STATISTICS IN MEDICINE
*Statist. Med.* 2002; **21**:2563–2599 (DOI: 10.1002/sim.1216)

# TUTORIAL IN BIOSTATISTICS
# Likelihood methods for measuring statistical evidence

Jeffrey D. Blume*,†

*Center for Statistical Sciences, Brown University, Providence, RI 02912, U.S.A.*

SUMMARY

Focused on interpreting data as statistical evidence, the evidential paradigm uses likelihood ratios to measure the strength of statistical evidence. Under this paradigm, re-examination of accumulating evidence is encouraged because (i) the likelihood ratio, unlike a *p*-value, is unaffected by the number of examinations and (ii) the probability of observing strong misleading evidence is naturally low, even for study designs that re-examine the data with each new observation. Further, the controllable probabilities of observing misleading and weak evidence provide assurance that the study design is reliable without affecting the strength of statistical evidence in the data. This paper illustrates the ideas and methods associated with using likelihood ratios to measure statistical evidence. It contains a comprehensive introduction to the evidential paradigm, including an overview of how to quantify the probability of observing misleading evidence for various study designs. The University Group Diabetes Program (UGDP), a classic and still controversial multi-centred clinical trial, is used as an illustrative example. Some of the original UGDP results, and subsequent re-analyses, are presented for comparison purposes.


KEY WORDS: Law of Likelihood; statistical evidence; misleading evidence; bump function; tepee function; University Group Diabetes Program (UGDP)

## 1. INTRODUCTION

Science looks to statistics for an objective measure of the strength of evidence in a given body of observations. This tutorial describes how and why likelihood ratios measure the strength of statistical evidence and discusses why that evidence is seldom misleading. To illustrate this methodology, a brief re-analysis of a well-known clinical trial is presented.

At present, there are two generally accepted approaches to interpreting data – one frequentist and one Bayesian – but neither of these approaches explicitly answer the question: 'What do the data say?' [1–4]. It has been argued elsewhere that both methodologies answer different questions, specifically 'What should I do?' and 'What should I believe?', respectively [1, 5].

* Correspondence to: Jeffrey D. Blume, Center for Statistical Sciences, Brown University Box G-H., Providence, RI 02912, U.S.A.
† E-mail: jblume@stat.brown.edu

*Received March 2001*
*Accepted November 2001*

Another viewpoint is that frequentist and Bayesian approaches both include more information in their interpretation of the data than just what the data themselves supply. Bayesian inference includes prior information/belief, and frequentist inference includes information about the sample space 'for example, the number of planned looks at the data'.

Likelihood based methodology, as presented here, essentially splits the difference between the frequentist and Bayesian approaches. The evidential paradigm (dubbed Evidentialism by Vineland and Hodge [6]) provides a framework for presenting and evaluating likelihood ratios as measures of statistical evidence for one hypothesis over another. Although likelihood ratios figure prominently in both frequentist and Bayesian methodologies, they are neither the focus nor the endpoint of either methodology. The key difference is that, under the evidential paradigm, the measure of the strength of statistical evidence is decoupled from the probability of observing misleading evidence. As a result, controllable probabilities of observing misleading or weak evidence provide quantities analogous to the type I and type II error rates of hypothesis testing, but do not themselves represent the strength of the evidence in the data.

This tutorial is intended for readers with a moderate amount of statistical training. However, readers who have only taken a semester or year-long graduate level introductory course in statistics or biostatistics will find the majority of material accessible. The basic concepts of this paradigm are presented and illustrated in Sections 2 and 3.1. The remaining subsections of Section 3 are more technical and detail how often likelihood ratios are misleading. These subsections may be skipped without loss of continuity. In Section 4, the University Group Diabetes Program (UGDP), a classic and still controversial multi-centred clinical trial, is re-examined with likelihood methods. The final section contains closing comments, computer code is explained in Appendix D, and some UGDP data are provided in Appendix E.

## 2. THE EVIDENTIAL PARADIGM

### 2.1. The Law of Likelihood

An experiment or observational study produces observations which, under a probability model, represent statistical evidence. The Law of Likelihood is a starting point for interpreting such evidence [1, 3, 7]. It provides a mechanism for interpreting those observations, in the context of a probability model, as statistical evidence for one hypothesis *vis-à-vis* another.

> *The Law of Likelihood.* If the first hypothesis, $H_1$, implies that the probability that a random variable $X$ takes the value $x$ is $P_1(x)$, while the second hypothesis, $H_2$, implies that the probability is $P_2(x)$, then the observation $X=x$ is evidence supporting $H_1$ over $H_2$ if and only if $P_1(x)>P_2(x)$, and the likelihood ratio, $P_1(x)/P_2(x)$, measures the strength of that evidence [1, 7].

Here $P_1(x)=P(x|H_1)$ is the probability of observing $x$ given that $H_1$ is true and $P_2(x)=P(x|H_2)$ is the probability of observing $x$ given that $H_2$ is true. The ratio of these conditional probabilities, $P(x|H_1)/P(x|H_2)$, is the likelihood ratio. Both $H_1$ and $H_2$ are 'simple' hypotheses because they specify a single numeric value for the probability of observing $x$. A 'composite' hypothesis, such as $H_c=\{H_1 \text{ or } H_2\}$, specifies a set of numeric values for $P(x|H_c)$, in this

 

case $\{P(x|H_1)$ or $P(x|H_2)\}$. Likelihood ratios involving composite hypotheses are, without additional assumptions, undefined because $P(x|H_c)$ does not specify a single numeric value. The Law's silence on composite hypotheses is discussed further in Section 2.6.

The Law of Likelihood explains that statistical evidence for one simple hypothesis *vis-à-vis* another is measured by the likelihood ratio. The hypothesis that is better supported is the hypothesis that assigns a higher probability to the observed events. That is, the data better support the hypothesis that more accurately predicted the observed data. If both hypotheses place the same probability on the observed events, then the observations do not support one hypothesis over the other.

This concept of statistical evidence 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. Why would it be wrong to say that the data represent evidence against $H_1$ if $P(x|H_1)$ is small? The reason is $P(x|H_1)$, while small, might be the largest among the hypotheses under consideration, making $H_1$ the hypothesis best supported by the data. A more in-depth discussion of this point is given by Royall (reference [1], Section 3.3) and Goodman and Royall [8].

An important implication of the Law of Likelihood is that statistical evidence and uncertainty have *different* mathematical forms. Uncertainty is measured by 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 statistical evidence. This is a critical insight; the measure of the strength of evidence and the frequency with which such evidence occurs are distinct mathematical quantities. R.A. Fisher, who was often cryptic in his writings, clearly understood this: "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" (reference [9], p. 93).

### *2.2. The strength of statistical evidence*

The likelihood ratio for $H_1$ versus $H_2$ measures the strength of evidence for the first hypothesis over the second hypothesis. 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. If the likelihood ratio is greater than one, the evidence favours $H_1$ over $H_2$ and if the likelihood ratio is less than one, the evidence favours $H_2$ over $H_1$. Likelihood ratios may 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 have been suggested to distinguish between weak, moderate and strong evidence [1, 10]. (Others have proposed similar benchmarks in the literature [3, 11, 12].) Observations resulting in a likelihood ratio of 8 or more (or less than $1/8$) represent at least moderate or fairly strong evidence and those with a likelihood ratio of 32 or more (or less than $1/32$) represent strong evidence. Note that while a likelihood ratio of 8 represents fairly strong evidence, so does a likelihood ratio of 7.5 or 10 (albeit to a lesser or greater degree).

 

Table I. Properties of the diagnostic test for disease $D$.

| Disease $D$ | Test result | |
|---|---|---|
| | Positive | Negative |
| Present | 0.94 | 0.06 |
| Absent | 0.02 | 0.98 |

The scale of statistical evidence is not discrete; evidence does not jump from category to category.

A likelihood ratio of $k$-strength represents statistical evidence of the same strength in different problems. This is because a likelihood ratio supporting $H_1$ over $H_2$ by a factor of $k$ represents evidence precisely strong enough to increase the prior probability ratio of $H_1$ to $H_2$ by a factor of $k$, regardless of what $H_1$ and $H_2$ represent and regardless of their initial probabilities [10]. This is expressed mathematically by

$$\frac{P(H_1|x)}{P(H_2|x)} = k\,\frac{P(H_1)}{P(H_2)} \tag{1}$$

where $k = P(x|H_1)/P(x|H_2)$ is the likelihood ratio, $P(H_1)/P(H_2)$ is the prior probability ratio, and $P(H_1|x)/P(H_2|x)$ is the posterior probability ratio (that is, the probability ratio *after* observing the data $x$). Note that the prior probabilities, $P(H_1)$ and $P(H_2)$, do not depend on the data and must be predetermined (often subjectively).

Sometimes the data, while representing strong statistical evidence for $H_1$ over $H_2$, do not represent *strong enough* evidence to make $H_1$ appear more probable than $H_2$. For example, suppose that $H_2$ is initially 20 times more likely than $H_1$. An experiment is conducted and the data yield a likelihood ratio of $k = 10$ supporting $H_1$ over $H_2$. Therefore, the data represent fairly strong evidence supporting $H_1$ over $H_2$, even though $H_2$ remains more probable than $H_1$ (but now by a factor of two, instead of 20).

Equation (1) shows why the same statistical evidence may convince some that $H_1$ is more probable than $H_2$, but not others. This is because everyone has different prior probabilities for the hypotheses depending on their own experience and beliefs. Hence it is worth while reiterating that statistical evidence is not measured by the relative belief that one hypothesis is more likely than another either before or after the experiment. That is, statistical evidence is not measured by the prior or posterior probability ratio. Rather, statistical evidence is measured by the data dependent factor that uniformly modifies all beliefs, no matter what their initial magnitude. That factor is the likelihood ratio.

### 2.3. *Example: a diagnostic test*

Consider a diagnostic test for disease $D$ that has 94 per cent sensitivity and 98 per cent specificity. The properties of this test are listed in Table I. A test result from this diagnostic test represents statistical evidence about the presence of disease $D$. Here, a positive test result does not prove that disease is present, but represents statistical evidence supporting $H_{D+}$ (disease present) over $H_{D-}$ (disease absent) because $0.94 = P(T+|D+) > P(T+|D-) = 0.02$. The statistical evidence may be characterized as 'strong' because the likelihood ratio is $P(T+|D+)/P(T+|D-) = 0.94/0.02 = 47$.

 

Whether or not this positive test result presents *convincing* evidence that the disease is present depends on the disease prevalence in the population, $P(D+)$. In this context, equation (1) becomes

$$\frac{P(D+|T+)}{P(D-|T+)} = 47\,\frac{P(D+)}{P(D-)} \tag{2}$$

where $P(D-) = 1 - P(D+)$ and $P(D-|T+) = 1 - P(D+|T+)$. When the disease prevalence is less than 2 per cent, the prior probability ratio, $P(D+)/P(D-) < 0.02/0.98 = 1/49$, is very small and the posterior probability ratio, $P(D+|T+)/P(D-|T+) < 47/49$, is less than one. Therefore, a positive test result for this rare disease does not represent *strong enough* evidence to convince a physician that the disease is present, even though that positive result represents strong evidence for $H_{D+}$ over $H_{D-}$.

For example, upon observing a positive test, a prior probability of disease $P(D+) = 0.015$ increases to $0.417 = P(D+|T+)$ and the prior probability of no disease, $P(D-) = 0.985$, is reduced to $0.583 = P(D-|T+)$. It is still more probable that the disease is absent, but nevertheless wrong to claim that this positive result is evidence that the disease is absent. Why? Because observing a positive result increases the probability that the disease is present from 0.015 to 0.417 and decreases the probability that the disease is absent from 0.985 to 0.583, or equivalently, because the likelihood ratio supports $H_{D+}$ over $H_{D-}$ by a factor of 47.

An application of Bayes theorem shows that a positive test result always increases the physician's belief that the disease is present if the likelihood ratio supporting $H_{D+}$ over $H_{D-}$ is greater than one. However that same positive result will not be convincing if it fails to overwhelm the prior probability that the disease is absent. Even further, any medical *action* that the physician takes is based not only on his or her beliefs about the presence of disease, but also on the possible benefits and costs (monetary and otherwise) associated with the treatment. If available, a simple, non-invasive, inexpensive treatment may be prescribed as a precautionary measure, even though the physician might not believe that the disease is present.

The point is that action, belief and evidence are distinct concepts. It is not enough to answer the question 'What do the data say?' by explaining what one believes or what one should do. Likewise for statistical methodology; the techniques for answering questions such as 'What should I do?' (frequentist) and 'What should I believe?' (Bayesian) cannot be used to answer the question 'What do the data say?' [1, 4, 5, 8]. When speaking of statistical 'inference', it is necessary to distinguish between these three questions, not only because the methodologies required to address them are different, but also because each question is important in its own right.

### *2.4. Likelihood functions*

The diagnostic test example, although very instructive, is somewhat simplistic in that its probability model (given by Table I) consists of only two simple hypotheses, $H_{D+}$ and $H_{D-}$, and only a single likelihood ratio needs to be reported. Under more complex probability models, hypotheses often specify a particular value for a parameter of interest (such as a mean, odds ratio or probability), making the total number of simple hypotheses much greater than two. In this case, many likelihood ratios should be reported, namely one for each pair of simple hypotheses, and simply listing all of them becomes quite cumbersome. A more concise

 

way of reporting the evidence is to report and/or display the likelihood function, which is the mathematical representation of the statistical evidence in the data [13].

