Review of Martin, Ryan (2019) False confidence, non-additive beliefs, and valid statistical inference. International Journal of Approximate Reasoning, 113, 39-73.

by
Michael Evans
University of Toronto

This is a substantial article on an important topic, namely, the foundations of statistics. It needs to be stated up front that the reviewer advocates a very different approach to the foundations than is being put forward in this paper but there are still many points of agreement.

To comprehend the need for such foundations it is necessary to understand why the subject of statistics exists in the first place. In our opinion this arises due to two general problems that arise in applied science. Consider a context where there is a real-world object of interest ψ whose specific value is unknown. Furthermore, data x has been collected which is felt to be relevant towards addressing one or both of the following problems: (i) provide an estimate $\psi(x)$ of the true value of ψ together with an assessment of the accuracy of $\psi(x)$ and (ii) provide a measure of the evidence in x, either in favor of or against, a hypothesized value ψ_0 for ψ together with an assessment of how strong this evidence is. A theory of statistical reasoning is then required that answers (i) and (ii) in a fully logical and well-supported way.

Part of the thesis of this paper is that, not only is this an important problem that cannot be ignored, but none of the current familiar candidates for such a theory are satisfactory and so a new approach is put forward as a possible solution. The proposal is interesting, and there is agreement concerning the need for new directions, but in the end there is much that needs to be settled before one could claim that the approach advocated in the paper achieves this goal.

One constant theme in statistical research is that the answers to (i) and (ii) should be based upon the *evidence* in the data. The goal for a theory then is to express this evidence as unambiguously and convincingly as possible. So in considering a theory it is relevant to ask how the concept of evidence is being handled. The treatment in this paper could be considered as a generalization of the fiducial approach where the expression of the evidence results in the provision of a probability distribution for the unknown ψ on the space Ψ of possible values. The probabilities given by this distribution are measuring the degrees of belief concerning the truth of the values of ψ. Similarly, some treatments of Bayesian inference consider the posterior distribution of ψ to be the expression of the evidence concerning ψ and for the same reason.

It is natural to ask if indeed such distributions are the proper expression of evidence. In the Bayesian context, with a proper prior, this seems wrong. This conclusion is supported by the very basic *principle of evidence*: if the posterior probability of ψ is greater (less) than the prior probability of ψ, then there is evidence in favor of (against) ψ being the true value. It seems hard to

argue with the validity of such a principle. In particular, it is based on how the data is changing beliefs, which is what evidence does, and it is not intrinsically dependent on how strong or weak our initial beliefs are. Note too that the degrees of belief approach to determining evidence requires a cut-off value and this is implicit and obvious via the principle of evidence. There is an additional role for the posterior with the principle of evidence because there is still the issue of how weak or strong our beliefs are concerning what the evidence says. These considerations lead to answers for both (i) and (ii), see Evans (2015). So evidence is measured by change in beliefs and the strength or accuracy of the inferences is assessed separately. In our view the fiducial approach and its generalizations do not treat evidence properly and in fact confound evidence with belief and these are quite different concepts.

That evidence and its strength can be fairly easily expressed in the proper Bayesian context is, in our view, a strong argument for formulating statistical problems with the basic ingredients of a Bayesian problem, namely, the statistical model and the prior. In fact we would take this as virtually a definition of what is meant by a core statistical problem. Of course, not all problems faced by a statistician will satisfy these requirements, but there is tremendous value in realizing that a sound theory of statistical reasoning can indeed be developed for the core. Such a theory must also influence what is done in problems outside this domain.

The fiducial approach and its generalizations has a model as an ingredient as well as a pivotal function $U(x, \psi)$ whose distribution is known. This paper adds the specification of a random set. There are issues associated with the fiducial approach that have never been resolved and the treatment in this paper can claim to be a significant attempt at resolving some of these. For the remainder of the review I will refer to both as the fiducial approach.

No matter whether one takes a Bayesian, fiducial or some other approach there is a significant problem that needs to be addressed as part of any theory of statistical reasoning that is going to be applied to scientific problems. This concerns the choice of the ingredients necessary for the application of the theory. It needs to be noted too that these choices are always subjective. First, there needs to be an argument as to why the particular ingredients chosen are suitable. Often no such argument is put forward beyond, for example, the model being intuitively reasonable and similarly for the prior. Actually, what is needed are elicitation algorithms that, based on expert knowledge lead to such choices. Surely this is the correct approach to advocate for any problem we are going to consider core and any attempt at default choices takes us outside. Second, every ingredient that is chosen must be falsifiable in the sense that the (objective) data can indicate that a poor choice has been made. There are general procedures for checking a model and for checking for prior-data conflict but it is unclear how one checks the additional ingredients required for the fiducial approach. For example, how does one assess if a particular pivotal or random set is or is not contradicted by the data. Also, if there are multiple possible choices of these ingredients that materially affect the inferences how does one choose among them? Certainly this is a place where much more needs to be developed for

fiducial and similar considerations also apply to approaches that rely on adding a loss function as an ingredient to the problem. A theory that doesn't allow for falsifiability of the ingredients is not appropriate for scientific contexts in our view. Again the basic Bayesian approach together with the principle of evidence has distinct advantages on both criteria particularly when bias, as subsequently discussed, is taken into account.

The issue of bias in the chosen ingredients has nothing to do with whether or not they are contradicted by the data. When considering bias we are asking if the ingredients have been chosen such that the inferences drawn for (i) or (ii) are foregone conclusions. In the Bayesian formulation described here, this bias can be precisely measured as the prior probability of obtaining evidence against a ψ value when it is true and as the prior probability of obtaining evidence in favor of ψ when it is meaningfully false. If either of these prior probabilities are large, then one has to be concerned about the validity of the conclusions drawn. Fortunately there is a way to control such biases as they converge to 0 as the amount of data increases. But note that referring meaningfully to bias requires the principle of evidence. The paper under review discusses the important new concepts of false confidence and validity. These concepts are similar, at least in spirit, to bias and, as discussed in Evans and Guo (2019), this leads to connections with confidence and frequentism.

The methods discussed in this paper are interesting and novel. We could see these, or some variation thereof, being applied in what are called here noncore problems. But in the same spirit as expressed in the paper, we believe that the field of statistics needs to identify a class of core problems and an associated theory of statistical reasoning that can handle core problems in a logical and sound way. Not only does this put the field on a sound foundation but it becomes clear when we, perhaps out of necessity, stray outside so that suitable qualifications can be supplied concerning the validity of whatever answers are derived for (i) and (ii). For us being as clear as possible about the concept of statistical evidence points the way forward but there is no denying that currently there is little agreement in the statistical community about the path to take.

References

Evans, M. (2015) Measuring Statistical Evidence Using Relative Belief. Monographs on Statistics and Applied Probability 144, CRC Press, Taylor & Francis Group.

Evans, M. and Guo, Y. (2019) Measuring and controlling bias for some Bayesian inferences and the relation to frequentist criteria. arXiv:1903.01696.