Statistical Science
2014, Vol. 29, No. 2, 227–239
DOI: 10.1214/13-STS457
© Institute of Mathematical Statistics, 2014

On the Birnbaum Argument for the Strong Likelihood Principle[1]

Deborah G. Mayo

Abstract. An essential component of inference based on familiar frequentist notions, such as p-values, significance and confidence levels, is the relevant sampling distribution. This feature results in violations of a principle known as the strong likelihood principle (SLP), the focus of this paper. In particular, if outcomes $\mathbf{x}^*$ and $\mathbf{y}^*$ from experiments E_1 and E_2 (both with unknown parameter θ) have different probability models $f_1(\cdot)$, $f_2(\cdot)$, then even though $f_1(\mathbf{x}^*;\theta) = cf_2(\mathbf{y}^*;\theta)$ for all θ, outcomes $\mathbf{x}^*$ and $\mathbf{y}^*$ may have different implications for an inference about θ. Although such violations stem from considering outcomes other than the one observed, we argue this does not require us to consider experiments other than the one performed to produce the data. David Cox [*Ann. Math. Statist.* **29** (1958) 357–372] proposes the Weak Conditionality Principle (WCP) to justify restricting the space of relevant repetitions. The WCP says that once it is known which E_i produced the measurement, the assessment should be in terms of the properties of E_i. The surprising upshot of Allan Birnbaum's [*J. Amer. Statist. Assoc.* **57** (1962) 269–306] argument is that the SLP appears to follow from applying the WCP in the case of mixtures, and so uncontroversial a principle as sufficiency (SP). But this would preclude the use of sampling distributions. The goal of this article is to provide a new clarification and critique of Birnbaum's argument. Although his argument purports that [(WCP and SP) entails SLP], we show how data may violate the SLP while holding both the WCP and SP. Such cases also refute [WCP entails SLP].

Key words and phrases: Birnbaumization, likelihood principle (weak and strong), sampling theory, sufficiency, weak conditionality.

1. INTRODUCTION

It is easy to see why Birnbaum's argument for the strong likelihood principle (SLP) has long been held as a significant, if controversial, result for the foundations of statistics. Not only do all of the familiar frequentist error-probability notions, p-values, significance levels and so on violate the SLP, but the Birnbaum argument purports to show that the SLP follows from principles that frequentist sampling theorists accept:

> The likelihood principle is incompatible with the main body of modern statistical theory and practice, notably the Neyman–Pearson theory of hypothesis testing and of confidence intervals, and incompatible in general even with such well-known concepts as standard error of an estimate and significance level. [Birnbaum (1968), page 300.]

The incompatibility, in a nutshell, is that on the SLP, once the data $\mathbf{x}$ are given, outcomes other than $\mathbf{x}$ are irrelevant to the evidential import of $\mathbf{x}$. "[I]t is clear that reporting significance levels violates the LP [SLP], since significance levels involve averaging over sample points other than just the observed $\mathbf{x}$." [Berger and Wolpert (1988), page 105.]

Deborah G. Mayo is Professor of Philosophy, Department of Philosophy, Virginia Tech, 235 Major Williams Hall, Blacksburg, Virginia 24061, USA (e-mail: mayod@vt.edu).

[1] Discussed in 10.1214/14-STS470, 10.1214/14-STS471, 10.1214/14-STS472, 110.1214/14-STS473, 10.1214/14-STS474 and 10.1214/14-STS475; rejoinder at 10.1214/14-STS482.

1.1 The SLP and a Frequentist Principle of Evidence (FEV)

Birnbaum, while responsible for this famous argument, rejected the SLP because "the likelihood concept cannot be construed so as to allow useful appraisal, and thereby possible control, of probabilities of erroneous interpretations" [Birnbaum (1969), page 128]. That is, he thought the SLP at odds with a fundamental frequentist principle of evidence.

Frequentist Principle of Evidence (general): Drawing inferences from data requires considering the relevant error probabilities associated with the underlying data generating process.

David Cox intended the central principle invoked in Birnbaum's argument, the Weak Conditionality Principle (WCP), as one important way to justify restricting the space of repetitions that are relevant for informative inference. Implicit in this goal is that the role of the sampling distribution for informative inference is not merely to ensure low error rates in repeated applications of a method, but to avoid misleading inferences in the case at hand [Mayo (1996); Mayo and Spanos (2006, 2011); Mayo and Cox (2010)].

To refer to the most familiar example, the WCP says that if a parameter of interest θ could be measured by two instruments, one more precise then the other, and a randomizer that is utterly irrelevant to θ is used to decide which instrument to use, then, once it is known which experiment was run and its outcome given, the inference should be assessed using the behavior of the instrument actually used. The convex combination of the two instruments, linked via the randomizer, defines a mixture experiment, E_{mix}. According to the WCP, one should condition on the known experiment, even if an unconditional assessment improves the long-run performance [Cox and Hinkley (1974), pages 96–97].

While conditioning on the instrument actually used seems obviously correct, nothing precludes the Neyman–Pearson theory from choosing the procedure "which is best on the average over both experiments" in E_{mix} [Lehmann and Romano (2005), page 394]. They ask the following: "for a given test or confidence procedure, should probabilities such as level, power, and confidence coefficient be calculated conditionally, given the experiment that has been selected, or unconditionally?" They suggest that "[t]he answer cannot be found within the model but depends on the context" (ibid). The WCP gives a rationale for using the conditional appraisal in the context of informative parametric inference.

1.2 What Must Logically Be Shown

However, the upshot of the SLP is to claim that the sampling theorist must go all the way, as it were, given a parametric model. If she restricts attention to the experiment producing the data in the mixture experiment, then she is led to consider just the data and not the sample space, once the data are in hand. While the argument has been stated in various forms, the surprising upshot of all versions is that the SLP appears to follow from applying the WCP in the case of mixture experiments, and so uncontroversial a notion as sufficiency (SP). "Within the context of what can be called classical frequency-based statistical inference, Birnbaum (1962) argued that the conditionality and sufficiency principles imply the [strong] likelihood principle" [Evans, Fraser and Monette (1986), page 182].

Since the challenge is for a sampling theorist who holds the WCP, it is obligatory to consider whether and how such a sampling theorist can meet it. While the WCP is not itself a theorem in a formal system, Birnbaum's argument purports that the following is a theorem:

[(WCP and SP) entails SLP].

If true, any data instantiating both WCP and SP could not also violate the SLP, on pain of logical contradiction. We will show how data may violate the SLP while still adhering to both the WCP and SP. Such cases also refute [WCP entails SLP], making our argument applicable to attempts to weaken or remove the SP. Violating SLP may be written as not-SLP.

We follow the formulations of the Birnbaum argument given in Berger and Wolpert (1988), Birnbaum (1962), Casella and Berger (2002) and Cox (1977). The current analysis clarifies and fills in important gaps of an earlier discussion in Mayo (2010), Mayo and Cox (2011), and lets us cut through a fascinating and complex literature. The puzzle is solved by adequately stating the WCP and keeping the meaning of terms consistent, as they must be in an argument built on a series of identities.

1.3 Does It Matter?

On the face of it, current day uses of sampling theory statistics do not seem in need of going back 50 years to tackle a foundational argument. This may be so, but only if it is correct to assume that the Birnbaum argument is flawed somewhere. Sampling theorists who feel unconvinced by some of the machinations of the argument must admit some discomfort at the lack of resolution of the paradox. If one cannot show the relevance of error probabilities and sampling distributions to inferences once the data are in hand, then the uses of frequentist sampling theory, and resampling methods, for inference purposes rest on shaky foundations.

The SLP is deemed of sufficient importance to be included in textbooks on statistics, along with a version of Birnbaum's argument that we will consider:

> It is not uncommon to see statistics texts argue that in frequentist theory one is faced with the following dilemma: either to deny the appropriateness of conditioning on the precision of the tool chosen by the toss of a coin, or else to embrace the strong likelihood principle, which entails that frequentist sampling distributions are irrelevant to inference once the data are obtained. This is a false dilemma. ...The 'dilemma' argument is therefore an illusion. [Cox and Mayo (2010), page 298.]

If we are correct, this refutes a position that is generally presented as settled in current texts. But the illusion is not so easy to dispel, thus this paper.

Perhaps, too, our discussion will illuminate a point of agreement between sampling theorists and contemporary nonsubjective Bayesians who concede they "have to live with some violations of the likelihood and stopping rule principles" [Ghosh, Delampady and Sumanta (2006), page 148], since their prior probability distributions are influenced by the sampling distribution. "This, of course, does not happen with subjective Bayesianism. ...the objective Bayesian responds that objectivity can only be defined relative to a frame of reference, and this frame needs to include the goal of the analysis." [Berger (2006), page 394.] By contrast, Savage stressed:

> According to Bayes's theorem, $P(\mathbf{x}|\theta)$ ... constitutes the entire evidence of the experiment ...[I]f $\mathbf{y}$ is the datum of some other experiment, and if it happens that $P(\mathbf{x}|\theta)$ and $P(\mathbf{y}|\theta)$ are proportional functions of θ (that is, constant multiples of each other), then each of the two data $\mathbf{x}$ and $\mathbf{y}$ have exactly the same thing to say about the value of θ. [Savage (1962a), page 17, using θ for his λ and P for Pr.]

2. NOTATION AND SKETCH OF BIRNBAUM'S ARGUMENT

2.1 Points of Notation and Interpretation

Birnbaum focuses on informative inference about a parameter θ in a given model M, and we retain that context. The argument calls for a general term to abbreviate: the inference implication from experiment E and result $\mathbf{z}$, where E is an experiment involving the observation of $\mathbf{Z}$ with a given distribution $f(\mathbf{z};\theta)$ and a model M. We use the following:

> $\mathsf{Infr}_E[\mathbf{z}]$: the parametric statistical inference from a given or known $(E,\mathbf{z})$.

(We prefer "given" to "known" to avoid reference to psychology.) We assume relevant features of model M are embedded in the full statement of experiment E. An inference method indicates how to compute the informative parametric inference from $(E,\mathbf{z})$. Let

> $(E,\mathbf{z}) \Rightarrow \mathsf{Infr}_E[\mathbf{z}]$: an informative parametric inference about θ from given $(E,\mathbf{z})$ is to be computed by means of $\mathsf{Infr}_E[\mathbf{z}]$.

The principles of interest turn on cases where $(E,\mathbf{z})$ is given, and we reserve "$\Rightarrow$" for such cases. The abbreviation $\mathsf{Infr}_E[\mathbf{z}]$, first developed in Cox and Mayo (2010), could allude to any parametric inference account; we use it here to allow ready identification of the particular experiment E and its associated sampling distribution, whatever it happens to be. $\mathsf{Infr}_{E_{\text{mix}}}(\mathbf{z})$ is always understood as using the convex combination over the elements of the mixture.

Assertions about how inference "is to be computed given $(E,\mathbf{z})$" are intended to reflect the principles of evidence that arise in Birnbaum's argument, whether mathematical or based on intuitive, philosophical considerations about evidence. This is important because Birnbaum emphasizes that the WCP is "not necessary on mathematical grounds alone, but it seems to be supported compellingly by considerations ...concerning the nature of evidential meaning" of data when drawing parametric statistical inferences [Birnbaum (1962), page 280]. In using "$=$" we follow the common notation even though WCP is actually telling us when $\mathbf{z}_1$ and $\mathbf{z}_2$ *should* be deemed inferentially equivalent for the associated inference.

By noncontradiction, for any $(E,\mathbf{z})$, $\mathsf{Infr}_E[\mathbf{z}] = \mathsf{Infr}_E[\mathbf{z}]$. So to apply a given inference implication means its inference directive is used and not some competing directive at the same time. Two outcomes $\mathbf{z}_1$ and $\mathbf{z}_2$ will be said to have the same inference implications in E, and so are inferentially equivalent within E, whenever $\mathsf{Infr}_E[\mathbf{z}_1] = \mathsf{Infr}_E[\mathbf{z}_2]$.

2.2 The Strong Likelihood Principle: SLP

The principle under dispute, the SLP, asserts the inferential equivalence of outcomes from distinct experiments E_1 and E_2. It is a universal if-then claim:

> SLP: For any two experiments E_1 and E_2 with different probability models $f_1(\cdot)$, $f_2(\cdot)$ but with the same unknown parameter θ, if outcomes $\mathbf{x}^*$ and $\mathbf{y}^*$ (from E_1 and E_2, resp.) give rise to proportional likelihood functions ($f_1(\mathbf{x}^*; \boldsymbol{\theta}) = cf_2(\mathbf{y}^*; \boldsymbol{\theta})$ for all θ, for c a positive constant), then $\mathbf{x}^*$ and $\mathbf{y}^*$ should be inferentially equivalent for any inference concerning parameter θ.

A shorthand for the entire antecedent is that $(E_1, \mathbf{x}^*)$ is an SLP pair with $(E_2, \mathbf{y}^*)$, or just $\mathbf{x}^*$ and $\mathbf{y}^*$ form an SLP pair (from $\{E_1, E_2\}$). Assuming all the SLP stipulations, for example, that $\boldsymbol{\theta}$ is a shared parameter (about which inferences are to be concerned), we have the following:

> SLP: If $(E_1, \mathbf{x}^*)$ and $(E_2, \mathbf{y}^*)$ form an SLP pair, then $\mathsf{Infr}_{E_1}[\mathbf{x}^*] = \mathsf{Infr}_{E_2}[\mathbf{y}^*]$.

Experimental pairs E_1 and E_2 involve observing random variables $\mathbf{X}$ and $\mathbf{Y}$, respectively. Thus, $(E_2, \mathbf{y}^*)$ or just $\mathbf{y}^*$ asserts "E_2 is performed and $\mathbf{y}^*$ observed," so we may abbreviate $\mathsf{Infr}_{E_2}[(E_2, \mathbf{y}^*)]$ as $\mathsf{Infr}_{E_2}[\mathbf{y}^*]$. Likewise for $\mathbf{x}^*$. A generic $\mathbf{z}$ is used when needed.

2.3 Sufficiency Principle (Weak Likelihood Principle)

For informative inference about θ in E, if T_E is a (minimal) sufficient statistic for E, the Sufficiency Principle asserts the following:

> SP: If $T_E(\mathbf{z}_1) = T_E(\mathbf{z}_2)$, then $\mathsf{Infr}_E[\mathbf{z}_1] = \mathsf{Infr}_E[\mathbf{z}_2]$.

That is, since inference within the model is to be computed using the value of $T_E(\cdot)$ and its sampling distribution, identical values of T_E have identical inference implications, within the stipulated model. Nothing in our argument will turn on the minimality requirement, although it is common.

2.3.1 *Model checking.* An essential part of the statements of the principles SP, WCP and SLP is that the validity of the model is granted as adequately representing the experimental conditions at hand [Birnbaum (1962), page 280]. Thus, accounts that adhere to the SLP are not thereby prevented from analyzing features of the data, such as residuals, in checking the validity of the statistical model itself. There is some ambiguity on this point in Casella and Berger (2002):

> Most model checking is, necessarily, based on statistics other than a sufficient statistic. For example, it is common practice to examine residuals from a model... Such a practice immediately violates the Sufficiency Principle, since the *residuals* are not based on sufficient statistics. (Of course such a practice directly violates the [strong] LP also.) [Casella and Berger (2002), pages 295–296.]

They warn that before considering the SLP and WCP, "we must be comfortable with the model" [*ibid*, page 296]. It seems to us more accurate to regard the principles as inapplicable, rather than violated, when the adequacy of the relevant model is lacking. Applying a principle will always be relative to the associated experimental model.

2.3.2 *Can two become one?* The SP is sometimes called the weak likelihood principle, limited as it is to a single experiment E, with its sampling distribution. This suggests that if an arbitrary SLP pair, $(E_1, \mathbf{x}^*)$ and $(E_2, \mathbf{y}^*)$, could be viewed as resulting from a single experiment (e.g., by a mixture), then perhaps they could become inferentially equivalent using SP. This will be part of Birnbaum's argument, but is neatly embedded in his larger gambit to which we now turn.

2.4 Birnbaumization: Key Gambit in Birnbaum's Argument

The larger gambit of Birnbaum's argument may be dubbed *Birnbaumization*. An experiment has been run, label it as E_2, and $\mathbf{y}^*$ observed. Suppose, for the parametric inference at hand, that $\mathbf{y}^*$ has an SLP pair $\mathbf{x}^*$ in a distinct experiment E_1. Birnbaum's task is to show the two are evidentially equivalent, as the SLP requires.

We are to imagine that performing E_2 was the result of flipping a fair coin (or some other randomizer given as irrelevant to θ) to decide whether to run E_1 or E_2. Cox terms this the "enlarged experiment" [Cox (1978), page 54], E_B. We are then to define a statistic T_B that stipulates that if $(E_2, \mathbf{y}^*)$ is observed, its SLP pair $\mathbf{x}^*$ in the unperformed experiment is reported;

$$T_B(E_i, \mathbf{Z}_i) = \begin{cases} (E_1, \mathbf{x}^*), & \text{if } (E_1, \mathbf{x}^*) \text{ or } (E_2, \mathbf{y}^*), \\ (E_i, \mathbf{z}_i), & \text{otherwise.} \end{cases}$$

Birnbaum's argument focuses on the first case and ours will as well.

Following our simplifying notation, whenever E_2 is performed and $\mathbf{Y} = \mathbf{y}^*$ observed, and $\mathbf{y}^*$ is seen to admit an SLP pair, then label its particular SLP pair $(E_1, \mathbf{x}^*)$. Any problems of nonuniqueness in identifying SLP pairs are put to one side, and Birnbaum does not consider them. Thus, when $(E_2, \mathbf{y}^*)$ is observed, T_B reports it as $(E_1, \mathbf{x}^*)$. This yields the Birnbaum experiment, E_B, with its statistic T_B. We abbreviate the inference (about θ) in E_B as

$$\mathsf{Infr}_{E_B}[\mathbf{y}^*].$$

The inference implication (about θ) in E_B from $\mathbf{y}^*$ under Birnbaumization is

$$(E_2, \mathbf{y}^*) \Rightarrow \mathsf{Infr}_{E_B}[\mathbf{x}^*],$$

where the computation in E_B is always a convex combination over E_1 and E_2. But also,

$$(E_1, \mathbf{x}^*) \Rightarrow \mathsf{Infr}_{E_B}[\mathbf{x}^*].$$

It follows that, within E_B, $\mathbf{x}^*$ and $\mathbf{y}^*$ are inferentially equivalent. Call this claim

$$[B] : \mathsf{Infr}_{E_B}[\mathbf{x}^*] = \mathsf{Infr}_{E_B}[\mathbf{y}^*].$$

The argument is to hold for any SLP pair. Now [B] does not yet reach the SLP which requires

$$\mathsf{Infr}_{E_1}[\mathbf{x}^*] = \mathsf{Infr}_{E_2}[\mathbf{y}^*].$$

But Birnbaum does not stop there. Having constructed the hypothetical experiment E_B, we are to use the WCP to condition back down to the known experiment E_2. But this will not produce the SLP as we now show.

