Statistica Sinica **22** (2012), 1147-1198
doi:http://dx.doi.org/10.5705/ss.2009.125

THE STRENGTH OF STATISTICAL EVIDENCE FOR COMPOSITE HYPOTHESES: INFERENCE TO THE BEST EXPLANATION

David R. Bickel

University of Ottawa

Abstract: A general function to quantify the weight of evidence in a sample of data for one hypothesis over another is derived from the law of likelihood and from a statistical formalization of inference to the best explanation. For a fixed parameter of interest, the resulting weight of evidence that favors one composite hypothesis over another is the likelihood ratio using the parameter value consistent with each hypothesis that maximizes the likelihood function over the parameter of interest. Since the weight of evidence is generally only known up to a nuisance parameter, it is approximated by replacing the likelihood function with a reduced likelihood function on the interest parameter space. The resulting weight of evidence has both the interpretability of the Bayes factor and the objectivity of the p-value. In addition, the weight of evidence is coherent in the sense that it cannot support a hypothesis over any hypothesis that it entails. Further, when comparing the hypothesis that the parameter lies outside a non-trivial interval to the hypothesis that it lies within the interval, the proposed method of weighing evidence almost always asymptotically favors the correct hypothesis under mild regularity conditions. Even at small sample sizes, replacing a simple hypothesis with an interval hypothesis substantially reduces the probability of observing misleading evidence. Sensitivity of the weight of evidence to hypotheses' specification is mitigated by making them imprecise. The methodology is illustrated in the multiple comparisons setting of gene expression microarray data, and issues with simultaneous inference and multiplicity are addressed.

Key words and phrases: Bayes factor, Bayesian model selection, coherence, direct likelihood, hypothesis testing, evidential support, foundations of statistics, likelihoodism, model selection, strength of statistical evidence.

1. Introduction

1.1. Decision-theory and evidential inference

1.1.1. Classical inference frameworks

Current needs to evaluate evidence over thousands of hypotheses in genomics and data mining reopen the question of how to quantify the strength of evidence. Some of the most pronounced differences between inferences made by methods based on coverage or error frequencies and by other statistical methods occur in the realm of multiple comparisons, giving new importance to old debates on the foundations of statistics.

Each of the two main frameworks of statistical inference rests on solid decision-theoretic foundations. In the most-developed frequentist framework, that of Neyman and Pearson, the practice of deciding to reject only those hypotheses with valid p-values falling below a fixed significance level strictly controls the rate of Type I errors. In the strict Bayesian framework, that of Estienne (1903a,b, 1904a,b), F. P. Ramsey (cited in Jefreys (1948)), de Finetti (1970), and Savage (1954), the concept of coherent decision-making leads to probability as a measure of belief in the sense that it has a one-to-one correspondence with how much an intelligent agent would wager on its truth given the available information and a fixed loss function, prior distribution, and model. The methods of both frameworks find direct applications to problems requiring some degree of automatic decision-making. For example, the Neyman-Pearson framework provides rules deciding when a clinical trial is successful or when to stop an unsuccessful trial, and the Bayes-Estienne framework enables e-mail filters to decide which messages are unwanted.

The methods of these decision-theoretic frameworks have been adapted to problems requiring reports of the strength of the evidence in the data supporting one hypothesis over another rather than automated decisions to reject one hypothesis in favor of another. Bayes factors have long been advocated as measures of the strength of statistical evidence (e.g., Jefreys (1948); Kass and Raftery (1995)). Accordingly, Osteyee and Good (1974) considered the logarithm of the Bayes factor the "weight of evidence" for one hypothesis over another. This seems reasonable since the Bayes factor is equal to the posterior odds divided by the prior odds. (*Weight of evidence* is instead used herein as an abbreviation for *strength of statistical evidence*.)

Likewise, p-values from methods designed to control the rate of Type I (false positive) errors are routinely interpreted in the scientific literature as measures of evidence favoring alternative hypotheses over null hypotheses. Although the comparison of a p-value to a previously fixed level of significance to make a decision on rejecting a null hypothesis is common in clinical trials, in less regulated fields, the p-value is more often interpreted as a measure of evidence or support that a sample of data provides about a statistical hypothesis. Wright (1992) put it simply, "The smaller the p-value, the stronger the evidence against the null hypothesis." This use by Fisher of the p-value to quantify the degree of consistency of the data with the null hypothesis is called *significance testing* to sharply distinguish it from its use by Neyman to decide whether to reject the null hypothesis at a previously fixed Type I error rate (Cox (1977)). Among the examples of significance testing to be found in scientific disciplines as diverse as biomedicine, basic neuroscience, and physics may be found the common but theoretically unjustified practice of taking a sufficiently high p-value as evidence that there is "no effect" (Spicer and Francisco (1997); Pasterkamp et al. (2003)) and many statisticians' interpretation of a sufficiently low p-value as strong evidence against the null hypothesis; e.g., Fraser, Reid, and Wong (2004). Even the critics of significance testing acknowledge that it serves its purpose in some situations (Spjøtvoll (1977); Goodman and Royall (1988)).

1.1.2. Objectivity and interpretability

In spite of the uncontested value of methods of the Neyman-Pearson and Bayes-Estienne frameworks in the decision-making roles for which they are optimal, their application to quantifying the strength of statistical evidence remains controversial. For neither the p-value nor the Bayes factor qualifies as a general measure of evidence if the strength of statistical evidence in a particular data set for one given hypothesis over another under a specified family of probability distributions must meet both of these criteria:

1. the *objectivity* condition, that the strength of evidence does not vary from one researcher to another;
2. the *interpretability* condition, that the strength of evidence has the same practical interpretation for any sample size.

The first condition rules out Bayes factors that depend on subjective or default priors, and the second condition rules out the p-value (Bickel (2011b)), as will be seen later in this subsection. By contrast, the likelihood ratio satisfies both of the necessary conditions for a measure of the strength of statistical evidence. In a philosophical study of the foundations of statistical theory, I. Hacking proposed the *law of likelihood* in terms of data d and hypotheses h and i: "d supports h better than i whenever the likelihood ratio of h to i given d exceeds 1" (Hacking (1965, p.71, italics added)). The usual restatement of the law follows. At each value of θ, the D-dimensional parameter, $f(\bullet;\theta)$ denotes the probability density or probability mass function of the random n-tuple X of which the fixed n-tuple of observations x is a realization. $L(\bullet) = L(\bullet; x) = f(x; \bullet)$, a function on the parameter space Θ, is called the *likelihood function.* In the evidential framework of statistical inference, the likelihood ratio $L(\theta'; x)/L(\theta''; x)$ is the strength of statistical evidence in $X = x$ that supports $\theta = \theta'$ over $\theta = \theta''$, and if $L(\theta'; x)/L(\theta''; x) > 1$, there is more evidence for $\theta = \theta'$ than for $\theta = \theta''$ (Royall (2000a)). Both hypotheses under consideration are simple in the sense that each corresponds to a single parameter value, a point in Θ. In this case of two simple hypotheses, the logarithm of the Bayes factor equals $\log(L(\theta'; x)/L(\theta''; x))$, which Edwards (1992) called the support for $\theta = \theta'$ over $\theta = \theta''$.

More generally, the Bayes factor has been used to compute a ratio of posterior probabilities of two hypotheses that are composite, that is, corresponding to multiple parameter values. In the strict Bayes-Estienne framework, however, since the prior probability of each hypothesis varies from one decision maker to another, the ratio of posterior probabilities violates the objectivity condition of a measure of evidence. In the applied data analysis, Bayesians rarely make the effort required to elicit prior distributions from experts to adequately reflect their levels of uncertainty about parameter values, perhaps because it is justifiable in very few practical situations. The arguably less subjective practice of automatically assigning 50% prior probability to each hypothesis reduces the ratio of posterior probabilities to the Bayes factor. Although the principle of insufficient reason behind that practice still has its defenders (Williamson (2005)), the well known problems with partitioning the parameter set into equally probable subsets remain (Kass and Wasserman (1996)). The Bayes factor also requires a prior distribution if either hypothesis corresponds to more than one parameter value or if there is a nuisance parameter. Although default priors are much more convenient than their frankly subjective counterparts and seem to offer more objectivity (Berger (2004)), there is no consensus on how to select one of the many available rules for generating default priors, and yet small-sample inference can be sensitive to such selection (Kass and Wasserman (1996)). Arguments for priors based on group invariance are compelling but do not apply to all situations, whereas generally applicable and widely used reference priors are functions of which parameters are of interest (Bernardo (1979)), thereby eroding Bayes-Estienne foundations unless the prior levels of an agent's beliefs should depend on which parameter that agent intends to use in decision making. Regardless of the specific algorithm selected, the automatic generation of priors introduces a problem of interpreting the resulting posterior probabilities since the prior probabilities do not correspond to any scientist's actual levels of belief, as a more rigorous application of Bayes-Estienne decision theory would require. Consequently, a default prior often serves to determine what an ideal agent whose beliefs were encoded by that prior would believe upon observing the data (Bernardo (1997)). If a prior is instead chosen in order to derive credible sets that match confidence intervals, using Bayesian calculations for Neyman-Pearson inference, objectivity and an unambiguous interpretation of probability are thereby purchased at the price of abandoning strict Bayes-Estienne decision theory, except in special cases.

The interpretability condition for a measure of evidence is defined at the end of this subsection after introducing its foundational concepts. With the likelihood ratio as the measure of the strength of evidence, the analog of a Type I error rate plays key roles in sample size planning and in the choice of a method of eliminating nuisance parameters without itself quantifying the strength of evidence

(Strug, Rohde, and Corey (2007); Blume (2008).) This analog, the probability of observing misleading evidence, is defined as follows. Consider the strength of evidence in observed data generated by distribution P_θ in favor of the false hypothesis that the data were generated by a distribution in the set $\{P_{\theta'} : \theta' \neq \theta\}$. The *observation of misleading evidence* is the event that the strength of evidence for the false hypothesis exceeds a fixed threshold representing the boundary between weaker and stronger evidence, and the *probability of observing misleading evidence* is the limiting relative frequency of observations of misleading evidence under repeated sampling.

Ideally, the probability of observing misleading evidence would converge to 0 with increasing sample size. In other words, more information would increase the reliability of inferences made from the available evidence, at least asymptotically. Hypothesis testing with a fixed Type I error rate, say 0.05, as the threshold separating weaker evidence from stronger evidence, with the p-value as the level of evidence, fails in this regard. Indeed, under the null hypothesis, the probability that the p-value is less than that threshold is equal to the fixed error rate for all samples sizes and thus cannot vanish. Consequently, the result of a conventional hypothesis test, whether expressed as a p-value or as an accept/reject decision, cannot be evidentially interpreted without taking the sample size into consideration, which is why a given p-value is thought to provide stronger evidence against the null hypothesis if the sample is small than if it is large (Royall (1997)). For example, as Goodman and Royall (1988) explain, a p-value of 0.05 in many cases corresponds to a likelihood ratio indicating overwhelming evidence *in favor of* the null hypothesis for sufficiently large samples. For this reason, a candidate measure of evidence is considered *interpretable* if the associated probability of observing misleading evidence approaches 0 asymptotically.

1.2. Evidence for a composite hypothesis

In spite of meeting the two criteria of a measure of evidence, the classical law of likelihood is insufficient for statistical inference if either hypothesis is composite. This insufficiency threatens to severely limit the scope of likelihood-evidential inference since most statistical tests in common use compare a simple null hypothesis $\theta = \theta''$ to a composite alternative hypothesis such as $\theta > \theta''$ or $\theta \neq \theta''$.

In some areas of application, subject-matter knowledge may inform the replacement of a composite hypothesis $\theta \in \Theta'$ (for some $\Theta' \subseteq \Theta$) with a simple hypothesis $\theta = \theta'$ for computing $L(\theta')/L(\theta'')$ as the weight of evidence. For example, in linkage analysis, Strug and Hodge (2006) set θ' to the smallest plausible value of the recombination fraction θ for the purpose of using likelihood ratios instead of p-values that employ composite alternative hypotheses. In other domains, any selection of a simple hypothesis in place of a composite hypothesis would be unacceptably arbitrary or subjective. Nonetheless, there may sometimes be advantages in evidential inference to setting θ' to the parameter value as close as possible to θ'' such that $|\theta' - \theta''|$ remains high enough to be practically significant; this concept of scientific significance was previously applied to non-evidential gene expression data analyses (Bickel (2004); Van De Wiel and Kim (2007)). An alternative is to set θ' to some conventional value, e.g., the value corresponding to a two-fold expression difference (an expression ratio estimate of 1/2 or 2) remains a commonly used threshold with gene expression studies in spite of its arbitrary nature (Lewin et al. (2006)). Comparing the evidential strength of one simple hypothesis to another has the advantage that $P_{\theta''}\left(L\left(\theta'\right)/L\left(\theta''\right) \geq \Lambda\right)$, the probability of observing misleading evidence at level $\Lambda > 1$, is asymptotically bounded by the standard normal cumulative distribution function evaluated at $-\sqrt{2\log\Lambda}$ if L is smooth and if the parameter dimension D is fixed, or by the Chebyshev or Markov bound $1/\Lambda$ more universally (Royall (2000a)). In addition, limiting the parameter of interest to one of two values is convenient when planning the size of a study (Strug, Rohde, and Corey (2007)).

Nonetheless, the weight of evidence involving a composite hypothesis cannot in general be measured or even approximated by substituting a simple hypothesis selected prior to observing the data. However, a solution to the composite hypothesis problem does appear to lie in the use of a likelihood interval or more general likelihood set. The level-Λ likelihood set $\mathcal{E}\left(\Lambda\right)$ consists of all values of θ satisfying $L\left(\theta\right) \geq L\left(\widehat{\theta}\right)/\Lambda$, where $\widehat{\theta}$ is the maximum likelihood estimate. Nonmembership in a likelihood set determines which parameter values are considered "obviously open to grave suspicion" (Fisher (1973, pp.75-76)) if not inconsistent with the data (Barnard (1967); Hoch and Blume (2008)). Thus, whenever $L\left(\widehat{\theta}\right)/L\left(\theta''\right) > \Lambda$ and $\widehat{\theta} \neq \theta''$, one or more parameter values in $\mathcal{E}\left(\Lambda\right)$ are considered better supported than $\theta = \theta''$ by the data, and, for that reason, $L\left(\widehat{\theta}\right)/L\left(\theta''\right)$ measures the weight of evidence for the composite hypotheses $\theta \in \mathcal{E}\left(\Lambda\right)$ over the simple hypothesis $\theta = \theta''$. By the same reasoning, $L\left(\widehat{\theta}\right)/L\left(\theta''\right)$ measures the weight of evidence for the composite hypotheses $\theta \neq \theta''$ over the simple hypothesis $\theta = \theta''$.

More generally, a formal interpretation of the principle of inference to the best explanation entails that $W\left(\Theta', \Theta''\right) = \sup_{\theta' \in \Theta'} L\left(\theta'\right)/\sup_{\theta'' \in \Theta''} L\left(\theta''\right)$ uniquely quantifies the weight of evidence for the hypotheses that $\theta \in \Theta'$ over the hypothesis that $\theta \in \Theta''$ in the absence of prior hypothesis probabilities, where Θ' and Θ'' are subsets of the parameter space (Section 2). It follows that, if θ is the parameter of interest, $\theta \in \Theta'$ is better supported than $\theta \in \Theta''$ if and only if

$\sup_{\theta' \in \Theta'} L(\theta'; x) > \sup_{\theta'' \in \Theta''} L(\theta''; x)$, a conclusion a preliminary version of the present article (Bickel (2008)) and Zhang (2009b) independently derived from different axiomatic systems.

Zhang (2009b) also recorded asymptotic properties of $W(\Theta', \Theta'')$, applied it to several interesting examples, refuted objections against its adoption, and gave guidelines for its derivation from statistical reports in the absence of the original data. The most important practical difference between our two approaches emerges in the presence of a nuisance parameter. Zhang (2009b) follows Royall (1992) and He, Huang, and Liang (2007) in framing the nuisance parameter problem as a special case of the composite hypothesis problem, whereas Bickel (2008) maintains the complete separation between the two problems (see Section 2.5, below). Unique contributions of the present paper are summarized in Section 6.1.

A discrepancy between the performance of the likelihood ratio for two fixed simple hypotheses and the likelihood ratio maximized over a subset of the parameter space including parameter values arbitrarily close to that of a simple hypothesis was uncovered by the example of the multivariate normal family with a 5-dimensional mean as θ (Kalbfeisch (2000)). Asymptotically, for any fixed θ' and θ'' in $\Theta = \mathbb{R}^5$, there is a 2.1% upper bound on $P_{\theta''}(L(\theta')/L(\theta'') > 8)$, the probability of observing misleading evidence at level $\Lambda = 8$ (Royall (2000a)). By contrast, the probability that the level-8 likelihood set contains θ'', assuming it is the true value of θ, is less than 50% (Kalbfeisch (2000)). This means the asymptotic probability of observing misleading evidence for $\theta \in \mathbb{R}^5 \backslash \{\theta''\}$ over $\theta = \theta''$ exceeds the asymptotic probability of observing misleading evidence for $\theta = \theta'$ over $\theta = \theta''$ by a factor of 25 or more. This malady is not limited to the normal case, but is symptomatic of inadequate interpretability when a hypothesis representing practically the entire parameter space is pitted against a simple hypothesis. The universal upper bound on $P_{\theta''}(L(\theta')/L(\theta'') > 8)$ is 12.5%, the Chebyshev or Markov bound. That is more than a factor of 4 smaller than $P_{\theta''}(L(\widehat{\theta})/L(\theta'') > 8) = 52.7\%$ in the example of $D = 5$ and conditions under which $2\log(L(\widehat{\theta})/L(\theta''))$ is asymptotically distributed as χ^2 with D degrees of freedom.

Given such an asymptotic distribution, $L\left(\widehat{\theta}\right)/L(\theta'')$ does not meet the interpretability condition of Section 1.1 since

$$\forall_{\Lambda > 1} \lim_{n \to \infty} P_{\theta''}\left(L(\widehat{\theta})/L(\theta'') > \Lambda\right) = \alpha$$

for some $\alpha > 0$. Thus, $L\left(\widehat{\theta}\right)/L(\theta'')$ is no more interpretable than a p-value as the strength of evidence.

Interpretability is recovered by instead quantifying the strength of evidence for a composite hypothesis over an interval hypothesis, e.g., for $|\theta| > \theta_+$ over $|\theta| \leq \theta_+$ for some fixed $\theta_+ > 0$. The proof is in Section 2. In addition to satisfying the interpretability condition, weighing evidence for composite hypotheses has intrinsic scientific merit, as, for example, when assessing evidence for bioequivalence or differential gene expression. Section 2 also highlights connections between Hacking's law of likelihood, evidence sets, and evidence for or against composite hypotheses.

For some applications, the main drawback of replacing a simple hypothesis with an interval hypothesis is the dependence on the interval bounds. This is largely overcome by the extension of evidential inference to handle imprecise composite hypotheses in Section 3.

The proposed methodology is studied by simulation (Section 4) and illustrated by application to microarray gene expression data (Section 5). Imprecise composite hypotheses provide a natural formalization of the imprecision inherent in what is meant when a biologist says a gene is "differentially expressed"; this imprecision applies to differential protein and metabolite expression as well as to differential gene expression. (Looking over thousands of genes for differential expression poses an extreme multiple comparisons problem in the Neyman-Pearson framework. Because, unlike the p-value, the likelihood ratio as a measure of statistical evidence is not based on the control of a Type I error rate, it is not adjusted for multiple comparisons by enforcing control of a family-wise error rate or a false discovery rate (Section 2.4). While many statisticians see the ability to correct for multiple tests in this way as an important advantage of the p-value over the likelihood ratio alone (Korn and Freidlin (2006)), others maintain that the perceived need to correct for multiple comparisons exposes a shortcoming in the evidential interpretation of the p-value (Royall (1997)).)

Section 6 has a discussion and opportunities for further research.

2. Inference about Precise Hypotheses

2.1. Preliminaries

2.1.1. Basic notation

The symbols $\subset$ and $\subseteq$ designate proper subsets and (possibly improper) subsets, respectively.

Consider the fixed positive integer D and the parameter space $\Theta \subseteq \mathbb{R}^D$. For all $\theta \in \Theta$, the probability distribution of the observable random n-tuple $X \in \Omega \subseteq \mathbb{R}^n$ admits a probability density or mass function $f(\bullet;\theta)$ on Ω such that $\theta' \neq \theta'' \Rightarrow f(\bullet;\theta') \neq f(\bullet;\theta'')$. For $X = x$, the likelihood function on Θ is $L(\bullet) = L(\bullet;x) = f(x;\bullet)$. Unless specified otherwise, the propositions of this paper hold generally for all x in $\{y : y \in \Omega, \forall_{\theta' \in \Theta} f(y; \theta') > 0\}$. Both Θ and Ω are nonempty. If imprecise hypotheses are under consideration, the probability distributions that determine the values of L are incomplete (Section 3).

