On the Birnbaum Argument for the Strong Likelihood Principle

Deborah Mayo[1]
Department of Philosophy
Virginia Tech
mayod@vt.edu

The goal of this article is to provide a new clarification and critique of Birnbaum's (1962) well-known argument purporting to show that principles of sufficiency and (weak) conditionality entail the (strong) likelihood principle. The argument has long been held as a significant result for the foundations of statistics. Not only do all of the familiar frequentist sampling theory notions violate the strong likelihood principle (SLP), the Birnbaum argument purports to show that the SLP follows from principles that sampling theorists accept. We argue that Birnbaum's premises contain an equivocation of the vital terms leading to the unsoundness of the argument. Elucidating and extending existing critiques, we attempt to unravel the logical flaws that have enabled Birnbaum's paradox to persist for over fifty years.

Key terms: Birnbaumization, likelihood principle (weak and strong), mixture experiments, sufficiency, validity, weak conditionality.

[1] I am extremely grateful to Aris Spanos and David Cox for numerous useful corrections and suggestions on earlier drafts of this paper, and for discussions over many years on this and closely related issues. My understanding of Birnbaum was also greatly facilitated by the work of, and discussions with, the philosopher of science, Ronald Giere. Thanks to Giere, who worked with Birnbaum for many years, I also possess Birnbaum's original materials and notes.

On the Birnbaum Argument for the Strong Likelihood Principle

Deborah Mayo

> Without any intent to speak with exaggeration . . . it seems to me that this is really a historic occasion. This paper is a landmark in statistics. . . .
>
> I, myself, like other Bayesian statisticians, have been convinced of the truth of the likelihood principle for a long time. Its consequences for statistics are very great. . . .
>
> [T]his paper is really momentous in the history of statistics. It would be hard to point to even a handful of comparable events. (Savage 1962b, 307-308)

1 Introduction

It is easy to see why Birnbaum's argument for the *strong likelihood principle* (SLP) has long been held as a significant 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:[1]

> 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 according to the SLP, once the data **x** are known, outcomes other than **x** are irrelevant to the evidential import of **x**. While the argument has been stated in various forms, the surprising upshot of all versions is that the SLP appears to follow from applying principles that are uncontroversial for sampling theory statistics, such as the sufficiency principle (SP) and a very weak principle of conditioning, (WCP) (sections 3 and 4). As Casella and Berger remark:

> This violation of the Formal Likelihood Principle may seem strange because, by Birnbaum's theorem, we are then violating either the Sufficiency or the Conditionality Principle. (Casella and R. Berger 2002, 295)

In this paper, we will argue that the Birnbaum argument is unsound. We follow the formulations of the Birnbaum argument given in Berger and Wolpert (1988), Birnbaum (1962), Casella and R. Berger (2002), and D. R. Cox (1977). We modify and extend a brief earlier discussion in Mayo (2010), Mayo and Cox (2011), as well as link our treatment to the main criticisms of the Birnbaum argument that have been raised by others, notably, Cox (1977), Cox and Hinkley (1974), Durbin (1970), Evans, Fraser and Monette (1986) and Kalbfleisch (1975). While our criticism is in sync with portions of existing objections (discussed in section 6) we avoid certain weaknesses that have allowed Birnbaum's argument to stand. In particular, we do not propose to revise the principles that are taken to entail the SLP, but rather to show that Birnbaum's argument, in any of its forms, necessarily applies these principles in a self-contradictory manner. It requires, in effect, that the evidential import of a known result **x**' from experiment E' should, and also should not, be influenced by an unperformed experiment E". Our treatment also differs in allowing the relevant principles of evidence to be equivalence relations. A main goal will be to illuminate the core equivocations that have enabled puzzlement surrounding Birnbaum's argument to persist for over fifty years.

As it is often said, "the proof is surprisingly simple" (Bjornstad 1992, 467), apparently following with two equivalences as premises. Our analysis will also be simple: the series of equivalences in the two premises contain an equivocation of the vital terms leading to the unsoundness of the argument. Indeed, it is the simplicity of Birnbaum's argument that masks the subtle equivocations, the unraveling of which are not only necessary to grasp the issue, but are of deep interest in their own right. Our exposition approaches the argument in pieces, and from several different perspectives. By the end, whichever perspective the reader deems most illuminating, it is hoped that the problem will be glaringly clear.

The import of data in drawing parametric statistical inferences: Ev and Infr.

The Birnbaum argument combines uncontroversial mathematical principles with a presupposition that Birnbaum concedes 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, 491). As such, the argument as a whole rests on intuitive judgments of the nature of evidence for parametric inference, and is understandably of significant foundational interest.

Birnbaum's argument makes use of the notion Ev(E, **x**), "the evidence about the parameter arising from experiment E and result **x**," where this is permitted to range over any inference, conclusion, or report, and is intended to encompass any school of inference. For Birnbaum, E is an experiment involving the observation of **x** with a given distribution f(**x**), and Ev(E, **x**) is to capture the import of the evidence or resulting inference. Birnbaum's analysis has sometimes been discounted because of the vagueness of the idea of a single measure of evidence. But the criticism to be raised here does not turn on this. Birnbaum explicitly intended the idea to cover any form of parametric inference: Bayesian, likelihood, and frequentist assessments, where the parametric model is given. That said, we prefer to use a more perspicuous notation, first developed in Cox and Mayo (2010):

> $\mathrm{Infr}_{\mathrm{E}}(\mathbf{x})$: an inference drawn from data **x** from experiment E, in an associated methodology.

As with Ev, the parametric inference may take any form. In evoking the inferential output, Infr emphasizes the need to look to E for the associated inferential context, avoiding the presumption of a single measure of "the" evidence.

$\mathrm{Infr}_{\mathrm{E}}(\mathbf{x})$, like Ev, is generally used as an abbreviation that requires a completion to be a well-formed proposition (it is a *propositional function*). It may abbreviate

> An inference about θ from E with result **x** depends on__.

Another construal, following Berger (1985, 35), is to construe it as

> The "information" about θ that is obtained (or should be reported) upon observing **x** is __.

A central use of these notions is to discuss the equivalence or inequivalence of the evidential import or inference from different experiments. Birnbaum calls these *principles of evidence.* In using $\mathrm{Infr}_{\mathrm{E}}(\mathbf{x})$ to state general principles of evidence, the completion of $\mathrm{Infr}_{\mathrm{E}}(\mathbf{x})$ usually indicates which conclusions drawn from **x** from E ought to be considered equivalent to others. We denote by,

> $\mathrm{Infr}_{\mathrm{E'}}(\mathbf{x'})$ equiv $\mathrm{Infr}_{\mathrm{E''}}(\mathbf{x''})$: an inference from experiment E' with data **x**' is or ought to be the same as an inference from E" with outcome **x**".

Alternatively, in some cases the use of "=" is more natural. For example, to write that $\mathrm{Infr}_E(\mathbf{x})$ depends only on, or is entirely given by, some quantitative output, such as a significance level p, we write $\mathrm{Infr}_E(\mathbf{x}) = p$. Birnbaum (1972) observes:

> That Ev is used specifically only to establish relations Ev(E', **x**') = Ev(E", **x**"), which could alternatively always be written as (E', **x**')~(E", **x**") where ~ is an equivalence relation to be interpreted as 'is evidentially equivalent to', (Birnbaum 1972, 858[2]).

In accord with Birnbaum and the general literature on Ev, it is understood that "by listing E we allow Ev to depend on full knowledge of all aspects of the experiment and not on just the observed **x**" (Berger 1985, 35). In addition, we understand the model of experiment E to include a reference to the inference methodology to be applied. In a Bayesian formulation, $\mathrm{Infr}_E(\mathbf{x})$ might refer to the bearing of **x** on degree of belief in the hypotheses in question (Edwards, Lindman, and Savage 1963, 201); for a sampling theorist, $\mathrm{Infr}_E(\mathbf{x})$ might refer to the import of the data for an assessment of error probabilities (e.g., confidence levels, p-values). Our focus is on the implications for sampling theory, so the sampling distribution intended for interpreting evidence would be part of E.

While there would be an advantage to rigor in making explicit each feature of context, sample and parameter spaces, test statistic, methodology, and so on, the argument is more perspicaciously rendered by retaining this simple notation, with necessary qualifications to be added. It also enables a ready correspondence to Birnbaum's "Ev". The reader is free to substitute Ev throughout, as this will not alter our argument.

