Proceedings of the 3rd International Conference on Statistics: Theory and Applications (ICSTA'21)
~~Prague, Czech Republic~~ Virtual Conference – August, 2021
Paper No. 139
DOI: 10.11159/icsta21.139

The Concept of Statistical Evidence

Michael Evans
Dept. of Statistical Sciences
University of Toronto
Toronto, Canada
mevans@utstat.utoronto.ca

Abstract - The concept of statistical evidence has proven to be somewhat elusive in the development of the discipline of Statistics. Still there is a conviction that appropriately collected data contains evidence concerning the answers to questions of scientific interest. We discuss some of the attempts at making the concept of evidence precise and, in particular, present an approach based upon measuring how beliefs change from a priori to a posteriori. Of necessity this is Bayesian in nature as a proper prior is required that reflects beliefs about where the truth lies before the data is observed. Bayesian inference is often criticized for its subjective nature. It is possible, however, to deal with this subjectivity in a scientifically sound manner. In part, this is done by assessing and controlling the bias the prior and model induce into inferences and this depends intrinsically on being clear about statistical evidence. In addition, the model and the prior are falsifiable through model checking and checking for prior-data conflict. Both the assessment of bias and the falsification steps are essentially frequentist in nature so this provides a degree of unity between sometimes conflicting philosophies. This approach to statistical reasoning can be seen as dealing with the inevitable subjectivity required in the choice of ingredients to an analysis so that a statistical analysis can approach the goal of objectivity that is central to scientific work.

***Keywords*:** principle of evidence, relative belief, bias, coverage probability

1. Introduction

The concept of statistical evidence is central to the subject of statistics but there is a lack of clarity on what it is and this causes controversy. As an example of this, consider the current discussions concerning p-values. In fact, do p-values measure statistical evidence? It seems impossible to answer such a question when we are unclear about what statistical evidence is, see, however, Example 2.

The centrality of the concept of statistical evidence has long been recognized without a generally accepted answer being presented. Alan Birnbaum's work in the 1960's, as in [1], was concerned with trying to characterize the concept via equivalence relations and has generated quite a bit of controversy ever since. Also, Royall's book on likelihood methods [2] has this central concern as does the relative belief approach discussed in [3], where the topic is approached from a Bayesian perspective. These references are closest to the spirit of this talk but there are many books and papers where statistical evidence is the primary concern, see [4]-[8]. Also, of some considerable importance are the many treatments and discussions in the philosophy of science literature where it is known as confirmation theory, see [9].

It is reasonable to ask what the ultimate goal is in being as precise as possible about statistical evidence. As expressed here, this is somewhat ambitious as it is to develop a theory of statistical reasoning that works, in that it provides answers to statistical problems, is logically sound, in other words free of paradoxes, and is scientifically sound, namely, conforms to the values that are recognized as representing good science such as objectivity. Furthermore, we want that theory to be based on a clear characterization of statistical evidence.

This is necessarily considerably idealized and there is no pretention that such a theory will fully satisfy the needs of the working statistician who encounters all kinds of messy problems that require considerable flexibility. Yet, without a central core to the subject, and that is what such a theory aims to provide, doubts exist concerning the validity of any statistical methodology. Also, our concern is with a "theory of statistical reasoning" as opposed to "theory of statistical inference". By this we mean that there is more required for a successful theory than just stating the rules of inference.

This paper is concerned with some developments related to [3]. This is outlined here and it is shown how this can resolve certain difficulties. One outcome is to find complementary roles for Bayes, for inference, and for frequentism, for design and assessing the merits/reliability of a study. This points to a possible resolution between what are at times quite different approaches to statistical reasoning.

2. The Problem

Consider then a scientific context where there are questions concerning an object of interest Ψ and data x has been collected (assumed properly) believed to contain *evidence* concerning the answers to one or both of the following problems.

E *estimation:* Provide an estimate $\psi(x)$ of Ψ together with an assessment of its accuracy based on the evidence.
H *hypothesis assessment:* Quote the evidence in favour of or against some specified value ψ_0 of Ψ together with an assessment of the strength of the evidence.

It is to be noted here that, in considering **H**, it is stated that it is desirable that evidence in favour of ψ_0 being true can be stated as well as evidence against. How then are we to reason from the data to answer **E** and/or **H**?