2.1.2. Hypothesis types

Definition 1. For any nonempty subset Θ' of Θ, the hypothesis that $\theta \in \Theta'$ is *simple* if Θ' has only one element; otherwise, the hypothesis that $\theta \in \Theta'$ is *composite*. Any simple or composite hypothesis $\theta \in \Theta'$ is *intrinsically simple* if θ, conditional on $\theta \in \Theta'$, is a random D-tuple of some probability space $(\Theta', \mathcal{A}', p')$. Any composite hypothesis that is not intrinsically simple is *complex*.

As will become clear in Example 3, distinguishing composite hypotheses that are intrinsically simple from those that are complex facilitates inference about a random θ with frequency distribution p, not to be confused with any prior distribution π that represents uncertainty from the subjective viewpoint of an intelligent agent rather than objective variability actually occurring in the world. Thereby differentiating the random-effects or *physical* distribution p from the epistemological or *mental* distribution π (Williamson (2005)) plays a crucial role in the framework of Section 2.2. Whereas physical probabilities model limiting relative frequencies or proportions of real objects, mental probabilities instead model hypothetical levels of belief.

If θ is physically random with marginal probability space $(\Theta, \mathcal{A}, p)$, then the hypothesis $\theta \in \Theta'$ will not be of interest unless $\Theta' \in \mathcal{A} \backslash \emptyset$. To succinctly represent the hypotheses of potential interest, let $\Phi = \{\varphi(\Theta') : \Theta' \in \mathcal{A} \backslash \emptyset\}$ denote a parameter set isomorphic to $\mathcal{A} \backslash \emptyset$, where $\varphi : \mathcal{A} \backslash \emptyset \to \Phi$ is an invertible $\mathcal{A}$-measurable function. Consider the family $\left\{\overline{P}_\phi : \phi \in \Phi\right\}$ of probability distributions of X that admit probability density or mass functions $\left\{\overline{f}(\bullet; \phi) : \phi \in \Phi\right\}$ satisfying

$$\overline{f}(x; \phi') = \int f(x; \theta')\, dp\left(\theta' | \theta \in \varphi^{-1}(\phi')\right) \tag{2.1}$$

for all $\phi' \in \Phi$. Then the likelihood function on Φ is $\overline{L}(\bullet) = \overline{L}(\bullet; x) = \overline{f}(x; \bullet)$. Thus, every intrinsically simple hypothesis $\theta \in \Theta'$ under the first family of sampling distributions corresponds to a simple hypothesis $\phi = \varphi(\Theta')$ under the new family. By contrast, a complex hypothesis cannot be reduced to a simple hypothesis.

Whether $\theta \in \Theta'$ means "θ that is in Θ'" or "the hypothesis that θ is in Θ'" may be determined from the context. In the sequel, every subset of Θ is nonempty and corresponds to either a simple hypothesis or a composite hypothesis of potential interest. Accordingly, 2^Θ denotes the set of all non-empty subsets of Θ if θ is fixed or $\mathcal{A} \backslash \emptyset$ if θ is random, in which case "all $\Theta' \subseteq \Theta$" stands for "all $\Theta' \in \mathcal{A} \backslash \emptyset$."

2.2. Explanatory theory of evidence

Section 2.2.1 formalizes the concept of explanatory power that will be used in Section 2.2.2 to define the weight of evidence. For the sake of applications, Section 2.2.3 expresses the weight of evidence more concisely.

2.2.1. Inference to the best explanation

Let $\operatorname{ex}(\Theta') = \operatorname{ex}(\Theta'; x)$ denote the *explanatory power* of $\theta \in \Theta'$ with respect to $X = x$, that is, the ability of $\theta \in \Theta'$ to explain why x was observed as opposed to some other realization of X. The function ex on $2^{\Theta} \times \Omega$ will be restricted by weak conditions needed for use with the weight of evidence. To motivate the condition that pertains specifically to simple hypotheses, measures of explanatory power proposed by Popper (2002, Appendix IX) and Niiniluoto (2004) will exemplify the concept.

Example 1. Niiniluoto (2004) recorded two functions that quantify the ability of a simple hypothesis to explain data. The one that does not necessitate assigning probabilities to hypotheses is now generalized to continuous parameter values. Let $\overline{f}(x) = \int f(x;\theta)\, d\pi(\theta)$, where π is a non-Dirac measure on Θ and $f(\bullet;\theta)$ is a probability mass function. Then

$$\operatorname{ex}\left(\left\{\theta'\right\}; x\right) = \frac{f(x;\theta') - \overline{f}(x)}{1 - \overline{f}(x)} \tag{2.2}$$

is the explanatory power of $\theta = \theta'$ with respect to $X = x$.

Example 2. Also with discrete data and parameters in mind, Popper (2002, pp.416, 420-421) considered

$$\begin{aligned}\operatorname{ex}\left(\left\{\theta'\right\}; x\right) &= \frac{f(x;\theta') - \overline{f}(x)}{f(x;\theta') + \overline{f}(x)};\\ \operatorname{ex}\left(\left\{\theta'\right\}; x\right) &= \log_2\left(\frac{f(x;\theta')}{\overline{f}(x)}\right)\end{aligned}$$

as two possible values of explanatory power that are equally applicable to continuous data and parameters.

Since ex will only serve to rank hypotheses, the measure π defining $\overline{f}(x)$ in the examples need neither be specified nor included in the simple-hypothesis axiom to be included in the definition of ex. Intrinsically simple hypotheses are replaced with the equivalent simple hypotheses for application of that axiom.

A strong idealization of the principle of inference to the best explanation stipulates that the simple hypothesis of highest explanatory power be inferred

(Niiniluoto (2004)). The complex-hypothesis axiom of the explanatory function ex weakens that idealization by stipulating only that the ability of $\theta \in \Theta'$ to explain $X = x$ cannot exceed that of $\theta \in \Theta''$, where Θ'' contains a parameter value of highest explanatory power, unless either $\theta \in \Theta'$ or $\theta \in \Theta''$ is intrinsically simple.

For application to both simple hypotheses and composite hypotheses according to the above sketch, ex satisfies the conditions imposed by the following recursive definition.

Definition 2. A function ex on $2^{\Theta} \times \Omega$ is an *explanatory function* if it satisfies the following axioms.

1. $\mathrm{ex}\left(\{\theta'\};x\right) : \Theta \to \mathbb{R}^1$ increases monotonically with the likelihood function L as $\theta' \in \Theta$ varies.
2. For all $\Theta' \subseteq \Theta$ and $\Theta'' \subseteq \Theta$ such that each of the hypotheses $\theta \in \Theta'$ and $\theta \in \Theta''$ is either simple or complex,
$$\arg \sup_{\theta \in \Theta' \cup \Theta''} \mathrm{ex}\left(\{\theta\};x\right) \in \Theta'' \implies \mathrm{ex}\left(\Theta';x\right) \leq \mathrm{ex}\left(\Theta'';x\right). \tag{2.3}$$
3. For all $\Theta' \subseteq \Theta$ such that $\theta \in \Theta'$ is an intrinsically simple hypothesis,
$$\mathrm{ex}\left(\Theta';x\right) = \overline{\mathrm{ex}}\left(\left\{\varphi\left(\Theta'\right)\right\};x\right), \tag{2.4}$$
where $\overline{\mathrm{ex}}$ is an explanatory function on $\{\{\phi\} : \phi \in \Phi\} \times \Omega$, i.e., $\overline{\mathrm{ex}}$ satisfies Axioms 1 and 2 with $\{\{\phi\} : \phi \in \Phi\}$ in place of 2^{Θ}, $\phi = \varphi(\Theta')$ in place of $\theta \in \Theta'$, etc. (Axiom 3 does not apply to $\overline{\mathrm{ex}}$ since the fact that ϕ is fixed rather than random means each hypothesis that $\phi = \varphi(\Theta')$ is simple, not intrinsically simple.)

(2.4) says the explanatory power of an intrinsically simple hypothesis is equal to that of the equivalent simple hypothesis about the parameter in the family of distributions induced by (2.1).

Violation of (2.3) would mean there is a simple or complex hypothesis that explains the data better than a simple or complex hypothesis that contains the best explanation.

2.2.2. Evidential functions

Let $W(\Theta', \Theta'') = W(\Theta', \Theta''; x)$ denote the *weight of evidence* in $X = x$ that supports $\theta \in \Theta'$ over $\theta \in \Theta''$. In the terminology of Section 2.2.1, the evidential function W is now defined in terms of the explanatory function ex that yields $\mathrm{ex}(\Theta'; x)$ as the power of the hypothesis $\theta \in \Theta'$ to explain why $X = x$.

Definition 3. A function W on $2^{\Theta} \times 2^{\Theta} \times \Omega$ is an *evidential function* with respect to an explanatory function ex if it satisfies the following axioms.

1. For all $\theta' \in \Theta$ and $\theta'' \in \Theta$,
$$W\left(\left\{\theta'\right\}, \left\{\theta''\right\}; x\right) = \frac{L\left(\theta'; x\right)}{L\left(\theta''; x\right)}. \tag{2.5}$$
2. For all $\Theta', \Theta'', \Theta''' \subseteq \Theta$,
$$\begin{aligned} W\left(\Theta', \Theta''; x\right) = W\left(\Theta'', \Theta'; x\right) \iff & \quad W\left(\Theta', \Theta'''; x\right) = W\left(\Theta'', \Theta'''; x\right) \\ \iff & \quad W\left(\Theta''', \Theta'; x\right) = W\left(\Theta''', \Theta''; x\right). \end{aligned} \tag{2.6}$$
3. For all $\Theta', \Theta'' \subseteq \Theta$,
$$W\left(\Theta', \Theta''; x\right) \leq W\left(\Theta'', \Theta'; x\right) \iff \text{ex}\left(\Theta'; x\right) \leq \text{ex}\left(\Theta''; x\right). \tag{2.7}$$
4. For all $\Theta', \Theta'' \subseteq \Theta$ such that $\theta \in \Theta'$ and $\theta \in \Theta''$ are intrinsically simple hypotheses,
$$W\left(\Theta', \Theta''; x\right) = \overline{W}\left(\left\{\varphi\left(\Theta'\right)\right\}, \left\{\varphi\left(\Theta''\right)\right\}; x\right), \tag{2.8}$$
where $\overline{W}$ is any evidential function on $\{\{\phi\} : \phi \in \Phi\}^2 \times \Omega$.

According to (2.5) , the likelihood ratio $W\left(\left\{\theta'\right\}, \left\{\theta''\right\}\right)$ is the weight of evidence in $X = x$ that supports $\theta = \theta'$ over $\theta = \theta''$; this *special law of likelihood* is restricted to the special case of simple hypotheses (Section 1.1.2). (2.5) calibrates the weight of evidence for one simple hypothesis over another. The special law of likelihood does not in itself specify how to weigh evidence for or against a complex hypothesis (Royall (2000b); Blume (2002)) unless all parameter values represented by the complex hypothesis have the same likelihood (Royall (1997, pp.17-18)).

By contrast, Definition 3 does apply to composite hypotheses. Specifically, the principle of inference to the best explanation idealized by (2.3) extends the special law of likelihood to complex hypotheses, whereas intrinsically simple hypotheses are replaced with simple hypotheses in accordance with (2.1).

Following Jefreys (1948) with the weight of evidence in place of the Bayes factor and with a slight change of wording, the number of achieved bans ($b = \log_{10} W\left(\Theta', \Theta''\right)$) indicates weak evidence ($0 < |b| < 1/2$), moderate evidence ($1/2 \leq |b| < 1$), strong evidence ($1 \leq |b| < 3/2$), very strong evidence ($3/2 \leq |b| < 2$), or decisive evidence ($|b| \geq 2$) supporting $\theta \in \Theta'$ over $\theta \in \Theta''$ if $b > 0$, or supporting $\theta \in \Theta''$ over $\theta \in \Theta'$ if $b < 0$.

Example 3. Let $\Theta = \{1, \ldots, 101\}$ correspond to 101 urns, each containing black balls and white balls. An urn is selected randomly, with known probability $p(i) = 1/101$ of selecting the ith urn. A ball is then randomly drawn with an equal probability of drawing any ball from the selected urn θ, as in Kyburg and Teng (2006, p.216). The proportion of black balls in the first urn is $10^{-0.8}$, and the proportion of black balls in each other urn is 10^{-2}. Consider the simple hypothesis that $\theta = 1$ and the composite hypothesis that $\theta \in \{1, 2\}$. The latter is not the complex hypothesis that the ball was drawn either from the first urn or the second urn but rather is the intrinsically simple hypothesis that the ball was randomly selected either from the first urn with 50% probability or from the second urn with 50% probability. Thus, (2.8) pertains and, if a black ball is drawn, then

$$\begin{aligned} W(\{1\}, \{1,2\}; \text{black}) &= \overline{W}(\{\phi_1\}, \{\phi_{1,2}\}; \text{black}) \\ &= \frac{\overline{f}(\text{black}; \phi_1)}{\overline{f}(\text{black}; \phi_{1,2})} \\ &= \frac{f(\text{black}; 1)}{(50\%)\, f(\text{black}; 1) + (50\%)\, f(\text{black}; 2)} \\ &= \frac{10^{-0.8}}{(10^{-0.8} + 10^{-2})/2} \approx 2 \in \left(0, 10^{1/2}\right), \end{aligned}$$

where $\phi = \phi_1$ and $\phi = \phi_{1,2}$ are the two hypotheses in the new parameterization in the notation of Section 2.1.2. In words, drawing a black ball *weakly* supports the hypothesis that the ball was drawn from the first urn over the hypothesis that the ball was randomly selected either from the first urn with 50% probability or from the second urn with 50% probability. Again applying Definition 3 gives

$$W(\{1\}, \{2, \ldots, 100\}; \text{black}) = \frac{10^{-0.8}}{10^{-2}} = 10^{1.2} \geq 10^1,$$

showing that drawing a black ball *strongly* supports the hypothesis that the ball was randomly selected from the first urn over the hypothesis that it was selected randomly from one of the other urns.

Popper (2002, p.430) anticipated a special case of Definition 3 by noting that the explanatory power can be interpreted as a measure "of the weight of the evidence in favor of" the hypothesis. From that perspective, the weight of evidence for one hypothesis over another may be deemed synonymous with the explanatory power of the former hypothesis relative to the latter hypothesis, thereby obviating normalization by $\overline{f}(x)$. However, the simple identification of the weight of evidence with relative explanatory power breaks down in the presence of a nuisance parameter (Section 2.5.2).

The evidential functions on $2^{\Theta} \times 2^{\Theta} \times \Omega$ should not be confused with the *evidence functions* on $\Theta \times \Theta \times \Omega$ that Lele (2004) studied. Definition 3 may extend the latter to composite hypotheses by substituting each evidence function for the likelihood ratio in (2.5) and by making the analogous modification to the likelihood axiom of Definition 2.

2.2.3. General law of likelihood

Sections 2.2.1 and 2.2.2 lead to two practical equations for weighing evidence favoring one hypothesis over another.

Proposition 1. General law of likelihood. *For any explanatory function* ex*, let* W *denote the evidential function with respect to* ex*. Then the weight of evidence in* $X = x$ *that supports* $\theta \in \Theta'$ *over* $\theta \in \Theta''$ *is*

$$W\left(\Theta', \Theta''; x\right) = \frac{\sup_{\theta' \in \Theta'} L\left(\theta'; x\right)}{\sup_{\theta'' \in \Theta''} L\left(\theta''; x\right)} \tag{2.9}$$

for all $\Theta' \subseteq \Theta$ *and* $\Theta'' \subseteq \Theta$ *such that* $\theta \in \Theta'$ *and* $\theta \in \Theta''$ *is each either a simple hypothesis or a complex hypothesis but is*

$$W\left(\Theta', \Theta''; x\right) = \frac{\int_{\Theta'} L\left(\theta'; x\right) dp\left(\theta' | \theta \in \Theta'\right)}{\int_{\Theta''} L\left(\theta''; x\right) dp\left(\theta'' | \theta \in \Theta''\right)} \tag{2.10}$$

for all $\Theta' \subseteq \Theta$ *and* $\Theta'' \subseteq \Theta$ *such that* $\theta \in \Theta'$ *and* $\theta \in \Theta''$ *are intrinsically simple.*

Proof. In the case that $\theta \in \Theta'$ and $\theta \in \Theta''$ are intrinsically simple, (2.8), (2.1), and (2.5) together entail (2.10). The remainder of the proof derives (2.9) for the case that $\theta \in \Theta'$ and $\theta \in \Theta''$ is each either a simple hypothesis or a complex hypothesis. By (2.3),

$$\theta' = \arg\sup_{\theta \in \Theta'} \operatorname{ex}\left(\{\theta\}\right) \implies \operatorname{ex}\left(\Theta'\right) = \operatorname{ex}\left(\left\{\theta'\right\}\right)$$

for all $\Theta' \subseteq \Theta$. Then, according to Definition 2 and (2.7),

$$\theta' = \arg\sup_{\theta \in \Theta'} L\left(\theta\right) \implies W\left(\Theta', \left\{\theta'\right\}\right) = W\left(\left\{\theta'\right\}, \Theta'\right),$$

which, by (2.6), in turn yields

$$\theta' = \arg\sup_{\theta \in \Theta'} L\left(\theta\right) \implies W\left(\Theta', \Theta''\right) = W\left(\left\{\theta'\right\}, \Theta''\right)$$

and, similarly,

$$\theta'' = \arg\sup_{\theta \in \Theta''} L\left(\theta\right) \implies W\left(\Theta', \Theta''\right) = W\left(\Theta', \left\{\theta''\right\}\right)$$

for all $\Theta', \Theta'' \subseteq \Theta$. Combining results,

$$W\left(\Theta', \Theta''\right) = \quad W\left(\left\{\arg \sup_{\theta \in \Theta'} L\left(\theta\right)\right\}, \left\{\arg \sup_{\theta \in \Theta''} L\left(\theta\right)\right\}\right)$$

and thus, from (2.5),

$$W\left(\Theta', \Theta''\right) = \frac{L\left(\arg \sup_{\theta \in \Theta'} L\left(\theta\right)\right)}{L\left(\arg \sup_{\theta \in \Theta''} L\left(\theta\right)\right)}.$$

The proof does not depend on the exact form of the explanatory power of a simple hypothesis but only requires that it monotonically increase with the likelihood (Definition 2). See Foster (2004) for a defense of that requirement. The connection to the principle of inference to the best explanation largely answers the objection that an explanatory rationale for (2.9) “has no strong logical grounding” (Lehmann (2006)).

2.3. Implications of the theory

2.3.1. Properties of the weight of evidence

The “coherence” of the weight of evidence in the technical sense of Schervish (1996) and Lavine and Schervish (1999) follows trivially from Proposition 1.

Proposition 2. Coherence. *For any explanatory function* ex*, let* W *denote the evidential function with respect to* ex*. Given any simple or complex hypotheses* $\theta \in \Theta'$ *and* $\theta \in \Theta''$*,*

$$\forall_{\Theta'', \Theta''' \subseteq \Theta} \forall_{\Theta' \subseteq \Theta''} W\left(\Theta', \Theta'''; x\right) \leq W\left(\Theta'', \Theta'''; x\right). \tag{2.11}$$

The coherence property prevents attributing more evidence to a simple or complex hypothesis than to an implication of that hypothesis (Schervish (1996); Lavine and Schervish (1999)).

While a ratio of posterior probabilities satisfies coherence, it generally violates the principle of inference to the best explanation.

