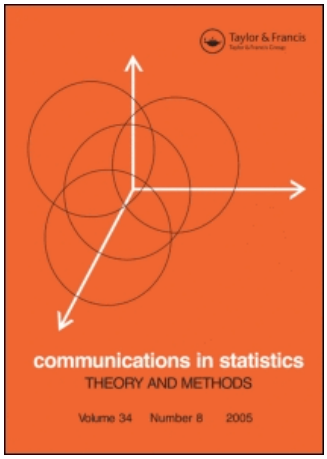

Communications in Statistics - Theory and Methods
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713597238

How Often Likelihood Ratios are Misleading in Sequential Trials

Jeffrey D. Blume [a]
[a] Center for Statistical Sciences, Brown University, Providence, Rhode Island, USA

Online Publication Date: 01 January 2008
To cite this Article: Blume, Jeffrey D. (2008) 'How Often Likelihood Ratios are Misleading in Sequential Trials', Communications in Statistics - Theory and Methods, 37:8, 1193 - 1206
To link to this article: DOI: 10.1080/03610920701713336
URL: http://dx.doi.org/10.1080/03610920701713336

Communications in Statistics—Theory and Methods, 37: 1193–1206, 2008

ISSN: 0361-0926 print/1532-415X online
DOI: 10.1080/03610920701713336

Inference

How Often Likelihood Ratios are Misleading in Sequential Trials

JEFFREY D. BLUME

Center for Statistical Sciences, Brown University, Providence, Rhode Island, USA

How often would investigators be misled if they took advantage of the likelihood principle and used likelihood ratios—which need not be adjusted for multiple looks at the data—to frequently examine accumulating data? The answer, perhaps surprisingly, is not often. As expected, the probability of observing misleading evidence does increase with each additional examination. However, the amount by which this probability increases converges to zero as the sample size grows. As a result, the probability of observing misleading evidence remains bounded—and therefore controllable—even with an infinite number of looks at the data. Here we use boundary crossing results to detail how often misleading likelihood ratios arise in sequential designs. We find that the probability of observing a misleading likelihood ratio is often much less than its universal bound. Additionally, we find that in the presence of fixed-dimensional nuisance parameters, profile likelihoods are to be preferred over estimated likelihoods which result from replacing the nuisance parameters by their global maximum likelihood estimates.

Keywords Brownian motion; Law of likelihood; Misleading evidence; Sequential trials.

Mathematics Subject Classification Primary 62A01; Secondary 62L05.

1. Introduction

Consider using a likelihood ratio, instead of a p-value, to measure the strength of statistical evidence for one hypothesis over another. Because the likelihood function is unaffected by the number of examinations of the data—unlike a p-value, re-examination of accumulating evidence with a likelihood ratio does not diminish its strength—frequent examination of accumulating data is encouraged. For some, there is a lingering concern that the repeated examinations will drastically increase the chance of observing misleading evidence. But here we allay those concerns. Using boundary crossing results, we show that although the probability

Received December 21, 2006; Accepted July 31, 2007
Address correspondence to Jeffrey D. Blume, Center for Statistical Sciences, Brown University, Providence, RI 02912, USA; E-mail: jblume@stat.brown.edu

of generating misleading evidence under a sequential design is greater than that under a fixed sample size design, the probability remains naturally bounded and controllable, even with an infinite number of looks at the data. We also detail how often strong misleading evidence is observed in sequential trials for study design purposes.

We pay particular attention to the "unscrupulous investigator" scenario, where an investigator examines his data with each new observation, stopping only when the data support his favorite hypothesis over the true one. This scenario yields the maximum chance of generating misleading evidence in a sequential design. Two classes of sequential designs are considered. In the first class, there are no limits or restrictions on the sample size; while in the second, the sample size is constrained by predetermined upper and lower bounds for the number of observations. It will be shown that in either class, the probability of generating strong misleading evidence is naturally bounded and controllable. Furthermore, these results generalize when fixed-dimensional nuisance parameters are present: they apply in large samples for nearby alternatives when profile likelihood ratios are used, but not when estimated likelihood ratios are used.

1.1. *Background*

The mathematical representation of statistical evidence is the likelihood function (Berger and Wolpert, 1988; Birnbaum, 1962) and the Law of Likelihood explains how the evidence should be measured and interpreted (Edwards, 1972; Hacking, 1965; Royall, 1997). If $x_1, \ldots, x_n$ are realizations of random variables $X_1, \ldots, X_n$ iid $f(X_i; \theta_0)$, then $L(\theta) \propto \prod f(x_i; \theta)$ is the likelihood function and the likelihood ratio, $L(\theta_1)/L(\theta_2)$ measures the strength of the evidence supporting the hypothesis that $H_1 : \theta = \theta_1$ over $H_2 : \theta = \theta_2$ (Hacking, 1965; Royall, 1997). The reader is referred to Royall (1997) for a comprehensive discussion of this approach. An article by Blume (2002) provides an introduction and a recent overview of the key concepts. See also Blume (2005, 2007) for recent developments.

The probability of observing strong evidence that supports a false hypothesis, say $\theta \neq \theta_0$, over the true one, θ_0, by a factor of k or greater is $P_0(L(\theta)/L(\theta_0) \geq k)$, which we refer to as the probability of observing k-strength misleading evidence or just the probability of misleading evidence. The probability of weak evidence is $P_0(1/k < L(\theta)/L(\theta_0) < k)$ and the probability of strong evidence is $P_0(L(\theta_0)/L(\theta) \geq k)$. Here $k > 1$ has meaning: with benchmarks of $k = 8$ and 32 indicating fairly strong and strong evidence, respectively (Edwards, 1972; Jeffreys, 1961; Royall, 1997, 2000). 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 themselves represent the strength of the evidence in the data.

