A New Proof of the Likelihood Principle

By Greg Gandenberger

Abstract

I present a new proof of the Likelihood Principle that avoids two responses to a well-known proof due to Birnbaum ([1962]). I also respond to arguments that Birnbaum's proof is fallacious, which if correct would apply to this new proof as well. On the other hand, I urge caution in interpreting proofs of the Likelihood Principle as arguments against the use of frequentist statistical methods.

1 Introduction

Allan Birnbaum showed in his ([1962]) that the Likelihood Principle follows from the conjunction of the Sufficiency Principle and the Weak Conditionality Principle.[1,2] The

[1]Birnbaum calls the principles he uses simply the Sufficiency Principle and the Conditionality Principle. Dawid ([1977]) distinguishes between weak and strong versions of the Sufficiency Principle, but this distinction is of little interest: to my knowledge, no one accepts the Weak Sufficiency Principle but not the Strong Sufficiency Principle. The distinction between weak and strong versions of the Conditionality Principle (due to Basu [1975]) is of much greater interest: the Weak Conditionality Principle is much more intuitively obvious, and Kalbfleisch's influential response to Birnbaum's proof (discussed in Section 3) involves rejecting the Strong Conditionality Principle but not the Weak Conditionality Principle.

[2]The Conditionality Principle Birnbaum states in his ([1962]) is actually the Strong Conditionality Principle, but the proof he gives requires only the weak version. Birnbaum strengthens his proof in a later paper

Sufficiency Principle and the Weak Conditionality Principle are intuitively appealing, and some common frequentist practices appear to presuppose them. Yet frequentist methods violate the Likelihood Principle, while likelihoodist methods and Bayesian conditioning do not.[3] As a result, many statisticians and philosophers have regarded Birnbaum's proof as a serious challenge to frequentist methods and a promising 'sales tactic' for Bayesian methods.[4]

Most frequentist responses to Birnbaum's proof fall into one of three categories: (1) proposed restrictions on its premises, (2) allegations that it is fallacious, and (3) objections to the framework within which it is expressed. In this paper I respond to objections in categories (1) and (2). In Section 2, I give a new proof of the Likelihood Principle that avoids responses in category (1). In Section 3, I explain that analogues of the minimal restrictions on Birnbaum's premises that are adequate to block his proof are not adequate to block the new proof. The responses in category (2) apply to this new proof as well, but I argue in Section 4 that those responses are mistaken. Arguments in category (3) have been less influential because standard frequentist theories presuppose the same framework that Birnbaum uses. For objections to theories that use a different framework, see (Berger and Wolpert [1988], pp. 47–64).

([1972]) by showing that the logically weaker Mathematical Equivalence Principle can take the place of the Weak Sufficiency Principle. I present Birnbaum's proof using the Weak Sufficiency Principle rather than the Mathematical Equivalence Principle because the former is easier to understand and replacing it with the latter does not address any important objections to the proof.

[3]Throughout this paper, unless otherwise specified 'frequentist' and 'frequentism' refer to 'statistical frequentist' views about statistical inference that emphasise the importance of using methods with good long-run operating characteristics, rather than to 'probability frequentist' views according to which probability statements should be understood as statements about some kind of long-run frequency. The earliest statistical frequentists were also probability frequentists, but the historical and logical connections between statistical frequentism and probability frequentism are complex. I use the word 'frequentist' despite its ambiguity because it is the most widely recognised label for statistical frequentism and because recognised alternatives also have problems (Grossman [2011b], pp. 68–70).

[4]See for instance (Birnbaum [1962], p. 272; Savage [1962], p. 307; Berger and Wolpert [1988], pp. 65–6; Mayo [1996], p. 391, fn. 17; and Grossman [2011b], p. 8), among many others. Birnbaum himself continued to favour frequentist methods even as he refined his proof of the Likelihood Principle ([1970b]). He claims that the fact that frequentist principles conflict with the Likelihood Principle indicates that our concept of evidence is 'anomalous' ([1964]). I regard the frequentist principles that conflict with the Likelihood Principle (such the Weak Repeated Sampling Principle from Cox and Hinkley [1974], pp. 45–6) not as plausible constraints on the notion of evidence, but rather as articulating potential reasons to use frequentist methods despite the fact that they fail to track evidential meaning in accordance with the intuitions that lead to the Likelihood Principle. This view differs from Birnbaum's only in how it treats words like 'evidence' and 'evidential meaning,' but this verbal change seems to me to clarify matters that Birnbaum obscures.

I see little hope for a frequentist in trying to show that Birnbaum's proof and the new proof given here are unsound, but one can question the use of these proofs as objections to frequentist methods. The Likelihood Principle as Birnbaum (e.g. [1962], p. 271) and I formulate it says roughly that two experimental outcomes are evidentially equivalent if they have proportional likelihood functions—that is, if the probabilities[5] that the set of hypotheses under consideration assign to *those outcomes* are proportional as functions of those hypotheses. Frequentist methods violate the Likelihood Principle because their outputs can vary with the probabilities some of the hypotheses under consideration ascribe to *unobserved* outcomes, such as outcomes 'more extreme than' the ones observed in the case of p-values.

The Likelihood Principle implies that frequentist methods should not be used if one assumes that a method of inference should not be used if it can produce different outputs given evidentially equivalent inputs. However, this assumption is not as innocuous as it might seem. It is at best a slight oversimplification: frequentist methods are useful even if Bayesian methods are in some sense 'correct' because frequentist methods often provide adequate, computationally efficient approximations to Bayesian methods. In addition, although idealised Bayesian conditioning conforms to the Likelihood Principle, the Bayesian methods that statisticians actually use in fact violate the Likelihood Principle as well, albeit in subtle ways that generally have little or no effect on the results of their analyses.[6] More importantly, the assumption presupposes an 'ev-

[5] In continuous cases, we would be dealing with probability densities rather than probabilities.