The SLP has long been controversial. The controversy grows out of fundamental philosophical differences about the nature of statistical inference. Our discussion does not depend on accepting any statistical philosophy, nor even on endorsing the principles evoked, at least those that are not purely mathematical[3]. It has limited aims: to deny that SLP violations entail a violation either of the sufficiency (SP) or weak conditionality (WCP) principles. If we are correct, this refutes a position that is generally presented as settled:

> These two principles, namely the conditionality principle (WCP) and sufficiency principle (SP) together have a far reaching implication. Birnbaum (1962) proved that they imply one must then follow the [strong] likelihood principle (SLP), which requires that inference be based on the likelihood alone, ignoring the sample space. (Ghosh, Delampady, and Samanta 2006, 38[4])

Since the crux of the present problem is essentially logical, it is perhaps not inappropriate for a philosopher to tackle it. While extensions beyond sampling theory fall outside of the current restricted goal, our formulation points to a very general flaw. Contemporary nonsubjective or default Bayesian methods are also thought to violate the SLP.[5] A more streamlined statistical formulation, while beyond the scope of this paper would, we suspect, enable this critique to be extended to justify SLP violations within a Bayesian formulation.

Sections 2, 3, and 4 discuss the (strong) likelihood principle (SLP), and the principles of sufficiency (SP) and weak conditionality (WCP), respectively. Section 5 analyzes two variations of Birnbaum’s argument, and section 6 reviews the central criticisms by others, and how the current treatment both illuminates and strengthens them. Section 7 gives a brief concluding overview.

2 The (Strong) Likelihood Principle (SLP) and Its Violations

The strong likelihood principle is a universal “if then” claim:

> *SLP*: For any two experiments E’ and E” with different probability models f’, f” but with the same unknown parameter θ, if the likelihood of outcomes **x’*** and **x”*** (from E’ and E” respectively) are proportional to each other, then **x**’* and **x”*** should have the identical evidential import for any inference concerning parameter θ**.**

SLP pairs. When the antecedent of the SLP holds, **x’*** and **x”*** are said to have “the same likelihood function,” i.e., f’(**x**; θ) = **c**f”(**x”**, θ) for all θ, **c** a positive constant. In such cases, we abbreviate by saying **x’*** and **x”*** are *SLP pairs*, and the asterisk * will be used to indicate this.

So we can abbreviate the SLP as follows:

SLP: for any two experiments, E’ and E”, if **x’*** and **x”*** are *SLP pairs* (from E’ and E” respectively) then Infr$_{E'}$(**x’***) equiv Infr$_{E''}$(**x**”*).

Notation: The hatch marks ‘ and “ indicate from which of the two experiments **x** arose. **x’** tells us E’ was performed and **x’** observed, but it will do no harm to equivalently write:

(E’, **x’**): E’ was run and **x’** observed,

since such notation is common in Birnbaum’s argument. To indicate the pair of experiments giving rise to SLP pairs, we can index the experiment using E^j where j = 1 or j = 2, symbolizing with single and double hatch marks:

(E^j ,$\mathbf{x}^j$): E^j was run and $\mathbf{X} = \mathbf{x}^j$ observed.

Principles of evidence *prescribe* an equivalence, and in so doing they *proscribe* whatever violates the equivalence. It will often be useful to bring out explicitly what is being proscribed by a given principle of evidence. Obviously, the SLP proscribes SLP violations. What are these?

SLP violations. We may characterize an SLP violation, in a given methodology, as any inferential context (in that methodology) where the antecedent of the SLP is true and the consequent is false (examples to follow). The SLP pairs that are of central interest are those leading to SLP violations. Inference within the subjective Bayesian methodology is said to uphold the SLP (so long as its assumptions hold):

> The only contribution of the data is through the likelihood function . . . In particular, if we have two pieces of data **x**’ and **x**” with the same likelihood function . . . the inferences about θ from the two data sets should be the same. This is not usually true in the orthodox theory, and its falsity in that theory is an example of its incoherence. (Lindley 1976, 361; **x**’ and **x**” replace x^1 and x^2; θ replaces his q)

2.1 SLP violation with binomial, negative binomial

Example 1. *Binomial versus negative binomial*. Consider independent Bernoulli trials, with the probability of success at each trial an unknown constant θ, but produced by different procedures, E’, E” respectively. Experiment E’ has a pre-assigned number n of

Bernoulli trials, say, 20, and R is the number of successes observed. In E”, trials continue until a pre-assigned number r, say, 6, of successes have occurred, with the number N trials recorded. The sampling distribution of R is *binomial*:

$f(R; \theta) = ({}_{n}C_{r})\, \theta^{r}(1-\theta)^{n-r}$

while the sampling distribution of N is negative *binomial*

$f(N; \theta) = ({}_{n-1}C_{r-1})\, \theta^{r}(1-\theta)^{n-r}$.

If two outcomes from E’ and E”, respectively, have the same number of successes and failures, then they have the “same” likelihood, in the sense that they are proportional to $\theta^{r}(1-\theta)^{n-r}$. The two outcomes, **x’*** and **x”***, are SLP pairs. But the difference in the sampling distributions of the statistics R and N of experiments E’ and E”, respectively, entails a difference in p-values or other error probability assessments. Accordingly, their evidential appraisals differ according ot sampling theory. Thus, **x**’* and **x**”* are SLP pairs leading to an SLP violation.

An SLP violation with binomial (E’) and negative binomial (E”):

(E’, r=6) and (E”, n=20) have proportional likelihoods,

but $\mathrm{Infr}_{E'}$(**x**’*= 6) is *not* equiv to $\mathrm{Infr}_{E''}$(**x**”*=20).

Although the difference is generally slight in this case, its simplicity makes it a good example to fix these ideas.

Loss of relevant information if the index is erased. A fundamental role a principle of inference must play in sampling theory is to indicate the relevant sampling distribution for inference. In making inferences about θ on the basis of data **x** in sampling theory**,** relevant information would be lost if the report removed the index from E and stated only:

> Data **x** consisted of r successes in n Bernoulli trials, generated from *either* a binomial experiment with n fixed at 20, or a negative binomial experiment with r fixed at 6—erasing the index indicating the actual source of data.

2.2 SLP violation with fixed normal testing and optional stopping: E’, E”

Example 2. *Fixed versus sequential sampling*. Suppose **X**’ and **X**” are samples from distinct experiments E’ and E”, both distributed as $N(\mu,\sigma^{2})$, with σ known, and p-values are to be calculated for the hypotheses:

$$H_0: \mu = 0 \text{ vs. } H_1: \mu > 0.$$

In E’ the sample size is fixed, whereas in E” the sampling rule is to continue sampling until:

$$\bar{X} > c = 1.96\sigma/(n^{.5}).$$

Suppose that E” is first able to stop with n = 169 trials. Then the likelihood of **x**” is proportional to the likelihood of a result that could have occurred from E’, where n was fixed in advance to be 169, and **x**’ is $1.96\sigma/(n^{.5})$. Although the corresponding p-values would be different, the two results would be inferentially equivalent according to the SLP.[6]

SLP violation with fixed normal testing and optional stopping: E’, E”:

> (E’, $1.96\sigma/169^{.5}$) and (E”, n = 169) have proportional likelihoods, yet
> $\text{Infr}_{E’}(1.96\sigma/169^{.5})$ is *not* equiv to $\text{Infr}_{E”}(n = 169)$.

By contrast, the SLP equates the evidential import of the two. This is a striking difference. For some sampling theorists, this example alone “taken in the context of examining consistency with $\theta = 0$, is enough to refute the strong likelihood principle” (Cox 1977, p. 54), since, with probability 1, it will stop with a “nominally” significant result even though $\theta = 0$. It contradicts what he calls the *strong repeated sampling principle*.

2.3 Inapplicability versus violations of principles: 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). It would seem, therefore, that accounts that adhere to the SLP are not thereby prevented from considering outcomes other than the one observed, prior to conducting the experiment. Nor are they prevented from analyzing features of the data such as residuals, which are relevant to the question of checking the statistical model itself. There is some ambiguity on this point in G.Casella and R. Berger (2002):

> Most data analysts perform some sort of “model checking” when analyzing a set of data. 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 R. Berger 2002, 295-6)

They rightly warn that before considering the SLP and WCP, “we must be comfortable with the model” (296). It seems to us more accurate to regard the principles as inapplicable, rather than violated, when the model is not taken as an adequate representation of the given parametric inference.[7]

After all, another way to state the SLP is that, given the adequacy of underlying models for experiments E’ and E”, all of the information they supply for informative inference is contained in their respective likelihoods. Birnbaum restricts the informative inference to parametric inference within a model, and we do as well. Therefore, we stipulate that the statistical models underlying any SLP pairs that arise for examination are accepted as adequately representing the experimental conditions, and, in particular, that they accord with the sample space and sampling distributions referred to in f’(**x**’; θ) and f”(**x**”; θ), respectively. Likewise E’, E” and their corresponding distributions should retain their meanings throughout any argument about their evidential import.

3 Sufficiency Principles

One of the principles to which the Birnbaum result refers is the sufficiency principle (SP).[8]

Sufficient statistic: Let data **x**= $(x_1,x_2,\ldots,x_n)$ be a realization of random sample **X**, following a distribution f(.); a *statistic* T(**X**) is a *sufficient* statistic if the following relation holds:

$$f(\mathbf{x};\theta) = f_T(t;\theta)\, f_{x|T}(\mathbf{x}|\, t),$$

where $f_{x|T}(\mathbf{x}|\, t)$ does not depend on the unknown parameter θ.

