Derivation of Bayesian and frequentist inference from evidentiary first principles with applications to propagating uncertainty about statistical methods

David R. Bickel

Derivation of Bayesian and frequentist inference from evidentiary first principles with applications to propagating uncertainty about statistical methods

July 6, 2022

David R. Bickel

Informatics and Analytics

University of North Carolina at Greensboro

The Graduate School

241 Mossman Building, CAMPUS

Greensboro, NC 27402-6170

dbickel@uncg.edu

Abstract

Unquantified uncertainty about probabilistic model assumptions tends to inflate claims of statistical significance, potentially leading to reported results that cannot be replicated. The same bias occurs at a higher level when Bayesian inference or frequentist inference is chosen to achieve significance in the absence of a way to propagate the uncertainty about which statistical paradigm to select.

The objective of this article is to correct that bias at both levels on the basis of evidentiary first principles for statistical inference. It is found that just as Bayesian inference is warranted when a prior distribution is considered alongside the data as evidence, frequentist inference in the form of confidence intervals and their generalization to confidence distributions is warranted when a hypothesis testing procedure is considered as a piece of evidence. Hierarchical evidence in the same framework enables reporting results reflecting uncertainty about which of those pieces of evidence to admit as well as uncertainty about model assumptions. That is illustrated by a method of averaging Bayesian hypothesis testing with frequentist hypothesis testing and by a method of averaging confidence intervals and/or credible intervals.

Keywords: approximate confidence distribution; fiducial inference; foundations of statistics; general principle of maximum entropy; p-hacking; replication crisis

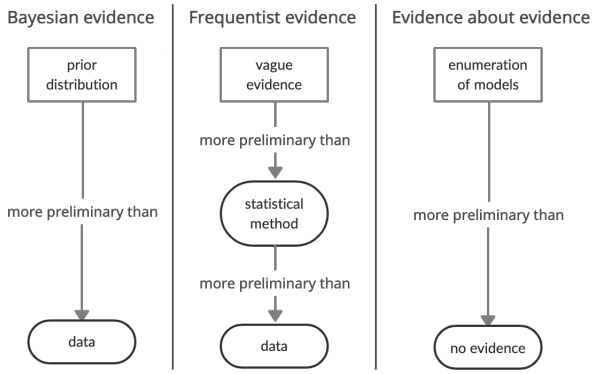

Figure 1: Each of the first two flowcharts is a different possibility of the body of total evidence admitted for reporting scientific results. The third flowchart is a higher-level body of total evidence for propagating uncertainty about the first two bodies of total evidence to the results. Each rectangle at the top of a flowchart abbreviates a different possibility for background evidence. The other pieces of evidence are less preliminary within the flowchart.

1 Why unify statistical inference?

The war between the Bayesian and frequentist camps of statistical inference continues [32]. The current cease-fire hides unresolved controversies that inflict unintended damage on the sciences. For example, in the special issue of *The American Statistician* on the replication crisis in science [43], the proposed solutions depend on various stances toward the Bayesian and frequentist positions. At two extremes of the associated debate on hypothesis testing, some Bayesians argue for calibrating p-values using posterior probabilities or related quantities, whereas some frequentists argue for the continued use of uncalibrated p-values.

The lack of a consensus among statisticians leaves scientists without clear guidance. Meanwhile, science must carry on. Results and conclusions must be reported, ideally with their uncertainty quantified. But Bayesian and frequentist methods of uncertainty quantification can yield very different results. One approach is to use the statistical paradigm and model that make the conclusions appear the most certain or statistically significant. That is practiced in molecular evolution whenever a Bayesian method is chosen over a frequentist method in order to report a more certain phylogenetic tree [15, p. 431]. The general form of that practice, selecting models and methods of inference to maximize statistical significance, has been condemned as p-hacking and blamed for the inability to replicate research results [1].

This article proposes a solution in the form of evidentiary first principles for statistical inference; the main special cases are summarized in Figure 1. That results in generally applicable methods of propagating uncertainty about Bayesian and/or frequentist models.

2 General theory: Proof distributions from pieces of evidence

Let Pr be the joint probability distribution of an observable data set denoted by Data and an unknown quantity or vector of scientific interest denoted by U. The *body of total evidence* is a sequence of K pieces of evidence, where each *piece of evidence* is a proposition about Pr. Each piece of evidence is considered more preliminary than the ones after it in the sequence. For example, if a Bayesian prior distribution B is considered more preliminary than an observed data set denoted by data, then the body of total evidence is this sequence of two pieces of evidence:

1. The marginal distribution of U is B, a known prior distribution.

2. The probability that Data = data is 100%.

As a proposition about Pr, each piece of evidence may be represented as a subset of the set of all possible joint probability distributions of Data and U. Then let Ev_k be the set representing the kth piece of evidence for $k = 1, \ldots, K$.

Ev_1, being the most preliminary piece of evidence, is called the *background evidence*. Each member of the background evidence may be interpreted as practical information encoded in a language [45, chapter 9]. In terms of measure theory, Ev_1 is a set of measures that are not necessarily probability measures. For example, if the marginal distribution of U is uniform on the real line, then it does not integrate to 1. By contrast, Ev_k must be a set of probability measures for each $k = 2, \ldots, K$. All measures have the same domain (measure space) for a given application.

Putting the example in the subset representation, the body of total evidence is the ordered pair $(\mathrm{Ev}_1, \mathrm{Ev}_2)$, where Ev_1 is the set of all joint distributions such that the marginal distribution of U is B, and Ev_2 is the set of all joint distributions such that the probability that Data = data is 100%. In short,

$$\mathrm{Ev}_1 = \{\mathrm{Pr} : \mathrm{Pr}\,(U \in \bullet) = B\}\,; \tag{1}$$

$$\mathrm{Ev}_2 = \{\mathrm{Pr} : \mathrm{Pr}\,(\mathrm{Data} = \mathrm{data}) = 1\}\,. \tag{2}$$

Two or more pieces of evidence are inconsistent with each other if their intersection is the empty set. For the sake of enough generality to recover frequentist inference as a special case in Section 3, it is not required that the pieces of evidence be consistent with each other. What is required is that each piece of evidence influence the marginal distribution of U to be used for scientific inference as much as possible while respecting which pieces of evidence are more preliminary than others. That requirement is made precise as follows.

