1. Introduction

1.1. Decision-theoretic inference and evidential inference

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 most-developed Bayesian framework, that of F. P. Ramsey (cited in Jeffreys (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 increases monotonically with how much the rational decision-maker 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-Ramsey 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., Jeffreys (1948); Kass and Raftery (1995)). Accordingly, Osteyee and Good (1974) called 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 if the two hypotheses considered are mutually exclusive and jointly exhaustive.

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

*Content last modified on 3 November 2008; uploaded to COBRA on 3 December 2008.

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 a low p-value as strong evidence against the null hypothesis; e.g., Fraser *et al.* (2004). Even the critics of significance testing acknowledge that it serves its purpose in some situations (Spjøtvoll, 1977; Goodman and Royall, 1988).

In spite of the uncontested value of methods of the Neyman-Pearson and Bayes-Ramsey 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 given a family of probability distributions must meet both of these necessary criteria:

- the *coherence* condition, that strength of evidence is always consistent with the rules of logic;
- the *objectivity* condition, that the strength of evidence does not vary from one researcher to another.

Schervish (1996) and Lavine and Schervish (1999) point out that a candidate measure of the strength of evidence is illogical or *incoherent* if it can assign more support to a hypothesis than to a hypothesis it implies; candidates that cannot do so are considered coherent. For example, an incoherent candidate might say an observation of parents' eye colors supports the hypothesis that their child will have brown eyes over the hypothesis that she will have either blue eyes or brown eyes. The incoherence of the p-value as a measure of support is apparent when comparing one-sided and two-sided p-values under the same model (Schervish, 1996; Royall, Statistical Evidence: A Likelihood Paradigm, 1997a). For a scalar parameter θ, say the null hypothesis $\theta = 0$ is tested with $\theta \neq 0$ and $\theta > 0$ as the alternative hypotheses. If the p-value of the two-sided test is twice that of the one-sided test, then significance testing would attach more evidence to the hypothesis that $\theta > 0$ than to the hypothesis that $\theta \neq 0$ relative to the same null hypothesis; this is incoherent since $\theta > 0 \Rightarrow \theta \neq 0$. If the significance level lies between the two p-values, then $\theta > 0$ but not $\theta \neq 0$ would be accepted over the null hypothesis.

That the Bayes factor is likewise incoherent as a measure of evidence (Lavine and Schervish, 1999) is evident in the case of nested hypotheses. Consider the observation x of a discrete random variable X. Based on prior predictive mass

functions P_1 and P_2 corresponding to hypotheses $\theta = 0$ and $-1 < \theta < 1$, respectively, the Bayes factor

$$P_1 (X = x) / P_2 (X = x)$$

will be greater than 1 if the maximum likelihood estimate is sufficiently close to 0 and if the prior density of θ is nonzero for all $\theta \in (-1, 1)$. In this case, the logarithm of the Bayes factor as the weight of evidence would attribute more support to $\theta = 0$ than to $-1 < \theta < 1$, and yet $\theta = 0 \Rightarrow -1 < \theta < 1$.

Even so, the Bayes factor may instead be used to compute a ratio of posterior probabilities of the hypotheses in question, and such a ratio would satisfy the coherence condition (Lavine and Schervish, 1999). In the strict Bayes-Ramsey 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. Much of applied Bayesian analysis is less strict, and the effort required to elicit prior distributions from experts to adequately reflect their levels of uncertainty about parameter values is rarely made, perhaps because it is justifiable in very few practical situations. The less subjective practice of automatically assigning 50% prior probability to each hypothesis sacrifices coherence by reducing the ratio of posterior probabilities to the Bayes factor. 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). Further, the automatic generation of priors introduces a problem of interpretation since it implicitly rejects probability as a level of belief as defined by Bayes-Ramsey decision theory unless the default priors in fact approximate someone's levels of belief. Consequently, a default prior often serves to determine what a hypothetical individual 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 frequentist inference, objectivity is again purchased at the price of coherence.

By contrast, the likelihood ratio satisfies both of the necessary conditions for a measure of the strength of statistical evidence; it is coherent in the above sense (Lavine and Schervish, 1999) without resorting to levels of belief, hypothetical or otherwise. 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). The law is usually restated as follows. At each value of θ, the p-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. $L(\bullet) = L(\bullet; x) = f(\bullet; \theta)$, a function on the paramter space Θ, is called the *likelihood function*. In the evidential framework

of statistical inference, the likelihood ratio $L\left(\theta';x\right)/L\left(\theta'';x\right)$ is the strength of the statistical evidence in $X=x$ that supports $\theta=\theta'$ over $\theta=\theta''$, and if $L\left(\theta';x\right)/L\left(\theta'';x\right)>1$, there is more evidence for $\theta=\theta'$ than for $\theta=\theta''$ (Royall, On the probability of observing misleading statistical evidence, 2000b). 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 Bayes factor weight of evidence (Section 1.1) equals $\log\left(L\left(\theta';x\right)/L\left(\theta'';x\right)\right)$, which Edwards (1992) called the support for $\theta=\theta'$ over $\theta=\theta''$.

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 *et al.*, 2007; Blume, How often likelihood ratios are misleading in sequential trials, 2008b). This analog, the probability of observing misleading evidence, is defined as follows. Consider the strength of evidence in a random sample of data drawn from a reference distribution and the strength of evidence in that sample against the reference distribution in favor of the hypothesis that the generating distribution is in a given set of other distributions. 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 relative frequency of observations of misleading evidence under infinitely 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 at a fixed Type I error rate fails in this regard since measuring the strength of evidence by the p-value results in the same probability of observing misleading evidence for all samples sizes. 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, Statistical Evidence: A Likelihood Paradigm, 1997a). 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.

1.2. Evidence for a composite hypothesis

The classical law of likelihood is insufficient for statistical inference if either hypothesis is composite, that is, if it corresponds to multiple parameter values, each an element of some $\Theta'\subseteq\Theta$. This insufficiently 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 can inform the replacement of a composite hypothesis $\theta \in \Theta'$ with a simple hypothesis $\theta = \theta'$ in order to compute $L\left(\theta'\right)/L\left(\theta''\right)$ as the strength of statistical 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 $\left|\theta' - \theta''\right|$ 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 $\Phi\left(-\sqrt{2\log\Lambda}\right)$ if L is smooth and if p is fixed, where Φ is the standard normal cumulative density function, or by $1/\Lambda$ more universally (Royall, On the probability of observing misleading statistical evidence, 2000b). In addition, limiting the parameter of interest to one of two values is convenient when planning the size of a study (Strug *et al.*, 2007).

Nonetheless, the strength of statistical evidence involving a composite hypothesis cannot in general be measured or even approximated by substituting a simple hypothesis selected prior to observing the data. This can be seen in the problem of quantifying the strength of evidence favoring the composite hypothesis that θ, the mean of a normal distribution of unknown variance, is greater than θ'' to the simple hypothesis that it equals θ''. Replacing the standard deviation σ with $\widehat{\sigma}(\theta)$, its value that maximizes the likelihood when the mean is θ, gives the *profile likelihood* $L_{\text{profile}}(\theta) = L(\theta, \widehat{\sigma}(\theta))$ instead of the likelihood $L(\theta, \sigma)$. The use of some $\theta' > \theta''$ in the simple hypothesis $\theta = \theta'$ as a surrogate for the composite hypothesis $\theta > \theta''$ leads to $L_{\text{profile}}\left(\theta'\right)/L_{\text{profile}}\left(\theta''\right) = \left(\widehat{\sigma}\left(\theta'\right)/\widehat{\sigma}\left(\theta''\right)\right)^{-n}$ as the approximate strength of statistical evidence for $\theta = \theta'$ over $\theta = \theta''$. No matter what fixed value was chosen for θ', some sample x may be observed that is sufficiently far from typical samples of both θ' and θ'' that there is arbitrarily little approximate evidence for either simple hypothesis over the other: $\forall_{\delta>0} \lim_{\theta/\sigma\to\infty} P\left(\left|1 - \left(\widehat{\sigma}\left(\theta'\right)/\widehat{\sigma}\left(\theta''\right)\right)^{-n}\right| < \delta\right) = 1$. Since $L_{\text{profile}}\left(\theta'\right)/L_{\text{profile}}\left(\theta''\right)$ approaches the value representing no evidence as $\theta - \theta''$ increases with σ, θ', and θ'' fixed, it completely fails to approximate the strength of statistical evidence for $\theta > \theta'$ over $\theta = \theta''$.

A general solution to the composite hypothesis problem is implicit in the use of a likelihood interval or more general likelihood set. The level-Λ likelihood set $\mathcal{E}(\Lambda)$ consists of all values of θ satisfying $L(\theta) \geq L\left(\widehat{\theta}\right)/\Lambda$, where $\widehat{\theta}$ is the

maximum likelihood estimate. Membership in a likelihood set determines which parameter values are considered consistent 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 strength of statistical 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 strength of statistical evidence for the composite hypotheses $\theta \neq \theta''$ over the simple hypothesis $\theta = \theta''$.

A discrepancy between the performance of the likelihood ratio for two fixed simple hypotheses and the likelihood ratio maximized over a subset of 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 (Kalbfleisch, 2000). Asymptotically, for any fixed θ' and θ'' in $\Theta = \mathbb{R}^5$, there is a 2.1% upper bound on $P_{\theta''}\left(L\left(\theta'\right)/L\left(\theta''\right) > 8\right)$, the probability of observing misleading evidence at level $\Lambda = 8$ (Royall, On the probability of observing misleading statistical evidence, 2000b). By contrast, the probability that the level-8 likelihood set contains θ'', assuming it is the true value of θ, is less than 50% (Kalbfleisch, 2000). This means the asymptotic probability of observing misleading evidence for $\theta \in \mathbb{R}^5 \setminus \left\{\theta''\right\}$ 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 $P_{\theta''}\left(L\left(\theta'\right)/L\left(\theta''\right) > 8\right)$ is 12.5%, more than a factor of 4 smaller than $P_{\theta''}\left(L\left(\widehat{\theta}\right)/L\left(\theta''\right) > 8\right) = 52.7\%$ in the example of $p = 5$ and conditions under which $2\log\left(L\left(\widehat{\theta}\right)/L\left(\theta''\right)\right)$ is asymptotically distributed as χ^2 with p degrees of freedom.

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

$$\forall_{\Lambda>1} \lim_{n\to\infty} P_{\theta''}\left(L\left(\widehat{\theta}\right)/L\left(\theta''\right) > \Lambda\right) > 0.$$

Thus, $L\left(\widehat{\theta}\right)/L\left(\theta''\right)$ 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, which highlights connections between Hacking's law of likelihood, evidence sets, and evidence for or against composite hypotheses.

The main drawback of replacing a simple hypothesis with an interval hypothesis is the sensitive 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. 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, Statistical Evidence: A Likelihood Paradigm, 1997a). This issue is discussed further in Section 6, which concludes with a sketch of opportunities for further research.