Example 4. Let $\Theta = \left\{\theta^{(1)}, \ldots, \theta^{(101)}\right\}$ correspond to 101 distinct cosmological theories, each providing a different physical explanation of astronomical observations represented by x. The outcome $X = x$ would occur with probability $10^{-0.8}$ on the big bang theory $\left(\theta = \theta^{(1)}\right)$ and 10^{-2} on each of the other 100 theories, including the steady state theory $\left(\theta = \theta^{(2)}\right)$ (cf. Efron (2004)). If the theories were judged equally plausible before the measurements were made, each would have equal prior probability. Then the Bayes factor would ascribe more evidential weight to the big bang than to the hypothesis that either the big bang or the steady state theory is true:

$$W_{\mathrm{BF}}\left(\left\{\theta^{(1)}\right\}, \left\{\theta^{(1)}, \theta^{(2)}\right\}; x\right) = \frac{10^{-0.8}}{\left(10^{-0.8} + 10^{-2}\right)/2} \approx 2,$$

formally violating the coherence property. The ratio of posterior probabilities is coherent:

$$\frac{\pi\left(\theta=\theta^{(1)}|x\right)}{\pi\left(\theta\in\left\{\theta^{(1)},\theta^{(2)}\right\}|x\right)}=\frac{10^{-0.8}}{10^{-0.8}+10^{-2}}$$

However, the posterior odds fails to ascribe more weight to the big bang than to its denial, revealing a conflict between the principle of insufficient reason and the principle of inference to the best explanation:

$$\frac{\pi\left(\theta=\theta^{(1)}|x\right)}{\pi\left(\theta\neq\theta^{(1)}|x\right)}=\frac{10^{-0.8}}{(100)\left(10^{-2}\right)}=10^{-0.8}<1.$$

Few scientists would let a plethora of less adequate explanations prevent them from making an inference to the best explanation, the merits of Bayesianism in other settings notwithstanding. By contrast, the general law of likelihood indicates that there is strong evidence that the big bang occurred:

$$W\left(\left\{\theta^{(1)}\right\},\left\{\theta^{(2)},\ldots,\theta^{(101)}\right\};x\right)=\frac{10^{-0.8}}{10^{-2}}=10^{1.2}.$$

The Bayesian approach treats the theories of Example 4 exactly as if they were the randomly selected urns of Example 3, as seen in the mathematical equality of the results. Bayesianism has long been criticized for its inability to distinguish between frequencies of parameter values and levels of belief about parameter values (e.g., Kardaun et al. (2003)). While it is now generally acknowledged that no prior distribution can encode the state of zero information (Kass and Wasserman (1996); Bernardo (1997)), it is still claimed that a constant likelihood function does do so (Edwards (1992, Sec. 4.5); Schweder and Hjort (2002)).

In order to establish two more properties of the weight of evidence, the probability of observing misleading evidence mentioned in Section 1.1 is now defined more generally.

Definition 4. For any $\Theta'\subseteq\Theta$, $\Theta''\subseteq\Theta$, and $\Lambda>1$, the *probability of observing misleading evidence* in $X=x$ that supports $\theta\in\Theta'$ over $\theta\in\Theta''$ at level Λ with respect to some θ in Θ'' is

$$\alpha_\theta\left(\Lambda;\Theta',\Theta''\right)=P_\theta\left(W\left(\Theta',\Theta'';X\right)\geq\Lambda\right),$$

where X has probability density or mass function $f\left(\bullet;\theta\right)$.

As argued in Section 1.1, the weight of evidence is difficult to interpret unless the probability of observing misleading evidence approaches 0 asymptotically. That interpretability condition is satisfied in the case that one of two mutually exclusive hypotheses is a composite hypothesis corresponding to a parameter interval. The proof is facilitated by first noting that the weight of evidence almost always asymptotically selects the correct hypothesis:

Proposition 3. Consistency. *For any $\Theta'' \subset \Theta$ such that its interior* $\operatorname{int}\Theta''$ *contains θ,*

$$\lim_{n\to\infty} P_\theta\left(W\left(\Theta\backslash\Theta'', \Theta''\right) < 1\right) = \lim_{n\to\infty} P_\theta\left(W\left(\Theta'', \Theta\backslash\Theta''\right) > 1\right) = 1 \qquad (2.12)$$

under regularity conditions ensuring the weak consistency of $\widehat{\theta}^{(n)}$, the maximum likelihood estimate of θ.

Proof. The weak consistency of $\widehat{\theta}^{(n)}$ implies $\lim_{n\to\infty} P_\theta\left(\widehat{\theta}^{(n)} \in \operatorname{int}\Theta''\right) = 1$. (2.12) then follows from Proposition 1.

Proposition 4. Interpretability. *For any $\Theta'' \subset \Theta$ such that its interior* $\operatorname{int}\Theta''$ *contains θ,*

$$\lim_{n\to\infty} \alpha_\theta\left(\Lambda; \Theta\backslash\Theta'', \Theta''\right) = 0$$

for all $\Lambda > 1$ under regularity conditions ensuring the weak consistency of the maximum likelihood estimate of θ.

Proof. By Proposition 3,

$$\Lambda > 1 \implies \lim_{n\to\infty} P_\theta\left(W\left(\Theta\backslash\Theta'', \Theta''\right) \geq \Lambda\right) = 0.$$

2.3.2. Likelihood sets

The concept of the likelihood set is closely related to that of the strength of evidence for composite hypotheses, as sketched in Section 1.2.

Definition 5. Given some fixed $\Lambda > 1$ and $\Theta' \subseteq \Theta$, the *likelihood set of level* Λ for $X = x$ with respect to Θ' is

$$\mathcal{E}\left(\Lambda\right) = \mathcal{E}\left(\Lambda; x, \Theta'\right) = \left\{\theta'' : \theta'' \in \Theta, L\left(\theta''; x\right) \geq \sup_{\theta' \in \Theta'} \frac{L\left(\theta'; x\right)}{\Lambda}\right\}.$$

Definition 6. Given some fixed $\beta \in \mathbb{R}^1$ and $\Theta' \subseteq \Theta$, the *β-ban likelihood set* Θ' is $\mathcal{E}\left(10^\beta\right)$, its likelihood set of level 10^β.

Remark 1. Likewise, the β-bit likelihood set and the β-nat evidence set could be defined by substituting $\Lambda = 2^\beta$ and $\Lambda = e^\beta$, respectively. MacKay (2002) discusses the history of calling logarithmic "units" bits, bans, or nats, according to the base of the logarithm.

The likelihood set is used to distinguish parameter values supported by the data from parameter values less consistent with the data (Fisher (1973); Barnard (1967); Hoch and Blume (2008)). Such usage implicitly invokes a method of measuring the strength of evidence of a composite hypothesis in the same way as rejecting the hypothesis of a parameter value falling outside a $1-\alpha$ confidence interval implicitly invokes a hypothesis test with a Type I error rate of α. This practice is more precisely understood in terms of the weight of evidence for a composite hypothesis over its negation.

Proposition 5. *If $\mathcal{E}(\Lambda)$ is the likelihood set of level Λ for $X = x$ with respect to Θ', then*

$$W\left(\mathcal{E}(\Lambda), \Theta' \backslash \mathcal{E}(\Lambda); x\right) > \Lambda.$$

Proof. The result follows immediately from Proposition 1 and Definition 5.

In short, the practice of considering a parameter value insufficiently supported by the data if it falls outside a likelihood set receives some justification from measuring the strength of evidence for a composite hypothesis by its best-supported parameter value. However, since that practice is equivalent to weighing evidence for a simple hypothesis against that of a composite hypothesis in which it is essentially nested, it lacks interpretability in the sense of Sections 1.1 and 2.3.1 Non-interpretable procedures can be unsuitable for sequential data analysis (Section 2.4.4).

2.3.3. Bioequivalence illustration

Suppose θ is some scalar difference between two treatments that are considered bioequivalent if $\theta_- < \theta < \theta_+$ for two values θ_- and θ_+, which are often set by a regulatory agency. The bioequivalence testing problem is naturally framed as that of measuring the strength of evidence for $\theta \in (\theta_-, \theta_+)$ over $\theta \notin (\theta_-, \theta_+)$. In a Neyman-Pearson approach to bioequivalence, $\theta \in (\theta_-, \theta_+)$ is accepted if an interval of a sufficient level of confidence is a subset of (θ_-, θ_+). Choi, Cafo, and Rohde (2008) similarly consider there to be strong evidence of bioequivalence if a likelihood interval $\mathcal{E}(\Lambda)$ of sufficiently high level Λ is a subset of (θ_-, θ_+).

The latter approach is justified by the following implication of the explanatory theory of evidence (Section 2.2). In order to accommodate multidimensional parameters, the implication is stated in terms of equivalence sets and likelihood sets rather than equivalence intervals and likelihood intervals. Quantifying the strength of evidence for equivalence, $\theta \in \Theta'$, over nonequivalence, $\theta \notin \Theta'$, for some $\Theta' \subseteq \Theta$ corresponds to finding the likelihood set of highest level that is a subset of Θ':

Proposition 6. *The weight of evidence in $X = x$ that supports $\theta \in \Theta'$ over $\theta \notin \Theta'$ exceeds Λ if and only if $\mathcal{E}(\Lambda)$, the likelihood set of level Λ, is a subset of Θ'.*

Proof. From $\mathcal{E}(\Lambda) \subseteq \Theta'$, the definition of a likelihood set gives

$$\forall_{\theta'' \notin \Theta'} \exists_{\theta' \in \Theta'} L\left(\theta''; x\right) \Lambda < L\left(\theta'; x\right),$$

requiring that $\sup_{\theta' \in \Theta'} \inf_{\theta'' \notin \Theta'} L(\theta'; x) / L(\theta''; x) > \Lambda$, the left-hand side of which equals $W(\Theta', \Theta \backslash \Theta'; x)$ by Proposition 1, proving sufficiency. To prove necessity, assume there is a value θ'' that is in $\mathcal{E}(\Lambda)$ but not in Θ'. Given $W(\Theta', \Theta \backslash \Theta'; x) > \Lambda$, Proposition 1 yields $\sup_{\theta' \in \Theta'} L(\theta'; x) > \Lambda L(\theta''; x)$ since $\theta'' \in \Theta \backslash \Theta'$. Because $\theta'' \in \mathcal{E}(\Lambda)$, we have $\sup_{\theta' \in \Theta} L(\theta'; x) \leq \Lambda L(\theta''; x)$, producing a contradiction.

2.4. Multiplicity

2.4.1. Simultaneous inference

In a typical problem commonly encountered in high-dimensional biology, there are multiple focus subparameters $\theta_1, \ldots, \theta_D$ with the corresponding hypotheses $\theta_1 \in \Theta'_1, \ldots, \theta_D \in \Theta'_D$ such that $\theta = \langle \theta_1, \ldots, \theta_D \rangle$ and $\Theta'_1 \times \cdots \times \Theta'_D \subset \Theta_1 \times \cdots \times \Theta_D = \Theta$. A necessary and sufficient condition for multiple hypotheses to hold simultaneously is that the parameter of interest is in the intersection of their representative sets. For example, $\theta_1 \in \Theta'_1$ and $\theta_2 \in \Theta'_2$, i.e., $\langle \theta_1, \theta_2 \rangle \in \Theta'_1 \times \Theta_2$ and $\langle \theta_1, \theta_2 \rangle \in \Theta_1 \times \Theta'_2$, if and only if $\langle \theta_1, \theta_2 \rangle \in (\Theta'_1 \times \Theta_2) \cap (\Theta_1 \times \Theta'_2) = \Theta'_1 \times \Theta'_2$. In the same way, whether one or more of multiple hypotheses holds is equivalent to whether the parameter of interest is in the union of their representative sets. The simultaneous inference problem is thereby reduced to a composite hypothesis problem to which the laws of likelihood apply without modification.

According to the models most widely used in bioinformatics, each focus subparameter generates data independent of the data of the other focus subparameters:

$$f(x; \theta, \gamma) = \prod_{i=1}^{D} f_i(x_i; \theta_i, \gamma), \tag{2.13}$$

where $\langle x_1, \ldots, x_D \rangle = x$ and $f_i(\bullet; \theta_i)$ is the probability density or mass function of X_i for $i \in \{1, \ldots, D\}$. The likelihood function on Θ_i is $L_i = L_i(\bullet; x_i)$. Then the weight of evidence for $\theta_i \in \Theta'_i$ over $\theta_i \in \Theta''_i$ is simply

$$\begin{aligned} W_i\left(\Theta'_i, \Theta''_i\right) &= W\left(\Theta_1 \times \cdots \times \Theta'_i \times \cdots \times \Theta_D, \Theta_1 \times \cdots \times \Theta''_i \times \cdots \times \Theta_D\right) \\ &= \frac{\sup_{\theta'_i \in \Theta'_i} L_i(\theta'_i; x_i)}{\sup_{\theta''_i \in \Theta''_i} L_i(\theta''_i; x_i)}, \end{aligned}$$

according to (2.9). Likewise, the weight of evidence for $\theta_1 \in \Theta'_1$ and $\theta_2 \in \Theta'_2$ over $\theta_1 \in \Theta'_1$ alone is

$$W\left(\Theta'_1 \times \Theta'_2 \times \Theta_3 \times \cdots \times \Theta_D, \Theta'_1 \times \Theta_2 \times \cdots \times \Theta_D\right) = \frac{\sup_{\theta'_2 \in \Theta'_2} L_2(\theta'_2; x)}{\sup_{\theta_2 \in \Theta_2} L_2(\theta_2; x)},$$

and so on. The weight of evidence for $\theta_1 \in \Theta_1'$ and $\theta_2 \in \Theta_2'$ over $\theta_1 \notin \Theta_1'$ and $\theta_2 \notin \Theta_2'$ is at least as large as that for $\theta_1 \in \Theta_1'$ and $\theta_2 \in \Theta_2'$ over $\theta_1 \notin \Theta_1'$ or $\theta_2 \notin \Theta_2'$, that is, with $\overline{\Theta_1'} = \Theta \backslash \Theta_1'$,

$$\frac{\sup_{\theta_1' \in \Theta_1'} L_1\left(\theta_1'\right)}{\sup_{\theta_1'' \notin \Theta_1'} L_1\left(\theta_1''\right)} \frac{\sup_{\theta_2' \in \Theta_2'} L_2\left(\theta_2'\right)}{\sup_{\theta_2'' \notin \Theta_2'} L_2\left(\theta_2''\right)} \geq \frac{\sup_{\theta_1' \in \Theta_1'} L_1\left(\theta_1'\right) \sup_{\theta_2' \in \Theta_2'} L_2\left(\theta_2'\right)}{\sup_{\left\langle \theta_1'', \theta_2'' \right\rangle \in \left(\overline{\Theta_1'} \times \Theta\right) \cup \left(\Theta \times \overline{\Theta_2'}\right)} L_1\left(\theta_1''\right) L_2\left(\theta_2''\right)}.$$

Example 5. Supposing D murder trials take place on a certain day, let $\theta_i = 0$ if the ith defendant is neither guilty of manslaughter nor murder, $\theta_i = 1$ if guilty of manslaughter, and $\theta_i = 2$ if guilty of murder. Since the evidence presented in each trial does not depend on that of other trials, the weight of evidence that defendant i is guilty of murder is

$$W_i\left(\{2\}, \{0, 1\}\right) = \frac{L_i\left(2; x_i\right)}{L_i\left(0; x_i\right) \vee L_i\left(1; x_i\right)}.$$

The independence condition is not always as appropriate as in the example: it would produce erroneous results in Example 9.

2.4.2. Multiple-comparison adjustments

It is often maintained that multiple comparisons such as those made in the analysis of microarray data call for adjustments to reported levels of evidence that would be obtained for single comparisons. Such adjustments are almost invariably justified by a desire to control a false discovery rate or other generalized Type I error rate. For example, Korn and Freidlin (2006) regard the repeated application of the law of likelihood as highly dangerous since it treats the number of comparisons performed as evidentially irrelevant. Indeed, because the special law of likelihood quantifies the strength of evidence associated with each comparison rather than controlling a rate of false positives, the strength of evidence for one hypothesis over another remains the same irrespective of the number of comparisons made (Blume (2002)). More generally, while the approach based on the laws of likelihood accounts for data dependence between comparisons (Section 2.4.1), it is not modified to control error rates. In fact, the rationale for such control applies even under the independence of the data associated with each comparison.

Example 6. Since Korn and Freidlin (2006) liken the problem of multiple comparisons to that of selective reporting, consider a drug company that replicates N independent microarray experiments each yielding n measured ratios of expression between paired treatment and control mice for each of D genes under essentially the same conditions. For the jth experiment, the company calculates the weight of evidence in expression ratios $X_i^{(j)} = x_i^{(j)}$ for $\theta_i \in \Theta'$, the hypothesis that the ith gene is differentially expressed, over $\theta_i \in \Theta''$, the hypothesis that it is equivalently expressed between treatment and control. However, the company only reports to the regulatory agency which genes have decisive evidence of differential expression within each experiment along with the details of the statistical model and selection process. For any given gene, the process of selection clearly has no impact on the probability of observing misleading evidence. Let $y_i^{(j)} = 1$ if the ith gene has decisive evidence of differential expression in the jth experiment and $y_i^{(j)} = 0$ otherwise. The cumulative weight of evidence in the censored or reduced data for the ith gene under the simplifying assumption of independence (2.13) is

$$\begin{aligned} w_i\left(\Theta'_i, \Theta''_i\right) &= \frac{\sup_{\theta'_i \in \Theta'} \prod_{j=1}^{N} P_{\theta'_i}\left(1_{[100,\infty)}\left(W\left(\Theta', \Theta''; X_i^{(j)}\right)\right) = y_i^{(j)}\right)}{\sup_{\theta''_i \in \Theta''} \prod_{j=1}^{N} P_{\theta''_i}\left(1_{[100,\infty)}\left(W\left(\Theta', \Theta''; X_i^{(j)}\right)\right) = y_i^{(j)}\right)} \\ &= \frac{\sup_{\theta'_i \in \Theta'} \left(\alpha_{\theta'}\left(100; \Theta', \Theta''\right)\right)^{N_1} \left(1 - \alpha_{\theta'}\left(100; \Theta', \Theta''\right)\right)^{N-N_1}}{\sup_{\theta''_i \in \Theta''} \left(\alpha_{\theta''}\left(100; \Theta', \Theta''\right)\right)^{N_1} \left(1 - \alpha_{\theta''}\left(100; \Theta', \Theta''\right)\right)^{N-N_1}}, \end{aligned}$$

where N_1 is the number of experiments for which $y_i^{(j)} = 1$, in the terminology of Definition 4. As $N \to \infty$, $N_1 \to N$ if $\theta_i \in \Theta'$ or $N_1 \to 0$ if $\theta_i \in \Theta''$, with the implication that $1_{(1,\infty)}\left(w_i\left(\Theta'_i, \Theta''_i\right)\right)$ is a weakly consistent estimator of $1_{\Theta'}(\theta)$ by a variant of Proposition 3. From this perspective of estimating $1_{\Theta'}(\theta)$, the loss in efficiency due to the selection-induced data reduction is not addressed by the control of an error rate.

The evidential interpretation of p-value adjusted for multiple comparisons has its roots in Fisher's disjunction: if the p-value is low, then either an event of low probability has occurred or the null hypothesis is false (Fisher (1925); Johnstone (1986); Barnard (1967)). Without some adjustment, a low p-value can instead occur with high probability given enough tests. Thus, even when the p-value is understood as a measure of evidence, the multiple testing problem is formulated in terms of error rate control. If a single hypothesis is tested at a given significance level α, then α is the probability of making a Type I error under the null hypothesis. However, if multiple hypotheses are each tested at level α, then the probability of at least one Type I error under the truth of all null hypotheses is greater than α except in the trivial case of complete dependence between test statistics. This probability is called the family-wise error rate (FWER). Consequently, a plethora of methods have been developed to control the FWER for various assumptions while retaining as much power to reject the null hypothesis as possible. The control of FWERs has been criticized for admitting many false negatives in order to avoid all false positives in most samples, and newer criteria for judging significance gain power by allowing more false positives. Such criteria include control of the probability that false positives exceed a given number or proportion (Van der Laan, Dudoit, and Pollard (2004)). A less conservative multiple comparison procedure controls the false discovery rate (FDR), the expectation value of the ratio of the number of Type I errors to the number of rejected null hypotheses (Benjamini and Hochberg (2000); Benjamini et al. (2001); Benjamini and Yekutieli (2005); Yekutieli et al. (2006); Benjamini and Liu (1999)). The smallest FDR at which a hypothesis is rejected (Storey (2002)) is offered in many microarray data analysis programs as a corrected or adjusted p-value; e.g., Pollard et al. (2005). All of these approaches replace control of the test-wise error rate with control of a different Type I error rate, and all may lead to a corrected p-value for each null hypothesis considered (Van der Laan, Dudoit, and Pollard (2004)).