Sufficiency Principle (general): If random sample **X**, in experiment E, has probability density f(**x**; θ), the assumptions of the model are valid, and T is minimal sufficient for θ, then if T(**x**’) = T(**x**”), then $\mathrm{Infr}_{E'}(\mathbf{x}') = \mathrm{Infr}_{E''}(\mathbf{x}'')$.

Since the sufficiency principle holds for different inference schools, any application must take into account the inference method under discussion (Cox and Mayo 201, 286-7).

> *Sufficiency Principle (in sampling theory):* If a random sample **X**, in experiment E, arises from f(**x**;θ), and the assumptions of the model are valid, then all the information about θ contained in the data may be obtained from considering its minimal sufficient statistic T and its *sampling distribution* $f_T(\mathbf{t};\theta)$ (in experiment E).

Note that the general sufficiency principle still holds for sampling theory; parametric inference in that context always employs the associated sampling distribution for statistic T within E. We suggest that "some confusion over the role of sufficiency" (Cox and Mayo 2010, 289) is intermingled in the Birnbaum argument. It will be useful to rehearse some elementary points.

Binomial and negative binomial experiments E' and E". Consider our comparison of binomial and negative binomial trials in Example 1. Suppose we have a binomial experiment E' with the number of trials n = 20. The sufficient statistic is $T_{E'}(\mathbf{x}) = R$, the number of successes in 20 trials. Let $\mathbf{x'}_a$ and $\mathbf{x'}_b$ be two outcomes from binomial experiment E' with exactly 6 successes but in a different order. We have $T_{E'}(\mathbf{x'}_a) = T_{E'}(\mathbf{x'}_b)$, so from the sufficiency principle (SP):

$$\mathrm{Infr}_{E'}(\mathbf{x'}_a) = \mathrm{Infr}_{E'}(\mathbf{x'}_b).$$

The sufficiency of R follows from the fact that

$$f(\mathbf{x}|R = r; \theta) = f(\mathbf{x}; \theta)/f(R; \theta) = 1/b,$$

which is a known uniform distribution free of θ with range b = $({}_nC_r)$.

Compare this with the negative binomial experiment E". Within the negative binomial experiment E", the sufficient statistic is N, the number of trials required before the rth success. The sufficiency of N follows from the fact that

$$f(\mathbf{x}|N = n; \theta) = f(\mathbf{x};\theta)/f(N{=}n; \theta) = 1/m,$$

which is a known uniform distribution free of θ with range m = $({}_{n-1}C_{r-1})$.

Let **x"** be an outcome of E" where r=6 was fixed, and suppose the 6th success occurred on the 20th trial. The SP does not entail that $\mathrm{Infr}_{E''}(\mathbf{x''})$ equiv $\mathrm{Infr}_{E'}(\mathbf{x'}_a)$, even if the two strings have their 6 successes in the same order. In the case of the negative binomial, the ordering of the sample determines whether there will be a next trial, whereas in the binomial case, the ordering is irrelevant. E' and E" refer to two different experiments, while SP refers to *one* experiment, with its single sampling distribution, whatever it may be.

That is why the SP is sometimes called the *weak* LP, which asserts that within a single experiment E, the likelihood contains all the relevant information for parametric inference, given E is an adequate model. The interesting and controversial claim is the strong likelihood principle (SLP).

The SLP has at times been defined in a way that is correct only for the weak version:

> Ev(E,**x**) should depend on E and **x** only through the likelihood function f(x|θ) for the observed **x**. (Berger 1985, 35)

This statement overlooks the fact that the SLP refers to a pair of experiments, not a single experiment. Moreover, in sampling theory, even within a single experiment E, $\mathrm{Infr}_E(\mathbf{x})$ depends on more than the likelihood. Otherwise it would suffice to report the series of trials erasing whether it came from the binomial or negative binomial. We have seen that this does not suffice.

But this may suggest that if it were feasible to take an arbitrary pair of experiments E', E" giving rise to potential SLP pairs, and to turn the two into a single experiment (e.g., by a mixture), that perhaps a pair of results could become evidentially equivalent using SP. This will be part of Birnbaum's gambit (section 5).

4 Weak Conditionality Principle

The second principle of evidence on which Birnbaum's argument rests is the *weak conditionality principle* (WCP). This principle, Birnbaum notes, follows not from mathematics alone but from intuitively plausible views of "evidential meaning." To understand the interpretation of the WCP that gives it its plausible ring, we consider its development in "what 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).

4.1 The basis for the WCP

Example 3. *Two measuring instruments of different precisions.* We flip a fair coin to decide which of two instruments, E' or E", to use in observing a normally distributed random sample **X** to make inferences about mean θ. *E'* has a known variance of 10^{-4}, while that of E" is known to be 10^4. The experiment is a mixture: E-mix. The fair coin or other randomizer may be characterized as observing an indicator statistic J, taking values 1 or 2 with probabilities .5, independent of the process under investigation. The full data indicates first the result of the coin toss, and then the measurement: (E^j, $\mathbf{x}^j$).[9]

The sample space of E-mix with components E^j, j = 1, 2, consists of the union of {(j, **x'**): j = 0, possible values of **X'**} and {(j, **x**"): j = 1, possible values of **X**"}.

In testing a null hypothesis such as θ = 0, the same **x** measurement would correspond to a much smaller p-value were it to have come from E′ than if it had come from E": denote them as p′(**x**) and p″(**x**), respectively. However, the overall significance level of the mixture, the convex combination of the p-value: [p′(**x**) + p″(**x**)]/2, would give a misleading report of the precision or severity of the actual experimental measurement (See Cox and Mayo 2010, 296).

Suppose that we know we have observed a measurement from E" 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).

In effect, an individual unlucky enough to use the imprecise tool gains a more informative assessment because he might have been lucky enough to use the more precise tool! (Birnbaum 1962, 491; Cox and Mayo 2010, 296). Once it is known whether E′ or E″ has produced **x**, the p-value or other inferential assessment should be made conditional on the experiment actually run.

> *Weak Conditionality Principle (WCP):* If a mixture experiment is performed, with components E', E" determined by a randomizer (independent of the parameter of interest), then once (E', **x'**) is known, inference should be based on E' and its sampling distribution, not on the sampling distribution of the convex combination of E' and E".

4.2 Understanding the WCP

The WCP includes a prescription and a proscription for the proper evidential interpretation of **x'**, once it is known to have come from E':

> The evidential meaning of any outcome (E', **x'**) of any experiment E having a mixture structure is the same as: the evidential meaning of the corresponding outcome **x'** of the corresponding component experiment E', *ignoring otherwise the over-all structure of the original experiment* E (Birnbaum 1962, 489 E_h and x_h replaced with E' and x' for consistency).

While the WCP seems obvious enough, it is actually rife with equivocal potential. To avoid this, we spell out its three assertions.

First, it applies once we know which component of the mixture has been observed, and what the outcome was (E^j $\mathbf{x}^j$). (Birnbaum considers mixtures with just two components).

Second, there is the prescription about evidential equivalence. Once it is known that E^j has generated the data, given that our inference is about a parameter of E^j, inferences are appropriately drawn in terms of the distribution in E^j—the experiment known to have been performed.

Third, there is the proscription. In the case of informative inferences about the parameter of E^j our inference should not be influenced by whether the decision to perform E^j was determined by a coin flip or fixed all along. Misleading informative inferences might result from averaging over the convex combination of E^j and an experiment known not to have given rise to the data. The latter may be called the unconditional (sampling) distribution.

Behavioristic versus informative (or evidential) construal. The same issue arises with mixtures of experiments for N-P tests with predesignated error probabilities. Lehmann and Romano (2010) ask:

> Should the procedure be chosen which is best on the average over both experiments, or should the best conditional procedure be preferred; and, 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? . . . The underlying question is of course the same: Is a conditional or unconditional point of view more appropriate? (p. 394, chapter 10)

They suggest that "the answer cannot be found within the model but depends on the context" (ibid). For example, if the overall experiment is to be performed many times, the average performance may be of principal interest, and an unconditional approach suitable. This might be called a *behavioristic* context.

That a sampling theorist might interpret some contexts behavioristically, calling for the unconditional sampling distribution, is consistent with her applying the WCP in what Birnbaum calls contexts of "informative inference":

> For purposes of informative inference, if [**X** = **x'**] is observed with the first instrument, then the report [(E', **x**')] seems to be an appropriate and complete description of the statistical evidence obtained; and the 'more complete' report [that E' might not have been selected] . . . seems to differ from it only by the addition of recognizably redundant elements irrelevant to the evidential meaning and evidential interpretation of this outcome of [E']. (Birnbaum 1962, 491)

The redundant elements, Birnbaum continues, are the other experiments which might have been carried out, but in fact were not (ibid.).