3. The Ingredients

So far all we have available to answer this question is the data x. Perhaps it would be best to just use the data alone, but it seems more is needed and these are of necessity choices made by someone and, in particular, often by the statistician. It is then reasonable to ask what characteristics do we want these ingredients to have? The following seem necessary, but there may well be others.

$\mathbf{I}_1$ *Minimal*: The minimal ingredients needed to get a valid characterization/measure of statistical evidence.
$\mathbf{I}_2$ *Bias*: An assessment can be made to determine to what extent the chosen ingredients produce foregone conclusions to **E** or **H**.
$\mathbf{I}_3$ *Falsifiable*: Any chosen ingredient specified can be assessed against the (objective) data to see if it is contradicted.

Certainly, objectivity is the ideal goal in scientific endeavours but subjective choices need to be made when carrying out a statistical analysis. So $\mathbf{I}_2$ and $\mathbf{I}_3$ are concerned with assessing and controlling the effects of subjectivity as much as possible. Also, we will always assume here that the data x, if collected properly, is objective.

For the developments here the following two subjective ingredients will be used.

Model: $\{f_\theta: \theta \in \Theta\}$ a collection of conditional probability distributions for x given θ such that the object of interest $\psi = \Psi(\theta) \in \Psi$ (Ψ also denotes the set of possible values) is specified by the true distribution that gave rise to x.
Prior: π a probability distribution on Θ.

Discussion of $\mathbf{I}_2$ is deferred until later in the paper. Both ingredients, however, satisfy $\mathbf{I}_3$ via model checking for the model and via checking for prior-data conflict for the prior, see [10]. In the case of the prior, falsification means that an indication has been found that the true value lies in the tails of the prior. Methodology for modifying a prior when such a conflict occurs is discussed in [11]. There is also another component to our discussion that is not really a choice, as it is largely dictated by the measurement process that produced the data.

Meaningful Difference δ: the *difference that matters* so that $d_\Psi(\psi_1, \psi_2) \leq \delta$, for some distance measure d_Ψ, means that ψ_1 and ψ_2 are, for practical purposes, indistinguishable.

The need for specifying δ has long been recognized as can be seen from [12], a reference from 1919, and it plays a role here.

The model and prior specify a joint prior probability measure P for $\omega = (\theta, x)$ with joint density given by $\pi(\theta) f_\theta(x)$. The goal now is use P together with x to answer **E** and **H**.

4. The Rules of Statistical Inference

The following three rules or principles of statistical inference are used here with the first two being the most important. These are stated for a probability model (Ω, F, P) where it is desired to infer about the occurrence of the event $A \in F$ after observing that $C \in F$ has occurred and it is assumed for now that $P(C) > 0$.

$\mathbf{R}_1$ *Principle of conditional probability*: Belief about the truth of A, as expressed $P(A)$, is replaced by $P(A|C)$.
$\mathbf{R}_2$ *Principle of evidence*: The observation of C is evidence in favor of A when $P(A|C) > P(A)$, is evidence against A when $P(A|C) < P(A)$ and is evidence neither for nor against A when $P(A|C) = P(A)$.
$\mathbf{R}_3$ *Principle of relative belief*: When comparisons are made the evidence is ordered by the relative belief ratio $RB(A|C) = P(A|C)/P(A)$.

The principle of evidence has a long history in the philosophy of science although it is not always called that. Consider the following quote from Popper [13] found in Appendix ix. "If we are asked to give a criterion of the fact that the evidence y supports or corroborates a statement x, the most obvious reply is: that y increases the probability of x."

The relative belief ratio in $\mathbf{R}_3$ can be replaced by any valid measure of statistical evidence. By *valid* we mean that there is a cut-off such that values of the measure with respect to the cut-off characterize evidence for or against, as the value 1 does for the relative belief ratio, via the principle of evidence. For example, the Bayes factor is such a measure and with the same cut-off. It can be argued, however, that the relative belief ratio has the nicest properties and is the easiest to use.