The probability of observing misleading evidence is bounded above by $1/k$ (Barnard, 1947; Birnbaum, 1962; Smith, 1953). This bound is "universal" because it applies under any probability model. Moreover, this bound holds even with repeated examinations of accumulating data (Robbins, 1970): if both $f(X_n)$ and $g(X_n)$ are probability density functions and X_n is a vector of n observations, then

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

Thus, an experimenter who plans to examine his data with each new observation, stopping only when the data support H_g over H_f, will be eternally frustrated with probability at least $1 - 1/k$. This is an important scientific safeguard; an investigator searching for evidence to support his favorite hypothesis over the correct hypothesis is likely (with probability greater than $1 - 1/k$) *never* to find such evidence.

The universal bound is a crude device for controlling the probability of misleading evidence. The actual probability is often much less. For example, in a fixed sample size experiment, the limiting frequency of observing strong misleading evidence (for sufficiently smooth likelihood functions) is $\Phi[-|c|\sqrt{n}/2 - \ln k/|c|\sqrt{n}]$ where c is the distance between the null and alternative hypothesis in information units (Pratt, 1977; Royall, 2000). This probability has been named the "Bump" function because of its shape. Thus, the limiting maximum probability of observing misleading evidence over all alternatives is only $\Phi[-\sqrt{2\ln k}]$.

1.2. *The Probability That Observed Evidence is Misleading*

After the data have been collected, the strength of the evidence is determined by the likelihood ratio. Whether the evidence is weak or strong is clear from the numerical value of the likelihood ratio; however, it remains unknown if the evidence is misleading or not. Suppose we collect observations $x_1, x_2, \ldots, x_n \sim \text{iid } f(x_i, \theta)$ such that $L_n(\theta_1)/L_n(\theta_0) = k^* > 1$. Because $k^* > 1$, the data support H_1 over H_0. Consequently, these data only represent misleading evidence if they were generated from $f(x_i, \theta_0)$ (i.e., if H_0 is true). Therefore, the probability that the observed evidence is misleading is

$$P(H_0 \mid x_1, x_2, \ldots, x_n) = \left[1 + rk^*\right]^{-1} \tag{2}$$

by Bayes theorem, where $r = P(H_1)/P(H_0)$ is a ratio of (unknown) probabilities. As the strength of evidence (k^*) increases, the probability that the observed evidence is misleading decreases. When $r = 1$ the probability is 0.1111 and 0.0303 for $k^* = 8$ and $k^* = 32$, respectively. For a simple numerical example, the reader is referred to Blume (2002, Sec. 3.1). Additionally, (2) says that if the statistical evidence from two different experiments is identical—their likelihood functions are proportional—then both have exactly the same potential to be misleading.

The stopping rule is indeed relevant for determining how often a study design will yield misleading evidence. But once data are collected, the stopping rule is completely irrelevant; it does not affect the measure of evidence or the potential for observed evidence to be misleading. This works because we distinguish between (1) the measure of the strength of statistical evidence, (2) the probability that the measure will be misleading, and (3) the probability that an observed measure is misleading. Failure to do so has resulted in irresolvable controversies, such as those surrounding the proper use and interpretation of p-values or those concerning adjustments for multiple comparisons and multiple looks at data (see Blume and Peipert, 2003; Goodman, 1998; Goodman and Royall, 1988; Royall, 1986, 1997).

2. Sequential Trials

The defining quality of a sequential experimental design is the continual evaluation of statistical evidence during the course of study, for the purpose of determining

if the study should continue. The study stops, not after collecting a predetermined number of observations, but when the observations themselves represent sufficiently strong evidence. Here we investigate the study design that maximizes the chance of generating misleading evidence. This design is sometimes referred to as the "unscrupulous investigator" scenario; so named because the investigator examines his data with each new observation, stopping only when the data support his favorite hypothesis over the true one. We show that even in that case, the probability of generating misleading evidence is low and controllable.

Let $X_1, X_2, \ldots, X_n, \ldots$ be i.i.d. $f(X_i; \mu)$ and let $L_n(\mu) = \prod f(x_i; \mu)$ be the corresponding likelihood function for n observations. In the subsections that follow, we assume that observations are normally distributed with known variance σ^2 (these results apply more generally and extensions are considered in later sections). We consider the stopping rule whose purpose is to generate strong evidence for a fixed H_1 over a fixed H_0.

Let N be the stopping time defined by

$$N = \min\left\{n : n \geq 1, \frac{L_n(\mu_1)}{L_n(\mu_0)} \geq k\right\} \tag{3}$$

for some $k > 1$. We first consider the probability that this stopping time is between (m_0) and (m). We refer to this study as a "biased interval design." (Strong evidence supporting H_0 over H_1 is effectively ignored; hence the "bias"). From this it is easy to obtain the special case where no limits or restrictions on the sample size exist, which we will refer to as "biased open design".