In essence, a likelihood function is an expression of the conditional probabilities $P(x|H_1)$, $P(x|H_2),\ldots$ as a single function $P(x|H_i)$ for $i=1,2,\ldots$ . The notation used to represent a likelihood function is $L(H_i|x)$ or $L(H_i)$, rather than $P(x|H_i)$, to emphasize that the observed data are fixed and the hypothesis $H_i$ varies [14]. A plot of the likelihood function ($L(H_i)$ versus $H_i$) reveals which hypotheses are better supported by the data because these hypotheses will have a larger $P(x|H_i)$ relative to other hypotheses.

To illustrate, consider an experiment that generates evidence about the (unknown) probability of obtaining heads on a toss of a biased coin. The experiment consists of flipping the biased coin $n$ times and recording the number of times it lands heads. Because each toss of the coin is independent, the probability model is binomial and the probability of observing $x$ heads out of $n$ tosses is $P(x|\theta)=c(n,x)\theta^x(1-\theta)^{n-x}$, where $c(n,x)=n!/(n-x)!x!$ is the binomial coefficient and $\theta$ is the unknown probability of heads. Here each simple hypothesis specifies a value for the unknown parameter $\theta$ (for example, $H_0{:}\theta=0.5$).

If the coin is tossed 50 times and 14 heads are observed, the likelihood function is given by $L(\theta)=c(n,x)\theta^{14}(1-\theta)^{50-14}$. The likelihood ratio for $H_1$ versus $H_2$ is given by $L(H_1)/L(H_2)=P(x=14|H_1)/P(x=14|H_2)$. For example, these data support $\theta=0.3$ over $\theta=0.5$ by a factor of $L(0.3)/L(0.5)=0.3^{14}(0.7)^{36}/0.5^{14}(0.5)^{36}=143$. The value of $\theta$ that maximizes the likelihood function is called the maximum likelihood estimator (MLE) and will be denoted by $\hat{\theta}$. The MLE deserves mention because it is the best supported hypothesis ($L(\hat{\theta})/L(\theta)\geqslant 1$ for all $\theta$). However, not too much emphasis should be placed on the MLE because other hypotheses are (essentially) equally well supported. For example, the MLE in the biased coin example is $\hat{\theta}=14/50=0.28$, but there is only very weak evidence to support $\theta=0.28$ over $\theta=0.3$ because $L(0.28)/L(0.3)=1.05$.

Likelihood functions are graphed to provide a visual impression of the evidence over the parameter space. For presentation purposes, likelihood functions are standardized by their maximum value (a constant). The scaling constant for the likelihood function can be chosen arbitrarily because only ratios of likelihood functions measure the statistical evidence. In the biased coin example, the standardized likelihood function is

$$\frac{L(\theta)}{\max_\theta L(\theta)}=\frac{L(\theta)}{L(\hat{\theta})}=\frac{\theta^{14}(1-\theta)^{36}}{0.28^{14}(1-0.28)^{36}} \tag{3}$$

Figure 1 displays the standardized likelihood function. The $y$-axis is not labelled to emphasize that only ratios of points on the likelihood function have evidential meaning. Note that the best supported value, the MLE $\hat{\theta}=0.28$, is at the crest of the likelihood function. The usefulness of the likelihood function is that it 'shows' all the likelihood ratios and provides a visual impression of the evidence about $\theta$. Appendix D1 shows how to reproduce this plot in the statistical package S-plus.

Likelihood functions play a prominent role in another well-known statistical principle that is often confused with the Law of Likelihood. That principle is the likelihood principle, which preceded the Law of Likelihood by several years and outlined the conditions under which two experiments produce equivalent statistical evidence [13]. The condition is a simple one: two experiments produce equivalent statistical evidence if their likelihood functions are proportional by a fixed constant, or equivalently, if all the likelihood ratios are equal for the

 

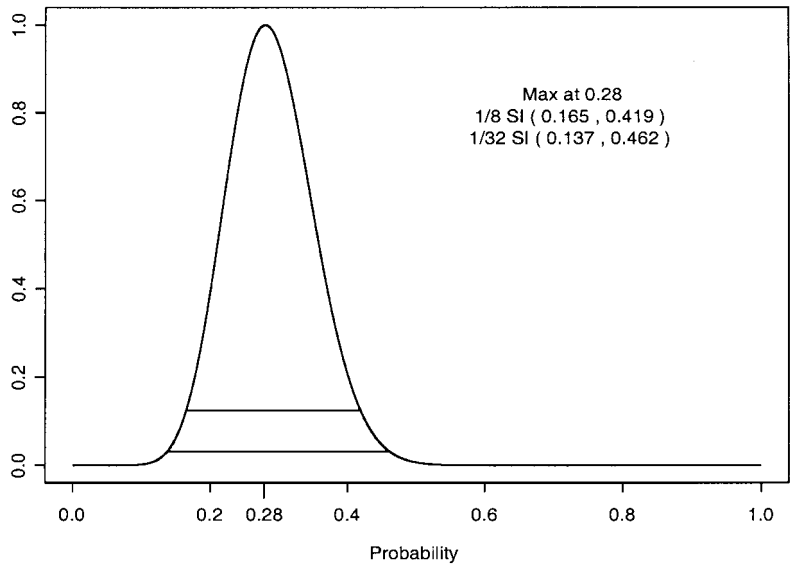


Figure 1. Likelihood function for the probability of heads.

two experiments (see reference [14] for a review). The Law of Likelihood is concerned with how the data should be interpreted as statistical evidence for one hypothesis over another and should not be confused with the likelihood principle.

### 2.5. *Support intervals*

Drawn within the likelihood function in Figure 1 are two lines parallel to the $x$-axis. These lines represent likelihood support intervals, identifying all the parameter values for $\theta$ that are consistent with the data at a certain level. The values for $\theta$ that are most consistent with the data are those values under the crest of the likelihood function. Hence, a $1/k$ likelihood support interval (SI) is defined as the set of $\theta$s where the standardized likelihood function is greater than $1/k$ [1]. That is, a $1/k$ likelihood support interval is

$$\left\{\text{all } \theta \text{ where } \frac{L(\theta)}{\max_\theta L(\theta)} \geqslant \frac{1}{k}\right\} = \left\{\text{all } \theta \text{ where } \frac{L(\hat{\theta})}{L(\theta)} \leqslant k\right\} \tag{4}$$

In words, any theta within the $1/k$ SI is supported by the data because the best supported hypothesis, $\hat{\theta}$, is only better supported by a factor of $k$ or less. Thus if $k = 8$ there is only weak evidence supporting the MLE over any theta in the interval. Likelihood support intervals present a concise summarization of the evidence about $\theta$ without having to graph the likelihood function or report numerous likelihood ratios.

Remaining consistent with the aforementioned benchmarks, $k = 8$ and 32 are used to construct moderate and strong support intervals. In Figure 1, the 1/8 SI for the probability of heads is 0.165 to 0.419 (the 1/8 SI is the line corresponding to a height of 1/8 on the $y$-axis). Hypotheses within the interval may be better supported over others within the interval, but the level of support is weak and less than a factor of 8. For those hypothesized values outside

 

the interval, there always exists another probability, namely $\hat{\theta} = 0.28$, that is better supported by a factor of more than 8. For these reasons 1/32 SIs are often called 'stronger' support intervals over 1/8 SIs. In the current example, hypotheses suggesting the probability of heads is between 0.137 and 0.462 (the 1/32 SI is the line corresponding to a height of 1/32 on the $y$-axis) are considered consistent with the data at a stronger level. These support intervals indicate that the coin is biased.

Support intervals and *exact* confidence intervals are indeed related but this relationship is only mathematical, because the interpretation and construction of these intervals is quite different (see reference [1] for an in-depth discussion). For example, a $1/k$ SI for the mean of a normal distribution is $\bar{x} \pm \sqrt{(2 \log k)}\sigma/\sqrt{n}$ while a $(1-\alpha)100$ per cent confidence interval (CI) is $\bar{x} \pm Z_{\alpha/2}\sigma/\sqrt{n}$. When $k = 8$, $\sqrt{(2 \log k)} = 2.039$ and the $1/k$ SI has the form of a 95.9 per cent confidence interval (a 95 per cent CI corresponds to a 1/6.67 SI). However the form of the confidence interval depends on more than just the type I error. If three looks at the data are planned, that same 95.9 per cent confidence interval becomes $\bar{x} \pm 2.47\sigma/\sqrt{n}$ because $\alpha$ is divided by three to maintain the overall type I error. Advantageously, likelihood support intervals depend only on the data, so they keep the same form regardless of the number of looks.

### 2.6. Composite hypotheses

The Law of Likelihood explains how to measure evidence between two simple hypotheses and is therefore silent with regard to composite hypotheses. This silence is a natural consequence of the Law itself; it avoids the inherent subjectivity in summarizing evidence over a composite space.

Consider again the biased coin example from the previous section. How might the evidence for the composite hypothesis $H_c{:}\theta < 0.5$ versus the simple hypothesis $H_s{:}\theta = 0.5$ be measured? The Law provides no direction here, because the likelihood ratio, $L(\theta < 0.5)/L(0.5) = P(x = 14|H_c)/P(x = 14|H_s)$, depends on the undefined quantity $L(\theta < 0.5) = P(x = 14|\theta < 0.5) = P(x = 14|H_c)$ that specifies a set of probabilities rather than just one. The problem here is that even if we were willing to specify a rule for choosing a single probability out of the set, the final measure of support will depend on that rule.

For example, the support for $H_c$ over $H_s$, say $k_c$, may be defined by: (i) taking the maximum, $k_c = \max_{\theta<0.5}[L(\theta)/L(0.5)] = L(0.28)/L(0.5) = 150$, which indicates very strong support for $H_c$ over $H_s$, (ii) taking the minimum, $k_c = \min_{\theta<0.5}[L(\theta)/L(0.5)] = L(0)/L(0.5) = 0$, which indicates overwhelming support for $H_s$ over $H_c$, or (iii) taking a weighted average, $k_c = \int_\theta (L(\theta)/L(0.5))w(\theta)\,\mathrm{d}\theta$ where $w(\theta)$ are weights such that $\int_\theta w(\theta) = 1$, which indicates that the support will be between 0 and 150 depending on the weights. Thus the support for $H_s$ over $H_c$ changes along with the rule.

Alternatively, situations involving composite hypotheses can be transformed so that the Law of Likelihood applies. The idea is Bayesian—change the probability model by specifying a prior distribution for the simple hypotheses. The prior provides a mechanism for reducing the composite hypothesis to a simple hypothesis and then the Law of Likelihood applies to the likelihood ratio under the extended probability model. (This new likelihood ratio is called a Bayes factor and Appendix A provides the details.) This Bayesian approach does not provide a measure of evidence where the Law of Likelihood cannot; it simply changes the problem so that the Law of Likelihood is applicable. The drawback of this approach is that there are many 'Bayesian' solutions with no objective criteria to choose between them [10].

 

The point here is not that choosing a method of summarization (that is, a rule) is bad, but that it is unnecessary and often arbitrary. Pairwise reasoning between two simple hypotheses is not alien to statistics and, in fact, is a major component of current statistical thinking. For example, a power curve displays the probability of rejecting a simple null hypothesis at every possible simple alternative hypothesis. (When examining the power of a test, no attempt is made to summarize the power over a composite space, although Bayesians sometimes determine the 'pre-posterior risk' which is similar in some respects [15].) Comprehensively describing the power for rejecting some fixed null hypothesis requires the presentation of the entire power curve. Analogously, the entire spectrum of evidential support is conveyed concisely in a graph of the likelihood function. No further reduction or summarization of the evidence is necessary.

## 3. MISLEADING EVIDENCE

### 3.1. Definition and implications

It is possible to observe strong evidence for $H_2$ over $H_1$ when, in actuality, $H_1$ is correct. *Misleading evidence is defined as strong evidence in favour of the incorrect hypothesis over the correct hypothesis.* 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.

Recall the diagnostic test example discussed in Section 2.3 whose properties were given in Table I. For this test, a positive test result is always strong evidence supporting $H_{D+}$ over $H_{D-}$ because $P(T+|D+)/P(T+|D-) = 47$, but an observed positive test represents misleading evidence if the disease is truly absent. What makes this test a good test is that it rarely produces this type of misleading evidence; the probability of observing a (misleading) likelihood ratio that supports $H_{D+}$ is $P(T+|D-) = 0.02$.

However, the probability of observing misleading evidence is irrelevant once data have been collected. While this probability provides assurance that the study design is reliable, it does not affect the strength of statistical evidence in observed data (as measured by the likelihood ratio) or the probability that the *observed* evidence is misleading. In fact, if two different studies produce the same statistical evidence for one hypothesis over another (that is, they have identical likelihood ratios), then the probability that the observed evidence is misleading is exactly the same for both studies.

To illustrate, consider another diagnostic test whose properties are given in Table II. The sensitivity of this test is only 47 per cent, much worse than the first test, but its specificity is slightly better at 99 per cent. A positive test result from this test also represents strong statistical evidence supporting $H_{D+}$ over $H_{D-}$ by a factor of $0.47/0.01 = 47$, but if the disease is absent, the second test is only half as likely to produce (misleading) evidence in favour of $H_{D+}$ (because $P(T+|D-) = 0.01$).

A positive result from either test represents evidence supporting $H_{D+}$ over $H_{D-}$ by a factor of 47. However, the second test produces misleading positive results half as often. Why, then, is it unimportant to know which test produced the observed positive result? Put another way, is an observed positive result on the second test 'less likely to be misleading' or 'more reliable' in some sense, or does it warrant more 'confidence' [10]. The answer is no.

