0) = \exp\left\{\frac{(\mu_1 - \mu_0)}{\sigma^2} S_n - \frac{n}{2\sigma^2}(\mu_1^2 - \mu_0^2)\right\}$$

In terms of Brownian motion, the likelihood ratio for H_1 versus H_0 is

$$L(t, W(t); \mu_1, \mu_0) = \exp\left\{\frac{(\mu_1 - \mu_0)}{\sigma^2} W(t) - \frac{t}{2\sigma^2}(\mu_1^2 - \mu_0^2)\right\}$$

Notice that the likelihood ratio for continuous time (Brownian motion case) is similar to the likelihood ratio in discrete time (Normal case) with $W(t)$ and t replacing S_n and n.

A.3 Stopping Times

This section is divided into four parts. The first part introduces stopping times based on a likelihood ratio between two pre-specified simple hypotheses. These stopping times are shown to be the time at which some random process crosses a linear boundary. We observe that, under certain conditions, Brownian motion provides a 'good' approximation to the discrete time process, but unfortunately these conditions are seldom met. In the second part, boundary crossing probabilities associated with a Brownian motion stopping time are derived. Siegmund's modification of the Brownian approximation, which gives better numerical accuracy in discrete time, is presented in the third part. The basic idea is to adjust the Brownian motion process by accounting for the expected overshoot of the boundary in discrete time. The fourth part discusses the implications for generating misleading evidence under an Open design.

A.3.1 Part I: Introduction

Let $\{W(t), 0 \leq t < \infty\}$ be a Brownian motion process with drift μ. Without loss of generality take the variance per unit time to be one. Now a Brownian motion stopping time τ that equals the first time t when the likelihood ratio for $H_1 : \mu = \mu_1 > \mu_0$ versus $H_0 : \mu = \mu_0$ is equal to k is defined as

$$\begin{aligned} \tau &= \inf\{t : L(t, W(t); \mu_1, \mu_0) = k\} \\ &= \inf\{t : W(t) - \frac{1}{2}t(\mu_1 + \mu_0) = b\} \\ &= \inf\{t : W(t) = b + \eta t\} \end{aligned} \tag{A.1}$$

where $b = \ln k/(\mu_1 - \mu_0)$ and $\eta = (\mu_1 + \mu_0)/2$. If $\mu = \mu_0$ is the true mean, then stopping time τ represents the first time strong misleading evidence for μ_1 over μ_0 is obtained. As shown in equation (A.1), τ can also be interpreted as the first time the random process $W(t)$ touches the linear boundary $b + \eta t$. Note that $W(t)$ cannot overshoot the boundary because the process is continuous.

Now consider i.i.d. normal random variables $X_1, X_2, \ldots$ with mean μ. Without loss of generality take the variance to be one. The discrete time analogue to equation (A.1) is the stopping time $\tilde{\tau}$ that equals the first time when the likelihood ratio for H_1 versus H_0 is greater than or equal to k and is defined as

$$\begin{aligned} \tilde{\tau} &= \inf\{n : L(n, S_N; \mu_1, \mu_0) \geq k\} \\ &= \inf\{n : S_n - \frac{1}{2}n(\mu_1 + \mu_0) \geq b\} \\ &= \inf\{n : S_n \geq b + \eta n\} \end{aligned} \tag{A.2}$$

where $b = \ln k/(\mu_1 - \mu_0)$ and $\eta = (\mu_1 + \mu_0)/2$. If $\mu = \mu_0$ is the true mean, then

stopping time $\tilde{\tau}$ represents the smallest sample size where strong misleading evidence for μ_1 is obtained. Note that $\hat{\tau}$ is equal to N (equation (2.5)) from Chapter 2. Here $\tilde{\tau}$ represents the first time the random walk S_n touches or crosses the linear boundary $b + \eta n$.

Both τ and $\tilde{\tau}$ indicate the time at which their respective processes cross a linear boundary with intercept b and slope η. This similarity suggests the natural approximation $P(\tilde{\tau} \leq n) \approx P(\tau \leq n)$. We now argue that, under some limiting conditions, this approximation holds.