[6] I say that a method of inference *violates* a sufficient condition for evidential equivalence such as the Likelihood Principle if in some possible situation it would produce different outputs depending on which datum it receives among a set of data that the principle implies are evidentially equivalent without a difference in utilities, prior opinions, or background knowledge. I say that a method that does not violate a given condition of this kind *conforms to* it. In theory, subjective Bayesians conform to the Likelihood Principle by updating their belief in H upon learning E by the formula $P_{new}(H) = P_{old}(H|E) = \frac{P_{old}(H)P(E|H)}{\sum_i P(E|H_i)P_{old}(H_i)}$, where i ranges over an index set of the set of hypotheses under consideration. (The sum becomes an integral in the case of continuous hypothesis spaces.) Thus, $P_{new}(H)$ depends on E only through the likelihoods $P(E|H_i)$, as conforming to the Likelihood Principle requires. In practice, subjective Bayesians typically use methods that depend on the sampling distribution to estimate an expert's $P_{old}(H)$, such as methods that involve fitting a prior distribution that is conjugate to the sampling distribution. Objective Bayesians use priors that depend on the sampling distribution in order to achieve some aim such as maximising a measure of the degree to which the posterior distribution depends on the data rather than the prior, as in the reference Bayesian approach (Berger [2006] p. 394). Some contemporary Bayesians (e.g. the authors of Gelman et

identialist' epistemology that some statisticians and philosophers reject. For instance, frequentist pioneers Jerzy Neyman and Egon Pearson claim that frequentist methods should be interpreted not in evidential terms but simply as decision rules warranted by their long-run performance (e.g. [1933], pp. 290–1).[7] The use of the Likelihood Principle as an objection to frequentist methods simply begs the question against this view. Many frequentists regard the Neyman-Pearson approach as too 'behavioristic' for use in science (e.g. Fisher [1955]), but there are 'conditional frequentist' approaches (initiated by Kiefer [1977]) that attempt to address this problem by ensuring that the measure of long-run performance used is relevant to the particular application in question. I do not claim that evidentialism is false or that conditional frequentism is viable, but only that those who would use the Likelihood Principle as an argument against frequentist methods need to account for such views.

Proofs of the Likelihood Principle have implications for the philosophy of statistics and perhaps for statistical practice even if they do not warrant the claim that frequentist methods should not be used. The Likelihood Principle does imply that evidential frequentism—the view that frequentist methods track the evidential meaning of data—is false.[8] This conclusion is relevant to debates internal to frequentism that plausibly hinge on whether frequentist methods should be understood as tracking evidential meaning, as decision rules justified by their operating characteristics, or in some other way (Mayo [1985]). Topics of such debates include the use of accept/reject rules rather than p values, predesignation rules, stopping rules, randomised tests, and the use of

al. [2003], pp. 157–96) also endorse model-checking procedures that violate the Likelihood Principle more drastically. It is worth noting that neither subjective nor objective Bayesians violate the Likelihood Principle in a different sense of 'violates' than the one used here, even when checking their models: they do not allow information not contained in the likelihood of the observed data to influence the inferences they draw conditional on a model (Gelman [2012]). But they generally do allow the sampling distribution of the experiment (for instance, whether the experiment is binomial or negative binomial) to influence their choice of a model, and thereby potentially influence the conclusions they reach.

[7] Incidentally, but there is evidence that Pearson was never fully committed to this view (Mayo [1992]).

[8] I take it that what makes a method frequentist is that it would provide some kind of guarantee about long-run performance in repeated applications to the same experiment with varying data, which requires that its outputs depend on the probabilities of unobserved sample points in violation of the Likelihood Principle. I also take it that a method that violates a true sufficient condition for evidential equivalence thereby fails to track evidential meaning.

conditional procedures.

A few technical notes are in order before proceeding. I assume that inferences are being performed in the context of a statistical model of the form $(\mathcal{X}, \Theta, \mathbf{P})$, where $\mathcal{X}$ is a finite sample space of (possibly vector-valued) points $\{\mathrm{x}\}$, Θ a possibly uncountable parameter space of (possibly vector-valued) points $\{\theta\}$, and $\mathbf{P}$ a family (not necessarily parametric) of probability distributions P_θ over $\mathcal{X}$ indexed by θ. Against standard practice in the philosophy of science, I follow Birnbaum (e.g. [1962], pp. 269–70) and others writers in the literature about the Likelihood Principle in using the term 'experiment' to refer to any data-generating process with such a model, even when that process is not manipulated.[9] I assume that $\mathbf{P}$ contains all of the hypotheses of interest. The model may also include a prior probability distribution over Θ and/or a utility/loss function defined on the Cartesian product of Θ and a set of possible outputs. Following (Grossman [2011a], p. 561), I assume that the choice of experiments is not informative about θ.[10]

The assumption that sample spaces are finite restricts the scope of the proof given here, but this fact is not a serious problem for at least two reasons. First, infinite sample spaces are merely convenient idealisations. Real measuring devices have finite precision, and real measurable quantities are bounded even if our knowledge of their bounds is vague.[11] Second, the view that the Likelihood Principle holds for experiments with finite sample spaces but not for infinite sample spaces is implausible, unappealing, and insufficient to save evidential frequentism. Thus, a proof of the Likelihood Princi-

[9] This broad use of the term 'experiment' is not ideal, but there is no alternative that is obviously better. A nontechnical term such as 'observational situation' fails to convey the presence of a statistical model. Grossman's term 'merriment' ([2011b], p. 63) is less apt to give rise to misconceptions but not widely recognised.