2.1. *Biased Interval Designs*

Variables beyond the experimenter's control, such as time, money, and participant availability, often restrict the sample size of a study. For these situations an interval design can be used. An interval design collects at least m_0 but no more than m observations and has a stopping rule that determines if sampling continues as long as the sample size is within the appropriate range. The stopping rule N provides the maximum chance of generating misleading evidence for H_1 in such designs. The probability that the stopping time (3) is between m_0 and m observations is $P_0(m_0 \leq N \leq m)$.

Lemma 2.1 (Biased Interval Design). *Suppose $X_1, X_2, \ldots$ are i.i.d. normal random variables with 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*

$$\begin{aligned} P_0(m_0 \leq N \leq m) \cong\ & \Phi\left[-\left(\frac{\ln k}{|c|} + \rho\right)m^{-\frac{1}{2}} - \frac{|c|}{2}m^{\frac{1}{2}}\right] \\ & + \Phi\left[-\left(\frac{\ln k}{|c|} + \rho\right)(m_0 - 1)^{-\frac{1}{2}} - \frac{|c|}{2}(m_0 - 1)^{\frac{1}{2}}\right] \\ & + \frac{\exp\{-\rho|c|\}}{k}\left\{\Phi\left[\left(\frac{\ln k}{|c|} + \rho\right)(m_0 - 1)^{-\frac{1}{2}} - \frac{|c|}{2}(m_0 - 1)^{\frac{1}{2}}\right]\right. \\ & \left. - \Phi\left[\left(\frac{\ln k}{|c|} + \rho\right)m^{-\frac{1}{2}} - \frac{|c|}{2}m^{\frac{1}{2}}\right]\right\}, \end{aligned} \tag{4}$$

where $c = (\mu_1 - \mu_0)/\sigma$ *is the distance between* H_1 *and* H_0 *in standard deviation units and* $\rho \cong 0.583$.

The result follows from established properties of stopping times (Blume, 2007, supplement). The symbol $\cong$ indicates approximate equality. In (4) ρ is a constant that approximates the expected overshoot of the stopping boundary, depends on the normality assumption, and must be evaluated numerically (Siegmund, 1985, p. 50). Under these conditions, the numerical accuracy of this approximation is quite good (Siegmund, 1985, Table 3.5, p. 50, 57).

For $m_0 \leq 1$, Lemma 2.1 provides the maximum probability that the biased interval design generates misleading evidence. For $m_0 > 1$, Lemma 2.1 provides only an approximation to the biased interval design probability because it ignores the sample paths that had crossed the boundary before m_0, but were forced to continue only to cross the boundary again after m_0. Under the null hypothesis, the sample paths have negative drift and are therefore rare. As a result, this approximation is fine for our purposes.

2.2. *Biased Open Designs*

A biased open design is an open design (Wald, 1947) that continues sampling—possibly forever—until strong evidence for H_1 over H_0 is obtained. The probability that a biased open design generates strong misleading evidence for H_1 over H_0 can be expressed in terms of the probability that the stopping time N is finite, i.e., $P_0(N < \infty) = P_0\left(\frac{L_n(\mu_1)}{L_n(\mu_0)} \geq k; \text{ for any } n = 1, 2, \ldots\right)$. This is easily seen to be the limiting case of $P_0(1 \leq N \leq m)$ as $m \to \infty$, yielding the remark below.

Remark 2.1 (Biased Open Design)**.** Suppose $X_1, X_2, \ldots$ are i.i.d. normal random variables with 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

$$P_0(N < \infty) \cong \exp\{-\rho|c|\}/k \tag{5}$$

where $c = (\mu_1 - \mu_0)/\sigma$ is the distance between H_1 and H_0 in standard deviation units and $\rho \cong 0.583$.

The function on the RHS of Eq. (2.1) will be called the "tepee" function. This probability has been studied in the sequential analysis literature (Siegmund, 1985, §10.1), but is discussed here because of its previously unnoticed implications in this context. Additionally, it is well known that $P_1(N < \infty) = 1$ for all μ_1.

2.3. *Illustrations*

Figure 1 displays the probability that a biased open design produces strong misleading evidence for μ_1 over μ_0 when $k = 8$ (Eq. (2.1) is curve "*c*"), along with the corresponding bump function for reference. Note that the bump function "moves" inwards as the sample sizes increases (compare curves "*a*" and "*b*"). By contrast, the tepee function does not depend on sample size.

The bump function represents the probability that a fixed sample size design produces strong misleading evidence for μ_1 over μ_0 on a particular observation. The

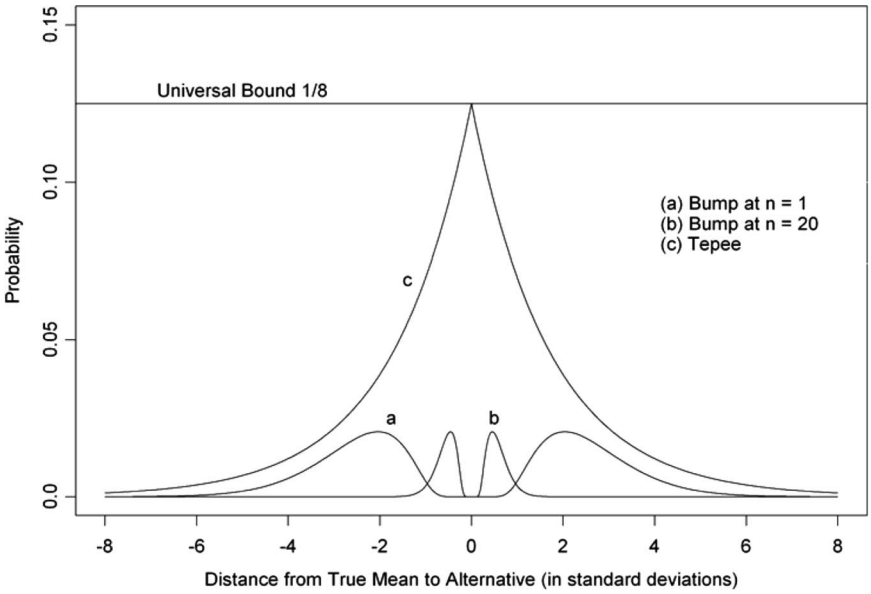

Figure 1. The bump (with $n = 1, 20$) and Tepee functions.