The goal for moving from Ev_1, the most preliminary piece of evidence, to Ev_2, the next most preliminary piece of evidence, is to identify the distributions in Ev_2 that would not attribute more certainty than warranted by the evidence $(\mathrm{Ev}_1, \mathrm{Ev}_2)$. That is accomplished by finding the distributions Pr_2 in Ev_2 and Pr_1 in Ev_1 that maximize the differential entropy of $\rho_{1,2}$, the probability density function of Pr_2 with respect to Pr_1, assuming that the density exists. The density exists if $\mathrm{Pr}_1 \gg \mathrm{Pr}_2$, which means Pr_1 dominates Pr_2. That differential entropy is

$$S\left(\mathrm{Pr}_2 \,||\, \mathrm{Pr}_1\right) = \begin{cases} -\int \rho_{1,2}\left(u, \mathrm{data}\right) \log \rho_{1,2}\left(u, \mathrm{data}\right) d\,\mathrm{Pr}_1\left(u, \mathrm{data}\right) & \text{if } \mathrm{Pr}_1 \gg \mathrm{Pr}_2 \\ -\infty & \text{if } \mathrm{Pr}_1 \not\gg \mathrm{Pr}_2 \end{cases}, \tag{3}$$

which is the additive inverse of the relative entropy or Kullback-Leibler divergence between Pr_2 and Pr_1. The set of those entropy-maximizing distributions in Ev_2 is symbolized by $\mathrm{Ev}_1\,\mathrm{Ev}_2$. They are called the *distributions of proof* since their probabilities may be interpreted as degrees of proof on the basis of the evidence $(\mathrm{Ev}_1, \mathrm{Ev}_2)$. More concisely,

$$\mathrm{Ev}_1\,\mathrm{Ev}_2 = \arg \sup_{\mathrm{Pr}_2 \in \mathrm{Ev}_2} \sup_{\mathrm{Pr}_1 \in \mathrm{Ev}_1} S\left(\mathrm{Pr}_2 \,||\, \mathrm{Pr}_1\right), \tag{4}$$

which is short for

$$\mathrm{Ev}_1\,\mathrm{Ev}_2 = \left\{\mathrm{Pr}_2' \in \mathrm{Ev}_2 : \sup\left\{S\left(\mathrm{Pr}_2' \,||\, \mathrm{Pr}_1\right) : \mathrm{Pr}_1 \in \mathrm{Ev}_1\right\} = \sup\left\{S\left(\mathrm{Pr}_2 \,||\, \mathrm{Pr}_1\right) : \mathrm{Pr}_2 \in \mathrm{Ev}_2, \mathrm{Pr}_1 \in \mathrm{Ev}_1\right\}\right\},$$

where sup is the least upper bound.

Similarly, the goal for moving from $\mathrm{Ev}_1\,\mathrm{Ev}_2$, the set of the distributions of proof given $(\mathrm{Ev}_1, \mathrm{Ev}_2)$, to Ev_3, the next piece of evidence, is to identify $\mathrm{Ev}_1\,\mathrm{Ev}_2\,\mathrm{Ev}_3$, the set of the distributions of proof given $(\mathrm{Ev}_1, \mathrm{Ev}_2, \mathrm{Ev}_3)$. In analogy with equation (4), that is

$$\mathrm{Ev}_1\,\mathrm{Ev}_2\,\mathrm{Ev}_3 = \arg \sup_{\mathrm{Pr}_3 \in \mathrm{Ev}_3} \sup_{\mathrm{Pr}_{1,2} \in \mathrm{Ev}_1\,\mathrm{Ev}_2} S\left(\mathrm{Pr}_3 \,||\, \mathrm{Pr}_{1,2}\right).$$

The process continues in the same way with each successive piece of evidence until achieving $\mathrm{Ev}_1 \cdots \mathrm{Ev}_K$, the set of the distributions of proof given the body of total evidence:

$$\mathrm{Ev}_1 \cdots \mathrm{Ev}_K = \arg \sup_{\mathrm{Pr}_K \in \mathrm{Ev}_K} \sup_{\mathrm{Pr}_{1,\ldots,K-1} \in \mathrm{Ev}_1 \cdots \mathrm{Ev}_{K-1}} S\left(\mathrm{Pr}_K \,||\, \mathrm{Pr}_{1,\ldots,K-1}\right).$$

Distributions of proof combine these previous generalizations of the principle of maximum en-

tropy [26, 29, 27, 28]:

- Maximizing differential entropy over a sequence of linear constraints that need not be mutually consistent [23, 22]
- Maximizing over both arguments of $S(\bullet||\bullet)$ rather than only over the first argument [18, 9]