2. Evidential inference about precise hypotheses

2.1. Preliminaries

The symbols $\subset$ and $\subseteq$ designate proper subsets and (possibly improper) subsets, respectively. Consider the fixed positive integer p and the parameter space $\Theta \subseteq \mathbb{R}^p$. For all $\theta \in \Theta$, the distribution of the observable random variable $X \in \Omega \subseteq \mathbb{R}^n$ has a probability measure admitting 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 $\left\{y : y \in \Omega, \forall_{\theta' \in \Theta} f(y;\theta') > 0\right\}$.

Operators $\wedge$ and $\vee$ will be used for the minimum and maximum, respectively. In addition, for any set S and functions $g' : S \to \mathbb{R}$ and $g'' : S \to \mathbb{R}$, $\forall_{u \in S} (g' \wedge g'')(u) = g'(u) \wedge g''(u)$ and $\forall_{u \in S} (g' \vee g'')(u) = g'(u) \vee g''(u)$.

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.

In the sequel, every subset of Θ is nonempty and thus corresponds to either a simple hypothesis or a composite hypothesis.

2.2. Evidential theory

Restated in these terms, the original law of likelihood governs the special case of comparing two simple hypotheses:

Axiom 1 Special law of likelihood. *For all $\theta' \in \Theta$ and $\theta'' \in \Theta$, the strength of the statistical evidence in $X = x$ that supports $\theta = \theta'$ over $\theta = \theta''$ is*

$$\mathrm{ev}\left(\left\{\theta'\right\},\left\{\theta''\right\}\right) = \mathrm{ev}\left(\left\{\theta'\right\},\left\{\theta''\right\};x\right) = \frac{L\left(\theta';x\right)}{L\left(\theta'';x\right)}. \tag{1}$$

Mathematically, the special law of likelihood in effect defines what is meant by the strength of evidence for one simple hypothesis over another. It does not specify how to measure the strength of evidence for or against a composite hypothesis (Royall, Rejoinder to comments on R. Royall, 'On the probability of observing misleading statistical evidence', 2000c; Blume, Likelihood methods for measuring statistical evidence, 2002a). Such measurement is made possible by precisely defining what is meant by the strength of evidence when a composite hypothesis is involved:

Axiom 2 General law of likelihood. *For all $\Theta' \subseteq \Theta$ and $\Theta'' \subseteq \Theta$, the strength of the statistical evidence in $X = x$ that supports $\theta \in \Theta'$ over $\theta \in \Theta''$ is $\mathrm{ev}\left(\Theta',\Theta''\right) = \mathrm{ev}\left(\Theta',\Theta'';x\right)$, where ev is the function satisfying the following two coherence and non-accumulation conditions as well as Axiom 1 and*

$$\forall_{\Theta',\Theta'',\Theta'''\subseteq\Theta}\, \mathrm{ev}\left(\Theta',\Theta'''\right) \leq \mathrm{ev}\left(\Theta'',\Theta'''\right) \Leftrightarrow \mathrm{ev}\left(\Theta''',\Theta';x\right) \geq \mathrm{ev}\left(\Theta''',\Theta''\right).$$

Condition 1 Coherence. *If $\mathrm{ev}\left(\Theta',\Theta'';x\right)$ is the strength of the statistical evidence in $X = x$ that supports $\theta \in \Theta'$ over $\theta \in \Theta''$ for all $\Theta' \subseteq \Theta$ and $\Theta'' \subseteq \Theta$, then $\forall_{\Theta'',\Theta'''\subseteq\Theta}\forall_{\Theta'\subseteq\Theta''}\, \mathrm{ev}\left(\Theta',\Theta''';x\right) \leq \mathrm{ev}\left(\Theta'',\Theta''';x\right)$.*

Condition 2 Non-accumulation. *If $\mathrm{ev}\left(\Theta',\Theta'';x\right)$ is the strength of the statistical evidence in $X = x$ that supports $\theta \in \Theta'$ over $\theta \in \Theta''$ for all $\Theta' \subseteq \Theta$ and $\Theta'' \subseteq \Theta$, then $\forall_{\Theta',\Theta''\subseteq\Theta}\exists_{\theta'\in\Theta'}\, \mathrm{ev}\left(\left\{\theta'\right\},\Theta'';x\right) \geq \mathrm{ev}\left(\Theta',\Theta'';x\right)$.*

The non-accumulation condition prevents attributing more evidence to a composite hypothesis than to any of its constituent simple hypotheses, as the posterior probability does by integrating the likelihood function of each hypothesis with respect to the same prior distribution. The coherence condition prevents the logical fallacy of attributing more evidence to a hypothesis than to an implication of that hypothesis (Schervish, 1996; Lavine and Schervish, 1999), as noted in Section 1.1. It is now clear that neither the p-value nor the Bayes factor qualifies as a coherent measure of evidence:

Proposition 1 *Let $\mathrm{ev_p}\left(\Theta_A,\left\{\theta_0\right\};x\right)$ be 1 minus a p-value for a test of null hypothesis $\theta = \theta_0$ against alternative hypothesis $\theta \in \Theta_A$. Condition 1 contradicts the assertion that, for all $\theta_0 \in \Theta$ and $\Theta_A \subseteq \Theta\backslash\left\{\theta_0\right\}$, $\mathrm{ev_p}\left(\Theta_A,\left\{\theta_0\right\};x\right)$ is the strength of the statistical evidence in $X = x$ that supports $\theta \in \Theta_A$ over $\theta = \theta_0$.*

Proof. Consider the two-sided alternative hypothesis $\theta \neq \theta_0$ and the one-sided alternative hypotheses $\theta < \theta_0$ and $\theta > \theta_0$ for $\Theta = \mathbb{R}$. Under commonly used models,

$$1 - \mathrm{ev_p}\left(\mathbb{R}\backslash\left\{\theta_0\right\},\left\{\theta_0\right\};x\right) = 2 \inf_{\Theta_A\in\{(-\infty,\theta_0),(\theta_0,\infty)\}} \left(1 - \mathrm{ev_p}\left(\Theta_A,\left\{\theta_0\right\};x\right)\right)$$

and thus either

$$\mathrm{ev_p}\left(\left(-\infty,\theta_0\right),\left\{\theta_0\right\};x\right) > \mathrm{ev_p}\left(\mathbb{R}\backslash\left\{\theta_0\right\},\left\{\theta_0\right\};x\right)$$

or

$$\mathrm{ev_p}\left(\left(\theta_0,\infty\right),\left\{\theta_0\right\};x\right) > \mathrm{ev_p}\left(\mathbb{R}\backslash\left\{\theta_0\right\},\left\{\theta_0\right\};x\right)$$

even though $(-\infty,\theta_0)\subset\mathbb{R}\backslash\{\theta_0\}$ and $(\theta_0,\infty)\subset\mathbb{R}\backslash\{\theta_0\}$, in violation of Condition 1. ■

Proposition 2 *Let* $\mathrm{ev_{BF}}\left(\Theta',\Theta'';x\right)$ *be the Bayes factor*

$$\mathrm{ev_{BF}}\left(\Theta',\Theta'';x\right)=\frac{\int_{\Theta'}f\left(x;\theta'\right)d\mu'\left(\theta'\right)}{\int_{\Theta''}f\left(x;\theta''\right)d\mu''\left(\theta''\right)},$$

where μ' *and* μ'' *are the probability measures representing the prior distributions of* θ *on* Θ' *and* Θ''*, respectively. Condition 1 contradicts the assertion that, for all* $\Theta'\subseteq\Theta$ *and* $\Theta''\subseteq\Theta$*,* $\mathrm{ev_{BF}}\left(\Theta',\Theta'';x\right)$ *is the strength of the statistical evidence in* $X=x$ *that supports* $\theta\in\Theta'$ *over* $\theta\in\Theta''$*.*

Proof. If there is a unique maximum likelihood value, $\widehat{\theta}$, and if μ'' on Θ has support outside $\theta=\widehat{\theta}$, then

$$\begin{aligned}\mathrm{ev_{BF}}\left(\left\{\widehat{\theta}\right\},\Theta;x\right) &= \frac{L\left(\widehat{\theta};x\right)}{\int_{\Theta}L\left(\theta'';x\right)d\mu''\left(\theta''\right)}\\ &> 1.\end{aligned}$$

Therefore, $\forall_{\Theta'''\subseteq\Theta}\,\mathrm{ev_{BF}}\left(\left\{\widehat{\theta}\right\},\Theta''';x\right) > \mathrm{ev_{BF}}\left(\Theta,\Theta''';x\right)$ even though $\left\{\widehat{\theta}\right\}\subset\Theta$, against Condition 1. ■

Following Jeffreys (1948) with the strength of statistical evidence in place of the Bayes factor and with a slight change of wording, the number of achieved bans ($b=\log_{10}\mathrm{ev}\left(\Theta',\Theta'';x\right)$) 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 $b<0$. The next result facilitates the measurement of the strength of statistical evidence for or against composite hypotheses.

Proposition 3 *For all* $\Theta'\subseteq\Theta$ *and* $\Theta''\subseteq\Theta$*, the strength of the statistical evidence in* $X=x$ *that supports* $\theta\in\Theta'$ *over* $\theta\in\Theta''$ *is*

$$\mathrm{ev}\left(\Theta',\Theta''\right)=\mathrm{ev}\left(\Theta',\Theta'';x\right)=\frac{\sup_{\theta'\in\Theta'}L\left(\theta';x\right)}{\sup_{\theta''\in\Theta''}L\left(\theta'';x\right)}. \tag{2}$$

Proof. Equation (2) obviously satisfies Condition 1 (as Lavine and Schervish

(1999) observed), Condition 2, and, whenever $\Theta' = \left\{\theta'\right\}$ and $\Theta'' = \left\{\theta''\right\}$, equation (1). To prove that the solution is unique, first assume

$$\operatorname{ev}\left(\Theta'_*, \Theta''_*\right) = \sup_{\theta' \in \Theta'_*} \inf_{\theta'' \in \Theta''_*} \operatorname{ev}\left(\left\{\theta'\right\}, \left\{\theta''\right\}\right) \tag{3}$$

holds for some $\Theta'_*, \Theta''_* \subseteq \Theta$. Then Condition 1 implies that

$$\begin{aligned} \forall_{\overline{\theta}' \in \Theta \setminus \Theta'_*} \operatorname{ev}\left(\Theta'_* \cup \left\{\overline{\theta}'\right\}, \Theta''_*\right) &\geq \operatorname{ev}\left(\Theta'_*, \Theta''_*\right) \vee \operatorname{ev}\left(\left\{\overline{\theta}'\right\}, \Theta''_*\right) \\ &= \sup_{\theta' \in \Theta'_* \cup \left\{\overline{\theta}'\right\}} \operatorname{ev}\left(\left\{\theta'\right\}, \Theta''_*\right) \end{aligned}$$

and, likewise, that

$$\forall_{\overline{\theta}'' \in \Theta \setminus \Theta''_*} \operatorname{ev}\left(\Theta'_*, \Theta''_* \cup \left\{\overline{\theta}''\right\}\right) \leq \inf_{\theta'' \in \Theta''_* \cup \left\{\overline{\theta}''\right\}} \operatorname{ev}\left(\Theta'_*, \left\{\theta''\right\}\right).$$