Now, adding redundant elements to the report (E', **x'**) would essentially have no impact on the inference, except perhaps in being less simple. Thus, to fully state the WCP we must always insert its implicit third (proscriptive) feature; namely, we ought not to allow such "irrelevant" elements to alter the evidential construal of (E', **x'**) by averaging over the components, as in an unconditional mixture analysis.

4.3 Avoiding ambiguity: behavioristic and informative contexts

A number of ambiguities can now be clarified.

The "conditional or unconditional" question in the context of a mixture experiment $\text{Infr}_{\text{E-mix}}$ asks: once the experiment is chosen and the result is known, how should we determine $\text{Infr}_{\text{E-mix}}$? Should we use the unconditional or the conditional sampling distribution?[10] Referring to Example 3, the measuring machines with different precisions:

(i) Under an unconditional significance level assessment,

$\text{Infr}_{\text{E-mix}}(\text{E'}, \mathbf{x'}) = .5\text{p'}+.5\text{p''}$.

(ii) Under a conditional significance level assessment, $\text{Infr}_{\text{E-mix}}(\text{E'}, \mathbf{x'}) = \text{p'}$.

By contrast, $\text{Infr}_{\text{E'}}(\mathbf{x}\text{'}) = \text{p'}$ and $\text{Infr}_{\text{E''}}(\mathbf{x}\text{''}) = \text{p''}$, respectively.

Unless p" = 0, these two are not equivalent. So without the qualification "under an unconditional (or conditional) assessment," (in (i) and (ii)) we would inconsistently have:

$\text{Infr}_{\text{E-mix}}(\text{E'},\mathbf{x}\text{'}) \neq \text{Infr}_{\text{E-mix}}(\text{E'},\mathbf{x}\text{'})$.

Note: It seems preferable to use "=" here because it is understood that a particular quantity will be reported, but nothing turns on this.[11]

A second ambiguity. G. Casella and R. Berger (2002) write:

> The [weak] Conditionality principle simply says that if one of two experiments is randomly chosen and the chosen experiment is done, yielding data **x**, the information about θ depends only on the experiment performed. . . . *The fact that this experiment was performed, rather than some other, has not increased, decreased, or changed knowledge of θ.* (p. 293, emphasis added)

I have emphasized the last line in order to underscore a possible equivocation. Casella and Berger's intended meaning is the correct claim:

(i) Given that it is known that measurement **x**' is observed as a result of using tool E', then it does not matter (and it need not be reported) whether or not E' was chosen by a random toss (that might have resulted in using tool E") or had been fixed all along.

Of course we do not know what measurement would have resulted had the unperformed measuring tool been used.

Compare (i) to a false and unintended reading:

(ii) If some measurement **x** is observed, then it does not matter (and it need not be reported) whether it came from a precise tool E' or imprecise tool E".

The idea of detaching **x**, and reporting that "**x** came from somewhere I know not where," will not do. For one thing, we need to know the experiment in order to compute the sampling inference. For another, E' and E" may be like our weighing procedures with very different precisions. It is analogous to being given the likelihood of the result in Example 1, withholding whether it came from a negative binomial or a binomial.

Claim (i), by contrast, may well be warranted, not on purely mathematical grounds, but as the most appropriate way to report the precision of the result attained, as when the WCP applies. The essential difference in claim (i) is that it is known that (E, **x**'), enabling its inferential import to be determined.

The linguistic similarity of (i) and (ii) may explain the equivocation that vitiates the Birnbaum argument.

4.4 Is the WCP an equivalence?

A central question is whether the WCP is a proper equivalence relation, holding in both directions (Evans et al.1986; Durbin 1970). Weighing against viewing it as an equivalence is this: it makes no sense to say that one should replace the unconditional with the conditional assessment (once it is known which component of a mixture was performed), and *at the same time* to maintain that the unconditional and conditional assessments are evidentially equivalent. The WCP prescribes conditioning on the experiment known to have produced the data, *and not the other way around*. It is only because these do not yield equivalent appraisals that the WCP may serve to avoid counterintuitive assessments (e.g., those that would otherwise be permitted by those famous weighing machines). It is their inequivalence, in short, that gives Cox's WCP its normative proscriptive force:

> WCP proscription: Once (E', **x**') is known, $\text{Infr}_{E'}$(**x**') should be computed using not the unconditional sampling distribution over E' and E", but, rather, the sampling distribution of E'.

Yet there is an equivalence within the WCP, and so long as it is consistently interpreted, it raises no problems.[12] This turns out to be the linchpin to disentangling the Birnbaum argument.

To hold the WCP for a given context is to judge that the information that E' was determined by a flip is a redundancy, equivalent to conjoining to the known outcome (E', **x**') what is effectively a tautology:

"either E' was chosen from a randomizer, or E' was fixed all along."

We might write this:

Knowing that (E', **x**') occurred,

Infr$_{E'}$(**x**') equiv [Infr$_{E'}$(**x**') and (Either E' was chosen by flipping, or E' was fixed)],

where it is given that the flipping in no way alters the construal of (E', **x**').

Or again, once **x**' is observed from E', an informative inference about θ should not be influenced by any not-done experiments; *it should be equivalent to what it would have been* had E' been fixed at the start, namely, *a function of the sampling distribution of E' alone.*[13] If the randomizer was genuinely irrelevant for the informative inference from (E', **x**'), then one is free to conjoin to a given result (E', **x**'), "and the reason E' was performed is irrelevant."

Birnbaum defines the WCP as applicable "if an experiment E is (mathematically equivalent to) a mixture" (p. 491) of components. He emphasizes that the equivalence is *mathematical* because it is to hold even if no mixture was actually performed. Viewing the WCP as endorsing a genuine "two-way" equivalence requires viewing any known experimental result as equivalent, evidentially, to its being a component of a corresponding mixture, even though it is known that in fact E was not chosen by a mixture. While this may seem unsettling, no untoward evidential interpretations result so long as the proscriptive part of the WCP remains, and is not contradicted (say, by allowing the imaginary mixture to influence the interpretation of the known "component"). We will come back to this in considering Birnbaum's response to Durbin (section 6).

5 Birnbaum's Argument

We can now consider the Birnbaum argument for the SLP defined in section 2. In abbreviated form it states:

> *SLP*: for any two experiments, E’ and E”, if $\mathbf{x}'^*$ and $\mathbf{x}''^*$ are *SLP pairs* (from E’ and E” respectively), then $\mathrm{Infr}_{E'}(\mathbf{x}'^*)$ equiv $\mathrm{Infr}_{E''}(\mathbf{x}''^*)$.

Begin with any case where the antecedent of the SLP holds. The task is to show that $\mathbf{x}'^*$ and $\mathbf{x}''^*$ ought to be deemed evidentially equivalent.

Premise one:

Suppose we have observed (E’, $\mathbf{x}'^*$) with an SLP pair (E”, $\mathbf{x}''^*$). Then we are to view (E’, $\mathbf{x}'^*$) as having resulted from getting heads on the toss of a fair coin, where tails would have meant performing E” (any other irrelevant randomizer would do). This is sometimes called the “enlarged experiment.” Now construct the Birnbaum test statistic T-B defined in terms of the enlarged experiment:

$$\text{T-B}(E^j, \mathbf{x}^{j}*) = (E', \mathbf{x}'^*), \text{ if } \mathbf{x}'= \mathbf{x}'^* \text{ or } j = 2 \text{ and } \mathbf{x}'' = \mathbf{x}''^*.$$

$$\text{Else, report the outcome } (E^j, \mathbf{x}^j).$$

In words: in the case of a member of an SLP pair, statistic T-B has the effect of erasing the index j. Inference based on T-B is to be computed averaging over the performed and unperformed experiments E’ and E”. This is the *unconditional formulation* of the enlarged experiment. This gives premise one:

> (1) For any (E’, $\mathbf{x}'^*$), the result of construing its evidential import in terms of the unconditional formulation is that:
> $\mathrm{Infr}_{E\text{-}B}(\mathbf{x}'^*)$ equiv $\mathrm{Infr}_{E\text{-}B}(\mathbf{x}''^*)$.

The likelihood functions of (E’, $\mathbf{x}'^*$) and (E”, $\mathbf{x}''^*$) are proportional for all θ, being $.5f(\mathbf{x}'^*; \theta)$ and $.5f(\mathbf{x}''^*; \theta)$, respectively. T-B is a sufficient statistic *within* the enlarged experiment, treated unconditionally. However E’ and E” are different models of the experiment producing the two likelihoods, and the enlarged model associated with T-B is yet a third model of the experiment. T-B erases the source of SLP pairs, and inferences from E-B are based on the convex combination of E’ and E”. Premise two now alludes to the WCP:

Premise two:

(2) Once it is known that E*’* produced the outcome **x**’*, the inference should be computed just as if it were known all along that E$^{’}$ was going to be performed, i.e., one should use the conditional formulation, ignoring any mixture structure:

$\text{Infr}_{\text{E-B}}(\mathbf{x}\text{'*})$ equiv $\text{Infr}_{\text{E'}}(\mathbf{x}\text{'*})$.

