Statistical Science
2014, Vol. 29, No. 2, 247–251
DOI: 10.1214/14-STS472
Main article DOI: 10.1214/13-STS457

Discussion: Foundations of Statistical Inference, Revisited

Ryan Martin and Chuanhai Liu

Abstract. This is an invited contribution to the discussion on Professor Deborah Mayo's paper, "On the Birnbaum argument for the strong likelihood principle," to appear in *Statistical Science*. Mayo clearly demonstrates that statistical methods violating the likelihood principle need not violate either the sufficiency or conditionality principle, thus refuting Birnbaum's claim. With the constraints of Birnbaum's theorem lifted, we revisit the foundations of statistical inference, focusing on some new foundational principles, the inferential model framework, and connections with sufficiency and conditioning.

Key words and phrases: Birnbaum, conditioning, dimension reduction, inferential model, likelihood principle.

1. INTRODUCTION

Birnbaum's theorem (Birnbaum, 1962) is arguably the most controversial result in statistics. The theorem's conclusion is that a framework for statistical inference that satisfies two natural conditions, namely, the sufficiency principle (SP) and the conditionality principle (CP), must also satisfy an exclusive condition, the likelihood principle (LP). The controversy lies in the fact that LP excludes all those standard methods taught in Stat 101 courses. Professor Mayo successfully refutes Birnbaum's claim, showing that violations of LP need not imply violations of SP or CP. The key to Mayo's argument is a correct formulation of CP; see also Evans (2013). Her demonstration resolves the controversy around Birnbaum and LP, helping to put the statisticians' house in order.

The controversy and confusion surrounding Birnbaum's claim has perhaps discouraged researchers from considering questions about the foundations of statistics. We view Professor Mayo's paper as an invitation for statisticians to revisit these fundamental questions, and we are grateful for the opportunity to contribute to this discussion.

Though LP no longer constrains the frequentist approach, this does not mean that pure frequentism is necessarily correct. For example, reproducibility issues[1] in large-scale studies is an indication that the frequentist techniques that have been successful in classical problems may not be appropriate for today's high-dimensional problems. We contend that something more than the basic sampling model is required for valid statistical inference, and appropriate conditioning is one aspect of this. Here we consider what a new framework, called *inferential models* (IMs), has to say concerning the foundations of statistical inference, with a focus on something more fundamental than CP and SP. For this, we begin in Section 2 with a discussion of valid probabilistic inference as motivation for the IM framework. A general efficiency principle is presented in Section 3 and we give an IM-based dimension reduction strategy that accomplishes what the classical SP and CP set out to do. Section 4 gives some concluding remarks.

Ryan Martin is Assistant Professor, Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago, 851 S. Morgan St., Chicago, Illinois 60607, USA (e-mail: rgmartin@uic.edu). Chuanhai Liu is Professor, Department of Statistics, Purdue University, 250 North University St., West Lafayette, Indiana 47907-2067, USA (e-mail: chuanhai@purdue.edu).

[1] "Announcement: Reducing our irreproducibility," *Nature* **496** (2013), DOI:10.1038/496398a.

2. VALID PROBABILISTIC INFERENCE

2.1 A Validity Principle

The sampling model for observable data X, depending on an unknown parameter θ, is the familiar starting point. Our claim is that the sampling model alone is not sufficient for valid probabilistic inference. By "probabilistic inference" we mean a framework whereby any assertion/hypothesis about the unknown parameter has a (predictive) probability attached to it after data is observed. The sampling model facilitates a comparison of the chances of the event $X = x$ on different probability spaces. For probabilistic inference, there must be a known distribution available, after data is observed. The random element that corresponds to this distribution shall be called a predictable quantity, and probabilistic inference is obtained by predicting this predictable quantity after seeing data. The probabilistic inference is "valid" if the predictive probabilities are suitably calibrated or, equivalently, have a fixed and known scale for meaningful interpretation. Without risk of confusion, we shall call the prediction of the predictable quantity valid if it admits valid probabilistic inference. We summarize this in the following *validity principle* (VP).

VALIDITY PRINCIPLE. Probabilistic inference requires associating an unobservable but predictable quantity with the observable data and unknown parameter. Probabilities to be used for inference are obtained by valid prediction of the predictable quantity.