One natural question concerns what to do in the continuous case and our answer is to use limits. In general, the relative belief ratio at the value $\psi = \Psi(\theta)$ is defined as follows, where the events A_ε converge nicely to the point ψ as $\epsilon \to 0$,

$$RB_\Psi(\psi \mid x) = \lim_{\epsilon} \frac{P(A_\epsilon \mid x)}{P(A_\epsilon)} = \frac{\pi_\Psi(\psi \mid x)}{\pi_\Psi(\psi)}$$

with $\pi_\Psi(\cdot \mid x)$ the posterior density and π_Ψ the prior density of Ψ. This limit exists whenever the prior density is positive and continuous at ψ. There is a relationship between the relative belief ratio and the Bayes factor, see [3], but there is no need for "spike" priors and, moreover, in the continuous case, when a Bayes factor is defined via the same kind of limit, this limit is given by the limiting relative belief ratio.

5. Inferences

The application of the rules of inference to the ingredients comprised of the model, prior and data leads to answers to E and **H**. Since it is simpler, consider first problem **H**.

H: Assess $H_0: \Psi(\theta) = \psi_0$ via $RB_\Psi(\psi_0 \mid x)$, so evidence in favour of (against) H_0 is found if $> (<) 1$.

It is still necessary to indicate how strong this evidence is, one way or the other. There is no universal scale on which evidence is measured, so $RB_\Psi(\psi_0 \mid x)$ needs to be calibrated and that is context dependent. For this we use the ordering of $\psi \in \Psi$ given by $\mathbf{R}_3$, where ψ_1 is not preferred to ψ_2 whenever $RB_\Psi(\psi_1 \mid x) \leq RB_\Psi(\psi_2 \mid x)$. Then, with $\Pi_\Psi(\cdot \mid x)$ denoting the posterior probability measure of Ψ, we call the posterior probability $\Pi_\Psi(RB_\Psi(\psi \mid x) \leq RB_\Psi(\psi_0 \mid x) \mid x)$ the *strength* of the evidence. When there is evidence in favour, and this is a big probability, then there is little belief that the true value has an even more evidence in its favour, so there is strong evidence in favour of H_0. When there is evidence against, and this is a small probability, then there is strong evidence against H_0 via the same reasoning. There is no reason why the strength of the evidence needs to be measured by one number, however, and sometimes this needs to be augmented by additional posterior probabilities. Notice that the basic inference only relies on $\mathbf{R}_1$ and $\mathbf{R}_2$ while the measure of strength requires all three principles.

For **E** it is necessary to invoke all three principles of inference for the basic inference although the measure of accuracy only requires the first two.

E: Based on the ordering, estimate ψ by $\psi(x) = \text{argsup}\, RB_\Psi(\psi \mid x)$.

The error in the estimate is then assessed via the "size" of the plausible region given by $Pl_\Psi(x) = \{\psi: RB_\Psi(\psi \mid x) > 1\}$, the set of values having evidence in favour of being the true value, and its posterior content. Note that $Pl_\Psi(x)$ only depends on $\boldsymbol{R}_2$, which implies that all valid estimates have the same accuracy. As such, alternative choices could be made that involve, for example, some smoothing as long as the principle being used always produced values in $Pl_\Psi(x)$. It is also possible to quote a γ-credible region $C_{\Psi,\gamma}(x) = \{\psi: RB_\Psi(\psi \mid x) > k_\gamma\}$ for accuracy assessment, provided $\gamma \leq \Pi_\Psi(Pl_\Psi(x) \mid x)$ as otherwise the region would include values for which there is evidence against them being the true value. This establishes a close link with likelihood methods and frequentism as, for the full model parameter the estimate $\theta(x)$ is the MLE and $Pl_\Psi(x)$ is a particular likelihood region. The relative belief approach has a number of good properties. For example, since RB_Ψ is invariant under reparameterizations, all inferences are invariant and possess many other optimal properties, see [3].

Consider now a well-known example that has presented difficulties for the frequentist approach.

Example 1. *Fieller's problem*.

This problem has an extensive literature where [14]-[19] are notable examples. For this we have *i.i.d.* samples $x = (x_1, \ldots, x_m) \sim N(\mu, \sigma_0^2)$ independent of $y = (y_1, \ldots, y_n) \sim N(\nu, \sigma_0^2)$ where $(\mu, \nu) \in R^2$ is unknown but the common variance is known. This latter assumption can be weakened to an unknown variance but this doesn't change the basic nature of the problem where a 0.95-confidence interval is required for the ratio of means $\psi = \Psi(\mu, \nu) = \mu/\nu$. A common approach is to use the following pivotal to obtain the region,

$$(\bar{x} - \bar{y}\psi)/\sigma_0\sqrt{\frac{1}{m} + \frac{\psi^2}{n}} \sim N(0,1).$$

The issue that arises is that sometimes the absurd interval given by all of R^1 is obtained. A confidence interval is called *absurd* if can be equal to the whole relevant parameter space or the null set with positive probability for some θ values.