More generally, once ($\mathbf{x}^{j}$*) is known to have come from E^{j}, j = 1 or 2, premise two is

$\text{Infr}_{\text{E-B}}(\mathbf{x}^{j}*))$ equiv $\text{Infr}_{\text{E'}}(\mathbf{x}^{j}*)$.

From premises one and two it is concluded, for any arbitrary SLP pair **x**’*, **x”***,

$\text{Infr}_{\text{E'}}(\mathbf{x}\text{'*})$ equiv $\text{Infr}_{\text{E''}}(\mathbf{x}\text{''*})$.

The SLP is said to follow. This is an unsound argument.

5.1 Validity and soundness

Deductive validity is a matter of form: to say that an argument is formally deductively valid is to say that if its premises are all true, then the truth of the conclusion would have to follow.[14] However, it is easy to render any argument formally valid (e.g., by adding premises). In order to infer the truth of the conclusion of a valid argument, formal validity does not suffice. What is needed in addition is the truth of the premises.

So let us consider the truth of the two premises of Birnbaum’s argument.

Premise one is true provided that $\text{Infr}_{\text{E-B}}(\mathbf{x}\text{'*})$ is the inference from (E’, **x**’*) averaging over the unconditional sampling distribution of statistic T-B. In other words, in premise one, E-B is modeled as reporting the value of sufficient statistic T-B in the enlarged experiment. In effect it reports just the likelihood of **x***, which enters inference in terms of the convex combination of E’ and E”.

For premise two to be true, $\text{Infr}_{\text{E-B}}(\mathbf{x}^{j}*)$ must refer to the inference from (E^{j}, $\mathbf{x}^{j*}$) modeled in terms of the sampling distribution of E^{j} alone. Whenever **x**’* and **x**”* are SLP violations, the distributions of E’ and of E” both differ from that of T-B. The experiment E-B on which inference is to be based has different meanings in each premise. The argument is invalid.[15] It would be (loosely) akin to arguing:

(1) Defining the best author of 2010 as whoever won the Booker Prize,
the best author of 2010 = my mother.

(2) Defining the best author of 2010 as the one whose book was read by the most book clubs,
the best author of 2010 = Peter Stern.

Therefore, my mother = Peter Stern.

Even rendering (1) and (2) true, the conclusion can be false.

5.2 Second formulation: allowing both “if then” premises to be true

We can formulate the argument so that both premises are true “if then” statements[16] incorporating the stipulated sampling distributions:

As before, suppose an arbitrary member of an SLP pair (E’, E”) is observed, e.g., (E’, $\mathbf{x}'*$) is observed. The question is to its evidential import.

(1) If $\mathrm{Infr}_{E\text{-}B}(\mathbf{x'*})$ is computed unconditionally, averaging over the sampling distributions of T-B, then
$\mathrm{Infr}_{E\text{-}B}(\mathbf{x}'*)$ equiv $\mathrm{Infr}_{E\text{-}B}(\mathbf{x}''*)$.[17]

(2) If $\mathrm{Infr}_{E\text{-}B}(E^j,\mathbf{x}^j*)$ is computed conditionally, using the sampling distribution of E^j, then
$\mathrm{Infr}_{E\text{-}B}(\mathbf{x}^j*)$ equiv $\mathrm{Infr}_{E'}(\mathbf{x}^j*)$ for j= 1, 2.

Construed as “if then” claims, the premises can both be true, but then we cannot validly infer the SLP:

$$\mathrm{Infr}_{E'}(\mathbf{x}^{'}*) \text{ equiv } \mathrm{Infr}_{E''}(\mathbf{x}^{''}*).$$

We would need contradictory antecedents to hold.

The formal invalidity is proved by any SLP violation since, in that case, the premises are true and the conclusion is false. SLP violation pairs are readily available (e.g., Examples 1 and 2), and no contradiction results. In fact, we have demonstrated something stronger: whenever we deal with an SLP violation pair, the two “if then” premises when true yield a false conclusion.

6. Critical Discussion

A number of critical discussions of the SLP exist.[18] Birnbaum himself rejected the SLP, and endorsed what he called the “confidence concept,” which was based on control of error probabilities (see section 7).

The general tactic for criticisms is to find fault with either premise one, the application of the sufficiency principle (SP) (or its equivalent), or else to find fault with premise two, the application of the WCP. Typically, revised versions of these principles are given, and an order of application recommended, blocking the SLP. The current treatment differs in allowing these principles to be defined and applied exactly as Birnbaum’s argument directs. We considered two forms of the argument. Our criticism, in either formulation, is simply that it is impossible consistently to combine the unconditional and conditional directives in the same problem, and yet purport that the conclusion follows. Considering other lines of criticism will enable us to strengthen our arguments further.

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 **x**’* that is a member of an SLP pair (E’, E”) with given, different probability models f’, f”, that he will be able to derive from uncontroversial principles SP and WCP, that **x’*** and **x”*** should have the identical evidential import for an inference concerning parameter θ.

Now the WCP, Birnbaum grants, does not follow from formal mathematics but from intuitive, plausible judgments about the “nature of evidential meaning” when drawing “informative” parametric inferences. To make the application of the WCP as plausible as possible, therefore, Birnbaum refers to the type of clear-cut mixture experiment of the sort that arises in Cox’s (1958) famous example of measuring instruments E’, E” with different precisions. That is why his argument is given in terms of the “weak” conditionality principle, recognizing that broader conditioning principles are not so readily acceptable, especially to sampling theorists. The onus is on Birnbaum to demonstrate that this restricted WCP compels us to accept the SLP. In return, we grant Birnbaum the most generous formulation of his premises.

6.1 Is E-B a mixture experiment?

Readers who have come this far may well wonder why we have allow Birnbaum to take us into a land where a known, arbitrary outcome **x** from E, not generated from a mixture, is treated as if it were generated by a mixture, if it happens to have an SLP pair in an unperformed experiment, and then, faced with this imaginary “enlarged” experiment, we are to “condition” back down to known experiment E. Given that his argument has stood for over fifty years, we wish to give it the maximal run for its money, and not try to block its premises, however questionable its key moves may appear from the point of view of informative inference.

What then to say about enlarging the known experiment (in premise one)? Is the Birnbaum experiment E-B with statistic T-B a mixture experiment? By Birnbaum’s own definition of a mixture, there needs to be two components, the first an application of the randomizer, whose distribution is independent of the parameter about which the inference pertains. (It is not necessary even to know this probability.)

However, one cannot perform the following: Toss a fair coin. If it lands heads, perform an experiment E’ that yields a member of an SLP pair (**x**’*); if tails, observe an experiment that yields the other member of the SLP pair (**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 **x**’*. There is a single experiment, performed on either E’ or E”, and it is stipulated that we know which was performed and what its outcome was.[19] An ordinary mixture has to permit any of the possible outcomes in the respective sample spaces of E’ and E”, and Birnbaum’s experiment includes only the SLP pairs.

As Cox observes in relation to the SLP violation from a binomial and negative binomial (section 2.1, Example 1), “the ‘enlarged’ experiment is peculiar to the particular **x**’* and **x”*** (replacing his y’, z’) and that to apply the conclusion generally we need an ‘an enlarged’ experiment for every pair (r, n)” (Cox 1978, 54).

This is correct, and it is the linchpin of an important objection. From a pre-data perspective, the convex combination would seem to require averaging over all of these pairs of outcomes from this pair of experiments. Some have described the Birnbaum experiment as unperformable, or at most a “mathematical mixture” rather than an “experimental mixture” (Kalbfleisch 1975). Birnbaum himself calls it a “mathematical” and “hypothetical” mixture.

6.2 Birnbaumization

We may give Birnbaum more leeway, so as to see E-B as “performed” as follows: An outcome has been observed that is known to be a member of an SLP pair {E’, E”}; let us label it as (E”, **x**”*) this time. We are then to imagine that (E”, **x**”*) was the result of flipping a fair coin (or some other randomizer). The Birnbaum statistic T-B instructs us to convert the known result x”* to: (E’, **x’***) has been observed, and it could have come from E’ or E”. The inference uses the unconditional sampling distribution, averaging over the two experiments E’ and E”.

We may have a term for Birnbaum’s gambit for dealing with SLP pairs: *Birnbaumization*. Birnbaumizing the observed (E”, **x**”*) yields

> **x*** has been observed and it might have come from E’ or E” and we are not told which.

The more common contemporary formulation of his argument, equivalently, defines the statistic T-B as mapping both **x’*** and **x”*** into **x’***.[20]

Example 4. Let E” be optional stopping as in the SLP violation of Example 2. E” is run, and statistical significance at level p is obtained after 169 trials. So it is known that:

> (E”, **x”***) and the stopping rule ends at 169 trials.

This result has an SLP pair, namely, fixed sample size testing of the null with n fixed at 169, and a result with statistical significance level p observed. We report T-B:

> T-B(E”, **x**”*) = (E’, **x’***): statistical significance with fixed n = 169.