Consider the case when $\mu = 0$. Define a discrete stopping time as $\tilde{\tau}(c) = \inf\{n : S_n \geq c\}$ and an analogous continuous stopping time as $\tau(c) = \inf\{t : W(t) \geq c\}$ where $c = \sqrt{m}\xi$ for some $\xi > 0$. From the assumed convergence of $m^{-\frac{1}{2}} S_{[mt]}$ to $W(t)$ (see section A.2), it follows as $m \to \infty$ that

$$\begin{aligned} P\{\tilde{\tau}(c) \leq n\} &= P\left\{\max_{1 \leq k \leq n} S_k \geq c\right\} \\ &= P\left\{\max_{0 \leq s \leq t} m^{-1/2} S_{[ms]} \geq \xi\right\} \to P\left\{\sup_{0 \leq s \leq t} W(s) \geq \xi\right\} \\ &= P\{\tau(\xi) \leq t\} \end{aligned}$$

where $n = [mt]$. Furthermore, it can be shown if $\{W(t), 0 \leq t < \infty\}$, is a Brownian Motion process with drift zero, then for any m the process $\{W(mt)/\sqrt{m}, 0 \leq t < \infty\}$, is also a Brownian Motion process with drift zero. Using this fact we have

$$\begin{aligned} P\{\tau(\xi) \leq t\} &= P\left\{\sup_{0 \leq s \leq t} m^{-1/2} W(ms) \geq \xi\right\} \\ &= P\left\{\sup_{0 \leq s \leq mt} W(s) \geq m^{1/2}\xi\right\} \end{aligned}$$

$$= \quad P\{\tau(m^{1/2}\xi) \leq mt\}$$

Provided both n and c are large and c is proportional to $n^{1/2}$, the previous two results suggest the natural approximation

$$P\{\tilde{\tau}(c) \leq n\} \approx P\{\tau(c) \leq n\} \tag{A.3}$$

Thus under these conditions, we can approximate the distribution of $\tilde{\tau}$ with that of τ [37, p242].

If $E(X_i) = \mu \neq 0$ we need to impose further conditions to ensure convergence in distrubtion of $m^{-1/2}S_{[mt]}$. Because $E(S_m) = m\mu$, it is necessary that μ be proportional to $m^{-1/2}$ as $m \to \infty$ to ensure the convergence in law of $m^{-\frac{1}{2}}S_{[mt]}$. See [37, p242-243] for an outline of a similar argument showing equation (A.3) holds in this case. We note that a complete justification requires in addition to c becoming infinitely large at rate $m^{1/2}$ that μ tend to 0 at rate $m^{-1/2}$, so that μc tends to a finite stable limit.

These requirements are often hard to meet, suggesting Brownian motion approximations, by themselves, are often poor. Siegmund states:

> "...Brownian approximations are frequently not especially accurate. The most obvious reason is that the test statistics under consideration may not be approximately normally distributed, but even for exactly normal data, the discreteness of the time scale can have a substantial effect. (In some cases these two factors cancel and make the Brownian approximation look very good, but this should be regarded as a fortunate accident and not something to be expected)." [37, p49]

As we will see in section A.3.3, one can improve the accuracy of Brownian motion approximations by adjusting for the expected overshoot of the boundary in discrete time. This approximation produces better numerical accuracy and is not restricted by the limiting assumptions discussed in this section. Next we derive the boundary crossing probabilities associated with the Brownian motion stopping time τ.

A.3.2 Part II: Boundary Crossing Probabilities

In this section we derive $P_\mu(\tau \leq m)$ for the purposes of approximating $P_\mu(\widetilde{\tau} \leq m)$, where m is some fixed number, $\widetilde{\tau} = \inf\{n : S_n \geq b + \eta n\}$ and $\tau = \inf\{t : t \geq 0, W(t) \geq b + \eta t\}$. Note that $P_\mu(\tau \leq m)$ is the probability that $W(t)$ touches the linear boundary $b + \eta t$. Only a sketch of the necessary steps are provided here, see [36], [37] for complete details.