For the context being considered here there are also the priors given by $\mu \sim N(\mu_0, \tau_{10}^2)$, independent of $\nu \sim N(\nu_0, \tau_{20}^2)$ where, in practice, an algorithm is easily specified to elicit the hyperparameters based upon what is known about the measurement process. A natural question then is whether or not a plausible region can be absurd. Putting $C = \{x: RB_\Psi(\psi \mid x) = 1 \text{ a. e. } \Pi_\Psi\}$ and M equal to the prior probability measure of the data x, the following general result can be proved, see [20]. Therefore, a plausible region is effectively never absurd.

Theorem The plausible region for $\psi = \Psi(\theta)$ (i) never satisfies $Pl_\Psi(x) = \Psi$ and (ii) satisfies $Pl(x) = \phi$ with prior probability 0 when $M(C) = 0$.

It is still possible for the plausible region to equal an *exclusive region*, namely, $Pl(x) = (-\infty, a(x)) \cup (b(x), \infty)$ but these seem logically necessary. For example, if a 0.95-confidence interval for μ contains 0, this implies an exclusive confidence region for $1/\mu$ and that seems correct. For Fieller's problem, $RB_\Psi(\psi \mid x, y)$ can be obtained in closed form and it is easy to obtain $Pl(x)$ and $\Pi_\Psi(Pl(x) \mid x)$ but this doesn't answer the confidence question which we will return to shortly.

As a specific numerical example, suppose $\mu = 20, \nu = 10, \sigma_0^2 = 1$ so $\psi = 20/10 = 2$ and $\delta = 0.1$. With $m = n = 10$ the data was generated from this model yielding mss equal to $(\bar{x}, \bar{y}) = (20.188, 10.699)$. The 0.95-confidence region based upon the pivotal equals $(1.770, 2.016)$. Suppose now that the elicitation of the priors led to $(\mu_0, \tau_{10}) = (17.5, 2.912), (\nu_0, \tau_{20}) = (8.75, 2.328)$. This results in the estimate $\psi(x) = 1.89$ with plausible region $Pl_\Psi(x, y) = (1.76, 2.03)$. If our concern is with the hypothesis that the true value of ψ is 2, then remembering the difference that matters δ, the hypothesis $H_0: \Psi(\theta) \in (2 - \delta, 2 + \delta)$ is assessed, which has $RB_\Psi(2 \mid x, y) = 3.44$ with strength equal to 0.44. So, there is evidence in favour, but only of middling strength at best. Figure 1 is a plot of the prior and posterior densities of ψ on the left and of the relative belief ratio on the right. The prior has a very long tail in this case but the posterior concentrates quite significantly.

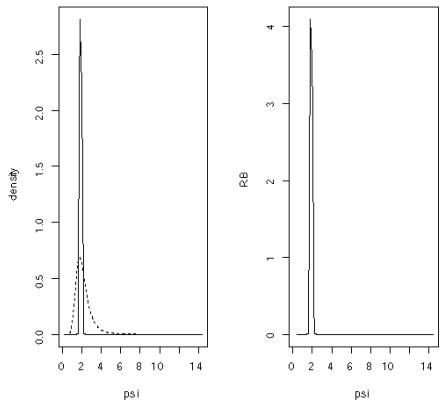

Fig. 1: Plots of the prior (- -) and posterior (-) densities of ψ on the left and of the relative belief ratio on the right in Example 1.

6. Bias

Bias calculations are considered necessary as part of assessing the quality of a study. For example, would you accept the results of a statistical analysis that reported evidence against (in favour of) $H_0: \Psi(\theta) = \psi_0$ if the prior probability of obtaining such evidence was ≈ 1 when H_0 is true (false)? This is what we mean by bias. As part of assessing bias, let $M(\cdot \mid \psi)$ denote the prior predictive probability measure for the data x given that $\Psi(\theta) = \psi$.

Consider first measuring bias for problem **H**, namely, for the hypothesis $H_0: \Psi(\theta) = \psi_0$.