 

Table II. Properties of the second diagnostic test for disease $D$.

| Disease $D$ | Test result | |
|---|---|---|
| | Positive | Negative |
| Present | 0.47 | 0.53 |
| Absent | 0.01 | 0.99 |

An observed positive result on the second test is equivalent, as statistical evidence about the presence or absence of disease, to an observed positive result from the first test. The strength of evidence for $H_{D+}$ over $H_{D-}$ is identical for both tests because their likelihood ratios are equal, but more to the point, an observed positive result from the first test is not any more likely to be misleading than an observed positive result from the second test. Why? Because an observed positive result is misleading if, and only if, the subject does not have the disease and $P(D-|T+)$ is the same for both tests! From Bayes theorem we have

$$P(D-|T+) = \frac{P(T+|D-)P(D-)}{P(T+|D-)P(D-)+P(T+|D+)P(D+)} = \frac{P(D-)}{P(D-)+\frac{P(T+|D+)}{P(T+|D-)}P(D+)} = \frac{P(D-)}{P(D-)+47P(D+)} \tag{5}$$

for both tests. Notice that $P(D-|T+)$ only depends on the likelihood ratio and the prevalence. Hence the propensity for an observed positive result to be misleading does not depend on which test produced the result (see reference [1], Section 4.5 for further discussion). Overall the implication is clear: the probability of observing misleading evidence does not affect the strength of statistical evidence in the data or the probability that *observed* evidence is misleading.

It is critical to distinguish between the probability of observing misleading evidence and the probability that observed evidence is misleading. (Incidentally, confusion over adjusting $p$-values and type I errors for multiple comparisons or multiple looks at the data is, in part, due to the failure to distinguish between these two probabilities [1, 16].) Even though the probability of observing misleading evidence does not affect our measure of statistical evidence, it plays an important role in the planning of experiments along with the probability of observing weak evidence. Therefore, the remainder of Section 3 is devoted to introducing the probabilities of misleading and weak evidence and establishing that, in general, strong misleading evidence is seldom observed.

### *3.2. Probabilities of misleading, weak evidence and the type* I, II *errors of hypothesis testing*

Study designs are evaluated in terms of their operational characteristics, for example, the frequency with which a particular study design will produce misleading or weak evidence. These design characteristics provide assurance that the design is reliable, but do not affect the statistical evidence in the observed data. When discussing the probabilities of misleading and weak evidence it is instructive to contrast them with the familiar type I and II errors of hypothesis testing.

 

The classical Neyman–Pearson hypothesis testing theory and the evidential paradigm both 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. Conversely, the Law of Likelihood says 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 $L_n(\mu) = \prod_{i=1}^{n} f(x_i;\mu)$ 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

$$P_0\left(\frac{L_n(\mu_1)}{L_n(\mu_0)} \geqslant k_{\alpha,n}\right) = \alpha \quad \text{where } \alpha \text{ is fixed over } n \tag{6}$$

By contrast the probability of observing misleading evidence for $\mu_1$ over $\mu_0$ is

$$P_0\left(\frac{L_n(\mu_1)}{L_n(\mu_0)} \geqslant k\right) = M(n,k) \quad \text{where } k \text{ is fixed over } n \tag{7}$$

While the type I error rate and the probability of observing misleading evidence appear to have the same mathematical form, they actually are very different. In hypothesis testing, the type I error rate is fixed at $\alpha$ and the strength of evidence at which the test rejects, $k_{\alpha,n}$, depends on $\alpha$ and $n$. This often leads to confusion because two tests which reject at the same $\alpha$-level represent different strengths of evidence depending on their respective sample sizes [17, 18]. Just the opposite is true for the probability of observing misleading evidence; the strength of the evidence, $k$, is fixed and the resulting probability, $M(n,k)$, depends on $k$ and the sample size $n$.

To make these differences transparent, consider the case when the data are normally distributed with mean $\mu_0$ and known variance $\sigma^2$. Further suppose that $n$ observations are collected in a standard fixed sample size experiment. The probabilities of observing misleading and weak evidence depend on the sample size and the distance between the two fixed simple hypotheses $H_1{:}\mu = \mu_1$ and $H_0{:}\mu = \mu_0$. The strength of evidence for $H_1$ versus $H_0$ provided by the observations $x_1, \ldots, x_n$ is given by

$$\frac{L_n(\mu_1)}{L_n(\mu_0)} = \exp\left\{\frac{n(\mu_1 - \mu_0)}{\sigma^2}\left(\frac{S_n}{n} - \frac{\mu_1 + \mu_0}{2}\right)\right\} \tag{8}$$

where $S_n = x_1 + \cdots + x_n$.

When the true mean is $\mu_0$, the probability that the observations will support $\mu_1$ over $\mu_0$ by a factor of $k$ or more is straightforward to calculate [1, 10, 19].

$$M(n,k) = P_0\left(\frac{L_n(\mu_1)}{L_n(\mu_0)} \geqslant k\right) = \Phi\left[-\frac{\Delta\sqrt{n}}{2\sigma} - \frac{\sigma \ln k}{\Delta\sqrt{n}}\right] \tag{9}$$

where $\Delta = |\mu_1 - \mu_0|$ and $\Phi(.)$ is the standard normal cumulative distribution function.

 

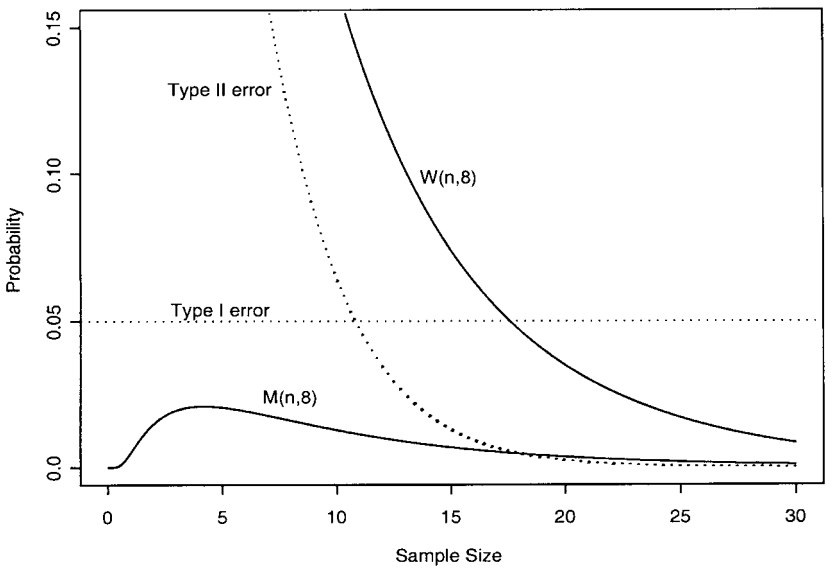


Figure 2. Probabilities of weak and misleading evidence.

On the other hand, the type I error rate in hypothesis testing is held constant as the sample size varies, that is, the above probability $M(n,k)$ always equals $\alpha$. Thus, for a hypothesis test, we can calculate $k_{\alpha,n}$ such that $M(n,k)=\alpha$ demonstrating that the strength of the evidence at which the $\alpha$-sized test rejects $H_0$ varies with $n$ and is given by

$$k_{\alpha,n}=\exp\left\{Z_\alpha\frac{\Delta\sqrt{n}}{\sigma}-\frac{n\Delta^2}{2\sigma^2}\right\} \tag{10}$$

Further, $k_{\alpha,n}$ is often less than one, indicating that a hypothesis test rejects $H_0$ when the evidence favours $H_0$ 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 $\alpha=0.025$ when the two hypotheses of interest are one standard deviation apart, that is, $\Delta=\sigma$. Now the hypothesis test rejects $H_0$ in favour 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 lead to rejection of $H_0$, even though factors greater than 1 indicate support for $H_0$ over $H_1$.

The probability of observing weak evidence and the type II error rate also have similar forms, but again represent different quantities. The probability of observing weak evidence favouring either hypothesis when $H_0$ is the correct hypothesis can be expressed as [1]

$$W(n,k)=P_0\left(\frac{1}{k}<\frac{L_n(\mu_1)}{L_n(\mu_0)}<k\right)=\Phi\left[\frac{\Delta\sqrt{n}}{2\sigma}+\frac{\sigma\ln k}{\Delta\sqrt{n}}\right]-\Phi\left[\frac{\Delta\sqrt{n}}{2\sigma}-\frac{\sigma\ln k}{\Delta\sqrt{n}}\right] \tag{11}$$

Note that this is the same as the corresponding probability when $H_1$ is correct. Figure 2 plots the probability of observing misleading and weak evidence ($M(n,8)$ and $W(n,8)$, respectively) as well as the type I and II error rates, as a function of the sample size. Note that all four

 

curves are quite distinct. Hence the type I error probability is not the probability of observing misleading evidence and the type II error probability is not the probability of observing weak evidence. Note that, as shown in Figure 2, both probabilities of misleading and weak evidence converge to zero as the sample size increases [1].

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 $\min(\alpha+k\beta)$, 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, that is, reject when the likelihood ratio is greater than $k$ [20].

### 3.3. The universal bound

The frequency with which misleading evidence is observed is, in general, low. For any fixed sample size and any pair of probability distributions, the probability of observing misleading evidence of strength $k$ or greater is always less than or equal to $1/k$ [1, 10, 13, 21]. Mathematically, if both $f(X)$ and $g(X)$ are probability density functions and $X$ is distributed according to $f(X)$ then

$$P_f\left(\frac{g(X)}{f(X)} \geqslant k\right) \leqslant \frac{1}{k} \tag{12}$$

This bound has been named the universal bound [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. In fact, the probability of observing very strong misleading evidence (that is, evidence supporting an incorrect hypothesis over the correct hypothesis by a factor of 32 or more) cannot exceed $1/32 = 0.031$.

In the case of multiple looks at the data, the likelihood function is unaffected. Conversely, the probability of observing misleading evidence increases with each look (from two to infinity) *but also remains bounded* by the universal bound (for a proof see Robbins [22]). 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

$$P_f\left(\frac{g(X_n)}{f(X_n)} \geqslant k; \text{ for any } n = 1, 2, \ldots\right) \leqslant \frac{1}{k} \tag{13}$$

The probability of observing misleading evidence remains bounded even though it increases with each look because the amount by which the probability increases converges to zero as the sample size grows [23, 24]. Further, when $f(X_n)$ is the true density, the likelihood ratio itself, $g(X_n)/f(X_n)$, converges to zero as the sample size grows (by the law of large numbers), making it unlikely that it will be greater than $k$ in moderate to large samples.

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 $1 - 1/k$. Practically, this implies that an investigator searching for evidence to support his favourite 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, deliberately or otherwise, to collect strong misleading evidence.

 

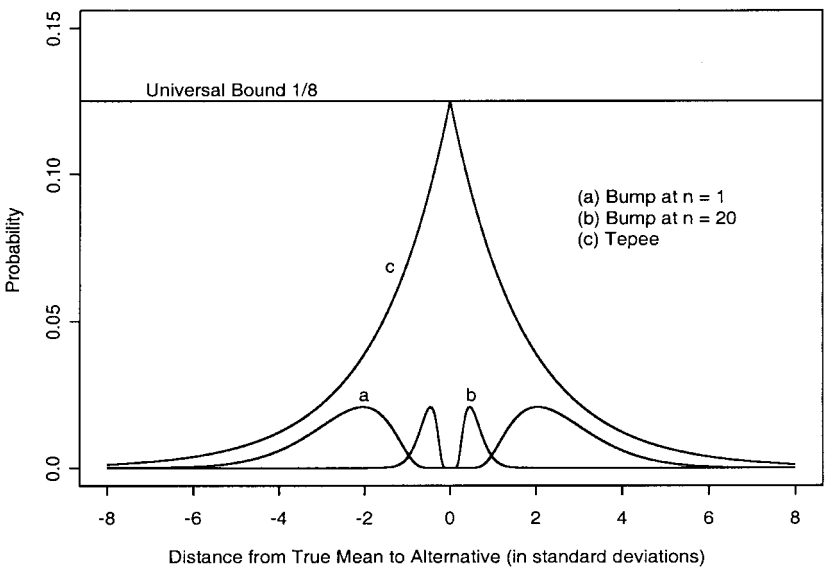


Figure 3. The bump function (at $n = 1, 20$) and tepee function.

### *3.4. In depth: the probability of observing misleading evidence*

The universal bound provides a crude measure of the frequency of observing misleading evidence during either a fixed sample size or sequential study. However the probability of observing misleading evidence rarely, if ever, achieves the universal bound in fixed sample size studies. Here we detail how often misleading evidence is observed when the underlying distribution is normal and note that these results apply in some generality for non-normal distributions when the sample size is large.

Consider generating evidence about the mean of a normally distributed random variable by selecting $n$ observations and examining the likelihood function. That is, $X_1, X_2, \ldots, X_n$ are normally distributed random variables with mean $\mu_0$ and known variance $\sigma^2$. As discussed in Section 3.2, the probability of observing misleading evidence depends on the sample size and the distance between the two fixed simple hypotheses $H_1{:}\mu = \mu_1$ and $H_0{:}\mu = \mu_0$, which was defined as $\Delta = |\mu_1 - \mu_0|$. This probability is given by equation (9) and was examined as a function of $n$ holding $\Delta$ constant.