The frequentist approach aims at developing procedures, such as confidence intervals and testing rules, having long-run frequency properties. Expressions like "95% confidence" have no predictive probability interpretation after data is observed, so frequentist methods are not probabilistic in our sense. Nevertheless, certain frequentist quantities, such as p-values, may be justifiable from a valid probabilistic inference point of view (Martin and Liu, 2014b).

When genuine prior information is available and can be summarized as a usual probability model, the corresponding Bayesian inference is both probabilistic and valid [see Martin and Liu (2014a), Remark 4]. When no genuine prior information is available, and a default prior distribution is used, the validity property is questionable. Probability matching priors, Bernstein–von Mises theorems, etc., are efforts to make posterior inference valid, in the sense above, at least approximately. The standard interpretation of these results is, for example, that Bayesian credible intervals have the nominal frequentist coverage probability asymptotically; see Fraser (2011). In that case, the remarks above concerning frequentist methods apply.

Fiducial inference was introduced by Fisher (1930) to avoid using artificial priors in scientific inference. Subsequent work includes structural inference (Fraser, 1968), the Dempster–Shafer theory of belief functions (Shafer, 1976; Dempster, 2008), generalized inference (Chiang, 2001; Weerahandi, 1993) and generalized fiducial inference (Hannig, 2009, 2013). Fiducial distributions are defined by expressing the parameter as a data-dependent function of a pivotal quantity. This results in a bona fide posterior distribution only in Fraser's structural models and, in those cases, it corresponds to a Bayesian posterior (Lindley, 1958; Taraldsen and Lindqvist, 2013). Therefore, the fiducial distribution is meaningful when the corresponding Bayesian prior is meaningful in the sense above. More on fiducial from the IM perspective is given below.

2.2 IM Framework

The IM framework, proposed recently by Martin and Liu (2013), has its roots in fiducial and Dempster–Shafer theory; see also Martin, Zhang and Liu (2010). At a fundamental level, the IM approach is driven by VP. Here is a quick overview.

Write the sampling model/data-generating mechanism as

$$X = a(\theta, U), \quad U \sim \mathsf{P}_U, \tag{1}$$

where $X \in \mathbb{X}$ is the observable data, $\theta \in \Theta$ is the unknown parameter, and $U \in \mathbb{U}$ is an unobservable auxiliary variable with known distribution P_U. Following VP, the goal is to use the data X and the distribution for U for meaningful probabilistic inference on θ without assuming a prior. The following three steps describe the IM construction.

A-STEP. Associate the observed data $X = x$, the parameter and the auxiliary variable via (1) and construct the set-valued mapping, given by

$$\Theta_x(u) = \{\theta : x = a(\theta, u)\}, \quad u \in \mathbb{U}.$$

The fiducial approach considers the distribution of $\Theta_x(U)$ as a function of $U \sim \mathsf{P}_U$. The IM framework, on the other hand, predicts the unobserved U using a random set.

P-STEP. Predict the unobservable U with a random set $\mathcal{S}$. The distribution $\mathsf{P}_{\mathcal{S}}$ of $\mathcal{S}$ is required to be valid in the sense that

$$f(U) \geq_{\text{st}} \mathsf{Unif}(0, 1), \tag{2}$$

where $f(u) = \mathsf{P}_{\mathcal{S}}(\mathcal{S} \ni u)$ and $\geq_{\text{st}}$ means "stochastically no smaller than."

C-STEP. Combine $\Theta_x(\cdot)$ and $\mathcal{S}$ as $\Theta_x(\mathcal{S}) = \bigcup_{u \in \mathcal{S}} \Theta_x(u)$. For a given assertion $A \subseteq \Theta$, evaluate the evidence in x for and not against the claim "$\theta \in A$" via the belief and plausibility functions:

$$\mathsf{bel}_x(A) = \mathsf{P}_{\mathcal{S}}\{\Theta_x(\mathcal{S}) \subseteq A\},$$
$$\mathsf{pl}_x(A) = \mathsf{P}_{\mathcal{S}}\{\Theta_x(\mathcal{S}) \cap A \neq \varnothing\}.$$

In cases where $\Theta_x(\mathcal{S})$ is empty with positive $\mathsf{P}_{\mathcal{S}}$-probability, some adjustments to the above formulas are needed; see Martin and Liu (2013).