[10] I do not follow Grossman ([2011a], p. 561) in assuming that utilities are either independent of the observation or unimportant. This restriction is needed when the Likelihood Principle is formulated as a claim about what kinds of methods should be used, but not when it is formulated as a sufficient condition for two outcomes to be evidentially equivalent.

[11] For instance, we might model the entry of customers into a bank as a Poisson process, but we would not take seriously the implication of this model that with positive probability ten billion customers will enter the bank in one second. The model neglects constraints such as the sizes of the bank and of the world population that become relevant only far out in the tails of the distribution.

ple for experiments with finite sample spaces would be sufficient for present purposes even if there were actual experiments with truly infinite sample spaces. Moreover, it seems likely that the proof given here could be extended to experiments with continuous sample spaces, as Berger and Wolpert extend Birnbaum's proof ([1988], pp. 32–6). Attempting to extend the proof given here in this way would be an interesting technical exercise, but for the reasons just discussed it would not be either philosophically or practically illuminating.

Birnbaum's proof is more elegant than the proof given here, and its premises are easier to grasp. On the other hand, the new proof is safe against natural responses to Birnbaum's proof that have been influential. Thus, while Birnbaum's proof may have more initial persuasive appeal than the proof given here, the new proof is better able to withstand critical scrutiny.

2 The New Proof

I show in this section that the Likelihood Principle follows from the conjunction of the Experimental Conditionality Principle and what I call the Weak Ancillary Realisability Principle. In the next section I display the advantages this proof has over Birnbaum's.

Both the Experimental Conditionality Principle and the Weak Ancillary Realisability Principle appeal to the notion of a *mixture experiment*. A mixture experiment consists of using a random process to select one of a set of component experiments and then performing the selected experiment, where the component experiments share a common index set Θ that is independent of the selection process. For instance, one might flip a coin to decide which of two thermometers to use for a measurement that will be used to test a particular hypothesis, and then perform that measurement and the associated test. A non-mixture experiment is called *minimal*. A mixture experiment with two minimal, equiprobable components is called *simple*.

Roughly speaking, the Experimental Conditionality Principle says that the outcome

of a mixture experiment is evidentially equivalent to the corresponding outcome of the component experiment actually performed. The Weak Ancillary Realisability Principle says that the outcome of a minimal experiment is evidentially equivalent to the corresponding outcome of a two-stage mixture experiment with an isomorphic sampling distribution.

The Experimental Conditionality Principle can be expressed more formally as follows:

> **The Experimental Conditionality Principle.** For any outcome x of any component E' of any mixture experiment E, $\text{Ev}(E, (E', x)) = \text{Ev}(E', x)$.

where '$\text{Ev}(E, x)$' refers to the 'evidential meaning' of outcome x of experiment E, and '$(E, (E', x))$' refers to outcome x of component E' of mixture experiment E. 'Evidential meaning' is an undefined primitive notion that principles like those discussed in this paper are intended partially to explicate. In words, the Experimental Conditionality Principle says that the evidential meaning of the outcome of an experiment does not depend on whether that experiment is performed by itself or as part of a mixture.

The Experimental Conditionality Principle has considerable intuitive appeal: denying it means accepting that the appropriate evidential interpretation of an experiment can depend on whether another experiment that was not performed had a chance of being performed. If this claim does not seem odd in the abstract, then consider it in the case of the following example:

> **Example 1.** Suppose you work in a laboratory that contains three thermometers, T_1, T_2, and T_3. All three thermometers produce measurements that are normally distributed about the true temperature being measured. The variance of T_1's measurements is equal to that of T_2's but much smaller than that of T_3's. T_1 belongs to your colleague John, so he always

gets to use it. T_2 and T_3 are common lab property, so there are frequent disputes over the use of T_2. One day, you and another colleague both want to use T_2, so you toss a fair coin to decide who gets it. You win the toss and take T_2. That day, you and John happen to be performing identical experiments that involve testing whether the temperature of your respective indistinguishable samples of some substance is greater than 0°C or not. John uses T_1 to measure his sample and finds that his result is just statistically significantly different from 0°. John celebrates and begins making plans to publish his result. You use T_2 to measure your sample and happen to measure exactly the same value as John. You celebrate as well and begin to think about how you can beat John to publication. 'Not so fast,' John says. 'Your experiment was different from mine. I was bound to use T_1 all along, whereas you had only a 50% chance of using T_2. You need to include that fact in your calculations. When you do, you'll find that your result is no longer significant.'

According to radically 'behaviouristic' forms of frequentism, John may be correct. You performed a mixture experiment by flipping a coin to decide which of two thermometers to use, and thus which of two component experiments to perform. The uniformly most powerful level α test[12] for that mixture experiment does *not* consist of performing the uniformly most powerful level α test for whichever component experiment is actually performed. Instead, it involves accepting probability of Type I error greater than α when T_3 is used in exchange for a probability of Type I error less than α when T_2 is used, in such a way that the probability of Type I error for the mixture experiment as a whole remains α (see Cox [1958], p. 360).