Consider the reverse situation. Hold the sample size constant at $n$ observations and let the probability be a function of $\Delta = c\sigma$, where $c$ represents the distance between the two hypotheses measured in standard deviations. The resulting probability of observing misleading evidence is

$$P_0\left(\frac{L_n(\mu_1)}{L_n(\mu_0)} \geqslant k\right) = \Phi\left[-\frac{|c|\sqrt{n}}{2} - \frac{\ln k}{|c|\sqrt{n}}\right] \tag{14}$$

This probability has been named the bump function by Royall [10] because of its graphical appearance. Figure 3 graphs this probability with $k = 8$, for $n = 1, 20$ (curves 'a' and 'b', respectively). Note that the $x$-axis is in units of standard deviations. The universal bound of

 

$1/8 = 0.125$ is far greater than the maximum probability of observing misleading evidence, which is given by $\Phi[-\sqrt{(2\ln k)}] = 0.021$ [10].

There essentially is no chance of finding misleading evidence for alternatives near zero, that is, for $\Delta \approx 0$. This happens because $\mu_1$ and $\mu_0$ specify distributions so similar that only very extreme observations will produce strong evidence supporting $\mu_1$ over $\mu_0$ and those extreme observations are improbable under $\mu_0$. 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 $\mu_0$ are more likely to occur, because those observations are not too extreme under $H_0$. After reaching its maximum, the bump function decreases until there is essentially no chance of observing strong misleading evidence for $\mu_1$. At this point, observations which would support $\mu_1$ over $\mu_0$ (for large values of $\Delta$) again are very improbable under $\mu_0$.

For designs that are sequential in nature, a different function represents the probability of observing misleading evidence. Consider the same normal model above, but now our study is designed to continue sampling (possibly forever) until strong evidence for $H_1$ over $H_0$ is obtained. The data are examined after each observation is collected and the study terminates if, and only if, the data support $H_1{:}\mu = \mu_1$ over $H_0{:}\mu = \mu_0$ by a factor of $k$ or more. Even though this design is *severely biased* in favour of $H_1$, the universal bound still applies. The probability that such a biased sequential study design will generate misleading evidence supporting $H_1$ over $H_0$ is approximately

$$P_0\left(\frac{L_n(\mu_1)}{L_n(\mu_0)} \geqslant k;\ \text{for any } n = 1, 2, \ldots\right) \cong \frac{\exp\{-\rho\Delta/\sigma\}}{k} = \frac{\exp\{-\rho c\}}{k} \tag{15}$$

where the subscript on the probability denotes the true mean, $\Delta = |\mu_1 - \mu_0| = c\sigma$, and $\rho \cong 0.583$ is a constant (see references [23, 24] for details).

This probability has been named the tepee function because of its graphical appearance. Figure 3 graphs this probability, equation (15) when $k = 8$ (curve 'c'). Under the sequential design, 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 when $n = 1$. 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 $\mu_1$ over $\mu_0$. 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.

The probability increases steadily as the distance between the two hypotheses approaches zero. One explanation for this is that, as the sample size increases, the probability at each alternative builds up. That is, there exists some probability for each fixed sample size, and this probability is given by the bump function. Alternatives close to $H_0$ accumulate more of this probability because the bump function effectively 'moves inwards toward zero' as the sample size increases. Thus these 'closer' alternatives accumulate a large amount of probability as the entire bump function moves across them. For the same reason, alternatives 'far' from $H_0$ do not accumulate much more probability than what is initially specified on the first observation.

Finally, both the bump function and tepee function apply in moderate to large samples when the underlying distribution is non-normal and there exists a single parameter of interest [10, 23, 24]. The probability of generating misleading evidence in sequential study designs that restrict the sample size are also investigated in references [23] and [24].

 

### 3.5. Nuisance parameters

In multi-parameter models, such as $X_1, X_2, \ldots, X_n$ i.i.d. $f(X_i; \theta, \gamma)$, the Law of Likelihood explains how to measure evidence for the simple hypotheses $H_0{:}(\theta, \gamma) = (\theta_0, \gamma_0)$ *vis-à-vis* $H_1{:}(\theta, \gamma) = (\theta_1, \gamma_1)$. It is the likelihood ratio

$$\prod_{i=1}^{n} \frac{f(X_i; \theta_1, \gamma_1)}{f(X_i; \theta_0, \gamma_0)} = \frac{L_n(\theta_1, \gamma_1)}{L_n(\theta_0, \gamma_0)} \tag{16}$$

which measures the support for one probability distribution, identified by $(\theta_1, \gamma_1)$, versus another, identified by $(\theta_0, \gamma_0)$. Under this model the hypothesis $H_0{:}\theta = \theta_0$ is a composite hypothesis in the following form: $H_0{:}(\theta, \gamma) = (\theta_0, \gamma)$. It is now readily seen that $H_0$ specifies a family of probability distributions because $\gamma$ is unknown. For a fixed value of gamma, say $\gamma_0 = \gamma_1 = \gamma$, the likelihood ratio in equation (16) measures the relative support for $\theta_1$ versus $\theta_0$ but still depends on $\gamma$. Of course, it would be preferable if the likelihood ratio was free of $\gamma$. The problem is that $\gamma$ cannot, in general, be removed to provide a likelihood function of $\theta$ alone [10]. Thus, *ad hoc* solutions are required.

One approach is to use a conditional or marginal likelihood that is free of the nuisance parameter. A common example is the likelihood function for an odds ratio based on the conditional distribution given the margin totals. Another example is the likelihood function for a regression coefficient based on the marginal distribution in a normal linear regression model. Both approaches are appealing because the universal bound, bump and tepee functions all still characterize the probability of observing misleading evidence.

Alternatively one might use a profile or estimated likelihood function. The profile likelihood function maximizes the joint likelihood with respect to the nuisance parameter at each value of the parameter of interest. For fixed $\theta$, the profile likelihood function is defined as $\max_\gamma L_n(\theta, \gamma) = L_n(\theta, \hat{\gamma}(\theta)) = L_{pn}(\theta)$. The approach is to use the profile likelihood function for $\theta$ as if it were a true likelihood function. Along the same line of reasoning, the estimated likelihood function for $\theta$ can be used as if it were a true likelihood function. For fixed $\theta$, the estimated likelihood is defined as $L_n(\theta, \hat{\gamma}_n) = L_{en}(\theta)$ where $\hat{\gamma}_n$ is any consistent estimator of $\gamma$. For example, the overall MLE might be used in place of the nuisance parameter.

If both $\theta$ and $\gamma$ are fixed dimensional parameters and if $f$ is a smooth function, then the profile likelihood will behave like a true likelihood in large samples. That is, for a profile likelihood ratio, the limiting probability of observing misleading evidence is given by the bump function and the tepee function [10, 23, 24]. The same is not true for estimated likelihood ratios, where the probability of observing misleading evidence can be much greater [10, 23, 24].

There are important special circumstances when the likelihood ratio (equation (16)) will be free of $\gamma$. If the likelihood function factors $(L_n(\theta, \gamma) \propto L_{n1}(\theta) L_{n2}(\gamma))$, then the parameters $\theta$ and $\gamma$ are said to be orthogonal [25]. Then the likelihood ratio in equation (16) is only a function of $\theta$ and the support for $\theta_1$ over $\theta_0$ is measured by examining that likelihood ratio. This orthogonalization may sometimes be achieved through reparameterization but is most easily attained, if such a reparameterization exists, through the profile likelihood function [26]. Thus, it is not necessary to determine the reparameterization that orthogonalizes the parameters because the profile likelihood automatically provides the answer.

Of course, profiling does not always work. Profile likelihoods can behave poorly in small samples and situations where the number of parameters is large compared to the sample

 

size [10, 27]. However there are times when they also give very intuitive answers; consider profiling out the variance in a normal model, which yields a profile likelihood for the mean that is proportional to a $t$-distribution (see Appendix B). Furthermore, there are many unexplored adjustments to profile likelihoods that may increase the profile likelihood's applicability in samples of moderate size [28].

## 4. EXAMPLE: THE UNIVERSITY GROUP DIABETES PROGRAM

### *4.1. Background*

The computer code and data required to recreate the likelihood functions presented in this section have been moved to the appendices in order to facilitate ease in reading. The data pertaining to the analysis that follows are included in Appendix E and the S-plus functions for graphing likelihoods are included in Appendix D. It should be noted that the S-plus functions are not specific to this example and may be used whenever data of this type is encountered.

The University Group Diabetes Program (UGDP) was a multi-centred, placebo controlled, randomized, double-masked clinical trial with long-term follow-up. Participants were actively recruited from 1961 to 1966 and followed through 1975. The primary objective of the UGDP was to evaluate the effect of hypoglycaemic agents on vascular complications of adult-onset diabetes. Tolbutamide, a drug that lowers blood glucose levels, was one of four treatments evaluated in the UGDP, but was discontinued ahead of schedule in 1969.

At the time, medical wisdom conjectured that lowering blood glucose levels would lower the eight-year mortality from cardiovascular disease appreciably. The UGDP was designed to examine this relationship. Two of the inclusion criteria for enrolment into the study were a diagnosis of adult-onset diabetes within the last 12 months of enrolment and a life expectancy greater than five years. Smoking history was not collected. Readings of ECGs, fundus photos and X-rays as well as cause of death coding (decided by a central committee) were all completed in a masked manner. Randomization of participants to treatment was stratified by clinic. Balance of treatment assignments across clinics was forced by blocking within each clinic.

Monte Carlo 5 per cent monitoring bounds were used to monitor both overall and cardiovascular mortality over time. No beneficial effect of Tolbutamide with respect to overall mortality was observed ($p > 0.05$ [29]). However, the cardiovascular mortality monitoring bounds suggested that participants on Tolbutamide were at increased risk over the placebo group for mortality from cardiovascular causes ($p < 0.05$ [29]). Fixed sample size chi-square tests were used to assess the overall and cardiovascular mortality difference between the Tolbutamide and placebo groups at $p = 0.17$ and $p = 0.005$, respectively [29]. The overall mortality results, combined with the excess mortality from cardiovascular causes on the Tolbutamide arm, compelled the UGDP investigators to discontinue the use of Tolbutamide:

> "The findings of this study indicate that the combination of diet and Tolbutamide therapy is no more effective than diet alone in prolonging life. Moreover, the findings suggest that Tolbutamide and diet may be less effective than diet alone or than diet and insulin at least in so far as cardiovascular mortality is concerned. For this reason, use of Tolbutamide has been discontinued in the UGDP" [29].

 

When termination of the Tolbutamide arm of the UGDP was disclosed to the medical and scientific community, questions were raised that challenged the findings of the UGDP, its study design, and the competency of the analysis that showed that the increased cardiovascular mortality was attributable to Tolbutamide. Published papers suggested that the difference in cardiovascular death observed between the Tolbutamide and placebo arms was due to some unobserved factor or an imbalance of baseline cardiovascular risk factors between the groups. An extensive list of these criticisms along with a response can be found in references [30] and [31]. (See also references [29, 32–35].) A subsequent analysis of the UGDP mortality results by Cornfield [36] confirmed the UGDP investigators' analysis and showed the Tolbutamide and placebo groups to be adequately balanced on known risk factors.

### *4.2. Early termination of the Tolbutamide arm*

The UGDP investigators observed that participants on Tolbutamide were not benefiting in regard to overall mortality, but were at increased risk for cardiovascular mortality. Because it is unethical to allow a trial to continue for the purpose of demonstrating a harmful effect of treatment, the investigators discontinued the Tolbutamide arm of the trial. The aim of the re-analysis in this section is to measure the statistical evidence about the cardiovascular mortality difference. Specifically, it is of interest to know if the data support the conclusion that the participants on the Tolbutamide arm are at a greater risk of dying from cardiovascular causes than participants on the placebo arm. In the literature, some have suggested this is not the case [33, 34]. Such a finding is especially important if there is no benefit in overall mortality for Tolbutamide over placebo.

Each of the twelve UGDP clinics generated data about cardiovascular death on the Tolbutamide and placebo arms. Table E1 gives the raw data – the number of cardiovascular deaths and participants on each treatment arm by clinic. Under a binomial probability model these data represent evidence about the probability of cardiovascular death on the placebo arm, $\theta_p$, and the probability of cardiovascular death on the Tolbutamide arm, $\theta_t$. (The binomial model was discussed in detail in the biased coin example of Section 2.4.) The dotted lines in Figures 4 and 5 are the clinic-specific likelihood functions, which display the evidence about $\theta_p$ and $\theta_t$, respectively. Notice that when there are no cardiovascular deaths the resulting likelihood function has an interesting shape: it is maximized at $\theta = 0$ and monotonically decreases as $\theta$ increases.

In addition to the clinic-specific likelihood functions, Figures 4 and 5 also display two likelihood functions (the solid lines) that represent the combined evidence about the probability of cardiovascular death. In both figures the thinner of the two likelihood functions represents the evidence under a model that pools all of the data and assumes a homogeneous binomial model for all participants. Under this model (labelled 'a'), the 1/32 SI for $\theta_p$ is 1.9 to 9.9 per cent and for $\theta_t$ is 7.5 to 19.7 per cent. The best supported hypothesis for $\theta_t$ is 13 per cent, which is not in the 1/32 SI for $\theta_p$, suggesting that the probability of a cardiovascular death is different between the two treatment arms.