The IM output is the pair of functions $(\mathsf{bel}_x, \mathsf{pl}_x)$ and, when applied to an assertion A about the parameter θ of interest, these provide a (personal) probabilistic summary of the evidence in data $X = x$ supporting the truthfulness of A. Property (2) guarantees that the IM output is valid. Our focus in the remainder of the discussion will be on the A-step and, in particular, auxiliary variable dimension reduction. Such concerns about dimensionality are essential for efficient inference on θ. These are also closely tied to the classical ideas of sufficiency and conditionality.

We end this section with a few remarks on fiducial. If one takes the predictive random set $\mathcal{S}$ in the P-step as a singleton, that is, $\mathcal{S} = \{U\}$, where $U \sim \mathsf{P}_U$, then bel_x and pl_x are equal and equal to the fiducial distribution. In this sense, fiducial provides probabilistic inference. However, the singleton predictive random set is not valid in the sense of (2), so fiducial inference is generally not valid, violating the second part of VP. One can also reconstruct the fiducial distribution by choosing a valid predictive random set $\mathcal{S}$ so that $\mathsf{bel}_x(A)$ equals the fiducial probability of A for all suitable $A \subseteq \Theta$. But this would generally require that $\mathcal{S}$ depend on the observed $X = x$, and the resulting inference suffers from a selection bias, or double-use of the data, resulting in unjustifiably large belief probabilities.

3. EFFICIENCY AND DIMENSION REDUCTION

3.1 An Efficiency Principle

It is natural to strive for efficient statistical inference. In the context of IMs, we want $\mathsf{pl}_X(A)$ to be as stochastically small as possible, as a function of X, when the assertion A about θ is false. To connect this to classical efficiency, $\mathsf{pl}_X(A)$ can be interpreted like the p-value for testing $H_0 : \theta \in A$, so stochastically small plausibility when A is false corresponds to the high power of the test. We state the following *efficiency principle* (EP).

EFFICIENCY PRINCIPLE. Subject to the validity constraint, probabilistic inference should be made as efficient as possible.

EP is purposefully vague: it allows for a variety of techniques to be employed to increase efficiency. The next section discusses one important technique related to auxiliary variable dimension reduction.

3.2 Improved Efficiency via Dimension Reduction

In the classical setting, sufficiency reduces the data to a good summary statistic. In the IM context, however, the dimension of the auxiliary variable, not the data, is of primary concern. For example, in the case of iid sampling, the dimension of U is the same as that of X, which is usually greater than that of θ. In such cases, it is inefficient to predict a high-dimensional auxiliary variable for inference on a lower dimensional parameter. The idea is to reduce the dimension of U to that of θ. This auxiliary variable dimension reduction will indirectly result in some transformation of the data.

How to reduce the dimension of U? We seek a new auxiliary variable V, of the same dimension of θ, such that the baseline association (1) can be rewritten as

$$T(X) = b(\theta, V), \quad V \sim \mathsf{P}_V, \tag{3}$$

for some functions T and b, and distribution P_V. Here P_V may actually depend on some features of the data X. Such a dimension reduction is general, but Martin and Liu (2014a) consider an important case, which we summarize here. Suppose we have two one-to-one mappings, $x \mapsto (T(x), H(x))$ and $u \mapsto (\tau(u), \eta(u))$, with the requirement that $\eta(U) = H(X)$. Since $H(X)$ is observable, so too must be the feature $\eta(U)$ of U. This point has two important consequences: first, a feature of U that is observed need not be predicted, hence a dimension reduction; second, the feature of U that is observed naturally provides some information about the part that remains unobserved, so conditioning should help improve prediction. By construction, the baseline association (1) is equivalent to

$$T(X) = b\big(\theta, \tau(U)\big) \quad \text{and} \quad H(X) = \eta(U), \tag{4}$$

and this suggests an association of the form (3), where $V = \tau(U)$ and P_V is the conditional distribution of $\tau(U)$ given $\eta(U) = H(X)$.