3 Bayesian and frequentist special cases

3.1 Posterior distributions from prior distributions

With the general framework of Section 2 in place, Bayesian inference will now be derived from the definition of differential entropy (3) and from the example of equations (1) and (2). By substitution into equation (4), the set of the distributions of proof given the body of total evidence is

$$\begin{aligned}
\mathrm{Ev}_1\,\mathrm{Ev}_2 &= \arg \sup_{\mathrm{Pr}_2 \in \{\mathrm{Pr}:\mathrm{Pr}(\mathrm{Data}=\mathrm{data})=1\}} \sup_{\mathrm{Pr}_1 \in \{\mathrm{Pr}:\mathrm{Pr}(U\in\bullet)=B\}} -\int \rho_{1,2}(u,\mathrm{data}) \log \rho_{1,2}(u,\mathrm{data})\, d\,\mathrm{Pr}_1(u,\mathrm{data}) \\
&= \arg \sup_{\mathrm{Pr}_2:\mathrm{Pr}_2(\mathrm{Data}=\mathrm{data})=1} \sup_{\mathrm{Pr}_1:\mathrm{Pr}_1(U\in\bullet)=B} -\int \rho_{1,2}(u,\mathrm{data}) \log \rho_{1,2}(u,\mathrm{data})\, d\,\mathrm{Pr}_1(u,\mathrm{data}) \\
&= \arg \sup_{\mathrm{Pr}_2:\mathrm{Pr}_2(\mathrm{Data}=\mathrm{data})=1} -\int \rho_{1,2}(u,\mathrm{data}) \log \rho_{1,2}(u,\mathrm{data})\, d\,\mathrm{Pr}_1(u,\mathrm{data}),
\end{aligned}$$

where ρ_2 is the probability density function of Pr_2 with respect to a dominating measure Pr_1 such that $\mathrm{Pr}_1(U \in \bullet) = B$. Then $\mathrm{Ev}_1\,\mathrm{Ev}_2 = \{\mathrm{Pr}_{\mathrm{data}}\}$; the unique distribution of proof is $\mathrm{Pr}_{\mathrm{data}}$, which is the joint probability measure of U and Data such that $\mathrm{Pr}_{\mathrm{data}}(\mathrm{Data} = \mathrm{data}) = 1$ and such that B_{data}, the marginal distribution of U from $\mathrm{Pr}_{\mathrm{data}}$, satisfies

$$B_{\mathrm{data}}(\bullet) = \mathrm{Pr}_{\mathrm{data}}(U \in \bullet) = \mathrm{Pr}_1(U \in \bullet|\,\mathrm{Data} = \mathrm{data}) \tag{5}$$

[44, 22].

From the right-hand side, it can be seen that B_{data} is the posterior distribution of U indicated by Bayes's theorem. All of the usual Bayesian results can then be derived from B_{data}, including credible intervals, expected-utility-maximizing decisions, expected-loss-minimizing estimates, and the posterior probabilities of hypotheses about U.

3.2 Approximate confidence distributions from p-values

Just as Section 3.1 formalizes what evidence is needed to use Bayesian inference to draw scientific conclusions, this subsection clarifies what evidence would be needed to accept or reject scientific hypotheses at various levels of proof using frequentist inference. Instead of admitting a prior distribution as evidence, this section admits a method of generating p-values and confidence intervals.

This frequentist special case of the general framework of Section 2 starts with a p-value testing the null hypothesis that U is equal to a particular scalar or vector u on the basis of data, the observed data set. Such a p-value is written as $p_{\text{data}}(u)$ and has the defining property that the conditional distribution of $p_{\text{Data}}(u)$ is uniform between 0 and 1 ($p_{\text{Data}}(u) \sim \text{Uniform}_{0,1}$) given $U = u$. Recall that data, typically a vector or matrix of measurements, is a realization of Data, a random data set. An observed 95% confidence set (a confidence interval for scalar u) $u_{\text{data}}(95\%)$ is the set of all possible values of U that are not rejected at the $100\% - 95\%$ significance level, that is,

$$u_{\text{data}}(95\%) = \{u : p_{\text{data}}(u) \geq 0.05\}, \tag{6}$$

and similarly for confidence levels other than 95%. The frequentist probability that the random confidence set $u_{\text{Data}}(95\%)$ covers the true value of u is 95%.

For example, if U is the difference between the mean blood pressure of a treatment group and a control group, then the typical null hypothesis to test on the basis of a record data of clinical measurements would be that $U = 0$, and the p-value testing that hypothesis would be denoted by $p_{\text{data}}(0)$. If there were no difference between the treatment and control groups and if the study were repeated enough times, the histogram of such p-values would be roughly flat between 0 and 1, with about 5% of them below 0.05. There is a 95% probability that the 95% confidence interval $u_{\text{Data}}(95\%)$ includes the true mean difference.

The body of total evidence that recovers much of frequentist inference is this sequence of three pieces of evidence:

1. The background evidence includes every joint measure such that the distribution of $p_{\text{Data}}(u)$ dominates $\text{Uniform}_{0,1}$ for every value u of the unknown quantity of interest, and such that the marginal distribution of Data is a measure that dominates D, where D is defined in the next piece of evidence in this numbered list. That background evidence is broad enough to represent very little information in itself.

2. In addition to $p_{\text{Data}}(u) \sim \text{Uniform}_{0,1}$ for every u, the marginal distribution of Data is D, an

unknown prior predictive distribution.

3. The probability that Data = data is 100%.

In the mathematical notation, the body of total evidence is an ordered triple $(\mathrm{Ev}_1, \mathrm{Ev}_2, \mathrm{Ev}_3)$ satisfying these equations:

$$\mathrm{Ev}_1 \subset \{\mathrm{Pr} : \mathrm{Pr}(\mathrm{Data} \in \bullet) \gg D, \mathrm{Pr}(p_{\mathrm{Data}}(u) \in \bullet) \gg \mathrm{Uniform}_{0,1} \text{ for all } u\};$$

$$\mathrm{Ev}_2 = \{\mathrm{Pr} : \mathrm{Pr}(\mathrm{Data} \in \bullet) = D, \mathrm{Pr}(p_{\mathrm{Data}}(u) \in \bullet) = \mathrm{Uniform}_{0,1} \text{ for all } u\};$$

$$\mathrm{Ev}_3 = \{\mathrm{Pr} : \mathrm{Pr}(\mathrm{Data} = \mathrm{data}) = 1\}.$$