A potential problem with the homogeneous model is that it ignores any extra variability that may be present due to clinic-to-clinic differences. In fact, the clinic-specific likelihood functions appear more dispersed on the Tolbutamide arm. This extra variation can be accounted for under a beta-binomial model, which is essentially an extended binomial model that allows for correlated participant outcomes. Under the beta-binomial model, the likelihood functions

 

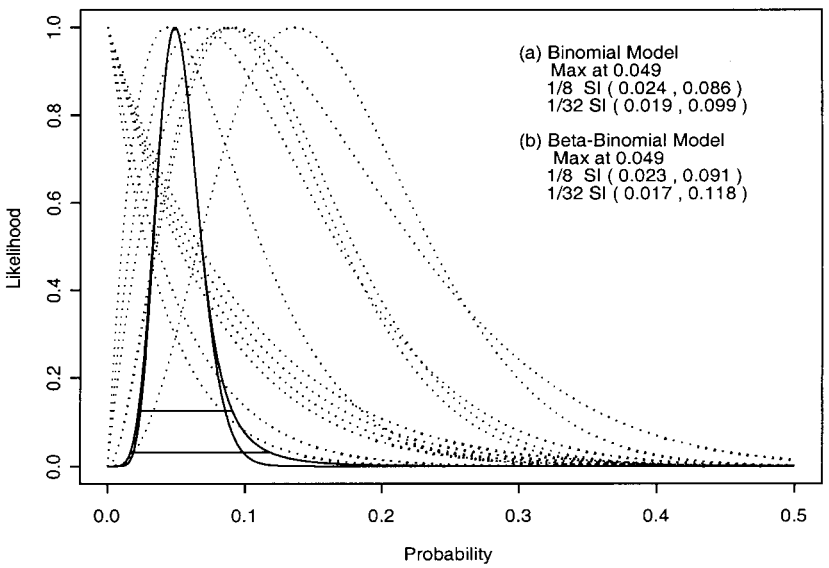


Figure 4. Probability of cardiovascular death (placebo) by clinic and overall.

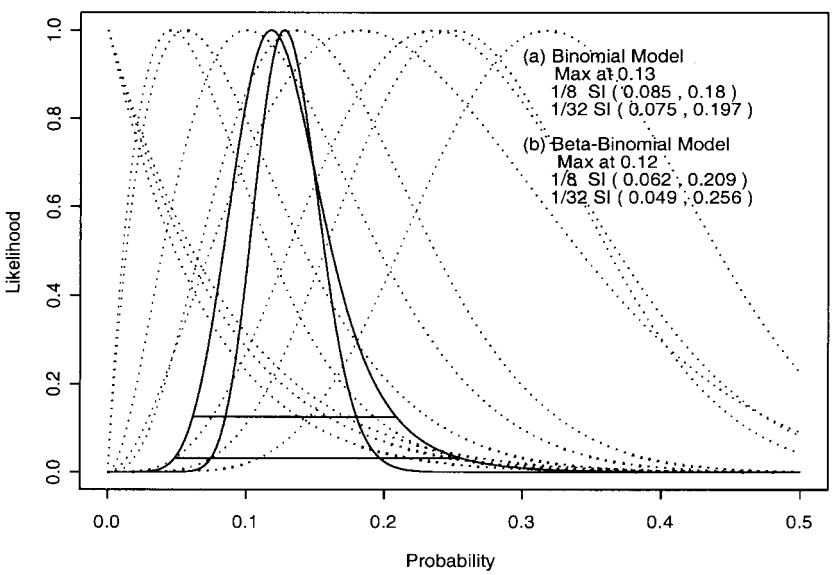


Figure 5. Probability of cardiovascular death (Tolbutamide) by clinic and overall.

representing the combined evidence about the probability of cardiovascular death are wider, albeit only slightly so on the placebo arm. Under the beta-binomial model (labelled 'b'), the 1/32 SI for $\theta_p$ is 1.7 to 11.8 per cent and for $\theta_t$ is 4.9 to 25.6 per cent. There appears to be

 

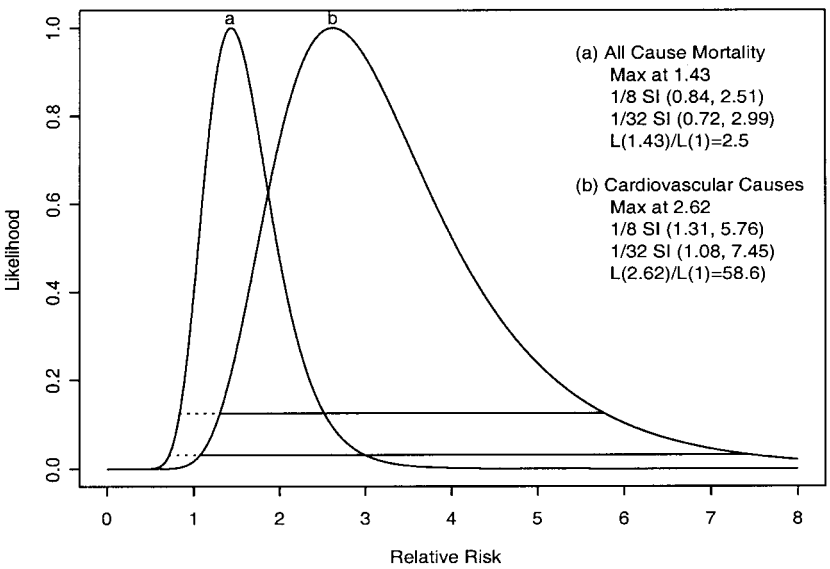


Figure 6. Relative risk of death (Tolbutamide versus placebo).

considerably more variability on the Tolbutamide arm, as evidenced by the noticeably wider beta-binomial likelihood function.

The likelihood function for the beta-binomial model is

$$L(\theta,\phi)=c\left(\sum_i n_i,x\right)\frac{\prod_{i=0}^{x-1}(\theta+\gamma i)\prod_{i=0}^{n-x-1}(1-\theta+\gamma i)}{\prod_{i=0}^{n-1}(1+\gamma i)} \tag{17}$$

where $x$ is the number of success out of $n$ trials, $c(n,x)$ is the usual binomial coefficient, $\theta$ is the probability of success, $\gamma=\phi/(1-\phi)$ and $\phi>0$ is a nuisance parameter that allows for overdispersion [37]. The average number of successes observed under both models is $n\theta$, but the variability is different: $n\theta(1-\theta)(1+(n-1)\phi)$ for the beta-binomial model compared to $n\theta(1-\theta)$ for the binomial model. The presence of $\phi$ is a mixed blessing here. It adds flexibility to the model, but is unknown and must also be estimated. The approach taken here is to eliminate this nuisance parameter by profiling it out of the likelihood function (see Section 3.5). The resulting profile likelihood function for $\theta$ must be evaluated numerically (the S-plus function for doing this is given in Appendix D3).

To compare the two parameters $\theta_t$ and $\theta_p$, it is helpful to examine the evidence for their ratio or difference, or for the odds ratio $\theta_t(1-\theta_p)/\theta_p(1-\theta_t)$. The relative risk, $\theta_t/\theta_p$, is often used to compare two different probabilities and this is the approach taken here. The conditional likelihood function for the relative risk may not be widely known. Appendix C presents a detailed derivation, while Appendix D2 provides the S-plus function for graphing the likelihood function.

Figure 6 displays the likelihood functions for the relative risk (RR) of death from all causes and death from cardiovascular causes, comparing the Tolbutamide group to the placebo group. For the all-cause mortality curve, the best supported hypothesis is RR = 1.43, and is

 

Table III. Relative risk versus hazard rates.

| | Comparison of death risks: Tolbutamide versus placebo | | |
|---|---|---|---|
| | Point estimate | 1/8 SI | 1/32 SI |
| *Cardiovascular causes* | | | |
| Relative risk | 2.62 | 1.31–5.76 | 1.08–7.45 |
| Hazard rate | 2.66 | 1.29–5.94 | 1.05–7.76 |
| *All causes* | | | |
| Relative risk | 1.43 | 0.84–2.51 | 0.72–2.99 |
| Hazard rate | 1.47 | 0.87–2.60 | 0.76–3.14 |

supported over the hypothesis of $RR = 1$ by a factor of only 2.5. Hypotheses that the all cause relative risk is approximately one are consistent with the data. For the cardiovascular causes mortality curve, the best supported hypothesis is $RR = 2.62$. An $RR = 2.62$ is supported over an $RR = 1$ by a factor of 58.6, indicating strong evidence that the relative risk is 2.62 compared to hypotheses suggesting that the relative risk is about 1. The 1/8 and 1/32 SIs are $(1.31, 5.76)$ and $(1.08, 7.45)$, respectively. The 1/32 support interval indicates that the data support an estimated increase in the risk of cardiovascular death of 8 per cent to 645 per cent for the Tolbutamide participants over the Placebo participants. In terms of cardiovascular death, this evidence suggests a harmful effect of Tolbutamide. (Diamond and Forrester [38] present a Bayesian analysis which concludes that the data do not suggest an increase in cardiovascular risk for Tolbutamide participants. Their results vary, however, depending upon the choice of the prior distribution and therefore emphasize the fact that such an analysis is not based solely on 'what the data say'.)

An analysis based on the probability of a cardiovascular death, such as relative risk or odds ratio, assumes a common probability of death for all participants. This assumption is not satisfied in the UGDP because participants have varying follow-up times. However, the study is noticeably balanced with respect to follow-up time, for example, participants were at risk for a total of 5033.6 quarter-years in the placebo group and 4922.2 in the Tolbutamide group with similar balance in various subgroups such as clinic [29, 32]. Alternatively, one could use follow-up times by examining the difference in the hazard rates between the placebo and Tolbutamide groups. The likelihood function for the hazard rate $\lambda$ with $D$ deaths and $P$ person years at risk is proportional to $(\lambda P)^D \exp\{-\lambda P\}$ [1, 39]. Table III compares the two analysis approaches. Because the UGDP is well balanced between treatments in terms of person years, both approaches yield very similar results.

### 4.3. Checking for Imbalance

Can the differential risk of dying from cardiovascular causes be explained by baseline imbalances in the compositions of the treatment groups, as some critics of the UGDP have suggested [33–35]? There is always the possibility that this mortality difference is due to some unknown factor, but the process of randomization provides a reason why this is most likely not the case. Thus, to examine the balance of some covariate across treatments, one compares the proportion of participants with that covariate between the two treatment groups. That is, examine the relative risk for possessing that covariate between Tolbutamide and placebo.

 

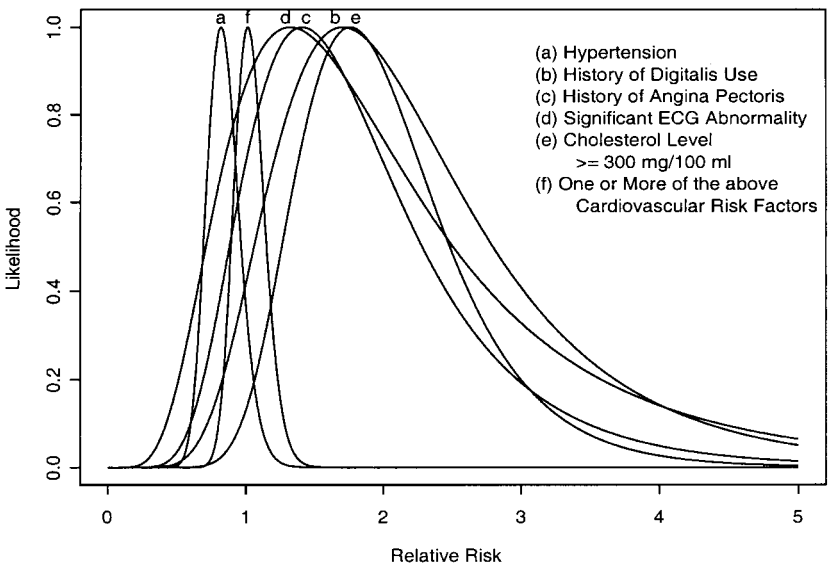


Figure 7. Relative risk of baseline cardiovascular risk factors (Tolbutamide versus placebo).

Examination of the likelihood functions for the relative risk of *six separate baseline cardiovascular risk factors* shows the Tolbutamide and placebo groups to be fairly balanced (Figure 7). The six cardiovascular risk factors are: (a) hypertension; (b) history of digitalis use; (c) history of angina pectoris; (d) significant ECG abnormality; (e) cholesterol level $\geqslant 300$ mg/100 ml; and (f) one or more of the previous cardiovascular risk factors. Hypotheses suggesting that $RR = 1$ are within the 1/8 support interval for each risk factor. The likelihood curve for cholesterol (labelled 'e') appears to favour relative risks greater than one. However, the 1/8 support interval is $(1, 3.24)$ and the 1/32 support interval is $(0.85, 3.91)$, indicating that the evidence is suggestive at best. The best supported hypothesis of $RR = 1.76$ is supported over the hypothesis of $RR = 1$ by a factor of 7.6 ($L(1.76)/L(1) = 7.6$). Thus, the Tolbutamide and placebo groups appear adequately balanced with respect to baseline cardiovascular risk factors. A similar analysis of other baseline risk factors such as fasting blood glucose level $\geqslant 110$ mg/100 ml, relative body weight $\geqslant 1.25$, visual acuity (either eye) $\leqslant 20/200$, serum creatinine level $\geqslant 1.5$ mg/100 ml, arterial calcification, age $> 53$ years (median age), gender, and race indicate good balance on these factors as well. All of the 1/8 support intervals contain hypotheses suggesting $RR = 1$, including 1 itself.