A central quantity in the derivation of $P_\mu(\tau \leq m)$ is the joint probability

$$P_\mu(\tau < m, W(m) \leq c) = \exp\{2b(\mu - \eta)\}\Phi[(c - 2b)/\sqrt{m} - \mu\sqrt{m}] \qquad \text{(A.4)}$$

where $c = b + \eta m$ and $W(t)$ is a Brownian motion process with drift μ and, without loss of generality, variance per unit time one, [37, eqn (3.14)]. Next we have

$$\begin{aligned} P_\mu(\tau \leq m) &= P_\mu(W(m) \geq c) + P_\mu(\tau < m, W(m) \leq c) \\ &= \Phi[-b/\sqrt{m} + (\mu - \eta)\sqrt{m}] \\ &\quad + \exp\{2b(\mu - \eta)\}\Phi[-b/\sqrt{m} - (\mu - \eta)\sqrt{m}] \qquad \text{(A.5)} \end{aligned}$$

Equation (A.5) is the probability of generating strong evidence for $H_1 : \mu = \mu_1$ over $H_0 : \mu = \mu_0$ on or before the m^{th} observation, when μ is the true mean. If

$\mu = \mu_0$, the evidence is misleading and equation (A.5) reduces to

$$P_0\{\tau \leq m\} = \Phi\left[-\frac{\ln k}{\Delta\sqrt{m}} - \frac{\Delta\sqrt{m}}{2}\right] + \frac{1}{k}\Phi\left[-\frac{\ln k}{\Delta\sqrt{m}} + \frac{\Delta\sqrt{m}}{2}\right] \tag{A.6}$$

where $\Delta = |\mu_1 - \mu_0|$. Likewise, if $\mu = \mu_1$ is the true mean, then equation (A.5) is the probability of observing strong evidence for H_1 over H_0 on or before the m^{th} observation. In this case equation (A.5) reduces to

$$P_1\{\tau \leq m\} = \Phi\left[-\frac{\ln k}{\Delta\sqrt{m}} + \frac{\Delta\sqrt{m}}{2}\right] + k\Phi\left[-\frac{\ln k}{\Delta\sqrt{m}} - \frac{\Delta\sqrt{m}}{2}\right] \tag{A.7}$$

where $\Delta = |\mu_1 - \mu_0|$.

A.3.3 Part III: Adjusting for Discrete Time

In section A.3.1 we noted that some asymptotic conditions were required for the Brownian motion process to provide a 'good' approximation to the discrete time case (i.e. for equation (A.3) to hold). Because these conditions are not often met, Brownian motion approximations, by themselves, are not often very accurate. Thus it is not always best to simply substitute $P_\mu(\tau \leq m)$ for $P_\mu(\tilde{\tau} \leq m)$. Here we present Siegmund's modified Brownian approximation to $P_\mu(\tilde{\tau} \leq m)$, for the purposes of giving better numerical accuracy in discrete time. The plan of attack is to derive $P_\mu(\tilde{\tau} \leq m)$ using certain Brownian motion approximations while accounting for the expected overshoot of the boundary in discrete time. Only a sketch of the necessary steps are provided here, see [37, §10.3], [35], [38] for complete details.

A central quantity in the derivation of $P_\mu(\tilde{\tau} \leq m)$ is the joint probability

$$P_\mu(\tilde{\tau} < m, S_m \leq c) \cong \exp\{2(b+\rho)(\mu-\eta)\}\Phi[(c-2(b+\rho))/\sqrt{m} - \mu\sqrt{m}] \tag{A.8}$$

where $c = b + \eta m$ and ρ is $E_\mu[S_{\tilde{\tau}} - (b + \eta\tilde{\tau})]$ the expected overshoot of the boundary by S_n at time $\tilde{\tau}$. For the normal case ρ is numerically evaluated (using numerical integration techniques) to three decimal places as $\rho \cong 0.583$ [37, p50], [38, p603].