But Condition 2 implies that

$$\begin{aligned} \forall_{\overline{\theta}' \in \Theta \setminus \Theta'_*} \operatorname{ev}\left(\Theta'_* \cup \left\{\overline{\theta}'\right\}, \Theta''_*\right) &\leq \sup_{\theta' \in \Theta'_*} \operatorname{ev}\left(\left\{\theta'\right\}, \Theta''_*\right) \vee \operatorname{ev}\left(\left\{\overline{\theta}'\right\}, \Theta''_*\right) \\ &= \sup_{\theta' \in \Theta'_* \cup \left\{\overline{\theta}'\right\}} \operatorname{ev}\left(\left\{\theta'\right\}, \Theta''_*\right) \end{aligned}$$

and, similarly, that

$$\forall_{\overline{\theta}'' \in \Theta \setminus \Theta''_*} \operatorname{ev}\left(\Theta'_*, \Theta''_* \cup \left\{\overline{\theta}''\right\}\right) \geq \inf_{\theta'' \in \Theta''_* \cup \left\{\overline{\theta}''\right\}} \operatorname{ev}\left(\Theta'_*, \left\{\theta''\right\}\right).$$

Therefore, the assumption that equation (3) holds for some $\Theta'_*, \Theta''_* \subseteq \Theta$ implies that

$$\operatorname{ev}\left(\Theta', \Theta''\right) = \sup_{\theta' \in \Theta'} \inf_{\theta'' \in \Theta''} \operatorname{ev}\left(\left\{\theta'\right\}, \left\{\theta''\right\}\right) \tag{4}$$

holds for all $\Theta', \Theta'' \subseteq \Theta$ such that $\Theta'_* \subset \Theta'$ and $\Theta''_* \subset \Theta''$. Now equation (3) holds for $\Theta'_* = \left\{\theta'\right\} \subseteq \Theta$ and $\Theta''_* = \left\{\theta''\right\} \subseteq \Theta$ for all $\theta' \in \Theta$ and $\theta'' \in \Theta$. Thus, equation (4) holds for all $\Theta', \Theta'' \subseteq \Theta$. From the substitution provided by equation (1), $\forall_{\Theta', \Theta'' \subseteq \Theta} \operatorname{ev}\left(\Theta', \Theta''\right) = \sup_{\theta' \in \Theta'} \inf_{\theta'' \in \Theta''} L\left(\theta'; x\right) / L\left(\theta''; x\right)$. ■

The probably of observing misleading evidence mentioned in Section 1.1 is now defined more generally.

Definition 2 *If, for all $\Theta' \subseteq \Theta$ and $\Theta'' \subseteq \Theta$, $\operatorname{ev}\left(\Theta', \Theta''; x\right)$ is the strength of the statistical evidence in $X = x$ that supports $\theta \in \Theta'$ over $\theta \in \Theta''$, then 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*

$$\alpha_{\theta''}\left(\Lambda; \Theta', \Theta''\right) = P\left(\operatorname{ev}\left(\Theta', \Theta''; X_{\theta''}\right) \geq \Lambda\right),$$

where the random variable $X_{\theta''}$ has probability density or mass function $f\left(\bullet; \theta''\right)$.

In the case that one of two mutually exclusive hypotheses is a composite hypotheses corresponding to a parameter interval, the probability of observing misleading evidence approaches 0 as the sample size increases. As argued in Section 1.1, that property is needed to interpret a level of evidence apart from the sample size. Accordingly, the strength of statistical evidence almost always asymptotically selects the correct hypothesis:

Proposition 4 Consistency. *Suppose X is a random variable with probability mass or density function $f(\bullet;\theta)$. Under regularity conditions ensuring the weak consistency of the maximum likelihood estimate of θ,*

$$\forall_{\theta_-\in\mathbb{R},\theta_+\in(\theta_-,\infty)} \lim_{n\to\infty} P\left(\mathrm{ev}\left(\left[\theta_-,\theta_+\right],\Theta\backslash\left[\theta_-,\theta_+\right];X\right)>1\right)=1$$

if $\theta\in(\theta_-,\theta_+)$ or

$$\forall_{\theta_-\in\mathbb{R},\theta_+\in(\theta_-,\infty)} \lim_{n\to\infty} P\left(\mathrm{ev}\left(\left(\theta_-,\theta_+\right),\Theta\backslash\left(\theta_-,\theta_+\right);X\right)<1\right)=1$$

if $\theta\in\Theta\backslash[\theta_-,\theta_+]$.

Proof. In the case that $\theta\in(\theta_-,\theta_+)$, from $\forall_{\delta>0}\lim_{n\to\infty}P\left(\left|\widehat{\theta}-\theta\right|<\delta\right)=1$, it follows that $\forall_{\theta_-\in\mathbb{R},\theta_+\in(\theta_-,\infty)}\lim_{n\to\infty}P\left(\widehat{\theta}\in[\theta_-,\theta_+]\right)=1$. Proposition 3 indicates that $\widehat{\theta}\in[\theta_-,\theta_+]\Rightarrow\mathrm{ev}\left(\left[\theta_-,\theta_+\right],\Theta\backslash\left[\theta_-,\theta_+\right];X\right)>1$. Similarly, in the case that $\theta\in\Theta\backslash[\theta_-,\theta_+]$, from $\forall_{\delta>0}\lim_{n\to\infty}P\left(\left|\widehat{\theta}-\theta\right|<\delta\right)=1$, it follows that $\forall_{\theta_-\in\mathbb{R},\theta_+\in(\theta_-,\infty)}\lim_{n\to\infty}P\left(\widehat{\theta}\in\Theta\backslash\left(\theta_-,\theta_+\right)\right)=1$. Proposition 3 indicates that $\widehat{\theta}\in\Theta\backslash\left(\theta_-,\theta_+\right)\Rightarrow\mathrm{ev}\left(\left(\theta_-,\theta_+\right),\Theta\backslash\left(\theta_-,\theta_+\right);X\right)<1$. ■

Proposition 5 Interpretability. *Suppose X is a random variable with probability mass or density function $f(\bullet;\theta)$. Under regularity conditions ensuring the weak consistency of the maximum likelihood estimate of θ,*

$$\forall_{\theta_-\in\mathbb{R},\theta_+\in(\theta_-,\infty),\theta''\in(\theta_-,\theta_+),\Lambda>1} \lim_{n\to\infty}\alpha_{\theta''}\left(\Lambda;\Theta\backslash\left[\theta_-,\theta_+\right],\left[\theta_-,\theta_+\right]\right)=0.$$

Proof. By Propositions 4 and 3,

$$\lim_{n\to\infty}P\left(\mathrm{ev}\left(\Theta\backslash\left[\theta_-,\theta_+\right],\left[\theta_-,\theta_+\right];X_{\theta''}\right)<\Lambda\right)=1$$

for all $\theta_-\in\mathbb{R}$, $\theta_+\in(\theta_-,\infty)$, $\theta''\in(\theta_-,\theta_+)$, and $\Lambda>1$. Definition 2 completes the proof. ■

2.3. Likelihood sets and composite hypotheses

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 3 *Given some fixed* $\Lambda > 1$*, the* likelihood set of level Λ *for* $X = x$ *is*

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

Definition 4 *Given some fixed* $\beta \in \mathbb{R}$ *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 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 (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 general law of likelihood:

Proposition 6 *Given* $\mathcal{E}(\Lambda)$*, the likelihood set of level* Λ *for* $X = x$*,*

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

Proof. The result follows immediately from Proposition 3 and Definition 3. ■

In short, the practice of considering a parameter value insufficiently supported by the data if it falls outside a likelihood set receives formal justification from measuring the strength of evidence for a composite hypothesis by its best-supported parameter value.

2.3.1. Bioequivalence

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 *et al.* (2007) 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 general law of likelihood. In order to accomodate multidimensional parameters, the implication is stated in terms of equivalence intervals and likelihood intervals rather than equivalence sets and likelihood sets. 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 7 *The strength of the statistical 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 $\sup_{\theta' \in \Theta'} \inf_{\theta'' \notin \Theta'} L\left(\theta'; x\right) / L\left(\theta''; x\right) > \Lambda$, the left-hand side of which equals $\mathrm{ev}\left(\Theta', \Theta \backslash \Theta'; x\right)$ by Proposition 3, proving sufficiency. To prove necessity, assume there is a value θ'' that is in $\mathcal{E}(\Lambda)$ but not in Θ'. Given $\mathrm{ev}\left(\Theta', \Theta \backslash \Theta'; x\right) > \Lambda$, Proposition 3 yields $\sup_{\theta' \in \Theta'} L\left(\theta'; x\right) > \Lambda L\left(\theta''; x\right)$ since $\theta'' \in \Theta \backslash \Theta'$. Because $\theta'' \in \mathcal{E}(\Lambda)$, we have $\sup_{\theta' \in \Theta} L\left(\theta'; x\right) \leq \Lambda L\left(\theta''; x\right)$, producing a contradiction. ■

2.4. 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 $\theta \in \Theta \subseteq \mathbb{R}^{p}$; both ν and p are fixed positive integers. The likelihood function corresponding to each probability density or mass function $f(\bullet; \theta, \gamma)$ on Ω is $L(\bullet) = L(\bullet; x) = f(x; \bullet)$ on $\Theta \times \Gamma$.

The problem of measuring the strength of statistical evidence in the presence of a nuisance parameter has been posed as a problem of approximating the strength of statistical 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 likelihood function with L_{profile}, the profile likelihood function on Θ, defined by $\forall_{\theta \in \Theta} L_{\text{profile}}(\theta) = L_{\text{profile}}(\theta; x) = \sup_{\gamma \in \Gamma} L(\theta, \gamma; x)$. Under the special law of likelihood, the profile likelihood ratio $L_{\text{profile}}\left(\theta'; x\right) / L_{\text{profile}}\left(\theta''; x\right)$ serves as a widely applicable approximation to the strength of statistical 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

$$\mathrm{ev}_{\text{profile}}\left(\Theta', \Theta''\right) = \mathrm{ev}_{\text{profile}}\left(\Theta', \Theta''; x\right) = \frac{\sup_{\theta' \in \Theta'} L_{\text{profile}}\left(\theta'; x\right)}{\sup_{\theta'' \in \Theta''} L_{\text{profile}}\left(\theta''; x\right)}. \tag{5}$$

Example 1 The normal family. *The proposed methodology will be illustrated with the comparison of the hypotheses $|\theta| > \theta_{+}$ and $|\theta| \leq \theta_{+}$ for some $\theta_{+} \geq 0$ on the basis of $x = \left(x^{(1)}, ..., x^{(n)}\right)$, a sample of n independent observations from*

a normal distribution with unknown mean $\theta \in \mathbb{R}$ and variance $\gamma = \sigma^2 \in (0, \infty)$. Hence, the density function satisfies

$$f\left(x; \theta, \sigma^2\right) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{x^{(j)} - \theta}{\sigma}\right)^2\right). \tag{6}$$

Since, as noted in Section 1.2,

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

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

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

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 2.3.1),

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

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, On the probability of observing misleading statistical evidence, 2000b), and, if the nuisance parameter is orthogonal to the interest parameter, it is equal to the likelihood ratio (Royall, Statistical Evidence: A Likelihood Paradigm, 1997a). Instead of seeing the profile likelihood as an approximation, it could be derived from Proposition 2 by framing the nuisance parameter problem as an instance of the composite hypothesis problem. That interpretation of the profile likelihood would be problematic, however, since there are models for which it fails to approximate the strength of statistical evidence unless the sample is sufficiently large (Royall, On the probability of observing misleading statistical evidence, 2000b).

