On the Birnbaum Argument for the Strong Likelihood Principle

Deborah G. Mayo[1]

Abstract

An essential component of inference based on familiar frequentist notions p-values, significance and confidence levels, is the relevant sampling distribution (hence the term *sampling theory*). This feature results in violations of a principle known as the *strong likelihood principle* (SLP), the focus of this paper. In particular, if outcomes **x*** and **y*** from experiments E_1 and E_2 (both with unknown parameter θ), have different probability models f_1, f_2, then even though $f_1(\mathbf{x}^*; \theta) = cf_2(\mathbf{y}^*; \theta)$ for all θ, outcomes **x*** and **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 (1958) 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 the particular E_i.

The surprising upshot of Allan Birnbaum's (1962) 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 directly refute [WCP entails SLP].

Key words: strong likelihood principle, mixture experiment, Birnbaumization, weak conditionality principle, sufficiency principle

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 notions, p-values, significance levels, and so on violate the SLP, the Birnbaum argument claims the SLP follows from principles that frequentist sampling theorists accept!

The SLP says, in effect, that given the data and the model, all the information (for parametric inference) is in the likelihood function:

> **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 respectively) determine the same likelihood function ($f_1(\mathbf{x}^*; \theta) = cf_2(\mathbf{y}^*; \theta)$ for all θ), then $\mathbf{x}^*$ and $\mathbf{y}^*$ should be inferentially equivalent**.**

[1] Department of Philosophy, Major Williams Hall 235, Virginia Tech, Blacksburg VA 24061.

Now an essential component of frequentist inference is the relevant sampling distribution (hence the term *sampling theory*). But this feature renders it incompatible with the SLP.

> 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, 300)

The incompatibility, in a nutshell, is that on the SLP, once the data **x** are given, outcomes other than **x** are irrelevant to the evidential import of **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 **x**."(Berger and Wolpert 1988, 105).

Although the sampling theorist considers 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 (1958) 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 the particular E_i.

The surprising upshot of Allan Birnbaum's (1962) argument is that the SLP appears to follow from applying the WCP in the case of experimental 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 directly refute [WCP entails SLP].

We follow the formulations of the Birnbaum argument given in Berger and Wolpert (1988), Birnbaum (1962), and Casella and R. Berger (2002), and D. R. Cox (1977). The current analysis clarifies and fills in the 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.

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 must be flawed somewhere. Even those 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.

Our discussion should also serve to 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, 148), since their prior probability distributions are influenced by the sampling distribution. "This, of course, does not happen with subjective Bayesianism" (J. Berger 2006, 394). As Savage stressed:

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

1. 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 **z**, where E is an experiment involving the observation of **Z** with a given distribution $f(\mathbf{z};\theta)$ and a model M. We use:

> $\text{Infr}_E(\mathbf{z})$: the parametric statistical inference from a given (E, **z**).

An inference method indicates how to compute the parametric inference from given $(E, \mathbf{z})$. Let:

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

The abbreviation $\text{Infr}_E[\mathbf{z}]$, 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.

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 $\text{Infr}_E[\mathbf{z}_1] = \text{Infr}_E[\mathbf{z}_2]$. To apply a given inference rule means its particular inference directive is used, as defined by ⇒, not some competing directive at the same time. This ensures non-contradiction, for any $(E, \mathbf{z})$: $\text{Infr}_E[\mathbf{z}] = \text{Infr}_E[\mathbf{z}]$.

2.2 *The SLP and its violations*

The principle under dispute, the SLP, involves 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 $f_1(.)$, $f_2(.)$ but with the same unknown parameter θ, if outcomes $\mathbf{x}^*$ and $\mathbf{y}^*$ (from E_1 and E_2 respectively) determine the same likelihood function ($f_1(\mathbf{x}^*; \theta) = cf_2(\mathbf{y}^*; \theta)$ for all θ), then $\mathbf{x}^*$ and $\mathbf{y}^*$ should be inferentially equivalent**.**

A shorthand for the entire antecedent is that (E_1, **x***) is a *SLP pair* with (E_2, **y***), or just **x*** and **y*** form an SLP *pair* (from $\{E_1,E_2\}$). Experimental pairs E_1 and E_2 involve observing random variables **X** and **Y**, respectively. Thus (E_1, **y***) or just **y*** asserts "E_2 is performed and **y*** observed". We may abbreviate $\text{Infr}_{E_2}[E, \mathbf{y}^*]$ as $\text{Infr}_{E_2}[\mathbf{y}^*]$. Likewise for **x***. Assuming the SLP stipulations, e.g., that θ is a shared parameter, we have:

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

An SLP violation occurs when $\mathbf{x}^*$ and $\mathbf{y}^*$ form an SLP pair, but $\text{Infr}_{E_1}[\mathbf{x}^*] \neq \text{Infr}_{E_2}[\mathbf{y}^*]$.