2.5 Why Appeal to Hypothetical Mixtures?

Before turning to that, we address a possible query: why suppose the argument makes any appeal to a hypothetical mixture? (See also Section 5.1.) The reason is this: The SLP does not refer to mixtures. It is a universal generalization claiming to hold for an arbitrary SLP pair. But we have no objection to imagining [as Birnbaum does (1962), page 284] a universe of all of the possible SLP pairs, where each pair has resulted from a θ-irrelevant randomizer (for the given context). Then, when $\mathbf{y}^*$ is observed, we pluck the relevant pair and construct T_B. Our question is this: why should the inference implication from $\mathbf{y}^*$ be obtained by reference to $\mathsf{Infr}_{E_B}[\mathbf{y}^*]$, the convex combination? Birnbaum does not stop at [B], but appeals to the WCP. Note the WCP is based on the outcome $\mathbf{y}^*$ being given.

3. SLP VIOLATION PAIRS

Birnbaum's argument is of central interest when we have SLP violations. We may characterize an SLP violation as any inferential context where the antecedent of the SLP is true and the consequent is false:

> SLP violation: $(E_1, \mathbf{x}^*)$ and $(E_2, \mathbf{y}^*)$ form an SLP pair, but $\mathsf{Infr}_{E_1}[\mathbf{x}^*] \neq \mathsf{Infr}_{E_2}[\mathbf{y}^*]$.

An SLP pair that violates the SLP will be called an *SLP violation pair* (from E_1, E_2, resp.).

It is not always emphasized that whether (and how) an inference method violates the SLP depends on the type of inference to be made, even within an account that allows SLP violations. One cannot just look at the data, but must also consider the inference. For example, there may be no SLP violation if the focus is on point against point hypotheses, whereas in computing a statistical significance probability under a null hypothesis there may be. "Significance testing of a hypothesis... is viewed by many as a crucial element of statistics, yet it provides a startling and practically serious example of conflict with the [SLP]." [Berger and Wolpert (1988), pages 104–105.] The following is a dramatic example that often arises in this context.

3.1 Fixed versus Sequential Sampling

Suppose $\mathbf{X}$ and $\mathbf{Y}$ are samples from distinct experiments E_1 and E_2, both distributed as $\mathsf{N}(\theta, \sigma^2)$, with σ^2 identical and known, and p-values are to be calculated for the null hypothesis H_0: $\theta = 0$ against H_1: $\theta \neq 0$.

In E_2 the sampling rule is to continue sampling until $\overline{y}_n > c_\alpha = 1.96\sigma/\sqrt{n}$, where $\overline{y}_n = \frac{1}{n}\sum_{i=1}^n y_i$. In E_1, the sample size n is fixed and $\alpha = 0.05$.

In order to arrive at the SLP pair, we have to consider the particular outcome observed. Suppose that E_2 is run and is first able to stop with $n = 169$ trials. Denote this result as $\mathbf{y}^*$. A choice for its SLP pair $\mathbf{x}^*$ would be $(E_1, 1.96\sigma/\sqrt{169})$, and the SLP violation is the fact that the p-values associated with $\mathbf{x}^*$ and $\mathbf{y}^*$ differ.

3.2 Frequentist Evidence in the Case of Significance Tests

> "[S]topping 'when the data looks good' can be a serious error when combined with frequentist measures of evidence. For instance, if one used the stopping rule [above]... but analyzed the data as if a *fixed* sample had been taken, one could *guarantee* arbitrarily strong frequentist 'sig-

> nificance' against H_0 ... ." [Berger and Wolpert (1988), page 77.]

From their perspective, the problem is with the use of frequentist significance. For a detailed discussion in favor of the irrelevance of this stopping rule, see Berger and Wolpert (1988), pages 74–88. For sampling theorists, by contrast, this example "taken in the context of examining consistency with $\theta = 0$, is enough to refute the strong likelihood principle" [Cox (1978), page 54], since, with probability 1, it will stop with a 'nominally' significant result even though $\theta = 0$. It contradicts what Cox and Hinkley call "the weak repeated sampling principle" [Cox and Hinkley (1974), page 51]. More generally, the frequentist principle of evidence (FEV) would regard small p-values as misleading if they result from a procedure that readily generates small p-values under H_0.[2]

For the sampling theorist, to report a 1.96 standard deviation difference known to have come from optional stopping, just the same as if the sample size had been fixed, is to discard relevant information for inferring inconsistency with the null, while "according to any approach that is in accord with the strong likelihood principle, the fact that this particular stopping rule has been used is irrelevant." [Cox and Hinkley (1974), page 51.][3] The actual p-value will depend of course on when it stops. We emphasize that our argument does not turn on accepting a frequentist principle of evidence (FEV), but these considerations are useful both to motivate and understand the core principle of Birnbaum's argument, the WCP.

4. THE WEAK CONDITIONALITY PRINCIPLE (WCP)

From Section 2.4 we have [B] $\mathsf{Infr}_{E_B}[\mathbf{x}^*] = \mathsf{Infr}_{E_B}[\mathbf{y}^*]$ since the inference implication is by the constructed T_B. How might Birnbaum move from [B] to the SLP, for an arbitrary pair $\mathbf{x}^*$and $\mathbf{y}^*$?

There are two possibilities. One would be to insist informative inference ignore or be insensitive to sampling distributions. But since we know that SLP violations result because of the difference in sampling distributions, to simply deny them would obviously render his argument circular (or else irrelevant for sampling theory). We assume Birnbaum does not intend his argument to be circular and Birnbaum relies on further steps to which we now turn.

4.1 Mixture ($E_{\mathbf{mix}}$): Two Instruments of Different Precisions [Cox (1958)]

The crucial principle of inference on which Birnbaum's argument rests is the weak conditionality principle (WCP), intended to indicate the relevant sampling distribution in the case of certain mixture experiments. The famous example to which we already alluded, "is now usually called the 'weighing machine example,' which draws attention to the need for conditioning, at least in certain types of problems" [Reid (1992), page 582].

We flip a fair coin to decide which of two instruments, E_1 or E_2, to use in observing a Normally distributed random sample $\mathbf{Z}$ to make inferences about mean θ. E_1 has variance of 1, while that of E_2 is 10^6. We limit ourselves to mixtures of two experiments.

In testing a null hypothesis such as $\theta = 0$, the same $\mathbf{z}$ measurement would correspond to a much smaller p-value were it to have come from E_1 rather than from E_2: denote them as $p_1(\mathbf{z})$ and $p_2(\mathbf{z})$, respectively. The overall (or unconditional) significance level of the mixture E_{mix} is the convex combination of the p-values: $[p_1(\mathbf{z}) + p_2(\mathbf{z})]/2$. This would give a misleading report of how precise or stringent the actual experimental measurement is [Cox and Mayo (2010), page 296]. [See Example 4.6, Cox and Hinkley (1974), pages 95–96; Birnbaum (1962), page 280.]

Suppose that we know we have observed a measurement from E_2 with its much larger variance:

> The unconditional test says that we can assign this a higher level of significance than we ordinarily do, because if we were to repeat the experiment, we might sample some quite different distribution. But this fact seems irrelevant to the interpretation of an observation which we know came from a distribution [with the larger variance]. [Cox (1958), page 361.]

The WCP says simply: *once it is known which E_i has produced $\mathbf{z}$, the p-value or other inferential assessment should be made with reference to the experiment actually run.*

[2] Mayo and Cox (2010), page 254:

FEV: $\mathbf{y}$ is (strong) evidence against H_0, if and only if, were H_0 a correct description of the mechanism generating $\mathbf{y}$, then, with high probability this would have resulted in a less discordant result than is exemplified by $\mathbf{y}$.

[3] Analogous situations occur without optional stopping, as with selecting a data-dependent, maximally likely, alternative [Cox and Hinkley (1974), Example 2.4.1, page 51]. See also Mayo and Kruse (2001).

4.2 Weak Conditionality Principle (WCP) in the Weighing Machine Example

We first state the WCP in relation to this example.

We are given $(E_{\text{mix}}, \mathbf{z}_i)$, that is, $(E_i, \mathbf{z}_i)$ results from mixture experiment E_{mix}. WCP exhorts us to condition to be relevant to the experiment actually producing the outcome. This is an example of what Cox terms "conditioning for relevance."

> WCP: Given $(E_{\text{mix}}, \mathbf{z}_i)$, condition on the E_i producing the result
>
> $$(E_{\text{mix}}, \mathbf{z}_i) \Rightarrow \mathsf{Infr}_{E_i}[(E_{\text{mix}}, \mathbf{z}_i)] = p_i(\mathbf{z}) = \mathsf{Infr}_{E_i}[\mathbf{z}_i].$$
>
> Do not use the unconditional formulation
>
> $$(E_{\text{mix}}, \mathbf{z}_i) \nRightarrow \mathsf{Infr}_{E_{\text{mix}}}[(E_{\text{mix}}, \mathbf{z}_i)] = [p_1(\mathbf{z}) + p_2(\mathbf{z})]/2.$$
>
> The concern is that
>
> $$\mathsf{Infr}_{E_{\text{mix}}}[(E_{\text{mix}}, \mathbf{z}_i)] = [p_1(\mathbf{z}) + p_2(\mathbf{z})]/2 \neq p_i(\mathbf{z}).$$

There are three sampling distributions, and the WCP says the relevant one to use whenever $\mathsf{Infr}_{E_{\text{mix}}}[\mathbf{z}_i] \neq \mathsf{Infr}_{E_i}[\mathbf{z}_i]$ is the one known to have generated the result [Birnbaum (1962), page 280]. In other cases the WCP would make no difference.

4.3 The WCP and Its Corollaries

We can give a general statement of the WCP as follows:

A mixture E_{mix} selects between E_1 and E_2, using a θ-irrelevant process, and it is given that $(E_i, \mathbf{z}_i)$ results, $i = 1, 2$. WCP directs the inference implication. Knowing we are mapping an outcome from a mixture, there is no need to repeat the first component of $(E_{\text{mix}}, \mathbf{z}_i)$, so it is dropped except when a reminder seems useful:

(i) Condition to obtain relevance:

$$(E_{\text{mix}}, \mathbf{z}_i) \Rightarrow \mathsf{Infr}_{E_i}[(E_{\text{mix}}, \mathbf{z}_i)] = \mathsf{Infr}_{E_i}(\mathbf{z}_i).$$

In words, $\mathbf{z}_i$ arose from E_{mix} but the inference implication is based on E_i.

(ii) Eschew unconditional formulations:

$$(E_{\text{mix}}, \mathbf{z}_i) \nRightarrow \mathsf{Infr}_{E_{\text{mix}}}[\mathbf{z}_i],$$

whenever the unconditional treatment yields a different inference implication,

that is, whenever $\mathsf{Infr}_{E_{\text{mix}}}[\mathbf{z}_i] \neq \mathsf{Infr}_{E_i}[\mathbf{z}_i]$.

NOTE. $\mathsf{Infr}_{E_{\text{mix}}}[\mathbf{z}_i]$ which abbreviates $\mathsf{Infr}_{E_{\text{mix}}}[(E_{\text{mix}}, \mathbf{z}_i)]$ asserts that the inference implication uses the convex combination of the relevant pair of experiments.

We now highlight some points for reference.

4.3.1 *WCP makes a difference*. The cases of interest here are where applying WCP would alter the unconditional implication. In these cases WCP makes a difference.

Note that (ii) blocks computing the inference implication from $(E_{\text{mix}}, \mathbf{z}_i)$ as $\mathsf{Infr}_{E_{\text{mix}}}[\mathbf{z}_i]$, whenever $\mathsf{Infr}_{E_{\text{mix}}}[\mathbf{z}_i] \neq \mathsf{Infr}_{E_i}[\mathbf{z}_i]$ for $i = 1, 2$. Here E_1, E_2 and E_{mix} would correspond to three sampling distributions.

WCP requires the experiment and its outcome to be given or known: If it is given only that $\mathbf{z}$ came from E_1 or E_2, and not which, then WCP does not authorize (i). In fact, we would wish to block such an inference implication. For instance,

$$(E_1 \text{ or } E_2, \mathbf{z}) \nRightarrow \mathsf{Infr}_{E_1}[\mathbf{z}].$$

Point on notation: The use of "$\Rightarrow$" is for a given outcome. We may allow it to be used without ambiguity when only a disjunction is given, because while E_1 entails (E_1 or E_2), the converse does not hold. So no erroneous substitution into an inference implication would follow.

4.3.2 *Irrelevant augmentation*: *Keep irrelevant facts irrelevant* (*Irrel*). Another way to view the WCP is to see it as exhorting us to keep what is irrelevant to the sampling behavior of the experiment performed irrelevant (to the inference implication). Consider Birnbaum's (1969), page 119, idea that a "trivial" but harmless addition to any given experimental result $\mathbf{z}$ might be to toss a fair coin and augment $\mathbf{z}$ with a report of heads or tails (where this is irrelevant to the original model). Note the similarity to attempts to get an exact significance level in discrete tests, by allowing borderline outcomes to be declared significant or not (at the given level) according to the outcome of a coin toss. The WCP, of course, eschews this. But there is a crucial ambiguity to avoid. It is a harmless addition only if it remains harmless to the inference implication. If it is allowed to alter the test result, it is scarcely harmless.

A holder of the WCP may stipulate that a given $\mathbf{z}_i$ can always be augmented with the result of a θ-irrelevant randomizer, provided that it remains irrelevant to the inference implication about θ in E_i. We can abbreviate this irrelevant augmentation of a given result $\mathbf{z}_i$ as a conjunction: (E_i & Irrel),

(Irrel): $\mathsf{Infr}_{E_i}[(E_i \ \& \ \mathsf{Irrel}, \mathbf{z}_i)] = \mathsf{Infr}_{E_i}[\mathbf{z}_i]$, $i = 1, 2$.

We illuminate this in the next subsection.

4.3.3 *Is the WCP an equivalence?* "It was the adoption of an unqualified equivalence formulation of conditionality, and related concepts, which led, in my 1962 paper, to the monster of the likelihood axiom" [Birnbaum (1975), page 263]. He admits the contrast with "the one-sided form to which applications" had been restricted [Birnbaum (1969), page 139, note 11]. The question of whether the WCP is a proper equivalence relation, holding in both directions, is one of the most central issues in the argument. But what would be alleged to be equivalent?

Obviously not the unconditional and the conditional inference implications: the WCP makes a difference just when they are inequivalent, that is, when $\mathsf{Infr}_{E_{\text{mix}}}[\mathbf{z}_i] \neq \mathsf{Infr}_{E_i}[\mathbf{z}_i]$. Our answer is that the WCP involves an inequivalence as well as an equivalence. The WCP prescribes conditioning on the experiment known to have produced the data, and not the other way around. It is their inequivalence that gives Cox's WCP its normative proscriptive force. To assume the WCP identifies $\mathsf{Infr}_{E_{\text{mix}}}[\mathbf{z}_i]$ and $\mathsf{Infr}_{E_i}[\mathbf{z}_i]$ leads to trouble. (We return to this in Section 7.)

However, there is an equivalence in WCP (i). Further, once the outcome is given, the addition of θ-irrelevant features about the selection of the experiment performed are to remain irrelevant to the inference implication:

$$\mathsf{Infr}_{E_i}\big[(E_{\text{mix}}, \mathbf{z}_i)\big] = \mathsf{Infr}_{E_i}\big[(E_i \ \& \ \mathsf{Irrel}, \mathbf{z}_i)\big].$$

Both are the same as $\mathsf{Infr}_{E_i}[\mathbf{z}_i]$. While claiming that $\mathbf{z}$ came from a mixture, even knowing it came from a nonmixture, may seem unsettling, we grant it for purposes of making out Birnbaum's argument. By (Irrel), it cannot alter the inference implication under E_i.

5. BIRNBAUM'S SLP ARGUMENT

5.1 Birnbaumization and the WCP

What does the WCP entail as regards Birnbaumization? Now WCP refers to mixtures, but is the Birnbaum experiment E_B a mixture experiment? Not really. One cannot perform the following: Toss a fair coin (or other θ-irrelevant randomizer). If it lands heads, perform an experiment E_2 that yields a member of an SLP pair $\mathbf{y}^*$; if tails, observe an experiment that yields the other member of the SLP pair $\mathbf{x}^*$. We do not know what outcome would have resulted from the unperformed experiment, much less that it would be an outcome with a proportional likelihood to the observed $\mathbf{y}^*$. There is a single experiment, and it is stipulated we know which and what its outcome was. Some have described the Birnbaum experiment as unperformable, or at most a "mathematical mixture" rather than an "experimental mixture" [Kalbfleisch (1975), pages 252–253]. Birnbaum himself calls it a "hypothetical" mixture [Birnbaum (1962), page 284].

While a holder of the WCP may simply deny its general applicability in hypothetical experiments, given that Birnbaum's argument has stood for over fifty years, we wish to give it maximal mileage. Birnbaumization may be "performed" in the sense that T_B can be defined for any SLP pair $\mathbf{x}^*$, $\mathbf{y}^*$. Refer back to the hypothetical universe of SLP pairs, each imagined to have been generated from a θ-irrelevant mixture (Section 2.5). When we observe $\mathbf{y}^*$ we pluck the $\mathbf{x}^*$ companion needed for the argument. In short, we can Birnbaumize an experimental result: Constructing statistic T_B with the derived experiment E_B is the "performance." But what cannot shift in the argument is the stipulation that E_i be given or known (as noted in Section 4.3.1), that i be fixed. Nor can the meaning of "given $\mathbf{z}^*$" shift through the argument, if it is to be sound.

Given $\mathbf{z}^*$, the WCP precludes Birnbaumizing. On the other hand, if the reported $\mathbf{z}^*$ was the value of T_B, then we are given only the disjunction, precluding the computation relevant for i fixed (Section 4.3.1). Let us consider the components of Birnbaum's argument.