It remains to discuss how the dimension reduction strategy described above related to EP. The following theorem gives one relatively simple illustration of the improved efficiency via dimension reduction.

THEOREM 1. *Suppose that the baseline association* (1) *can be rewritten as* (4) *and that* $\tau(U)$ *and* $\eta(U)$ *are independent. Then inference based on* $T(X) = b(\theta, \tau(U))$ *alone, by a valid prediction of* $\tau(U)$*, ignoring* $\eta(U)$*, is at least as efficient as inference from* (4) *by a valid prediction of* $(\tau(U), \eta(U))$.

See the corollary to Proposition 1 in Martin and Liu (2014a, full-length version); see also Liu and Martin (2015). Therefore, reducing the dimension of the auxiliary variable cannot make inference less efficient. The point in that paper is that reducing the dimension actually improves efficiency, hence EP.

In the standard examples, for example, regular exponential families, our dimension reduction above corresponds to classical sufficiency; see Example 1. Outside the standard examples, the IM dimension reduction gives something different from sufficiency, in particular, the former often leads directly to further dimension reduction compared to the latter; see Example 2. That the IM dimension reduction naturally contains some form of conditioning is an advantage. The absence of conditioning in the standard definition of sufficiency is one possible reason why conditional inference has yet to become part of the mainstream. The IM framework also has advantages beyond dimension reduction and conditioning. In particular, the IM output gives valid prior-free probabilistic inference on θ.

3.3 "Dimension Reduction Entails SP and CP"

This section draws some connections between the IM dimension reduction above and SP and CP. First, it is clear that following the auxiliary variable dimension reduction strategy described above entails CP. In the Cox example, the randomization that determines which measurement instrument will be used corresponds to an auxiliary variable whose value is observed completely. So, our auxiliary variable dimension reduction strategy implies conditioning on the actual instrument used, hence CP. For SP, Theorem 1 gives some insight. That is, when a sufficient statistic has dimension the same as θ, one can take $T(X)$ as that sufficient statistic and select independent $\tau(U)$ and $\eta(U)$. In general, our dimension reduction and efficiency considerations are more meaningful than sufficiency and SP. The examples below illustrate this point further.

EXAMPLE 1. Suppose X_1, X_2 are independent $\mathsf{N}(\theta, 1)$ samples, and write the association as $X_i = \theta + U_i$, where U_1, U_2 are independent standard normal. In this case, there are lots of candidate mappings (τ, η) to rewrite the baseline association in the form (4). Two choices are $\{\tau(u) = u_1, \eta(u) = u_2 - u_1\}$ and $\{\tau(u) = \bar{u}, \eta(u) = (u_1 - \bar{u}, u_2 - \bar{u})\}$. At first look, the second choice, corresponding to sufficiency, seems better than the first. However, the dimension-reduced associations (3) based on these two choices are exactly the same. This means, first, there is nothing special about sufficiency in light of proper conditioning (Evans, Fraser and Monette, 1986; Fraser, 2004). Second, it suggests that, at least in simple problems, the dimension-reduced association (3) does not depend on the choice of (τ, η), that is, it only depends on the sufficient statistic, hence SP. The message here holds more generally, though a rigorous formulation remains to be worked out.

EXAMPLE 2. Consider independent exponential random variables X_1, X_2, the first with mean θ and the second with mean θ^{-1}. In this case, the minimal sufficient statistic, (X_1, X_2), is two-dimensional while the parameter is one-dimensional. Martin and Liu (2014a) take the baseline association as $X_1 = \theta U_1$ and $X_2 = \theta^{-1} U_2$, where U_1, U_2 are independent standard exponential. They employ a novel partial differential equations technique to identify a function η of (U_1, U_2) whose value is fully observed, so that only a scalar auxiliary variable needs to be predicted. Their solution is equivalent to that based on the conditional distribution of the maximum likelihood estimate given an ancillary statistic (Fisher, 1973; Ghosh, Reid and Fraser, 2010). The message here is that the IM-based auxiliary variable dimension reduction strategy does something similar to the classical strategy of conditioning on ancillary statistics, but it does so in a mostly automatic way.