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. There may be no SLP violation if the focus is on point against point hypotheses, while in computing a 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, 104-5). The following is a dramatic example.

Fixed versus sequential sampling: Suppose **X** and **Y** are samples from distinct experiments E_1 and E_2, both distributed as N(μ, σ^2), with σ identical and known, and p-values are to be calculated for the null hypothesis H_0: $\mu = 0$ against H_1: $\mu \neq 0$. In E_2 the sampling rule is: continue sampling until: $|\bar{y}_n| > c_\alpha = 1.96\sigma/(n^{1/2})$. In E_1, the sample size n is fixed, and $\alpha = 0.05$. Suppose that E_2 is run and stops with $n = 169$ trials; this is $\mathbf{y}^*$. A choice for its SLP pair $\mathbf{x}^*$ is E_1 with $n = 169$, that happens to yield significance (E_1, $1.96\sigma/13$). The SLP violation is the fact that: $\text{Infr}_{E_1}[1.96\sigma/13] \neq \text{Infr}_{E_2}[n = 169]$. Note that to arrive at the SLP pair we have to consider the *particular* outcome observed in E_2.

> [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 'significance' against H_0. (Berger and Wolpert 1988, 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, 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, 54), since it contradicts what Cox and Hinkley call "the *weak repeated sampling principle*" (Cox and Hinkley 1974, 51). Under the sampling theory philosophy, 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 H_0 (Mayo and Cox 2006, 2010; Mayo 1996; Mayo and Spanos 2006).

2.3 Sufficiency Principle (Weak Likelihood Principle)

The Sufficiency Principle (SP) is often called the *weak* likelihood principle, limited as it is to a single experiment E, with its sampling distribution. If T_E is a (minimal) sufficient statistic for E, the *Sufficiency Principle* asserts:

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

Since inference within the model is to be computed using the value of T_E 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.

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, 491). Thus, accounts that adhere to the SLP are not thereby prevented from analyzing features of the data such as residuals, which are relevant to questions of checking the statistical model itself. There is some ambiguity on this point in Casella and R. Berger (2002):

> Most model checking is, necessarily, based on statistics other than a sufficient statistic... 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 R. Berger 2002, 295-6)

We regard the principles as inapplicable, rather than violated, with inadequate models.

Can two become one? The SP suggests that if an SLP pair **x***, **y*** could be seen as coming 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, embedded in his larger gambit to which we now turn. We call the larger gambit *Birnbaumization*.

2.4 Birnbaumization: Key gambit in the strategy

An experiment has been run, label it as E_2, and $\mathbf{y}^*$ observed. Suppose $\mathbf{y}^*$ has an SLP pair $\mathbf{x}^*$ in a distinct experiment E_1. Birnbaum must show the two are evidentially equivalent.

We are to imagine E_2 was the result of a type of mixture experiment: we flipped 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, 54), E_B. Because it is not an actual mixture, Birnbaum calls it a "mathematical" mixture. We are to define a statistic T_B that stipulates: If $\mathbf{y}^*$ is observed, its SLP pair $\mathbf{x}^*$ in the unperformed experiment is reported.

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

When $\mathbf{y}^*$ is observed, T_B reports $\mathbf{x}^*$. In effect the report is: the result could have come from E_1, *or* E_2. It is reported just as if we do not know which experiment generated the result. Thus, the inference in E_B under Birnbaumization for $\mathbf{y}^*$, as for $\mathbf{x}^*$is:

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

which is to be computed using the convex combination over E_1 and E_2 (the two experiments that might have generated $\mathbf{y}^*$. It follows that, *within* E_B, $\mathbf{x}^*$ and $\mathbf{y}^*$ are inferentially equivalent.