tepee function is the analogous probability for the biased open design. For large alternatives, the tepee function provides values that are only slightly larger than the bump function. For distant alternatives, (greater than 3 or 4 standard deviations), the bump function shows there is little chance of a single observation representing strong misleading evidence for μ_1 over μ_0. Moreover, the tepee function shows that, even if we continue to sample until such strong misleading evidence is obtained, this probability cannot be substantially increased.

For alternatives closer to the true hypothesis, the tepee function increases steadily as the distance between the two hypotheses approaches zero. This happens because as the sample size increases, the probability at each alternative builds up. Small values of the alternative (i.e., those close to the true value) accumulate this probability as the bump function "moves inwards toward zero" when the sample size increases. Thus large amounts of probability accumulate on near alternative as the entire bump function moves across them. By contrast, the bulk of the bump function is already past alternatives far from the true value (i.e., large values of "c") so these alternatives do not accumulate much more probability than what is initially specified by the bump function on the first observation.

Note that $\lim_{c\to 0} P_0(N < \infty) = \frac{1}{k}$ but $P_0(N < \infty) = 0$ at $c = 0$. That is, the tepee function is really discontinuous at $c = 0$; and therefore never achieves the universal bound. This discontinuity is not reflected in Figure 1.

Figure 2 shows the probability of collecting misleading evidence under a biased interval design when $m_0 = 3$ and $m = 15$ (curve "d"). The tepee function (curve "c") and the bump functions at $n = 3, 15$ (curves "a" and "b" respectively) are plotted for reference. Notice that the biased interval design produces a bump-like shape for the probability of generating misleading evidence. Many "bump-like" shapes are possible and are governed simply by m_0 and m. Thus, the function representing

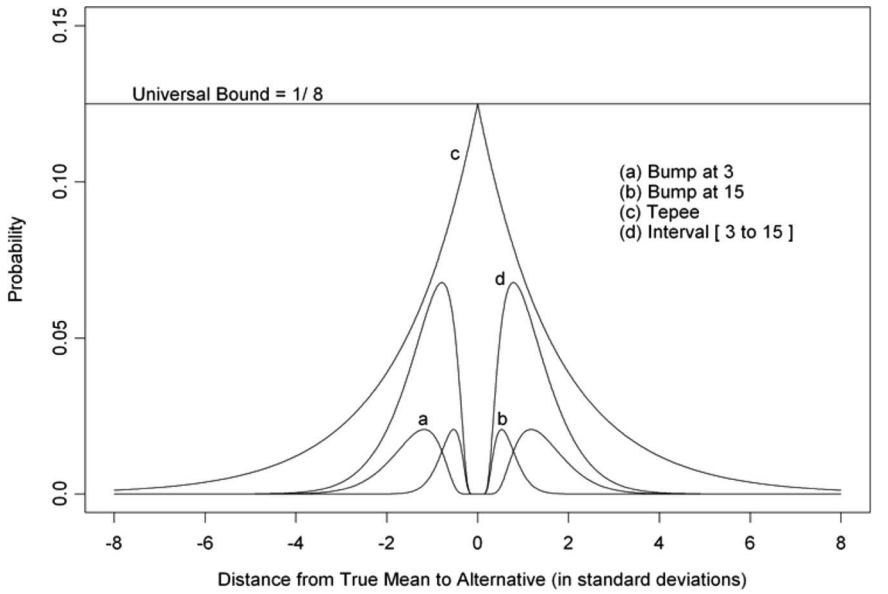

Figure 2. The biased interval design when $m_0 = 3$ and $m = 15$, Tepee and Bump functions.

the probability that a biased interval design generates misleading evidence will be referred to as the extended bump function.

When the distance between the two hypotheses is large, the probability of generating misleading evidence is, for the most part, a direct function of m_0. Most of the probability of collecting misleading evidence for alternatives far from the true hypothesis, although small, accumulates early (before m_0 observations). Conversely, most of the probability of collecting misleading evidence for alternatives close to the true hypothesis accumulates later (after m observations). Thus, the valley of the extended bump function is caused by truncating the sample size at m observations. Like the tepee function, the shape of the extended bump function can be explained by imagining a bump function moving over the x-axis. But now the bump function does not begin moving until $n = 3$ (see curve "a"). When $n = 15$ it can no longer move inward (see curve "b").

3. Single Parameter Exponential Family Distributions

Now we relax our normality assumption by assuming that our observations have a distribution that belongs to the single parameter exponential family. We will see that when using a biased open design, a generalized tepee function represents the probability of generating strong misleading evidence about the value of the parameter. The result is local, depending on the distance between the two hypotheses expressed in canonical form. For the biased interval design, a generalized extended bump function provides a good, but local, approximation to that probability for all parametric models in the single parameter exponential family with third moment equal to zero. For models with a nonzero third moment, an adjustment for skewness is required. Note that the expected excess over the

boundary, ρ_+, needs to be calculated for each distribution (see Siegmund, 1985, §10.4).