A logistic regression model can be used to estimate each participant's risk of cardiovascular death. This estimate of risk adjusts for the various factors included in the logistic model. Using this approach, the UGDP study investigators modelled the probability of a cardiovascular death with 17 covariates [29]. These covariates included the 13 baseline cardiovascular, medical and demographic risk factors examined in this paper along with a single indicator variable for each of the four treatment groups in the UGDP (two of which are not considered in this paper). Cornfield used the logistic regression model (setting all four treatment indicators to zero) to tabulate participants by their estimated probability of cardiovascular risk at baseline (reference [36], Table E5).

 

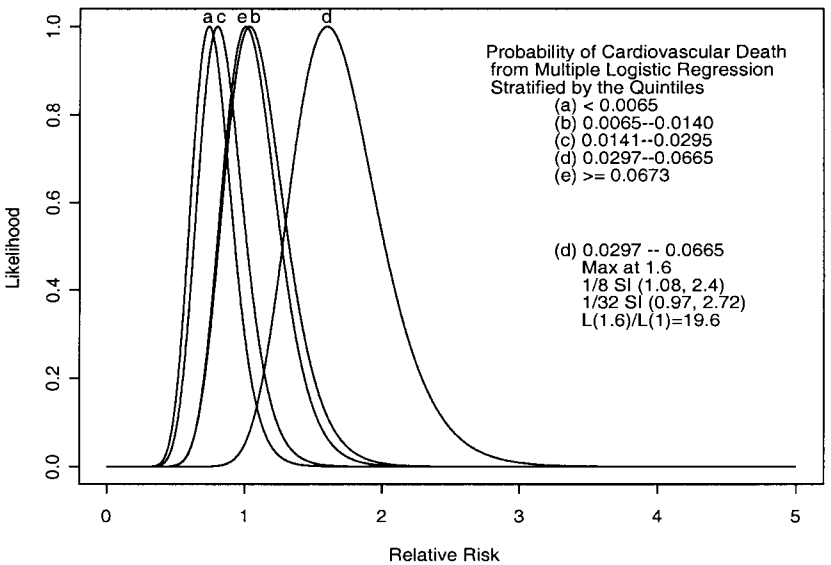


Figure 8. Relative risk of adjusted probability cardiovascular death (Tolbutamide versus placebo).

Categories are determined by the estimated quintiles of the risk distribution for all 823 participants (using participants from all treatments at baseline). Figure 8 displays the likelihood functions for relative risk *of the estimated baseline risk of cardiovascular death* for the Tolbutamide versus placebo group. For each curve except 'd', the hypotheses suggesting that the relative risk is near 1, including 1 itself, are an element of the 1/8 support intervals. Thus the data support that the Tolbutamide and placebo groups are adequately balanced on the proportion of participants who are at those levels of risk.

There is moderate evidence that the Tolbutamide group has a higher proportion of participants with an estimated risk between 0.0297 and 0.0665 (4th quintile). For the curve labelled 'd' in Figure 8, hypotheses suggesting the relative risk is near 1, including 1 itself, are an element of the 1/32 SI $(0.97, 2.79)$ but not nearly as many are members of the 1/8 SI $(1.08, 2.41)$. This indicates that the hypothesis $\mathrm{RR} = 1$ is consistent with the data but only at a moderate level. In fact the best supported hypothesis is $\mathrm{RR} = (54/204)/(34/205) = 1.596$ indicating that the Tolbutamide group had roughly 60 per cent more participants with this level of cardiovascular risk than the placebo group. The likelihood ratio for the best supported hypothesis to the hypothesis of $\mathrm{RR} = 1$ is 19.6 $(L(1.6)/L(1) = 19.6)$, indicating moderately strong support for the best supported hypothesis over the hypothesis of equal risk.

On this topic Cornfield noted:

> "The average number of risk factors present among those assigned to placebo was 1.65 as compared with 1.92 among those receiving Tolbutamide, an excess of about one-fourth a risk factor. All in all, the luck of the draw does not seem to have been too bad" [36].

The likelihood analysis supports this conclusion.

 

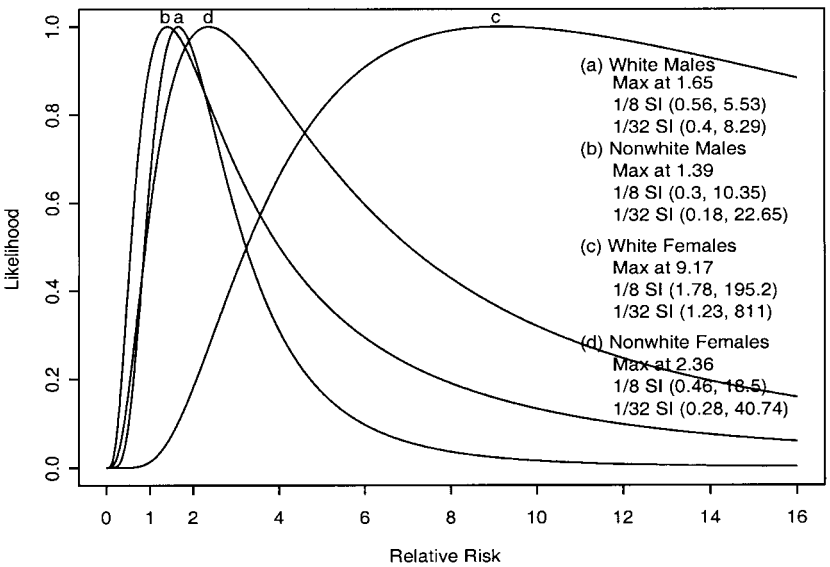


Figure 9. Relative risk of death from cardiovascular causes (Tolbutamide versus placebo).

### *4.4. Analyses within subgroups*

A subgroup is a subset of the study population distinguished by a particular characteristic or set of characteristics [30]. Subgroups are subject to the same limitations and retain the same randomization properties as any post-stratification. However, the term subgroup is usually reserved for a homogenous post-stratification with a small sample size. As the number of strata increases, the number of participants in each stratum decreases. This leads to the so-called subgroup inference problem, because the number of participants in each subgroup often is too small to make reliable inferences regarding that particular subgroup. The subgroup inference problem is often erroneously cited as a consequence of the total number of strata (that is, groups) and can be avoided with careful planning, so that large enough sample sizes are achieved in relevant subgroups.

Likelihood based methods permit the construction of any number of subgroups and allow for the evaluation of their evidence without adjustment. (Note, however, that their scientific validity still depends on the characteristics used to define the subgroup.) Even small subgroups provide statistical evidence. The width of the likelihood function, which depends upon the sample size in the subgroup, will indicate the proper amount of variability in the evidence. The law of large numbers implies that, with a large enough sample size in each stratum, inferences about an unknown strata-specific parameter will be correct regardless of the total number of strata. Therefore, in subgroup analyses, concern should be for the size of the subgroup and not the total number of subgroups.

As an example, examine the data in the four different race (white versus non-white) by gender subgroups. Figure 9 presents the evidence about the relative risk for cardiovascular death in these four subgroups and is essentially a picture of the gender by race by treatment interaction (see Table E3, Appendix E). There is very strong evidence for increased risk

 

of cardiovascular mortality for white women on Tolbutamide over white women on placebo RR = 9.17, $L(9.17)/L(1) = 80$, 1/32 SI (1.23, 811). The data suggest that white women on Tolbutamide have over nine times the risk of cardiovascular death than white women on placebo. There is only moderate evidence, however, for increased risk for overall mortality for white women RR = 3.57, $L(3.57)/L(1) = 16.9$, 1/32 SI (0.9, 28.71). In the other three subgroups, participants on Tolbutamide do not appear to be at any increased risk of cardiovascular death compared to participants on placebo.

This result appears to be new. Kilo *et al*. [40, 41] noted that in the placebo group the risk ratio of cardiovascular death for males to females was 5.3:1 (11.1 per cent for males versus 2.1 per cent for females). They argue that this 'anomalously' low cardiovascular death risk for women is uncharacteristic for female diabetics and, as a consequence, causes the corresponding Tolbutamide risk to appear high in comparison. The likelihood analysis does not support this argument because only the white women are at increased risk. Furthermore, the risk ratio of cardiovascular death in the placebo group for white males to non-white males to white females to non-white females is 4.63:5.43:0.78:1 (10.64 per cent:12.5 per cent:1.8 per cent:2.3 per cent). Thus, the 'anomalously' low cardiovascular death risk for women cannot be the culprit because the non-white women, who have a similarly low cardiovascular death risk when compared to the males, are not at increased risk for cardiovascular death in the Tolbutamide group, 1/32 SI (0.28, 40.74).

From a medical perspective, such a gender by race by treatment interaction hardly seems plausible. Yet the evidence suggests that such an interaction exists. While the evidence indicates such a difference, the data may not represent sufficiently strong evidence to modify peoples' beliefs about the existence of such a gender by race by treatment interaction. Regardless of one's prior disposition however, this evidence indicates that a dangerous trend existed in the UGDP participant population. The Data Safety Monitoring Committee (DSMC) would have had to monitor the situation carefully, had it been allowed to continue. An investigation as to how white women in this trial differed from other participants might have proved useful. One interesting possibility might have been for the DSMC to have recommended the discontinuation of Tolbutamide for white women only.

Overall, this analysis of the UGDP data indicates an increased risk of cardiovascular death for participants on Tolbutamide, while demonstrating no imbalance beyond that expected by randomization on relevant cardiovascular risk factors between the Tolbutamide and placebo groups. Furthermore the excess cardiovascular risk seems only to apply to white women.

## 5. COMMENTS

The Law of Likelihood explains how to objectively measure the strength of evidence for one hypothesis over another. Its dependence on two simple hypotheses is not a flaw, but a strength. Consequently, arbitrary methods of summarizing evidence over a composite space are avoided. Further, the probability of observing misleading evidence is naturally controlled, even with multiple looks at the data. This makes the probability of generating weak evidence the relevant quantity in sample size calculations. More importantly, the measure of the strength of evidence is mathematically separated from the probability of observing such evidence, allowing each quantity to be dealt with in its distinct role.

 

A strong case can be argued for monitoring clinical trials with likelihood functions. Constant monitoring of statistical evidence does not diminish its strength because the likelihood is unaffected by the number of examinations. The data speak for themselves, through the likelihood function, independently of the probability of observing misleading evidence. The probability of observing misleading evidence of $k$-strength or greater remains bounded by $1/k$ for any number of looks under very general conditions. Furthermore, in common situations the probability is often much less than the universal bound. However, because this probability is irrelevant in the analysis stage, the Data Safety Monitoring Committee may constantly monitor a clinical trial by looking at a repeatedly likelihood function without fear of causing or observing a spurious association.

## APPENDIX A: TRANSFORMING COMPOSITE HYPOTHESES INTO SIMPLE ONES

A Bayesian analysis essentially transforms the composite hypotheses into simple ones and then applies the Law of Likelihood to the resulting simple hypotheses. This is done by changing the probability model. To illustrate, suppose our probability model is $X \sim f(\mathbf{X}|\theta)$ where $\theta \in \Theta$ is a real-valued parameter of interest (in the biased coin example $f(\mathbf{X}|\theta)$ was a binomial distribution and $\Theta = [0,1]$). Identify $H_s{:}\theta = \theta_s$ as a simple hypothesis and $H_c{:}\theta \in \Theta_c \subseteq \Theta$ as a composite hypothesis. Upon observing $\mathbf{X} = \mathbf{x}$, the Law of Likelihood explains that the measure of evidence for $H_c$ over $H_s$ is the likelihood ratio given by $P_c(\mathbf{x})/P_s(\mathbf{x})$. However this is the problem because $P_c(\mathbf{x}) = f(\mathbf{X} = \mathbf{x}|\theta \in \Theta_c)$ does not exist (the probability model $f$ is indexed by a fixed parameter $\theta$, not a range of $\theta$s).

Assuming a prior or weighting distribution $g(\theta)$ and using $P_c(\mathbf{x}) = \int_{\Theta_c} f(\mathbf{x}|\theta) g(\theta)\, d\theta$ is equivalent to applying the Law of likelihood under a *different* probability model. Initially $\mathbf{X} \sim f(\mathbf{X}|\theta)$ but after specifying $g(\theta)$ the probability model is changed to $\mathbf{X} \sim h(\mathbf{X}) = \int_{\Theta} f(\mathbf{X}|\theta)\, g(\theta)\, d\theta$. Consequently, $P_c(\mathbf{X})$ is a simple hypothesis and the ratio $P_c(\mathbf{x})/P_s(\mathbf{x})$ measures the evidence for $H_c$ over $H_s$ under the model $\mathbf{X} \sim h(\mathbf{X})$. (Note that $P_c(\mathbf{x})/P_s(\mathbf{x})$ is called a Bayes factor.) Once again, the purported measure of statistical evidence depends upon the unspecified weighting distribution $g(\theta)$.