Considering the p-value as a measure of statistical evidence that must be adjusted to continue to measure statistical evidence under multiple comparisons has been formally justified as follows. In significance testing, the observed p-value is viewed as the probability that a true null hypothesis would be rejected under repeated sampling in the hypothetical case that the observed test statistic happened to lie on the boundary of the rejection region (Cox (1977)). Here, the rejection region is purely hypothetical since no decision to reject or not reject the null hypothesis is made on the basis of any error rate actually selected before observation, as the Neyman-Pearson framework would require. That significance testing interpretation of the p-value lies behind defining the adjusted p-value of a null hypothesis as the lowest Type I error rate of a test at which the null hypothesis would be rejected (Shafer (1995)). This overall Type I error rate is usually a family-wise error rate, a generalization thereof, or a false discovery rate (Van der Laan, Dudoit, and Pollard (2004)). This formalism of defining a corrected p-value in terms of controlling an error rate is combined with the motivation behind reporting a corrected p-value rather than a decision on the rejection of the hypothesis, namely, the corrected p-value quantifies the strength of evidence against the null hypothesis (Wright (1992)). Evidentially interpreting a p-value corrected in order to control a hypothetical Type I error rate exemplifies what Goodman (1998) and Johnstone (1986) noted of significance testing in general: Neymanian theory fuels Fisherian practice.

The argument that p-values must be corrected to control a Type I error rate would obtain even in the absence of information about the distribution of interest in data from other distributions. This raises the question of whether an adjusted p-value or an unadjusted quantity such as a raw p-value or likelihood ratio better measures the weight of evidence with respect to one of several comparisons. Example 5 may clarify the issue. In weighing the evidence for and against the hypothesis that a defendant is guilty, should the jury take into account the number of defendants currently under trial for the same crime elsewhere in the country, perhaps to control a rate of false convictions, or is that information irrelevant to task of assessing the strength of evidence for guilt over innocence in the trial at hand? As Mayo and Cox (2006) argued, while controlling family-wise error rates may prove advantageous in certain contexts in which the goal of data analysis is to determine a course of action, the uncorrected p-value is more appropriate in contexts where inductive reasoning or evidence evaluation is the aim. Such clarification of the purpose behind data analysis is crucial, for confusing the weight of statistical evidence with how that evidence should be used can have undesired consequences. Since Fisher (1973, pp.95-96, 103-106), a primary argument for measuring evidential strength rather than computing optimal decisions has relied on the unpredictability of the use to which evidence will be put. While the nuisance parameter problem may often make complete separation between evidence and application impossible even when guided by the explanatory theory of evidence (Section 2.5), such distinction remains an ideal worth approaching, at least in basic science.

A non-decision-theoretic context suggesting adjustment of p-values is that in which it is believed that "most of the individual null hypotheses are essentially correct" (Cox (2006, p.88)), thereby to some extent combining the strength of evidence in the data with that of one's prior confidence. The same purpose is served more precisely and frankly by assigning prior probability to each of the null hypotheses in proportion to such confidence (Westfall, Johnson, and Utts (1997)).

The observation that correcting p-values for selection has decision-theoretic rather than inferential or evidential rationales does not mean an evidential rationale for such correction will never be formulated. That would be accomplished either by arguing without appeal to the control of error rates, to optimality, or to other decision-theoretic concepts or by demonstrating that the problem of evidence cannot be separated from the problem of decision. For related discussions on the distinction between the decision problem and the inference problem, see Fisher (1973), Edwards (1992, Appendix I), Hald (2007), Montazeri, Yanofsky and Bickel (2010), and Bickel (2011a).

Evidential inference based directly on the law of likelihood is only beginning to find applications in extreme multiple comparison situations. Taking a first step, Strug and Hodge (2006) studied the implications of evidential inference as an alternative to Neyman-Pearson error rate control in linkage analysis. They find that although consideration of error rates informs study design, their use in correcting p-values distorts the strength of evidence.

2.4.3. Empirical Bayes

The error-control rationale for adjusting p-values is distinct from the rationale behind empirical Bayes methods formulated in order to "borrow strength" or available information from distributions besides the distribution corresponding to the comparison at hand. The latter rationale motivates some applications to genomic expression data since it is believed that measurements of the expression of some genes are informative for inference about the expression of other genes. It is also consistent with the uncontested applicability of Bayes's theorem in the presence of a distribution of parameter frequencies (Fisher (1973); Wilkinson (1977); Edwards (1992); Kyburg and Teng (2006); (Hald, 2007, p.36); Fraser (2011)), a situation in which few would insist on corrections to control the FWER or FDR when the problem is one of inference rather than decision.

Typical empirical Bayes methods rely on modeling parameter values as random variables of a physical distribution p intended to reflect actual variability rather than levels of belief. While that approach often leads to competitive performance (Yanofsky and Bickel (2010); Montazeri, Yanofsky and Bickel (2010)) or even optimality under some class of loss functions, its relevance to objectively weighing evidence has received little attention. Section 5.2 explores the use of a successful empirical Bayes method for inference under the special law of likelihood in the context of microarray data analysis.

2.4.4. Sequential data analysis

The consideration of stopping times in settings involving sequential analysis, like that of error rates in settings involving multiple comparisons, is relevant to study design (Berger and Wolpert (1988)) but not to measuring the strength of evidence under the likelihood principle (Blume (2008)). An "unscrupulous investigator" may attempt to conclude that $\theta \in \Theta'$ by supplementing a sample of n_- independent and identically distributed (IID) observations $x_1, \ldots, x_{n_-}$ with additional IID observations $x_{n_-+1}, \ldots, x_n$ just until $W(\Theta', \Theta'') \geq \Lambda$, called the *stopping condition*, where $\Theta' \cap \Theta'' = \emptyset$, $\Theta' \cup \Theta'' = \Theta$, and Λ is the desired level of evidence ($\Lambda > 1$).

Let $X^{(n)} = \langle X_1, \ldots, X_n \rangle$ denote the n-tuple of IID random variables of which $x_1, \ldots, x_n$ are realizations, and assume that $\theta = E_\theta(X_1)$ for some $\theta \in \Theta''$ and that $\sigma^2 = \mathrm{Var}_\theta(X_1)$ is known and finite. The probability that the investigator can ever successfully support the false hypothesis depends on the construction of Θ''.

If $\theta = \theta''$ is the hypothesis the investigator endeavors to reject, then $\Theta'' = \{\theta''\}$. For any finite n_- and Λ, the number of additional observations needed to satisfy the stopping condition is almost surely finite according to the law of the iterated logarithm (Robbins (1970)). In other words, the probability of eventually observing misleading evidence is 1. By implication, as more data are obtained indefinitely, a level-Λ likelihood interval that does not contain θ will almost always occur. An anonymous reviewer pointed out this objection to likelihood sets as defined in Section 2.3.2.

The ability of the investigator to sample until achieving the desired conclusion regardless of the initial study size n_- is a consequence of the noninterpretability of $W\left(\mathbb{R}^1\backslash\left\{\theta''\right\},\left\{\theta''\right\}\right)$ that was noted in Sections 1.1 and 2.3.1. It will now be seen that the use of an interpretable weight of evidence solves the problem.

Proposition 7. *For any $\Theta'' \subset \Theta$ such that its interior* $\operatorname{int}\Theta''$ *contains θ, any $\alpha \in (0,1]$, and any $\Lambda > 1$, there exists a counting number n_- such that*

$$P_\theta\left(\exists n \in \{n_- + 1, n_- + 2, \ldots\} : W\left(\Theta\backslash\Theta'', \Theta''; X^{(n)}\right) \geq \Lambda\right) \leq \alpha$$

under regularity conditions ensuring the strong consistency of $\widehat{\theta}^{(n)} = \widehat{\theta}\left(X^{(n)}\right)$, the maximum likelihood estimate of θ, where $X^{(n)}$ is a random n-tuple on a basic probability space $(\Omega, \Sigma, P_\theta)$.

Proof. $P_\theta\left(\lim_{n\to\infty}\widehat{\theta}^{(n)} \in \operatorname{int}\Theta''\right) = 1$ by the strong consistency of $\widehat{\theta}^{(n)}$, implying

$$\lim_{n_-\to\infty} P_\theta\left(\forall n \in \{n_- + 1, n_- + 2, \ldots\} : \widehat{\theta}^{(n)} \in \Theta''\right) = 1;$$

$$\lim_{n_-\to\infty} P_\theta\left(\exists n \in \{n_- + 1, n_- + 2, \ldots\} : W\left(\Theta\backslash\Theta'', \Theta''; X^{(n)}\right) > 1\right) = 0.$$

If the data are censored, e.g., by the drug company of Example 6, just until satisfying the stopping condition, then the observation is $\left\langle X^{(N)}, N\right\rangle = \left\langle x^{(n)}, n\right\rangle$, where the reported sample size n is a realization of the random quantity N. In that case, the likelihood function for the purposes of weighing evidence or, alternatively, for performing Bayesian inference, is specified by

$$L\left(\theta; \left\langle x^{(n)}, n\right\rangle\right) = f\left(x^{(n)}; \theta\right) P_\theta\left(N = n\right)$$

rather than simply by $L\left(\theta; x^{(n)}\right) = f\left(x^{(n)}; \theta\right)$, as in the absence of censoring. To see that in the discrete-data case, note that $P_\theta\left(X^{(N)} = x^{(n)}, N = n\right) = P_\theta\left(X^{(N)} = x^{(n)} | N = n\right) P_\theta\left(N = n\right)$. The factor $P_\theta\left(N = n\right)$ automatically accounts for the stopping rule without any ad hoc adjustments.

2.5. Nuisance parameters

2.5.1. Elimination of nuisance parameters

Suppose the family of distributions is parameterized by a free nuisance parameter $\gamma \in \Gamma \subseteq \mathbb{R}^\nu$ as well as by the free interest parameter $\theta \in \Theta \subseteq \mathbb{R}^D$ such that neither θ nor γ is a function of the other parameter; both ν and D are fixed positive integers. The likelihood function corresponding to each probability density or mass function $f(\bullet;\theta,\gamma)$ on Ω is $\ell(\bullet)=\ell(\bullet;x)=f(x;\bullet)$ on $\Theta\times\Gamma$.

The problem of measuring the weight of evidence in the presence of a nuisance parameter has been posed as a problem of approximating the weight of evidence that would be in the data were the value of the nuisance parameter known (Tsou and Royall (1995)). The nuisance parameter is often eliminated by replacing the unknown likelihood function ℓ on $\Theta\times\Gamma$ with a known *reduced likelihood function* L on Θ such as an integrated likelihood function, a conditional likelihood function, a marginal likelihood function, an estimated likelihood function, or a profile likelihood function.

Applying that approach to composite hypotheses, a reduced likelihood function is chosen to approximate $W_\gamma(\Theta',\Theta'')=\sup_{\theta'\in\Theta'}\ell(\theta',\gamma)/\sup_{\theta''\in\Theta''}\ell(\theta'',\gamma)$, the weight of evidence for $\theta\in\Theta'$ over $\theta\in\Theta''$, since $W_\gamma(\Theta',\Theta'')$ is unknown without knowledge of γ. Some of the reduced likelihood functions provide better approximations than others, depending on the family of distributions, as will be seen in Section 2.5.2. Once the nuisance parameters have been eliminated, the reduced likelihood function L on Θ takes the place of the likelihood function, yielding $\sup_{\theta'\in\Theta'}L(\theta')/\sup_{\theta''\in\Theta''}L(\theta'')$ to approximate $W_\gamma(\Theta',\Theta'')$.

The elimination of nuisance parameters is exemplified here with the *profile likelihood function* L_{profile}, defined by $\forall_{\theta\in\Theta}L_{\text{profile}}(\theta)=L_{\text{profile}}(\theta;x)=\sup_{\gamma\in\Gamma}$ $\ell(\theta,\gamma;x)$. Under the special law of likelihood, the profile likelihood ratio L_{profile} $(\theta';x)/L_{\text{profile}}(\theta'';x)$ serves as a widely applicable approximation to the weight of evidence in $X=x$ for $\theta=\theta'$ over model $\theta=\theta''$. Likewise, the strength of evidence in $X=x$ that supports $\theta\in\Theta'$ over $\theta\in\Theta''$ may be approximated by

$$W_{\text{profile}}\left(\Theta',\Theta''\right)=W_{\text{profile}}\left(\Theta',\Theta'';x\right)=\frac{\sup_{\theta'\in\Theta'}L_{\text{profile}}(\theta';x)}{\sup_{\theta''\in\Theta''}L_{\text{profile}}(\theta'';x)}, \tag{2.14}$$

provided that each hypothesis is either simple or complex.

Example 7. The proposed methodology is illustrated with the comparison of the hypotheses $|\theta|>\theta_+$ and $|\theta|\le\theta_+$ for some $\theta_+\ge 0$ on the basis of $x=\left(x^{(1)},\ldots,x^{(n)}\right)$, a sample of n independent observations from a normal distribution with unknown mean $\theta\in\mathbb{R}^1$ and variance $\gamma=\sigma^2\in(0,\infty)$. Hence, the density function satisfies

$$f\left(x;\theta,\sigma^2\right)=\prod_{j=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\left(\frac{x^{(j)}-\theta}{\sigma}\right)^2\right). \tag{2.15}$$

Since

$$L_{\text{profile}}\left(\theta'\right)/L_{\text{profile}}\left(\theta''\right)=\left(\widehat{\sigma}\left(\theta'\right)/\widehat{\sigma}\left(\theta''\right)\right)^{-n},$$

the strength of evidence for $|\theta| > \theta_+$ over $|\theta| \leq \theta_+$ is

$$W_{\text{profile}}\left(\mathbb{R}^1 \backslash \left[-\theta_+, \theta_+\right], \left[-\theta_+, \theta_+\right]\right) = \begin{matrix} \inf_{\theta'' \in [-\theta_+, \theta_+]} \left(\frac{\widehat{\sigma}}{\widehat{\sigma}(\theta'')}\right)^{-n} & \left|\widehat{\theta}\right| > \theta_+ \\ \sup_{\theta' \in [-\theta_+, \theta_+]} \left(\frac{\widehat{\sigma}(\theta')}{\widehat{\sigma}}\right)^{-n} & \left|\widehat{\theta}\right| \leq \theta_+ \end{matrix},$$

where $\widehat{\theta}$ and $\widehat{\sigma} = \widehat{\sigma}\left(\widehat{\theta}\right)$ are the maximum likelihood estimates of θ and σ. In bioequivalence applications (Section 2.3.3),

$$W_{\text{profile}}\left(\left(-\theta_+, \theta_+\right), \mathbb{R}^1 \backslash \left(-\theta_+, \theta_+\right)\right) = 1/W_{\text{profile}}\left(\mathbb{R}^1 \backslash \left(-\theta_+, \theta_+\right), \left(-\theta_+, \theta_+\right)\right)$$

approximates the evidence for equivalence.

The profile likelihood has several advantages as an approximation: it resembles a likelihood ratio under certain conditions and has a low asymptotic probability of misleading evidence (Royall (2000a)), and, if the nuisance parameter is orthogonal to the interest parameter, it is equal to the likelihood ratio (Royall (1997)). For some models, the nuisance parameter may instead be eliminated by use of a conditional or marginal likelihood (Royall (1997)) as approximations of the likelihood function without nuisance parameters. The latter is convenient and reliable for cases in which a test statistic carries most of the relevant information in the data (Schweder and Hjort (2002); Johnson (2005); Royall (1997)).

Alternatively, provided a probability distribution or other measure of γ that is suitable for evidential inference, the nuisance parameter could be eliminated by integration. Methods have been proposed for specifying a nuisance parameter distribution or other measure to integrate the likelihood not only for Bayesian statistics (Kass and Raftery (1995); Berger, Liseo, and Wolpert (1999); Kass and Raftery (1995); Berger, Liseo, and Wolpert (1999); Clyde and George (2004)) but also for Neyman-Pearson statistics (Severini (2007, 2010)). In fact, the nuisance parameter measure need not be a pure prior distribution since it may depend on data (Kalbfeisch and Sprott (1970); Aitkin (1991); Dempster (1997); Severini (2007, 2010)).

This flexibility of choice in the method for eliminating nuisance parameters allows researchers to tailor data analyses to particular applications such as that of Example 9, underscoring the fact that the motivating objectivity condition of Section 1.1 by no means reduces statistical inference to a series of automatic calculations. On the other hand, different approaches to eliminating nuisance parameters can yield similar results. For example, likelihoods integrated with respect to certain distributions approximate the profile likelihood (Severini (2007)).

2.5.2. Other interpretations of profile likelihood

Instead of seeing the profile likelihood as one of many possible approximations of the unknown likelihood function of the interest parameter, the profile likelihood could be derived from Proposition 1 by framing the nuisance parameter problem as an instance of the composite hypothesis problem as follows. Royall (1992), He, Huang, and Liang (2007), and Zhang (2009b), contrary to Section 2.5.1, identified the weight of evidence for $\langle\theta,\gamma\rangle \in \Theta' \times \Gamma$ over $\langle\theta,\gamma\rangle \in \Theta'' \times \Gamma$ with the weight of evidence for $\theta \in \Theta'$ over $\theta \in \Theta''$, thus assuming that the latter can be precisely known without knowledge of γ. Under that conflation of the problem of composite hypotheses with the problem of nuisance parameters, (2.14) would exactly specify the weight of evidence for $\theta \in \Theta'$ over $\theta \in \Theta''$, as it does in the axiomatic system of Zhang (2009b).

However, there are families of distributions in which the profile likelihood can fail to meaningfully measure the weight of evidence (Royall (2000a); He, Huang, and Liang (2007)). For that reason, Royall (1992) and He, Huang, and Liang (2007) represented the weight of evidence as an interval of profile likelihood ratios, and the weight was represented as a single likelihood ratio in Sections 2.2 and 2.5.1 that is unknown if there is an unknown nuisance parameter. The elimination of a nuisance parameter γ, whether by profiling, integration, or other means, only approximates $W_\gamma(\Theta',\Theta'')$.

The profile likelihood ratio (2.14) would much more plausibly measure the explanatory power of $\theta \in \Theta'$ relative to $\theta \in \Theta''$ than it would measure the weight of evidence for $\theta \in \Theta'$ over $\theta \in \Theta''$, provided that each hypothesis is either simple or complex, as seen in the following examples. More generally, the relative ability $\mathrm{ex}(\Theta',\Theta'')$ of $\theta \in \Theta'$ compared to $\theta \in \Theta''$ to explain $X = x$, is $\sup_{\gamma'\in\Gamma} \mathrm{ex}_{\gamma'}(\Theta') / \sup_{\gamma''\in\Gamma} \mathrm{ex}_{\gamma''}(\Theta'')$, where $\mathrm{ex}_{\gamma'}$ is an explanatory function for each $\gamma' \in \Gamma$ such that $\mathrm{ex}_{\gamma'}(\{\theta'\},\{\theta''\}) = \ell(\theta',\gamma')/\ell(\theta'',\gamma')$ for all $\theta',\theta'' \in \Theta$ and $\gamma' \in \Gamma$, assuming γ is fixed. The weight of evidence nonetheless remains $W_\gamma(\Theta',\Theta'')$, a function of γ.

Example 8. For any single observation x of a normal variate X of unknown mean γ and variance θ, the profile likelihood would ascribe infinite weight of evidence in that observation to the hypothesis that $\theta = 0$ over any $\theta \neq 0$, which is clearly untenable (Royall (2000a); He, Huang, and Liang (2007)). However, the hypothesis that $\theta = 0$, if true, would explain the observation much better than would any other simple hypothesis about θ, resonating with interpreting the profile likelihood ratio as a measure of relative explanatory power.

The Neyman-Scott problem also precludes viewing profile likelihood as evidence (Royall (1992); He, Huang, and Liang (2007)) but accords with viewing it as explanatory power. Less pathological problems point to the same conclusion.