Notice that equation (A.8) is simply equation (A.4) with b replaced by $b+\rho$. This suggests replacing b by $b+\rho$ in equation (A.5) to obtain a better approximation for $P_\mu(\tilde{\tau} \leq m)$. Siegmund notes, "Hence these corrected approximations are tantamount to using the exact formulae for Brownian motion applied to a boundary $b' + \eta m$, where $b' - b[= \rho]$ is the expected excess of the discrete process over the boundary"[37, p50]. The implication is that adjusting a Brownian motion process for discrete time amounts to shifting the boundary by the expected overshoot of S_n over the boundary.

A slightly more accurate approximation to $P_\mu(\tilde{\tau} \leq m)$ can be found by a derivation similar to that which was used to obtain equation (A.5) [37, p 50]. Now we have

$$\begin{aligned} P_\mu(\tilde{\tau} \leq m) &= P_\mu(S_m \geq c) + P_\mu(\tilde{\tau} < m, S_m \leq c) \\ &\cong \Phi[-b/\sqrt{m} + (\mu - \eta)\sqrt{m}] \qquad (A.9) \\ &\quad + \exp\{2(b+\rho)(\mu-\eta)\}\Phi[-(b+2\rho)/\sqrt{m} - (\mu-\eta)\sqrt{m}] \end{aligned}$$

This equation does not have the interpretation of equation (A.5) with b replaced by $b + \rho$. However, the difference between the two expressions is $o(m^{-1/2})$ [37, p50], [35, p709].

Equation (A.9) is the probability of observing strong evidence for $H_1 : \mu = \mu_1$ over $H_0 : \mu = \mu_0$ on or before the m^{th} observation. If $\mu = \mu_0$ the evidence is

misleading and equation (A.9) reduces to

$$P_0\{\tilde{\tau} \leq m\} \cong \Phi\left[-\frac{\ln k}{\Delta\sqrt{m}} - \frac{\Delta\sqrt{m}}{2}\right] + \frac{\exp\{-\rho\Delta\}}{k}\Phi\left[-\left(\frac{\ln k}{\Delta} + 2\rho\right)\frac{1}{\sqrt{m}} + \frac{\Delta\sqrt{m}}{2}\right] \tag{A.10}$$

where $\Delta = |\mu_1 - \mu_0|$. Likewise, if $H_1 : \mu = \mu_1$ is the correct hypothesis, then equation (A.9) is the probability of observing strong evidence for H_1 over H_0 on or before the m^{th} observation. In this case equation (A.9) reduces to

$$P_{\mu_1}\{\tilde{\tau} \leq m\} \cong \Phi\left[-\frac{\ln k}{\Delta\sqrt{m}} + \frac{\Delta\sqrt{m}}{2}\right] + k\exp\{\rho\Delta\}\Phi\left[-\left(\frac{\ln k}{\Delta} + 2\rho\right)\frac{1}{\sqrt{m}} - \frac{\Delta\sqrt{m}}{2}\right] \tag{A.11}$$

where $\Delta = |\mu_1 - \mu_0|$.

The improved accuracy of equation (A.9) over (A.5) is demonstrated in Tables 3.5 and 10.1 of [37, p50]. In every instance the modified approximation (A.9) provides an exceptionally better approximation than (A.5) to the exact value which was numerically computed in [32] and [33].

A.3.4 Part IV: Implications for Misleading Evidence

If $\mu = \mu_0$ is the true mean, then both stopping times τ and $\tilde{\tau}$ represent the first time misleading evidence for μ_1 over μ_0 is observed. If we let m go to infinity in equations (A.6) and (A.10) we get the probability of eventually generating misleading evidence for H_1 over H_0. That is, we only stop sampling when we have strong evidence for H_1 over H_0 (this is an Open design). The probability that we will stop sampling at some time (instead of sampling infinitely) is

$$\lim_{m\to\infty} P_0(\tau \leq m) = 1/k \tag{A.12}$$

by taking the limit of equation (A.6). Likewise in the discrete time case, the limit of equation (A.10) is

$$\lim_{m\to\infty} P_0(\tilde{\tau} \leq m) \cong \exp\{-\rho\Delta\}/k \tag{A.13}$$

where $\Delta = |\mu_1 - \mu_0|$. Equation (A.13) is called the Tepee function and is discussed in detail in Chapter 2, section 2.3.