## APPENDIX B: PROFILE LIKELIHOOD FOR A MEAN

Suppose $X_1, X_2, \ldots, X_n \sim$ i.i.d. $N(\theta, \sigma^2)$. The profile likelihood function for $\theta$ is $\max_{\sigma^2} L(\theta, \sigma^2) = L(\theta, \widehat{\sigma^2}(\theta))$ where $\widehat{\sigma^2}(\theta) = \sum_{i=1}^{n} (X_i - \theta)^2/n$. The following algebra shows that the profile likelihood is proportional to the likelihood function for a mean under the Student's $t$ distribution.

$$L(\theta, \widehat{\sigma^2}(\theta)) \propto \frac{1}{[\widehat{\sigma^2}(\theta)]^{\frac{n}{2}}} \exp\left\{ -\frac{1}{2} \frac{\sum_{i=1}^{n} (X_i - \theta)^2}{\widehat{\sigma^2}(\theta)} \right\}$$

$$= \exp\left\{ -\frac{n}{2} \right\} \left[ \frac{\sum_{i=1}^{n} (X_i - \theta)^2}{n} \right]^{-\frac{n}{2}}$$

 

$$\propto \left[\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n}+(\bar{X}-\theta)^2\right]^{-\frac{n}{2}}$$

$$=\left[1+\frac{n(\bar{X}-\theta)^2}{(n-1)\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}}\right]^{-\frac{n}{2}}$$

$$=\left[1+\frac{t^2}{v}\right]^{-\frac{v+1}{2}}$$

where $v=n-1$, $s^2=\sum_{i=1}^{n}(X_i-\bar{X})^2/(n-1)$ and $t=\sqrt{n}(\bar{X}-\theta)/s$.

Royall (reference [1], pp. 133–134) suggests modifying the exponent of the above profile likelihood by replacing $n$ with $n-1$, so that for small sample sizes, the probability of observing misleading evidence remains below the maximum of the bump function.

## APPENDIX C: CONDITIONAL LIKELIHOOD FOR THE RELATIVE RISK

This derivation follows Royall (reference [1], p. 165). The likelihood function for the relative risk is derived in the same fashion as the conditional likelihood function for the odds ratio (that is, we condition on the margin totals). However, the underlying distributions of events observed in the two groups are assumed to be negative binomial rather than binomial. The unconditional joint likelihoods are the same regardless of the underlying assumption. Furthermore, the likelihood principle implies that the choice of the underlying distribution (in this case binomial or negative binomial) is immaterial because the likelihood ratios are the same for both distributions [1, 14]. To rid ourselves of the nuisance parameter in the unconditional joint likelihood, we condition on the total number of successes which yields the conditional likelihood function for the relative risk. Note, however, that the conditional joint likelihoods are different (one is a function of the odds ratio and the other is a function of the relative risk) depending on the parameter of interest. Consider the following $2\times 2$ table:

| | Events | | Total number of observations |
|---|---|---|---|
| | Successes | Failures | |
| Group 1 | $a$ | $b$ | $m$ |
| Group 2 | $c$ | $d$ | $n$ |

Of interest is the RR$=(a/m)/(c/n)$, where $m$ is the total number of i.i.d. Bernoulli($\theta_1$) observations required to obtain $b$ failures and $n$ is the total number of i.i.d. Bernoulli($\theta_2$) observations required to obtain $d$ failures. Let $\gamma=\theta_1/\theta_2$ be the relative risk for a successful event in group 1 versus group 2.

 

Under the above set-up we have that, $M \sim$ negative binomial$(b, 1-\theta_1)$ and $N \sim$ negative binomial$(d, 1-\theta_2)$ where

$$P(M=m)=\binom{m-1}{b-1}(1-\theta_1)^b\theta_1^{m-b} \quad \text{and} \quad P(N=n)=\binom{n-1}{d-1}(1-\theta_2)^d\theta_2^{n-d}$$

Then the likelihood function for the relative risk, $\gamma$, is

$$L(\gamma) \propto \left[\sum_{j=b}^{m+n-d}\binom{m+n-j-1}{d-1}\binom{j-1}{b-1}\gamma^{j-m}\right]^{-1}$$

*Proof*

$$\begin{aligned}
L(\gamma) &= P(M=m|M+N=m+n)=\frac{P(M=m, M+N=m+n)}{P(M+N=m+n)} \\
&= \frac{P(M=m)P(N=n)}{\sum_j P(M+N=m+n|M=j)P(M=j)} \\
&= \frac{\binom{m-1}{b-1}(1-\theta_1)^b\theta_1^{m-b}\binom{n-1}{d-1}(1-\theta_2)^d\theta_2^{n-d}}{\sum_{j=b}^{m+n-d}\binom{m+n-j-1}{d-1}(1-\theta_2)^d\theta_2^{m+n-j-d}\binom{j-1}{b-1}(1-\theta_1)^b\theta_1^{j-b}} \\
&\propto \left[\sum_{j=b}^{m+n-d}\binom{m+n-j-1}{d-1}\binom{j-1}{b-1}\gamma^{j-m}\right]^{-1}
\end{aligned}$$

QED

## APPENDIX D: COMPUTER CODE

The likelihood functions presented here were drawn and analysed in the statistical software package S-plus. S-plus can be somewhat cumbersome for programming, so I have included functions that will automatically draw the likelihood functions and give their MLE, 1/8 and 1/32 support intervals. To import these functions, copy and paste them into the command window at the command prompt '>'. These functions and others are available in electronic form from the author.

I have included three functions: bin.lik for binomial likelihoods, betabin.lik for profile beta-binomial likelihoods, and rr.lik for conditional likelihoods for the relative risk. All the functions have some common input variables: hilim is the upper limit of the X-axis; lolim is the lower limit of the X-axis; acc is a multiplier that can increase or decrease the number of points used to evaluate the likelihood; main1 is the title of the graph; and like.only suppresses the command for a new plot window and draws the likelihood on the existing plot. An example of how to call each S-plus function is given along with the function.

 

*D1. Binomial Likelihood*

Figure 1 displays the likelihood function for the probability of success, when 14 successes out of 50 trials were observed. To produce Figure 1 type

```
bin.lik(x=14,n=50)
```

at the command prompt. The function bin.lik is given below.

```
bin.lik_function(x,n,like.only=F,acc = 1, lolim=0,hilim=1,
       main1 =  "Likelihoods:  Binomial Model") {

p <- seq(lolim, hilim, length = 1000 * acc)
like <- exp(x*log(p) + (n-x) * log(1 - p))
like <- like/max(like)
if (like.only==F) {
      graphsheet(orientation = "portrait", height = 6, width = 7.5,
      pointsize = 12)
      plot(p, like, type = "n", xlab = "Probability", ylab = " ",
      main = main1)}
p1 <- p[like >= 1/8]
p2 <- p[like >= 1/32]
i1 <- rep(1/8, length(p1))
i2 <- rep(1/32, length(p2))

lines(p,like,type="l")
lines(p1, i1, type = "l")
lines(p2, i2, type = "l")

if (like.only==F) {
      whr <- if(p[like == max(like)] <= (lolim + hilim
             )/2) quantile(p, 0.8) else quantile(p, 0.2)
      text(whr, 0.95, paste("Max at", signif(c(p[like ==
             max(like)]), digits = 2)),cex=.8)
      text(whr, 0.91, paste("1/8 SI (", round(min(p1),
             digits = 2), ",", round(max(p1), digits = 2),
             ")"),cex=.8)
      text(whr, 0.87, paste("1/32 SI (", round(min(p2),
             digits = 2), ",", round(max(p2), digits = 2),
             ")"),cex=.8)}
}
```

*D2. Conditional Likelihood for Relative Risk*

Figure 6 displays the likelihood function for the relative risk of cardiovascular death. Here there were 26 deaths out of 204 participants on in the treatment group and 10 deaths out of 205 participants in the placebo group (see Table E2 in Appendix E). To produce Figure 6 type

```
rr.lik(x=26,m=204,y=10,n=205,hilim=8)
```

 

the command prompt. The likelihood for the relative risk of all causes can be added by typing

```
rr.lik(x=30,m=204,y=21,n=205,like.only=T,hilim=8)
```

at the next command prompt. The function `rr.lik` is given below.

```
rr.lik_function(x,m,y,n,lolim=0,hilim=6, like.only=F,acc=1,
main1="Conditional Likelihood: Relative Risk  (Probability Ratio)"){

## conditional likelihood on total number of successes; negative Binomial
## x successes out of m trials and y successes out of n trials
## lolim is the lower limit of the relative risk axis
## hilim is the lower limit of the relative risk axis

if ((m - x) * (n - y) == 0)
     "Conditional model requires at least one failure in each group."
else {
      z_seq(lolim, hilim, len = 1000 * acc)
      like_matrix(1,nrow=length(z), ncol=1)
      j_seq(m-x,m+y,1) ## summation index
      Mx_matrix(z, nrow = length(j), ncol = length(z), byrow=T)
      Mx_Mx^(j - m)
      co_choose(m+n-j-1,(n-y)-1)*choose(j-1,(m-x)-1)
      like_1/(t(Mx) %*% co)

      if(y== 0) like_like * exp(co[length(j)])
      else like_like/max(like)

rrhat_z[like == max(like)]
rrhat_round(rrhat, digits = 2)

if(like[length(z)] == max(like)) rrhat_NA
if(like [1] == max(like)) rrhat_NA

L1_round(min(max(like)/like[abs(z - 1) == min(abs(z - 1))],1000000),
digits = 1)

z2_z[like >= 1/8]
z3_z[like >= 1/32]
i2_rep(1/8, length(z2))
i3_rep(1/32, length(z3))

if (like.only==F)
{graphsheet(orientation = "portrait", height = 6, width = 7.5, pointsize = 12)
plot(z, like, type = "n", xlab = "Relative Risk ", ylab ="Likelihood",
main=mainl) }
```

 

```
lines(z, like, lty=1,col=1)

if (like.only==F) {
lines(z2, i2, type = "l",col=1)
lines(z3, i3, type = "l",col=1)
text(z[.85*1000*acc], 0.94, paste("Max at ", rrhat),cex=0.8)

text(z[.85*1000*acc], 0.9, paste("1/8 SI (",
if(min(z2) == z[1])  "NA" else round(min(z2), digits = 2), ",",
if(max(z2) == z[1000]) "NA" else round(max(z2), digits = 2), ")"),cex=0.8)

text(z[.85*1000*acc], 0.86, paste("1/32 SI (",
if(min(z3) == z[1])  "NA" else round(min(z3), digits =2), ",",
if(max(z3) == z[1000]) "NA" else round(max(z3), digits = 2), ")"),cex=0.8)

text(z[.85*1000*acc], 0.81, paste("L(",  rrhat, ")/L(1)=", L1),cex=0.8)}
}
}
```

*D3. Beta-Binomial Likelihood*

Figures 4 and 5 display the likelihood functions for the probability of cardiovascular death under placebo and Tolbutamide, respectively. This function is slightly more complicated than the previous two because it takes a data matrix as input. The data matrix has two columns: in the first column is the number of successes and the second column is the total number of trials. The number of rows equals the number of different experiments, in this case centres. For the raw data see Table E1 in Appendix E. To produce a data matrix for Figure 4 define a vector for the number of deaths at each clinic for the placebo group as

```
cvd.pl_c(1, 2, 3, 1, 2, 0, 0, 0, 1, 0, 0, 0)
```

and the same for the number of participants at each clinic for the placebo group

```
cvn.p1_c(15, 22, 22, 23, 23, 19, 24, 13, 11, 10, 12, 11)
```

Be sure to keep clinics in the same position across the vectors. To produce Figure 6 type

```
betabin.lik(cbind(cvd.pl,cvn.pl),simple.model=T,twostage.model=T,
hilim=0.5,acc=10)
```

The function betabin.lik is given below.

```
betabin.lik_function(x, lolim = 0,  hilim = 1, simple.model = F,
        twostage.model = F, separate.lines = T,
        main1 =  "Likelihoods: Binomial-Beta Model\n", acc = 1) {

#   FUNCTION: betabin.lik
# Original author: Richard Royall -- June 24/96
#
```

 

```
#   Data matrix x : k successes (col 1) in m trials (col 2)
#   Binomial-beta model: Xi is binom(mi,pi). p's iid B(a,b)
#    E(X/k) =a/(a+b) (par of interest), gamma(nuisance)=a+b
#
#   Plots LFs for E(success proportion) under various models:
#    (1) Each proportion Binomial with its own probability (pi)
#      (separate.lines=T)
#    (2) Each proportion Binomial with common probability (pi=p)
#      (simple.model=T)
#    (3) PROFILE LF under two-stage (Binomial-beta) model
#           (with 1/8 and 1/32 support intervals)
#      (twostage.model=T)
#
     p <- seq(lolim, hilim, length = 100 * acc)
     pl <- seq(lolim, hilim, length = 1000)
     LM <- NULL
     for(i in 1:nrow(x)) {
           lik <- dbinom(x[i, 1],x[i, 2], pl)
           lik <- lik/max(lik)
           LM <- cbind(LM, lik)
     }
     matplot(pl, LM, type = "n", xlab = "Probability", ylab = "Likelihood",
             main = main1, cex=1,yaxt="n")
     axis(side=2,at=seq(0,1,.2),cex=1,srt=90,adj=0.5,mgp=0)
     if(twostage.model == T) {
           pb <- p[2:(length(p) - 1)]
           llikpro <- rep(0, length(pb))
           for(i in 1:length(pb)) {
                 g <- 100000
                 alpha <- pb[i] * g
                 beta <- (1 - pb[i]) * g
                 llikproi <- sum(lgamma(alpha + x[, 1]) +
                       lgamma(beta + x[, 2] - x[, 1]) -
                       lgamma(alpha + beta + x[, 2]) -
                       lgamma(alpha) - lgamma(beta) + lgamma(
                       alpha + beta))
                 repeat {
                       newg <- g/10
                       newalpha <- pb[i] * newg
                       newbeta <- (1 - pb[i]) * newg
                       newllikproi <- sum(lgamma(newalpha +
                         x[, 1]) + lgamma(newbeta + x[, 2] -
                         x[, 1]) - lgamma(newalpha + newbeta +
                         x[, 2]) - lgamma(newalpha) - lgamma(
                         newbeta) + lgamma(newalpha +
                         newbeta))
```