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

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

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

But Birnbaum does not stop there; we are to use a (weak) conditionality principle to "condition back down" to the known experiment E_2. But this will not produce the SLP as we now show.

3. The Weak Conditionality Principle (WCP)

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 "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, 582).

3.1 Mixture (E_{mix}) Two instruments of different precisions Cox (1958)

We flip a fair coin to decide whether to use a very precise or imprecise tool: E_1 or E_2. The *WCP says simply: Once it is known which E_i produced* **z***, the p-value or other inferential assessment should be made with reference to the experiment actually run*.

Example: in observing a normally distributed **Z** in testing a null hypothesis $\theta = 0$, let E_1 have variance of 1, while that of E_2 is 10^6. The same **z** measurement corresponds to a much smaller p-value if from E_1 rather than E_2: $p_1(\mathbf{z})$ and $p_2(\mathbf{z})$, respectively. (See Birnbaum 1962, 491).

The *overall* (or *unconditional*) significance level of the mixture E_{mix} is the convex combination of the p-values: $[ap_1(\mathbf{z}) + bp_2(\mathbf{z})]$. This would give a misleading report of how stringent the actual experimental measurement is (Cox and Mayo 2010, 296).

Suppose that we know we have 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, 361).

So WCP appears obviously correct. Yet Birnbaum's result purports that WCP (+ sufficiency) entails SLP.

> 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, 298)

But the illusion is not so easy to dispel; thus this paper. Let us state the WCP.

3.2 Weak Conditionality Principle (WCP) in the measuring example

WCP: Given $(E_{mix},\mathbf{z}_i)$: Condition on the E_i producing the result:

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

Do not use the unconditional formulation:

$$(E_{mix}, \mathbf{z}_i) \nRightarrow \text{Infr}_{E_{mix}}[\mathbf{z}_i] = [\text{a}p_1(\mathbf{z}) + \text{b}p_2(\mathbf{z})].$$

The concern is that:

$$\text{Infr}_{E_{mix}}[\mathbf{z}_i] = [\text{a}p_1(\mathbf{z}) + \text{b}p_2(\mathbf{z})] \neq p_{i.}$$

The WCP says: eschew unconditional formulations whenever $\text{Infr}_{E_{mix}}[\mathbf{z}_i] \neq \text{Infr}_{E_i}[\mathbf{z}_i]$. There are *three* sampling distributions and the WCP says the relevant one to use is the one known to have generated the result (Birnbaum 1962, 491).

3.3 The general 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.

(i) *Condition to obtain the relevant sampling distribution*:

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

(ii) *Eschew unconditional formulations*:

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

whenever the unconditional treatment yields a different inference implication:

i.e., whenever $\text{Infr}_{E_{mix}}[\mathbf{z}_i] \neq \text{Infr}_{E_i}[\mathbf{z}_i]$.

Recall: $\text{Infr}_{E_{mix}}[\mathbf{z}]$ alludes to the associated convex combination of the relevant pair of experiments. We now highlight some points for reference.

WCP makes a difference. The cases of interest here are where applying WCP would alter the unconditional implication. Here, WCP makes a difference to the parametric inference. Note that (ii) blocks computing the inference implication from $(E_{mix}, \mathbf{z}_i)$ as $\text{Infr}_{E_{mix}}[\mathbf{z}_i]$ when $\text{Infr}_{E_{mix}}[\mathbf{z}_i] \neq \text{Infr}_{E_i}[\mathbf{z}_i]$. Here E_1, E_2, and E_{mix} 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. We might write this:

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

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, 263). The question 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, i.e., when $\text{Infr}_{E_{mix}}[\mathbf{z}_i] \neq \text{Infr}_{E_i}[\mathbf{z}_i]$. 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. Nevertheless stipulation (i) of the WCP is an equivalence: If (E_i, $\mathbf{z}_i$) is known to have come from a θ-irrelevant mixture:

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

4. Birnbaum's Argument

4.1 Birnbaumization and the WCP

Some have described the Birnbaum experiment as unperformable, or at most a "mathematical mixture" rather than an "experimental mixture" (Kalbfleisch, 1975, 252-253). Birnbaum himself calls it a "hypothetical" mixture. While a holder of the WCP may deny it applies to hypothetical mixtures, since Birnbaum's argument has stood for over fifty years, we wish to give it maximal mileage. We may imagine a hypothetical universe of SLP pairs, each generated from a θ-irrelevant mixture. When we observe **y*** we pluck the x* companion needed for the argument. So, we can Birnbaumize a result: Constructing statistic T_B with derived experiment E_B is the "performance". But what cannot shift in the argument is that E_i be *given* (as noted in 3.3); that *i* be fixed.