For some models, the nuisance parameter may instead be eliminated by use of a marginal or conditional likelihood (Royall, Statistical Evidence: A Likelihood Paradigm, 1997a) as approximations of the likelihood function without nuisance parameters. Alternatively, provided a prior distribution of γ that is suitable for evidential inference, it could be eliminated by integration. (Methods have been proposed for specifying a nuisance parameter prior to integrate the likelihood not only for Bayesian inference (Kass and Raftery, 1995; Berger *et al.*, 1999; Clyde and George, 2004) but also for Neyman-Pearson inference (Severini, 2007).) While this flexibility in the method for eliminating nuisance parameters allows researchers to optimize performance for particular applications according to their best judgments, it thereby to some extent relaxes the

motivating objectivity condition of Section 1.1. On the other hand, different approaches to eliminating nuisance parameters can yield similar results; for example, likelihoods integrated with respect to certain priors approximate the profile likelihood (Severini, 2007).

3. Evidential 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.

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 *et al.*, 1995) and Bayesian inference (Zadeh, 2002). Similarly, Dollinger *et al.* (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. Additional terminology

Consider Θ, the parameter space of a family of probability density or mass functions that have support on all Ω.

Definition 5 *Any 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; $\widetilde{\Theta}\left(\theta'\right)$ is considered to be the extent to which θ' belongs to a fuzzy subset $\widetilde{\Theta}$ of Θ. Let $\mathcal{F}\left(\Theta\right)$ be the set of all fuzzy subsets of Θ.

For each $\theta' \in \Theta$ and $\lambda' \in [0,1]$, let $\widetilde{\Theta}_{\theta',\lambda'}$ denote the fuzzy subset of Θ such that $\widetilde{\Theta}_{\theta',\lambda'}\left(\theta'\right) = \lambda'$ and $\theta' \neq \theta'' \Rightarrow \widetilde{\Theta}_{\theta',\lambda'}\left(\theta''\right) = 0$. The index λ' may be dropped in the case of full membership: $\widetilde{\Theta}_{\theta'} = \widetilde{\Theta}_{\theta',1}$; this corresponds to the simple hypothesis that $\theta = \theta'$ (Definition 1).

3.2. Evidential theory of imprecise hypotheses

The statements below use the concept of the strength of evidence to extend the $\in$ symbol, operationally defining a hypothesis as a proposition that the parameter value corresponding to the data-generating distribution "is a member of" a fuzzy subset, which mathematically is a function rather than a set. First, the special law of likelihood is restated in the terminology of Section 3.1:

Axiom 3 *For all $\theta' \in \Theta$ and $\theta'' \in \Theta$, the strength of the statistical evidence in $X = x$ that supports $\theta = \theta'$ over $\theta = \theta''$, i.e., for $\theta \in \widetilde{\Theta}_{\theta'}$ over $\theta \in \widetilde{\Theta}_{\theta''}$, is*

$$\mathrm{ev}\left(\widetilde{\Theta}_{\theta'}, \widetilde{\Theta}_{\theta''}\right) = \mathrm{ev}\left(\widetilde{\Theta}_{\theta'}, \widetilde{\Theta}_{\theta''}; x\right) = \frac{L\left(\theta'; x\right)}{L\left(\theta''; x\right)}. \tag{7}$$

To overcome the criticism that fuzzy set theory lacks a way to unambiguously assign membership values other than 0 and 1 (Lindley, 2004), a fractional membership value is calibrated as the strength of evidence associated with two hypotheses that share a probability distribution:

Definition 6 *For all $\theta' \in \Theta$ and $\lambda' \in [0,1]$, the strength of the statistical evidence in $X = x$ that supports $\theta \in \widetilde{\Theta}_{\theta',\lambda'}$ over $\theta = \theta'$ is equal to* the degree to which θ' is a member of $\widetilde{\Theta}_{\theta',\lambda'}$*:*

$$\mathrm{ev}\left(\widetilde{\Theta}_{\theta',\lambda'}, \widetilde{\Theta}_{\theta'}\right) = \mathrm{ev}\left(\widetilde{\Theta}_{\theta',\lambda'}, \widetilde{\Theta}_{\theta'}; x\right) = \lambda'.$$

With that calibration, the special law of likelihood applies to imprecise hypotheses:

Proposition 8 *For all $\theta' \in \Theta$ and $\theta'' \in \Theta$, the strength of the statistical evidence in $X = x$ that supports $\theta \in \widetilde{\Theta}_{\theta'}$ over $\theta \in \widetilde{\Theta}_{\theta''}$ is*

$$\mathrm{ev}\left(\widetilde{\Theta}_{\theta',\lambda'}, \widetilde{\Theta}_{\theta'',\lambda''}\right) = \mathrm{ev}\left(\widetilde{\Theta}_{\theta',\lambda'}, \widetilde{\Theta}_{\theta'',\lambda''}; x\right) = \frac{\lambda' L\left(\theta'; x\right)}{\lambda'' L\left(\theta''; x\right)}$$

under the multiplicative convention that for all fuzzy subsets $\widetilde{\Theta}'$, $\widetilde{\Theta}''$, and $\widetilde{\Theta}'''$ of Θ,

$$\mathrm{ev}\left(\widetilde{\Theta}', \widetilde{\Theta}''; x\right) \mathrm{ev}\left(\widetilde{\Theta}'', \widetilde{\Theta}'''; x\right) = \mathrm{ev}\left(\widetilde{\Theta}', \widetilde{\Theta}'''; x\right).$$

Proof. By the multiplicative convention,

$$\begin{aligned}\operatorname{ev}\left(\widetilde{\Theta}_{\theta',\lambda'},\widetilde{\Theta}_{\theta'',\lambda''}\right) &= \operatorname{ev}\left(\widetilde{\Theta}_{\theta',\lambda'},\widetilde{\Theta}_{\theta'}\right)\operatorname{ev}\left(\widetilde{\Theta}_{\theta'},\widetilde{\Theta}_{\theta''}\right)\operatorname{ev}\left(\widetilde{\Theta}_{\theta''},\widetilde{\Theta}_{\theta'',\lambda''}\right)\\ &= \operatorname{ev}\left(\widetilde{\Theta}_{\theta',\lambda'},\widetilde{\Theta}_{\theta'}\right)\operatorname{ev}\left(\widetilde{\Theta}_{\theta'},\widetilde{\Theta}_{\theta''}\right)\frac{\operatorname{ev}\left(\widetilde{\Theta}_{\theta''},\widetilde{\Theta}_{\theta''}\right)}{\operatorname{ev}\left(\widetilde{\Theta}_{\theta'',\lambda''},\widetilde{\Theta}_{\theta''}\right)}.\end{aligned}$$

Axiom 3 and Definition 6 complete the proof. ■

The general law of likelihood is likewise extended to govern imprecise hypotheses:

Axiom 4 *For all fuzzy subsets $\widetilde{\Theta}'$ and $\widetilde{\Theta}''$ of Θ, the strength of the statistical evidence in $X = x$ that supports $\theta \in \widetilde{\Theta}'$ over $\theta \in \widetilde{\Theta}''$ is $\operatorname{ev}\left(\widetilde{\Theta}',\widetilde{\Theta}'';x\right)$, where* ev *is the function satisfying the following two coherence and non-accumulation conditions as well as equation (7) and*

$$\forall_{\widetilde{\Theta}',\widetilde{\Theta}'',\widetilde{\Theta}'''\in\mathcal{F}(\Theta)}\operatorname{ev}\left(\widetilde{\Theta}',\widetilde{\Theta}'''\right)\leq\operatorname{ev}\left(\widetilde{\Theta}'',\widetilde{\Theta}'''\right)\Leftrightarrow\operatorname{ev}\left(\widetilde{\Theta}''',\widetilde{\Theta}'\right)\geq\operatorname{ev}\left(\widetilde{\Theta}''',\widetilde{\Theta}''\right).$$

Condition 3 Imprecise coherence. *If, for all fuzzy subsets $\widetilde{\Theta}'$ and $\widetilde{\Theta}''$ of Θ, $\operatorname{ev}\left(\widetilde{\Theta}',\widetilde{\Theta}'';x\right)$ is the strength of the statistical evidence in $X = x$ that supports $\theta \in \widetilde{\Theta}'$ over $\theta \in \widetilde{\Theta}''$, then*

$$\forall_{\widetilde{\Theta}'',\widetilde{\Theta}'''\in\mathcal{F}(\Theta)}\forall_{\widetilde{\Theta}'\in\left\{\widetilde{\Theta}''\wedge\widetilde{\Theta}_*:\widetilde{\Theta}_*\in\mathcal{F}(\Theta)\right\}}\operatorname{ev}\left(\widetilde{\Theta}',\widetilde{\Theta}''';x\right)\leq\operatorname{ev}\left(\widetilde{\Theta}'',\widetilde{\Theta}''';x\right).$$

Condition 4 Imprecise non-accumulation. *If, for all fuzzy subsets $\widetilde{\Theta}'$ and $\widetilde{\Theta}''$ of Θ, $\operatorname{ev}\left(\widetilde{\Theta}',\widetilde{\Theta}'';x\right)$ is the strength of the statistical evidence in $X = x$ that supports $\theta \in \widetilde{\Theta}'$ over $\theta \in \widetilde{\Theta}''$, then*

$$\forall_{\widetilde{\Theta}',\widetilde{\Theta}''\in\mathcal{F}(\Theta)}\exists_{\theta'\in\Theta}\operatorname{ev}\left(\widetilde{\Theta}_{\theta'}\wedge\widetilde{\Theta}',\widetilde{\Theta}'';x\right)\geq\operatorname{ev}\left(\widetilde{\Theta}',\widetilde{\Theta}'';x\right).$$

The next two propositions extend equation (2) to the imprecise hypothesis case.