Example 9. In a scenario posed by an anonymous reviewer, exactly D insiders know a secret. The probability that the secret does ($x = 1$) or does not ($x = 0$) leak is

$$f\left(x; \theta_1, \ldots, \theta_D\right) = \left(1 - \prod_{i=1}^{D} \left(1 - \theta_i\right)\right)^{x} \left(\prod_{i=1}^{D} \left(1 - \theta_i\right)\right)^{1-x},$$

where $\theta_i = 1$ if the ith insider leaks the secret and $\theta_i = 0$ otherwise. If the secret leaks ($X = 1$), the likelihood function is given by

$$L\left(\theta_1, \ldots, \theta_D; 1\right) = \begin{cases} 1 & \text{if } \prod_{i=1}^{D} \left(1 - \theta_i\right) = 0, \\ 0 & \text{if } \prod_{i=1}^{D} \left(1 - \theta_i\right) = 1. \end{cases}$$

Thus, the leaking of a secret constitutes irrefutable evidence that at least one of the insiders leaked it over the hypothesis $\langle \theta_1, \ldots, \theta_D \rangle = \langle 0, \ldots, 0 \rangle$ that none of them leaked it:

$$W\left(\{0,1\}^D \Big\backslash \left\{\langle 0, \ldots, 0 \rangle\right\}, \left\{\langle 0, \ldots, 0 \rangle\right\}\right) = \frac{1}{0}.$$

However, the evidence against any given suspect is much weaker. Quantifying $W_i\left(\{1\}, \{0\}\right)$, the weight of evidence that $\theta_i = 1$ over $\theta_i = 0$, treats θ_i as the interest parameter and $\gamma = \langle \theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_D \rangle$ as the nuisance parameter. In this case, eliminating the latter by means of the profile likelihood yields

$$W_i\left(\{1\}, \{0\}\right) = \frac{\sup_{\gamma \in \{0,1\}^{D-1}} L\left(\theta_1, \ldots, \theta_{i-1}, 1, \theta_{i+1}, \ldots, \theta_D; 1\right)}{\sup_{\gamma \in \{0,1\}^{D-1}} L\left(\theta_1, \ldots, \theta_{i-1}, 0, \theta_{i+1}, \ldots, \theta_D; 1\right)} = 1,$$

which is reasonable for sufficiently large D. For small D, the integration method of eliminating nuisance parameters is more reasonable since it allows $W_i\left(\{1\}, \{0\}\right)$ to be close to but greater than 1. This can be accomplished without recourse to subjective or conventional priors by modeling each θ_i as an independent Bernoulli random variable of limiting relative frequency $p\left(\theta_i\right) \in (0, 1)$, entailing that $p(1)$ or $p(0)$ is the frequentist probability that the ith insider does or does not reveal the secret. Then the integration method gives

$$\begin{aligned} &W_i\left(\{1\}, \{0\}\right) \\ &\quad = \frac{\int L\left(\theta_1, \ldots, \theta_{i-1}, 1, \theta_{i+1}, \ldots, \theta_D; 1\right) dp\left(\theta_1\right) \cdots dp\left(\theta_{i-1}\right) dp\left(\theta_{i+1}\right) \cdots dp\left(\theta_D\right)}{\int L\left(\theta_1, \ldots, \theta_{i-1}, 0, \theta_{i+1}, \ldots, \theta_D; 1\right) dp\left(\theta_1\right) \cdots dp\left(\theta_{i-1}\right) dp\left(\theta_{i+1}\right) \cdots dp\left(\theta_D\right)} \\ &\quad = \frac{1}{1 - p(0)^{D-1}} > 1, \end{aligned}$$

approaching the result of the profile likelihood as $D \to \infty$, but $W_i\left(\{1\}, \{0\}\right) = 1/p(1)$ for $D = 2$.

Thus, the integrated likelihood is more reasonable than the profile likelihood for weighing the evidence that a given insider revealed a secret since the use of the latter would mean the revelation has no evidence to that effect even if there were only two insiders. The profile likelihood may nonetheless quantify explanatory power, in which case the hypothesis that a given insider revealed the secret would not explain its revelation better than would the hypothesis that he or she did not reveal it.

3. Inference about imprecise hypotheses

Since the boundary between one composite hypothesis and another is often arbitrary to a large extent, the effect of specifying that boundary will be mitigated by making it imprecise or, more technically, fuzzy. An objection against the use of fuzzy logic is that problems solved using fuzzy set theory can be solved using probability theory instead (Laviolette (2004)). However, whereas in the context of statistical inference, probability is usually seen in terms of the representation of uncertainty, there is no uncertainty associated with hypothesis specification as envisioned here. Because the specification of hypotheses does not depend on frequencies of events or levels of belief, fuzzy set membership functions rather than probability distributions will be used to specify hypotheses in order to avoid confusion. This approach is in line with traditional interpretations of degrees of set membership (Klir (2004); Nguyen and Walker (2000)) as opposed to reinterpreting them as degrees of uncertainty as per Singpurwalla and Booker (2004). By keeping vagueness or imprecision distinct from uncertainty, fuzzy set theory enables a clearer presentation of the proposed methodology than would be possible with the probability calculus alone. Thus, the proposed methodology remains objective in the sense that the strength of evidence for a given hypothesis over another given hypothesis does not depend on any researcher's prior levels of belief even though each given hypothesis may have an imprecise specification.

The use of vague hypotheses to broaden the framework of Section 2 has a different motivation than related work on the interface between statistics and fuzzy logic. Fuzzy set theory has been used to specify vague hypotheses for generalizations of both Neyman-Pearson hypothesis testing (Romer, Kandel, and Backer (1995)) and Bayesian inference (Zadeh (2002)). Similarly, Dollinger, Kulinskaya, and Staudte (1996) suggested measuring evidence by the extent to which a test statistic falls in a fuzzy rejection region determined by a fixed Type I error rate; this leads to fuzzy hypothesis tests and fuzzy confidence intervals. Fuzzy hypothesis tests and fuzzy confidence intervals have also been formulated to overcome a flaw in previous methods involving discrete distributions (Geyer and Meeden (2005)).

3.1. Incomplete likelihood

A measure P of total mass $c = \int dP$ is a *complete*, *incomplete*, or *strictly incomplete* probability distribution of *completeness* c if $c = 1$, $0 < c \leq 1$ or $0 < c < 1$, respectively (Rényi (1970, p.569)). Consider the family

$$\left\{P_{\langle\theta',c'\rangle} : \theta' \in \Theta, c' \in (0,1]\right\} \tag{3.1}$$

of incomplete probability distributions on Ω such that $P_{\langle\theta',c'\rangle}(\bullet) = c'P_{\langle\theta',1\rangle}(\bullet)$, where θ', γ', and c' are the interest parameter value, nuisance parameter value, and level of completeness that uniquely specify $P_{\langle\theta',c'\rangle}$. Denote each complete distribution $P_{\langle\theta',1\rangle}$ by $P_{\theta'}$. The true sampling distribution of X is P_θ with θ unknown.

The *incomplete likelihood function* $\widetilde{L}(\bullet) = \widetilde{L}(\bullet; x)$ on $\Theta \times (0,1]$ satisfies $\widetilde{L}(\theta, c) = f(x; \theta, c)$ for all $\langle\theta, c\rangle \in \Theta \times (0,1]$, where $f(\bullet; \theta, c)$ is an incomplete probability mass or density function of $P_{\langle\theta,c\rangle}$. Thus, $\widetilde{L}(\theta; x) = \widetilde{L}(\bullet, 1; x)$ is the Fisherian or *complete likelihood function.* For all $\theta \in \Theta$ and $c \in (0,1]$, the identity $\widetilde{L}(\theta, c; x) = c\widetilde{L}(\theta; x)$ follows from the parameterization (3.1) since it requires that $f(x; \theta, c) = cf(x; \theta, 1)$.

3.2. Imprecise hypotheses

In order to concisely represent hypothesis imprecision in terms of incomplete probability distributions, the subsection employs concepts from fuzzy set theory.

Definition 7. Any measurable function that maps Θ to $[0,1]$ is a *fuzzy subset* of Θ.

Following Nguyen and Walker (2000), this definition makes no distinction between a fuzzy subset and its membership function; $\widetilde{\Theta}'(\theta)$ is considered to be the extent to which θ belongs to a fuzzy subset $\widetilde{\Theta}'$ of Θ, summarized as $\theta\tilde{\in}\widetilde{\Theta}'$. The $\tilde{\in}$ symbol plays the role of the $\in$ symbol in order to specify a hypothesis in terms of membership in a fuzzy subset, which is literally a function rather than a set of parameter values. The meaning of "the hypothesis $\theta\tilde{\in}\widetilde{\Theta}'$ is true to extent $\widetilde{\Theta}'(\theta)$" depends on whether θ is random according to a physical distribution, as will be seen in the remainder of this subsection. Each such $\widetilde{\Theta}'$ corresponding to a hypothesis must be a member of $\mathcal{F}(\Theta)$, the set of all fuzzy subsets of Θ such that $\widetilde{\Theta}' \in \mathcal{F}(\Theta) \implies \exists\theta \in \Theta : \widetilde{\Theta}'(\theta) = 1$.

If θ is random with sampling distribution p and if $\Theta' \in \mathcal{A}$, then the generalized probability of $\theta\tilde{\in}\widetilde{\Theta}'$, conditional on some event $X \in \Omega'$, is defined as

$$\tilde{\mathbf{P}}\left(\theta\tilde{\in}\widetilde{\Theta}'|X \in \Omega'\right) = \mathbf{E}\left(\widetilde{\Theta}'(\theta)|X \in \Omega'\right),$$

where $\tilde{\mathbf{P}}$ generalizes the probability measure $\mathbf{P}$, an extension of P_θ and p, and where $\mathbf{E}$ is the usual expectation operator $\mathbf{E}(\bullet) = \int \bullet d\mathbf{P}$. By construction, $\tilde{\mathbf{P}}$ obeys Bayes's rule:

$$\tilde{\mathbf{P}}\left(X \in \Omega' | \theta \tilde{\in} \widetilde{\Theta}'\right) = P_\theta\left(X \in \Omega'\right) \frac{\mathbf{E}\left(\widetilde{\Theta}'(\theta) | X \in \Omega'\right)}{\mathbf{E}\left(\widetilde{\Theta}'(\theta)\right)}.$$

Accordingly, each $\tilde{\mathbf{P}}\left(X \in \bullet | \theta \tilde{\in} \widetilde{\Theta}'\right)$ such that $\widetilde{\Theta}' \in \mathcal{F}(\Theta)$ is assumed to admit the generalized probability density or mass function $\overline{\widetilde{f}}\left(\bullet; \widetilde{\varphi}\left(\widetilde{\Theta}'\right)\right)$ satisfying

$$\overline{\widetilde{f}}\left(x; \phi'\right) = \mathbf{E}\left(f(x;\theta)\right) \frac{\mathbf{E}\left(\widetilde{\varphi}^{-1}(\phi')(\theta) | X = x\right)}{\mathbf{E}\left(\widetilde{\varphi}^{-1}(\phi')(\theta)\right)} \tag{3.2}$$

for all $x \in \Omega$ and $\phi' \in \widetilde{\Phi}$, where $\widetilde{\Phi}$ is a parameter set isomorphic to $\mathcal{F}$ by the invertible map $\widetilde{\varphi} : \mathcal{F} \to \widetilde{\Phi}$. Then the generalized likelihood function on $\widetilde{\Phi}$ for purposes of quantifying evidential weight and explanatory power is $\overline{\widetilde{L}}(\bullet) = \overline{\widetilde{L}}(\bullet; x) = \overline{\widetilde{f}}(x; \bullet)$. Thus, each composite hypothesis $\theta \tilde{\in} \widetilde{\Theta}'$ corresponds to a simple hypothesis $\phi = \widetilde{\varphi}\left(\widetilde{\Theta}'\right)$.

For the case of fixed θ, every imprecise hypothesis is equivalent to a precise hypothesis. Let $\xi_{\widetilde{\Theta}'}(\theta) = \left\langle \theta, \widetilde{\Theta}'(\theta) \right\rangle$ and $\Xi\left(\widetilde{\Theta}'\right) = \left\{\xi_{\widetilde{\Theta}'}(\theta) : \theta \in \Theta\right\}$ for all $\widetilde{\Theta}' \in \mathcal{F}(\Theta)$. Every parameter value ξ in $\Xi\left(\widetilde{\Theta}'\right)$ indexes P_ξ, a member of the family of incomplete probability distributions (3.1). Each imprecise hypothesis $\theta \tilde{\in} \widetilde{\Theta}'$ is called *simple*, *intrinsically simple*, or *complex* if the precise hypothesis $\xi_{\widetilde{\Theta}'}(\theta) \in \Xi\left(\widetilde{\Theta}'\right)$ is simple, intrinsically simple, or complex, respectively.

These calibrations of $\theta \tilde{\in} \widetilde{\Theta}'$ by distribution completeness values overcome the objection against fuzzy set theory that it fails to unambiguously assign fractional membership values (Lindley (2004)). The calibrations facilitate the extension of evidential theory to imprecise hypotheses by automatically attenuating the weight of evidence and explanatory power according to the imprecision.

3.3. Extended theory of evidence

For fuzzy subsets $\widetilde{\Theta}', \widetilde{\Theta}'' \in \mathcal{F}(\Theta)$, let $\widetilde{\widetilde{W}}\left(\widetilde{\Theta}', \widetilde{\Theta}''\right) = \widetilde{\widetilde{W}}\left(\widetilde{\Theta}', \widetilde{\Theta}''; x\right)$ denote the weight of evidence in $X = x$ that supports $\theta \tilde{\in} \widetilde{\Theta}'$ over $\theta \tilde{\in} \widetilde{\Theta}''$. The function $\widetilde{\widetilde{W}}$ is defined by transforming each imprecise hypothesis concerning complete probability distributions to an equivalent precise hypothesis concerning incomplete probability distributions in accordance with Section 3.2.

Definition 8. A function $\widetilde{W}$ on $\mathcal{F}(\Theta)\times\mathcal{F}(\Theta)\times\Omega$ is the *extended evidential function* with respect to an explanatory function ex if it satisfies the following conditions.

1. For all $\widetilde{\Theta}',\widetilde{\Theta}''\in\mathcal{F}(\Theta)$ such that $\theta\tilde{\in}\widetilde{\Theta}'$ and $\theta\tilde{\in}\widetilde{\Theta}''$ is each either a simple hypothesis or a complex hypothesis,
$$\widetilde{W}\left(\widetilde{\Theta}',\widetilde{\Theta}'';x\right)=W\left(\Xi\left(\widetilde{\Theta}'\right),\Xi\left(\widetilde{\Theta}''\right);x\right),\tag{3.3}$$
where W is any evidential function on $2^{\Theta\times(0,1]}\times 2^{\Theta\times(0,1]}\times\Omega$.
2. For all $\widetilde{\Theta}',\widetilde{\Theta}''\in\mathcal{F}(\Theta)$ such that $\theta\tilde{\in}\widetilde{\Theta}'$ and $\theta\tilde{\in}\widetilde{\Theta}''$ are intrinsically simple hypotheses, let $\overline{W}$ be any evidential function on $\left\{\{\phi\}:\phi\in\widetilde{\Phi}\right\}^2\times\Omega$ defined with $\overline{\widetilde{L}}$ as the likelihood function on $\widetilde{\Phi}$. Then
$$\widetilde{W}\left(\widetilde{\Theta}',\widetilde{\Theta}'';x\right)=\overline{W}\left(\left\{\widetilde{\varphi}\left(\widetilde{\Theta}'\right)\right\},\left\{\widetilde{\varphi}\left(\widetilde{\Theta}''\right)\right\};x\right).\tag{3.4}$$

The general law of likelihood given by Proposition 1 is now extended to govern imprecise hypotheses:

Proposition 8. Extended law of likelihood. *For any explanatory function* ex*, let W denote the evidential function $2^{\Theta\times(0,1]}\times 2^{\Theta\times(0,1]}\times\Omega$ with respect to* ex*. Further, let $\widetilde{W}$ denote the extended evidential function with respect to* ex*. Then the weight of evidence in $X=x$ that supports $\theta\tilde{\in}\widetilde{\Theta}'$ over $\theta\tilde{\in}\widetilde{\Theta}''$ is*
$$\widetilde{W}\left(\widetilde{\Theta}',\widetilde{\Theta}'';x\right)=\frac{\sup_{\theta'\in\Theta}\widetilde{\Theta}'(\theta')\,\widetilde{L}(\theta';x)}{\sup_{\theta''\in\Theta}\widetilde{\Theta}''(\theta'')\,\widetilde{L}(\theta'';x)}\tag{3.5}$$
for all fuzzy subsets $\widetilde{\Theta}',\widetilde{\Theta}''\in\mathcal{F}(\Theta)$ such that $\theta\tilde{\in}\widetilde{\Theta}'$ and $\theta\tilde{\in}\widetilde{\Theta}''$ is each either a simple hypothesis or a complex hypothesis, but is
$$\widetilde{W}\left(\widetilde{\Theta}',\widetilde{\Theta}'';x\right)=\frac{\int\widetilde{\Theta}'(\theta')\,d\mathbf{P}(\theta'|X=x)\,/\int\widetilde{\Theta}'(\theta')\,dp(\theta')}{\int\widetilde{\Theta}''(\theta'')\,d\mathbf{P}(\theta''|X=x)\,/\int\widetilde{\Theta}''(\theta'')\,dp(\theta'')}\tag{3.6}$$
for all $\Theta'\subseteq\Theta$ and $\Theta''\subseteq\Theta$ such that $\theta\tilde{\in}\widetilde{\Theta}'$ and $\theta\tilde{\in}\widetilde{\Theta}''$ are intrinsically simple.

Proof. The case that $\theta\tilde{\in}\widetilde{\Theta}'$ and $\theta\tilde{\in}\widetilde{\Theta}''$ are intrinsically simple hypotheses is addressed first. (3.4), (2.5) , and (3.2) yield
$$\widetilde{W}\left(\widetilde{\Theta}',\widetilde{\Theta}'';x\right)=\frac{\mathbf{E}\left(\widetilde{\Theta}'(\theta)\,|X=x\right)/\mathbf{E}\left(\widetilde{\Theta}'(\theta)\right)}{\mathbf{E}\left(\widetilde{\Theta}''(\theta)\,|X=x\right)/\mathbf{E}\left(\widetilde{\Theta}''(\theta)\right)},$$
from which (3.6) immediately follows. Next consider the case that $\theta\tilde{\in}\widetilde{\Theta}'$ and $\theta\tilde{\in}\widetilde{\Theta}''$ is each either a simple hypothesis or a complex hypothesis. The hypotheses $\theta\tilde{\in}\widetilde{\Theta}'$

and $\theta \tilde{\in} \widetilde{\Theta}''$ are thus shorthand for $\left\langle \theta, \widetilde{\Theta}'\left(\theta\right) \right\rangle \in \Xi\left(\widetilde{\Theta}'\right)$ and $\left\langle \theta, \widetilde{\Theta}''\left(\theta\right) \right\rangle \in \Xi\left(\widetilde{\Theta}''\right)$, respectively (Section 3.2). By Definition 8 and (2.9),

$$\widetilde{W}\left(\widetilde{\Theta}', \widetilde{\Theta}''; x\right) = \frac{\sup_{\langle \theta, C\rangle \in \Xi\left(\widetilde{\Theta}'\right)} L\left(\langle \theta, C\rangle; x\right)}{\sup_{\langle \theta, C\rangle \in \Xi\left(\widetilde{\Theta}''\right)} L\left(\langle \theta, C\rangle; x\right)}.$$

Since $\langle \theta, C\rangle \in \Xi\left(\widetilde{\Theta}'\right)$ if and only if $\left\langle \theta, \widetilde{\Theta}'\left(\theta\right) \right\rangle \in \Xi\left(\widetilde{\Theta}'\right)$, we have

$$\widetilde{W}\left(\widetilde{\Theta}', \widetilde{\Theta}''; x\right) = \frac{\sup_{\theta \in \Theta} L\left(\left\langle \theta, \widetilde{\Theta}'\left(\theta\right) \right\rangle; x\right)}{\sup_{\theta \in \Theta} L\left(\left\langle \theta, \widetilde{\Theta}''\left(\theta\right) \right\rangle; x\right)}$$

in terms of the likelihood function $L\left(\langle \bullet \rangle; x\right)$ on $\Theta \times (0, 1]$. By the equivalence of $L\left(\left\langle \theta, \widetilde{\Theta}'\left(\theta\right) \right\rangle, x\right)$ and $\widetilde{L}\left(\theta, \widetilde{\Theta}'\left(\theta\right); x\right)$,

$$\widetilde{W}\left(\widetilde{\Theta}', \widetilde{\Theta}''; x\right) = \frac{\sup_{\theta' \in \Theta} \widetilde{L}\left(\theta', \widetilde{\Theta}'\left(\theta'\right); x\right)}{\sup_{\theta'' \in \Theta} \widetilde{L}\left(\theta'', \widetilde{\Theta}''\left(\theta''\right); x\right)}$$

in terms of the (possibly reduced) incomplete likelihood function $\widetilde{L}\left(\bullet; x\right)$ on $\Theta \times (0, 1]$. Using the identity $\widetilde{L}\left(\theta, c; x\right) = c\widetilde{L}\left(\theta; x\right)$ of Section 3.1 for substitution completes the proof of (3.5).

In the presence of a nuisance parameter, the reduced likelihood function $\widetilde{L}\left(\bullet, C\right)$ is formed by eliminating the nuisance parameter in order to approximate the weight of evidence, analogous to the precise hypothesis case of Section 2.5. Then each $\widetilde{L}\left(\theta, C\right)$ is a function of the distributions indexed by the same interest parameter value and with the same level of completeness but not a function of other members of the family of incomplete probability distributions. The method of nuisance parameter elimination must also preserve $\widetilde{L}\left(\theta, C; x\right) = C\widetilde{L}\left(\theta, 1; x\right)$ for all $\theta \in \Theta$ and $C \in (0, 1]$. The application of (3.5) in the presence of a nuisance parameter is illustrated in Section 5.1.