3.1. *Sequential Designs Revisited*

Observations $X_1, X_2, \ldots$ 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 one-to-one strictly increasing function of θ, so μ can be considered to be a function of θ or θ a function of μ, say $\theta(\mu)$. From exponential family theory we have $E_\theta[X_i] = \mu$ and $\mathrm{Var}_\theta[X_i] = \psi''(\theta)$, where the prime denotes differentiation with respect to θ. As before, let the likelihood function based on n observations be $L_n(\mu)$ and the sum of n observations as $S_n = X_1 + X_2 + \cdots + X_n$.

In what follows, it will be convenient to assume that f is standardized so that $\int xf(x)dx = 0$ and $\int x^2 f(x)dx = 1$ (a simple transformation on the X's will provide this case). Note that this standardization implies $\theta(0) = \psi'(0) = 0(=\psi(0))$. To further reduce algebraic computations, we will assume that $\mu_0 < 0 < \mu_1$ and $\psi(\theta_0) = \psi(\theta_1)$ in addition to the above standardization. This last assumption may always be satisfied by redefining X_i as $X_i - (\psi(\theta_1) - \psi(\theta_0))/(\theta_1 - \theta_0)$ and changing the corresponding value of μ (e.g., see Siegmund, 1985, p. 240).

Lemma 3.1 (Biased Interval Design, Bias)**.** *Suppose $X_1, X_2, \ldots$ are i.i.d. with standardized density $f_\mu(X) = \exp\{\theta X - \psi(\theta)\}f(X)$. Define the stopping time to be $N = \inf\{n : n \geq 1, L_n(\mu_1)/L_n(\mu_0) \geq k\}$. Then*

$$
\begin{aligned}
&P_{\mu_0}(m_0 \leq N \leq m) \\
&\quad = P_{\mu_0}(S_{m_0-1} \geq c_0) + P_{\mu_0}(S_m \geq c) + \exp\{-\Delta(b+\rho_+)\} \\
&\qquad \times \left\{ \Phi\left[-(c_0 + \kappa/3 - 2(b+\rho_+))(m_0 - 1 + c_0\kappa/3)^{-\frac{1}{2}} - \frac{\Delta}{2}(m_0 - 1 + c_0\kappa/3)^{\frac{1}{2}} \right] \right. \\
&\qquad \left. - \Phi\left[-(c + \kappa/3 - 2(b+\rho_+))(m + c\kappa/3)^{-\frac{1}{2}} - \frac{\Delta}{2}(m + c\kappa/3)^{\frac{1}{2}} \right] \right\} \\
&\qquad + o(m^{-1/2}) + o((m_0 - 1)^{-1/2}) - o(\Delta). \qquad (6)
\end{aligned}
$$

Here $c_0 = b + (m_0 - 1)\eta$, $c = b + m\eta$, $\eta = (\psi(\theta_1) - \psi(\theta_0))/\Delta$, $b = \ln k/\Delta$, $\Delta = |\theta_1 - \theta_0|$, $\kappa = E_0[X_1^3]$ *and the distribution of X_1 is from the standardized exponential family.*

The result follows from well-established properties of stopping times (Blume, 2007, supplement). Here the subscript on the probability denotes the true mean. The constant, ρ_+, represents the expected excess over the boundary under the "standardized" exponential family model. For unstandardized distributions, $\Delta' = \left|\theta_1\sqrt{\psi''(\theta_1)} - \theta_0\sqrt{\psi''(\theta_0)}\right|$ is used in place of Δ, but not in the error term. Note that under our standardization $\eta = 0$ because $\psi(\theta_1) = \psi(\theta_0)$.

As with the normal case, we have the following immediate special case (Siegmund, 1985, §10.1).

Remark 3.1 (Biased Open Design, bias). Suppose $X_1, X_2, \ldots$ are i.i.d. with standardized density $f_\mu(X) = \exp\{\theta X - \psi(\theta)\} f(X)$. Define the stopping time to be $N = \inf\{n : n \geq 1, L_n(\mu_1)/L_n(\mu_0) \geq k\}$. Then

$$P_0(N < \infty) = \exp\{-\rho_+ \Delta\}/k + o(\Delta) \tag{7}$$

where $\Delta = |\theta_1 - \theta_0|$.

The approximation is valid up to terms that are $o(\Delta^2)$ (Siegmund, 1985, Remark 10.6). Numerical experimentation suggests that Theorem 3.1 provides a reasonably accurate approximation for moderately large Δ. However, Remark 3.1 can be a highly local result, as we later illustrate. It is still true that $P_1(N < \infty) = 1$ for all μ_1.

If the observations come from a normal distribution with mean μ_0, we have that $\eta = (\mu_1 + \mu_0)/2$. Then the specified standardization implies that the problem can be recast as one where the observations have mean $-\Delta/2$ so that $\eta = 0$. As such, the probability of generating misleading evidence is equivalent to the probability that $S_n, n = 1, 2, \ldots$, now with negative drift $-\Delta/2$, crosses a positive horizontal boundary $b + \rho_+$.