H: *bias against* $= M(RB_\Psi(\psi_0 \mid X) \leq 1 \mid \psi_0)$, *bias in favour* $= \sup_{\psi: d_\Psi(\psi, \psi_0) \geq \delta} M(RB_\Psi(\psi_0 \mid X) \geq 1 \mid \psi)$

So, the bias against is the prior probability of not getting evidence in favour of H_0 when it is true while the bias in favour is the maximum prior probability of not getting evidence against H_0 when it is meaningfully false. Note also that these bias calculations only depend on the principle of evidence and not the specific valid measure of evidence used.

Now consider an example of perhaps the sharpest disagreement between Bayesian and frequentist approaches.

Example 2. *Jeffreys-Lindley paradox.*

For this $\bar{x} \sim N(\mu, \sigma_0^2/n)$ and $\mu \sim N(\mu_0, \tau_0^2)$ with the problem being to assess the hypothesis $H_0: \mu = \mu_0$. In this case the relative belief ratio and the Bayes factor, defined via the spiked prior, are equal. It is the case that $RB(\mu_0 \mid x) \to \infty$ as $\tau_0^2 \to \infty$, so with an increasingly diffuse prior, it seems that there is overwhelming evidence in favour of H_0. The standard p-value for this problem is given by the two-sided z-test and suppose that the associated p-value satisfies

$$p\text{-}value = 2\left(1 - \Phi\left(\frac{\sqrt{n}|\bar{x} - \mu_0|}{\sigma_0}\right)\right) \approx 0.$$

There is then a sharp discrepancy between the frequentist and the Bayesian using a diffuse prior. Still, when the strength of the evidence is measured, $\Pi_\Psi\left(RB(\mu \mid x) \leq RB\left(\mu_0 \middle| x\right) \middle| x\right) \to p\text{-}value$ as $\tau_0^2 \to \infty$, so the evidence becomes weak even though the relative belief ratio is large.

The weakness of the evidence in favour is a partial response to this problem, but it is the bias in favour calculation that resolves it for us. For it is easy to show that the bias against converges to 0 and the bias in favour converges to 1 as $\tau_0^2 \to \infty$. In other words, if a very diffuse prior is used it is virtually certain that evidence in favour will be obtained even when the hypothesis is false. This tell us that there is little point in conducting or reporting the results of such a study. But there is a way out of this dilemma because it can also be shown that for a fixed prior both biases converge to 0 as the amount of data increases and so bias can be controlled by design. The lesson in this is that we should choose priors via elicitation, don't choose arbitrarily diffuse priors in an attempt to be "conservative", and design to avoid bias.

There is another interesting aspect that arises from this simple example. Certainly, it is obvious that *p-value* is not a valid measure of evidence, at least according to our definition. Some simple algebra shows, however, that with $z_0 = |\bar{x} - \mu_0|/\sigma_0$ and $r = n\tau_0^2/\sigma_0^2$ then the difference in tail probabilities

$$p* = 2\big(1 - \Phi(z_0)\big) \; - \; 2(1 - \Phi(\left\{\log(1+r) + \frac{z_0^2}{1+r}\right\}^{\frac{1}{2}})$$

is a valid measure of evidence via the cut-off 0. For example, with *p-value* $= 0.05$, Table 1 shows that such a value is sometime evidence against and sometimes is evidence in favour and this depends primarily on the value of r.

Table 1: Evidence as determined by the two-sided z-test p-value with the location-normal.

n	r	p_*	evidence
10	1	-0.047	against
10	100	0.019	in favour
10	500	0.037	in favour
50	50	0.005	in favour

Also, the cut-off for determining evidence for or against occurs at $z_0^2 = [(1+r)\log(1+r)]/r \to \infty$ as $r \to \infty$, which implies that the associated *p-value* $\to 0$. This general conclusion is not overly reliant on the particular prior chosen, see [21], and suggests that smaller p-values are necessary to determine evidence against as n increases, since r increases with n.

Interesting results are also obtained when considering bias in the estimation problem.