Proposition 9 *For all fuzzy subsets $\widetilde{\Theta}'$ and $\widetilde{\Theta}''$ of Θ, the strength of the statistical evidence in $X = x$ that supports $\theta \in \widetilde{\Theta}'$ over $\theta \in \widetilde{\Theta}''$ is*

$$\operatorname{ev}\left(\widetilde{\Theta}',\widetilde{\Theta}'';x\right)=\sup_{\theta'\in\Theta}\inf_{\theta''\in\Theta}\operatorname{ev}\left(\widetilde{\Theta}'\wedge\widetilde{\Theta}_{\theta'},\widetilde{\Theta}''\wedge\widetilde{\Theta}_{\theta''};x\right).$$

Proof. The proof is analogous to that of equation (2), replacing, for example, $\Theta'_*\cup\left\{\overline{\theta}'\right\}$ for all $\overline{\theta}'\in\Theta\backslash\Theta'_*$ with $\widetilde{\Theta}'_*\vee\widetilde{\Theta}_{\overline{\theta}',\lambda'}$ for all

$$\overline{\theta}'\in\left\{\theta':\theta'\in\Theta,\widetilde{\Theta}'_*\left(\theta'\right)<1\right\}$$

and $\lambda'\in\left(\widetilde{\Theta}'_*\left(\overline{\theta}'\right),1\right]$. ■

Proposition 10 *For all fuzzy subsets $\widetilde{\Theta}'$ and $\widetilde{\Theta}''$ of Θ, the strength of the statistical evidence in $X = x$ that supports $\theta \in \widetilde{\Theta}'$ over $\theta \in \widetilde{\Theta}''$ is*

$$\mathrm{ev}\left(\widetilde{\Theta}', \widetilde{\Theta}''; x\right) = \frac{\sup_{\theta' \in \Theta} \widetilde{\Theta}'\left(\theta'\right) L\left(\theta'; x\right)}{\sup_{\theta'' \in \Theta} \widetilde{\Theta}''\left(\theta''\right) L\left(\theta''; x\right)}. \tag{8}$$

Proof. By Propositions 8 and 9,

$$\begin{aligned} \mathrm{ev}\left(\widetilde{\Theta}', \widetilde{\Theta}''; x\right) &= \sup_{\theta' \in \Theta} \inf_{\theta'' \in \Theta} \mathrm{ev}\left(\widetilde{\Theta}_{\theta', \widetilde{\Theta}'(\theta')}, \widetilde{\Theta}_{\theta'', \widetilde{\Theta}''(\theta'')}; x\right) \\ &= \sup_{\theta' \in \Theta} \inf_{\theta'' \in \Theta} \frac{\widetilde{\Theta}'\left(\theta'\right) L\left(\theta'; x\right)}{\widetilde{\Theta}''\left(\theta''\right) L\left(\theta''; x\right)}. \end{aligned}$$

■

4. Simulation study

To quantify the impact of replacing a simple hypothesis with a small-interval composite hypotheses in evidential inference, a series of simulations were carried out for the case of normal distributions (Example 1). $M = 10^5$ 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 $x_1, ..., x_M$, 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

$$\widehat{\alpha}_n^{\Theta''}(b) = \frac{1}{M} \sum_{i=1}^{M} I_{[10^b, \infty)}\left(\mathrm{ev}_{\mathrm{profile}}\left(\mathbb{R} \backslash \Theta'', \Theta''; x_i\right)\right) \tag{9}$$

with $\Theta'' = \{0\}$ for the composite-*simple* hypothesis pair or with $\Theta'' = [-1/10, 1/10]$ for the composite-*composite* hypotheses pair; $\varepsilon \in \mathcal{S} \Rightarrow I_{\mathcal{S}}(\varepsilon) = 1$ and $\varepsilon \notin \mathcal{S} \Rightarrow I_{\mathcal{S}}(\varepsilon) = 0$. 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 5, but not for $\Theta'' = \{0\}$, with the exception of smaller samples.

Figure 6 focuses on the comparison between and approximate evidence 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

In this section, the strength of evidence is compared to the approximate strength of evidence in 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 equation (6), replacing θ with θ_i, the logarithm of 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.

As in the simulation study of the last section, equation (5) 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), and Bochkina and Richardson (2007) redefined what is meant by "differential expression" by employing some biologically relevant value $\theta_+ > 0$. Accordingly, equation (5) 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 $\log 2$. (Lewin *et al.* (2006) and Bochkina and Richardson (2007) similarly derived posterior probabilities that $|\theta_i| > \log 2$, and Bickel

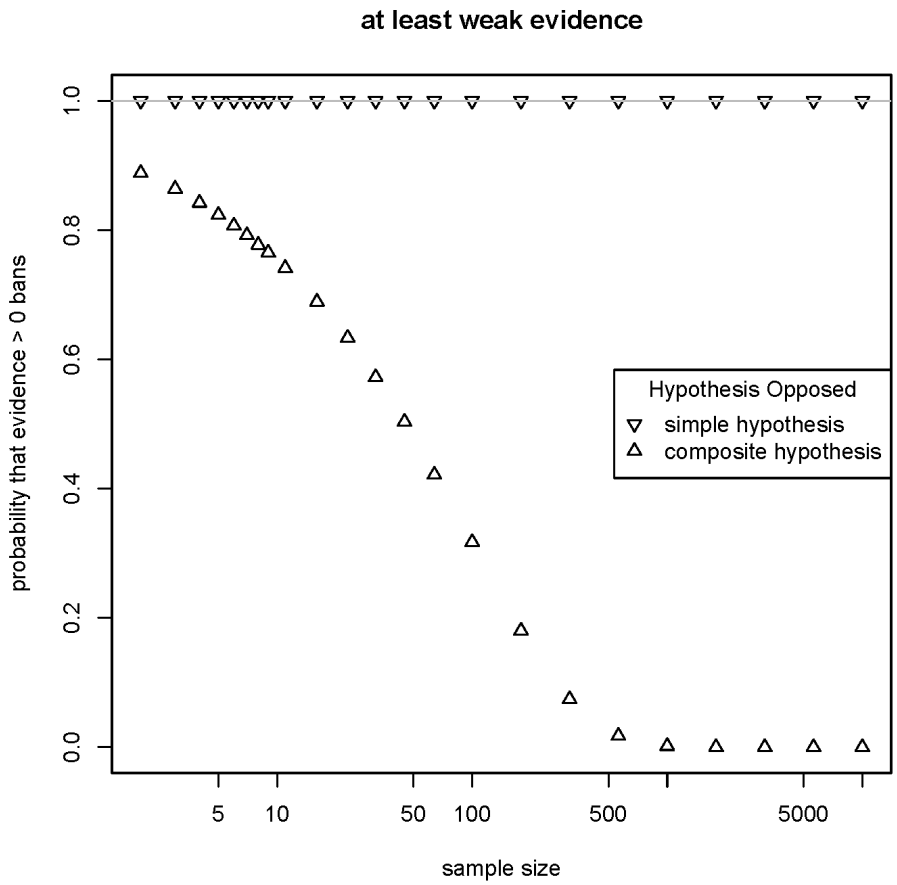

FIG 1. *Probabilities* $\widehat{\alpha}_n^{\{0\}}(1/\infty)$ *and* $\widehat{\alpha}_n^{[-1/10,1/10]}(1/\infty)$ *of observing any misleading* ***positive*** *evidence for the hypothesis that* $\theta \neq 0$ *over the "simple" hypothesis that* $\theta = 0$ *and for the hypothesis that* $|\theta| >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

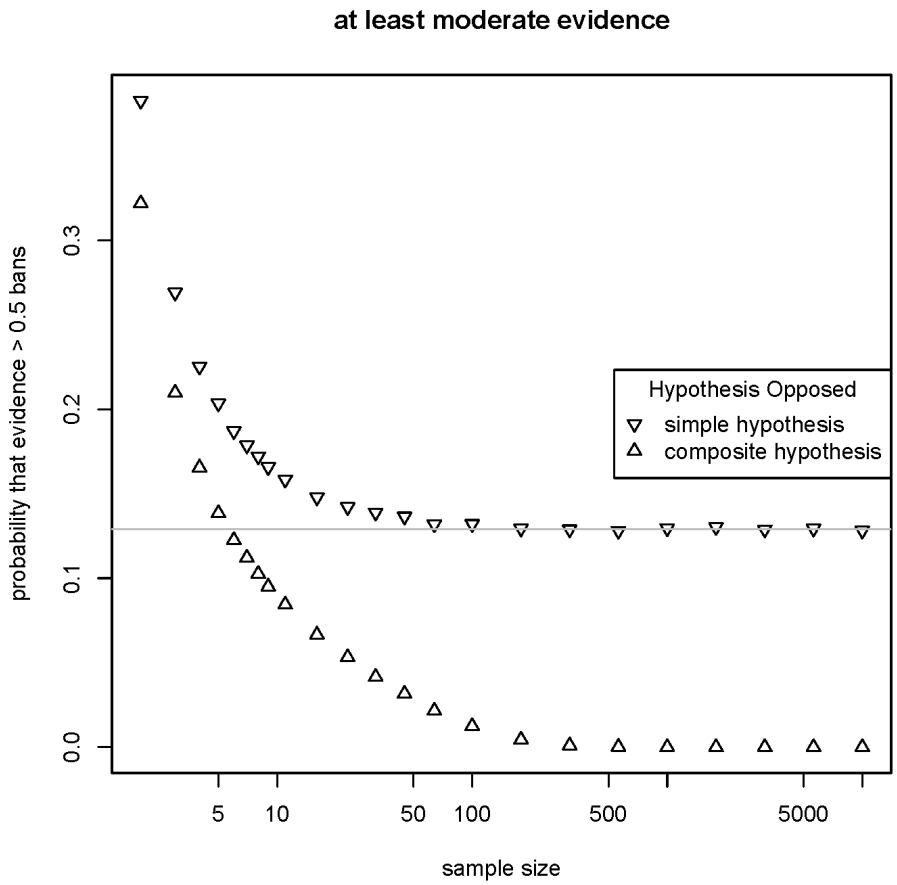

FIG 2. *Probabilities* $\widehat{\alpha}_n^{\{0\}}(1/2)$ *and* $\widehat{\alpha}_n^{[-1/10,1/10]}(1/2)$ *of observing misleading* ***moderate or stronger*** *evidence for the hypothesis that* $\theta \neq 0$ *over the "simple" hypothesis that* $\theta = 0$ *and for the hypothesis that* $|\theta| > 1/10$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