Most statisticians, including most frequentists, reject this line of reasoning. It

[12] A uniformly most powerful test of significance level α is a test that maximises the probability of rejecting the null hypothesis under each simple component of the alternative hypothesis among all tests that would reject the null hypothesis with probability no greater than α if the null hypothesis were true.

seems suspicious for at least three reasons. First, the claim that your measurement warrants different conclusions from John's seems bizarre. They are numerically identical measurements from indistinguishable samples of the same substance made using measuring instruments with the same stochastic properties. The only difference between your procedures is that John was 'bound' to use the thermometer he used, whereas you had a 50% chance of using a less precise thermometer. It seems odd to claim that the fact that you could have used a instrument other than the one you actually used is relevant to the interpretation of the measurement you actually got using the instrument you actually used. Second, the claim that John was 'bound' to use T_1 warrants scrutiny. Suppose that he had won that thermometer on a bet he made ten years ago that he had a 50% chance of winning, and that if he hadn't won that bet, he would have been using T_3 for his measurements. According to his own reasoning, this fact would mean that his result is not statistically significant after all.[13] The implication that one might have to take into account a bet made ten years ago that has nothing to do with the system of interest to analyse John's experiment is hard to swallow. In fact, this problem is much deeper than the fanciful example of John winning the thermometer in a bet would suggest. If John's use of T_1 as opposed to some other thermometer with different stochastic properties was a nontrivial result of *any* random process at any point in the past that was independent of the temperature being measured, then the denial of Weak Conditionality Principle as applied to this example implies that John analysed his data using a procedure that fails to track evidential meaning.[14] Third, at the time of your analysis you *know* which thermometer you received. How could it be

[13] One could argue that events like outcomes of coin tosses are determined by the laws of physics and relevant initial conditions anyway, so that both you and John were bound to use the thermometer you actually did use, but applying the same argument to any appeal to randomness that does not arise from genuine indeterminism would undermine frequentist justifications based on sampling distributions in almost all cases. In addition, this argument could be avoided by replacing the coin toss in the example with a truly indeterministic process such as (let us suppose) radioactive decay.

[14] This kind of reasoning makes plausible the claim that most if not all real experiments are components of mixture experiments that we cannot hope to identify, much less model, and thus that assuming something like the Experimental Conditionality Principle is necessary for performing any data analysis at all (Kalbfleisch [1975], p. 254).

better epistemically to fail to take that knowledge into account?

It is sometimes said (e.g. Wasserman [2012]) that an argument for the Weak Conditionality Principle like the one just given involves a hasty generalisation from a single example. However, the purpose of the example is merely to make vivid the intuition that features of experiments that could have been but were not performed are irrelevant to the evidential meaning of the outcome of the experiment that actually was performed. The intuition the example evokes, rather than the example itself, justifies the principle.

The Experimental Conditionality Principle does go beyond the example just discussed in that it applies to mixture experiments with arbitrary probability distributions over arbitrarily (finitely) many components. The intuition the example evokes has nothing to do with the number of component experiments in the mixture or the probability distribution over those experiments, so this extension is innocuous. Also, the Experimental Conditionality Principle does not require that the component experiment actually performed be minimal. Thus, it can be applied to a nested mixture experiment, and by mathematical induction it implies that the outcome of a nested mixture experiment with any finite number of 'stages' is evidentially equivalent to the corresponding outcome of the minimal experiment actually performed. There does not seem to be any reason to baulk at applying the principle repeatedly to nested mixture experiments in this way. Thus, the Experimental Conditionality Principle is not a hasty generalisation because insofar as it generalises beyond the example just discussed it does so in an unobjectionable way.

It is sometimes useful to express conditionality principles in terms of ancillary statistics. An ancillary statistic for an experiment is a statistic that has the same distribution under each of the hypotheses under consideration. A statistic that indexes the outcomes of a random process used to decide which component of a mixture experiment to perform is ancillary for that mixture experiment. Other ancillary statistics

arise not from the mixture structure of the experiment but from the set of hypotheses under consideration. Kalbfleisch ([1975]) calls a statistic that indexes the outcomes of an overt random process used to decide which component experiment to perform *experimental* ancillaries, and a statistic that is ancillary only because of features of the set of hypotheses under consideration a *mathematical* ancillary. those that arise from the set of hypotheses under consideration *mathematical* ancillaries. The Experimental Conditionality Principle permits conditioning on experimental ancillaries but does not address conditioning on mathematical ancillaries. The distinction between experimental and mathematical ancillaries is extra-mathematical in the sense that it is not given by the model $(\mathcal{X}, \Theta, \mathbf{P})$. As a result, it is difficult if not impossible to formulate that distinction precisely. This fact is not an objection to my proof in the relevant dialectical context because I use the distinction between experimental and mathematical ancillaries only to address a response to Birnbaum's proof due to Kalbfleisch [1975] that requires it. (See Section 3).

The Weak Ancillary Realisability Principle essentially says that one may replace one binary mathematical ancillary in a minimal experiment with an experimental ancillary without changing the evidential meanings of the outcomes. The following reasoning helps motivate this principle. Consider a hypothetical experiment with the sampling distribution given by table 1 below, where the cells of the table correspond to a partition of the sample space.

	v_1	v_2	v_3	v_4
h	$P_\theta(x_1)$	$P_\theta(x_2)$	$P_\theta(x_3)$	$P_\theta(x_4)$
t	$P_\theta(x_5)$	$P_\theta(x_6)$	$P_\theta(x_7)$	$P_\theta(x_8)$

Table 1: The sampling distribution of a hypothetical experiment used to motivate the Weak Ancillary Realisability Principle

For all $i = 1, \ldots, 8$, $P_\theta(x_i)$ is known only as a function of θ. However, it is known that $P_\theta(h) = P_\theta(v_1) = \frac{1}{2}$. Such a sampling distribution could arise by many processes, such as (1) the roll of an appropriately weighted eight-sided die with faces labelled