Given **z***, the WCP precludes Birnbaumizing. On the other hand, if the reported **z*** was the value of T_B, then we are given only the disjunction, precluding the computation relevant for *i* fixed. Let us consider the components of Birnbaum's argument.

4.2 Main elements of Birnbaum's argument

It is given that $\mathbf{y}^*$ is observed from E_2, and it has an SLP pair $\mathbf{x}^*$. Birnbaum must show SLP: $\text{Infr}_{E_2}[\mathbf{y}^*] = \text{Infr}_{E_1}[\mathbf{x}^*]$. He will apply, in some order, both (1) Birnbaumization and (2) WCP, to the known $\mathbf{y}^*$. A tension immediately results:

(1) If the inference is by Birnbaumization E_B:

$\mathbf{y}^* \Rightarrow \text{Infr}_{E_B}[\mathbf{x}^*] = \text{Infr}_{E_B}[\mathbf{y}^*]$.

Likewise for $\mathbf{x}^*$. T_B is a sufficient statistic for E_B (the conditional distribution of **Z**-the sample relating to E_{mix}—given T_B, is independent of θ).

(2) If the inference is by WCP:

$\mathbf{y}^* \not\Rightarrow \text{Infr}_{E_B}[\mathbf{y}^*]$, rather

$\mathbf{y}^* \Rightarrow \text{Infr}_{E_2}[\mathbf{y}^*]$ and $\mathbf{x}^* \Rightarrow \text{Infr}_{E_1}[\mathbf{x}^*]$.

To apply (1) is at odds with applying (2). We will not get:

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

The SLP only seems to follow from erroneously identifying:

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

4.3 The logical refutation of (SP and WCP entails SLP)

We can uphold both if-then claims in (1) and (2), while at the same time hold:

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

(3) is true in any case where **x*** and **y*** form a SLP violation pair. That is precisely the case when the antecedent of the SLP holds. Since whenever (3) holds, we have a counterexample to the SLP generalization, this shows 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.[1]

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 self-contradictions. By non-contradiction (for any E, $\mathbf{z}$):

$\text{Infr}_E[\mathbf{z}] = \text{Infr}_E[\mathbf{z}]$.

("⇒" is a function from outcomes to inference implications, and $\mathbf{z} = \mathbf{z}$, for any $\mathbf{z}$.)

Upholding and applying. As noted in Section 2.1, applying a rule means following its inference directive. We may uphold the stipulations in (1) and (2), but $\text{Infr}_{E_B}[\mathbf{y}^*]$ cannot have conflicting references within same argument, if it is to be sound. Note that SP is not blocked in (1). The SP is always relative to a model, here E_B.

5. 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 the 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 **y***, a member of an SLP pair with **x***, with different probability models $f_1(.)$, $f_2(.)$, that he will be able to derive from SP and WCP, that **x*** and **y*** should have the identical inference 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 refers to 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 **y*** its SLP pair could occur in repetitions, but the key point of the WCP is to deny that this fact should alter the

inference implication from the known **y***. It is Birnbaum who purports to give an argument that is relevant for a sampling theorist, and for "approaches which are independent of Bayes' principle" (1962, 495). Its implications for sampling theory is why it was dubbed "a landmark in statistics" (Savage 1962b, 307-8).

Let us look at the two statements about inference implications from a given (E_2, **y***), applying (1) and (2) in 4.2:

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

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

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

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

Now the SLP must be applicable to any SLP pair. However, to assume that (1) and (2) can be consistently applied for *any* **x***, **y*** pair would be to assume no SLP violations are possible, which would render Birnbaum's argument circular. So the choices are to regard Birnbaum's argument as unsound or circular (assuming what it purports to prove). 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 restricted to (pre-designated) point against point hypotheses.