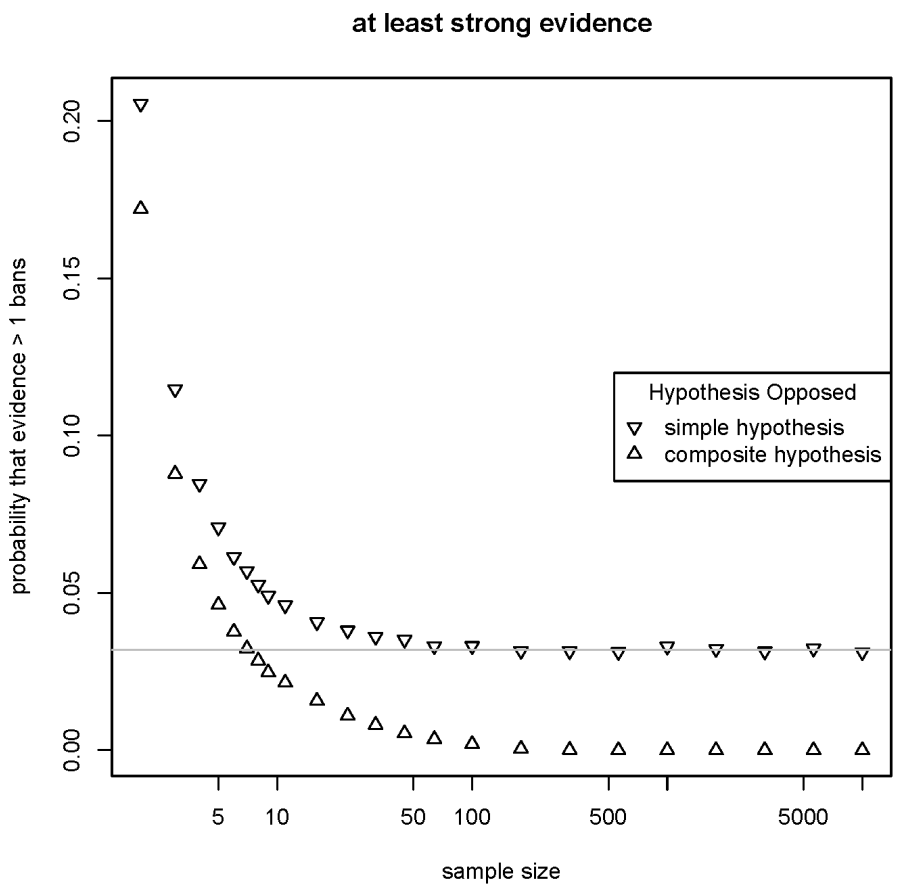

FIG 3. *Probabilities* $\widehat{\alpha}_n^{\{0\}}(1)$ *and* $\widehat{\alpha}_n^{[-1/10,1/10]}(1)$ *of observing misleading* ***strong, very strong, or decisive*** *evidence for the hypothesis that* $\theta \neq 0$ *over the "simple" hypothesis that* $\theta = 0$ *and for the hypothesis that* $|\theta|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

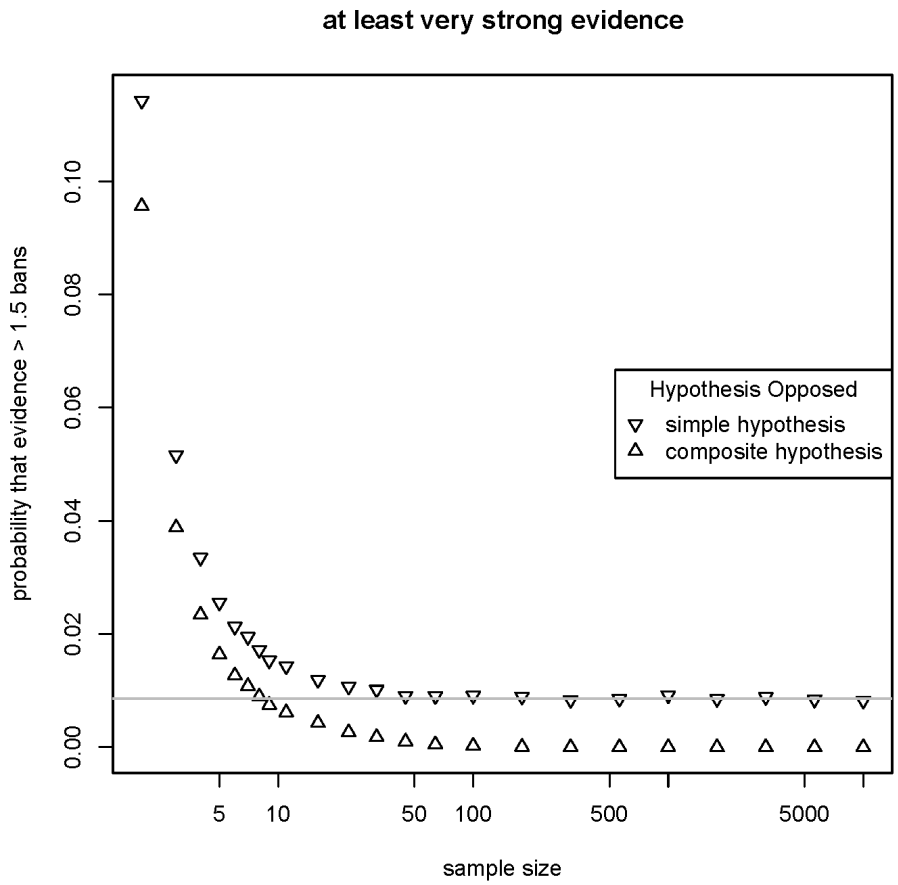

FIG 4. *Probabilities* $\widehat{\alpha}_n^{\{0\}}(3/2)$ *and* $\widehat{\alpha}_n^{[-1/10,1/10]}(3/2)$ *of observing misleading* ***very strong or decisive*** *evidence for the hypothesis that* $\theta \neq 0$ *over the "simple" hypothesis that* $\theta = 0$ *and for the hypothesis that* $|\theta| >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

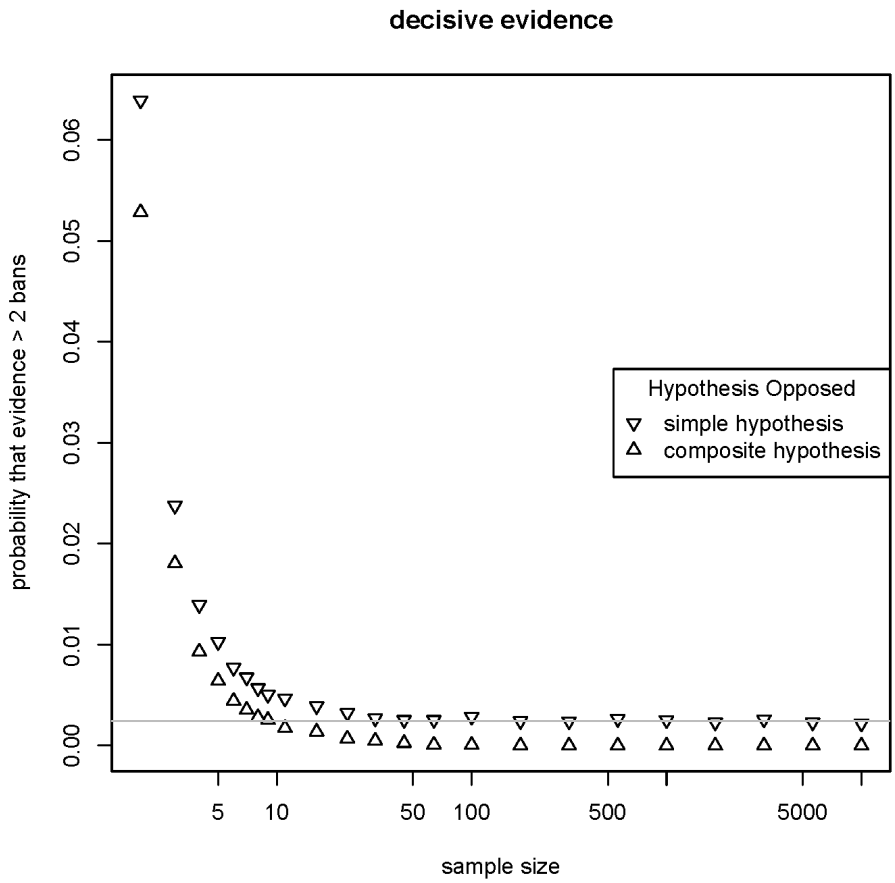

FIG 5. *Probabilities* $\widehat{\alpha}_n^{\{0\}}$ (2) *and* $\widehat{\alpha}_n^{[-1/10,1/10]}$ (2) *of observing misleading* ***decisive*** *evidence for the hypothesis that* $\theta \neq 0$ *over the "simple" hypothesis that* $\theta = 0$ *and for the hypothesis that* $|\theta| > 1/10$ *over the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

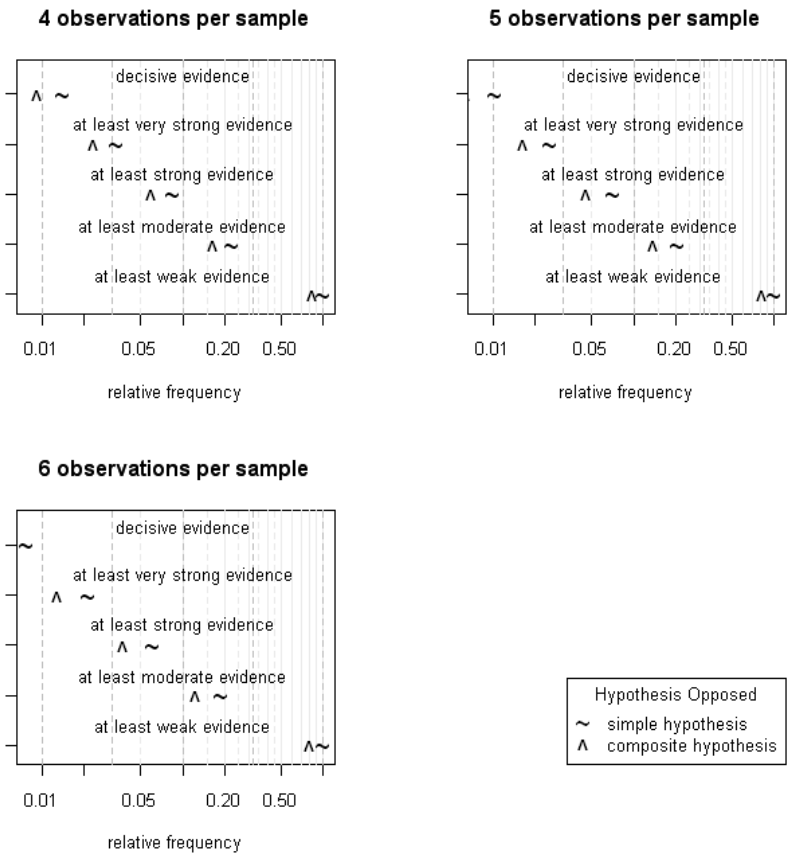

FIG 6. *Probabilities* $\widehat{\alpha}_n^{\{0\}}(b)$ *and* $\widehat{\alpha}_n^{[-1/10,1/10]}(b)$ *of observing misleading evidence for the hypothesis that* $\theta \neq 0$ *over the "simple" hypothesis that* $\theta = 0$ *and for the hypothesis that* $|\theta| > 1/10$ *over the "composite" hypothesis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

(2004) and Van De Wiel and Kim (2007) considered false discovery rates for which a "discovery" is defined in terms of a fold change threshold.)

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 $\theta_i \in \Theta' = \mathbb{R}\backslash\left[-\log\sqrt{2}, \log\sqrt{2}\right]$ 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 $\mathrm{ev}_{\mathrm{profile}}\left(\mathbb{R}\backslash\{0\}, \{0\}; x_i\right)$ than when using $\mathrm{ev}_{\mathrm{profile}}\left(\Theta', \Theta''; x_i\right)$ 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 are sensitive to the specification of θ_+, the value that determines the sharp boundary between equivalent expression $(|\theta_i| \leq \theta_+)$ 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

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

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 a two-fold or greater difference 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 $\mathrm{ev}_{\mathrm{profile}}\left(\widetilde{\Theta}', \widetilde{\Theta}''; x_i\right)$, the result of putting the profile likelihood function in place of likelihood function in equation (8), 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 $\mathrm{ev}_{\mathrm{profile}}\left(\Theta', \Theta''; x_i\right)$ with $\mathrm{ev}_{\mathrm{profile}}\left(\widetilde{\Theta}', \widetilde{\Theta}''; x_i\right)$ 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 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.