3.2. *Illustration: Bernoulli Distribution*

Consider observations $X_1, X_2, \ldots, X_n$ i.i.d. Bernoulli(p). The Bernoulli(p) distribution is a member of the single parameter exponential family, with canonical parameter $\theta = \log[p/(1-p)]$ and $\psi(\theta) = \log[1 + \exp(\theta)]$. Consider using a biased open design to generate evidence for $H_1 : p = p_1$ over $H_0 : p = p_0 = 1/2$. By Lemma 3.1 we have

$$P_0(N < \infty) = \frac{\exp\{-\rho_+ \Delta'\}}{k} + o(\Delta) \tag{8}$$

where $\Delta' = \left|\log\left[\frac{p_1}{1-p_1}\right]\sqrt{4p_1(1-p_1)} - \log\left[\frac{p_0}{1-p_0}\right]\sqrt{4p_0(1-p_0)}\right|$, and $\rho_+ \cong 0.32$. For small to moderate Δ, expression (8) provides a good approximation to the probability that a biased open design generates misleading evidence for p_1 over $p_0 = 1/2$.

The expected excess over the boundary, ρ_+, needs to be 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 $\mathrm{Var}_p[Z_i] = 1$. Notice that X_i is effectively transformed to a Bernoulli-type random variable with $\mathrm{Var}_p[2X_i - 1] = 4p(1-p)$ and that now ρ_+ is actually a function of p. This is the approach taken in Example 10.63 in Siegmund (1985, p. 227); however $X_i^* = 2X_i - 1$ is implicitly used instead of the original random variable. Although other methods are available (Siegmund, 1985, p. 107, 138), they have been shown to be less accurate (Siegmund, 1985, Table 10.1). Under the standardized Bernoulli model, numerical integration gives ρ_+ as approximately 0.32.

Figure 3 is a plot of the generalized tepee function that represents the probability that the biased open design generates strong ($k = 8$) misleading evidence for p_1 over $p_0 = 1/2$. The horizontal lines are the probability of observing misleading evidence for p_1 over $p_0 = 1/2$ on the 4th observation. Notice that

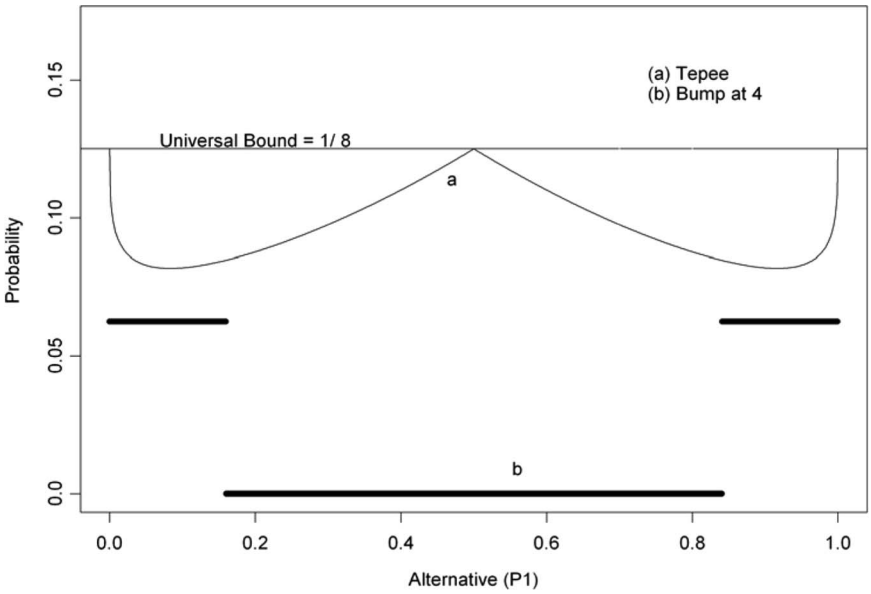

Figure 3. The generalized Tepee function for Bernoulli ($p = 1/2$).

when $p_1 = 0$ or 1, this generalized tepee function equals the universal bound $1/k$. This happens because with only three observations, the probability of observing misleading evidence can achieve the universal bound for such extreme p_1. For example, when $p_1 = 0$, the observations $X_1 + X_2 + X_3 = 0 = S_3$ give a likelihood ratio of $(1/0.5)^3 = 8$ in favor of $p_1 = 0$ over $p_0 = 1/2$. Under H_0, such evidence is misleading, and the probability of observing it is $P_0(S_3 = 0) = (0.5)^3 = 1/8$ equaling the universal bound.

3.3. *Illustration: Binomial Distribution*

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 a member of the single parameter exponential family, with canonical parameter $\theta = \log[p/(1-p)]$ and $\psi(\theta) = m \log[1 + \exp(\theta)]$. After standardizing as before, Theorem 3.1 applies with $\Delta' = \left|\log\left[\frac{p_1}{1-p_1}\right]\sqrt{4mp_1(1-p_1)}.\ -\log\left[\frac{p_0}{1-p_0}\right]\sqrt{4mp_0(1-p_0)}\right|$. We again use $\rho_+ \cong 0.32$. This is approximately correct; ρ_+ under the Binomial distribution is slightly smaller.