It is often supposed that the problem is that SP and WCP conflict, but that is not so. The conflict comes from WCP together with Birnbaumization—understood as both invoking the hypothetical mixture, and effectively erasing the information as to which experiment the data came. To paraphrase Cox's (1958, 361) objection to unconditional tests:

> Birnbaumization says that we can assign **y*** a different level of significance than we ordinarily do, because one may identify an SLP pair **x*** and construct statistic T_B. But this fact seems irrelevant to the interpretation of an observation which we know came from E_2.

6. 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 Evan, Fraser and Monette (1986). Allowing that any **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 denying SLP at any cost (please see the references[2]). We can achieve what they wish to show, without altering principles, and from WCP alone. Given **y***, WCP blocks Birnbaumization; given **y*** has been Birnbaumized, the WCP precludes conditioning.

We agree with Evans, Fraser, and Monette (1986, 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 too 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 $\text{Infr}_{E_{mix}}[\mathbf{x}^*] = \text{Infr}_{E_i}[\mathbf{y}^*]$. 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 $\text{Infr}_{E_B}[\mathbf{y}^*]$, the inference from the associated convex combination. Whether reference to hypothetical mixture E_B is relevant for inference from given **y*** is a distinct question. For an alternative criticism see Evans (2013).

7. 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, **y***) is given, Birnbaum asks us to consider that **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, **y***), once given. That is: $\text{Infr}_{E_i}[(E_{mix}, \mathbf{y}^*)] = \text{Infr}_{E_i}[\mathbf{y}^*]$. It follows that if $\text{Infr}_{E_1}[\mathbf{x}^*] \neq \text{Infr}_{E_2}[\mathbf{y}^*]$, the two remain unequal after the recognition that **y*** could have come from the mixture. What was an SLP violation, remains one.

Given **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 **y***, on pain of logical contradiction. If one does choose to Birnbaumize, and to construct T_B, admittedly, the known outcome **y*** yields the same value of T_B as would **x***. Using the sample space of E_B yields: [B]: $\text{Infr}_{\mathbf{E_B}}[\mathbf{x}^*] = \text{Infr}_{\mathbf{E_B}}[\mathbf{y}^*]$. This is based on the convex combination of the two experiments, and differs from both $\text{Infr}_{E_1}[\mathbf{x}^*]$ and $\text{Infr}_{E_2}[\mathbf{y}^*]$. 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, **x***) or (E_2, **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, **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 a different result. One may consistently apply both the E_B and the WCP directives to the same result, in the same model, only where WCP makes no difference. To claim for any **x* y***, the WCP never makes a difference would entail that there can be no SLP violations, which would make the argument circular.[3] Another possibility is to hold, as Birnbaum ultimately did, that the SLP is "clearly plausible" (Birnbaum 1968, 301) only in "the severely restricted case of a parameter space of just two points" where these are predesignated (Birnbaum 1969, 128). But that is to relinquish the general result.

Acknowledgements

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, at least since 2003. I appreciate the careful queries, and detailed suggested improvements on earlier drafts from anonymous referees. My understanding of Birnbaum was greatly facilitated by the work of philosopher, Ronald Giere, who worked with Birnbaum. I owe him gratitude, as well, for the gift of some of Birnbaum's original materials and notes.

References

Barndorff-Nielsen, O. (1975). Comments on Paper by J. D. Kalbfleisch. *Biometrika, 62* (2), 261–262.

Berger, J. O. (1986). Discussion on a Paper by Evans et al. [On Principles and Arguments to Likelihood]. *Can. J. Stat., 14* (3), 195–196.

———. (2006). The Case for Objective Bayesian Analysis. *Bayesian Analysis, 1* (3), 385–402.

Berger, J. O. & Wolpert, R. L. (1988). *The Likelihood Principle* (2nd ed.). *Vol. 6. Lecture Notes-Monograph Series*. Hayward, California: Institute of Mathematical Statistics.

Birnbaum, A. (1962). On the Foundations of Statistical Inference. In S. Kotz & N. Johnson (Eds), *Breakthroughs in Statistics*, *Vol. 1*, 478–518. New York: Springer-Verlag.

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

———. (1969). Concepts of Statistical Evidence. In S. Morgenbesser, P. Suppes & M. G. White, (Eds.), *Philosophy, Science, and Method: Essays in Honor of Ernest Nagel*, 112–143. New York: St. Martin's Press.