$x_1, \ldots, x_8$; (2) the flip of a fair coin with sides labelled h and t followed by the roll of either an appropriately weighted four-sided die with faces labelled $x_1, \ldots, x_4$ (if the coin lands heads) or an appropriately weighted four-sided die with faces labelled $x_5, \ldots, x_8$ (if the coin lands tails); or (3) the flip of a fair coin with sides labelled v_1, $\sim v_1$ followed by either the flip of an appropriately biased coin with sides labelled x_1 and x_5 (if the first coin lands on side v_1) or the roll of an appropriately weighted six-sided die with faces labelled $x_2, x_3, x_4, x_6, x_7, x_8$ (if the first coin lands tails). The intuition that the Weak Ancillary Realisability Principle aims to capture is that the outcome with likelihood function $P_\theta(x_1)$ has the same evidential meaning with respect to θ regardless of which of these kinds of process produces it: it makes no difference evidentially whether such an outcome arises from a one-stage process, a two-stage process in which the row is selected and then the column, or a two-stage process in which either the first column or its complement is selected and then the cell, provided that the overall sampling distribution is the same in the three cases. It is worth taking a moment to be sure that one has grasped this somewhat complicated example and satisfied oneself that the intuition I claim it evokes is indeed intuitive. The key step in the proof of the Likelihood Principle given below is the construction of a hypothetical experiment that has the same essential features as the experiment just described.

The formal statement of the Weak Ancillary Realisability Principle uses the following terminology and notation. For a given set of probability distributions $\mathbf{P} = \{P_\theta : \theta \in \Theta\}$ with support $\mathcal{X}$ and a given set $A \subset \mathcal{X}$, let ${}^A\mathbf{P}$ refer to the set of distributions obtained by replacing $P_\theta(X) \in \mathbf{P}$ with ${}^AP_\theta(X) = P_\theta(X|X \in A)$ for each $\theta \in \Theta$ and $X \in \mathcal{X}$, and let $A^C = \mathcal{X} \setminus A$. Call experimental models $(\mathcal{X}, \Theta, \{P^1_\theta\})$ and $(\mathcal{Y}, \Theta, \{P^2_\theta\})$ isomorphic under the one-to-one mapping $f : \mathcal{X} \to \mathcal{Y}$ if and only if $P^1_\theta(x) = P^2_\theta(f(x))$ for all $x \in \mathcal{X}$ and all $\theta \in \Theta$. The Weak Ancillary Realisability Principle can now be stated as follows.

The Weak Ancillary Realisability Principle. Let E be a minimal ex-

periment with model $(\mathcal{X}, \Theta, \mathbf{P})$ such that there is a set $A \subset \mathcal{X}$ such that for all $\theta \in \Theta$, $\Pr_\theta(X \in A) = p$ for some known $0 \leq p \leq 1$. Let E' consist of flipping a coin with known bias p for heads to decide between performing component experiment E_1 if the coin lands heads and E_2 if the coin lands tails, where E_1 is a minimal experiment with model isomorphic to $(A, \Theta, ^A\mathbf{P})$ under f_1, and E_2 is a minimal experiment with model isomorphic to $(A^C, \Theta, ^{A^C}\mathbf{P})$ under f_2. Under these conditions $\text{Ev}(E, x) = \text{Ev}(E', (E_1, f_1(x)))$ for all $x \in A$, and $\text{Ev}(E, x) = \text{Ev}(E', (E_2, f_2(x)))$ for all $x \in A^C$.

In words, if a minimal experiment's sample space can be partitioned into two sets of outcomes A and A^C such that the probability that the observed outcome is in A is known to be p, then the outcome of that experiment has the same evidential meaning as the corresponding outcome of an experiment that consists of first flipping a coin with bias p for heads and then performing a minimal experiment that is isomorphic to a minimal experiment over A if the coin lands heads and a minimal experiment that is isomorphic to a minimal experiment over A^C if it lands tails. Roughly speaking, this principle allows one to break a minimal experiment into two stages by turning a mathematical ancillary into an experimental ancillary.

The Weak Ancillary Realisability Principle does not require that the outcomes of the components E_1 and E_2 of E' be literally the same as the corresponding outcomes of E. An alternative approach to the one I have taken here would be to include this requirement and to adopt in addition a weak version of Dawid's Distribution Principle ([1977], p. 247) which says that corresponding outcomes of minimal experiments with isomorphic models are evidentially equivalent. I have chosen not to take this approach in order to make it easier to compare my proof to Birnbaum's, but considering it is instructive. The Distribution Principle this approach requires is weaker than Dawid's in that it only applies to minimal experiments and thus is compatible with Kalbfleisch's

response to Birnbaum's proof that is discussed in the next section. Adding to the Weak Ancillary Realisability Principle the requirement that the outcomes of the components E_1 and E_2 of E' be literally the same as the corresponding outcomes of E makes it possible to argue for the Weak Ancillary Realisability Principle as follows. It is always possible to form an appropriate E' for a given E provided that E can be repeated indefinitely many times. Simply flip a coin with probability p for heads. If the coin lands heads, repeat E until an outcome in A occurs, and then report that outcome. If the coin lands tails, repeat E until an outcome in A^C occurs, and then report that outcome. The claim that corresponding outcomes of some E and the E' formed from it in this way are evidentially equivalent is highly intuitive. The fact that the E' outcome occurred after some unspecified number of unspecified outcomes in the sample space of the component experiment not selected is uninformative because one already knows the probability of such outcomes (p or $1 - p$).