By substitution into equation (4), the set of the distributions of proof given the first two pieces of evidence is

$$\begin{aligned}\mathrm{Ev}_1\,\mathrm{Ev}_2 = \arg \sup_{\mathrm{Pr}_2:\mathrm{Pr}_2(\mathrm{Data}\in\bullet)=D,\mathrm{Pr}_2(p_{\mathrm{Data}}(u)\in\bullet)=\mathrm{Uniform}_{0,1}\text{ for all }u} \\ \sup_{\mathrm{Pr}_1:\mathrm{Pr}_1(\mathrm{Data}\in\bullet)\gg D,\mathrm{Pr}_1(p_{\mathrm{Data}}(u)\in\bullet)\gg\mathrm{Uniform}_{0,1}\text{ for all }u} -\int \rho_{1,2}(u,\mathrm{data})\log\rho_{1,2}(u,\mathrm{data})\,d\,\mathrm{Pr}_1(u,\mathrm{data}) \\ = \arg \sup_{\mathrm{Pr}_2:\mathrm{Pr}_2(\mathrm{Data}\in\bullet)=D,\mathrm{Pr}_2(p_{\mathrm{Data}}(U)\in\bullet)=\mathrm{Uniform}_{0,1}} -\int \rho_{1,2}(u,\mathrm{data})\log\rho_{1,2}(u,\mathrm{data})\,d\,\mathrm{Pr}_1(u,\mathrm{data}),\end{aligned} \tag{7}$$

where $\rho_{1,2}$ is the probability density function of Pr_2 with respect to any Pr_1 such that $\mathrm{Pr}_1(p_{\mathrm{Data}}(u) \in \bullet) \gg \mathrm{Uniform}_{0,1}$ for all u and such that $\mathrm{Pr}_1(\mathrm{Data} \in \bullet) \gg D$. Equation (7) says $\mathrm{Ev}_1\,\mathrm{Ev}_2$ is the set of joint distributions maximizing the differential entropy while holding the marginal distributions fixed at $\mathrm{Uniform}_{0,1}$ and D. Since the joint distribution maximizing the entropy while holding the marginal distributions of two random variables fixed is the product distribution corresponding to the independence of those random variables [17, pp. 309-310], it follows that for every $\mathrm{Pr}_\perp$ in $\mathrm{Ev}_1\,\mathrm{Ev}_2$, the joint distribution of $p_{\mathrm{Data}}(U)$ and Data is such that $p_{\mathrm{Data}}(U)$ and Data are independent, $\mathrm{Pr}_\perp(p_{\mathrm{Data}}(U) \in \bullet) = \mathrm{Uniform}_{0,1}$, and $\mathrm{Pr}_\perp(\mathrm{Data} \in \bullet) = D$. The set of those distributions is denoted by $\mathrm{Ev}_\perp$.

Since $\mathrm{Ev}_1\,\mathrm{Ev}_2 = \mathrm{Ev}_\perp$, it follows that the set of proof distributions given the body of total

evidence is

$$\begin{aligned}\mathrm{Ev}_1\,\mathrm{Ev}_2\,\mathrm{Ev}_3 &= \mathrm{Ev}_\perp\,\mathrm{Ev}_3 \\ &= \arg \sup_{\mathrm{Pr}_3:\mathrm{Pr}_3(\mathrm{Data}=\mathrm{data})=1} \sup_{\mathrm{Pr}_\perp\in\mathrm{Ev}_\perp} -\int \rho_{\perp,3}\left(u,\mathrm{data}\right)\log\rho_{\perp,3}\left(u,\mathrm{data}\right)d\,\mathrm{Pr}_\perp\left(u,\mathrm{data}\right),\end{aligned}$$

where $\rho_{\perp,3}$ is the probability density function of Pr_3 with respect to $\mathrm{Pr}_\perp$. By reasoning analogous to that behind equation (5), that results in $\mathrm{Pr}_{\perp,\mathrm{data}}$ as a distribution of proof, where

$$\mathrm{Pr}_{\perp,\mathrm{data}}\left(\bullet\right) = \mathrm{Pr}_\perp\left(\bullet\middle|\,\mathrm{Data}=\mathrm{data}\right)$$

for a $\mathrm{Pr}_\perp \in \mathrm{Ev}_\perp$. Then C_{data}, the marginal distribution of U from $\mathrm{Pr}_{\perp,\mathrm{data}}$, satisfies

$$C_{\mathrm{data}}\left(\bullet\right) = \mathrm{Pr}_{\perp,\mathrm{data}}\left(U\in\bullet\right) = \mathrm{Pr}_\perp\left(U\in\bullet\middle|\,\mathrm{Data}=\mathrm{data}\right). \tag{8}$$

Since $p_{\mathrm{Data}}\left(U\right) \sim \mathrm{Uniform}_{0,1}$ and since $p_{\mathrm{Data}}\left(U\right)$ and Data are independent under every $\mathrm{Pr}_\perp \in \mathrm{Ev}_\perp$, it follows that $p_{\mathrm{data}}\left(U\right) \sim \mathrm{Uniform}_{0,1}$ for $U \sim C_{\mathrm{data}}$. Finally, the probability that U is in the observed 95% confidence set $u_{\mathrm{data}}\left(95\%\right)$ of equation (6) is

$$\begin{aligned}C_{\mathrm{data}}\left(U\in u_{\mathrm{data}}\left(95\%\right)\right) &= C_{\mathrm{data}}\left(U\in\left\{u : p_{\mathrm{data}}\left(u\right)\geq 0.05\right\}\right) \\ &= C_{\mathrm{data}}\left(p_{\mathrm{data}}\left(U\right)\geq 0.05\right) = \mathrm{Uniform}_{0,1}\left(\left[0.05,1\right]\right) = 95\%.\end{aligned}$$

Since data is fixed rather than random, that probability is a degree of proof, not a frequentist probability.