E: *bias against* $= E_{\Pi_\Psi}\left(M(\psi \notin Pl_\psi(X) \mid \psi)\right) = E_{\Pi_\Psi}\big(M(RB_\psi(\psi \mid X) \le 1 \mid \psi)\big)$

bias in favour $= E_{\Pi_\Psi}\left(sup_{\psi: d_\Psi(\psi,\psi_0) \ge \delta} M(\psi_0 \notin Im_\psi(X) \mid \psi)\right) = E_{\Pi_\Psi}\left(sup_{\psi: d_\Psi(\psi,\psi_0) \ge \delta} M(RB_\psi(\psi_0 \mid X) \ge 1 \mid \psi)\right)$

So, the *bias against* for estimation is the prior probability that true value is not in the plausible region $Pl_\psi(x)$ which implies that $1 - E_{\Pi_\Psi}\big(M(\psi \notin Pl_\psi(X) \mid \psi)\big)$ is the prior coverage probability (confidence) of $Pl_\psi(x)$. The *bias in favour* is an upper bound on the prior probability that a meaningfully false value is not in the *implausible region*, the set of values for which there is evidence against. For the *bias against* typically there exists the value $\psi_0 = \text{argsup}\, M(RB_\psi(\psi \mid X) \le 1 \mid \psi)$ so $M(\psi \in Pl_\psi(X) \mid \psi) \ge 1 - M(RB_\psi(\psi_0 \mid X) \le 1 \mid \psi_0)$ and we have a lower bound on the (frequentist) confidence of $Pl_\psi(x)$. This allows us to give a frequentist interpretation to the plausible region in terms of coverage probabilities. This is an exact frequentist coverage probability when the object of interest Ψ is the full model parameter ɵ and, for a general Ψ, has a similar interpretation for frequentist confidence regions with mixed models. In effect, this is an exact lower bound on the frequentist coverage probability for the model given by $\{M(\cdot \mid \psi): \psi \in \Psi\}$ as the nuisance parameters have been integrated out. For the *bias in favour* the interpretation is similar to the idea of measuring the accuracy of a confidence region via the probability of covering a false value. Again, both biases converge to 0 with increasing amounts of data and so can be measured and controlled as part of ensuring that a statistical study produces reliable results. It is to be noted too that these frequentist properties hold for any prior.

Various optimality properties, in terms of bias, are established in [21] for this approach to characterizing/measuring statistical evidence. A satisfying overall conclusion from this is that, if bias assessments are held as being essential, then there are complementary roles for frequentism and Bayes. The role of bias is now considered for Fieller's problem.

Example 1. *Fieller's problem (continued).*

First consider the *bias against* measures. For $H_0: \psi_0 = 2$, then *bias against* $= 0.04$ so there is no bias against H_0. Also, $\psi_0 = \text{argsup}_\psi M(RB_\psi(\psi \mid X) \le 1 \mid \psi) = 2$ and $M(RB_\psi(\psi \mid X) \le 1 \mid 2) = 0.04$ and so the coverage probability of $Pl_\psi(x)$ is at least 0.96. Therefore, the desired coverage probability has been obtained and there will never be an absurd region reported. Figure 2 is a plot of $M(RB_\psi(2 \mid X) \le 1 \mid \psi)$ as a function of ψ. Note that with $m = n = 20$, then $M(RB_\psi(\psi \mid X) \le 1 \mid 2) = 0.028$, so $Pl_\psi(x)$ would then be a 0.972-confidence interval for ψ.

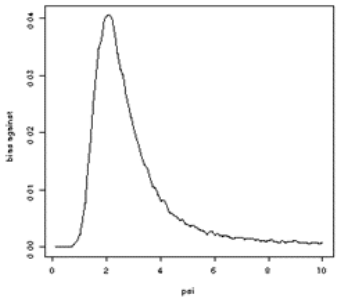

Fig. 2: Bias against ψ as a function of ψ in Example 1.

Tables 2 and 3 contain values of some of the *bias in favour* values for the **H** and **E** problems, respectively. It can be seen that obtaining small values for these quantities is much more demanding than controlling the *bias against*. This is strongly affected by the chosen δ as smaller values require ever more data to achieve a desired value for the bias in favour.

Table 2: Bias in favour for **H** in Fieller's problem.

$m = n$	*bias in favour* for $H_0: \Psi(\theta) = 2$
10	0.797
20	0.682
100	0.117
200	0.012

Table 3: Bias in favour for **E** in Fieller's problem.