The Weak Ancillary Realisability Principle is different from what one might call the Strong Ancillary Realisability Principle, which says across the board that the distinction between experimental and mathematical ancillaries is irrelevant to evidential meaning. In contrast, the Weak Ancillary Realisability Principle permits conditioning only on a *single*, *binary* ancillary (the indicator for A) from a *minimal* experiment. The proof given here would be redundant otherwise: the conjunction of the Strong Ancillary Realisability Principle and the Experimental Conditionality Principle obviously implies a Strong Conditionality Principle that permits conditioning on any ancillary, which has already been shown to imply the Likelihood Principle (Evans et al. [1986]).

The Strong Ancillary Realisability Principle may be true and does have some intuitive appeal, but the Weak Ancillary Realisability Principle is far easier to defend because one can construct a single simple illustration in which the principle seems obviously compelling that essentially covers all of the cases to which the principle applies. To demonstrate this point, I will start with an illustration that is a bit too sim-

ple and then add the necessary additional structure (see Figure 1). These illustrations make the same point as the example given in conjunction with table 1 above, but their concreteness makes them more vivid and thus perhaps more convincing.

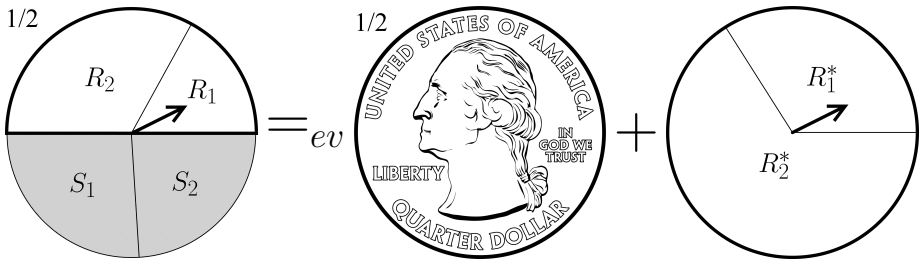

Figure 1: A pair of evidentially equivalent outcomes from corresponding instances of the pair of procedures described in Example 2.[16]

Example 2. Consider an ideal spinner divided into regions R_1, R_2, S_1, and S_2 such that one knows that R_1 and R_2 together occupy half of the spinner's area. Suppose that one wished to draw inferences about the relative sizes of the regions knowing only that one's data were generated using one of the following two procedures:

> **One-Stage Procedure.** Spin the spinner and report the result.
>
> **Two-Stage Procedure.** Flip a coin with bias $\frac{1}{2}$ for heads. If the coin lands heads, replace the spinner with one divided into two regions R_1^* and R_2^* such that for $i = 1$ or 2, R_i^* occupies the same fraction of the new spinner that R_i occupies of the half of the original spinner that R_1 and R_2 occupy together. If this spinner lands on R_i^*, report R_i as the result. If the coin lands tails, do likewise with the S regions.

[16]Quarter clipart courtesy FCIT: Portrait on a Quarter, retrieved February 2, 2013 from <etc.usf.edu/clipart/40200/40232/quart_front_40232.htm>.

Intuitively, knowing whether the one-stage or the two-stage procedure was performed would not help in drawing inferences about the relative sizes of the spinner regions from one's data. The difference between these two procedures does not matter for such inferences. Each procedure generates the same sampling distribution; the fact that one does so in one step while the other does so in two steps is irrelevant.

In Example 2, the fraction of the spinner that R_1 and R_2 are known to occupy and the bias of the coin used in the two-stage procedure are $\frac{1}{2}$. But the intuition that the example evokes has nothing to do with the fact that this number is $\frac{1}{2}$. Nor does it have anything to do with the fact that there are two R regions and two S regions. Thus, we can safely extend this intuition to the following more general example (see Figure 2).

Figure 2: A pair of evidentially equivalent outcomes from corresponding instances of the pair of procedures described in Example 3. The bias of the coin for heads p is the same as the fraction of the first spinner that the R regions occupy in total.

Example 3. Consider an ideal spinner divided into regions R_1, R_2, …, R_n and S_1, S_2, …, S_m such that one knows that R_1, R_2, …, R_n together occupy proportion p of the spinner for some particular $0 \leq p \leq 1$. Suppose that one wished to draw inferences about the relative sizes of the regions knowing only that one's data were generated using one of the following two procedures:

One-Stage Procedure. Spin the spinner and report the result.

Two-Stage Procedure. Flip a coin with bias p for heads. If the coin lands heads, replace the spinner with one divided into n regions R_1^*, R_2^*, ..., R_n^* such that for $i = 1, \ldots, n$, R_i^* occupies the same fraction of the new spinner that R_i occupies of the percentage p of the original spinner that $R_1, \ldots, R_n$ occupy together. If this spinner lands on R_i^*, report R_i as the result. If the coin lands tails, do likewise with the S regions.

Again, it seems obvious that knowing whether the one-stage or the two-stage procedure was performed would not help in drawing inferences about the relative sizes of the spinner regions from one's data. As long as the sampling distribution remains unchanged, whether an experiment is performed in one step or two is irrelevant to the evidential meanings of its outcomes.

The Weak Ancillary Realisability Principle does extend the intuition Example 3 evokes to experiments involving data-generating mechanisms that are not spinners. There is nothing special about spinners driving that intuition, so this extension is innocuous.

