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, 10.1214/14-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_{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
>
> $$(E_{\text{mix}}, \mathbf{z}_i) \Rightarrow \mathsf{Infr}_{E_i}\big[(E_{\text{mix}}, \mathbf{z}_i)\big] = p_i(\mathbf{z}) = \mathsf{Infr}_{E_i}[\mathbf{z}_i].$$
>
> Do not use the unconditional formulation
>
> $$(E_{\text{mix}}, \mathbf{z}_i) \not\Rightarrow \mathsf{Infr}_{E_{\text{mix}}}\big[(E_{\text{mix}}, \mathbf{z}_i)\big] = \big[p_1(\mathbf{z}) + p_2(\mathbf{z})\big]/2.$$
>
> The concern is that
>
> $$\mathsf{Infr}_{E_{\text{mix}}}\big[(E_{\text{mix}}, \mathbf{z}_i)\big] = \big[p_1(\mathbf{z}) + p_2(\mathbf{z})\big]/2 \neq p_i(\mathbf{z}).$$

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:

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

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

(ii) Eschew unconditional formulations:

$$(E_{\text{mix}}, \mathbf{z}_i) \not\Rightarrow \mathsf{Infr}_{E_{\text{mix}}}[\mathbf{z}_i],$$

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,

$$(E_1 \text{ or } E_2, \mathbf{z}) \not\Rightarrow \mathsf{Infr}_{E_1}[\mathbf{z}].$$

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,

$$(E_2, \mathbf{y}^*) \not\Rightarrow \mathsf{Infr}_{E_B}[\mathbf{y}^*],$$

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.