Note that equation (A.12) is the universal bound, independent of either μ_1 or μ_0. By contrast the Tepee function depends critically on Δ. Let the difference between these probabilities be $D = 1/k - \exp\{-\rho\Delta\}/k$. Now D is the probability that $W(t)$ crosses the boundary only to fall back below before the very next unit of discrete time elapses. For large Δ, $D \approx 1/k$ implies that the Brownian motion process may cross the boundary between discrete time points, but fails to stay above at any single discrete time point. In this case the slope of the boundary η is very steep, so that while $W(t)$ can cross the boundary very early it must continue to increase with at least slope η to remain above the boundary. But under μ_0, it is very improbable that $W(t)$ will maintain such a steep slope for any length of time.

The Tepee function has been derived elsewhere using Wald's Likelihood ratio identity (see [41]), by Wijsman [44, p82] and Siegmund [35, p709], [37, §10.1]. Wijsman and Siegmund also argue that the Tepee function holds when the X's are non-normal, however a new ρ must be computed (see Chapter 4).

A.4 Exponential Family Example

The approximation (A.3) is vaild in another sense. Suppose we have $\xi_1 < 0 < \xi_2$, then the probability that S_k, $k = 0, 1, 2, \ldots$ crosses $m^{1/2}\xi_1$ before $m^{1/2}\xi_2$ and

before time mt is approximately the probability that $W(s)$, $0 \leq s < \infty$, crosses ξ_1 before ξ_2 and before time t. This example is also due to Siegmund [37, appendix one].

Suppose we observe $X_1, X_2, \ldots$ independent with a single parameter exponential family probability density function $f_\theta(X) = \exp[\theta X - \psi(\theta)]f(X)$. Without loss of generality we can assume $\psi(0) = 0$, because a simple transformation on the x's will give us this case. Furthermore general exponential family theory demonstrates $E_\theta[X_i] = \psi'(\theta)$ and $Var_\theta(X_i) = \psi''(\theta)$, where the prime denotes differentiation with respect to θ.

Consider the stopping time based on the likelihood ratio between $H_1 : \theta = \theta_1$ and $H_0 : \theta = 0$ of the following form

$$N = \min\left\{n : n \geq 1, S_n - \frac{n\psi(\theta_1)}{\theta_1} \notin (\frac{\xi_1}{\theta_1}, \frac{\xi_2}{\theta_1})\right\} \tag{A.14}$$

where $S_n = x_1 + ... + x_n$, and $\xi_1 < 0 < \xi_2$. Thus the stopping time N is equal to the smallest sample size when the likelihood ratio for H_1 over H_0 is either greater than $\exp(\xi_2)$ or less than $\exp(\xi_1)$. Now consider the expected increment of the stopping time and we have by Taylor expansion (remember $\psi(0) = 0$)

$$\begin{aligned} E_0[X_1 - \frac{\psi(\theta_1)}{\theta_1}] &= \psi'(0) - \frac{[\psi(\theta_1) - \psi(0)]}{\theta_1} \\ &= -\frac{1}{2}\theta_1\psi''(0) + O(\theta_1^2) \end{aligned}$$

as well as

$$\begin{aligned} E_{\theta_1}[X_1 - \frac{\psi(\theta_1)}{\theta_1}] &= \psi'(\theta_1) - \frac{[\psi(\theta_1) - \psi(0)]}{\theta_1} \\ &= \psi'(\theta_1) - \psi'(0) - \frac{1}{2}\theta_1\psi''(0) + O(\theta_1^2) \end{aligned}$$