5.2 Birnbaum's Argument

$(E_2, \mathbf{y}^*)$ is given (and it has an SLP pair $\mathbf{x}^*$). The question is to its inferential import. Birnbaum will seek to show that

$$\mathsf{Infr}_{E_2}\big[\mathbf{y}^*\big] = \mathsf{Infr}_{E_1}\big[\mathbf{x}^*\big].$$

The value of T_B is $(E_1, \mathbf{x}^*)$. Birnbaumization maps outcomes into hypothetical mixtures E_B:

(1) If the inference implication is by the stipulations of E_B,

$$(E_2, \mathbf{y}^*) \Rightarrow \mathsf{Infr}_{E_B}\big[\mathbf{x}^*\big] = \mathsf{Infr}_{E_B}\big[\mathbf{y}^*\big].$$

Likewise for $(E_1, \mathbf{x}^*)$. T_B is a sufficient statistic for E_B (the conditional distribution of $\mathbf{Z}$ given T_B is independent of θ).

(2) If the inference implication is by WCP,

$$(E_2, \mathbf{y}^*) \nRightarrow \mathsf{Infr}_{E_B}[\mathbf{y}^*],$$

rather

$$(E_2, \mathbf{y}^*) \Rightarrow \mathsf{Infr}_{E_2}[\mathbf{y}^*]$$

and

$$(E_1, \mathbf{x}^*) \Rightarrow \mathsf{Infr}_{E_1}[\mathbf{x}^*].$$

Following the inference implication according to E_B in (1) is at odds with what the WCP stipulates in (2). Given $\mathbf{y}^*$, Birnbaumization directs using the convex combination over the components of T_B; WCP eschews doing so. We will not get

$$\mathsf{Infr}_{E_1}[\mathbf{x}^*] = \mathsf{Infr}_{E_2}[\mathbf{y}^*].$$

The SLP only seems to follow by the erroneous identity:

$$\mathsf{Infr}_{E_B}[\mathbf{z}_i^*] = \mathsf{Infr}_{E_i}[\mathbf{z}_i^*] \quad \text{for } i = 1, 2.$$

5.3 Refuting the Supposition that [(SP and WCP) entails SLP]

We can uphold both (1) and (2), while at the same time holding the following:

(3) $\mathsf{Infr}_{E_1}[\mathbf{x}^*] \neq \mathsf{Infr}_{E_2}[\mathbf{y}^*]$.

Specifically, any case where $\mathbf{x}^*$ and $\mathbf{y}^*$ is an SLP violation pair is a case where (3) is true. Since whenever (3) holds we have a counterexample to the SLP generalization, this demonstrates that SP and WCP and not-SLP are logically consistent. Thus, so are WCP and not-SLP. This refutes the supposition that [(SP and WCP) entails SLP] and also any purported derivation of SLP from WCP alone.[4]

SP is not blocked in (1). The SP is always relative to a model, here E_B. We have the following:

$\mathbf{x}^*$ and $\mathbf{y}^*$ are SLP pairs in E_B, and
$\mathsf{Infr}_{E_B}[\mathbf{x}^*] = \mathsf{Infr}_{E_B}[\mathbf{y}^*]$ (i.e., [B] holds).

One may allow different contexts to dictate whether or not to condition [i.e., whether to apply (1) or (2)], but we know of no inference account that permits, let alone requires, self-contradictions. By noncontradiction, for any $(E, \mathbf{z})$, $\mathsf{Infr}_E[\mathbf{z}] = \mathsf{Infr}_E[\mathbf{z}]$. ("$\Rightarrow$" is a function from outcomes to inference implications, and $\mathbf{z} = \mathbf{z}$, for any $\mathbf{z}$.)

Upholding and applying. This recalls our points in Section 2.1. Applying a rule means following its inference directive. We may uphold the if-then stipulations in (1) and (2), but to apply their competing implications in a single case is self-contradictory.

Arguing from a self-contradiction is unsound. The slogan that anything follows from a self-contradiction G and not-G is true, since for any claim C, the following is a logical truth: If G then (if not-G then C). Two applications of *modus ponens* yield C. One can also derive not-C! But since G and its denial cannot be simultaneously true, any such argument is unsound. (A sound argument must have true premises and be logically valid.) We know Birnbaum was not intending to argue from a self-contradiction, but this may inadvertently occur.

5.4 What if the SLP Pair Arose from an Actual Mixture?

What if the SLP pair $\mathbf{x}^*$, $\mathbf{y}^*$ arose from a genuine, and not a Birnbaumized, mixture. (Consider fixed versus sequential sampling, Section 3.1. Suppose E_1 fixes n at 169, the coin flip says perform E_2, and it happens to stop at $n = 169$.) We may allow that an unconditional formulation may be defined so that

$$\mathsf{Infr}_{E_{\text{mix}}}[\mathbf{x}^*] = \mathsf{Infr}_{E_{\text{mix}}}[\mathbf{y}^*].$$

But WCP eschews the unconditional formulation; it says condition on the experiment known to have produced $\mathbf{z}_i$:

$$(E_{\text{mix}}, \mathbf{z}_i^*) \Rightarrow \mathsf{Infr}_{E_i}[\mathbf{z}_i^*], \quad i = 1, 2.$$

Any SLP violation pair $\mathbf{x}^*$, $\mathbf{y}^*$ remains one: $\mathsf{Infr}_{E_1}[\mathbf{x}^*] \neq \mathsf{Infr}_{E_2}[\mathbf{y}^*]$.

6. DISCUSSION

We think a fresh look at this venerable argument is warranted. Wearing a logician's spectacles and entering the debate outside of the thorny issues from decades ago may be an advantage.

It must be remembered that the onus is not on someone who questions if the SLP follows from SP and WCP to provide suitable principles of evidence, however desirable it might be to have them. The onus is on Birnbaum to show that for any given $\mathbf{y}^*$, a member of an SLP pair with $\mathbf{x}^*$, with different probability models $f_1(\cdot)$, $f_2(\cdot)$, that he will be able to derive from SP and WCP, that $\mathbf{x}^*$ and $\mathbf{y}^*$ would have the identical inference

[4] By allowing applications of Birnbaumization and appropriate choices of the irrelevant randomization probabilities, SP can be weakened to "mathematical equivalence," or even (with compounded mixtures) omitted so that WCP would entail SLP. See Birnbaum (1972) and Evans, Fraser and Monette (1986).

implications concerning shared parameter θ. We have shown that SLP violations do not entail renouncing either the SP or the WCP.

It is no rescue of Birnbaum's argument that a sampling theorist wants principles in addition to the WCP to direct the relevant sampling distribution for inference; indeed, Cox has given others. It was to make the application of the WCP in his argument as plausible as possible to sampling theorists that Birnbaum begins with the type of mixture in Cox's (1958) famous example of instruments E_1, E_2 with different precisions.

We do not assume sampling theory, but employ a formulation that avoids ruling it out in advance. The failure of Birnbaum's argument to reach the SLP relies only on a correct understanding of the WCP. We may grant that for any $\mathbf{y}^*$ its SLP pair could occur in repetitions (and may even be out there as in Section 2.5). However, the key point of the WCP is to deny that this fact should alter the inference implication from the known $\mathbf{y}^*$. To insist it should is to deny the WCP. Granted, WCP sought to identify the relevant sampling distribution for inference from a specified type of mixture, and a known $\mathbf{y}^*$, but it is Birnbaum who purports to give an argument that is relevant for a sampling theorist and for "approaches which are independent of this [Bayes'] principle" [Birnbaum (1962), page 283]. Its implications for sampling theory is why it was dubbed "a landmark in statistics" [Savage (1962b), page 307].

Let us look at the two statements about inference implications from a given $(E_2, \mathbf{y}^*)$, applying (1) and (2) in Section 5.2:

$$(E_2, \mathbf{y}^*) \Rightarrow \mathsf{Infr}_{E_B}[\mathbf{x}^*],$$
$$(E_2, \mathbf{y}^*) \Rightarrow \mathsf{Infr}_{E_2}[\mathbf{y}^*].$$

Can both be applied in exactly the same model with the same given $\mathbf{z}$? The answer is yes, so long as the WCP happens to make no difference:

$$\mathsf{Infr}_{E_B}[\mathbf{z}_i^*] = \mathsf{Infr}_{E_i}[\mathbf{z}_i^*], \quad i = 1, 2.$$

Now the SLP must be applicable to an arbitrary SLP pair. However, to assume that (1) and (2) can be consistently applied for any $\mathbf{x}^*, \mathbf{y}^*$ pair would be to assume no SLP violations are possible, which really would render Birnbaum's argument circular. So from Section 5.3, the choices are to regard Birnbaum's argument as unsound (arguing from a contradiction) or circular (assuming what it purports to prove). Neither is satisfactory. We are left with competing inference implications and no way to get to the SLP. There is evidence Birnbaum saw the gap in his argument (Birnbaum, 1972), and in the end he held the SLP only restricted to (predesignated) point against point hypotheses.[5]

It is not SP and WCP that conflict; the conflict comes from WCP together with Birnbaumization—understood as both invoking the hypothetical mixture and erasing the information as to which experiment the data came. If one Birnbaumizes, one cannot at the same time uphold the "keep irrelevants irrelevant" (Irrel) stipulation of the WCP. So for any given $(E, \mathbf{z})$ one must choose, and the answer is straightforward for a holder of the WCP. To paraphrase Cox's (1958), page 361, objection to unconditional tests:

> Birnbaumization says that we can assign $\mathbf{y}^*$ a different level of significance than we ordinarily do, because one may identify an SLP pair $\mathbf{x}^*$ and construct statistic T_B. But this fact seems irrelevant to the interpretation of an observation which we know came from E_2. To conceal the index, and use the convex combination, would give a distorted assessment of statistical significance.

7. RELATION TO OTHER CRITICISMS OF BIRNBAUM

A number of critical discussions of the Birnbaum argument and the SLP exist. While space makes it impossible to discuss them here, we believe the current analysis cuts through this extremely complex literature. Take, for example, the most well-known criticisms by Durbin (1970) and Kalbfleish (1975), discussed in the excellent paper by Evans, Fraser and Monette (1986). Allowing that any $\mathbf{y}^*$ may be viewed as having arisen from Birnbaum's mathematical mixture, they consider the proper order of application of the principles. If we condition on the given experiment first, Kalbfleish's revised sufficiency principle is inapplicable, so Birnbaum's argument fails. On the other hand, Durbin argues, if we reduce to the minimal sufficient statistic first, then his revised principle of conditionality cannot be applied. Again Birnbaum's argument fails. So either way it fails.

Unfortunately, the idea that one must revise the initial principles in order to block SLP allows downplaying or dismissing these objections as tantamount to

[5] This alone would not oust all sampling distributions. Birnbaum's argument, even were it able to get a foothold, would have to apply further rounds of conditioning to arrive at the data alone.

denying SLP at any cost (please see the references[6]). We can achieve what they wish to show, without altering principles, and from WCP alone. Given $\mathbf{y}^*$, WCP blocks Birnbaumization; given $\mathbf{y}^*$ has been Birnbaumized, the WCP precludes conditioning.

We agree with Evans, Fraser and Monette (1986), page 193, "that Birnbaum's use of [the principles] ...are contrary to the intentions of the principles, as judged by the relevant supporting and motivating examples. From this viewpoint we can state that the intentions of S and C do not imply L." [Where S, C and L are our SP, WCP and SLP.] Like Durbin and Kalbfleisch, they offer a choice of modifications of the principles to block the SLP. These are highly insightful and interesting; we agree that they highlight a need to be clear on the experimental model at hand. Still, it is preferable to state the WCP so as to reflect these "intentions," without which it is robbed of its function. The problem stems from mistaking WCP as the equivalence $\mathsf{Infr}_{E_{\text{mix}}}[\mathbf{z}] = \mathsf{Infr}_{E_i}[\mathbf{z}]$ (whether the mixture is hypothetical or actual). This is at odds with the WCP. The puzzle is solved by adequately stating the WCP. Aside from that, we need only keep the meaning of terms consistent through the argument.

We emphasize that we are neither rejecting the SP nor claiming that it breaks down, even in the special case E_B. The sufficiency of T_B within E_B, as a mathematical concept, holds: the value of T_B "suffices" for $\mathsf{Infr}_{E_B}[\mathbf{y}^*]$, the inference from the associated convex combination. Whether reference to hypothetical mixture E_B is relevant for inference from given $\mathbf{y}^*$ is a distinct question. For an alternative criticism see Evans (2013).

8. CONCLUDING REMARKS

An essential component of informative inference for sampling theorists is the relevant sampling distribution: it is not a separate assessment of performance, but part of the necessary ingredients of informative inference. It is this feature that enables sampling theory to have SLP violations (e.g., in significance testing contexts). Any such SLP violation, according to Birnbaum's argument, prevents adhering to both SP and WCP. We have shown that SLP violations do not preclude WCP and SP.

The SLP does not refer to mixtures. But supposing that $(E_2, \mathbf{y}^*)$ is given, Birnbaum asks us to consider that $\mathbf{y}^*$ could also have resulted from a θ-irrelevant mixture that selects between E_1, E_2. The WCP says this piece of information should be irrelevant for computing the inference from $(E_2, \mathbf{y}^*)$ once given. That is, $\mathsf{Infr}_{E_i}[(E_{\text{mix}}, \mathbf{y}^*)] = \mathsf{Infr}_{E_i}[\mathbf{y}^*],\ i = 1, 2$. It follows that if $\mathsf{Infr}_{E_1}[\mathbf{x}^*] \neq \mathsf{Infr}_{E_2}[\mathbf{y}^*]$, the two remain unequal after the recognition that $\mathbf{y}^*$ could have come from the mixture. What was an SLP violation remains one.

Given $\mathbf{y}^*$, the WCP says do not Birnbaumize. One is free to do so, but not to simultaneously claim to hold the WCP in relation to the given $\mathbf{y}^*$, on pain of logical contradiction. If one does choose to Birnbaumize, and to construct T_B, admittedly the known outcome $\mathbf{y}^*$ yields the same value of T_B as would $\mathbf{x}^*$. Using the sample space of E_B yields [B]: $\mathsf{Infr}_{E_B}[\mathbf{x}^*] = \mathsf{Infr}_{E_B}[\mathbf{y}^*]$. This is based on the convex combination of the two experiments and differs from both $\mathsf{Infr}_{E_1}[\mathbf{x}^*]$ and $\mathsf{Infr}_{E_2}[\mathbf{y}^*]$. So again, any SLP violation remains. Granted, if only the value of T_B is given, using Infr_{E_B} may be appropriate. For then we are given only the disjunction: either $(E_1, \mathbf{x}^*)$ or $(E_2, \mathbf{y}^*)$. In that case, one is barred from using the implication from either individual E_i. A holder of WCP might put it this way: once $(E, \mathbf{z})$ is given, whether E arose from a θ-irrelevant mixture or was fixed all along should not matter to the inference, but whether a result was Birnbaumized or not should, and does, matter.

There is no logical contradiction in holding that if data are analyzed one way (using the convex combination in E_B), a given answer results, and if analyzed another way (via WCP), one gets quite a different result. One may consistently apply both the E_B and the WCP directives to the same result, in the same experimental model, only in cases where WCP makes no difference. To claim for any $\mathbf{x}^*$, $\mathbf{y}^*$, the WCP never makes a difference, however, would assume that there can be no SLP violations, which would make the argument circular.[7] Another possibility would be to hold, as Birnbaum ultimately did, that the SLP is "clearly plausible" [Birnbaum (1968), page 301] only in "the severely restricted case of a parameter space of just two points"

[6] In addition to the authors cited in the manuscript, see especially comments by Savage, Cornfield, Bross, Pratt, Dempster et al. (1962) on Birnbaum. For later discussions, see Barndorff-Nielsen (1975), Berger (1986), Berger and Wolpert (1988), Birnbaum (1970a, 1970b), Dawid (1986), Savage (1970) and references therein.

[7] His argument would then follow the pattern: If there are SLP violations, then there are no SLP violations. Note that (V implies not-V) is not a logical contradiction. It is logically equivalent to not-V. Then, Birnbaum's argument is equivalent to not-V: denying that $\mathbf{x}^*$, $\mathbf{y}^*$ can give rise to an SLP violation. That would render it circular.

where these are predesignated [Birnbaum (1969), page 128]. But that is to relinquish the general result.

ACKNOWLEDGMENTS

I am extremely grateful to David Cox and Aris Spanos for numerous discussions, corrections and joint work over many years on this and related foundational issues. I appreciate the careful queries and detailed suggested improvements on earlier drafts from anonymous referees, from Jean Miller and Larry Wasserman. My understanding of Birnbaum was greatly facilitated by philosopher of science, Ronald Giere, who worked with Birnbaum. I'm also grateful for his gift of some of Birnbaum's original materials and notes.

REFERENCES

BARNDORFF-NIELSEN, O. (1975). Comments on paper by J. D. Kalbfleisch. *Biometrika* **62** 261–262.

BERGER, J. O. (1986). Discussion on a paper by Evans et al. [On principles and arguments to likelihood]. *Canad. J. Statist.* **14** 195–196.

BERGER, J. O. (2006). The case for objective Bayesian analysis. *Bayesian Anal.* **1** 385–402. MR2221271

BERGER, J. O. and WOLPERT, R. L. (1988). *The Likelihood Principle*, 2nd ed. *Lecture Notes—Monograph Series* **6**. IMS, Hayward, CA.

BIRNBAUM, A. (1962). On the foundations of statistical inference. *J. Amer. Statist. Assoc.* **57** 269–306. Reprinted in *Breakthroughs in Statistics* **1** (S. Kotz and N. Johnson, eds.) 478–518. Springer, New York.

BIRNBAUM, A. (1968). Likelihood. In *International Encyclopedia of the Social Sciences* **9** 299–301. Macmillan and the Free Press, New York.

BIRNBAUM, A. (1969). Concepts of statistical evidence. In *Philosophy, Science, and Method: Essays in Honor of Ernest Nagel* (S. Morgenbesser, P. Suppes and M. G. White, eds.) 112–143. St. Martin's Press, New York.

BIRNBAUM, A. (1970a). Statistical methods in scientific inference. *Nature* **225** 1033.

BIRNBAUM, A. (1970b). On Durbin's modified principle of conditionality. *J. Amer. Statist. Assoc.* **65** 402–403.

BIRNBAUM, A. (1972). More on concepts of statistical evidence. *J. Amer. Statist. Assoc.* **67** 858–861. MR0365793

BIRNBAUM, A. (1975). Comments on paper by J. D. Kalbfleisch. *Biometrika* **62** 262–264.

CASELLA, G. and BERGER, R. L. (2002). *Statistical Inference*, 2nd ed. Duxbury Press, Belmont, CA.