There is one fact about the Weak Ancillary Realisability Principle that is important to the proof of the Likelihood Principle which Example 2 does not illustrate, namely that it can apply to a single experiment in more than one way. Call a partition of a sample space that is indexed by an ancillary statistic (such as $\{A, A^c\}$ in the statement of the Weak Ancillary Realisability Principle) an *ancillary partition*. An experiment can have multiple ancillary partitions to which the Weak Ancillary Realisability Principle applies. The proof of the Likelihood Principle presented below involves applying the Weak Ancillary Realisability Principle to two ancillary partitions of the same experiment, so it is worth considering an example of an experiment of this kind in order to confirm that it does not violate our intuitions (See Figure 3).

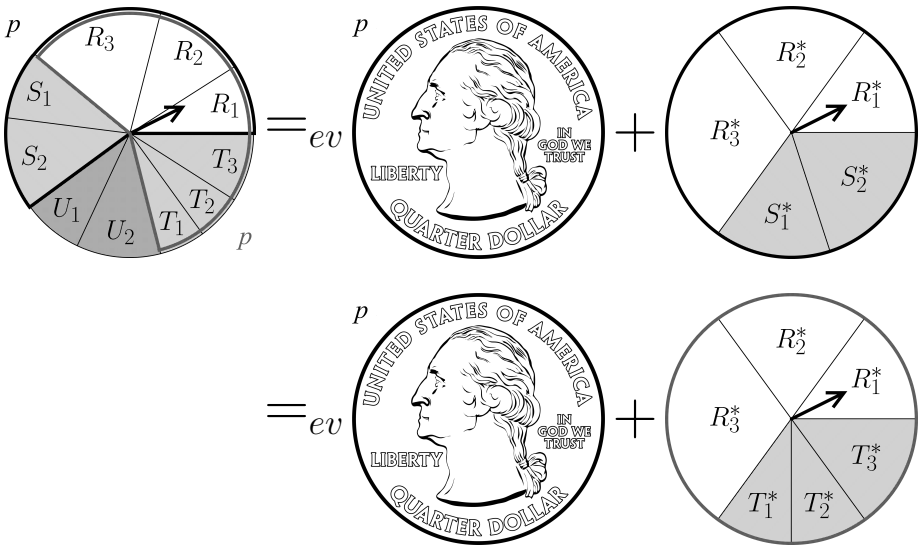

Figure 3: A pair of evidentially equivalent outcomes from corresponding instances of the pair of procedures described in Example 4. The bias of the coin for heads p is the same as the fraction of the first spinner that the R and S regions occupy in total, which is the same as the fraction that the R and T regions occupy.

Example 4. Consider an ideal spinner divided into regions R_1, R_2, …, R_n; $S_1, S_2, \ldots, S_m$; $T_1, T_2, \ldots, T_l$; and $U_1, U_2, \ldots, U_k$, such that for some $0 \leq p \leq 1$ one knows both that the R regions and the S regions together occupy proportion p of the spinner and that the R regions and the T regions together occupy proportion p of the spinner. Suppose that one wished to draw inferences about the relative sizes of the regions knowing only that one's data were generated using one of the following three procedures:

> **One-Stage Procedure.** Spin the spinner and report the result.
>
> **Two-Stage Procedure A.** Flip a coin with bias p for heads. If the coin lands heads, replace the spinner with one divided into $n + m$ regions R_1^*, R_2^*, …, R_n^*, S_1^*, S_2^*, …, S_m^* such that for all $i =$

$1, \ldots, n$, R_i^* occupies the same fraction of the new spinner that R_i occupies of the percentage p of the original spinner that $R_1, \ldots, R_n$ and $S_1, \ldots, S_m$ occupy together, and likewise for S_i^* and S_i for all $i = 1, \ldots, m$. If this spinner lands on R_i^* or S_i^*, report R_i or S_i as the result, respectively. If the coin lands tails, do likewise with the T and U regions.

Two-Stage Procedure B. Perform Two-Stage Procedure A but reverse the roles of S and T.

The intuition that corresponding outcomes of the One-Stage Procedure and of Two-Stage Procedure A are evidentially equivalent seems to be completely unaffected by the fact that one could also perform Two-Stage Procedure B, and vice versa. Thus, there is no reason not to apply the Weak Ancillary Realisability Principle twice to experiments with two ancillary partitions.

The Likelihood Principle can be expressed formally as follows:

The Likelihood Principle. Let E_1 and E_2 be experiments with a common parameter space Θ, and let x and y be outcomes of E_1 and E_2, respectively, such that $P_\theta(x) = cP_\theta(y)$ for all $\theta \in \Theta$ and some positive c that is constant in θ. Then $\text{Ev}(E_1, x) = \text{Ev}(E_2, y)$.

In words, two experimental outcomes with proportional likelihood functions for the same parameter are evidentially equivalent.

I can now prove the following result:

Theorem 1. *The Experimental Conditionality Principle and the Weak Ancillary Realisability Principle jointly entail the Likelihood Principle.*

Proof. Consider an arbitrary pair of experiments E_1 and E_2 with respective models $(\mathcal{X}, \Theta, \{P_\theta^1\})$ and $(\mathcal{Y}, \Theta, \{P_\theta^2\})$ such that $\mathcal{X} = \{x_0, x_1,$