———. (1970a). Statistical Methods in Scientific Inference. *Nature, 225* (5237), 1033.

———. (1970b). On Durbin's Modified Principle of Conditionality. *JASA, 65*, 402-403.

———. (1972). More on Concepts of Statistical Evidence. *JASA, 67* (340), 858–861.

———. (1975). Comments on Paper by J. D. Kalbfleisch. *Biometrika, 62* (2), 262–264.

Casella, G. & Berger, R. L. (2002). *Statistical Inference* (2nd ed.). Belmont, CA: Duxbury Press.

Cox, D. R. (1958). Some Problems Connected with Statistical Inference. *Ann. Math. Stat., 29* (2), 357–372.

———. (1977). The Role of Significance Tests (with Discussion). *Scand. J. Stat., 4* (2), 49–70.

———. (1978). Foundations of Statistical Inference: The Case for Eclecticism. *Aust. J. Stat., 20* (1), 43–59.

Cox, D. R. & Hinkley, D. V. (1974). *Theoretical Statistics*. London: Chapman and Hall.

Cox, D. R. & Mayo D. G. (2010). Objectivity and Conditionality in Frequentist Inference. In D. G. Mayo & A. Spanos, (Eds.), *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science*, 276–304. Cambridge: Cambridge University Press.

Dawid, A. P. (1986). Discussion on a Paper by Evans et al. [On Principles and Arguments to Likelihood]. *Can. J. Stat., 14* (3), 196–197.

Durbin, J. (1970). On Birnbaum's Theorem on the Relation Between Sufficiency, Conditionality and Likelihood. *JASA, 65* (329), 395–398.

Evans, M. J., Fraser, D. A. S. & Monette, G. (1986). On Principles and Arguments to Likelihood. *Can. J. Stat.*, *14* (3), 181–194.

Evans, M. (2013). What Does the Proof of Birnbaum's Theorem Prove. Unpublished manuscript.

Ghosh, J. K., Delampady, M. & Samanta, T. (2006). *An Introduction to Bayesian Analysis*. New York: Springer.

Kalbfleisch, J. D. (1975). Sufficiency and Conditionality. *Biometrika, 62* (2), 251–259.

Lehmann, E. L. & Romano, J. P. (2005). *Testing Statistical Hypotheses* (3rd ed.). New York: Springer.

Mayo, D. G. (1996). *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press.

———. (2010). An Error in the Argument From Conditionality and Sufficiency to the Likelihood Principle. In D. G. Mayo & A. Spanos (Eds.), *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science*, 305–314. Cambridge: Cambridge University Press.

Mayo, D. G. & Cox, D. R. (2010). Frequentist Statistics as a Theory of Inductive Inference. In D. G. Mayo & A. Spanos (Eds.), *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science*, 247–274. Cambridge: Cambridge University Press. First published (2006). In J. Rojo (Ed.), *The second Erich L. Lehmann symposium: Optimality*, *Lecture Notes-Monograph Series, 49*, 77–97. Beachwood, Ohio: Institute of Mathematical Statistics.

———. (2011). Statistical Scientist Meets a Philosopher of Science: A Conversation. *RMM, 2,* (Special Topic: Statistical Science and Philosophy of Science: Where do (should) they meet in 2011 and beyond?), D.G. Mayo, A. Spanos, & K. W. Staley (Guest Eds.), 103–114.

Mayo, D. G. & Spanos, A. (2006). Severe Testing as a Basic Concept in a Neyman–Pearson Philosophy of Induction. *Brit. J. Phil. Sci., 57* (2), 323–357.

Reid, N. (1992). Introduction to Fraser (1966) Structural Probability and a Generalization. In S. Kotz & N. L. Johnson, (Eds.), *Breakthroughs in Statistics*, 579–586. New York: Springer.

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

_____. (1962b). Discussion on a Paper by A. Birnbaum [On the Foundations of Statistical Inference]. *JASA, 57* (298), 307–308.

———. (1970). Comments on a Weakened Principle of Conditionality. *JASA, 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. *JASA, 57* (298), 307–326.

---

[1] 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 appears to entail SLP. See Evans, Fraser and Monet 1986.

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

[3] 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 **x***, **y*** can give rise to an SLP violation. That would render it circular.