6. Discussion

6.1. Multiple comparisons

Since it is often maintained that the thousands of comparisons made in the analysis of microarray data call for different methods of analysis than those used for single comparisons, it may seem that the control of a false discovery rate or other adjusted Type I error rate may be more appropriate than an evidential analysis such as that described here. Indeed, Korn and Freidlin (2006) voiced concerns about the direct application of the law of likelihood to the multiple comparisons situation. Because the 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, Likelihood methods for measuring statistical evidence, 2002a).

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 *et al.*, 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 offerred 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 *et al.*, 2004).

Considering the p-value as a measure of statistical evidence that must be adjusted to continue to measure statistical evidence under multiple comparisons

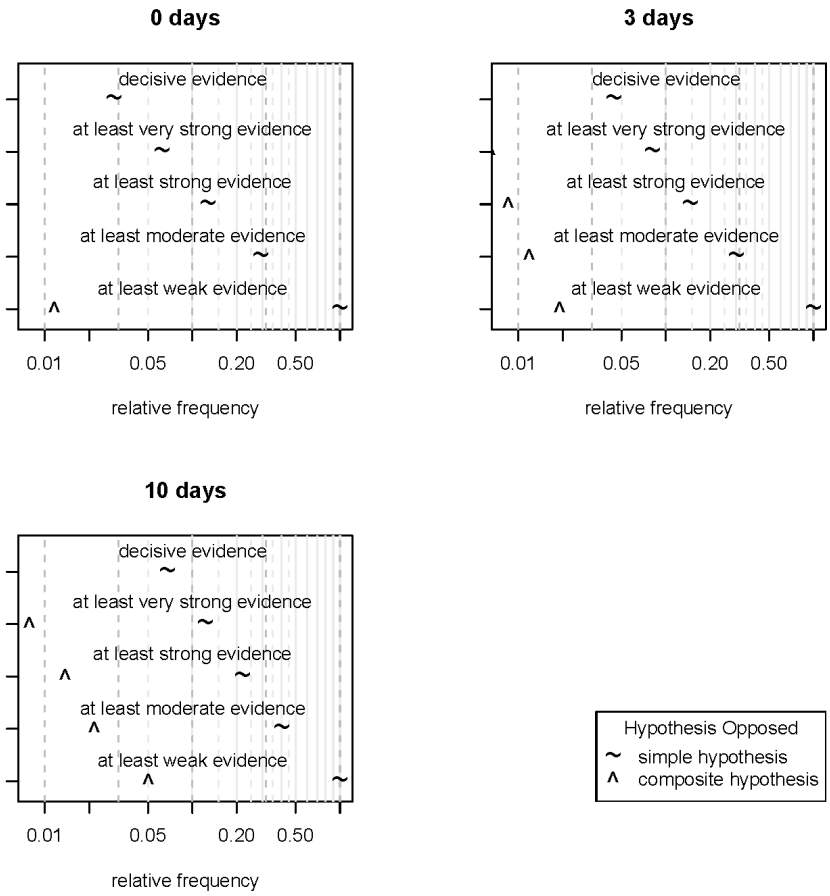

FIG 7. *Probabilities* $\widehat{\alpha}_n^{\{0\}}(b)$ *and* $\widehat{\alpha}_n^{[-\log\sqrt{2},\log\sqrt{2}]}(b)$ *of observing misleading evidence for the hypothesis that* $\theta_i \neq 0$ *over the "simple" hypothesis that* $\theta_i = 0$ *and for the hypothesis that* $|\theta_i| > \log\sqrt{2}$ *over the "composite" hypothesis that* $|\theta_i| \leq \log\sqrt{2}$*, respectively, for each of the evidence levels of Figures 1-5* ($b \in \{1/\infty, 1/2, 1, 3/2, 2\}$) *and at each of three stages of maturity (0, 3, and 10 days after the breaker stage of ripening). These proportions were computed using equation (9), 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.*

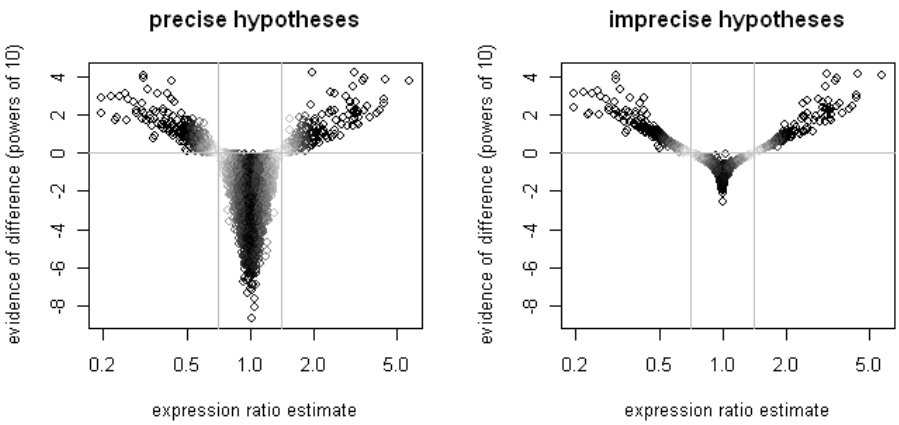

FIG 8. *Strength of statistical evidence for differential expression over equivalent expression plotted against the maximum likelihood estimate of the expression ratio for the tomato data at 10 days after the breaker stage of ripening. The vertical gray lines are drawn at the boundary that separates 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* $\left|2\widetilde{\Theta}'\left(\widehat{\theta}_i\right)-1\right|$ *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.* $\widetilde{\Theta}'\left(\widehat{\theta}_i\right)$ *estimates* $\widetilde{\Theta}'(\theta_i)$*, the degree to which the ith gene is differentially expressed.*

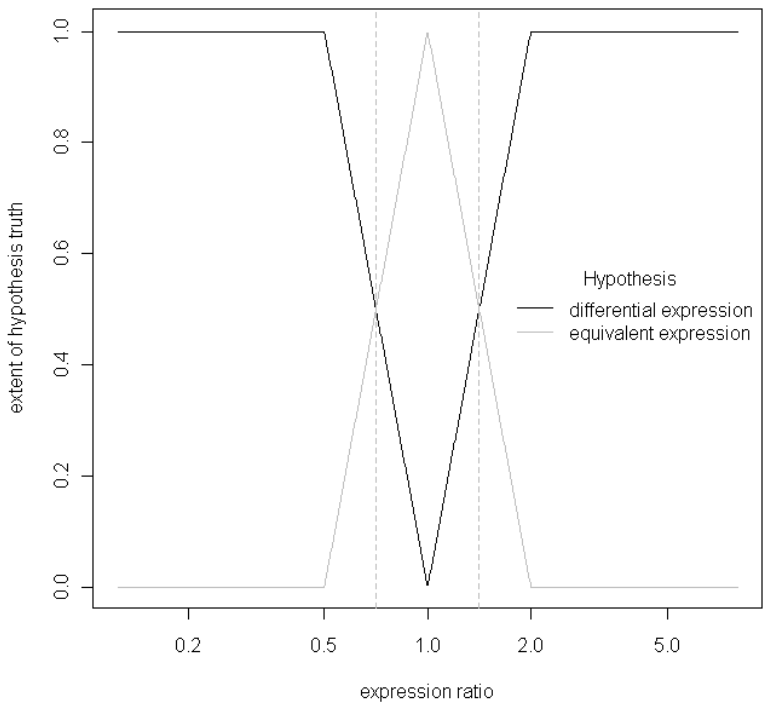

FIG 9. *The degree of the truth of each imprecise hypothesis plotted against* e^{θ}, *the geometric mean of the expression ratio in the population. The black curve represents* $\widetilde{\Theta}'$, *and the gray curve represents* $\widetilde{\Theta}''$. *The vertical lines correspond to the boundary between the precise hypotheses* Θ' *and* Θ''. *Degrees of truth are calibrated by Definition 6.*

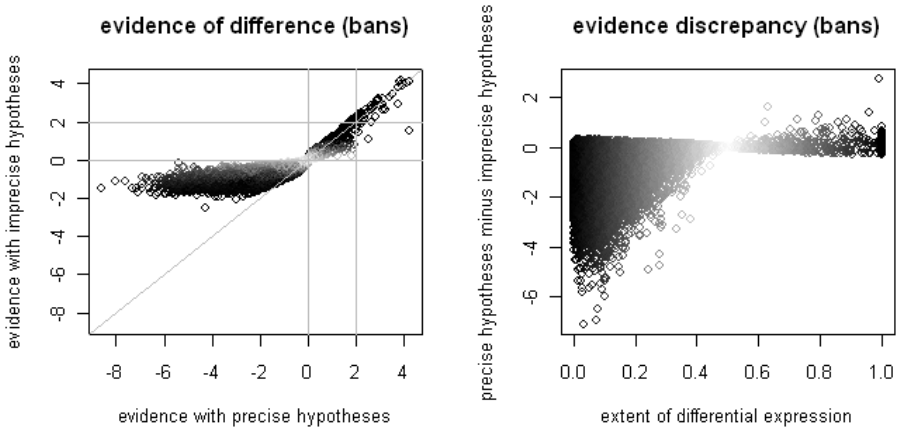

FIG 10. *Effects of replacing the precise hypotheses with the imprecise hypotheses for the data of Figure 8. The left-hand-side displays* $\mathrm{ev}_{\mathrm{profile}}\left(\widetilde{\Theta}', \widetilde{\Theta}''; x_i\right)$ *plotted against* $\mathrm{ev}_{\mathrm{profile}}\left(\Theta', \Theta'';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

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 happenned 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 (Shaffer, 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 *et al.*, 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. It is worth emphasizing that the error-control rationale for adjusting p-values is distinct from the rationale behind empirical Bayes and hierarchical Bayesian 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.

By contrast, 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 likelihood ratio better measures the strength of statistical evidence with respect to one of several comparisons. An analogy with legal evidence 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?

Evidential inference based directly on the law of likelihood is only beginning to find applications in extreme multiple comparison situations. Taking an important 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 argue that although consideration of error rates is important for study design, their use in correcting p-values interpreted evidentially distorts the strength of evidence. Evidential inference may also play an important role in the analysis of gene expression data, especially in light of the unexpected finding that with microarrays the control of the false discovery rate can yield less reproducible results that does a simple method unencumbered by multiple comparison procedures (Guo *et al.*, 2006).

6.2. Further development

As axioms, the laws of likelihood are not derived from more primitive assumptions but invite 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 2.3.1 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.

There remains ample opportunity for research improving the measurement of how much evidence favors composite hypotheses of interest. Methods that make the special law of likelihood less sensitive to model misspecification (Tsou

and Royall, 1995; Blume *et al.*, 2007) may be adapted to robust inference under the general law of likelihood. For large samples, nonparametric methods such as those of empirical likelihood and bootstrap likelihood might enable even more objective measurement of the strength of evidence.