Figure 4 shows the generalized tepee function for the Binomial ($m = 50$, $p_0 = 1/2$) distribution. It represents the probability that the biased open design generates strong ($k = 8$) misleading evidence for p_1 over $p_0 = 1/2$. The horizontal lines represent the probability of observing misleading evidence for p_1 over $p_0 = 1/2$ on the 1st binomial observation. The generalized tepee function has a similar appearance to the tepee function in the normal case, except for the tails, where it increases to the universal bound. However, in contrast to the Bernoulli model, there is essentially no chance of generating misleading evidence for extreme values of p_1 and the generalized tepee function incorrectly approximates this probability as quite large. This indicates that the local approximation can be poor if applied universally to all $\Delta = |\theta_1 - \theta_0|$.

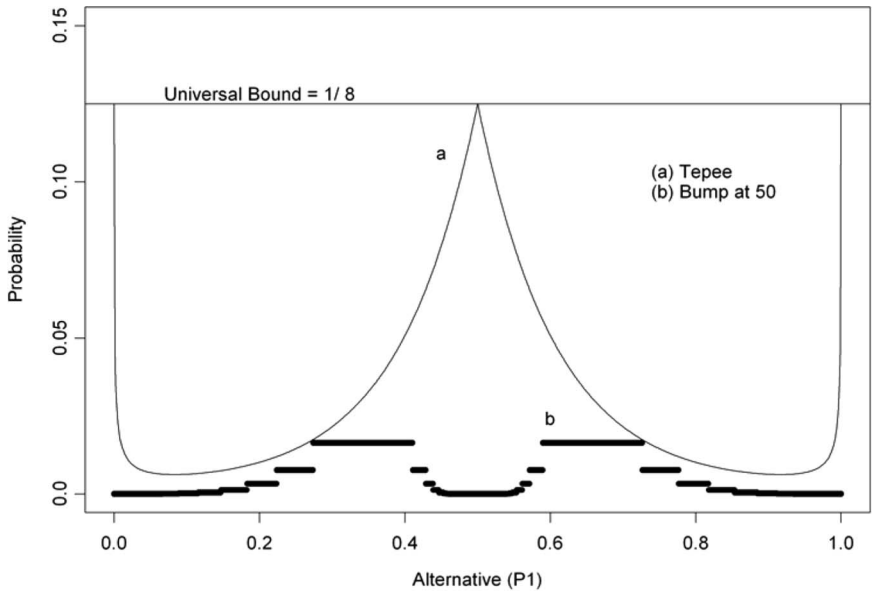

Figure 4. The Tepee function for binomial ($m = 50$, $p = 1/2$).

4. Removing Nuisance Parameters

The likelihood function sometimes depends on parameters that are not directly of interest themselves. We would like to eliminate the nuisance parameters in a way that does not increase our chances of observing misleading evidence. One approach is to use a conditional or marginal likelihood that is free of the nuisance parameter. Both are "true" likelihoods and are not more likely to be misleading. But in cases where such likelihoods do not exist, either a profile or estimated likelihood can be used to represent the evidence.

Suppose $X_1, X_2, \ldots, X_n$ i.i.d. $f(X_i; \theta, \gamma)$. The likelihood ratio $L_n(\theta_1, \gamma_1)/L_n(\theta_0, \gamma_0) = \prod_{i=1}^n f(X_i; \theta_1, \gamma_1)/f(X_i; \theta_0, \gamma_0)$ measures the support for one probability distribution, identified by (θ_1, γ_1) versus another identified by (θ_0, γ_0). For a fixed value of gamma, say $\gamma_0 = \gamma_1 = \gamma$, the likelihood ratio measures the relative support for θ_1 versus θ_0, but still depends on γ, which cannot in general, be removed. The profile likelihood function is $L_{pn}(\theta) = \max_\gamma L_n(\theta, \gamma) = L_n(\theta, \hat{\gamma}(\theta))$. The profile likelihood function maximizes the joint likelihood with respect to the nuisance parameter at each value of the parameter of interest. The estimated likelihood is $L_{en}(\theta) = L_n(\theta, \hat{\gamma_n})$, where $\hat{\gamma_n}$ is typically the global maximum likelihood estimate of γ.

In large samples and for alternatives in a neighborhood of the true value, both the tepee and extended bump functions represent the limiting probability of collecting strong misleading evidence in sequential designs. The results of Sec. 2 apply by replacing the variance with the inverse of the expected Fisher information. This is true for any smooth parametric model indexed only by the parameter of interest or when fixed dimensional nuisance parameters are present and a profile likelihood ratio is used. The main reason this is not true for estimated likelihoods is that the limiting distribution of the estimated log-likelihood ratio is not the same

as the true likelihood. As a result, the probability of generating misleading evidence with an estimated likelihood function is greater than that given by the tepee and extended bump functions. The fact that in large samples profile likelihood ratios behave just like "true" likelihood ratios is one key reason for prefering the profile likelihood approach to removing nuisance parameters.

The local analysis required consists of two major steps. First, we identify the limiting distribution of the log-likelihood ratio as the alternative approaches the true parameter value. The second step involves demonstrating that the limiting distribution of the stopping time under the class of smooth models is the same as that when the underlying distribution normal. The result is mainly a consequence of the fact that Brownian motion is the natural limiting process associated with sums of independent log-likelihood ratios. The results are presented below and an outline of these arguments can be found in Blume (2007, supplement).

For the class of smooth model, $X_1, X_2, \ldots, X_n$ be i.i.d. $f(X_i; \theta)$, where f is a smooth function of the real-valued parameter θ, we have the following.