Since, analogously, the degree of proof that U is in the observed confidence set of any level other than 95% is equal to that level, C_{data} meets the conditions of an *approximate confidence distribution* [5, 8] such as, in the case of scalar U, an asymptotic confidence distribution [38, 39]. An advantage of an approximate confidence distribution as opposed to an exact confidence distribution [e.g., 36, 34, 41, 13, 30, 35, 40] is that the former applies not only to continuous data but also to discrete data. For discrete data would require $p_{\mathrm{Data}}\left(u\right)$ to be only approximately uniform, and thus the 95% frequentist probability of coverage by the confidence set would also be approximate [12].

For example, suppose u is a scalar and that the function p_{data} is strictly monotonic increasing.

In that case, for any values u_1 and u_2 such that $u_1 < u_2$, equation (8) then gives

$$\begin{aligned} C_{\text{data}}\left(u_1 \leq U \leq u_2\right) &= C_{\text{data}}\left(p_{\text{data}}\left(u_1\right) \leq p_{\text{data}}\left(U\right) \leq p_{\text{data}}\left(u_2\right)\right). \\ &= p_{\text{data}}\left(u_2\right) - p_{\text{data}}\left(u_1\right), \end{aligned}$$

indicating that p_{data} is the cumulative distribution function of C_{data}. For that reason, p_{data} is called a "confidence distribution" or a "significance function" [38, 39]. The "confidence" terminology comes from the fact that whenever u_1 and u_2 are the limits of a 95% confidence interval interval, $C_{\text{data}}\left(u_1 \leq U \leq u_2\right) = 95\%$, and similarly for any level of confidence other than 95%.

The material in this section can be generalized from p-values to what statisticians call "pivots" and "pivotal quantities," which extend Gauss's concept of an error. Steps analogous to those above then lead to the independence of the pivotal quantity and Data. That independence implies that the proof distribution of U is a fiducial distribution [19], thereby establishing an extended principle of maximum entropy as what Evans [20, p. 91] called the lacking "new principle" needed to justify fiducial inference. This century has experienced the development of many fiducial-type distributions, many of which are approximate confidence distributions [25, 24, 46, 21, 42, 47, 2, 31, 14, 33].

4 Evidentiary model averaging and applications to p-hacking

4.1 Evidentiary model averaging

Typically, there is uncertainty not only about U, the random variable or random vector of interest, but also about the model assumptions behind the pieces of evidence. In the Bayesian case of Section 3.1, the body of total evidence depends on a prior distribution, whereas in the frequentist case of Section 3.2, the body of total evidence depends on a statistical testing procedure.

In both cases, there is also dependence on the *family*, the set of possible distributions of Data. Each value u of the unknown quantity of interest may correspond to multiple distributions in the family. In parametric statistics, u is is a vector of finite dimension and is called the *parameter of interest*, and the distribution in the family is in general indexed not only by the parameter of interest but also by a nuisance parameter. The nuisance parameter is not explicitly represented in Section 3 since it is integrated out with respect to a prior distribution in the Bayesian case and since the p-values in the frequentist case are valid for all values of the nuisance parameter. In nonparametric statistics, the parameter values are functions rather than finite-dimensional vectors.

In addition to uncertainty about the family of possible data distributions and uncertainty about

the prior distribution in the Bayesian case or the testing method in the frequentist case, there is often also uncertainty about whether to use Bayesian inference or frequentist inference. To concisely capture all of the potentially uncertain assumptions, each body of total evidence that is uncertain is called a *model of total evidence* and is indexed by a possible value of a random variable M. The number of models of total evidence is denoted by N.

That uncertainty about M can be propagated to statements about U by considering a body of total evidence at a hierarchical level above those of the models of total evidence. That is accomplished by applying the framework of Section 2 except with M in place of U and with a higher-level body of total evidence, which is called the *body of total evidence about the model*. The resulting proof distribution is then equivalent to a probability mass function written as pmf. Then an *average proof distribution* of U is

$$\Pr^{\{1,\ldots,N\}}(\bullet) = \operatorname{Prob}(U \in \bullet) = \sum_{m=1,\ldots,N} \operatorname{Prob}(M = m) \operatorname{Prob}(U \in \bullet | M = m) = \sum_{m=1,\ldots,N} \operatorname{pmf}(m) \Pr^{(m)}(\bullet), \tag{9}$$

where $\Pr^{(m)}$ is a distribution of proof from the mth model of total evidence for each $m = 1, \ldots, N$.

If the body of total evidence about the model is itself uncertain enough to matter, then the hierarchy can be analogously extended by adding a yet higher-level body of total evidence. The process may be repeated until the unquantified uncertainty is considered negligible, as is done with hierarchical Bayesian modeling.

In the simplest case, the body of total evidence about the model, not involving Data, is this sequence of pieces of evidence:

1. The only member of the background evidence is the counting measure $\Pr_{\#}$ on subsets of $\{1, \ldots, N\}$. That has the effect of putting equal weight on each model of total evidence.

2. This piece of evidence is vacuous in the sense that it is the set Ev_* of all distributions on subsets of $\{1, \ldots, N\}$.