7. Acknowledgments

I am grateful to Jeffrey Blume for a thought-provoking discussion on evidence and multiple comparisons. I also thank Xuemei Tang for providing the fruit development microarray data. This work was partially supported by Canada Foundation for Innovation, Ontario's Ministry of Research and Innovation, and the Faculty of Medicine of the University of Ottawa. Computations were performed in R (R Development Core Team, 2008), and the *Biobase* package of Bioconductor (Gentleman *et al.*, 2004) facilitated data management.

References

Alba, R., Payton, P., Fei, Z., McQuinn, R., Debbie, P., Martin, G. B., Tanksley, S. D. and Giovannoni, J. J., 'Transcriptome and selected metabolite analyses reveal multiple points of ethylene control during tomato fruit development', *Plant Cell*, vol. 17, no. 11, 2954–2965 (2005).

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

Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N. and Golani, I., 'Controlling the false discovery rate in behavior genetics research', *Behavioural brain research*, vol. 125, no. 1-2, 279–284 (2001).

Benjamini, Y. and Hochberg, Y., 'On the adaptive control of the false discovery rate in multiple testing with independent statistics', *Journal of Educational and Behavioral Statistics*, vol. 25, no. 1, 60–83 (2000).

Benjamini, Y. and Liu, W., 'A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence', *Journal of Statistical Planning and Inference*, vol. 82, no. 1-2, 163–170 (1999).

Benjamini, Y. and Yekutieli, D., 'Quantitative trait loci analysis using the false discovery rate', *Genetics*, vol. 171, no. 2, 783–790 (2005).

Berger, J. O., 'The case for objective Bayesian analysis', *Bayesian Analysis*, vol. 1, 1–17 (2004).

Berger, J. O., Liseo, B. and Wolpert, R. L., 'Integrated Likelihood Methods for Eliminating Nuisance Parameters', *Statistical Science*, vol. 14, no. 1, 1–28 (1999).

Bernardo, J. M., 'Noninformative priors do not exist: A discussion', *Journal of Statistical Planning and Inference*, vol. 65, 159–189 (1997).

Bickel, D. R., 'Degrees of differential gene expression: Detecting biologically significant expression differences and estimating their magnitudes', *Bioinformatics (Oxford, England)*, vol. 20, 682–688 (2004).

Blume, J. D., 'Likelihood methods for measuring statistical evidence', *Statistics in Medicine*, vol. 21, no. 17, 2563–2599 (2002).

Blume, J. D., 'How often likelihood ratios are misleading in sequential trials', *Communications in Statistics - Theory and Methods*, vol. 37, no. 8, 1193–1206 (2008).

Blume, J. D., Su, L., Olveda, R. M. and McGarvey, S. T., 'Statistical evidence for GLM regression parameters: A robust likelihood approach', *Statistics in Medicine*, vol. 26, no. 15, 2919–2936 (2007).

Bochkina, N. and Richardson, S., 'Tail posterior probability for inference in pairwise and multiclass gene expression data', *Biometrics*, vol. 63, no. 4, 1117–1125 (2007).

Choi, L., Caffo, B. S. and Rohde, C., 'A survey of the likelihood approach to bioequivalence trials', *Johns Hopkins University, Department of Biostatistics Working Papers*, vol. 134 (2007).

Clyde, M. and George, E. I., 'Model uncertainty', *Statistical Science*, vol. 19, no. 1, 81–94 (2004).

Cox, D. R., 'The role of significance tests', *Scandinavian Journal of Statistics*, vol. 4, 49–70 (1977).

Dollinger, M. B., Kulinskaya, E. and Staudte, R. G., *Information, Statistics and Induction in Science*, chap. Fuzzy hypothesis tests and confidence intervals, pp. 119–128 (Singapore: World Scientific, 1996).

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

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

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

Fraser, D. A. S., Reid, N. and Wong, A. C. M., 'Inference for bounded parameters', *Physical Review D*, vol. 69, no. 3 (2004).

Gentleman, Robert C, Carey, Vincent J., Bates, Douglas M., Bolstad, Ben, Dettling, Marcel, Dudoit, Sandrine, Ellis, Byron, Gautier, Laurent, Ge, Yongchao, Gentry, Jeff, Hornik, Kurt, Hothorn, Torsten, Huber, Wolfgang, Iacus, Stefano, Irizarry, Rafael, Leisch, Friedrich, Li, Cheng, Maechler, Martin, Rossini, Anthony J., Sawitzki, Gunther, Smith, Colin, Smyth, Gordon, Tierney, Luke, Yang, Jean Y. H. and Zhang, Jianhua, 'Bioconductor: Open software development for computational biology and bioinformatics', *Genome Biology*, vol. 5, R80 (2004).
URL: *http://genomebiology.com/2004/5/10/R80*

Geyer, C. J. and Meeden, G. D., 'Fuzzy and randomized confidence intervals and P-values', *Statistical Science*, vol. 20, no. 4, 358–366 (2005).

Goodman, S. N., 'Multiple comparisons, explained', *American Journal of Epidemiology*, vol. 147, no. 9, 807–812 (1998).

Goodman, S. N. and Royall, R., 'Evidence and scientific research', *American Journal of Public Health*, vol. 78, no. 12, 1568–1574 (1988).

Guo, Lei, Lobenhofer, Edward K., Wang, Charles, Shippy, Richard, Harris, Stephen C., Zhang, Lu, Mei, Nan, Chen, Tao, Herman, Damir, Goodsaid, Federico M., Hurban, Patrick, Phillips, Kenneth L., Xu, Jun, Deng, Xutao, Sun, Yongming Andrew, Tong, Weida, Dragan, Yvonne P. and Shi, Leming, 'Rat toxicogenomic study reveals analytical consistency across microarray platforms', *Nat Biotech*, vol. 24, no. 9, 1162–1169 (2006).

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

Hoch, J. S. and Blume, J. D., 'Measuring and illustrating statistical evidence in a cost-effectiveness analysis', *Journal of Health Economics*, vol. 27, no. 2, 476–495 (2008).

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

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

Kalbfleisch, J. D., 'Comment on R. Royall, 'On the probability of observing misleading statistical evidence", *Journal of the American Statistical Association*, vol. 95, no. 451, 770–771 (2000).

Kass, R. E. and Raftery, A. E., 'Bayes factors', *Journal of the American Statistical Association*, vol. 90, no. 430, 773–795 (1995).

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

Klir, G. J., 'Generalized information theory: Aims, results, and open problems', *Reliability Engineering and System Safety*, vol. 85, no. 1-3, 21–38 (2004).

Korn, E. L. and Freidlin, B., 'The likelihood as statistical evidence in multiple comparisons in clinical trials: No free lunch', *Biometrical Journal*, vol. 48, no. 3, 346–355 (2006).

Van der Laan, M. J., Dudoit, S. and Pollard, K. S., '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.*, vol. 3, 15 (2004).

Lavine, M. and Schervish, M. J., 'Bayes factors: What they are and what they are not', *American Statistician*, vol. 53, no. 2, 119–122 (1999).

Laviolette, M., 'Comment on N. D. Singpurwalla and J. M. Booker, 'Membership functions and probability measures of fuzzy sets", *Journal of the American Statistical Association*, vol. 99, no. 467, 879–880 (2004).

Lewin, A., Richardson, S., Marshall, C., Glazier, A. and Aitman, T., 'Bayesian modeling of differential gene expression', *Biometrics*, vol. 62, no. 1, 1–9 (2006).

Lindley, D. V., 'Comment on N.D. Singpurwalla and J.M. Booker, 'Membership functions and probability measures of fuzzy sets", *Journal of the American Statistical Association*, vol. 99, no. 467, 877–879 (2004).

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

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

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

Pasterkamp, R. J., Peschon, J. J., Spriggs, M. K. and Kolodkin, A. L., 'Semaphorin 7A promotes axon outgrowth through integrins and MAPKs', *Nature*, vol. 424, no. 6947, 398–405 (2003).

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

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

Romer, C., Kandel, A. and Backer, E., 'Fuzzy partitions of the sample space and fuzzy parameter hypotheses', *IEEE Transactions on Systems, Man and Cybernetics*, vol. 25, no. 9, 1314–1322 (1995).

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

Royall, R., 'On the probability of observing misleading statistical evidence', *Journal of the American Statistical Association*, vol. 95, no. 451, 760–768 (2000).

Royall, R., 'Rejoinder to comments on R. Royall, 'On the probability of observing misleading statistical evidence", *Journal of the American Statistical Association*, vol. 95, no. 451, 773–780 (2000).

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

Schervish, M. J., 'P Values: What They Are and What They Are Not', *American Statistician*, vol. 50, no. 3, 203–206 (1996).

Severini, Thomas A., 'Integrated likelihood functions for non-Bayesian inference', *Biometrika*, vol. 94, no. 3, 529–542 (2007).

Shaffer, J. P., 'Multiple hypothesis testing', *Annual Review of Psychology*, vol. 46, no. 1, 561–584 (1995).

Singpurwalla, N. D. and Booker, J. M., 'Membership functions and probability measures of fuzzy sets', *Journal of the American Statistical Association*, vol. 99, no. 467, 867–877 (2004).

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

Spicer, L. J. and Francisco, C. C., 'The adipose obese gene product, leptin: Evidence of a direct inhibitory role in ovarian function', *Endocrinology*, vol. 138, no. 8, 3374–3379 (1997).

Spjøtvoll, E., 'Comment on D. R. Cox, 'The role of significance tests", *Scandinavian Journal of Statistics*, vol. 4, 63–66 (1977).

Storey, J. D., 'A direct approach to false discovery rates', *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, vol. 64, no. 3, 479–498 (2002).

Strug, L. J. and Hodge, S. E., 'An alternative foundation for the planning and evaluation of linkage analysis. II. Implications for multiple test adjustments', *Human Heredity*, vol. 61, no. 4, 200–209 (2006).

Strug, L. J., Rohde, C. A. and Corey, P. N., 'An introduction to evidential sample size calculations', *American Statistician*, vol. 61, no. 3, 207–212 (2007).

Tsou, T.-S. and Royall, R., 'Robust likelihoods', *Journal of the American Statistical Association*, vol. 90, 316–320 (1995).

Van De Wiel, M. A. and Kim, K. In, 'Estimating the false discovery rate using nonparametric deconvolution', *Biometrics*, vol. 63, no. 3, 806–815 (2007).

Wright, S. P., 'Adjusted P-values for simultaneous inference', *Biometrics*, vol. 48, no. 4, 1005–1013 (1992).

Yekutieli, D., Reiner-Benaim, A., Benjamini, Y., Elmer, G. I., Kafkafi, N., Letwin, N. E. and Lee, N. H., 'Approaches to multiplicity issues in complex research in microarray analysis', *Statistica Neerlandica*, vol. 60, no. 4, 414–437 (2006).

Zadeh, L. A., 'Toward a perception-based theory of probabilistic reasoning with imprecise probabilities', *Journal of Statistical Planning and Inference*, vol. 105, no. 1, 233–264 (2002).