COX, D. R. (1958). Some problems connected with statistical inference. *Ann. Math. Statist.* **29** 357–372. MR0094890

COX, D. R. (1977). The role of significance tests. *Scand. J. Stat.* **4** 49–70. MR0448666

COX, D. R. (1978). Foundations of statistical inference: The case for eclecticism. *Aust. N. Z. J. Stat.* **20** 43–59. MR0501453

COX, D. R. and HINKLEY, D. V. (1974). *Theoretical Statistics*. Chapman & Hall, London. MR0370837

COX, D. R. and MAYO, D. G. (2010). Objectivity and conditionality in frequentist inference. In *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science* (G. Mayo and A. Spanos, eds.) 276–304. Cambridge Univ. Press, Cambridge.

DAWID, A. P. (1986). Discussion on a paper by Evans et al. [On principles and arguments to likelihood]. *Canad. J. Statist.* **14** 196–197.

DURBIN, J. (1970). On Birnbaum's theorem on the relation between sufficiency, conditionality and likelihood. *J. Amer. Statist. Assoc.* **65** 395–398.

EVANS, M. (2013). What does the proof of Birnbaum's theorem prove? Unpublished manuscript.

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

GHOSH, J. K., DELAMPADY, M. and SAMANTA, T. (2006). *An Introduction to Bayesian Analysis. Theory and Methods. Springer Texts in Statistics*. Springer, New York. MR2247439

KALBFLEISCH, J. D. (1975). Sufficiency and conditionality. *Biometrika* **62** 251–268. MR0386075

LEHMANN, E. L. and ROMANO, J. P. (2005). *Testing Statistical Hypotheses*, 3rd ed. *Springer Texts in Statistics*. Springer, New York. MR2135927

MAYO, D. G. (1996). *Error and the Growth of Experimental Knowledge*. Univ. Chicago Press, Chicago, IL.

MAYO, D. G. (2010). An error in the argument from conditionality and sufficiency to the likelihood principle. In *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science* (D. G. Mayo and A. Spanos, eds.) 305–314. Cambridge Univ. Press, Cambridge.

MAYO, D. G. and COX, D. R. (2010). Frequentist statistics as a theory of inductive inference. In *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science* (D. G. Mayo and A. Spanos, eds.) 247–274. Cambridge Univ. Press, Cambridge. First published in *The Second Erich L. Lehmann Symposium: Optimality* **49** (2006) (J. Rojo, ed.) 77–97. *Lecture Notes—Monograph Series*. IMS, Beachwood, OH.

MAYO, D. G. and COX, D. R. (2011). Statistical scientist meets a philosopher of science: A conversation. In *Rationality, Markets and Morals: Studies at the Intersection of Philosophy and Economics* **2** (D. G. Mayo, A. Spanos and K. W. Staley, eds.) (*Special Topic: Statistical Science and Philosophy of Science: Where do (should) They Meet in 2011 and Beyond*?) (October 18) 103–114. Frankfurt School, Frankfurt.

MAYO, D. G. and KRUSE, M. (2001). Principles of inference and their consequences. In *Foundations of Bayesianism* (D. Corfield and J. Williamson, eds.) **24** 381–403. *Applied Logic*. Kluwer Academic Publishers, Dordrecht.

MAYO, D. G. and SPANOS, A. (2006). Severe testing as a basic concept in a Neyman–Pearson philosophy of induction. *British J. Philos. Sci.* **57** 323–357. MR2249183

MAYO, D. G. and SPANOS, A. (2011). Error statistics. In *Philosophy of Statistics* **7** (P. S. Bandyopadhyay and M. R. Forster, eds.) 152–198. *Handbook of the Philosophy of Science*. Elsevier, Amsterdam.

REID, N. (1992). Introduction to Fraser (1966) structural probability and a generalization. In *Breakthroughs in Statistics* (S. Kotz and N. L. Johnson, eds.) 579–586. *Springer Series in Statistics*. Springer, New York.

SAVAGE, L. J., ed. (1962a). *The Foundations of Statistical Inference*: *A Discussion*. Methuen, London.

SAVAGE, L. J. (1962b). Discussion on a paper by A. Birnbaum [On the foundations of statistical inference]. *J. Amer. Statist. Assoc.* **57** 307–308.

SAVAGE, L. J. (1970). Comments on a weakened principle of conditionality. *J. Amer. Statist. Assoc.* **65** (329) 399–401.

SAVAGE, L. J., BARNARD, G., CORNFIELD, J., BROSS, I., BOX, G. E. P., GOOD, I. J., LINDLEY, D. V. et al. (1962). On the foundations of statistical inference: Discussion. *J. Amer. Statist. Assoc.* **57** 307–326.

Statistical Science
2014, Vol. 29, No. 2, 240–241
DOI: 10.1214/14-STS470
Main article DOI: 10.1214/13-STS457

Discussion of "On the Birnbaum Argument for the Strong Likelihood Principle"

A. P. Dawid

Abstract. Deborah Mayo claims to have refuted Birnbaum's argument that the Likelihood Principle is a logical consequence of the Sufficiency and Conditionality Principles. However, this claim fails because her interpretation of the Conditionality Principle is different from Birnbaum's. Birnbaum's proof cannot be so readily dismissed.

Key words and phrases: Conditionality principle, Birnbaum's theorem, likelihood principle, sufficiency principle, weak conditionality principle.

Deborah Mayo (2014) is not the first devoutly to wish that the (strong) Likelihood Principle [principle L of Birnbaum (1962)] was *not* a logical consequence of the Sufficiency Principle (Birnbaum's S) and the Conditionality Principle (Birnbaum's C). This concern arises because much of frequentist inference is in clear violation of L, while at the same time purporting to abide by S and C. This constitutes a self-contradiction, which frequentists are, however, loth to admit. Birnbaum himself appears to have been quite distraught at his own finding, and in the half-century since publication of his argument there has been a constant trickle of attempts to come to terms with it, including one or two of my own (Dawid, 1977; Dawid, 1983; Dawid, 1987; Dawid, 2011); a detailed account that I consider displays the underlying logic clearly can be found in Chapter II "Principles of Inference" of Dawid (2013).

Those who feel disquiet at the destructive implications of Birnbaum's theorem for their favored method of inference (be it frequentist or, for example, "objective Bayesian," which also violates L) have a number of strategies to try and ease that disquiet. If they accept the validity of the theorem, they might argue [along with Fraser (1963); Durbin (1970); Kalbfleisch (1975)] that S or C should not be taken as universally applicable—thus evading the consequent of the theorem by denying its antecedents. This is at least a logically sound ploy, although it reeks of adhockery. Also, the ploy may not be totally successful, since some of the "undesirable" implications of the theorem may survive weakening of its hypotheses: Dawid (1987) suggests that the principle of the irrelevance of the stopping rule is one such survivor.

A second possible strategy is to fully accept S and C and Birnbaum's argument—and thereby come to accept L. This is the path of enlightenment followed by conversion.

The third strategy involves accepting S and C, but still rejecting L. If that is your motivation (and you care about self-consistency), you have no option but to try and find fault with the logic of Birnbaum's theorem. This is Mayo's strategy. The only problem is that Birnbaum's theorem is indeed logically sound. That means that Mayo's attempt to argue the contrary must itself be unsound. Although there are many points at which I am deeply critical of her argument, I will content myself with drawing attention to her principal misunderstanding, which vitiates her entire enterprise: she simply has not grasped Birnbaum's conditionality principle C, conflating and confusing it with Cox's WCP, which is quite different.

According to Mayo, WCP requires that "one should condition on the known experiment," or (as she phrases it in Section 4.3) "eschew unconditional formulations." But Birnbaum describes his principle C as the requirement that

A. P. Dawid is Emeritus Professor of Statistics, Statistical Laboratory, Centre for Mathematical Sciences, Cambridge University, Wilberforce Road, Cambridge CB3 0WB, UK (e-mail: a.p.dawid@statslab.cam.ac.uk).

> the evidential meaning of any outcome of any mixture experiment is the same as that of the corresponding outcome of the corresponding component experiment, ignoring the over-all structure of the mixture experiment.

That is, Birnbaum's principle C requires *identity* of the inferences to be drawn (from the same data) in different circumstances. This imposes an equivalence relationship across such circumstances. Principle C has nothing to say about the form or nature of the inferences, and—importantly—unlike WCP is entirely nondirectional. Mayo has misconstrued it as synonymous with WCP, which would require that we should discard whatever inference we might have been contemplating in the mixture experiment and replace it by our favored inference in the component experiment. However, an equally (in)valid reading of C would be the contrary: that we should discard a contemplated component-experiment inference in favor of an inference formed for the mixture experiment. In fact, neither of these interpretations has anything to do with principle C and, typically—as indeed follows from Birnbaum's theorem and the fact that frequentist inference violates C—neither of them can be implemented consistently within a frequentist framework.

In her Section 4.3.3 Mayo does consider the relationship between WCP and equivalence principles, and quite correctly decides that WCP is not one of these. In Section 7 she opines, "The problem stems from mistaking WCP as the equivalence...." So at least she realises that WCP and Birnbaum's principle C are different. However the "problem" is just the contrary: she has mistaken Birnbaum's equivalence requirement C as the "nonequivalence" principle WCP.

Mayo has attempted to argue that L does not follow from S and WCP. Notwithstanding the shortfalls in her arguments, I agree with that conclusion. The trouble is, it has nothing to do with Birnbaum's theorem. Mayo has been attacking a straw man, and Birnbaum's result, S & C ⇒ L, remains entirely untouched by her criticisms.

REFERENCES

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

Dawid, A. P. (1977). Conformity of inference patterns. In *Recent Developments in Statistics* (*Proc. European Meeting Statisticians, Grenoble*, 1976) 245–256. North-Holland, Amsterdam. MR0471123

Dawid, A. P. (1983). Statistical inference: I. In *Encyclopedia of Statistical Sciences* **4** (S. Kotz, N. L. Johnson and C. B. Read, eds.) 89–105. Wiley, New York.

Dawid, A. P. (1987). Invited discussion of "On principles and arguments to likelihood," by M. Evans, D. A. S. Fraser and G. Monette. *Canad. J. Statist.* **14** 196–197.

Dawid, A. P. (2011). Basu on ancillarity. In *Selected Works of Debabrata Basu. Sel. Works Probab. Stat.* 5–8. Springer, New York. MR2799327

Dawid, A. P. (2013). Principles of statistics. Online lecture notes at http://www.flooved.com/reader/3470.

Durbin, J. (1970). On Birnbaum's theorem on the relation between sufficiency, conditionality and likelihood. *J. Amer. Statist. Assoc.* **65** 395–398.

Fraser, D. A. S. (1963). On the sufficiency and likelihood principles. *J. Amer. Statist. Assoc.* **58** 641–647. MR0153078

Kalbfleisch, J. D. (1975). Sufficiency and conditionality. *Biometrika* **62** 251–268. MR0386075

Mayo, D. (2014). On the Birnbaum argument for the strong likelihood principle. *Statist. Sci.* **29** 227–239.

Statistical Science
2014, Vol. 29, No. 2, 242–246
DOI: 10.1214/14-STS471
Main article DOI: 10.1214/13-STS457

Discussion of "On the Birnbaum Argument for the Strong Likelihood Principle"

Michael Evans

Abstract. We discuss Birnbaum's result, its relevance to statistical reasoning, Mayo's objections and the result in [*Electron. J. Statist.* **7** (2013) 2645–2655] that the proof of this result doesn't establish what is commonly believed.

Key words and phrases: Sufficiency, conditionality, likelihood, statistical evidence.

1. INTRODUCTION

The result established in Birnbaum (1962), that if one accepts the frequentist principles of sufficiency (S) and conditionality (C), then one must accept the likelihood principle (L), has been an issue in the foundations of statistics for 50 years. Many statisticians and philosophers of science accept Birnbaum's theorem as a logical fact because the proof is simple and, if they follow a pure likelihood or Bayesian prescription for inference, it doesn't violate the way they think statistical analyses should be conducted. Many frequentist statisticians reject the result basically because they don't like the consequence that frequentist evaluations of statistical methodologies are irrelevant.

In the end, an acceptable theory of inference has to be based on sound logic with no appeals to ex cathedra principles. Any principles used as part of forming such a theory have to have strong justifications and produce results that are free of paradoxes and contradictions. For example, the principle of conditional probability, which says we replace $P(A)$ as the measure of belief that event A is true by $P(A \mid C)$ after being told that event C has occurred, seems like a basic principle of inference that, with careful application, is sound.

Does the likelihood principle carry the same weight in a theory of inference as the principle of conditional probability? We don't think so and we will later argue that a somewhat weakened version is really just a consequence of the principle of conditional probability. Given that such principles can have a significant influence on what we view as correct statistical reasoning, it is important to examine the justifications for the likelihood principle, and Birnbaum's theorem is commonly cited as such, to see if these are correct.

Another principle cited in Mayo's paper is the *principle of frequentism.* So what is the justification for this principle? Generally, this seems to be based on the belief that it typically produces sensible statistical methods, although, as we will subsequently discuss, the story seems incomplete and unclear. If the principle of frequentism is correct, we need to have a good argument for it and a much more complete development of the theory.

The relevance of frequentism to Mayo's paper lies in the author's position that Birnbaum's argument is basically a violation of the principle. An argument is provided for why the joint application of S and C used in Birnbaum's proof constitute such a violation. I accept Mayo's reasoning. In fact, I think it is somewhat similar to the argument put forward in Evans, Fraser and Monette (1986) that the applications of S and C in the proof are incorrect because S discards as irrelevant precisely the information used by C to form the conditional model. So the justifications for S and C contradict one another in the proof and this doesn't seem right. This contradiction is avoided if one adopts the principle put forward in Durbin (1970), that we should restrict to ancillaries that are functions of a minimal sufficient statistic, and then Birnbaum's proof fails.

Michael Evans is Professor, Department of Statistical Sciences, University of Toronto, 100 St. George St., Toronto, Ontario M5S 3G3, Canada (e-mail: mevans@utstat.utoronto.ca).

As we will discuss in Section 3, however, the issue in Birnbaum's argument is not really with S and C together, but rather with C itself and with what is actually proved. A very broad hint that this is the case is provided in Evans, Fraser and Monette (1986) where, using the same style of argument as Birnbaum, it is "proved" that accepting C alone is equivalent to accepting L. So Durbin's point doesn't save the day even if we accept it. Actually, I don't think the arguments in Mayo's paper, or in Evans, Fraser and Monette (1986), completely dispense with Birnbaum's theorem either. They just reinforce the unsettling feeling that something is wrong somewhere.

Section 3 contains an outline of Evans (2013) that, for me at least, definitively settles the issue of what is wrong with Birnbaum's result and does this mathematically. As will be apparent from Section 2, however, it is clear that I believe that a proper prior is a necessary part of formally correct statistical reasoning. So why would a Bayesian want to invalidate Birnbaum's result? This is because the result, as usually stated, is not logically correct. Any valid theory has to have sound, logical foundations and so we don't want any faulty reasoning being used to justify that theory. In fact, Birnbaum's result even misleads, as we've heard it said that model checking and checking for prior-data conflict violate the likelihood principle and so should not be carried out. Both of these activities are a necessary part of a statistical analysis. For this is how we deal, at least in part, with the subjectivity inherent in a statistical analysis due to the choices made by a statistician. This point, at least with respect to model checking, is also made in Mayo's paper and I think it is an excellent one.

For proper Bayesians, a form of the likelihood principle is a consequence of the principle of conditional probability, a far more important principle. Applying the principle of conditional probability to the joint probability model for the model parameter and data after observing the data, we have that *probability statements about the model parameter* depend on the sampling model and data only through the likelihood (note the emphasis). Of course, the likelihood map is minimal sufficient so there is nothing surprising in this.

2. BIRNBAUM AND EVIDENCE

There is an aspect of Birnbaum's work in this area that is particularly noteworthy. This is his emphasis on trying to characterize statistical evidence concerning the true value of the model parameter as expressed by the function Ev. Consider the pairs (M, x), where $M = \{f_\theta : \theta \in \Theta\}$ is a set of probability distributions indexed by parameter $\theta \in \Theta$ and x is observed data coming from a distribution in M. Then Birnbaum (1962) writes $Ev(M_1, x_1) = Ev(M_2, x_2)$ to mean that the evidence in (M_1, x_1) is the same as the evidence in (M_2, x_2) whenever certain conditions are satisfied. We require here that M_1 and M_2 have the same parameter space, but this can be weakened to include models with parameter spaces that are bijectively equivalent.

The principles S, C and L are considered as possible partial characterizations of statistical evidence. For example, if (M_1, x_1) and (M_2, x_2) are related via S, then Birnbaum says that, for frequentist statisticians, $Ev(M_1, x_1) = Ev(M_2, x_2)$ and similarly for C. Birnbaum is careful to say that Ev does not characterize what statistical evidence is, it is a kind of "equivalence relation" (see Section 3).

In essence Birnbaum brings us to the heart of the matter in statistical inference. What is statistical evidence or, more appropriately, how do we measure it? It seems collectively we talk about it, but we rarely get down to details and really spell out how we are supposed to handle this concept. Perhaps the closest to doing this is the pure likelihood theory, as discussed, for example, in Royall (1997), but this is only a definition of relative evidence when comparing two values of the full model parameter. For marginal parameters, this approach uses the profile likelihood as the only general way to compare the evidence for different values and this is unsatisfactory from many points of view. For example, a profile likelihood function is not generally a likelihood function.

For a frequentist theory of statistical inference, as opposed to a theory of statistical decision, it seems essential that a general method for measuring statistical evidence be provided that can be applied in any particular problem. The p-value is often used as a frequentist measure of evidence against a hypothesis, but, for a variety of reasons, it does not seem to be appropriate. For example, we need a measure that can also provide evidence *for* something being true and not just evidence against, given that we have assumed that the true distribution *is* in M.

If we add a proper prior to the ingredients, then it seems we can come up with sensible measures of evidence. For evidence, as expressed by observed data in statistical problems, is what causes beliefs to change and so we can measure evidence by measuring change in belief. For example, if we are interested in the truth of the event A, and this has prior probability $P(A) > 0$, then after observing C, the principle of conditional probability leads to the posterior probability $P(A \mid C)$ as the appropriate expression of beliefs about A. Accordingly, we measure evidence by the change in belief from $P(A)$ to $P(A \mid C)$. A simple *principle of evidence* says that we have evidence for the truth of A when $P(A \mid C) > P(A)$, evidence against the truth of A when $P(A \mid C) < P(A)$ and no evidence one way or the other when $P(A \mid C) = P(A)$. This principle is common in discussions about evidence in the philosophy of science and it seems obviously correct.