Since the index of the experiment is erased, the report is:

> Statistical significance was obtained at trial 169 and the experiment might have come from optional stopping or fixed sample-size testing.

We obtain premise one: Using the unconditional assessment of Birnbaumization:

$\text{Infr}_{\text{E-B}}(\mathbf{x'}*)$ equiv $\text{Infr}_{\text{E-B}}(\mathbf{x''}*)$,

the convex combination of the two. The result known to have been observed receives more weight because it could have resulted from the not-performed experiment*, the opposite of what the WCP stipulates* in the companion premise two.

(If the result from E does not have an SLP pair, then just report it as (E, **z**)[21]).

So in this sense Birnbaumization is "performable," even without applying any randomizing device. For the sake of the argument, we can imagine that faced with a member of an SLP violation pair from E', someone might contemplate that "the result" could have come from its SLP pair, E" (where the "the result" would be **x*** without its index), and ask: *Ought one therefore to treat the SLP pairs identically?* Given the difference in sampling distributions, the sampling theorist would answer no. But Birnbaum invites us to consider that **x**'* resulted from a mixture experiment with components E' and E", and only statistic T-B reported. Why, from the perspective of informative parametric inference, should we remove the known information about which experiment produced the outcome? Doing so, we can see, would be desirable from the perspective of Birnbaum's argument: it replaces the two experiments with a single (enlarged) experiment, so as to make it possible to apply the sampling theorist's sufficiency principle (SP). By reporting statistic T-B, Birnbaumization has managed to treat members of SLP pairs identically: reporting the unconditional properties averaging over the two.[22]

To further support the possibility of at least visualizing, if not literally performing, Birnbaumization, consider Birnbaum's initial treatment. He imagines that for "any two (mathematical models of) experiments, having the common parameter space $\Theta = \{ \theta \}$," we can "consider the (hypothetical) mixture experiment [E-B] whose components are just [E' and E"] taken with equal probabilities. (1962, 496). So, in a sense, *all* of the possible SLP pairs are "out there" already, hypothetically. Even though an experimenter would not know which of the pair of hypothetical mixtures comes into play until after (E', **x**'*) is known, once it is, we pluck the pair of relevance and ignore any other possibilities. Birnbaum is not concerned with how many pairs there might be because he has no intention of actually computing the convex combinations. Thus, in answer to Cox, Birnbaum will just average over the pair of relevance, determined post-data. We grant all of this, and yet the purported conclusion will still not drop out.

6.3 Kalbfleisch (1975), Evans, Fraser and Monette (1986)

Kalbfleisch (1975) notes that the SP alone in premise one has rendered the SLP pair $\mathbf{x'}^*$, $\mathbf{x''}^*$ evidentially equivalent, and “since merely by defining a mixture experiment the sufficiency principle alone leads to similar conclusions, it would appear that intuitive and operational arguments against [SLP] can be made to apply to [SP] with equal force” (p. 252). He directs himself to revising the SP. But, there is no objection to the SP itself, if one accepts that any member of an SLP pair is remodelled in accordance with Birnbaumization. The problem is that even allowing Birnbaumizing the outcome, the argument does not reach the SLP conclusion. The SLP conclusion is:

$\mathrm{Infr}_{E'}(\mathbf{x'}^*)$ equiv $\mathrm{Infr}_{E''}(\mathbf{x''}^*)$,

where $\mathrm{Infr}_{E'}(\mathbf{x'}^*)$ and $\mathrm{Infr}_{E''}(\mathbf{x''}^*)$ each refer to the individual experiments E’and E”, adequately represented by the models and distributions attached to E’, E” respectively. This was Birnbaum’s own stipulation. The manner in which they are rendered equivalent by Birnbaumization no longer refers to these two experiments, but to some *third* (hypothetical) convex combination of the two, with a distinct sampling distribution.

Evans, Fraser and Monette (1986) offer a number of interesting reflections. Like Kalbfleish, they draw attention to how the Birnbaum argument plays on a lack of clarity in applying principles, SP and WCP. Our two formulations of the argument in 5.1 and 5.2, respectively, parallel the two variations of SP that they introduce.[23] They also offer two parallel variations on WCP. Among other contrasts, the current critique does not turn on altering the principles or restricting their use, or even considering their intended rationale. On the last point, Evans, *et al.* (1986) are quite right to claim that Birnbaum’s argument uses the principles (SP) and (WCP) in ways that are “contrary to the intentions of the principles” as judged by the examples that motivate them (p. 193). If the WCP is used as intended (in Cox’s weighing machine example) rather than as an equivalence relation, Birnbaum’s argument fails. We prefer to show how it fails, even accepting Birnbaum's recommended formulation. We return to this in considering Durbin.

6.4 Durbin (1970)

Paul Durbin (1970) takes issue with Birnbaum's application of the WCP. To employ our terms, his concern is that Birnbaumizing the E^j erases the indicator j showing which experiment was performed, so j is not part of the minimal sufficient statistic for E-B (in premise one). He argues:

> Since evidential meaning depends only on the minimal sufficient statistic it would seem reasonable to require that any analysis or interpretation of the results of the experiment should depend only on the value of the minimal sufficient statistics. This leads naturally to the requirement that the domain of applicability of (WCP) should be restricted to the components of the minimal sufficient statistic. (p. 396)

This leads to his modified principle of conditionality WCP', which could be applicable only if the outcome of the randomizer, j, is the value of a minimal sufficient statistic of a mixture experiment. The reason the sufficiency principle (SP) goes through in premise one turns on being able to erase the particular experiment and report that **x*** could have come from E' or E". Thus once the SP is applied in premise one, Durbin's WCP' is barred from applying in premise two.

Birnbaum's 1970 response to Durbin is interesting. He objects to the idea that "the order of applying the two concepts can make a difference."

> [WCP] and [SP] are equivalence relations, expressing respective general concepts, and in principle justifying substitutions in two directions. (Here no effects can follow from changes of 'order of application') (Birbaum 1970a, 403).

Yet he had admitted in his 1969 paper (p. 139, note 11) that "The formulation of the conditionality concept as one of equivalence," as in the WCP, "was proposed by this writer in (1962) as the natural explication of the concept, not withstanding the one-sided form to which applications of the concept had been restricted (substitution of simpler for less simple models of evidence)."

Admitting that his use of the WCP seems to have deviated from its intended “one-sided form”, and from the entire rationale for the WCP, which was in regard to actual, not hypothetical, mixtures, it is surprising that Birnbaum does not grapple with the deeper issue Durbin is raising. In Evans, et.al. (1986) as well, it is urged that Birnbaum’s argument uses principles of evidence “in contradiction to their original supporting examples.”(182) But it is more than that, and the main points of Durbin, and Evans et.al, could be pushed further to show that the problem is not a matter of order of application, nor gainsaying the intended purpose of the principles, nor even seeing both the SP and WCP as equivalences. The problem, we have seen, is that Birnbaum’s argument precludes both the WCP and SP being consistently applied (first formulation 5.1). We can state Birnbaum’s argument in terms of true “if then” claims (second formulation 5.2), but the SLP still cannot be inferred.

From premise two to premise one. It is illuminating to see what happens if we consider starting first from the WCP i.e., going from our premise two to premise one. The WCP begins with “given it is known experiment E’ was performed and **x**’ resulted,”–the first of the three assertions involved in the WCP (see section 4.2). The clause must therefore remain throughout the argument. However, its status in the Birnbaumized experiment is unclear: when (E’, **x’***) is “known” it means only that it came from either E’ or E” and we do not know which. Thus the inference based on statistic T-B is allowed to be influenced by the unperformed member of the pair. This would be deadly for applying the WCP, which says once (E’, **x’***) is known, inference is *not* to be influenced by an unperformed experiment (even if we had an actual rather than a hypothetical mixture)!

Once it is known that (E’, **x’***) is observed, there is no harm in seeing this as mathematically equivalent to a component of Birnbaum’s hypothetical mixture with components {E’, E”}, *provided that nothing in the analysis of (E’, **x’**) is influenced by what is stipulated to be irrelevant*. That is, WCP as an equivalence is akin to conjoining to (E’, **x’***) a claim that can cause no evidential alteration to the analysis. But Birnbaum simultaneously (in the companion premise) must renege on his stipulation, and proceed to insist that the interpretation of (E’,**x’***) be influenced (considerably!) by the addition of something that by definition cannot influence it.

In our argument, principles SP and WCP do not change, nor are they amended, nor does order matter. If applied in any order in a self-consistent manner, the argument fails. If applied inconsistently, the argument also fails.

Other discussants of Birnbaum bring out important points of unclarity as regards the experimental frameworks within Birnbaum's argument.[24] Space considerations prevent our going into them. They do not go as far as purporting, as we do here, the unsoundness of his argument, and unfortunately, have fallen prey to being dismissed (e.g., as assuming the falsity of the SLP) .