 

```
      if(newllikproi > llikproi) {
        llikproi <- newllikproi
        g <- newg
      }
      else {
        llikproi <- newllikproi
        g <- newg
        break
      }
}
repeat {
       newg <- g * 10^0.2
       newalpha <- pb[i] * newg
       newbeta <- (1 - pb[i]) * newg
       newllikproi <- sum(lgamma(newalpha +
         x[, 1]) + lgamma(newbeta + x[, 2] -
         x[, 1]) - lgamma(newalpha + newbeta +
         x[, 2]) - lgamma(newalpha) - lgamma(
         newbeta) + lgamma(newalpha +
         newbeta))
       if(newllikproi > llikproi) {
         llikproi <- newllikproi
         g <- newg
       }
       else {
         llikproi <- newllikproi
         g <- newg
         break
       }
 }
      repeat {
       newg <- g * 10^(-0.04)
       newalpha <- pb[i] * newg
       newbeta <- (1 - pb[i]) * newg
       newllikproi <- sum(lgamma(newalpha +
         x[, 1]) + lgamma(newbeta + x[, 2] -
         x[, 1]) - lgamma(newalpha + newbeta +
         x[, 2]) - lgamma(newalpha) - lgamma(
         newbeta) + lgamma(newalpha +
         newbeta))
       if(newllikproi > llikproi) {
         llikproi <- newllikproi
         g <- newg
       }
       else break
}
```

 

```
                  llikpro[i] <- llikproi
            }
###(END OF "for i")###
            likpro <- exp(llikpro - max(llikpro))
            lines(pb, likpro, lty =1)
            p1 <- pb[likpro >= 1/8]
            p2 <- pb[likpro >= 1/32]
            i1 <- rep(1/8, length(p1))
            i2 <- rep(1/32, length(p2))
            lines(p1, i1, type = "l")
            lines(p2, i2, type = "l")
            whr <- if(pb[likpro == max(likpro)] <= (lolim + hilim
                  )/2) quantile(p, 0.8) else quantile(p, 0.2)
            text(whr-0.05, 0.95, paste("Beta-Binomial Model"),cex=0.8,adj=-1)
            text(whr-0.05, 0.91, paste ("Max at", signif(c(pb[likpro ==
                  max(likpro)]), digits = 2)),cex=0.8,adj=-1)
            text(whr-0.05, 0.87, paste("1/8 SI (", round(min(p1),
                  digits = 3), ",", round(max(p1), digits = 3),
                  ")"), cex=0.8,adj=-1)
            text(whr-0.05, 0.83, paste("1/32 SI (", round(min(p2),
                  digits = 3), ",",  round(max(p2), digits = 3),
                  ")"),cex=0.8,adj=-1)
       }
###(END OF  "if twostage.model==T")###
       if (simple.model == T)  {
            likc <- dbinom(sum(x[, 1]),sum(x[, 2]), pl)
            likc <- likc/max(likc)
            lines(pl, likc, lty = 1)
            p1 <- pl[likc >= 1/8]
            p2 <- pl[likc >= 1/32]
            i1 <- rep(1/8,  length(p1))
            i2 <- rep(1/32, length(p2))
            lines(p1, i1, type = "l")
            lines(p2, i2, type = "l")

if(twostage.model==F) {
            whr <- if(pl[likc == max(likc)] <= (lolim + hilim
                  )/2) quantile(p, 0.8) else quantile(p, 0.2)}
            text(whr-0.05, 0.75, paste("Binomial Model"),cex=0.8,adj=-1)
            text(whr-0.05, 0.71, paste("Max at", signif(c(pl[likc ==
                  max(likc)]), digits = 2)),cex=0.8,adj=-1)
            text(whr-0.05, 0.67, paste("1/8 SI (", round(min(p1),
                  digits = 3), ",", round(max(p1), digits = 3),
                  ")"),cex=0.8,adj=-1)
            text(whr-0.05, 0.63, paste("1/32 SI (", round(min(p2),
                  digits = 3), ",", round(max(p2), digits = 3),
```

 

```
                ")"),cex=0.8,adj=-1)
    }
if(separate.lines == T) {
      matlines(pl, LM, type = "l", lty = 2, col = 1, cex = 1)
    }
}
```

## APPENDIX E: DATA

The data assembled in the following tables (Tables E1–E5) can be found in references [29, 32, 36]

Table E1. Distribution of cardiovascular deaths by clinic.

| Clinic | Tolbutamide | | Placebo | |
|---|---|---|---|---|
| | Deaths | Total participants | Deaths | Total participants |
| Baltimore | 1 | 22 | 0 | 24 |
| Cincinnati | 7 | 22 | 2 | 23 |
| Cleveland | 1 | 18 | 0 | 19 |
| Minneapolis | 6 | 24 | 2 | 22 |
| New York | 2 | 20 | 3 | 22 |
| Williamson | 3 | 22 | 1 | 23 |
| Birmingham | 2 | 11 | 0 | 13 |
| Boston | 4 | 17 | 1 | 15 |
| Chicago | 0 | 12 | 1 | 11 |
| St Louis | 0 | 11 | 0 | 10 |
| San Juan | 0 | 14 | 0 | 12 |
| Seattle | 0 | 11 | 0 | 11 |
| All clinics | 26 | 204 | 10 | 205 |

Table E2. Distribution of deaths.

| | Tolbutamide | | Placebo | |
|---|---|---|---|---|
| | Deaths | Total participants | Deaths | Total participants |
| All causes | 30 | 204 | 21 | 205 |
| Cardiovascular causes | 26 | 204 | 10 | 205 |

Table E3. Distribution of deaths by gender and race.

| | Tolbutamide | | | Placebo | | |
|---|---|---|---|---|---|---|
| | Deaths (by cause) | | Total participants | Deaths (by cause) | | Total participants |
| | All | Cardiovascular | | All | Cardiovascular | |
| White males | 9 | 7 | 40 | 8 | 5 | 47 |
| Non-white males | 4 | 4 | 23 | 5 | 2 | 16 |
| White females | 13 | 11 | 68 | 3 | 1 | 56 |
| Non-white females | 4 | 4 | 73 | 5 | 2 | 86 |

 

Table E4. Distribution of baseline cardiovascular risk factors.

| Baseline cardiovascular risk factor | Tolbutamide | | Placebo | |
|---|---|---|---|---|
| | Number with risk factor | Total participants | Number with risk factor | Total participants |
| (a) Hypertension | 60 | 199 | 74 | 201 |
| (b) History of digitalis use | 15 | 198 | 9 | 202 |
| (c) History of angina pectoris | 14 | 201 | 10 | 202 |
| (d) Significant ECG abnormality | 8 | 201 | 6 | 199 |
| (e) Cholesterol level $\geqslant$ 300 mg/100 ml | 30 | 199 | 17 | 198 |
| (f) One or more of the above risk factors | 92 | 192 | 88 | 186 |

Table E5. Distribution of the probability of cardiovascular death based on a multiple logistic regression.

| Probability of cardiovascular death | Tolbutamide | | Placebo | |
|---|---|---|---|---|
| | Number in quintile | Total participants | Number in quintile | Total participants |
| (a) $<$ 0.0065 | 36 | 204 | 49 | 205 |
| (b) 0.0065–0.0140 | 38 | 204 | 37 | 205 |
| (c) 0.0141–0.0295 | 35 | 204 | 44 | 205 |
| (d) 0.0297–0.0665 | 54 | 204 | 34 | 205 |
| (e) $\geqslant$ 0.0673 | 41 | 204 | 41 | 205 |

ACKNOWLEDGEMENTS

The author wishes to thank Marie Diener-West, Paul Rathouz, Richard Royall and two referees for their insightful comments.

REFERENCES

1. Royall RM. *Statistical Evidence: A Likelihood Paradigm.* Chapman and Hall: London, 1997.
2. Goodman SN. Toward evidence-based medical statistics: the p-value fallacy. *Annals of Internal Medicine* 1999; **130**(12):995–1004.
3. Edwards AWF. *Likelihood.* Cambridge University Press: London, 1971.
4. Goodman SN. P-values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. *American Journal of Epidemiology* 1993; **137**(5):485–496.
5. Royall RM. The likelihood paradigm for statistical evidence (with discussion). In *The Nature of Statistical Evidence*. University of Chicago Press: Chicago, IL (in press).
6. Vieland VJ, Hodge SE. Review of R. Royall (1997) statistical evidence: a likelihood paradigm. *Annals of Human Genetics* 1998; **63**:283–289.
7. Hacking I. *Logic of Statistical Inference.* Cambridge University Press: New York, 1965.
8. Goodman SN, Royall RM. Evidence and scientific research. *American Journal of Public Health* 1988; **78**(12):1568–1574.
9. Fisher RA. *Statistical Methods and Scientific Inference.* 2nd edn. Hafner: New York, 1959.
10. Royall RM. On the probability of observing misleading statistical evidence (with discussion). *Journal of the American Statistical Association* 2000; **95**(451):760–767.
11. Jeffreys H. *Theory of Probability*. 3rd edn. Oxford University Press: Oxford, 1961.
12. Kass RE, Raftery AE. Bayes factors. *Journal of the American Statistical Association* 1995; **90**:773–795.
13. Birnbaum A. On the foundations of statistical inference (with discussion). *Journal of the American Statistical Association* 1962; **53**:259–326.
14. Berger JO, Wolpert RL. *The Likelihood Principle.* Institute of Mathematical Statistics: Hayward, California, 1988.

 

15. Spiegelhalter DJ, Freedman LS, Blackburn PR. Monitoring clinical trials: conditional or predictive power? *Controlled Clinical Trials* 1986; **7**:8–17.
16. Goodman SN. Multiple comparisons, explained. *American Journal of Epidemiology* 1998; **147**:807–812.
17. Royall RM. The effect of sample size on the meaning of significant tests. *American Statistician* 1986; **40**: 313–315.
18. Peto R. Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPhearson K, Peto J, Smith PG. Design and analysis of randomized clinical trials requiring prolonged observation of each patient: Introduction and design. *British Journal of Cancer* 1976; **34**:585–612.
19. Pratt JW. 'Decisions' as statistical evidence and Birnbaum's confidence concept. *Synthese* 1977; **36**:59–69.
20. Cornfield J. Sequential trials, sequential analysis, and the likelihood principle. *American Statistician* 1966; **29**(2):18–33.
21. Smith CAB. The detection of linkage in human genetics. *Journal of the Royal Statistical Society*, *Series B* 1953; **15**:153–192.
22. Robbins H. Statistical methods related to the law of the iterated logarithm. *Annals of Mathematical Statistics* 1970; **41**:1397–1409.
23. Blume JD. *On the probability of observing misleading evidence in sequential trials*, PhD dissertation, Johns Hopkins University School of Public Health, 1999.
24. Blume JD. On the probability of observing misleading evidence in sequential trials. 2001 (submitted) Available at http://alexander.stat.Brown.edu/-Jblume/slides.
25. Anscombe FJ. Normal likelihood functions. *Annals of the Institute of Statistical Mathematics* 1964; **26**: 1–19.
26. Tsou TS, Royall RM. Robust likelihoods. *Journal of the American Statistical Association* 1995; **90**:316–320.
27. Wolpert RL. Discussion of On the Probability of Observing Misleading Statistical Evidence by RM Royall. *Journal of the American Statistical Association* 2000; **95**(451):760–767.
28. Barndorfff-Nielsen OE, Cox DR. *Inference and Asymptotics.* Chapman and Hall: New York, 1994.
29. The University Group Diabetes Program. A study of the effects hypoglycemic agents on vascular complications in patients with adult-onset diabetes. *Diabetes* 1970: **19** (suppl. 2):747–839.
30. Meinert CL, Tonascia S. *Clinical Trials: Design*, *Conduct*, *and Analysis.* Oxford University Press: New York, 1986.
31. Marks H. *The Progress of Experiment.* Cambridge University Press: Cambridge, 1997.
32. Committee for the Assessment of Biometric Aspects of Controlled Trials of Hypoglycemic Agents. Report of the committee for the assessment of biometric aspects of controlled trials of hypoglycemic agents. *Journal of the American Medical Association* 1975; **231**:583–608.
33. Schor S. The University Group Diabetes Program: a statistician looks at the mortality results. *Journal of the American Medical Association* 1971: **217**:1671–1675.
34. Feinstein AR. Clinical biostatistics VIII: an analytic appraisal of the University Group Diabetes Program (UGDP) study. *Clinical Pharmacology* 1971; **12**:167–191.
35. Seltzer HS. A summary of criticisms of the findings and conclusions of the University Group Diabetes Program (UGDP). *Diabetes* 1972; **21**:976–979.
36. Cornfield J. The University Group Diabetes Program: a further statistical analysis of the mortality findings. *Journal of the American Medical Association* 1971; **217**:1676–1687.
37. Prentice RL. Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors. *Journal of the American Statistical Association* 1986; **81**:321–327.
38. Diamond G, Forrester J. Clinical trials and statistical verdicts: probable grounds for appeal. *Annals of Internal Medicine* 1983; **98**(3):385–394.
39. Berry G. The analysis of mortality by the subject-years method. *Biometrics* 1983; **39**:173–184.
40. Kilo C, Miller J, Williamson J. The achilles heel of the university group diabetes program. *Journal of the American Medical Association* 1980; **245**(5):450–457.
41. Kilo C, Miller J, Williamson J. The crux of the UGDP: Spurious results and biologically inappropriate data analysis. *Diabetologia* 1980; **18**(3):179–185.

 