Of course, we also want to know how much evidence we have and this has led to a variety of different measures based on $(P(A), P(A \mid C))$. The Bayes factor $\mathrm{BF}(A \mid C) = P(A^c)P(A \mid C)P(A^c)/P(A)P(A^c \mid C)$ is one such measure, as $\mathrm{BF}(A \mid C) > 1$ if and only if $P(A \mid C) > P(A)$ and bigger values mean more evidence in favor of A being true. A central question associated with this, and other measures of evidence, is how to calibrate its values, as in when is $\mathrm{BF}(A \mid C)$ big and when is it small. Actually, we prefer measuring evidence via the relative belief ratio $\mathrm{RB}(A \mid C) = P(A \mid C)/P(A)$, as the associated mathematics and the calibration of its values are both simpler. The generalization to continuous contexts is effected by taking limits and then both measures agree. A full theory of inference, both estimation and hypothesis assessment, can be built based on this measure of evidence together with a very natural calibration. This is discussed in Baskurt and Evans (2013). Of course, many will not like this because it involves proper priors, and so is subjective and supposedly not scientific. Alternatively, some may complain that priors are somehow hard to come up with.

In reality, all of statistics, excepting the data when it is properly collected, is subjective. We choose models and we choose priors. What is important is that any choice we make, as part of a statistical analysis, be checkable against the objective data to ensure the choice at least make sense. We check the model by asking whether or not the data is surprising for each distribution in the model, and there are many well-known procedures for doing this. Perhaps not so familiar is that we can also check a proper prior by asking whether or not the true value is in a region of relatively low prior probability. Procedures for doing this consistently are developed in Evans and Moshonov (2006) and Evans and Jang (2011a). In fact, there are even logical approaches to modifying priors when prior-data conflict is found, as discussed in Evans and Jang (2011b). Moreover, with a suitable definition of evidence, we can measure a priori whether or not a prior is inducing bias into a problem; see Baskurt and Evans (2013). So subjectivity is not really the issue. We do our best to assess and control its effects, and maybe that is part of the role of statistics in science, but in the end it is an unavoidable aspect of any statistical investigation.

It is undoubtedly true that it is possible to write down complicated models for which it is extremely difficult, if not impossible, to prescribe an elicitation procedure in an application that leads to a sensible choice of a prior. But what does this say about our *choice* of model? It seems that we do not understand the effects of parameters in the model on the measurements we are taking sufficiently well to develop such a procedure. That is certainly possible, and perhaps even common, but it doesn't speak well for the modeling process and it shouldn't be held up as a criticism of what should be the gold standard for inference. An analogous situation arises with data collection where we know the gold standard is random sampling from the population(s) to which our inferences are to apply and, when we are interested in relationships among variables, controlled allocation of the values of predictors to sampled units. The fact that this is rarely, if ever, achieved doesn't cause us to throw out the baby with the dirty bath water. Gold standards serve as guides that we strive to attain and analyses that don't just need to be suitably qualified.

Our main point in this section is that the problem of measuring statistical evidence is the central issue in developing a theory of statistical inference. It seems that Birnbaum realized this and was searching for a way to accomplish this goal when he came upon what appeared to be a remarkable result.

3. WHAT'S WRONG WITH BIRNBAUM'S RESULT?

Perhaps everybody who has read the proof of Birnbaum's theorem is surprised at its simplicity. In fact, this is one of the reasons it is so convincing, as there does not appear to be a logical flaw in the proof. As Mayo has noted, however, there are reasons to be doubtful of, if not even reject, the result as being valid within the domain of any sensible theory of statistical inference. Still suspicions linger, as the formulation seems so simple.

As we will now explain, the result proved is not really the result claimed. If we want to treat Birnbaum's theorem and its proof as a piece of mathematics, then we have to be precise about the ingredients going into it. It is the imprecision in Birnbaum's formulation that leads to a faulty impression of exactly what is proved. This is more carefully explained in Evans (2013), but we can give a broad outline here.

Suppose we have a set D. A relation R on D is any subset $R \subset D \times D$. Meaningful relations express something and $(d_1, d_2) \in R$ means that d_1 and d_2 share some relevant property. Let $\mathcal{I}$ denote the set of all model–data pairs (M, x). So, for example, we can consider S as a relation on $\mathcal{I}$ by saying the pair $((M_1, x_1), (M_2, x_2)) \in S \subset \mathcal{I} \times \mathcal{I}$ whenever (M_1, x_1) and (M_2, x_2) have equivalent minimal sufficient statistics. Similarly, C and L are relations on $\mathcal{I}$.

An equivalence relation R on D is a relation that is reflexive: $(d, d) \in R$ for all $d \in D$, symmetric: $(d_1, d_2) \in R$ implies $(d_2, d_1) \in R$, and transitive: $(d_1, d_2), (d_2, d_3) \in R$ implies $(d_1, d_3) \in R$. It is reasonable to say that, whatever property is characterized by relation R, when R is an equivalence relation, then $(d_1, d_2) \in R$ means that d_1 and d_2 possess the property to the same degree. It is easy to prove that S and L are equivalence relations but C and $S \cup C$ are not equivalence relations; see Evans (2013).

Associated with an arbitrary relation R on D is the smallest equivalence relation on D containing R, which we will denote by $\bar{R}$. Clearly, $\bar{R}$ is the intersection of all equivalence relations containing R. But $\bar{R}$ can also be characterized in another way that is key to Birnbaum's proof.

LEMMA. *If R is a reflexive relation on D, then $\bar{R} = \{(x, y) : \exists n, x_1, \dots, x_n \in D$ with $x = x_1, y = x_n$ and $(x_i, x_{i+1}) \in R$ or $(x_{i+1}, x_i) \in R\}$.*

Note that S and C are both reflexive and, thus, $S \cup C$ is reflexive.

In Birnbaum's proof, he starts with $((M_1, x_1), (M_2, x_2)) \in L$, namely, these pairs have proportional likelihoods. Birnbaum constructs the mixture model (Birnbaumization) M^* and then argues that we have that $((M_1, x_1), (M^*, (1, x_1))) \in C$, $((M^*, (1, x_1)), (M^*, (2, x_2))) \in S$ and $((M^*, (2, x_2)), (M_2, x_2)) \in C$. Since $C \subset S \cup C$ and $S \subset S \cup C$, by the Lemma, this proves that $L \subset \overline{S \cup C}$ and this is all that Birnbaum's argument establishes. Since $S \cup C \subset L$ and L is an equivalence relation, we also have $L = \overline{S \cup C}$. As shown in Evans (2013), it is also true that $S \cup C$ is properly contained in L, so there is some content to the proof. In prose, Birnbaum's proof establishes the following: if we accept S, and we accept C, *and* we accept all the equivalences generated by these principles jointly, then we accept L. Certainly accepting S and C is not equivalent to accepting L since $S \cup C$ is a proper subset of L. We need the additional hypothesis and there doesn't appear to be any good reason why we should accept this as part of a theory of statistical inference. It is easy to construct relations R where $\bar{R}$ is meaningless. So we have to justify the additional pairs we add to a relation when completing it to be an equivalence relation.

It is interesting to note that the argument supposedly establishing the equivalence of C and L in Evans, Fraser and Monette (1986) also proceeds in the same way using the method of the Lemma. Since C is properly contained in L, this proof establishes that $\bar{C} = L$. So in fact, S is irrelevant in Birnbaum's proof. The problem with the principles S and C, as partial characterizations of statistical evidence, lies with C and the fact that it is not an equivalence relation. That C is not an equivalence relation is another way of expressing the well-known fact that, in general, a unique maximal ancillary doesn't exist.

The result $\bar{C} = L$ does have some content. To be a valid characterization of evidence in the context of $\mathcal{I}$, we will have to modify C so that it is an equivalence relation. The smallest equivalence relation containing C is L and this is unappealing, at least to frequentists, as it implies that repeated sampling properties are irrelevant for inference. Another natural candidate for a resolution is the largest equivalence relation contained in C that is compatible with all the equivalence relations based on maximal ancillaries. This is given by the equivalence relation based on the laminal ancillary. From Basu (1959), ancillary statistic a is a *laminal ancillary* if it is a function of every maximal ancillary and any other ancillary with this property is a function of a. The laminal ancillary is essentially unique. It is unclear how appealing this resolution would be to frequentists, but there don't seem to be any other natural candidates.

Many authors, including Mayo, refer to the weak conditionality principle which restricts attention to ancillaries that are physically part of the sampling. In such a case we would presumably write our models differently so as to reflect the fact that this sampling occurred in stages. In other words, the universe is different than $\mathcal{I}$, the one Birnbaum considered. There doesn't seem to be anything controversial about such a principle and it is well motivated by the two measuring instruments example and many others.

We don't believe, however, that the weak conditionality principle resolves the problem with conditionality more generally. For example, how does weak conditionality deal with situations like Example 2-2 in Fraser (2004) and many others like it? Conditioning on an ancillary seems absolutely essential if we are to obtain sensible inferences in such examples, but there doesn't appear to be any physical aspect of the sampling that corresponds to the relevant ancillary.

Many frequentist statisticians ignore conditionality, but, as noted in Fraser (2004), this is not logical. The theme in conditional inference is to find the right hypothetical sequence of repeated samples to compare the observed sample to. This takes us back to our question concerning the principle of frequentism: why are we considering repeated samples anyway? A successful frequentist theory of inference requires at least a resolution of the problems with conditionality. The lack of such a resolution leads to doubts as to the validity of the basic idea that underlies frequentism.

Issues concerning ancillaries are not irrelevant to Bayesians, as they have uses in model checking and checking for prior-data conflict. Notice that the principle of conditional probability does not imply that these activities need refer to any kind of posterior probabilities and it is perfectly logical for these to be based on prior probabilities. For example, model checking can be based on the distribution of an ancillary or the conditional distribution of the data given a minimal sufficient. Of course, C is not relevant for proper Bayesian probability statements about θ, as the principle of conditional probability implies that we condition on all of the data.

We acknowledge that it is possible that the problems with C might be fixable or even eliminated through a better understanding of what we are trying to accomplish in statistical analyses—these aren't just problems in mathematics. We can't resist noting, however, that the simple addition of a proper prior to the ingredients does the job, at least for inference.

4. CONCLUSIONS

Mayo's paper contains a number of insightful comments and, more generally, it helps to focus attention on what is the most important question in statistics, namely, what is the right way to formulate a statistical problem and carry out a statistical analysis? To a certain extent, Birnbaum's result has been an impediment in moving forward toward developing a theory of inference that has a solid foundation. It is good to have such underbrush removed from the discussion. We have to give great credit to Birnbaum, however, for his focus on what is important in achieving this goal, namely, the measurement of statistical evidence. That his theorem has lasted for so long is a testament to the difficulties involved in this task.

In general, we need a strong foundation for a theory of statistical inference rather than principles, often not clearly stated, that have only some vague, intuitive appeal. The only way we can determine whether or not an instance of statistical reasoning is correct lies within the context of a sound theory. That two statistical analyses based on the same data and addressing the same question can be deemed to be correct and yet come to different conclusions is not a contradiction. Statistics tells us that we simply must collect more data to resolve such differences. In our view, the role of statistics in science is to explain how to reason correctly in statistical contexts. Without a strong theory we can't do that.

REFERENCES

BASKURT, Z. and EVANS, M. (2013). Hypothesis assessment and inequalities for Bayes factors and relative belief ratios. *Bayesian Anal.* **8** 569–590. MR3102226

BASU, D. (1959). The family of ancillary statistics. *Sankhyā* **21** 247–256. MR0110115

BIRNBAUM, A. (1962). On the foundations of statistical inference. *J. Amer. Statist. Assoc.* **57** 296–326. MR0138176

DURBIN, J. (1970). On Birnbaum's theorem on the relation between sufficiency, conditionality and likelihood. *J. Amer. Statist. Assoc.* **65** 395–398.

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

EVANS, M. and JANG, G. H. (2011a). A limit result for the prior predictive applied to checking for prior-data conflict. *Statist. Probab. Lett.* **81** 1034–1038. MR2803740

EVANS, M. and JANG, G. H. (2011b). Weak informativity and the information in one prior relative to another. *Statist. Sci.* **26** 423–439. MR2917964

EVANS, M. and MOSHONOV, H. (2006). Checking for prior-data conflict. *Bayesian Anal.* **1** 893–914 (electronic). MR2282210

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

ROYALL, R. M. (1997). *Statistical Evidence. A Likelihood Paradigm. Monographs on Statistics and Applied Probability* **71**. Chapman & Hall, London. MR1629481

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 $\geq_{\mathrm{st}}$ 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

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

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

Statistical Science
2014, Vol. 29, No. 2, 252–253
DOI: 10.1214/14-STS473
Main article DOI: 10.1214/13-STS457
© Institute of Mathematical Statistics, 2014

Discussion: On Arguments Concerning Statistical Principles

D. A. S. Fraser

(i) *Statistical inference after Neyman–Pearson*. Statistical inference as an alternative to Neyman–Pearson decision theory has a long history in statistical thinking, with strong impetus from Fisher's research; see, for example, the overview in Fisher (1956). Some resulting concerns in inference theory then reached the mathematical statistics community rather forcefully with Cox (1958); this had focus on the two measuring-instruments example and on uses of conditioning that were compelling.

(ii) *Birnbaum and logical analysis in statistical inference*. Birnbaum (1962) introduced notation for the statistical inference available from an investigation with a model and data. This gave grounds to analyze how different methods or principles might influence the statistical inference. As part of this he discussed how sufficiency, likelihood and conditioning could differentially affect statistical inference. Much of his discussion centered on the argument from conditioning and sufficiency to likelihood, but a primary consequence was the attention attracted to conditioning and its role in inference. While this interest in conditioning was substantial for those concerned with the core of statistics, it has more recently been neglected or overlooked. Indeed, some recent texts, for example, Rice (2007), seem not to acknowledge conditioning in inference or even the measuring-instrument example.

(iii) *Mayo and statistical principles*. Mayo should be strongly commended for reminding us that the principles and arguments of statistical inference deserve very serious consideration and, we might add, could have very serious consequences (Fraser, 2014). Her primary focus is on the argument (Birnbaum, 1962) that the principles sufficiency and conditionality lead to the likelihood principle. This may not cover some recent aspects of conditioning (Fraser, Fraser and Staicu, 2010), but should strongly stimulate renewed interest in conditioning.

(iv) *Contemporary inference theory*. Many statistical models have continuity in how parameter change affects observable variables or, more specifically, how parameter change affects coordinate quantile functions, the inverses of the coordinate distribution functions. This continuity in its global effect is widely neglected in statistical inference. If this effect on quantile functions is accepted and used in the inference procedures, then in wide generality there is a well-determined conditioning (Fraser, Fraser and Staicu, 2010). And likelihood analysis then offers an exponential model approximation that is third-order equivalent to the given model, and this in turn provides third-order inference for any scalar component parameters of interest. Thus, the familiar conditioning conflicts are routinely avoided by acknowledging the important model continuity.

(v) *What is available?* The conditioning just described leads routinely to p-value functions $p(\psi)$ for any scalar component parameter $\psi = \psi(\theta)$ of the statistical model. A wealth of statistical inference methodology then immediately becomes available from such p-value functions. For example, a test for a value ψ_0 is given by the p-value $p(\psi_0)$, a confidence interval by the inverse $(\hat{\psi}_{\beta/2}, \hat{\psi}_{1-\beta/2}) = p^{-1}(1-\beta/2, \beta/2)$ of the p-value function, and a median estimate by the value $p^{-1}(0.5)$. But quite generally the needed p-value functions are not available from a likelihood function alone!

(vi) *What are the implications?* If continuity is included as an ingredient of many model-data combinations, then, as we have indicated, likelihood analysis produces p-values and confidence intervals, and these are not available from the likelihood function alone. This thus demonstrates that with such continuity-based conditioning the likelihood principle is not a consequence of sufficiency and conditioning principles. But if we omit the continuity then we are directly faced with the issue addressed by Mayo.

ACKNOWLEDGMENTS

Supported in part by the Natural Sciences and Engineering Research Council of Canada and Senior Scholars Funding at York University.

D. A. S. Fraser is Emeritus Professor, Department of Statistical Sciences, University of Toronto, 100 St. George St., Toronto, Ontario M5S 3G3, Canada (e-mail: dfraser@utstat.toronto.edu).

REFERENCES

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

COX, D. R. (1958). Some problems connected with statistical inference. *Ann. Math. Statist.* **29** 357–372. MR0094890

FISHER, R. A. (1956). *Statistical Methods and Scientific Inference*. Oliver and Boyd, Edinburgh.

FRASER, D. A. S. (2014). Why does statistics have two theories? In *Past, Present and Future of Statistical Science* (X. Lin, D. Banks, C. Genest, G. Molenberghs, D. Scott and J.-L. Wang, eds.) 237–252. CRC Press, Boca Raton, FL.

FRASER, A. M., FRASER, D. A. S. and STAICU, A.-M. (2010). Second order ancillary: A differential view from continuity. *Bernoulli* **16** 1208–1223. MR2759176

RICE, J. (2007). *Mathematical Statistics and Data Analysis*, 3rd ed. Brooks/Cole, Belmont, CA.

Statistical Science
2014, Vol. 29, No. 2, 254–258
DOI: 10.1214/14-STS474
Main article DOI: 10.1214/13-STS457

Discussion of "On the Birnbaum Argument for the Strong Likelihood Principle"

Jan Hannig

Abstract. In this discussion we demonstrate that fiducial distributions provide a natural example of an inference paradigm that does not obey Strong Likelihood Principle while still satisfying the Weak Conditionality Principle.

Key words and phrases: Generalized fiducial inference, strong likelihood principle violation, weak conditionality principle.

Professor Mayo should be congratulated on bringing new light into the veritable arguments about statistical foundations. It is well documented that p-values, confidence intervals and hypotheses tests do not satisfy the Strong Likelihood Principle (SLP). In the next section we will demonstrate that fiducial distributions provide a natural example of an inference paradigm that breaks SLP while still satisfying the Weak Conditionality Principle (WCP).