Proposition 4.1 (Smooth Models)**.** *In large samples and for a class of smooth parametric models with form* $f(X_i; \theta)$ *where* θ *is real-valued, Lemma* 2.1 *and Remark* 2.1 *describe the probability of generating misleading evidence for alternatives in a* $O(n^{-\frac{1}{2}})$ *neighborhood of the true value when the variance* σ^2 *is replaced with the inverse of the fisher information* $1/I(\theta_0)$.

Next, consider expanding the class of smooth models to include multiparameter distributions of the form $f(X_i; \theta, \gamma)$, where f is a smooth function. We are interested in the case were θ is taken to be real-valued and the nuisance parameter γ a finite-dimensional vector.

Proposition 4.2 (Profile Likelihoods)**.** *In large samples and for a class of smooth parametric models with form* $f(X_i; \theta, \gamma)$*, where* θ *is real-valued and the nuisance parameter* γ *is a finite-dimensional vector, Lemma* 2.1 *and Remark* 2.1 *describe the probability of generating misleading evidence under a profile likelihood for alternatives in a* $O\left(n^{-\frac{1}{2}}\right)$ *neighborhood of the true value, when the variance* σ^2 *is replaced with the inverse of the fisher information* $1/I_{\theta\theta}(1-\rho^2_{\theta\gamma})$ *where* $\rho^2_{\theta\gamma} = \frac{I^2_{\theta\gamma}}{I_{\theta\theta}I_{\gamma\gamma}}$.

5. Remarks

The results here represent a fairly close and conservative approximation to what happens in the "two-sided" sequential design that continues sampling until either H_1 or H_0 is better supported by a factor of k or more. For example, under a normal distribution with fixed variance we have $P_0[L(\mu_1)/L(\mu_0) \geq k$ before $L(\mu_0)/L(\mu_1) \geq k] \approx 1/(1 + k\exp\{\rho\,|\,\mu_1 - \mu_0|/\sigma\}) \leq 1/(1+k)$. This shows that very few of the (misleading) sample paths actually support H_0 over H_1 prior to representing strong misleading evidence. Hence, the probability of generating misleading evidence in the "two-sided" design is only slightly less than that described here.

It is not possible, deliberately or otherwise, to frequently generate misleading evidence with a likelihood ratio. Moreover, the probability of generating misleading evidence in a sequential design is low and controllable, so investigators should not be shy about examining their data in this fashion. When a fixed dimensional

nuisance parameter is present, consider using a profile likelihood. Unlike an estimated likelihood, the profile likelihood is not any more likely to be misleading than the true likelihood in large samples.

An important subtext is that even when a likelihood ratio is used to measure the statistical evidence, the stopping rule is relevant to the operating characteristics of the study design. This helps to clarify that the likelihood principle (LP) does not imply that stopping rules are totally irrelevant. Rather, the LP says that stopping rules are irrelevant when measuring the strength of statistical evidence and, subsequently, when determining how likely it is that the observed evidence is misleading. This is why collections of observations with equivalent (proportional) likelihood functions have exactly the same potential to be misleading, regardless of how they were generated.

Acknowledgment

The author would like to thank Richard Royall for his comments on this work.

References

Barnard, G. A. (1947). Review of Wald, A. *Sequential Analysis. Journal of the American Statistical Association* 42:658–664.

Berger, J. O., Wolpert, R. L. (1988). *The Likelihood Principle*. 2nd ed. In: Gupta, S. S., ed. Institute of Mathematical Statistics Lecture Note – Monograph Series, Vol. 6. Hayward, CA: IMS (Institute of Mathematical Statistics).

Birnbaum, A. (1962). On the foundations of statistical inference (with discussion). *Journal of the American Statistical Association* 53:259–326.

Blume, J. D. (2002). Likelihood methods for measuring statistical evidence. *Statistics in Medicine* 21:2563–2599.

Blume, J. D., Peipert, J. F. (2003). What your statistician never told you about *p*-values. *Journal of the American Association of Gynecologic Laparoscopists* 10:439–444.

Blume, J. D. (2005). How to choose a working model for measuring statistical evidence about a regression parameter. *International Statistical Review* 73:351–363.

Blume, J. D., Su, L., Acosta, L., McGarvey, S. (2007). Statistical evidence for GLM parameters: A robust likelihood approach. *Statistics in Medicine* 26:2919–2936.

Blume, J. D. (2007). Supplementary material for "How often likelihood ratios are misleading in sequential trials." See author's website.

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

Goodman, S. N. (1998). Multiple comparisons, explained. *American Journal of Epidemiology* 147:807–812.

Goodman, S. N., Royall, R. M. (1988). Evidence and scientific research. *American Journal of Public Health* 78:1568–1574.

Hacking, I. (1965). *Logic of Statistical Inference*. New York: Cambridge University Press.

Jeffreys, H. (1961). *Theory of Probability*. 3rd ed. Oxford: Oxford University Press.

Pratt, J. W. (1977). Decisions' as statistical evidence and Birnbaum's confidence concept. *Synthese* 36:59–69.

Robbins, H. (1970). Statistical methods related to the law of the iterated logarithm. *Annals of Mathematical Statistics* 41:1397–1409.

Royall, R. M. (1986). The effect of sample size on the meaning of significant tests. *American Statistician* 40:313–315.

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

Royall, R. (2000). On the probability of observing misleading statistical evidence (with discussion). *Journal of the American Statistical Association* 95:760–767.

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

Smith, C. A. B. (1953). The detection of linkage in human genetics. *Journal of the Royal Statistical Society, Series B* 15:153–192.

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