7. Concluding Remarks

We have analyzed Birnbaum's (1962) well-known argument purporting to show that principles of sufficiency (SP) and weak conditionality (WCP) entail the strong likelihood principle (SLP) from a number of perspectives. Our main focus was on the core logical flaws that have enabled Birnbaum's paradox to persist for over fifty years. While our criticism is in sync with portions of existing objections (discussed in section 6) we avoid certain weaknesses that have allowed Birnbaum's argument to stand. In particular, we do not propose to revise the principles that are taken to entail the SLP, but rather to show that Birnbaum's argument, in any of its forms, necessarily applies these principles in a self-contradictory manner. It requires, in effect, that the evidential import of a known result **x**' from experiment E' should, and *also should not,* be influenced by an unperformed experiment E". Our treatment also differs in allowing the relevant principles of evidence to be equivalence relations. However, the terms are forced to change within Birnbaum's argument. Some anticipated questions are now briefly considered.

Are we assuming sampling theory statistics? To this question, our answer is twofold: First, the reason the result has generated so much foundational interest, and the reason it is deemed what Savage termed a "breakthrough" is that it purports to demonstrate the SLP follows from principles to which the sampling theorist would assent. Second, our critical analysis does not depend on accepting any statistical philosophy, nor even on endorsing the principles evoked, at least those that are not purely mathematical[25]. Rather than assume sampling theory statistics, we use it to show the unsoundness of the general argument, which purports to show that WCP and SP entails the SLP—to allude to a common abbreviation. It is uncontroversial that sampling theory violates the SLP—outcomes other than the **x** observed are relevant for determining what it is warranted to infer from **x**. But this hardly means inferences are to be influenced by experiments other than the one known to have generated **x**. In retaining such SLP violations, however, there is no violation either of the sufficiency (SP) or weak conditionality (WCP) principles. If we are correct, this refutes a position that is generally presented as settled in current texts.

> 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)

The principles of evidence SP and WCP hold within a given statistical model of experiment E. Applied simultaneously to two opposed models of an experiment, conflicting results are just what should be expected.

Is it relevant for SLP violations in nonsubjective Bayesianism?

While extensions beyond sampling theory fall outside of the current restricted goal, our formulation points to a very general flaw. Contemporary nonsubjective or default Bayesian methods concede they "have to live with some violations of the likelihood and stopping rule principles" (Ghosh, Delampady, and Sumanta 2006, 148); we suspect the current critique can be extended to justify SLP violations within a Bayesian formulation. Such an extension may well illuminate a core point of agreement between nonsubjective Bayesians and sampling theorists. "The most common response of objective Bayesians" Jim Berger (2006, 394) remarks, is that 'objectivity' can only be defined relative to some frame of reference, and the natural frame for statistics is the statistical model:

> This, of course, does not happen with subjective Bayesianism. Again, 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, 394)

It will be argued, correctly, that there may be more than one way to model an experiment. This does not render the model arbitrary or lacking in justification.[26]

Did Birnbaum detect a flaw in his own argument?

One can only speculate from published work. Based on his responses to the first round of criticisms to Birnbaum (1962), leading him to restrict the SLP to point against point hypotheses (1968, 1969), and to introduce additional "axioms" (Birnbaum 1972), it appears that he did. [27] However, even stating these concessions, we cannot say Birnbaum is explicitly retracting his argument. He suggests that it is because "the likelihood concept cannot be construed so as to allow useful appraisal, and thereby possible control, of probabilities of erroneous interpretations (1969, 128) that he recommends the SLP be limited to "the severely restricted case of a parameter space of just two points." It seems he granted this much: an outcome can be replaced by the value of a sufficient statistic for an experiment E, whether it be a mixture or not, but Infr(E, x) must use the sampling distribution of E (in applying SP in sampling theory). Either that, or his argument would already assume that evidential import comes by way of likelihoods alone, which of course would render it circular.

Birnbaum (1969) came to accept a sampling theory notion of evidence, which he called the *confidence concept*, and which he showed violates the SLP (Birnbaum 1977)[28].

In his 1970 letter to *Nature* (commenting on Edwards 1969), Birnbaum writes:

> If there has been 'one rock in a shifting scene' …in recent decades, it has not been the likelihood concept, as Edwards suggests, but rather the concept by which confidence limits and hypothesis tests are usually interpreted".

Nevertheless, he still felt that this evidential construal of N-P theory was at odds with its more usual behavioristic construal, and hoped to encourage "communication among interested statisticians, scientific workers and philosophers and historians of science." (ibid.) We heartily endorse this conclusion.

References:

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

Berger, J. O. (1985). The frequentist viewpoint and conditioning. In L. LeCam and R. Olshen (Eds.), *Proceedings of the Berkeley conference in honor of Jack Kiefer and Jerzy Neyman* (pp. 15-44). Belmont, CA: Wadsworth.

Berger, J. O. (1986), Discussion on a paper by Evans et al. [On principles and arguments to likelihood]. *Canadian Journal of Statistics, 14,* 195-6.

Berger, J.O. (2006). The case for objective Bayesian analysis. Bayesian Analysis, 1, 385-402.

Berger, J. O., and Wolpert, R. L. (1988). *The likelihood principle*. Hayward, CA: California Institute of Mathematical Statistics.

Bernardo, J. M. (2005). Reference analysis. In D. K. Dey and C. R. Rao (Eds.), *Handbook of Statistics 25* (pp. 17-90). Amsterdam: Elsevier.

Birnbaum, A. (1962). On the foundations of statistical inference. In S. Kotz and N. Johnson (eds), *Breakthroughs in statistics,* (Vol.1, pp. 478-518). Springer Series in Statistics, New York: Springer-Verlag. Reprinted from *Journal of the American Statistical Association, 57,* 269–306.

Birnbaum, A. (1964). The anomalous concept of statistical evidence. Mimeographed technical report, IMM-NYU 332, *Courant Inst. Of Math. Sci*., NYU.

Birnbaum, A. (1968). Likelihood. In *International encyclopedia of the social sciences* (Vol. 9, pp. 299-301). New York: Macmillan and the Free Press.

Birnbaum, A. (1969). Concepts of Statistical Evidence. In S. Morgenbesser, P. Suppes, and M. G. White (Eds.), *Philosophy, science, and method: Essays in honor of Ernst Nagel* (pp. 112-143). New York: St. Martins.

Birnbaum, A. (1970a). On Durbin's modified principle of conditionality. *Journal of the American Statistical Association, 65*, 402-3.

Birbaum, A (1970b). Statistical Methods in Scientific Inference (letter to the editor). *Nature* 225, 1033.

Birnbaum, A. (1972), More on concepts of statistical evidence. *Journal of the American Statistical Association*, *67*, 858–61.

Birnbaum, A. (1975), Comments on Paper by J. D. Kalbfleisch. *Biometrika*, *62*(2), 262-264.

Bjornstad, J. (1992), Introduction to Birnbaum (1962): On the foundations of statistical inference. In S. Kotz and N. Johnson (eds), *Breakthroughs in statistics,* (Vol.1, pp. 461-477). Springer Series in Statistics, New York: Springer-Verlag.

Carlin, B. and Lewis, T. (2009). Bayesian Methods for Data Analysis, 3rd edition. Chapman Hall (CRC)

Casella, G., and Berger, R. L. (2002). *Statistical inference* (2nd ed.). Pacific Grove, CA: Duxbury Press.

Cox, D. R. (1958). Some problems connected with statistical inference. *The Annals of Mathematical Statistics, 29,* 357-372.

Cox, D. R. (1977). The role of significance tests [With discussion]. *Scandinavian Journal of Statistics*, *4,* 49–70.

Cox, D. R. (1978). Foundations of statistical inference: The case for eclecticism. *Australian Journal of Statistics*, *20*(1), 43-59. Knibbs Lecture, Statistical Society of Australia, 1977.

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

Cox, D. R., and Mayo, D. G. (2010). Objectivity and conditionality in frequentist inference. In D. Mayo and A. Spanos (Eds.), *Error and inference: Recent exchanges on experimental reasoning, reliability, and the objectivity and rationality of science* (pp. 276-304). Cambridge: Cambridge University Press.

Dawid, (1986), Discussion on paper by Evans et al. [On principles and arguments to likelihood]. *Canadian Journal of Statistics, 14,* 196-7.

Durbin, J. (1970). On Birnbaum's theorem on the relation between sufficiency, conditionality and likelihood. *Journal of the American Statistical Association, 65,* 395–8.

Edwards, A.W.F. (1969) Statistical methods in scientific inference. *Nature,* 222, 1233.

Edwards, W., Lindman, H., and Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70*, 193-242.

Evans, M. J., Fraser, D. A. S., and Monette, G. (1986). Likelihood. *Canadian Journal of Statistics*, *14*, 180-90.

Ghosh, J. K., Delampady, M., and Samanta, T. (2006). *An introduction to Bayesian analysis*. New York: Springer.