1. HISTORY OF FIDUCIAL INFERENCE

The origin of Generalized Fiducial Inference can be traced back to R. A. Fisher (Fisher, 1930, 1933, 1935) who introduced the concept of a fiducial distribution for a parameter and proposed the use of this fiducial distribution in place of the Bayesian posterior distribution. In the case of a one-parameter family of distributions, Fisher gave the following definition for a fiducial density $r(\theta)$ of the parameter based on a single observation x for the case where the cdf $F(x, \theta)$ is a decreasing function of θ:

$$r(\theta) = -\frac{\partial F(x, \theta)}{\partial \theta}. \tag{1.1}$$

For multiparameter families of distributions Fisher did not give a formal definition. Moreover, the fiducial approach led to confidence sets whose frequentist coverage probabilities were close to the claimed confidence levels but they were not exact in the frequentist sense. Fisher's proposal led to major discussions among the prominent statisticians of the 1930s, 40s and 50s (e.g., Dempster, 1966, 1968; Fraser, 1961a, 1961b, 1966, 1968; Jeffreys, 1940; Lindley, 1958; Stevens, 1950). Many of these discussions focused on the nonexactness of the confidence sets and also on the nonuniqueness of fiducial distributions. The latter part of the 20th century has seen only a handful of publications (Barnard, 1995; Dawid, Stone and Zidek, 1973; Salome, 1998; Dawid and Stone, 1982; Wilkinson, 1977) as the fiducial approach fell into disfavor and became a topic of historical interest only.

Since the mid-2000s, there has been a true resurrection of interest in modern modifications of fiducial inference. These approaches have become known under the umbrella name of *distributional inference*. This increase of interest came both in the number of different approaches to the problem and the number of researchers working on these problems, and manifested itself in an increasing number of publications in premier journals. The common thread for these approaches is a definition of inferentially meaningful probability statements about subsets of the parameter space without the need for subjective prior information.

These modern approaches include the Dempster–Shafer theory (Dempster, 2008; Edlefsen, Liu and Dempster, 2009) and its recent extension called *inferential models* (Martin, Zhang and Liu, 2010; Zhang and Liu, 2011; Martin and Liu, 2013a, 2013b, 2013c, 2013d). A somewhat different approach termed *confidence distributions* looks at the problem of obtaining an inferentially meaningful distribution on the pa-

Jan Hannig is Professor, Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, 330 Hanes Hall, Chapel Hill, North Carolina 27599-3260, USA (e-mail: jan.hannig@unc.edu).

rameter space from a purely frequentist point of view (Xie and Singh, 2013). One of the main contributions of this approach is the ability to combine information from disparate sources with deep implications for meta analysis (Schweder and Hjort, 2002; Singh, Xie and Strawderman, 2005; Xie, Singh and Strawderman, 2011; Hannig and Xie, 2012; Xie et al., 2013). Another more mature approach is called *objective Bayesian inference* that aims at finding nonsubjective model-based priors. An example of a recent breakthrough in this area is the modern development of reference priors (Berger, 1992; Berger and Sun, 2008; Berger, Bernardo and Sun, 2009; 2012; Bayarri et al., 2012). Another related approach is based on higher order likelihood expansions and implied data dependent priors (Fraser, Fraser and Staicu, 2010; Fraser, 2004, 2011; Fraser and Naderi, 2008; Fraser et al., 2010; Fraser, Reid and Wong, 2005). There is also important initial work showing how some simple fiducial distributions that are not Bayesian posteriors naturally arise within the decision theoretical framework (Taraldsen and Lindqvist, 2013).

Arguably, Generalized Fiducial Inference has been on the forefront of the modern fiducial revival. Starting in the early 1990s, the work of Tsui and Weerahandi (1989, 1991) and Weerahandi (1993, 1994, 1995) on *generalized confidence intervals* and the work of Chiang (2001) on the *surrogate variable method* for obtaining confidence intervals for variance components led to the realization that there was a connection between these new procedures and fiducial inference. This realization evolved through a series of works in the early 2000s (Hannig, 2009; Hannig, Iyer and Patterson, 2006; Iyer, Wang and Mathew, 2004; Patterson, Hannig and Iyer, 2004). The strengths and limitations of the fiducial approach are starting to be better understood; see especially Hannig (2009, 2013). In particular, the asymptotic exactness of fiducial confidence sets, under fairly general conditions, was established in Hannig (2013); Hannig, Iyer and Patterson (2006); Sonderegger and Hannig (2014). Generalized fiducial inference has also been extended to prediction problems in Wang, Hannig and Iyer (2012). Computational issues were discussed in Cisewski and Hannig (2012), Hannig, Lai and Lee (2014), and model selection in the context of Generalized Fiducial Inference has been studied in Hannig and Lee (2009); Lai, Hannig and Lee (2013).

2. GENERALIZED FIDUCIAL DISTRIBUTION AND THE WEAK CONDITIONALITY PRINCIPLE

Most modern incarnations of fiducial inference begin with expressing the relationship between the data, $\mathbf{X}$, and the parameters, ξ, as

$$\mathbf{X} = \mathbf{G}(\mathbf{U}, \xi), \tag{2.1}$$

where $\mathbf{G}(\cdot,\cdot)$ is termed the *data generating equation* (also called the association equation or structural equation) and $\mathbf{U}$ is the random component of this data generating equation whose distribution is free of parameters and completely known.

After observing the data $\mathbf{x}$ the next step is to use the known distribution of $\mathbf{U}$ and the inverse of the data (2.1) to define probabilities for the subsets of the parameter space. In particular, Generalized Fiducial Inference defines a distribution on the parameter space as the weak limit as $\varepsilon \to 0$ of the conditional distribution

$$\arg\min_{\xi} \|\mathbf{x} - \mathbf{G}(\mathbf{U}^\star, \xi)\| \mid \Big\{\min_{\xi} \|\mathbf{x} - \mathbf{G}(\mathbf{U}^\star, \xi)\| \le \varepsilon\Big\}, \tag{2.2}$$

where $\mathbf{U}^\star$ has the same distribution as $\mathbf{U}$. If there are multiple values minimizing the norm, the operator $\arg\min_\xi$ selects one of them (possibly at random). We stress at this point that the Generalized Fiducial Distribution is not unique. For example, different data generating equations can give a somewhat different Generalized Fiducial Distribution. Notice also that if $P(\min_\xi \|\mathbf{x} - \mathbf{G}(\mathbf{U}^\star, \xi)\| = 0) > 0$, which is the case for discrete distributions, the limit in (2.2) is the conditional distribution evaluated at $\varepsilon = 0$.

The conditional form of (2.2) immediately implies the Weak Conditional Principle for the limiting Generalized Fiducial Distribution. To demonstrate this, let us consider the two-instrument example of (Cox, 1958) (see also Section 4.1 of the discussed article). The data generating equation can be written in a hierarchical form:

$$M = 1 + I_{(0,1/2)}(U),$$
$$X = \theta + \sigma_M Z,$$

where $U \sim U(0,1)$ and $Z \sim N(0,1)$ are independent and the precisions $\sigma_1 << \sigma_2$ are known. If both $X = x$ the measurement made and $M = m$ the instrument used ($m = 1, 2$ for machine 1 and 2 respectively) are observed, the conditional distribution (2.2) is $N(x, \sigma_m^2)$, only taking into account the experiment actually performed. On the other hand, if only M is unobserved, then the conditional distribution (2.2) is the mixture $0.5N(\theta, \sigma_1^2) + 0.5N(\theta, \sigma_2^2)$. As claimed, the Generalized Fiducial Distribution follows WCP in this example.

3. GENERALIZED FIDUCIAL DISTRIBUTION AND THE STRONG LIKELIHOOD PRINCIPLE

In general, the Generalized Fiducial Distribution does not satisfy the Strong Likelihood principle. We first demonstrate this on inference for geometric distribution. To begin, we perform some preliminary calculations. Let X be a random variable with discrete distribution function $F(x, \xi)$. Let us assume for simplicity of presentation that for each fixed x, $F(x, \xi)$ is monotone in ξ and spans the whole $[0, 1]$. The inverse distribution function $F^{-1}(u, \xi) = \inf\{x : F(x, \xi) \geq u\}$ forms a natural data generating equation

$$X = F^{-1}(U, \xi), \quad U \sim (0, 1).$$

The minimizer in (2.2) is not unique, but any fiducial distribution will have a distribution function satisfying $1 - F(x, \xi) \leq H(\xi) \leq 1 - F(x_-, \xi)$ if $F(x, \xi)$ is decreasing in ξ and $F(x_-, \xi) \leq H(\xi) \leq F(x, \xi)$ if $F(x, \xi)$ is increasing. To resolve this nonuniqueness, Hannig (2009) and Efron (1998) recommend using the half correction which is the mixture distribution with distribution functions $H(\xi) = 1 - (F(x, \xi) + F(x_-, \xi))/2$ if $F(x, \xi)$ is decreasing in ξ or $H(\xi) = (F(x, \xi) + F(x_-, \xi))/2$ if $F(x, \xi)$ increasing.

Let us now consider observing a random variable $N = n$ following the Geometric(p) distribution. SLP implies that the inference based on observing $N = n$ should be the same as inference based on observing $X = 1$ where X is Binomial(n, p). However, the Geometric based Generalized Fiducial Distribution has a distribution function between $1 - (1 - p)^{n-1} \leq H_G(p) \leq 1 - (1 - p)^n$. The binomial based Generalized Fiducial Distribution uses bounds $1 - (1 - p)^n - np(1 - p)^{n-1} \leq H_B(p) \leq 1 - (1 - p)^n$. Thus, the effect of the stopping rule demonstrates itself in the Generalized Fiducial Inference through the lower bound that is much closer to the upper bound in the case of geometric distribution. (We remark that one cannot ignore the lower bound, as the upper bound is used to form upper confidence intervals and the lower bound is used for lower confidence intervals on p.) To conclude, the fiducial distribution in this example depends on both the distribution function of x and also on the distribution function of $x - 1$.

Let us now turn our attention to continuous distributions. In particular, assume that the parameter $\xi \in \Theta \subset \mathbb{R}^p$ is p-dimensional and that the inverse to (2.1) $\mathbf{G}^{-1}(\mathbf{x}, \xi) = \mathbf{u}$ exists. Then under some differentiability assumptions, Hannig (2013) shows that the generalized fiducial distribution is absolutely continuous with density

$$r(\xi) = \frac{f(\mathbf{x}, \xi) J(\mathbf{x}, \xi)}{\int_\Theta f(\mathbf{x}, \xi') J(\mathbf{x}, \xi')\, d\xi'}, \tag{3.1}$$

where $f(\mathbf{x}, \xi)$ is the likelihood and the function $J(\mathbf{x}, \xi)$ is

$$\begin{aligned} &J(\mathbf{x}, \xi) \\ &\quad = \sum_{\substack{\mathbf{i}=(i_1,\ldots,i_p) \\ 1 \leq i_1 < \cdots < i_p \leq n}} \left| \det\left(\frac{d}{d\xi} \mathbf{G}(\mathbf{u}, \xi)\Big|_{\mathbf{u}=\mathbf{G}^{-1}(\mathbf{x},\xi)} \right)_{\mathbf{i}} \right|, \end{aligned} \tag{3.2}$$

where $\frac{d}{d\xi}\mathbf{G}(\mathbf{u}, \xi)$ is the $n \times p$ Jacobian matrix of partial derivatives computed with respect of components of ξ. The sum in (3.2) spans over all p-tuples of indexes $\mathbf{i} = (1 \leq i_1 < \cdots < i_p \leq n) \subset \{1, \ldots, n\}$. Additionally, for any $n \times p$ matrix J, the sub-matrix $(J)_\mathbf{i}$ is the $p \times p$ matrix containing the rows $\mathbf{i} = (i_1, \ldots, i_p)$ of A. The form of (3.1) suggests that as long as the Jacobian $J(\mathbf{x}, \xi)$ does not separate into $J(\mathbf{x}, \xi) = f(\mathbf{x})g(\xi)$, in which case the Generalized Fiducial Distribution is the same as the Bayes posterior with $g(\xi)$ used as a prior, the Generalized Fiducial Distribution does not satisfy SLP due to the dependance on $d\mathbf{G}(\mathbf{u}, \xi)/d\xi$.

4. GENERALIZED FIDUCIAL DISTRIBUTION AND SUFFICIENCY PRINCIPLE

Whether the Generalized Fiducial Distribution satisfies the sufficiency principle depends entirely on what data generating equation is chosen. For example, let us assume that $\mathbf{Y} = (\mathbf{S}(\mathbf{X}), \mathbf{A}(\mathbf{X}))'$, where $\mathbf{S}$ is a p-dimensional sufficient and $\mathbf{A}$ is ancillary and $\mathbf{X}$ satisfies (2.1). Because $d\mathbf{A}/d\xi = 0$, the sum in (3.2) contains only one nonzero term:

$$J(\mathbf{x}, \xi) = \left| \det\left(\frac{d}{d\xi} \mathbf{S}(\mathbf{G}(\mathbf{u}, \xi))\Big|_{\mathbf{u}=G^{-1}(\mathbf{x},\xi)} \right) \right|. \tag{4.1}$$

Let $\mathbf{s} = \mathbf{S}(\mathbf{x})$ and $\mathbf{a} = \mathbf{A}(\mathbf{x})$ be the observed values of the sufficient and ancillary statistics respectively. To interpret the Generalized Fiducial Distribution assume that there is a unique ξ solving $\mathbf{s} = \mathbf{S}(\mathbf{G}(\mathbf{u}, \xi))$ for every $\mathbf{u}$ and denote this solution $Q_\mathbf{s}(\mathbf{u}) = \xi$. Also assume that the ancillary data generating equation $\mathbf{A}(\mathbf{G}(\mathbf{u}, \xi)) = \mathbf{A}(\mathbf{u})$ is not a function of ξ. A straightforward calculation shows that the fiducial density (3.1) with (4.1) is the conditional distribution of $Q_\mathbf{s}(\mathbf{U}^\star) \mid \mathbf{A}(\mathbf{U}^\star) = \mathbf{a}$. We conclude that this choice of data generating equation leads to inference based on sufficient statistics conditional on the ancillary. However, we still do not expect the SLP to hold in general even for this data generating equation.

Heuristically, this is because GFI is using not only the data observed, but also the data that based on the data generating equation could have been observed in the neighborhood of the observed data.

5. FINAL REMARKS

Let us close with discussing the example of Section 3.1. While the paper is not very clear on the exact specification of the events, it appears that for experiment 1 we observe the event

$$O_1 = \{\bar{y}_{169} > 1.96\sigma/\sqrt{169}\},$$

while for the experiment 2 we observe

$$O_2 = \{\bar{y}_k \le 1.96\sigma/\sqrt{k}, k = 1, \dots, 168, \\ \bar{y}_{169} > 1.96\sigma/\sqrt{169}\}.$$

Since $O_2 \subset O_1$, we see that the likelihood

$$P_\theta(O_2) = P_\theta(O_2 \mid O_1)P(O_1).$$

Consequently, we would have an SLP pair if and only if $P_\theta(O_2 \mid O_1)$ was a constant as a function of θ. However, this is not the case, as clearly $P_0(O_2 \mid O_1) > 0$ and $\lim_{\theta\to\infty} P_\theta(O_2 \mid O_1) = 0$. Consequently, we do not have an SLP pair.

ACKNOWLEDGMENTS

Supported in part by the National Science Foundation Grants 1007543 and 1016441.

REFERENCES

BARNARD, G. A. (1995). Pivotal models and the fiducial argument. *Internat. Statist. Rev.* **63** 309–323.

BAYARRI, M. J., BERGER, J. O., FORTE, A. and GARCÍA-DONATO, G. (2012). Criteria for Bayesian model choice with application to variable selection. *Ann. Statist.* **40** 1550–1577. MR3015035

BERGER, J. O. and BERNARDO, J. M. (1992). On the development of reference priors. In *Bayesian Statistics* 4 (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 35–60. Oxford Univ. Press, New York. MR1380269

BERGER, J. O., BERNARDO, J. M. and SUN, D. (2009). The formal definition of reference priors. *Ann. Statist.* **37** 905–938. MR2502655

BERGER, J. O., BERNARDO, J. M. and SUN, D. (2012). Objective priors for discrete parameter spaces. *J. Amer. Statist. Assoc.* **107** 636–648. MR2980073

BERGER, J. O. and SUN, D. (2008). Objective priors for the bivariate normal model. *Ann. Statist.* **36** 963–982. MR2396821

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

CISEWSKI, J. and HANNIG, J. (2012). Generalized fiducial inference for normal linear mixed models. *Ann. Statist.* **40** 2102–2127. MR3059078

COX, D. R. (1958). Some problems connected with statistical inference. *Ann. Math. Statist.* **29** 357–372. MR0094890

DAWID, A. P. and STONE, M. (1982). The functional-model basis of fiducial inference. *Ann. Statist.* **10** 1054–1074. MR0673643

DAWID, A. P., STONE, M. and ZIDEK, J. V. (1973). Marginalization paradoxes in Bayesian and structural inference. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **35** 189–233. MR0365805

DEMPSTER, A. P. (1966). New methods for reasoning towards posterior distributions based on sample data. *Ann. Math. Statist.* **37** 355–374. MR0187357

DEMPSTER, A. P. (1968). A generalization of Bayesian inference (with discussion). *J. R. Stat. Soc. Ser. B Stat. Methodol.* **30** 205–247. MR0238428

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

EDLEFSEN, P. T., LIU, C. and DEMPSTER, A. P. (2009). Estimating limits from Poisson counting data using Dempster–Shafer analysis. *Ann. Appl. Stat.* **3** 764–790. MR2750681

EFRON, B. (1998). R. A. Fisher in the 21st century (invited paper presented at the 1996 R. A. Fisher Lecture). *Statist. Sci.* **13** 95–122. MR1647499

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

FISHER, R. A. (1933). The concepts of inverse probability and fiducial probability referring to unknown parameters. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **139** 343–348.

FISHER, R. A. (1935). The fiducial argument in statistical inference. *Ann. Eugenics* **VI** 91–98.

FRASER, D. A. S. (1961a). On fiducial inference. *Ann. Math. Statist.* **32** 661–676. MR0130755

FRASER, D. A. S. (1961b). The fiducial method and invariance. *Biometrika* **48** 261–280. MR0133910

FRASER, D. A. S. (1966). Structural probability and a generalization. *Biometrika* **53** 1–9. MR0196840

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

FRASER, A. M., FRASER, D. A. S. and STAICU, A.-M. (2010). Second order ancillary: A differential view from continuity. *Bernoulli* **16** 1208–1223. MR2759176

FRASER, D. A. S. and NADERI, A. (2008). Exponential models: Approximations for probabilities. *Biometrika* **94** 1–9.

FRASER, D., REID, N. and WONG, A. (2005). What a model with data says about theta. *Internat. J. Statist. Sci.* **3** 163–178.