4. Simulation Study

To quantify the impact of replacing a simple hypothesis with a small-interval composite hypothesis in evidential inference, a series of simulations was carried out for the case of normal distributions (Example 7). $M = 10^5$ independent samples of independent standard normal observations were randomly generated for each of 23 sample sizes from $n = 2$ to $n = 10,000$. Given samples $x_1, \ldots, x_M$, each of size n, and a threshold of b bans of evidence for $\theta \neq 0$ over $\theta = 0$, the probability of observing misleading evidence was computed by

$$\widehat{\alpha}_n^{\Theta''}(b) = \frac{1}{M} \sum_{i=1}^{M} 1_{\left[10^b,\infty\right)} \left(W_{\text{profile}} \left(\mathbb{R}^1 \Big\backslash \Theta'', \Theta''; x_i \right) \right) \tag{4.1}$$

with $\Theta'' = \{0\}$ for the composite-*simple* hypothesis pair or with $\Theta'' = [-1/10, 1/10]$ for the composite-*composite* hypotheses pair. The levels of evidence were chosen to correspond to the probabilities of observing at least weak evidence ($b = 1/\infty$), at least moderate evidence ($b = 1/2$), at least strong evidence ($b = 1$), at least very strong evidence ($b = 3/2$), and decisive evidence ($b = 2$). Every observation of evidence favoring $\theta \neq 0$ or $|\theta| > 1/10$ at any level is misleading since the data were generated under $\theta = 0$.

The results are displayed as Figures 1−5, with one figure per level of evidence. Figure 1 highlights the most obvious discrepancy between the two choices of hypothesis pairs. Since the maximum likelihood estimate almost never equals 0, the evidence favors $\theta \neq 0$ over $\theta = 0$ with probability 1. By contrast, the evidence usually favors $|\theta| \leq 1/10$ over $|\theta| > 1/10$, except for small samples. At the higher evidence grades, Figures 2−5 also show that the probability of observing evidence for the incorrect hypothesis decreases as the sample size increases for $\Theta'' = [-1/10, 1/10]$, as expected from Proposition 4, but not for $\Theta'' = \{0\}$, with the exception of smaller samples.

Figure 6 displays probabilities of misleading evidence (4.1) for sample sizes common in experimental biology. Its plots for $n = 5$ and $n = 6$ are directly relevant to the application of the next section.

5. Application to Gene Expression Data

5.1. Evidence of differential expression

In this section, the theory of Sections 2 and 3 is illustrated with some of the tomato gene expression data described in Alba et al. (2005). Dual-channel microarrays were used to measure the mutant-to-wild-type expression ratios of 13,440 genes at the breaker stage of ripening and at 3 and 10 days thereafter. Each of the later two stages has six biological replicates ($n = 6$), but one of the biological replicates is missing at the breaker stage of ripening ($n = 5$).

For each of the three time points, there are two competing hypotheses per gene: the geometric mean of the expression ratio between mutant tomatoes and wildtype tomatoes is either 1 (the simple hypothesis corresponding to no mutation effect) or is not 1 (the composite hypothesis corresponding to a mutation effect). Since the data are approximately lognormal, the relevant family of distributions for each gene i is that of (2.15), replacing θ with θ_i, the logarithm of

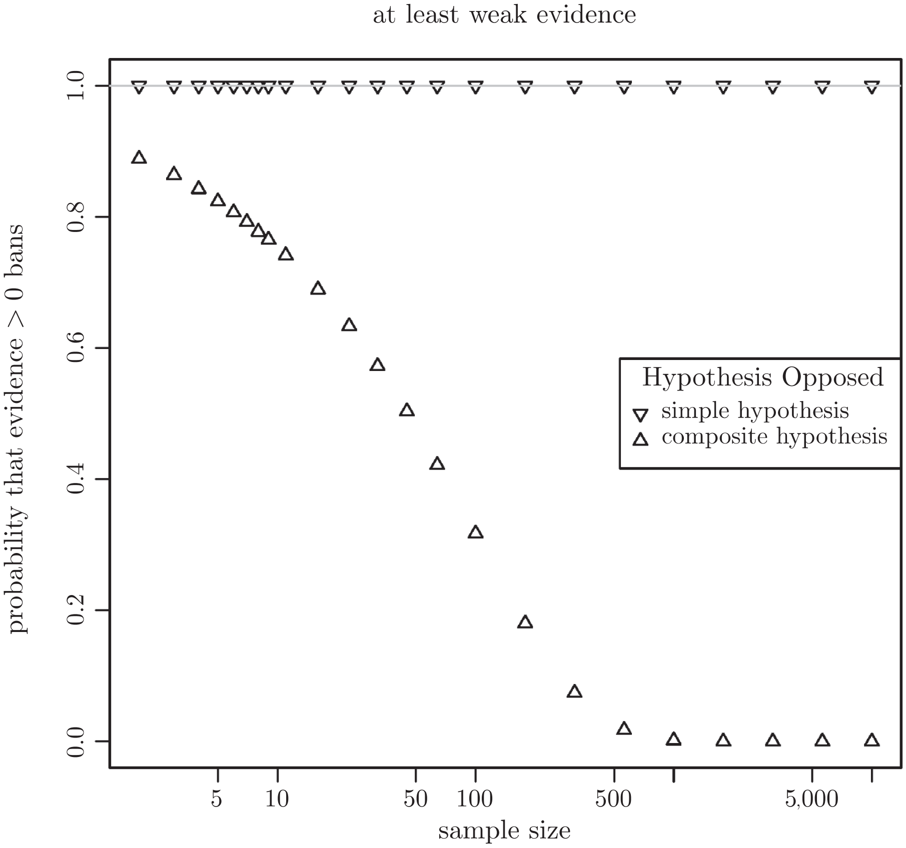

Figure 1. Probabilities $\widehat{\alpha}_n^{\{0\}}(1/\infty)$ and $\widehat{\alpha}_n^{[-1/10,1/10]}(1/\infty)$ of observing any misleading **positive** evidence for the hypothesis that $\theta \neq 0$ over the "simple" hypothesis that $\theta = 0$ and for the hypothesis that $|\theta| > 1/10$ over the "composite" hypothesis that $|\theta| \leq 1/10$, respectively.

geometric mean of the expression ratio of the ith gene, and replacing x with x_i, each component of which is the logarithm of an observed expression ratio of the ith gene. The maximum likelihood estimate of θ_i is $\widehat{\theta}_i$, the sample mean of the logarithms of the expression ratios for the ith gene. The commonly made independence assumption of Section 2.4, although known to be incorrect, remains a useful approximation in the absence of sufficiently large n to reliably estimate gene-gene interactions.

Like in the simulation study of the last section, (2.14) gives the strength of evidence for differential expression between the wild type and the mutant ($\theta_i \neq 0$) over equivalent expression ($\theta_i = 0$). Since, however, the expression ratio is not exactly 1, Bickel (2004), Lewin et al. (2006), Van De Wiel and Kim (2007), Bochkina and Richardson (2007), and McCarthy and Smyth (2009) redefined what is

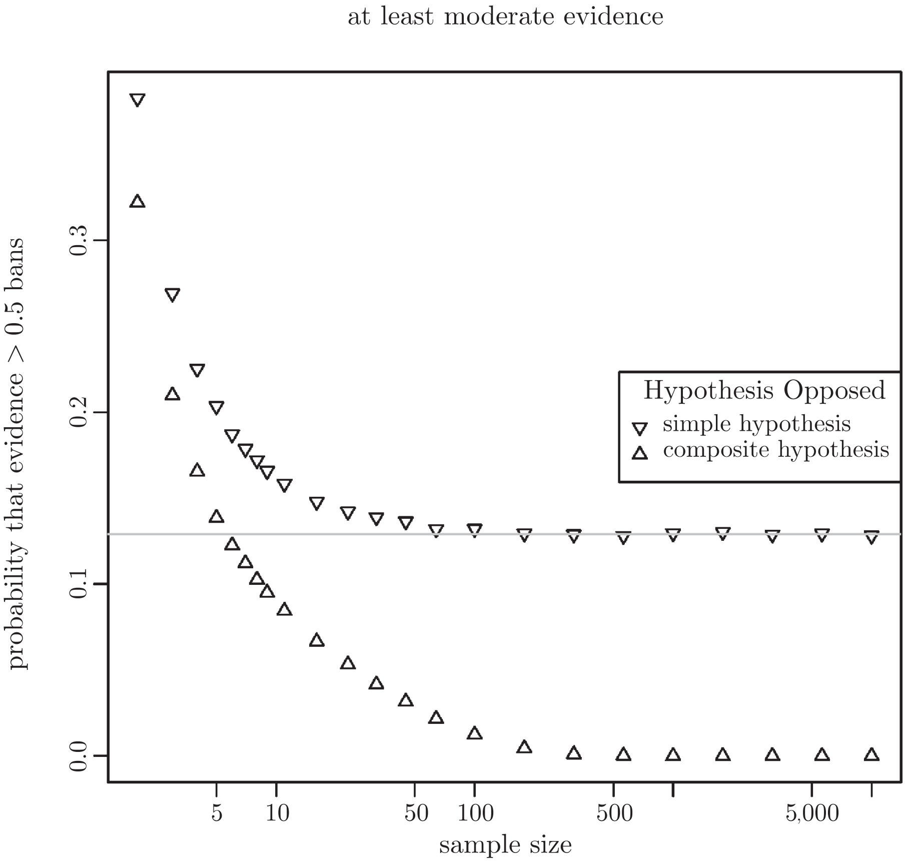

Figure 2. Probabilities $\widehat{\alpha}_n^{\{0\}}(1/2)$ and $\widehat{\alpha}_n^{[-1/10,1/10]}(1/2)$ of observing misleading **moderate or stronger** evidence for the hypothesis that $\theta \neq 0$ over the "simple" hypothesis that $\theta = 0$ and for the hypothesis that $|\theta| > 1/10$ over the "composite" hypothesis that $|\theta| \leq 1/10$, respectively. The horizontal gray line is drawn at $\lim_{n\to\infty,M\to\infty}\widehat{\alpha}_n^{\{0\}}(1/2)$ according to the χ^2 distribution with 1 degree of freedom; $\lim_{n\to\infty,M\to\infty}\widehat{\alpha}_n^{[-1/10,1/10]}(1/2) = 0$ by Proposition 4.

meant by "differential expression" by employing some biologically relevant value $\theta_+ > 0$. Accordingly, (2.14) also yields the strength of evidence for biologically significant differential expression between the wild type and the mutant ($|\theta_i| > \theta_+$) over biologically insignificant differential expression ($|\theta_i| \leq \theta_+$). Due to the importance of the twofold change in biochemistry, θ_+ is here set to $\frac{1}{2}\log 2$, the midpoint between 0 and $\log 2$. (Similarly, Lewin et al. (2006) and Bochkina and Richardson (2007) derived posterior probabilities that $|\theta_i| > \log 2$, and Bickel (2004), Van De Wiel and Kim (2007), and McCarthy and Smyth (2009) considered false discovery rates for which a "discovery" is defined in terms of fold

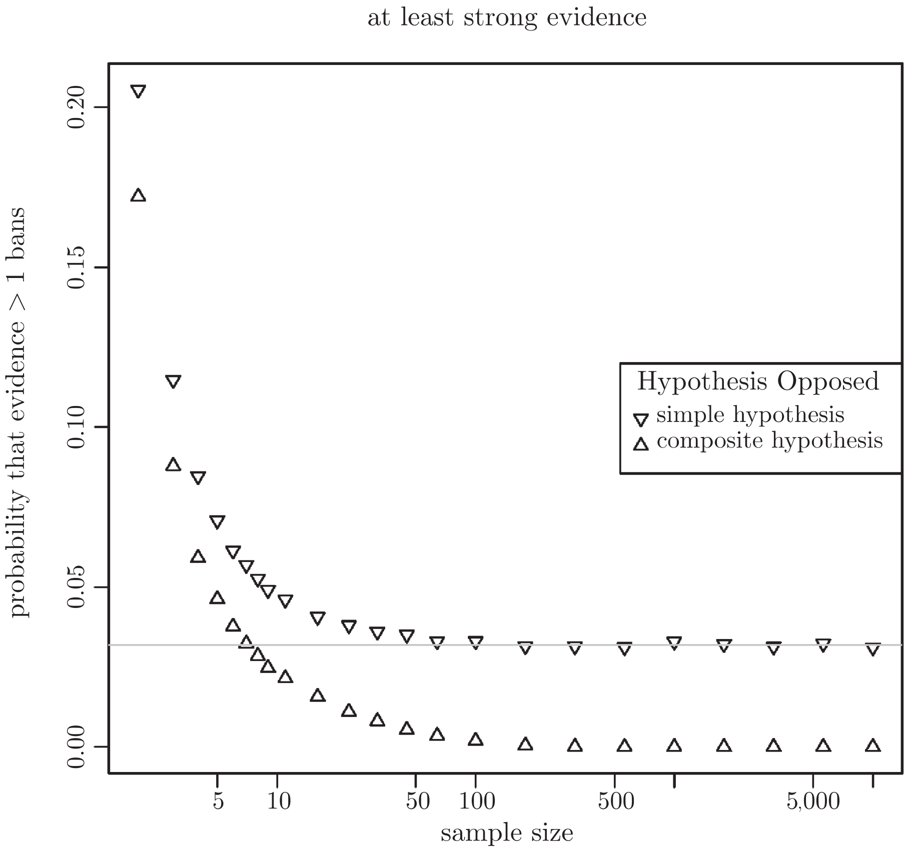

Figure 3. Probabilities $\widehat{\alpha}_n^{\{0\}}(1)$ and $\widehat{\alpha}_n^{[-1/10,1/10]}(1)$ of observing misleading **strong, very strong, or decisive** evidence for the hypothesis that $\theta \neq 0$ over the "simple" hypothesis that $\theta = 0$ and for the hypothesis that $|\theta| > 1/10$ over the "composite" hypothesis that $|\theta| \leq 1/10$, respectively.

change thresholds.)

As seen in Figure 7, the use of $|\theta_i| > \log\sqrt{2}$ rather than $|\theta_i| > 0$ as the hypothesis corresponding to differential expression leads to considering many fewer genes differentially expressed at each stage of maturity and at each level of evidence. Now the composite hypotheses for gene i are $\theta_i \in \Theta' = \mathbb{R}^1 \backslash [-\log\sqrt{2}, \log\sqrt{2}]$ and $\theta_i \in \Theta'' = \left[-\log\sqrt{2}, \log\sqrt{2}\right]$. There is an order of magnitude more genes counted as differentially expressed at each evidence grade when using $W_{\text{profile}}\left(\mathbb{R}^1 \backslash \{0\}, \{0\}; x_i\right)$ than when using $W_{\text{profile}}(\Theta', \Theta''; x_i)$ as the strength of evidence in x_i, the data for the ith gene.

The left-hand-side of Figure 8 stresses the main limitation of comparing two composite hypotheses: the results depend on the specification of θ_+, the value that determines the sharp boundary between equivalent expression $\left(|\theta_i| \leq \theta_+\right)$

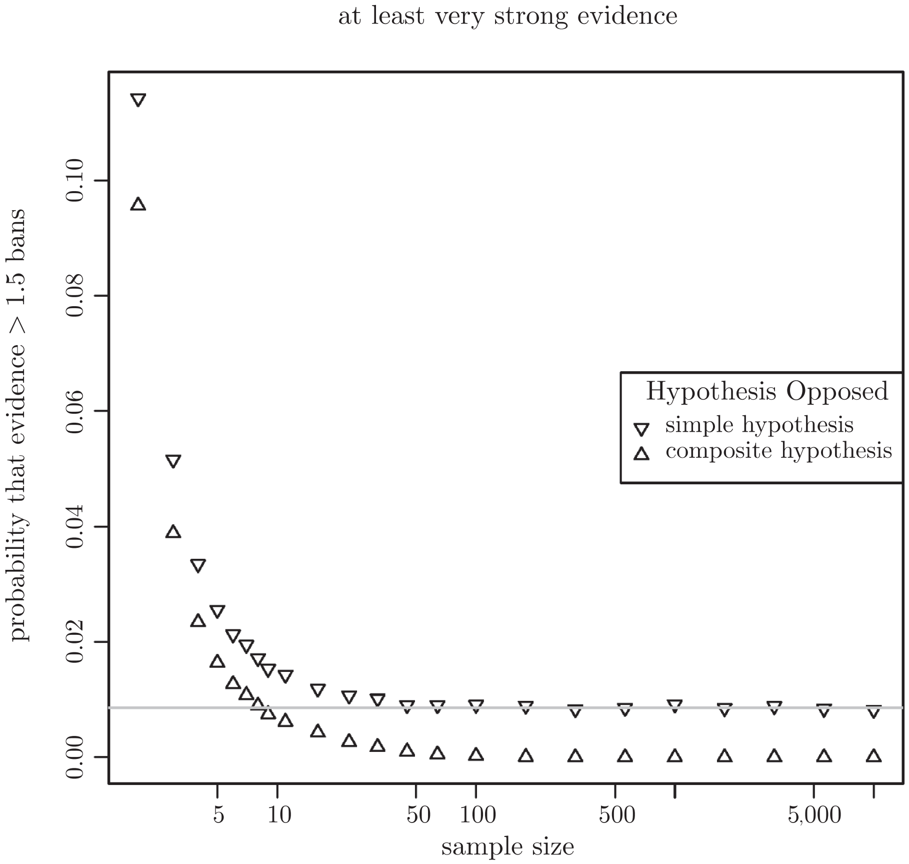

Figure 4. Probabilities $\widehat{\alpha}_n^{\{0\}}(3/2)$ and $\widehat{\alpha}_n^{[-1/10,1/10]}(3/2)$ of observing misleading **very strong or decisive** evidence for the hypothesis that $\theta \neq 0$ over the "simple" hypothesis that $\theta = 0$ and for the hypothesis that $|\theta| > 1/10$ over the "composite" hypothesis that $|\theta| \leq 1/10$, respectively.

and differential expression $(|\theta_i| > \theta_+)$; in this case, $\theta_+ = \log \sqrt{2}$. By instead allowing degrees of whether a gene is differentially expressed, the approach of Section 3 mitigates this effect. For correspondence with the above analyses with precise hypotheses, a gene is considered differentially expressed to extent

$$\widetilde{\Theta}'(\theta) = \begin{cases} \frac{|\theta|}{\log 2} & |\theta| \leq \log 2 \\ 1 & |\theta| > \log 2 \end{cases}$$

and equivalently expressed to extent $\widetilde{\Theta}''(\theta) = 1 - \widetilde{\Theta}'(\theta)$, as illustrated in Figure 9. Sokhansanj et al. (2004) instead considered a fuzzy subset on gene expression measurements that would only achieve full expression membership for infinite measurements. By contrast, $\widetilde{\Theta}'$ considers all genes with two-fold or greater dif-

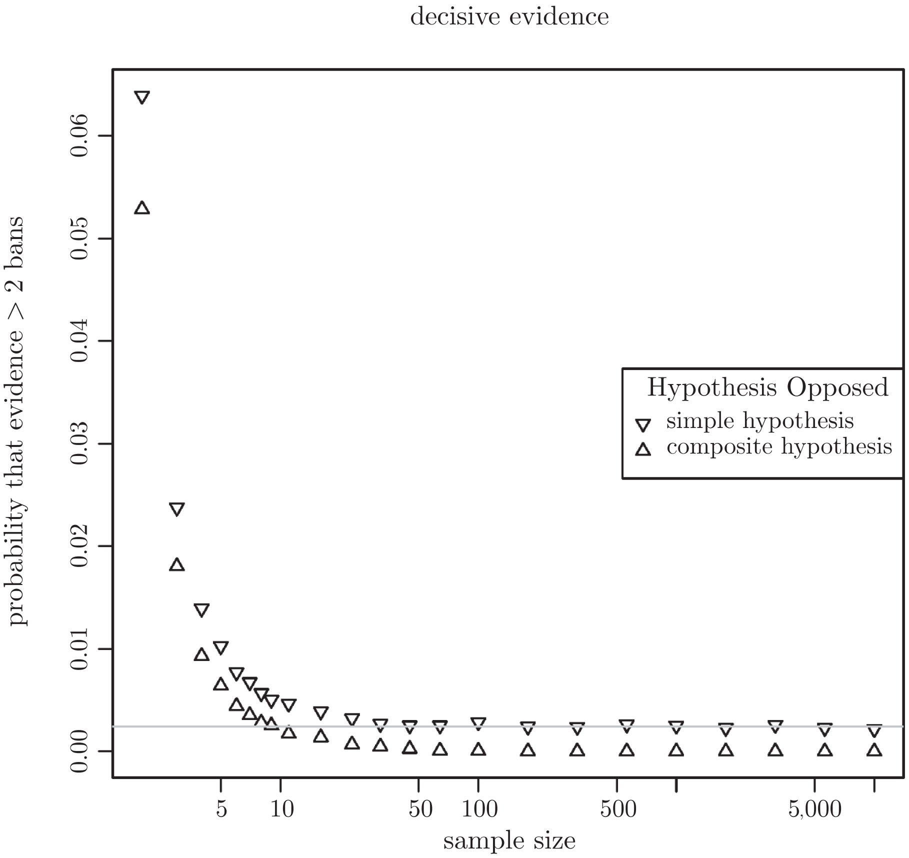

Figure 5. Probabilities $\widehat{\alpha}_n^{\{0\}}(2)$ and $\widehat{\alpha}_n^{[-1/10,1/10]}(2)$ of observing misleading **decisive** evidence for the hypothesis that $\theta \neq 0$ over the "simple" hypothesis that $\theta = 0$ and for the hypothesis that $|\theta| > 1/10$ over the "composite" hypothesis that $|\theta| \leq 1/10$, respectively.

ferential expression between populations to be fully differentially expressed.