More succinctly, that body of total evidence about the model is the ordered pair $(\mathrm{Ev}_{\#}, \mathrm{Ev}_*)$, where $\mathrm{Ev}_{\#} = \{\Pr_{\#}\}$. By substitution into equation (4), the set of the distributions of proof given $(\mathrm{Ev}_{\#}, \mathrm{Ev}_*)$ is

$$\begin{aligned} \mathrm{Ev}_{\#}\,\mathrm{Ev}_* &= \arg \sup_{\Pr_2 \in \mathrm{Ev}_*} \sup_{\Pr_1 \in \mathrm{Ev}_{\#}} -\int \rho_{1,2}(m) \log \rho_{1,2}(m)\, d\Pr_1(m) \\ &= \arg \sup_{\Pr_2 \in \mathrm{Ev}_*} -\int \rho_{\#,2}(m) \log \rho_{\#,2}(m)\, d\Pr_{\#}(m), \end{aligned}$$

where $\rho_{\#,2}$, as the probability density of $\Pr_2$ with respect to $\Pr_\#$, is the probability mass function according to $\Pr_2$. The equation simplifies to

$$\mathrm{Ev}_\# \, \mathrm{Ev}_* = \arg \sup_{\Pr_2 \in \mathrm{Ev}_*} -\sum \rho_{\#,2}\left(m\right) \log \rho_{\#,2}\left(m\right), \tag{10}$$

which is just maximum Shannon entropy. As is well known, the probability mass function on $\{1, \ldots, N\}$ that maximizes the Shannon entropy is $\mathrm{pmf}_=$, the probability mass function assigning equal probability to each member of the set, in this case each model of total evidence. Equation (10) then reduces to $\mathrm{Ev}_\# \, \mathrm{Ev}_* = \{\mathrm{pmf}_=\}$, which says $\mathrm{pmf}_=$ is the unique distribution of proof given $(\mathrm{Ev}_\#, \mathrm{Ev}_*)$. By equation (9),

$$\Pr^{\{1,\ldots,N\}}\left(\bullet\right) = \sum_{m=1,\ldots,N} \mathrm{pmf}_=\left(m\right) \Pr^{(m)}\left(\bullet\right) = \frac{1}{N} \sum_{m=1,\ldots,N} \Pr^{(m)}\left(\bullet\right), \tag{11}$$

which means the average proof distribution of U is the equal-weight mixture of the proof distributions of the models of total evidence.

4.2 Applications to *p*-hacking

Equation (11) suggests taking the mean of p-values and/or posterior probabilities over different methods and assumptions instead of p-hacking, the practice of selecting those achieving the most statistical significance. For example, Figure 2 displays probabilities of making a sign error averaged over Bayesian and frequentist methods weighted equally. The more general equation (9) can be used to weight the mean as required by the body of total evidence about the model.

Since 95% confidence intervals and 95% credible intervals are special cases of intervals having a 95% degree of proof, they can also be combined using equation (9). It can be used to average the approximate confidence distributions behind confidence intervals and/or the posterior distributions behind credible intervals. The result is a 95% proof interval such that there is a 95% degree of proof that the quantity of interest is in that interval, and similarly for degrees of proof higher or lower than 95%. For example, the special case of equation (11) is used in Figure 3 to propagate uncertainty about which Bayesian model of molecular evolution to use. Another example, lacking the above evidentiary foundation, was supported by other arguments [7].

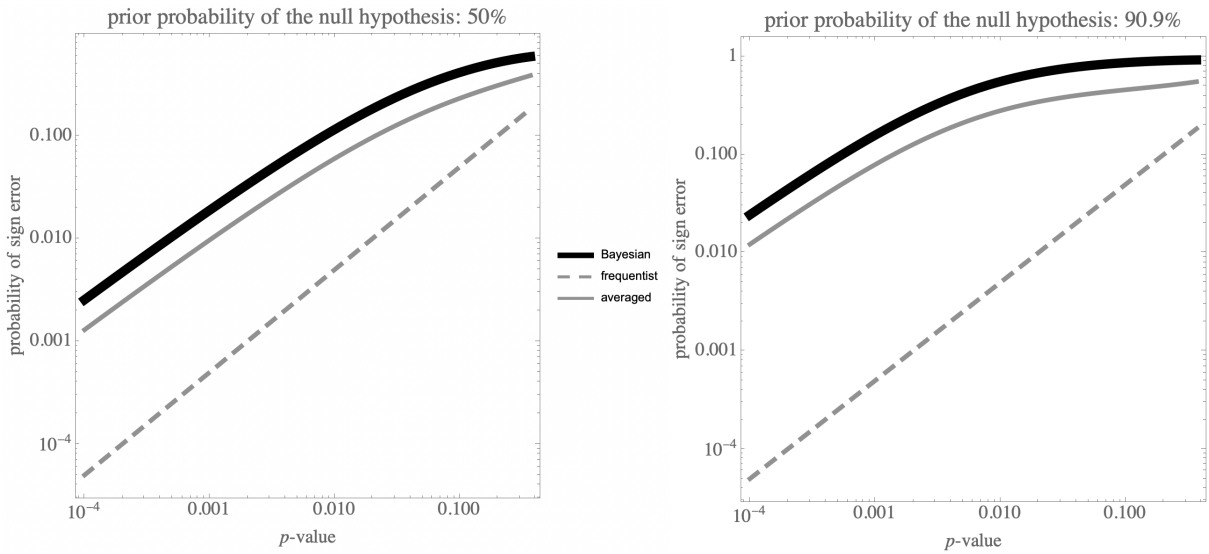

Figure 2: Probabilities that either $\widehat{u} > u_0$ when $U \leq u_0$ or $\widehat{u} < u_0$ when $U \geq u_0$, where $\widehat{u}$ is the observed estimate of U, and u_0 is the value that U would have if the null hypothesis were true, as functions of a two-sided p-value. The frequentist line is the probability of a sign error according to an approximate confidence distribution [6]. With that distribution as an approximate posterior distribution conditional on $U \neq u_0$ [6], the Bayesian curve is the posterior probability of a sign error using a Bayes factor calibration [37] and a 50% (left) or 90.9% [3] (right) prior probability that $U = u_0$. Since both probabilities are degrees of proof given their respective pieces of evidence (Section 3) and since there is considerable uncertainty about those pieces of evidence, the Bayesian and frequentist probabilities are "averaged" by taking their mean to obtain the overall degree of proof that there is a sign error (Section 4.1).