FRASER, D. A. S., REID, N., MARRAS, E. and YI, G. Y. (2010). Default priors for Bayesian and frequentist inference. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **72** 631–654. MR2758239

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

HANNIG, J., IYER, H. and PATTERSON, P. (2006). Fiducial generalized confidence intervals. *J. Amer. Statist. Assoc.* **101** 254–269. MR2268043

HANNIG, J., LAI, R. C. S. and LEE, T. C. M. (2014). Computational issues of generalized fiducial inference. *Comput. Statist. Data Anal.* **71** 849–858. MR3132011

HANNIG, J. and LEE, T. C. M. (2009). Generalized fiducial inference for wavelet regression. *Biometrika* **96** 847–860. MR2767274

HANNIG, J. and XIE, M.-G. (2012). A note on Dempster–Shafer recombination of confidence distributions. *Electron. J. Stat.* **6** 1943–1966. MR2988470

IYER, H. K., WANG, C. M. J. and MATHEW, T. (2004). Models and confidence intervals for true values in interlaboratory trials. *J. Amer. Statist. Assoc.* **99** 1060–1071. MR2109495

JEFFREYS, H. (1940). Note on the Behrens–Fisher formula. *Ann. Eugenics* **10** 48–51. MR0002080

LAI, R. C. S., HANNIG, J. and LEE, T. C. M. (2013). Generalized fiducial inference for ultra high dimensional regression. Available at arXiv:1304.7847.

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

MARTIN, R. and LIU, C. (2013a). Conditional inferential models: Combining information for prior-free probabilistic inference. Preprint.

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

MARTIN, R. and LIU, C. (2013c). Marginal inferential models: prior-free probabilistic inference on interest parameters. Preprint.

MARTIN, R. and LIU, C. (2013d). On a 'plausible' interpretation of p-values. Preprint.

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

PATTERSON, P., HANNIG, J. and IYER, H. K. (2004). Fiducial generalized confidence intervals for proportion of conformance. Technical Report 2004/11, Colorado State Univ., Fort Collins, CO.

SALOME, D. (1998). Staristical inference via fiducial methods. Ph.D. thesis, Univ. Groningen.

SCHWEDER, T. and HJORT, N. L. (2002). Confidence and likelihood. *Scand. J. Stat.* **29** 309–332. Large structured models in applied sciences; challenges for statistics (Grimstad, 2000). MR1909788

SINGH, K., XIE, M. and STRAWDERMAN, W. E. (2005). Combining information from independent sources through confidence distributions. *Ann. Statist.* **33** 159–183. MR2157800

SONDEREGGER, D. and HANNIG, J. (2014). Fiducial theory for free-knot splines. In *Contemporaly Developments in Statistical Theory, a Festschrift in Honor of Professor Hira L. Koul* (T. N. Sriraus, ed.) 155–189. Springer, Berlin.

STEVENS, W. L. (1950). Fiducial limits of the parameter of a discontinuous distribution. *Biometrika* **37** 117–129. MR0035955

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

TSUI, K.-W. and WEERAHANDI, S. (1989). Generalized p-values in significance testing of hypotheses in the presence of nuisance parameters. *J. Amer. Statist. Assoc.* **84** 602–607. MR1010352

TSUI, K.-W. and WEERAHANDI, S. (1991). Corrections: "Generalized p-values in significance testing of hypotheses in the presence of nuisance parameters". [*J. Amer. Statist. Assoc.* **84** (1989) 602–607. MR1010352]. *J. Amer. Statist. Assoc.* **86** 256. MR1137115

WANG, C. M., HANNIG, J. and IYER, H. K. (2012). Fiducial prediction intervals. *J. Statist. Plann. Inference* **142** 1980–1990. MR2903406

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

WEERAHANDI, S. (1994). Correction: "Generalized confidence intervals". [*J. Amer. Statist. Assoc.* **88** (1993) 899–905. MR1242940]. *J. Amer. Statist. Assoc.* **89** 726. MR1294096

WEERAHANDI, S. (1995). *Exact Statistical Methods for Data Analysis*. *Springer Series in Statistics*. Springer, New York. MR1316663

WILKINSON, G. N. (1977). On resolving the controversy in statistical inference. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **39** 119–171. MR0652326

XIE, M.-G. and SINGH, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter: A review. *Internat. Statist. Rev.* **81** 3–39. MR3047496

XIE, M., SINGH, K. and STRAWDERMAN, W. E. (2011). Confidence distributions and a unifying framework for meta-analysis. *J. Amer. Statist. Assoc.* **106** 320–333. MR2816724

XIE, M., LIU, R. Y., DAMARAJU, C. V. and OLSON, W. H. (2013). Incorporating external information in analyses of clinical trials with binary outcomes. *Ann. Appl. Stat.* **7** 342–368. MR3086422

ZHANG, J. and LIU, C. (2011). Dempster–Shafer inference with weak beliefs. *Statist. Sinica* **21** 475–494. MR2829843

Statistical Science
2014, Vol. 29, No. 2, 259–260
DOI: 10.1214/14-STS475
Main article DOI: 10.1214/13-STS457

Discussion of "On the Birnbaum Argument for the Strong Likelihood Principle"

Jan F. Bjørnstad

Abstract. The paper by Mayo claims to provide a new clarification and critique of Birnbaum's argument for showing that sufficiency and conditionality principles imply the likelihood principle. However, much of the arguments go back to arguments made thirty to forty years ago. Also, the main contention in the paper, that Birnbaum's arguments are not valid, seems to rest on a misunderstanding.

Key words and phrases: Likelihood, conditionality, sufficiency, Birnbaum's theorem.

The goal of this paper is to provide a new clarification and critique of Birnbaum's argument for showing that principles of sufficiency and conditionality entail the (strong) likelihood principle (LP).

I must admit I do not find that the paper provides such a new clarification of the criticism of Birnbaum's argument. Rather, much of the criticism in the paper goes back to arguments made in the 70s and 80s by several authors, for example, Durbin (1970), Kalbfleisch (1975), Cox (1978) and Evans, Fraser and Monette (1986). This critique has been discussed by several statisticians with an opposing view; see Berger and Wolpert (1988) and Bjørnstad (1991).

I will concentrate my discussion on what seems to be the most important contention in the paper, that the sufficient statistic in Birnbaum's proof erases the information as to which experiment the data came from and, hence, that the weak conditionality principle (WCP) cannot be applied; see, for example, Sections 5.2 and 7.

As I understand it, this is a misunderstanding of the proof. For one thing, it seems that only the observation x_2 from the experiment E_2 is considered in the mixture experiment instead of the correct (E_2, x_2). The observations in a mixture experiment are always of the form (E_h, x_h)—*never* as only x_h.

Other arguments leading up to this contention seem to rest on a misunderstanding of the sufficiency considerations in the proof, that given an observation from a certain experiment the result in an unperformed experiment is to be reported; see, for example, Sections 2.4 and 5.1. I find that this is definitely not the case. To be specific, the author considers the following proof in the discrete case:

Let (j, x_j) indicate that Experiment E_j was performed with data x_j, $j = 1, 2$. Assume the data values x_1^0 and x_2^0 have proportional likelihoods from experiments E_1 and E_2, respectively. Then the sufficient statistic in the mixture experiment used in Birnbaum's proof is given by

$$T(j, x_j) = (j, x_j) \quad \text{if } (j, x_j) \neq (1, x_1^0), (2, x_2^0),$$
$$T(1, x_1^0) = T(2, x_2^0) = (1, x_1^0).$$

Mayo claims that $T(1, x_1^0) = T(2, x_2^0)$ $(= c)$ implies that the weak conditionality principle (WCP) is violated. Now, one should note that the proof works with any $c \neq (1, x_1^0)$ and $c \neq (j, x_j)$, $j = 1, 2$ and all x_j. When E_2 is performed and x_2^0 is the result, then the evidence should only depend on E_2 and x_2^0, and not on a result of an unperformed experiment E_1. This is, of course, correct, but it does not depend on a result in E_1. (Actually E_1 is not an unperformed experiment either. We comment on this issue later.) By letting $c = (1, x_1^0)$

Jan F. Bjørnstad is Professor of Statistics, Department of Mathematics, University of Oslo and Head of Research, Division for Statistical Methods, Statistics Norway, P.O. Box 8131 Dep., N-0033 Oslo, Norway (e-mail: jab@ssb.no).

it seems so, but we see by choosing a $c \neq (1, x_1^0)$ it is not the case. So WCP is not violated. The sufficient statistic simply takes the same value for these two results of the mixture experiment. It has nothing to do with WCP.

So Birnbaum's proof does not require that the evidential support of a known result should depend on the result of an unperformed experiment. It follows that the main contention in the paper seems to rest on a lack of understanding of the basics of the proof of Birnbaum's theorem. In fact, it is possible to do the proof even more generally. One can show, for a given experiment [see, e.g., Cox and Hinkley (1974) and Bjørnstad (1996)], that if two likelihoods are proportional for two possible observations in the same experiment, there exists a minimal sufficient statistic with the same value for the two observations. This holds both for discrete and continuous models.

To make the sufficiency argument clearer, consider a mixture of a binomial experiment E_1 and a negative binomial experiment E_2 where the observations are $x_1 =$ number of successes in 12 trials and $x_2 =$ number of trials until 3 successes. If $x_1^0 = 3$ and $x_2^0 x_2 = 12$ then the likelihoods are proportional. A natural choice of the sufficient statistic T in the mixture experiment in Birnbaum's proof has

$$T(1, x_1^0) = T(2, x_2^0) = 3/12,$$

the proportion of successes in either case.

Clearly then, the value of T from experiment E_2 does not depend on the result from E_1.

As already mentioned, the author claims that the sufficient statistic T in the proof of Birnbaum's result has the effect of erasing the index of the experiment. Moreover, it is claimed that inference based on T is to be computed over the performed and unperformed experiments E_1 and E_2. As we have shown, this is simply not the case. It should also be mentioned that statistically the proof simply considers two instances of performing the mixture experiments resulting in proportional likelihoods and really has nothing to do with considering unperformed experiments.

Let me also mention that the author's premise in Section 5 is not correct. The starting point is *not* that we have an arbitrary outcome of one single experiment, but rather that two experiments have been performed about the same parameter resulting in proportional likelihoods. So Birnbaum does not enlarge a known single experiment but constructs a mixture of the two performed experiments. There is really *no unperformed* experiment here. In a sense, one may regard the paper by Mayo as actually not discussing the LP at all.

It should be clear that I find that the main contention in the paper does not hold when maintaining the original meaning of the principles of sufficiency, conditionality and likelihood. Other comments made in this paper referring to various authors in the 70s and 80s are a different matter. However, I do not find any *new* clarification of Birnbaum's fundamental theorem in this paper. For example, regarding sufficiency, it is necessary to restrict the application of sufficiency to nonmixture experiments, as Kalbfleisch (1975) did, in order to invalidate Birnbaum's result. Berger and Wolpert (1988) argue, I think, convincingly against such a restriction. See also Bjørnstad (1991).

Let me end this discussion by making clear the following fact: It is obviously clear that frequentistic measures may, and typically do, violate LP. This is true as far as it comes to analysis of the actual data we observe. But a major point here is that the LP does not say that one should not be concerned with how the methods do when used repeatedly. LP is simply *not about method evaluation*. Evaluation of methods is still important. So LP says in essence that frequentistic considerations are not *sufficient* for evaluating the uncertainty and reliability in the statistical analysis of the actual data; see also Bjørnstad (1996) for a discussion on this issue.

REFERENCES

BERGER, J. O. and WOLPERT, R. L. (1988). *The Likelihood Principle*, 2nd ed. *Lecture Notes—Monograph Series* **6**. IMS, Hayward, CA.

BJØRNSTAD, J. F. (1991). Introduction to Birnbaum (1962): On the foundations of statistical inference. In *Breakthroughs in Statistics* 1 (S. Kotz and J. Johnson, eds.) 461–477. *Springer Series in Statistics*. Springer, New York.

BJØRNSTAD, J. F. (1996). On the generalization of the likelihood function and the likelihood principle. *J. Amer. Statist. Assoc.* **91** 791–806. MR1395746

COX, D. R. (1978). Foundations of statistical inference: The case for eclecticism. *Aust. N. Z. J. Stat.* **20** 43–59. MR0501453

COX, D. R. and HINKLEY, D. V. (1974). *Theoretical Statistics*. Chapman & Hall, London. MR0370837

DURBIN, J. (1970). On Birnbaum's theorem in the relation between sufficiency, conditionality and likelihood. *J. Amer. Statist. Assoc.* **65** 395–398.

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

KALBFLEISCH, J. D. (1975). Sufficiency and conditionality. *Biometrika* **62** 251–268. MR0386075

Statistical Science
2014, Vol. 29, No. 2, 261–266
DOI: 10.1214/14-STS482
Main article DOI: 10.1214/13-STS457

Rejoinder: "On the Birnbaum Argument for the Strong Likelihood Principle"

Deborah G. Mayo

1. INTRODUCTION

I am honored and grateful to have so many interesting and challenging comments on my paper. I want to thank the discussants for their willingness to jump back into the thorny quagmire of Birnbaum's argument. To a question raised in the paper "Does it matter?", these discussions show the answer is yes. The enlightening connections to contemporary projects are especially valuable in galvanizing future efforts to address foundational issues in statistics.

As long-standing as Birnbaum's result has been, Birnbaum himself went through dramatic shifts in a short period of time following his famous (1962) result. More than of historical interest, these shifts provide a unique perspective on the current problem. Already in the rejoinder to Birnbaum (1962), he is worried about criticisms (by Pratt, 1962) pertaining to applying WCP to his constructed mathematical mixtures (what I call Birnbaumization), and hints at replacing WCP with another principle (Irrelevant Censoring). Then there is a gap until around 1968 at which point Birnbaum declares the SLP plausible "only in the simplest case, where the parameter space has but two" predesignated points [Birnbaum (1968), page 301]. He tells us in Birnbaum (1970a, page 1033) that he has pursued the matter thoroughly, leading to "rejection of both the likelihood concept and various proposed formalizations of prior information." The basis for this shift is that the SLP permits interpretations that "can be seriously misleading with high probability" [Birnbaum (1968), page 301]. He puts forward the "confidence concept" (Conf) which takes from the Neyman–Pearson (N–P) approach "techniques for systematically appraising and bounding the probabilities (under respective hypotheses) of seriously misleading interpretations of data" while supplying it an evidential interpretation [Birnbaum (1970a), page 1033]. Given the many different associations with "confidence," I use (Conf) in this Rejoinder to refer to Birnbaum's idea. Many of the ingenious examples of the incompatibilities of SLP and (Conf) are traceable back to Birnbaum, optional stopping being just one [see Birnbaum (1969)]. A bibliography of Birnbaum's work is Giere (1977). Before his untimely death (at 53), Birnbaum denies the SLP even counts as a principle of evidence (in Birnbaum, 1977). He thought it anomalous that (Conf) lacked an explicit evidential interpretation, even though, at an intuitive level, he saw it as the "one rock in a shifting scene" in statistical thinking and practice [Birnbaum (1970a), page 1033]. I return to this in Section 4 of this rejoinder.

2. BJØRNSTAD, DAWID AND EVANS

Let me begin by answering the central criticisms that, if correct, would be obstacles to what I purport to have shown in my paper. It is entirely understandable that leading voices in a long-lived controversy would assume that all of the twists and turns, avenues and roadways, have already been visited, and that no new flaw in the argument could enter to shake up the debate. I say to the reader that the surest sign that the issue is unsettled is that my critics disagree among themselves about the puzzle and even the key principles under discussion: the WCP, and in one case, the SLP itself. To remind us [Section 2.2]:

> SLP: For any two experiments E_1 and E_2 with different probability models f_1, f_2 but with the same unknown parameter θ, if outcomes $\mathbf{x}^*$ and $\mathbf{y}^*$ (from E_1 and E_2, resp.) determine the same likelihood function $[f_1(\mathbf{x}^*;\theta) = cf_2(\mathbf{y}^*;\theta)$ for all $\theta]$, then $\mathbf{x}^*$ and $\mathbf{y}^*$ should be inferentially equivalent for any inference concerning parameter θ.

A shorthand for the entire antecedent is that $(E_1, \mathbf{x}^*)$ is an *SLP pair* with $(E_2, \mathbf{y}^*)$, or just $\mathbf{x}^*$ and $\mathbf{y}^*$ form an SLP pair (from $\{E_1, \mathrm{E}_2\}$). Assuming all the SLP stipulations, we have

> SLP: If $(E_1, \mathbf{x}^*)$ and $(E_2, \mathbf{y}^*)$ form an SLP pair, then $\mathrm{Infr}_{E_1}[\mathbf{x}^*] = \mathrm{Infr}_{E_2}[\mathbf{y}^*]$.

Deborah G. Mayo is Professor of Philosophy, Department of Philosophy, Virginia Tech, 235 Major Williams Hall, Blacksburg, Virginia 24061, USA (e-mail: mayod@vt.edu).

Bjørnstad

According to Bjørnstad, “The starting point is *not* that we have an arbitrary outcome of one single experiment, but rather that two experiments have been performed about the same parameter resulting in proportional likelihoods.” I do not think Bjørnstad can actually mean to say the SLP cannot be applied until both members of the SLP pair are observed. So, for example, if in the sequential experiment one is able to stop (with a 0.05 p-value) at $n = 169$, resulting in $\mathbf{y}^*$, one may not regard it as evidentially equivalent to $\mathbf{x}^*$, the SLP pair with n fixed at 169, until and unless $\mathbf{x}^*$ is actually generated? The universal generalization of the SLP asserts that for an arbitrary $\mathbf{y}^*$, if it has an SLP pair $\mathbf{x}^*$, then $\mathbf{y}^*$ is equivalent in evidence to $\mathbf{x}^*$. Bjørnstad’s problematic reading results in his next remark: “So Birnbaum does not enlarge a known single experiment but constructs a mixture of the two performed experiments.” What is constructed in Birnbaum’s experiment E_B is a *hypothetical or mathematical mixture*, based on having observed $\mathbf{y}^*$ (from E_2). This is part of the key gambit I call *Birnbaumization* (Section 2.4). We are to consider the possibility that performing E_2 (which gave rise to $\mathbf{y}^*$) was the result of a θ-irrelevant randomizer (deciding between E_1 or E_2). Now I grant Birnbaum that we may imagine all the SLP pairs are “out there,” each pair assumed to have resulted from a θ-irrelevant randomizer, ripe for plucking whenever a member of an SLP pair is observed. (See Sections 2.5 and 5.1.) Yet even granting Birnbaum all of this, we still may not infer SLP (nor does it follow in the case where the mixture is actual).