$m = n$	*bias in favour* for estimating ψ
10	0.807
20	0.671
100	0.277
200	0.161

7. Conclusions

This paper has reviewed an approach to statistical reasoning that is an attempt to satisfy the desiderata for such a theory discussed at the beginning of the paper. The focus here has been on inferences driven primarily by the principle of evidence, which provides a characterization of statistical evidence, and the measurement and control of bias, again through the same characterization of statistical evidence. A happy consequence of this approach is a degree of unity between a Bayesian approach, which the inferences are as they involve a prior, and a frequentist approach, which here is primarily concerned with the reliability of the inferences obtained.

A fuller discussion would include much more commentary of what are conceived as the necessary steps involved in carrying out a statistical study. Based on our discussion here, however, the following steps seem necessary.

1. Choose a model $\{f_\theta: \theta \in \Theta\}$.
2. Elicit a prior π.
3. Specify the meaningful difference δ for the characteristic of interest $\psi = \Psi(\theta)$.
4. Measure the biases and determine the appropriate amount of data x to collect.
5. Check the model against the data and modify it if necessary.
6. Check the prior against the data and modify it if necessary.
7. Make inferences about ψ based on the principle of evidence.

References

[1] A. Birnbaum, "On the foundations of statistical inference (with discussion)". J. Amer. Stat. Assoc., 57, 269-332, 1962.
[2] R. Royall, *Statistical Evidence: A Likelihood Paradigm*. Chapman and Hall/CRC Press, 1997.

[3] M. Evans, *Measuring Statistical Evidence Using Relative Belief*. Chapman and Hall/CRC Press, 2015.
[4] G. Shafer, *A Mathematical Theory of Evidence*. Princeton University Press: Princeton, NJ, USA, 1976.
[5] B. Thompson, *The Nature of Statistical Evidence*. Lecture Notes in Statistics 189, Springer, 2007.
[6] M. Aitkin, *Statistical Inference: an Integrated Bayesian/Likelihood Approach*, Chapman and Hall/CRC Press 2010.
[7] R. Morey, J.-W. Romeijn and J. Rouder, "The philosophy of Bayes factors and the quantification of statistical evidence". J. Mathematical Psychology, 72, 6-18, 2016.
[8] V. J. Vieland and S-J. Seok, "Statistical evidence measured on a properly calibrated scale for multinomial hypothesis comparisons". Entropy, 18, 114, doi.org/10.3390/e18040114, 2016.
[9] W. Salmon, "Confirmation". Scientific American, 228, 5, 75-81, 1973.
[10] M. Evans and H. Moshonov, "Checking for prior-data conflict". Bayesian Analysis, 1, 4, 893-914, 2006.
[11] M. Evans and G-H. Jang, "Weak informativity and the information in one prior relative to another". Stat. Sci. 26, 3, 423-439, 2011.
[12] E. Boring, "Mathematical vs statistical significance". Psychological Bulletin, 16, 10, 335-338, 1919.
[13] K. Popper, *The Logic of Scientific Discovery*. Harper Torchbooks, 1968.
[14] R. C. Geary, "The frequency distribution of the quotient of two normal variates". J. Royal Statist. Soc. 97, 442-446, 1930.
[15] E. C. Fieller, "Some problems in interval estimation". J. of the Royal Statist. Soc. B, 16, 2, 175-186, 1954.
[16] D. V. Hinkley, "On the ratio of two correlated normal random variables", Biometrika 56, 635-639, 1969.
[17] T. Pham-Gia, N. Turkkan, and E. Marchand, " Density of the ratio of two normal random variables and applications". Comm. in Statist., Theory and Methods, 35, 9, 1569-1591, 2006.
[18] M. Ghosh, G. S. Datta, D. Kim and T. J. Sweeting, " Likelihood-based inference for the ratios of regression coefficients in linear models", Ann. Inst. Stat. Math., 58, 2006.
[19] D. A. S. Fraser, N. Reid, and W. Lin "When should modes of inference disagree? Some simple but challenging examples". Ann. Appl. Stat. 12 (2) 750-770, 2018.
[20] M. Evans, M. Liu, M. Moon, S. Sixta, S. Wei and S. Yang, "On some problems of confidence region construction". Manuscript, 2021.
[21] M. Evans, and Y. Guo, "Measuring and controlling bias for some Bayesian inferences and the relation to frequentist criteria". Entropy, 23(2), 190, doi:10.3390/e23020190, 2021.