$$= \frac{1}{2}\theta_1\psi''(0) + O(\theta_1^2)$$

Now suppose that $\theta_1 \downarrow 0$. Then, under both hypotheses, the expected increment of the random walk $S_n - n\psi(\theta_1)/\theta_1$ converges to zero at the same rate the stopping boundaries $\xi_1/\theta_1, \xi_2/\theta_1$ approaches $-\infty$, and ∞ respectively. Therefore it can be shown for $j = 0$ or 1 (The result follows from (A.3)),

$$P_{\theta_j}\{\theta_1[S_N - N\psi(\theta_1)] \geq \xi_2\} \rightarrow P\{W(t) \text{ touches } \xi_2 \text{ before touching } \xi_1\} \quad \text{(A.15)}$$

where $\{W(t), 0 \leq t < \infty\}$, is Brownian motion with drift $-(-1)^j\frac{1}{2}\psi''(0)$ and variance $\psi''(0)$.

This example demonstrates that the probability that a random walk is greater than some boundary is the probability that the limiting Brownian motion process touches one boundary before the other. The reader is enouraged to consult [37] for a more exact and in depth treatment of this subject.

Remove me

Bibliography

[1] Armitage, P. (1961). Discussion of [8]. *Journal of the Royal Statistical Society, Series B*, 23:30-31.

[2] Armitage, P. (1967). Some Developments in the Theory and Practice of Sequential Medical Trials. *Proceedings of the fifth Berkeley Symposium.* Math. Statist. Prob., 4:791-804.

[3] Armitage, P., McPherson, C.K. and Rowe, B.C. (1969). Repeated Significance Tests on Accumulating Data. *Journal of the Royal Statistical Society, Series A*, 132: 235-244.

[4] Barnard, G.A., Jenkins, G.M., and Winsten, C.B. (1962). Likelihood Inference and Time Series (with discussion). *Journal of the Royal Statistical Society, Series A*, 125: 321-372.

[5] Berger, J.O. and Wolpert, R.L. (1988). *The Likelihood Principle (2^{nd} edn).* Institute of Mathematical Statistics Lecture Note - Monograph Series, Vol. 6 (ed. S.S. Gupta), Hayward, Califronia: IMS.

[6] Berk, R. H. (1973). Some Asymptotic Aspects of Sequential Analysis. *Annals of Statistics*, 1: 1126-1138.

[7] Billingsley, P. (1968). *Convergence of Probability Measures*. New York: John Wiley and Sons.

[8] Birnbaum, A. (1962). On The Foundations of Statistical Inference (with discussion). *Journal of the American Statistical Association*, 53: 259-326.

[9] Cohen, J. (1994). The Earth is Round ($p < .05$). *American Psychologist*, 49: 997-1003.

[10] Cornfield, J. (1966). Sequential Trials, Sequential Analysis, and the Likelihood Principle. *American Statistician*, 29(2): 18-23.

[11] Cornfield, J. (1966). A Bayesian Test of Some Classical Hypotheses - with Application to Sequential Clinical Trials. *Journal of the American Statistical Association*, 61(315): 577-594.

[12] Edwards, A.W.F. (1969). Statistical Methods in Scientific Inference. *Nature*,227: 92.

[13] Edwards, A.W.F. (1972). *Likelihood*. London: Cambridge University Press.

[14] Edwards, W., Lindman, H., and Savage, L.J. (1963). Bayesian Statistical Inference for Psychological Research. *Psychological Review*, 70: 450-499.

[15] Einstein, A. (1956). *Investigations on the Theory of the Brownian Movement*. New York: Dover.

[16] Feller, W. (1940). Statistical Aspects of ESP. *Journal of Parapsychology*, 4: 271-298.

[17] Fisher, R.A. (1959). *Statistical Methods and Scientific Inference (2nd edn)*. New York: Hafner.

[18] Goodman, S.N. (1998). Multiple Comparisons, Explained. *American Journal of Epidemiology*, 147: 807-12.