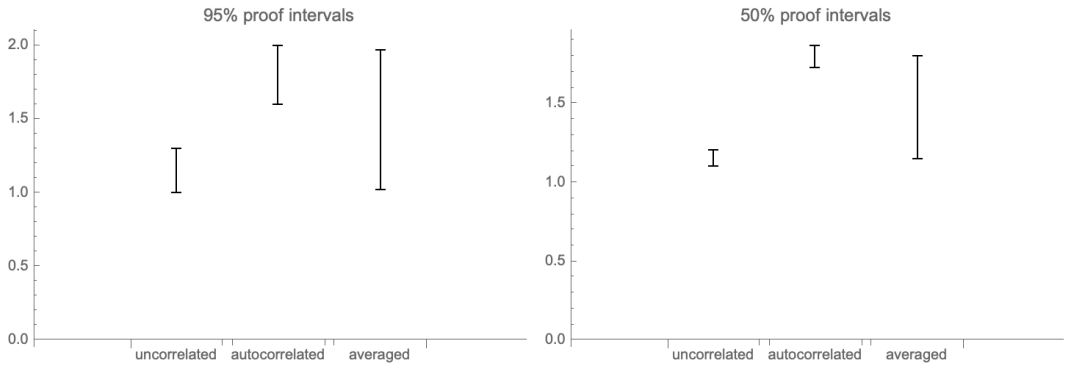

Figure 3: Intervals of cyanobacteria divergence times in billions of years ago as estimated from molecular and fossil data. The first two intervals on the left are the 95% credible intervals under a Cauchy 50% prior distribution under uncorrelated and autocorrelated models as reported in Betts et al. [4, Fig. 1]; the intervals do not overlap [16], indicating that neither interval in itself adequately quantifies the uncertainty. The plot on the right displays the corresponding 50% credible intervals assuming the posterior distributions are approximately normal. Under that approximation, the averaged equal-tail 95% and 50% proof intervals from equation (11) are also displayed. While the limits of the averaged 95% proof interval are very close to the maximum of the two upper limits and the minimum of the two lower limits, in agreement with a conservative method of uncertainty propagation [10], the averaged 50% proof interval is less conservative. By contrast, the method of taking the union of the credible intervals [11] results in sets that are not intervals when the credible intervals do not overlap.

Acknowledgments

The main result of Section 3.2 was derived at the University of Ottawa with partial support from the Natural Sciences and Engineering Research Council of Canada (RGPIN/356018-2009). The rest of this research was supported by the University of North Carolina at Greensboro.

References

[1] Andrade, C., 2021. HARKing, cherry-picking, P-hacking, fishing expeditions, and data dredging and mining as questionable research practices. The Journal of Clinical Psychiatry 82, 20f13804.

[2] Balch, M.S., 2012. Mathematical foundations for a theory of confidence structures. International Journal of Approximate Reasoning 53, 1003–1019. doi:10.1016/j.ijar.2012.05.006.

[3] Benjamin, D.J., Berger, J.O., Johannesson, M., Nosek, B.A., Wagenmakers, E.J., Berk, R., Bollen, K.A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C.D., Clyde, M., Cook, T.D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., Fehr, E., Fidler, F., Field, A.P., Forster, M., George, E.I., Gonzalez, R., Goodman, S., Green, E., Green, D.P., Greenwald, A.G., Hadfield, J.D., Hedges, L.V., Held, L., Hua Ho, T., Hoijtink, H., Hruschka, D.J., Imai, K., Imbens, G., Ioannidis, J.P.A., Jeon, M., Jones, J.H., Kirchler, M., Laibson, D., List, J., Little, R., Lupia, A., Machery, E., Maxwell, S.E., McCarthy, M., Moore, D.A., Morgan, S.L., Munafó, M., Nakagawa, S., Nyhan, B., Parker, T.H., Pericchi, L., Perugini, M., Rouder, J., Rousseau, J., Savalei, V., Schönbrodt, F.D., Sellke, T., Sinclair, B., Tingley, D., Van Zandt, T., Vazire, S., Watts, D.J., Winship, C., Wolpert, R.L., Xie, Y., Young, C., Zinman, J., Johnson, V.E., 2018. Redefine statistical significance. Nature Human Behaviour 2, 6–10.

[4] Betts, H.C., Puttick, M.N., Clark, J.W., Williams, T.A., Donoghue, P.C., Pisani, D., 2018. Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. Nature Ecology & Evolution 2, 1556–1562.

[5] Bickel, D.R., 2020. Confidence intervals, significance values, maximum likelihood estimates, etc. sharpened into Occam's razors. Communications in Statistics - Theory and Methods 49, 2703–2712.

[6] Bickel, D.R., 2021a. Null hypothesis significance testing interpreted and calibrated by estimating probabilities of sign errors: A Bayes-frequentist continuum. The American Statistician 75, 104–112.

[7] Bickel, D.R., 2021b. Propagating uncertainty about molecular evolution models and prior distributions to phylogenetic trees URL: `https://doi.org/10.5281/zenodo.5810696`. working paper, DOI: 10.5281/zenodo.5810696.

[8] Bickel, D.R., 2022a. Confidence distributions and empirical Bayes posterior distributions unified as distributions of evidential support. Communications in Statistics - Theory and Methods 51, 3142–3163.

[9] Bickel, D.R., 2022b. Maximum entropy derived and generalized under idempotent probability to address Bayes-frequentist uncertainty and model revision uncertainty: An information-theoretic semantics for possibility theory. Fuzzy Sets and Systems URL: `https://doi.org/10.1016/j.fss.2022.05.009`. DOI: 10.1016/j.fss.2022.05.009.

[10] Bickel, D.R., 2022c. Phylogenetic Trees and Molecular Evolution: A Hands-on Introduction with Uncertainty Quantification Corrected. Springer, New York. URL: `https://davidbickel.com/evolution/`. Forthcoming.

[11] Bickel, D.R., 2022d. Propagating clade and model uncertainty to confidence intervals of divergence times and branch lengths. Molecular Phylogenetics and Evolution 167, 107357.