The success in eliminating the undesirable discontinuity at the rigid boundary between hypotheses is evident from the right-hand-side of Figure 8, which displays $\widetilde{W}_{\text{profile}}\left(\widetilde{\Theta}', \widetilde{\Theta}''; x_i\right)$, the result of putting the profile likelihood function in place of likelihood function in (3.5), against $\exp\left(\widehat{\theta}_i\right)$, the maximum likelihood estimate of the expression ratio. Although the strength of evidence still changes sign at $\widehat{\theta}_i = \pm \log \sqrt{2}$, no trace remains of what resembles a phase transition at those points in the precise hypothesis case.

The replacement of $\widetilde{W}_{\text{profile}}(\Theta', \Theta''; x_i)$ with $\widetilde{W}_{\text{profile}}\left(\widetilde{\Theta}', \widetilde{\Theta}''; x_i\right)$ has high impact on inference for a large portion of the genes (Figure 10). Levels of evidence between 0 and 2 are most important for finding genes with evidence of differential

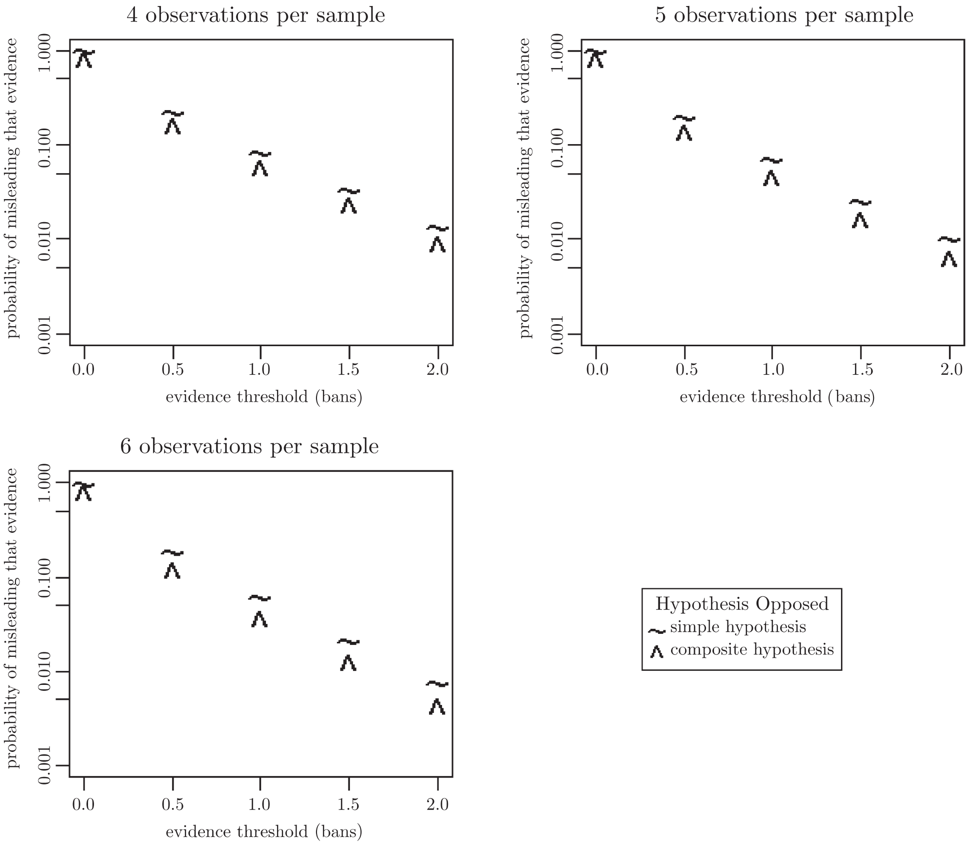

Figure 6. Probabilities $\widehat{\alpha}_n^{\{0\}}(b)$ and $\widehat{\alpha}_n^{[-1/10,1/10]}(b)$ of observing misleading evidence for the hypothesis that $\theta \neq 0$ over the "simple" hypothesis that $\theta = 0$ and for the hypothesis that $|\theta| > 1/10$ over the "composite" hypothesis that $|\theta| \leq 1/10$, respectively, for each of the evidence levels b of previous figures and at each of three sample sizes ($n \in \{4,5,6\}$).

expression since negative levels correspond to evidence for equivalent expression, and levels above 2 normally indicate decisive evidence for differential expression regardless of whether precise or imprecise hypotheses are specified.

5.2. Empirical Bayes inference

The "theoretical null" version of the empirical Bayes model of Efron (2007), when applied to data structured as in Section 5.1, assumes the Student t statistic $T(X_i)$ has probability density $f(\bullet;1)$ if gene i is differentially expressed, which occurs with probability $p(1)$, and $f(\bullet;0)$ if gene i is equivalently expressed, which occurs with probability $p(0)$, where $f(\bullet;0)$ is the Student t density with $n-1$ degrees of freedom. Thus, $\theta_i \in \{0,1\}$ has physical probability distribution p for each i (Section 2.4.3).

On the basis of the 10-day microarrays for the 7139 genes with complete data ($n = 6$), the probability mass function p was estimated by $\hat{p}$ and the probability density function $f(\bullet; 1)$ by $\hat{f}(\bullet; 1)$, with both estimators defined by the method of Efron (2007). Then

$$\hat{w}(1, 0; T(x_i)) = \frac{\hat{f}(T(x_i); 1)}{f(T(x_i); 0)}$$

is an approximate Bayes factor according to its role in approximating posterior probabilities by estimated local false discovery rates. Herein, $\hat{w}(1, 0; T(x_i))$ is instead employed as an estimate of the weight of evidence for $\theta_i = 1$ over $\theta_i = 0$ as defined by the special law of likelihood.

Figure 11 compares $\log_{10} W_{\text{profile}}\left(\mathbb{R}^1 \setminus \{0\}, \{0\}; x_i\right)$ and $\log_{10} W_{\text{profile}}(\Theta', \Theta''; x_i)$ of the fixed-parameter model (Section 5.1) to $\log_{10} \hat{w}(1, 0; T(x_i))$ of the random-parameter model defined in the present subsection. The discrepancies stem from the differences in model assumptions.

6. Closing Remarks

6.1. Highlights and discussion

Sections 2.1 and 2.2 axiomatically defined the evidential function W in order to uniquely weigh the evidence in observation x for a hypothesis $\theta \in \Theta'$ over another hypothesis $\theta \in \Theta''$. W applies not only to simple hypotheses, but also to composite hypotheses, including complex hypotheses about fixed parameter values and intrinsically simple hypotheses about random parameter values. Properly distinguishing between the nuisance parameter problem and the composite hypothesis problem in Section 2.5 avoids pathologies of the profile likelihood without resorting to the representation of evidence by intervals of profile likelihoods. The proposed framework compares favorably with Bayesianism in Example 4 because the former, but not the latter, satisfies the idealized principle of inference to the best explanation (Section 2.2.1). The evidential weight $W(\Theta', \Theta''; x)$ is consistent, coherent, and interpretable, as seen in Sections 2.3 and 4. These properties warrant consideration of a new approach to simultaneous inference, multiple comparisons, and sequential analysis (Section 2.4).

Incomplete probability distributions represent imprecision in hypotheses to mitigate the effect of hypothesis boundaries on the weight of evidence, as illustrated in the gene expression application (Sections 3 and 5). Nonetheless, making hypotheses imprecise sometimes insufficiently reduces the dependence of the weight of evidence on arbitrarily selected parameter values. In such settings, the use of two composite hypotheses separated by a non-arbitrary boundary entirely

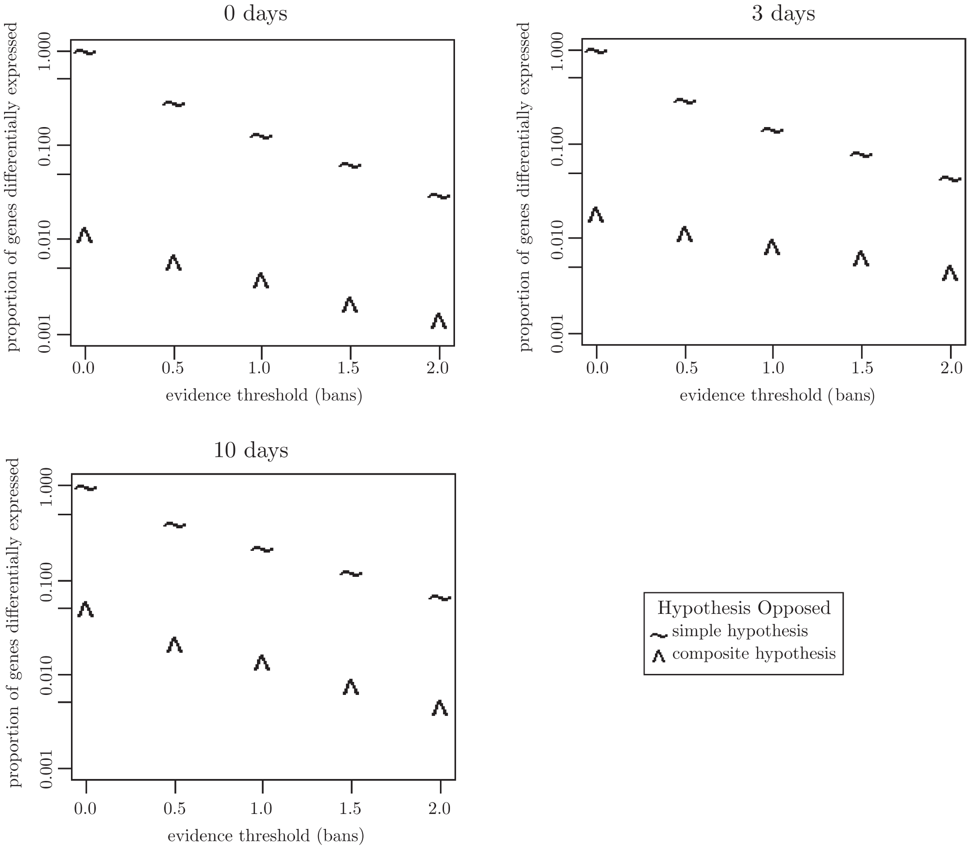

Figure 7. Proportions $\widehat{\alpha}_n^{\{0\}}(b)$ and $\widehat{\alpha}_n^{[-\log\sqrt{2},\log\sqrt{2}]}(b)$ of genes with differential expression evidence meeting or exceeding a fixed evidence threshold for the hypothesis that $\theta_i \neq 0$ over the "simple" hypothesis that $\theta_i = 0$ and for the hypothesis that $|\theta_i| > \log\sqrt{2}$ over the "composite" hypothesis that $|\theta_i| \leq \log\sqrt{2}$, respectively. Results are displayed for each of the evidence thresholds of the previous figures ($b \in \{1/\infty, 1/2, 1, 3/2, 2\}$) and at each of three stages of maturity (0, 3, and 10 days after the breaker stage of ripening). These proportions were computed using (4.1), but with x_i as the vector of the logarithms of the expression ratios for the ith gene and with M as the number of genes that have sufficient data for the computation of likelihood ratios.

eliminates such dependence. In the gene expression illustration of Section 5, the weight of evidence for biologically significant differential expression ($|\theta_i| > \theta_+$) versus biologically insignificant differential expression ($|\theta_i| \leq \theta_+$) would then be replaced by the weight of evidence for overexpression/upregulation ($\theta_i > 0$) versus underexpression/downregulation ($\theta_i < 0$), either superseding or complementing an application of decision theory to the latter two hypotheses (Bickel (2011a)).

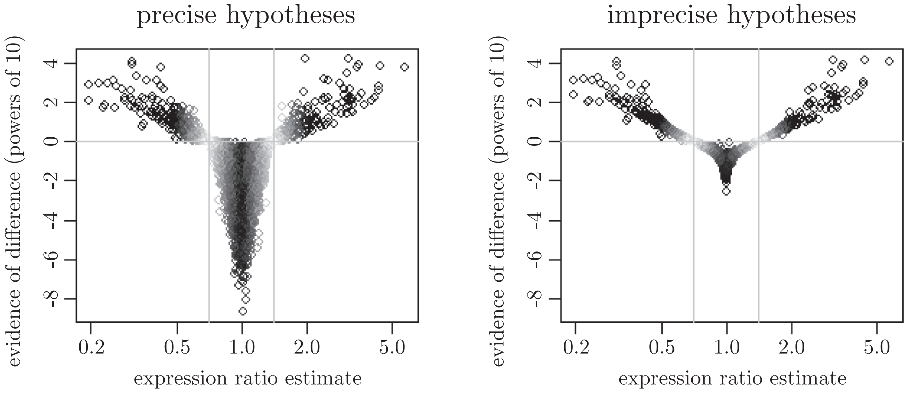

Figure 8. The weight of evidence for differential expression over equivalent expression plotted against the maximum likelihood estimate of the expression ratio for the tomato data at 10 days after the breaker stage of ripening. The vertical gray lines are drawn at the boundary that separates the two precise hypotheses, reflecting the idea that a gene is either differentially expressed or is equivalently expressed, with no possibility of something in between. By contrast, the imprecise hypotheses have no rigid boundary between differential expression and equivalent expression. Darker circles represent genes that correspond to higher values of $|2\widetilde{\Theta}'(\widehat{\theta}_i) - 1|$ and that thus seem to be more closely aligned with either one imprecise hypothesis or the other, whereas lighter circles correspond to more borderline genes. $\widetilde{\Theta}'(\widehat{\theta}_i)$ estimates $\widetilde{\Theta}'(\theta_i)$, the degree to which the ith gene is differentially expressed.

6.2. Opportunities for further research

6.2.1. Additional models and applications

The laws of likelihood offer an evidential framework that invites examination of their practical effects on statistical inference. The examination of normal variates of Section 4 concentrated on the probability of observing misleading evidence for a composite hypothesis over an interval hypothesis, finding that it is often much less than that for a composite hypothesis over a simple hypothesis. The microarray case study of Section 5 quantified the impact on evidential inference of replacing simple hypotheses with interval hypotheses and of replacing precise hypotheses with imprecise hypotheses.

The proposed framework may be further examined for other families of distributions and for other applications. In particular, the findings of Sections 2.3.3 and 3 suggest a fresh approach to bioequivalence studies in which researchers seek to determine whether the evidence favors an interval hypothesis over a composite hypothesis without requiring an artificially precise specification of the largest effect size considered equivalent.

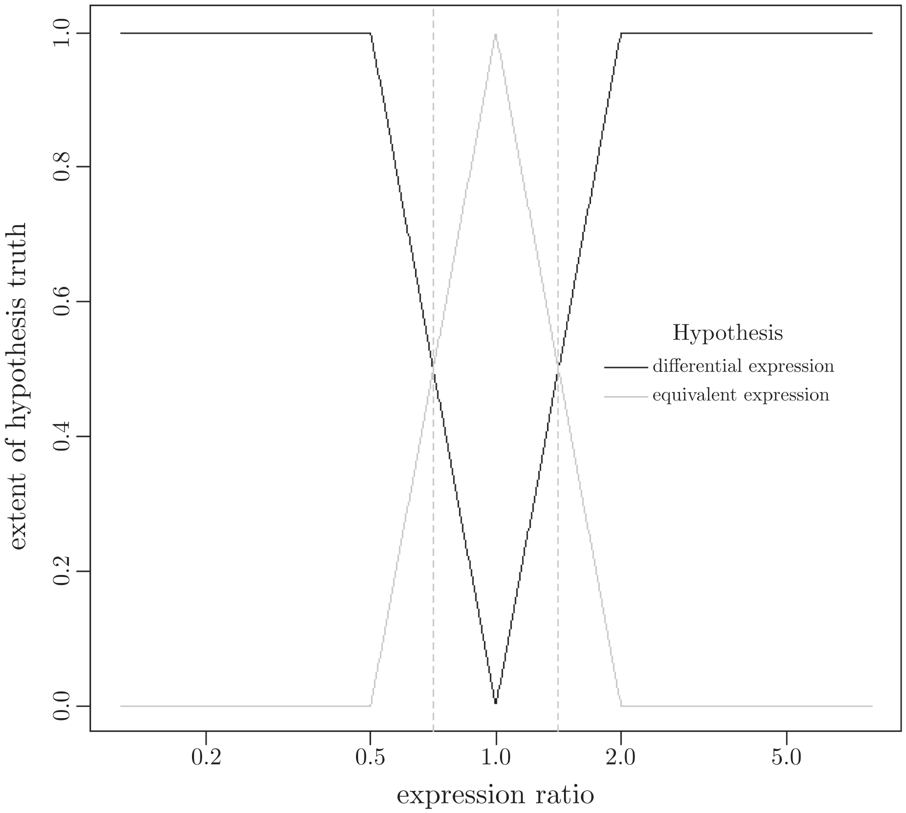

Figure 9. The degree of the truth of each imprecise hypothesis plotted against e^{θ}, the geometric mean of the expression ratio in the population. The black curve represents $\widetilde{\Theta}'$, and the gray curve represents $\widetilde{\Theta}''$. The vertical lines correspond to the boundary between the precise hypotheses Θ' and Θ''. Degrees of truth are calibrated by Definition 8.

6.2.2. Robust evidential inference

There remains ample opportunity for research to make evidential inference about composite hypotheses robust to unanticipated data distributions. Possible solutions may utilize robust adjusted likelihood functions, contamination mixture models, or nonparametric approaches, all of which require at least moderately large samples. Each strategy is discussed in turn.

A likelihood function adjustment designed to make the law of likelihood less sensitive to model misspecification (Royall and Tsou (2003); Blume et al. (2007)) might be used for robust inference under the general law of likelihood. The resulting *robust adjusted likelihood function* performs well under certain violations of the working model and yet retains full asymptotic efficiency if the working model is correct (Royall and Tsou (2003)). Since the adjustment improves both

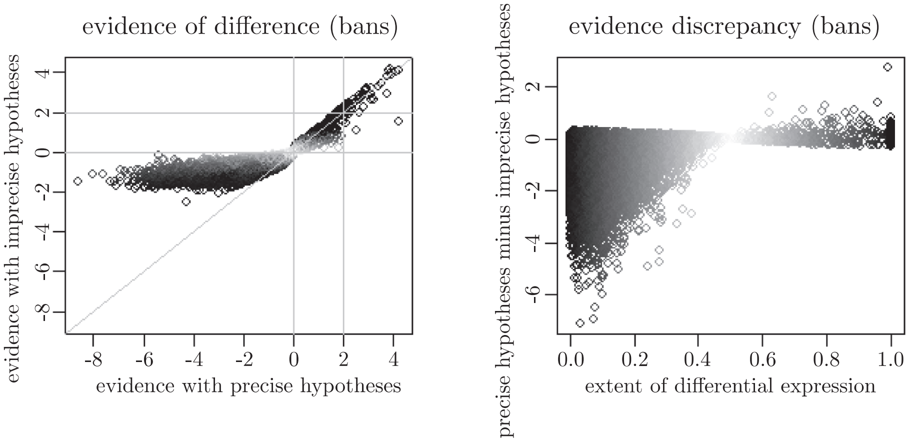

Figure 10. Effects of replacing the precise hypotheses with the imprecise hypotheses for the data of Figure 8. The left-hand-side displays $W_{\text{profile}}(\widetilde{\Theta}', \widetilde{\Theta}''; x_i)$ plotted against $W_{\text{profile}}(\Theta', \Theta''; x_i)$, and the right-hand-side has $W_{\text{profile}}(\Theta', \Theta''; x_i) -$ $W_{\text{profile}}(\widetilde{\Theta}', \widetilde{\Theta}''; x_i)$ against $\widetilde{\Theta}'(\widehat{\theta}_i)$, the estimated extent of differential expression. The grayscale is the same as that of Figure 8.