[19] Govindarajulu, Z. (1975). *Sequential Statistical Procedures*. New York: Academic Press.

[20] Jeffreys, H. (1961). *Theory of Probability (3rd edn)*. Oxford: Oxford University Press.

[21] Kalbfleisch, J.D. and Sprott, D.A. (1970). Application of Likelihood Methods to Models Involving Large Numbers of Parameters (with discussion). *Journal of The Royal Statistical Society, Series B*, 32, 175-208.

[22] Karr, A.F. (1993). *Probability*. New York: Springer-Verlag.

[23] Kass, R.E. and Raftery, A.E. Bayes Factors. *Journal of the American Statistical Association*, 90: 773-795.

[24] Lorden, G. (1973). Open-Ended Tests for Koopman-Darmois Families. *Annals of Statistics*, 1(4): 633-643.

[25] Parzen, E. (1962). *Stochastic Processes*. San Francisco: Holden-Day, Inc.

[26] Pratt, J.W. (1977). 'Decisions' as Statistical Evidence and Birnbaum's Confidence Concept. *Synthese*, 36: 59-69.

[27] Robbins, H. (1970). Statistical Methods Related To The Law Of The Iterated Logarithm. *Annals of Mathematical Statistics*, 41(5): 1397-1409.

[28] Robbins, H. and Siegmund, D. (1970). Boundary Crossing Probabilities For The Wiener Process and Sample Sums. *Annals of Mathematical Statistics*, 41(5): 1410-1429.

[29] Royall, R. (1997) *Statistical Evidence: A Likelihood Paradigm*. New York: Chapman and Hall.

[30] Royall, R. (1998). The Likelihood Paradigm for Statistical Evidence. *Presented at Ecological Society of American Symposium on "The Nature of Statistical Evidence"*, August 3, Baltimore, Maryland.

[31] Royall, R. (1999). On the Probability of Observing Misleading Statistical Evidence. *Journal of the American Statistical Association* (to appear).

[32] Samuel-Cahn, E. (1974). Repeated Significance Tests II, for Hypotheses About the Normal Distribution. *Communications in Statistics*, 3(8): 711-733.

[33] Samuel-Cahn, E. (1974). Repeated Significance Tests I and II. Generalizations. *Communications in Statistics*, 3(8): 735-744.

[34] Savage, L.J. (1954). *The Foundations of Statistics*. New York: Wiley.

[35] Siegmund, D. (1979). Corrected Diffusion Approximations In Certain Random Walk Problems. *Advances in Applied Probability*, 11: 701-719.

[36] Siegmund, D. (1982). Large Deviations For Boundary Crossing Probabilities. *The Annals of Probability*, 10(3): 581-588.

[37] Siegmund, D. (1985). *Sequential Analysis: Tests and Confidence Intervals*. New York: Springer-Verlag.

[38] Siegmund, D. (1985). Corrected Diffusion Approximations and Their Applications. *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*. Editors Lecam, L. and Olshen, R. vol. 2: 599-617. Wadsworth, Inc.

[39] Smith, C.A.B. (1953). The Detection of Linkage in Human Genetics. *Journal of the Royal Statistical Society, Series B*, 15: 153-192.

[40] Spiegelhalter, D.J., Freedman, L.S., Blackburn, P.R. (1986). Monitoring Clinical Trials: Conditional or Predictive Power?. *Controlled Clincial Trials*, 7:8-17.

[41] Wald, A. (1947). *Sequential Analysis*. New York: Wiley and Sons.

[42] Wax, N. (1954). *Noise and Stochastic Processes*. New York: Dover.

[43] Whitehead, J. (1992). *The Design and Analysis of Sequential Clinical Trials (2^{nd} edn)*. New York: Ellis Horwood.

[44] Wijsman, R. (1991). Chapter 4: Stopping Times. *Handbook Of Sequential Analysis*. Editors Ghosh, B.K. and Sen, P.K. New York: Marcel Dekker, Inc.

Remove me