[12] Bickel, D.R., Patriota, A.G., 2019. Self-consistent confidence sets and tests of composite hypotheses applicable to restricted parameters. Bernoulli 25, 47–74.

[13] Bityukov, S., Krasnikov, N., Nadarajah, S., Smirnova, V., 2011. Confidence distributions in statistical inference. AIP Conference Proceedings 1305, 346–353.

[14] Bowater, R.J., 2017. A defence of subjective fiducial inference. AStA Advances in Statistical Analysis 101, 177–197.

[15] Bromham, L., 2016. An Introduction to Molecular Evolution and Phylogenetics. Oxford University Press, Oxford.

[16] Bromham, L., 2019. Six impossible things before breakfast: Assumptions, models, and belief in molecular dating. Trends in Ecology & Evolution 34, 474–486.

[17] Cover, T., Thomas, J., 2007. Elements of Information Theory, Second Edition: Solutions to Problems. URL: https://bit.ly/3QtCVnW. accessed 16 June 2022.

[18] Csiszár, I., 1985. An extended maximum entropy principle and a Bayesian justification, in: Bernardo, J., DeGroot, M., Lindley, D.V., Smith, A. (Eds.), Bayesian Statistics 2. Elsevier Inc., Amsterdam, pp. 83–98.

[19] Dempster, A.P., 1963. On direct probabilities. Journal of the Royal Statistical Society: Series B (Methodological) 25, 100–110.

[20] Evans, M., 2015. Measuring Statistical Evidence Using Relative Belief. Chapman & Hall/CRC Monographs on Statistics & Applied Probability, CRC Press, New York.

[21] Gibson, G.J., Streftaris, G., Zachary, S., 2011. Generalised data augmentation and posterior inferences. Journal of Statistical Planning and Inference 141, 156–171.

[22] Giffin, A., Cafaro, C., Ali, S.A., 2016. Application of the maximum relative entropy method to the physics of ferromagnetic materials. Physica A: Statistical Mechanics and its Applications 455, 11–26.

[23] Giffin, A., Caticha, A., 2007. Updating probabilities with data and moments, pp. 74–84.

[24] Hannig, J., 2009. On generalized fiducial inference. Statistica Sinica 19, 491–544.

[25] Hannig, J., Iyer, H., Patterson, P., 2006. Fiducial generalized confidence intervals. Journal of the American Statistical Association 101, 254–269.

[26] Jaynes, E., 1957. Information theory and statistical mechanics. Physical Review 106, 620–630.

[27] Jaynes, E., 1982. On the rationale of maximum-entropy methods. Proceedings of the IEEE 70, 939–952.

[28] Jaynes, E., 2003. Probability Theory: The Logic of Science. Cambridge University Press, Cambridge.

[29] Jaynes, E.T., 1989. Where do we stand on maximum entropy? (1978), in: Rosenkrantz, R. (Ed.), E. T. Jaynes: Papers on Probability, Statistics and Statistical Physics. Springer Netherlands. volume 158 of *Synthese Library*, pp. 210–314.

[30] Kim, D., Lindsay, B.G., 2011. Using confidence distribution sampling to visualize confidence sets. Statistica Sinica 21, 923–948.

[31] Martin, R., Liu, C., 2013. Inferential Models: A Framework for Prior-Free Posterior Probabilistic Inference. Journal of the American Statistical Association 108, 301–313.

[32] Mayo, D., 2018. Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars. Cambridge University Press, Cambridge.

[33] Plante, A., 2020. A Gaussian alternative to using improper confidence intervals. Canadian Journal of Statistics 48, 773–801.

[34] Polansky, A.M., 2007. Observed Confidence Levels: Theory and Application. Chapman and Hall, New York.

[35] Schweder, T., Hjort, N., 2016. Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge.

[36] Schweder, T., Hjort, N.L., 2002. Confidence and likelihood. Scandinavian Journal of Statistics 29, 309–332.

[37] Sellke, T., Bayarri, M.J., Berger, J.O., 2001. Calibration of p values for testing precise null hypotheses. American Statistician 55, 62–71.

[38] Singh, K., Xie, M., Strawderman, W.E., 2005. Combining information from independent sources through confidence distributions. Annals of Statistics 33, 159–183.

[39] Singh, K., Xie, M., Strawderman, W.E., 2007. Confidence distribution (CD) – distribution estimator of a parameter. IMS Lecture Notes Monograph Series 2007 54, 132–150.

[40] Taraldsen, G., Lindqvist, B.H., 2018. Conditional fiducial models. Journal of Statistical Planning and Inference 195, 141–152.

[41] Tian, L., Wang, R., Cai, T., Wei, L.J., 2011. The highest confidence density region and its usage for joint inferences about constrained parameters. Biometrics 67, 604–10.

[42] Wang, C., Hannig, J., Iyer, H.K., 2012. Fiducial prediction intervals. Journal of Statistical Planning and Inference 142, 1980–1990.

[43] Wasserstein, R.L., Schirm, A.L., Lazar, N.A., 2019. Moving to a world beyond "p < 0.05". The American Statistician 73, 1–19.

[44] Williams, P.M., 1980. Bayesian conditionalisation and the principle of minimum information. The British Journal for the Philosophy of Science 31, 131–144.

[45] Williamson, J., 2010. In Defence of Objective Bayesianism. Oxford University Press, Oxford.

[46] Xiong, S., Mu, W., 2009. On construction of asymptotically correct confidence intervals. Journal of Statistical Planning and Inference 139, 1394–1404. doi:10.1016/j.jspi.2008.08.014.

[47] Zhao, S., Xu, X., Ding, X., 2012. Fiducial inference under nonparametric situations. Journal of Statistical Planning and Inference 142, 2779 – 2798. doi:10.1016/j.jspi.2012.03.023.