4. CONCLUDING REMARKS

Professor Mayo is to be congratulated for her contribution. Besides resolving the controversy surrounding Birnbaum's theorem, her paper is an invitation for a fresh discussion on the foundations of statistical inference. Though LP no longer constrains the frequentist approach, we have argued here that something more than the basic sampling model is required for valid statistical inference. The IM framework features the prediction of unobserved auxiliary variables as this "something more," and the idea of reducing the dimension of the auxiliary variable before prediction leads to improved efficiency, accomplishing what SP and CP are meant to do. We expect further developments for and from IMs in years to come.

ACKNOWLEDGMENTS

Supported in part by NSF Grants DMS-10-07678, DMS-12-08833 and DMS-12-08841.

REFERENCES

BIRNBAUM, A. (1962). On the foundations of statistical inference. *J. Amer. Statist. Assoc.* **57** 269–326. MR0138176

CHIANG, A. K. L. (2001). A simple general method for constructing confidence intervals for functions of variance components. *Technometrics* **43** 356–367. MR1943189

DEMPSTER, A. P. (2008). The Dempster–Shafer calculus for statisticians. *Internat. J. Approx. Reason.* **48** 365–377. MR2419025

EVANS, M. (2013). What does the proof of Birnbaum's theorem prove? *Electron. J. Stat.* **7** 2645–2655. MR3121626

EVANS, M. J., FRASER, D. A. S. and MONETTE, G. (1986). On principles and arguments to likelihood. *Canad. J. Statist.* **14** 181–199. MR0859631

FISHER, R. A. (1930). Inverse probability. *Math. Proc. Cambridge Philos. Soc.* **26** 528–535.

FISHER, R. A. (1973). *Statistical Methods and Scientific Inference*, 3rd ed. Hafner Press, New York.

FRASER, D. A. S. (1968). *The Structure of Inference*. Wiley, New York. MR0235643

FRASER, D. A. S. (2004). Ancillaries and conditional inference. *Statist. Sci.* **19** 333–369. MR2140544

FRASER, D. A. S. (2011). Is Bayes posterior just quick and dirty confidence? *Statist. Sci.* **26** 299–316. MR2918001

GHOSH, M., REID, N. and FRASER, D. A. S. (2010). Ancillary statistics: A review. *Statist. Sinica* **20** 1309–1332. MR2777327

HANNIG, J. (2009). On generalized fiducial inference. *Statist. Sinica* **19** 491–544. MR2514173

HANNIG, J. (2013). Generalized fiducial inference via discretization. *Statist. Sinica* **23** 489–514. MR3086644

LINDLEY, D. V. (1958). Fiducial distributions and Bayes' theorem. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **20** 102–107. MR0095550

LIU, C. and MARTIN, R. (2015). *Inferential Models*: *Reasoning with Uncertainty*. *Monographs in Statistics and Applied Probability Series*. Chapman & Hall, London. To appear.

MARTIN, R. and LIU, C. (2013). Inferential models: A framework for prior-free posterior probabilistic inference. *J. Amer. Statist. Assoc.* **108** 301–313. MR3174621

MARTIN, R. and LIU, C. (2014a). Conditional inferential models: Combining information for prior-free probabilistic inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* To appear. Full-length preprint version at arXiv:1211.1530. DOI:10.1111/rssb.12070

MARTIN, R. and LIU, C. (2014b). A note on p-values interpreted as plausibilities. *Statist. Sinica*. To appear. Available at arXiv:1211.1547. DOI:10.5705/ss.2013.087

MARTIN, R., ZHANG, J. and LIU, C. (2010). Dempster–Shafer theory and statistical inference with weak beliefs. *Statist. Sci.* **25** 72–87. MR2741815

SHAFER, G. (1976). *A Mathematical Theory of Evidence*. Princeton Univ. Press, Princeton, NJ. MR0464340

TARALDSEN, G. and LINDQVIST, B. H. (2013). Fiducial theory and optimal inference. *Ann. Statist.* **41** 323–341. MR3059420

WEERAHANDI, S. (1993). Generalized confidence intervals. *J. Amer. Statist. Assoc.* **88** 899–905. MR1242940