Giere, R. (1977). Allan Birnbaum's Conception of Statistical Evidence. Synthese 36 (1), 5-13.

Giere, R. (1977). Publications by Allan Birnbaum. Synthese 36 (1), 15-17.

Kalbfleisch, J. D. (1975). Sufficiency and conditionality. *Biometrika*, *62*(2), 251-259; Replies to comments, 268.

Lehmann, E. L., and Romano, J. P. (2010). *Testing statistical hypotheses* (3rd ed.). Springer Texts in Statistics, New York: Springer.

Lindley D. V. (1976). Bayesian statistics. In W. L. Harper and C.A. Hooker (Eds.), *Foundations of probability theory, statistical inference and statistical theories of science* (Vol. 2, pp. 353-362), Dordrect, The Netherlands: D. Reidel.

Mayo, D. G. (1996) *Error and the growth of experimental knowledge*. Chicago: University of Chicago Press.

Mayo, D. G. (2010a). An error in the argument from conditionality and sufficiency to the likelihood principle. In D. Mayo and A. Spanos (Eds.), *Error and inference: Recent exchanges on experimental reasoning, reliability, and the objectivity and rationality of science* (pp. 305-314). Cambridge: Cambridge University Press.

Mayo, D. G. and D. R. Cox (2010a), Frequentist Statistics as a Theory of Inductive Inference. In D Mayo and A. Spanos (Eds.), *Error and inference: Recent exchanges on experimental reasoning, reliability and the objectivity and rationality of science*, (pp. 247-275). Cambridge: Cambridge University Press. First published in *The second Erich L. Lehmann symposium: Optimality* (2006) Lecture Notes-Monograph Series, Vol. 49, Institute of Mathematical Statistics, (pp. 1-27).

Mayo, D. G. and D. R. Cox (2011), Statistical scientist meets a philosopher of science: a conversation with Sir David Cox. *Rationality, markets and morals (RMM), Special topic issue, Statistical science and philosophy of science: Where do* *(should) they meet in 2011 and beyond?* D. Mayo, A. Spanos, and K. Staley (Guest Eds.) Vol. 2, (pp. 103-114).

Mayo, D.G. and Kruse, M. (2001), "Principles of inference and their consequences". In D. Cornfield and J. Williamson (Eds.), *Foundations of Bayesianism* (pp. 381-403). Dordrecht, Kluwer Academic Publisher.

Mayo, D. and Spanos, A. (2011) Error statistics. In Prasanta S. Bandyopadhyay and Malcolm R. Forster (Volume Eds.); Dov M.Gabbay, Paul Thagard and John Woods (General Eds.) *Philosophy of statistics, Handbook of philosophy of science* (Vol 7, pp. 1-46). The Netherlands: Elsevier

Reid, N. (1992). Introduction to Fraser (1966) structural probability and a generalization. In S. Kotz and N. L. Johnson (Eds.). *Breakthroughs in statistics: Foundations and basic theory* (Vol. 1, pp. 579-587). New York: Springer.

Savage, L. (1962a), "Subjective probability and statistical practice". In G. A. Barnard and D. R. Cox (Eds.), *The foundations of statistical inference: A discussion* (pp. 9-35.) London: Methuen.

Savage, L. J. (1962b). Discussion on a paper by A. Birnbaum [On the foundations of statistical inference]. *Journal of the American Statistical Association*, *57,* 307–308.

Savage, L. J., Barnard, G., Cornfield, J., Bross, I, Box, G., Good, I., Lindley, D., Clunies-Ross, C., Pratt, J., Levene, H., Goldman, T., Dempster, A., Kempthorne, O, and Birnbaum, A. (1962). On the foundaitons of statistical inference: Discussion (of Birnbaum 1962). *Journal of the American Statistical Association,57*, 307-326.

Savage, L. J., (1970), Comments on a weakened principle of conditionality, *Journal of the American Statistical Association*, *65*(329), 399-401.

[1] Savage went on to predict that "once the likelihood principle is widely recognized, people will not long stop at that halfway house but will go forward and accept the implications of personalistic probability for statistics" (1962b, 308).

[2] We replace (E, x) and (E',x') with (E',x') and (E",x"), respectively, for consistency with our notation.

[3] Elsewhere I discuss the SLP in relation to the philosophy of statistics that I favor, e.g., Mayo 1996 (chapters 9-11), Mayo and Kruse 2001, Mayo and Spanos 2006.

[4] We substitute (SLP) for (LP) and (WCP) for (CP) in accordance with our definitions in this paper. We do not discuss broader conditionality principles, but Birnbaum's argument very explicitly only relies on the weakest form as we define it.

[5] Prior probability distributions employed in nonsubjective (or "reference") Bayesian accounts are influenced by the sampling distribution, and therefore strictly violate the SLP (e.g., Berger 2006; Bernardo 2005). Some may argue that the prior is part of the model and so this influence of the sampling distribution occurs only as part of arriving at the model.

[6] This application of the SLP to the case of optional stopping is often called the Stopping Rule Principle (SRP) (Berger and Wolpert 1988). Applying the stopping rule principle requires stipulating that the stopping rule was uninformative for the inference, as in the above example. See an early discussion in Savage (1962a). We do not discuss distinctions for discrete and continuous experimental models.

[7] Casella and Berger call the strong likelihood principle the "Formal Likelihood Principle," (Casella and Berger 2002, 292).

[8] Evans, Fraser and Monette (1986, 186) employ a principle of "mathematical equivalence". Here too the equivalence is within an experimental sampling distribution. Birnbaum (1972) had discussed an analogous principle.

[9] J, its distribution being independent of **X**, is an ancillary statistics. This discussion proceeds independently of any general theory of conditioning.

[10] Lehmann is clear that reporting the observed significance level (or significance probability) is acceptable, and even recommended, to permit individuals to apply their own criteria. Thus, for simplicity in the illustration, we can continue to talk of p-values or observed significance levels.

[11] Cox and Hinkley 1974, pp. 95-97, also illustrate the WCP in examples where unconditional N-P tests may do better on average than the conditional one, but for informative parametric inference the conditional test seems appropriate.

[12] Our treating WCP as an equivalence here distinguishes the current discussion from that of Mayo 2010.

[13] For that matter, as Birnbaum suggests (1969, 119), a "trivial but harmless" augmentation to any experiment might be to toss a fair coin and report heads or tails (where this was irrelevant to the original model). Given (E', **x'**),

$\text{Infr}_{E'}(\mathbf{x'})$ equiv [$\text{Infr}_{E'}(\mathbf{x'})$ and either a coin was tossed or it was not].

He intends the move in applying the WCP is to be just as innocuous as the report of an irrelevant coin toss.

[14] Were the premises true and the conclusion false, a logical contradiction (equivalent to Q & ~Q) would result.

[15] In any example that forms an SLP violation, we know that the sampling distribution of T-B differs from that of E' and E".

[16] Given the use of "conditional" for a different purpose throughout, I deliberately steer clear of also using it to describe an "if then" statement.

[17] This is an equivalent, but more abbreviated way of writing:

$\text{Infr}_{E\text{-}B}(E',\mathbf{x'})$ equiv $\text{Infr}_{E\text{-}B}(E'',\mathbf{x''})$.

[18] See Note 24.

[19] We sometimes hear sampling theory criticized because "two experiments with identical likelihoods could result in different p-values if the two experiments were designed differently." (Carlin and Louis 2009, 51). At first glance this may seem plausible: how can the "intended" design make a difference? But it is not known whether, had the unperformed and differently designed experiment been run, the result would have had an "identical likelihood" to the one observed. Moreover, at least for a sampling theorist, there may be a difference in what can be said about (actual or hypothetical) future trials of the differently designed experiments.

[20] That is, whether **x*** came from E' or E", we report it as having come from E', never E". But since it must be added that the convex combination will be appealed to in determining the inference, it must boil down to saying it came from one or the other. Equivalently, even though we know which experiment produced the result, we must just report its likelihood and average.

[21] The case where there is no SLP pair is not unproblematic, but we put that problem to one side.

[22] Note that in some cases every outcome from E' will have an SLP pair in E", but it need not. For example, for the binomial and negative binomial, each outcome has an SLP pair.

[23] They call them "S(as)" and "S(if)", p. 192. It would take us too far afield to discuss the numerous, rich, and interesting variations on principles found intheir work.

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

[25] Elsewhere the SLP is discussed in relation to a general philosophy of statistical inference that we call error statistics, e.g., Mayo 1996 (chapters 9-11), Mayo and Kruse 2001, Mayo and Spanos 2006.

[26] We do not take up the much-discussed issue of the lack of "unique ancillaries" on which to condition. Our analysis does not turn on this.

[27] In his 1962 responses, Birnbaum was already backing away from appealing to the WCP in his argument, suggesting that some kind of "censoring principle," proposed by Pratt (1962), could work instead.

[28] We do not think this gives an adequate construal of evidence; it is still too behavioristic. See Mayo 1996.