Neyman-Pearson and Bayesian uses of the likelihood function (Royall and Tsou (2003)), the adjustment is expected to improve evidential inference regarding composite hypotheses as well.

A more classical approach to making the likelihood function robust against potential outliers replaces the working model $\{f(\bullet; \theta) : \theta \in \Theta\}$ with a mixture model $\{(1-\varepsilon) f(\bullet; \theta) + \varepsilon g(\bullet; \gamma) : \theta \in \Theta\}$, where ε is the unknown probability of contamination and g is the contamination density or mass function parameterized by γ (Aitkin and Wilson (1980)). It may be advisable to extend this methodology to evidential inference about simple hypotheses before attempting to generalize it to handle precise and imprecise composite hypotheses.

The empirical likelihood version of (2.9) is

$$W(S', S'') = \frac{\sup_{F' \in S'} \mathcal{L}(F'; x)}{\sup_{F'' \in S''} \mathcal{L}(F''; x)},$$

where $\mathcal{L}(\bullet; x)$ is the nonparametric likelihood function (Owen (2001)) and where S' and S'' are broad sets of distributions corresponding to different hypotheses distinguished by their constraints, e.g., S' and S'' may be large families of distribution with means outside or inside some interval, respectively. Zhang (2009a) studied the simple hypothesis case $W(\{F'\}, \{F''\}) = \mathcal{L}(F'; x) / \mathcal{L}(F''; x)$. (3.5) may be analogously modified by replacing the parametric likelihood function with the nonparametric likelihood function and constraint satisfaction with partial constraint satisfaction indicated by the membership functions of fuzzy sets.

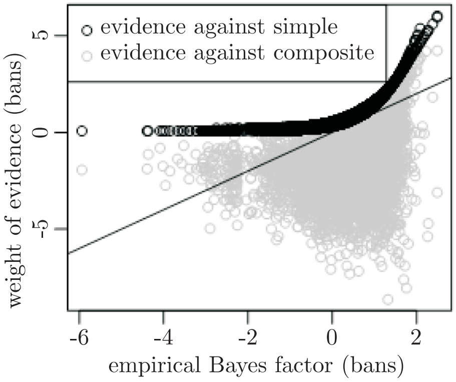

Figure 11. The weight of evidence for differential gene expression under the fixed-parameter model versus the Bayes factor under the empirical Bayes model for all genes with 6 ratios available. Each gray circle and each black circle represents a different gene. The diagonal is the line of equality.

Acknowledgements

I warmly thank an Associate Editor and three anonymous reviewers for helpful comments and especially for several examples, a criticism of my initial set of axioms, and guidance on plotting two of the figures. I am also grateful to Jeffrey Blume for an enlightening discussion on evidence, multiple comparisons, and Bayesianism. I thank Corey Yanofsky for proofreading the manuscript. Xuemei Tang provided the fruit development microarray data.

This work was partially supported by Agriculture and Agri-Food Canada and by the Faculty of Medicine of the University of Ottawa. Computations were performed in R (R Development Core Team (2008)) with the help of the Biobase package of Bioconductor (Gentleman et al. (2004)) and the locfdr package of CRAN (Efron (2007)).

Notes added in proof

Bickel (2011b) proposes another likelihood-based measure of the weight of evidence. Complementing the likelihood-evidential measures, systems of nested confidence intervals undergird information-theoretic frameworks for applying Bayesian methods (Bickel (2012a)) and frequentist methods, including multiple comparison procedures (Bickel (2012b)).

References

Aitkin, M. (1991). Posterior Bayes factors (with discussion). *J. Roy. Statist. Soc. Ser. B* **53**, 111-142.

Aitkin, M. and Wilson, G. T. (1980). Mixture models, outliers, and the EM algorithm. *Technometrics* **22**, 325-331.

Alba, R., Payton, P., Fei, Z., McQuinn, R., Debbie, P., Martin, G. B., Tanksley, S. D. and Giovannoni, J. J. (2005). Transcriptome and selected metabolite analyses reveal multiple points of ethylene control during tomato fruit development. *Plant Cell* **17**, 2954-2965.

Barnard, G. A. (1967). The use of the likelihood function in statistical practice. *Proc. 5th Berkeley Symp. on Math. Stat. Prob.* Vol. I, 27-40.

Benjamini, Y., Drai, D., Elmer, G., Kafkaf, N. and Golani, I. (2001). Controlling the false discovery rate in behavior genetics research. *Behavioural Brain Research* **125**, 279-284.

Benjamini, Y. and Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Educational and Behavioral Statistics* **25**, 60-83.

Benjamini, Y. and Liu, W. (1999). A stepdown multiple hypotheses testing procedure that controls the false discovery rate under independence. *J. Statist. Plann. Inference* **82**, 163-170.

Benjamini, Y. and Yekutieli, D. (2005). Quantitative trait loci analysis using the false discovery rate. *Genetics* **171**, 783-790.

Berger, J. O. (2004). The case for objective Bayesian analysis. *Bayesian Anal.* **1**, 1-17.

Berger, J. O., Liseo, B. and Wolpert, R. L. (1999). Integrated likelihood methods for eliminating nuisance parameters. *Statist. Sci.* **14**, 1-28.

Berger, J. O. and Wolpert, R. L. (1988). *The Likelihood Principle.* 2nd edition.

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. Ser. B* **41**, 113-147.

Bernardo, J. M. (1997). Noninformative priors do not exist: A discussion. *J. Statist. Plann. Inference* **65**, 159-189.

Bickel, D. R. (2004). Degrees of diferential gene expression: Detecting biologically signifcant expression diferences and estimating their magnitudes. *Bioinformatics* **20**, 682-688.

Bickel, D. R. (2008). The strength of statistical evidence for composite hypotheses with an application to multiple comparisons. COBRA Preprint Series, Working Paper 49; URL: `http://biostats.bepress.com/cobra/art49`.

Bickel, D. R. (2011a). Estimating the null distribution to adjust observed confidence levels for genome-scale screening. *Biometrics* **67**, 363-370.

Bickel, D. R. (2011b). A predictive approach to measuring the strength of statistical evidence for single and multiple comparisons. *Canad. J. Statist.* **39**, 610-631.

Bickel, D. R. (2012a). Controlling the degree of caution in statistical inference with the Bayesian and frequentist approaches as opposite extremes. *Electron. J. Statist.* **6** (2012), 686-709.

Bickel, D. R. (2012b). Game-theoretic probability combination with applications to resolving conflicts between statistical methods. *Internat. J. Approximate Reasoning*, DOI:10.1016/j.ijar.2012.04.002 (2012). (Online ahead of print).

Blume, J. D. (2002). Likelihood methods for measuring statistical evidence. *Statist. Medicine* **21**, 2563-2599.

Blume, J. D. (2008). How often likelihood ratios are misleading in sequential trials. *Comm. Statist. Theory Methods* **37**, 1193-1206.

Blume, J. D., Su, L., Olveda, R. M. and McGarvey, S. T. (2007). Statistical evidence for GLM regression parameters: A robust likelihood approach. *Statistics In Medicine* **26**, 2919-2936.

Bochkina, N. and Richardson, S. (2007). Tail posterior probability for inference in pairwise and multiclass gene expression data. *Biometrics* **63**, 1117-1125.

Choi, L., Cafo, B. and Rohde, C. (2008). A survey of the likelihood approach to bioequivalence trials. *Statist. Medicine* **27**, 4874-4894.

Clyde, M. and George, E. I. (2004). Model uncertainty. *Statistical Science* **19**, 81-94.

Cox, D. R. (1977). The role of significance tests. *Scand. J. Statist.* **4**, 49-70.

Cox, D. R. (2006). *Principles of Statistical Inference.* Cambridge University Press, Cambridge.

de Finetti, B. (1970). *Theory of Probability: a Critical Introductory Treatment.* 1st edition. John Wiley and Sons Ltd, New York.

Dempster, A. P. (1997). The direct use of likelihood for signifcance testing. *Statist. Comput.* **7**, 247-252.

Dollinger, M. B., Kulinskaya, E. and Staudte, R. G. (1996). Information, *Statistics and Induction in Science.* World Scientific, Singapore, 119-128.

Edwards, A. W. F. (1992). Likelihood. Johns Hopkins Press, Baltimore.

Efron, B. (2004). Bayesians, frequentists, and physicists. Phystat2003.

Efron, B. (2007). Size, power and false discovery rates. *Annals of Statistics* **35**, 1351-1377.

Estienne, J. E. (1903a). Essai sur l'art de conjecturer. *Revue d'artillerie* **61**, 405-449.

Estienne, J. E. (1903b). Essai sur l'art de conjecturer. *Revue d'artillerie* **62**, 73-117.

Estienne, J. E. (1904a). Essai sur l'art de conjecturer. *Revue d'artillerie* **64**, 5-39.

Estienne, J. E. (1904b). Essai sur l'art de conjecturer. *Revue d'artillerie* **64**, 65-97.

Fisher, R. A. (1925). *Statistical Methods for Research Workers.* Oliver and Boyd, London.

Fisher, R. A. (1973). *Statistical Methods and Scientifc Inference.* Hafner Press, New York.

Foster, J. (2004). In *Readings on Laws of Nature.* University of Pitsburg Press, Pitsburg.

Fraser, D. A. S. (2011). Is Bayes posterior just quick and dirty confidence? *Statist. Sci.* **26**, 299-316

Fraser, D. A. S., Reid, N. and Wong, A. C. M. (2004). Inference for bounded parameters. *Phys. Rev. D* **69**, 033002.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. H. and Zhang, J. (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology* **5**, R80.

Geyer, C. J. and Meeden, G. D. (2005). Fuzzy and randomized confidence intervals and pvalues. *Statist. Sci.* **20**, 358-366.

Goodman, S. N. (1998). Multiple comparisons, explained. *Amer. J. Epidemiology* **147**, 807-812.

Goodman, S. N. and Royall, R. (1988). Evidence and scientific research. *Amer. J. Public Health* 78, 1568-1574.

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

Hald, A. (2007). *A History of Parametric Statistical Inference from Bernoulli to Fisher*, 1713-1935. Springer, New York.

He, Y., Huang, W. and Liang, H. (2007). Axiomatic development of profile likelihoods as the strength of evidence for composite hypotheses. *Comm. Statist. Theory Methods***36**, 2695-2706.

Hoch, J. S. and Blume, J. D. (2008). Measuring and illustrating statistical evidence in a cost efectiveness analysis. *J. Health Economics* **27**, 476-495.

Jefreys, H. (1948). *Theory of Probability.* Oxford University Press, London.

Johnson, V. (2005). Bayes factors based on test statistics. *J. Roy. Statist. Soc. Ser B Statistical Methodology* **67**, 689-701.

Johnstone, D. J. (1986). Tests of significance in theory and practice (with discussion). *The Statistician* **35**, 491-504.

Kalbfeisch, J. D. (2000). Comment on "On the probability of observing misleading statistical evidence". *J. Amer. Statist. Assoc.* **95**, 770-771.

Kalbfeisch, J. D. and Sprott, D. A. (1970). Application of likelihood methods to models involving large numbers of parameters. *J. Roy. Statist. Soc. Ser. B* **32**, 175-208.

Kardaun, O. J. W. F., Salome, D., Schaafsma, W., Steerneman, A. G. M., Willems, J. C. and Cox, D. R. (2003). Refections on fourteen cryptic issues concerning the nature of statistical inference. *Internat. Statist. Rev. / Revue Internationale de Statistique* **71**, 277-303.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* **90,** 773-795.

Kass, R. E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association* **91**, 1343-1370.

Klir, G. J. (2004). Generalized information theory: Aims, results, and open problems. *Reliability Engineering and System Safety* **85**, 21-38.

Korn, E. L. and Freidlin, B. (2006). The likelihood as statistical evidence in multiple comparisons in clinical trials: No free lunch. *Biometrical J.* **48**, 346-355.

Kyburg, H. E. and Teng, C. M. (2001). *Uncertain Inference.* Cambridge University Press, Cambridge.

Kyburg, H. E. and Teng, C. M. (2006). Nonmonotonic logic and statistical inference. *Computational Intelligence* **22**, 26-51.

Lavine, M. and Schervish, M. J. (1999). Bayes factors: What they are and what they are not. *Amer. Statist.***53**, 119-122.

Laviolette, M. (2004). Comment on "Membership functions and probability measures of fuzzy sets". *J. Amer. Statist. Assoc.* **99**, 879-880.

Lehmann, E. (2006). On likelihood ratio tests. *IMS Lecture Notes Monograph Series* **49**, 1-8.

Lele, S. R. (2004). Evidence functions and the optimality of the law of likelihood. *The Nature of Scientific Evidence: Statistical, Philosophical, and Empirical Considerations*. University of Chicago Press, Chicago, pp.191-216.

Lewin, A., Richardson, S., Marshall, C., Glazier, A. and Aitman, T. (2006). Bayesian modeling of differential gene expression. *Biometrics* **62**, 1-9.

Lindley, D. V. (2004). Comment on "Membership functions and probability measures of fuzzy sets". *J. Amer. Statist. Assoc.* **99**, 877-879.

MacKay, D. J. (2002). *Information Theory, Inference and Learning Algorithms.* Cambridge University Press, Cambridge.

Mayo, D. G. and Cox, D. R. (2006). Frequentist statistics as a theory of inductive inference. *IMS Lecture Notes Monograph Series*, The Second Erich L. Lehmann Symposium Optimality.

McCarthy, D. J. and Smyth, G. K. (2009). Testing significance relative to a foldchange threshold is a TREAT. *Bioinformatics* **25**, 765-771.

Montazeri, Z., Yanofsky, C. M. and Bickel, D. R. (2010). Shrinkage estimation of efect sizes as an alternative to hypothesis testing followed by estimation in high-dimensional biology: Applications to diferential gene expression. *Statistical Applications in Genetics and Molecular Biology* **9**, 23.

Nguyen, H. T. and Walker, E. A. (2000). *A First Course in Fuzzy Logic.* CRC Press, London.

Niiniluoto, I. (2004). *Induction and Deduction in the Sciences.* Springer, New York.

Osteyee, D. B. and Good, I. J. (1974). Information, Weight of Evidence, the Singularity between Probability *Measures and Signal Detection.* Springer-Verlag, New York.

Owen, A. B. (2001). *Empirical Likelihood.* Chapman and Hall/CRC, London.

Pasterkamp, R. J., Peschon, J. J., Spriggs, M. K. and Kolodkin, A. L. (2003). Semaphorin 7a promotes axon outgrowth through integrins and mapks. *Nature* **424**, 398-405.

Pollard, K. S., Dudoit, S. and van der Laan, M. J. (2005). Multiple testing procedures: The multtest package and applications to genomics. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, 249-271.

Popper, K. (2002). *Logic of Scientifc Discovery.* Routledge, London.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rényi, A. (1970). *Probability Theory.* North-Holland, Amsterdam.

Robbins, H. (1970). Statistical methods related to the law of the iterated logarithm. *Ann. Math. Statist.* **41**, 1397-1409.

Romer, C., Kandel, A. and Backer, E. (1995). Fuzzy partitions of the sample space and fuzzy parameter hypotheses. *IEEE Transactions on Systems, Man and Cybernetics* **25**, 1314-1322.

Royall, R. (1992). The elusive concept of statistical evidence. *Bayesian Statist.* **4**, 405-418.

Royall, R. (1997). *Statistical Evidence: A Likelihood Paradigm.* CRC Press, New York.

Royall, R. (2000a). On the probability of observing misleading statistical evidence. *J. Amer. Statist. Assoc.* **95**, 760-768.

Royall, R. (2000b). Rejoinder to comments on R. Royall, "On the probability of observing misleading statistical evidence". *J. Amer. Statist. Assoc.* **95**, 773-780.

Royall, R. and Tsou, T. S. (2003). Interpreting statistical evidence by using imperfect models: Robust adjusted likelihood functions. *J. Roy. Statist. Soc. Ser. B* **65**, 391-404.

Savage, L. J. (1954). *The Foundations of Statistics.* John Wiley and Sons, New York.

Schervish, M. J. (1996). P values: What they are and what they are not. *American Statistician* **50**, 203-206.

Schweder, T. and Hjort, N. L. (2002). Confidence and likelihood. *Scandinavian Journal of Statistics* **29**, 309-332.

Severini, T. (2010). Likelihood ratio statistics based on an integrated likelihood. *Biometrika* **97**, 481-496.

Severini, T. A. (2007). Integrated likelihood functions for nonBayesian inference. *Biometrika* **94**, 529-542.

Shafer, J. P. (1995). Multiple hypothesis testing. *Ann. Rev. Psychology* **46**, 561-584.

Singpurwalla, N. D. and Booker, J. M. (2004). Membership functions and probability measures of fuzzy sets. *J. Amer. Statist. Assoc.***99**, 867-877.

Sokhansanj, B. A., Fitch, J. P., Quong, J. N. and Quong, A. A. (2004). Linear fuzzy gene network models obtained from microarray data by exhaustive search. *BMC Bioinformatics* **5**.

Spicer, L. J. and Francisco, C. C. (1997). The adipose obese gene product, leptin: Evidence of a direct inhibitory role in ovarian function. *Endocrinology* **138**, 3374-3379.

Spjøtvoll, E. (1977). Comment on "The role of significance tests". *Scand. J. Statist.* **4**, 63-66.

Storey, J. D. (2002). A direct approach to false discovery rates. *J. Roy. Statist. Soc. Ser. B: Statistical Methodology* **64**, 479-498.

Strug, L. J. and Hodge, S. E. (2006). An alternative foundation for the planning and evaluation of linkage analysis. ii. implications for multiple test adjustments. *Human Heredity* **61**, 200-209.

Strug, L. J., Rohde, C. A. and Corey, P. N. (2007). An introduction to evidential sample size calculations. *Amer. Statist.* **61**, 207-212.

Tsou, T. S. and Royall, R. (1995). Robust likelihoods. *J. Amer. Statist. Assoc.* **90**, 316-320.

Van De Wiel, M. A. and Kim, K. I. (2007). Estimating the false discovery rate using nonparametric deconvolution. *Biometrics* **63**, 806-815.

Van der Laan, M. J., Dudoit, S. and Pollard, K. S. (2004). Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Stat. Appl. in Genet. and Mol. Biol.* **3**, 15.

Westfall, P. H., Johnson, W. O. and Utts, J. M. (1997). A Bayesian perspective on the Bonferroni adjustment. *Biometrika* **84**, 419-427.

Wilkinson, G. N. (1977). On resolving the controversy in statistical inference (with discussion). *J. Roy. Statist. Soc. Ser. B* **39**, 119-171.

Williamson, J. (2005). *Bayesian Nets and Causality.* Oxford University Press, London.

Wright, S. P. (1992). Adjusted p-values for simultaneous inference. *Biometrics* **48**, 1005-1013.

Yanofsky, C. M. and Bickel, D. R. (2010). Validation of differential gene expression algorithms: Application comparing fold-change estimation to hypothesis testing. *BMC Bioinformatics* **11**, 63.

Yekutieli, D., ReinerBenaim, A., Benjamini, Y., Elmer, G. I., Kafkaf, N., Letwin, N. E. and Lee, N. H. (2006). Approaches to multiplicity issues in complex research in microarray analysis. *Statist. Neerlandica* **60**, 414-437.

Zadeh, L. A. (2002). Toward a perception-based theory of probabilistic reasoning with imprecise probabilities. *J. Statist. Plann. Inference* **105**, 233-264.

Zhang, Z. (2009a). Interpreting statistical evidence with empirical likelihood functions. *Biometrical J.* **51**, 710-720.

Zhang, Z. (2009b). A law of likelihood for composite hypotheses. arXiv:0901.0463.

Ottawa Institute of Systems Biology, Department of Biochemistry, Microbiology, and Immunology, Department of Mathematics and Statistics, University of Ottawa, 451 Smyth Road, Ottawa, Ontario K1H 8M5, Canada.

E-mail: dbickel@uottawa.ca

(Received May 2009; accepted August 2011)