Bjørnstad also criticizes me because he claims the SLP “is simply *not about method evaluation*.” His position is that there is an evidential appraisal, and quite separately an assessment of long-run performance. For a frequentist, or one who holds Birnbaum’s (Conf), evidential import is inseparable from an assessment of the relevant error probabilities. Not because we regard evidential import as all about long-runs, but because scrutinizing a given inference is bound up with a method’s ability to have alerted us to misleading interpretations.

Bjørnstad does “not find any *new* clarification of Birnbaum’s fundamental theorem in this paper” because he assumes I am channeling the attempts of Durbin (1970), Kalbfleisch (1975) and Evans, Fraser and Monette (1986), all of whom restrict or modify either SP or WCP to block the result. While I stand on the shoulders of these and other earlier treatments, a crucial difference is that, unlike them, I do not alter the principles involved. If one is out to demonstrate the logical flaw in an argument, as every good philosopher knows, one should scrupulously adhere to the premises and generously interpret the machinations of the arguer. This I do. Bjørnstad’s opinion is that “one may regard the paper by Mayo as actually not discussing the LP at all.” Or, alternatively, one may regard the position held by this critic to be mistaken about the SLP and Birnbaum’s argument.

Dawid

Professors Dawid and Evans disagree about the key principle invoked in Birnbaum’s argument, the WCP. Dawid views it as an equivalence relation, Evans says it is not. I follow Birnbaum in regarding the WCP as an equivalence, but, unlike both Dawid and Evans, I pin down what is to be meant in regarding WCP as an equivalence, or, for that matter, an inequivalence (see Section 4.3). First Dawid.

Dawid maintains that my WCP differs from the principle of conditionality Birnbaum uses in the SLP argument. Not so. I am working with the WCP stated in Birnbaum (1962, 1969), the very same one defined by Dawid:

> The evidential meaning of any outcome of any mixture experiment is the same as that of the corresponding outcome of the corresponding component experiment, ignoring the over-all structure of the mixture experiment.

Dawid’s definition is a portion of the one found in Birnbaum (1962), page 271. It assumes, of course, all of the other stipulations, for example, we are making “informative” inferences about θ, it is a θ-irrelevant mixture, the outcome is given, and all the rest. It is the definition used in countless variations of the SLP argument, and it is clearly captured in my Section 4.3. Perhaps I should have abbreviated it as CP; WCP comes from Berger and Wolpert’s (1988) manifesto, *The Likelihood Principle*. My intention was to underscore Birnbaum’s emphasis that the WCP concerns mixture experiments and is distinct from many other uses of “conditioning” in statistics [Birnbaum (1962), pages 282–283].

I wondered why Dawid thought I denied that WCP asserts an equivalence, until I noticed that Dawid lops off the end of my sentence from Section 7: I do not say “the problem stems from mistaking WCP as the equivalence” simpliciter, but rather it stems from the incorrect equivalence! The incorrect equivalence equates the inference from the given experiment with one that takes account of the (irrelevant) mixture structure. This is what Dawid is on about in describing invalid construals of WCP, so he can scarcely object. (See Section 5.2.)

As with any equivalence, there is an implicit inequivalence as a corollary. [See (i) and (ii) in Section 4.3.] Typically, in saying the evidential import of two outcomes are the same, one would not add "and be sure to ignore any features that would render them inequivalent." Birnbaum adds this warning because some treatments do not ignore the mixture structure. To put this another way, WCP includes the phrases "is the same as" as well as "ignoring." The problem is that Dawid is ignoring the word "ignoring" in the very definition he proffers. There is no difference between the phrases

- ignore the over-all structure of the mixture experiment

and

- eschew any construal that does not ignore the overall structure of the mixture experiment.

I also refer to this as irrelevance (Irrel) (Section 4.3.2) because Birnbaum describes the WCP as asserting the "irrelevance of (component) experiments not actually performed" [Birnbaum (1962), page 271].

Dawid opines that I am using the WCP in David Cox's (1958) famous weighing example, which he does not define; I am guessing he means to suggest I must be limiting myself to actual mixtures. That is to miss the genius of Birnbaum's argument. Birnbaum, quite deliberately, intends to capitalize on the persuasiveness of conditioning in Cox's famous example, but his ploy is to extend the argument to mathematical or hypothetical mixtures. (I am not saying it is an innocuous move, but that is a separate matter.) Even if Dawid chooses to view Cox's WCP as a nonequivalence, it is irrelevant; I am following Birnbaum in construing it as an equivalence, permitting, for example, $\mathbf{y}^*$, known to have come from a nonmixture, to be evidentially equivalent to the appropriate θ-irrelevant mixture as in Section 4.3. (Irrel) protects against illicit readings that Dawid warns against. SLP still will not follow.

So Dawid, Birnbaum and I are using the same definition of WCP. The onus is on Dawid to pinpoint where my characterization deviates from Birnbaum's. The only difference is that I have shown one cannot get to SLP, and Dawid gives no clue how to get around my criticism.

For Dawid to simply pronounce that "Birnbaum's theorem is indeed logically sound" and that therefore my argument "must itself be unsound" is question-begging, and will not do. Demonstrating unsoundness of my argument should be accomplished straightforwardly, as I have done regarding Birnbaum. That said, I fully agree with Dawid that one can view [(SP and WCP), entails SLP] as a theorem, but in order to *detach* the SLP, as is mandatory for Birnbaum, he is left with a "proof" that is either unsound or question-begging. Perhaps those who are long wedded to Birnbaum's argument are comfortable with merely assuming what was to have been shown. It is part of the mysterious "path of enlightenment followed by conversion" that Dawid mentions. That is no reason for others to allow "trust me, it is sound" to take the place of argument.

Evans

Given that Evans largely agrees with me, it may seem ungenerous to focus on apparent disagreements, but there is too great a danger in leaving some misimpressions regarding a problem already beset with decades of misunderstanding. Notably, it seems I have not convinced Evans of the logical error that Birnbaum makes. Instead Evans thinks the problem is with the conditionality principle WCP, and claims that frequentists need to fix it somehow. But it is not the principle, it is the "proof."

I have at least convinced Evans that there are cases where SP and WCP and not-SLP hold without logical contradiction (in Mayo, 2010). These cases may be called "counterexamples" to the argument whose conclusion is the SLP. They are also counterexamples to [WCP entails SLP], using the weaker notion of mathematical equivalence of Birnbaum (1972) that dispenses with (SP). Evans will take those counterexamples to show that WCP is not an equivalence relation, assuming a frequentist standpoint. Now it is true that any such counterexample may be seen to warn us against mistaking WCP as asserting the incorrect equivalence, noted in my rejoinder to Dawid. But that does not preclude WCP from asserting a correct equivalence. A more general issue I have with Evans' treatment is that it does not show where the source of the problem lies in arguments for SLP. Introducing his set-theoretic treatment into a simple argument, I am afraid, does not help to pinpoint where the argument goes wrong, but in fact leaves us with a very murky idea even as to his definition of WCP. The argument for the SLP begins with: We are given $\mathbf{y}^*$ from E_2, a member of an

SLP pair. Will Evans block introduction of the mathematical mixture in Birnbaumization? This would seem to cut off Birnbaum's argument too quickly. Were that sufficient, the debate would have surely ended with Kalbfleisch (1975). Note too, unlike Evans, my argument in the paper under discussion does not rely on assuming a frequentist principle at all, though obviously I avoid a formulation that rules it out in advance. To sum up this section, Evans uses my counterexamples to show a restricted WCP may be applied, while blocking SLP. Left as it is, it opens him to the criticism (the one Dawid raised!) that he is altering Birnbaum's WCP and restricting it to actual mixtures.

What a surprise, then, to hear Evans allege that "many authors, including Mayo, refer to the [WCP] which restricts attention to ancillaries that are physically part of the sampling." I do not know on what grounds Evans wants to distinguish actual and mathematical mixtures, but Birnbaum's argument for the SLP concerns mathematical or hypothetical mixtures. Birnbaum calls an experiment a mixture "if it is mathematically equivalent" to a mixture [Birnbaum (1962), page 279]. Further, Birnbaum (1962) emphasizes that earlier proofs [that WCP and SP imply SLP] were restricted to actual mixtures. "But in the above proof" he is able to get a result relevant for all classes of experiments by using an ancillary "constructed with the hypothetical mixture" $[E_B]$ [Birnbaum (1962), page 286]. So, I am not sure what Evans is alleging. In one place, Evans worries whether the WCP "resolves the problem with conditionality more generally," but this is a separate issue from Birnbaum's argument. Here the focus is on WCP solely for purposes of arriving at the SLP.

Although there was not space to discuss this in my paper, it is worth noting why merely blocking the SLP with a modified WCP fails to make progress with a further goal required of an adequate treatment. Consider how, in discussing Durbin's modified principle of conditionality, Birnbaum notes that "Durbin's formulation (C'), although weaker than [WCP], is nevertheless too strong (implies too much of the content of [SLP]) to be compatible with standard (non-Bayesian) statistical concepts and techniques" [Birnbaum (1970b), page 402]. Birnbaum (1975, page 264) raises the same problem with Kalbfleish's restriction to "minimal experiments" to which Evans' treatment is closely related. Evans does not show his modified conditionality principle avoids entailing "too much of the SLP." (This relates to Dawid's point about stopping rules in his comment.) For a frequentist account to satisfy Birbaum's (Conf), all cases that allow misleading interpretations with high probability should still show up as SLP violations.

To this end, my argument shows that any violation of SLP in frequentist sampling theory necessarily results in an illicit substitution in the formulation of Birnbaum's argument. To put the problem in general terms, $p = r$ does not follow from $p = q$ and $q = r$, if q shifts to q' within the argument, where $q \neq q'$ (fallacy of 4 terms). For specifics see Section 5. Thus, ours is in no danger of implying "too much" of the SLP: what was an SLP violation remains one. Now Evans may not be concerned with retaining those frequentist SLP violations, given he makes it very clear he embraces Bayesianism, but that is irrelevant to what an adequate treatment of Birnbaum's argument demands. I have seen some statistics textbooks leave the details of the SLP proof to the reader; I think it is time to give full credit to students who found it impossible to make a valid substitution in general. I explained why.

3. FRASER, HANNIG, MARTIN AND LIU

Let me turn to the second group of discussants. It is an honor to be "strongly commended" by Fraser for emphasizing the importance of "principles and arguments of statistical inference"; and I feel my efforts are worthwhile with Martin and Liu's noting my "demonstration resolves the controversy around Birnbaum and LP, helping to put the statisticians' house in order." I entirely agree with them that the "confusion surrounding Birnbaum's claim has perhaps discouraged researchers from considering questions about the foundations of statistics," at least from appealing to those foundations that reject the SLP. Let me underscore Fraser's point that the need for an inferential variation of (N–P) theory "reached the mathematical statistics community rather forcefully with Cox (1958); this had the focus on the two measuring-instruments example and on uses of conditioning that were compelling." Cox's (1958) example also appears in Hannig's discussion, and I will borrow his simple description of the case where the measurement but not the instrument M is observed. In that case, inference is based on the convex combination of the mixture components, consistent with WCP. This allows me to succinctly put an equivocation that I suspect may enter, in the case of SLP pairs, between the irrelevance of the mixture structure, given $(E_i, \mathbf{z}_i)$, and the irrelevance of the index i, given just the measurement. This equivocation may be behind the Birnbaum puzzle.

Fraser rightly reminds us that, "statistical inference as an alternative to (N–P) decision theory has a long history in statistical thinking" with strong impetus from Fisher. Still Birnbaum struggled to articulate a N–P theory as "an inference theory" (Birnbaum, 1977), and my view is that we had to solve "Birnbaum's problem" before doing so properly. Finding Birnbaum's argument unsound opens the door to foundations that are free from paying obeisance to the SLP. In this spirit Martin and Liu correctly view my paper as "an invitation for a fresh discussion on the foundations of statistical inference." Yet there is more than one way of proceeding. Tracing out the mathematical similarities and differences between the approaches of Fraser, Hanning, Martin and Liu is a task for which others are better equipped than I. All are said to violate SLP.

It is interesting to note, as Hannig does, that "since the mid 2000s, there has been a true resurrection of interest in modern modifications of fiducial inference" which had long fallen into disrepute. Fraser's has been one of the leading voices to persevere with innovative developments, and his own "confidence" idea is clearly in sync with Birnbaum. However, the differences that emerge in this group's discussions should not be downplayed. Hannig says that "the common thread for these approaches is a definition of inferentially meaningful probability statements about subsets of the parameter space without the need for subjective prior information," and Martin and Liu suggest that error probability accounts are appropriate only for decision procedures, as distinct from their "inferential models." Some might view these as attempts to build a concept of evidence as a kind of *probabilism* but without the priors. However, in the background of these contemporary developments lurks a suspicion that their SLP violations were picking up differences where no purely inferential difference was warranted. So long as Birnbaum's proof stood, this suspicion made sense.

Post-SLP, it is worth standing back and reflecting anew on these accounts. In this respect, this foundational project is just beginning because for 40 or 50 years, the questions of foundations were largely restricted to accounts that obeyed, or were close to obeying, the SLP. So, we have Birnbaum, alongside Fisher, being catapulted onto the contemporary foundational scene, squarely calling on us to address the still unresolved problem: how to obtain an account of statistical inference that also controls the probability of seriously misleading inferences. Better yet, the two goals should mesh into one.

4. POST-SLP FOUNDATIONS

Return to where we left off in the opening section of this rejoinder: Birnbaum (1969),

> The problem-area of main concern here may be described as that of determining precise *concepts of statistical evidence* (systematically linked with mathematical models of experiments), concepts which are to be *non-Bayesian, non-decision-theoretic*, and significantly *relevant to statistical practice*. [Birnbaum (1969), page 113.]

Given Neyman's behavioral decision construal, Birnbaum claims that "when a confidence region estimate is interpreted as representing statistical evidence about a parameter" [Birnbaum (1969), page 122], an investigator has necessarily adjoined a concept of evidence, (Conf) that goes beyond the formal theory. What is this evidential concept? The furthest Birnbaum gets in defining (Conf) is in his posthumous article Birnbaum (1977):

> (Conf) A concept of statistical evidence is not plausible unless it finds 'strong evidence for H_2 against H_1' with small probability (α) when H_1 is true, and with much larger probability ($1 - \beta$) when H_2 is true. [Birnbaum (1977), page 24.]

On the basis of (Conf), Birnbaum reinterprets statistical outputs from N–P theory as strong, weak, or worthless statistical evidence depending on the error probabilities of the test [Birnbaum (1977), pages 24–26]. While this sketchy idea requires extensions in many ways (e.g., beyond pre-data error probabilities and beyond the two hypothesis setting), the spirit of (Conf), that error probabilities quantify properties of methods which in turn indicate the warrant to accord a given inference, is, I think, a valuable shift of perspective. This is not the place to elaborate, except to note that my own twist on Birnbaum's general idea is to appraise evidential warrant by considering the capabilities of tests to have detected erroneous interpretations, a concept I call *severity*. That Birnbaum preferred a propensity interpretation of error probabilities is not essential. What matters is their role in picking up how features of experimental design and modeling alter a methods' capabilities to control "seriously misleading interpretations." Even those who embrace a version of probabilism may find a distinct role for a severity concept. Recall that Fisher always criticized the presupposition that a single use of mathematical probability must be competent for qualifying inference in all logical situations [Fisher (1956), page 47].

Birnbaum's philosophy evolved from seeking concepts of evidence in degree of support, belief or plausibility between statements of data and hypotheses to embracing (Conf) with the required control of misleading interpretations of data. The former view reflected the logical empiricist assumption that there exist context-free evidential relationships—a paradigm philosophers of statistics have been slow to throw off. The newer (post-positivist) movements in philosophy and history of science were just appearing in the 1970s. Birnbaum was ahead of his time in calling for a philosophy of science relevant to statistical practice; it is now long overdue!

> "Relevant clarifications of the nature and roles of statistical evidence in scientific research may well be achieved by bringing to bear in systematic concert the scholarly methods of statisticians, philosophers and historians of science, and substantive scientists..." [Birnbaum (1972), page 861].

REFERENCES

BERGER, J. O. and WOLPERT, R. L. (1988). *The Likelihood Principle*, 2nd ed. *Lecture Notes—Monograph Series* **6**. IMS, Hayward, CA.

BIRNBAUM, A. (1962). On the foundations of statistical inference. *J. Amer. Statist. Assoc.* **57** 269–306. Reprinted in *Breakthroughs in Statistics* **1** (S. Kotz and N. Johnson, eds.) 478–518. Springer, New York.

BIRNBAUM, A. (1968). Likelihood. In *International Encyclopedia of the Social Sciences* **9** 299–301. Macmillan and the Free Press, New York.

BIRNBAUM, A. (1969). Concepts of statistical evidence. In *Philosophy, Science, and Method: Essays in Honor of Ernest Nagel* (S. Morgenbesser, P. Suppes and M. G. White, eds.) 112–143. St. Martin's Press, New York.

BIRNBAUM, A. (1970a). Statistical methods in scientific inference. *Nature* **225** 1033.

BIRNBAUM, A. (1970b). On Durbin's modified principle of conditionality. *J. Amer. Statist. Assoc.* **65** 402–403.

BIRNBAUM, A. (1972). More on concepts of statistical evidence. *J. Amer. Statist. Assoc.* **67** 858–861. MR0365793

BIRNBAUM, A. (1975). Comments on "Sufficiency and conditionality" by J. D. Kalbfleisch. *Biometrika* **62** 262–264.

BIRNBAUM, A. (1977). The Neyman–Pearson theory as decision theory, and as inference theory; with a criticism of the Lindley–Savage argument for Bayesian theory. *Synthese* **36** 19–49. MR0652320

COX, D. R. (1958). Some problems connected with statistical inference. *Ann. Math. Statist.* **29** 357–372. MR0094890

DURBIN, J. (1970). On Birnbaum's theorem on the relation between sufficiency, conditionality and likelihood. *J. Amer. Statist. Assoc.* **65** 395–398.

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. (1956). *Statistical Methods and Scientific Inference*. Oliver and Boyd, Edinburgh.

GIERE, R. N. (1977). Allan Birnbaum's conception of statistical evidence. *Synthese* **36** 5–13. MR0494585

KALBFLEISCH, J. D. (1975). Sufficiency and conditionality. *Biometrika* **62** 251–268. MR0386075

MAYO, D. G. (2010). An error in the argument from conditionality and sufficiency to the likelihood principle. In *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science* (D. G. Mayo and A. Spanos, eds.) 305–314. Cambridge Univ. Press., Cambridge.

PRATT, J. W. (1962). On the foundations of statistical inference: Discussion. *J. Amer. Statist. Assoc.* **57** 314–316.