$\ldots, x_n\}$, $\mathcal{Y} = \{y_0, y_1, \ldots, y_m\}$, and $P_\theta^1(x_0) = cP_\theta^2(y_0)$ for all $\theta \in \Theta$ and some $c \geq 1$ that is constant in θ. There is no loss of generality in the assumption $c \geq 1$ because one can simply swap the labels of E_1, E_2, and their outcomes if $P_\theta^1(x_0) < P_\theta^2(y_0)$. x_0 is the outcome $x_0^\dagger$ of some unique minimal experiment $E_1^\dagger$ with sample space $\mathcal{X}^\dagger$ that is performed with some known probability q when E_1 is performed.[17] $E_1^\dagger$ is either E_1 iticself or a proper component of E_1.[18] x_0 just is $(E_1^\dagger, x_0^\dagger)$, so by the reflexivity of the evidential equivalence relation $\text{Ev}(E_1, x_0) = \text{Ev}(E_1, (E_1^\dagger, x_0^\dagger))$.[19] By the Experimental Conditionality Principle, $\text{Ev}(E_1, (E_1^\dagger, x_0^\dagger)) = \text{Ev}(E_1^\dagger, x_0^\dagger)$.[20]

Construct a hypothetical minimal experiment E_1^{CE} with sample space $\mathcal{X}^{CE}$ and sampling distribution given by table 2[21]. Although E_1^{CE} is minimal, I trust that no confusion will result from the use of expressions of the form (d, z_i) and (e, z_i) to refer to points in $\mathcal{X}^{CE}$ in accordance with table 2. The arrangement of sample points into rows and columns in table 2 only serves to display the relevant (mathematical) ancillary partitions of $\mathcal{X}^{CE}$. The outcomes in the first row that correspond to outcomes of $E_1^\dagger$ (that is, $\{(d, z_i) : \exists(x^\dagger \in \mathcal{X}^\dagger)(x_i = (E_1^\dagger, x^\dagger)\}$) constitute a set $A \subset \mathcal{X}$ such that $\Pr(X \in A) = p$ for all $X \in \mathcal{X}$ and some known $0 \leq p \leq 1$, namely $\frac{q}{2}$. Likewise, the outcomes in the first column (that is, $\{(d, z_0), (e, z_0)\}$ constitute a set $A \subset \mathcal{X}$ such that $\Pr(X \in A) = p$ for all $X \in \mathcal{X}$ and some known $0 \leq p \leq 1$, namely $\frac{1}{2}$.

[17] $E_1^\dagger$ is unique because experimental ancillaries involve overt randomisation. Thus, the problem of the nonuniqueness of maximal ancillaries that plagues frequentist attempts to incorporate conditioning on ancillary statistics in general (see Basu [1964] and subsequent discussion) does not arise. That q is known follows from the fact that the model specifies P_θ^1 for each $\theta \in \Theta\}$ and the stipulation that the process by which the component of a mixture experiment to be performed is selected is independent of θ.

[18] I am treating the 'component' relation for experiments as transitive.

[19] Evidential equivalence is assumed to be an equivalence relation.

[20] When E_1 is minimal, $E_1^\dagger = E_1$ and $x_0^\dagger = x_0$, so $\text{Ev}(E_1, x_0) = \text{Ev}(E_1^\dagger, x_0^\dagger)$ by reflexivity alone.

[21] The construction used in this table is a modified version of the construction Evans et al. use to show that the Likelihood Principle follows from the Strong Conditionality Principle alone ([1986], p. 188).

	z_0	z_1	z_2	z_3	$\ldots$	z_n
d	$\frac{1}{2}P_\theta^1(x_0)$	$\frac{1}{2}P_\theta^1(x_1)$	$\frac{1}{2}P_\theta^1(x_2)$	$\frac{1}{2}P_\theta^1(x_3)$	$\ldots$	$\frac{1}{2}P_\theta^1(x_n)$
e	$\frac{1}{2} - \frac{1}{2}P_\theta^1(x_0)$	$\frac{1}{2}P_\theta^1(x_0) - \frac{1}{2}\min_\theta P_\theta^1(x_0)$	$\frac{1}{2}\min_\theta P_\theta^1(x_0)$	0	$\ldots$	0

Table 2: Sampling distribution of E_1^{CE}

Let E_1^M be an experiment that consists of flipping a coin with bias $\frac{q}{2}$ for heads to choose between performing $E_1^\dagger$ if the coin lands heads and performing a minimal experiment with sampling distribution given by the distribution of E_1^{CE} conditional on the complement of the set of outcomes that correspond to outcomes of $E_1^\dagger$ if the coin lands tails. By the Weak Ancillary Realisability Principle, $\text{Ev}(E_1^{CE}, (d, z_0)) = \text{Ev}(E_1^M, (E_1^\dagger, x_0^\dagger))$. By the Experimental Conditionality Principle, $\text{Ev}(E_1^M, (E_1^\dagger, x_0^\dagger)) = \text{Ev}(E_1^\dagger, x_0^\dagger)$. From all of the equivalences established so far it follows that $\text{Ev}(E_1, x_0) = \text{Ev}(E_1^{CE}, (d, z_0))$.

Next construct a hypothetical Bernoulli experiment E^B with sample space (g, h) and sampling distribution given by $P_\theta^B(g) = P_\theta^1(x_0)$. Finally, construct a mixture experiment $E_1^{M_B}$ that consists of first flipping a coin with bias $\frac{1}{2}$ for heads to decide between performing E^B and performing a minimal experiment with the known sampling distribution given by the distribution of E_1^{CE} conditional on the complement of the first-column outcomes $\{(d, z_0), (e, z_0)\}$. By the Weak Ancillary Realisability Principle, $\text{Ev}(E_1^{CE}, (d, z_0)) = \text{Ev}(E_1^{M_B}, (E^B, g))$. By the Experimental Conditionality Principle, $\text{Ev}(E_1^{M_B}, (E^B, g)) = \text{Ev}(E^B, g)$. It follows that $\text{Ev}(E_1^{CE}, (d, z_0)) = \text{Ev}(E^B, g)$.

From $\text{Ev}(E_1, x_0) = \text{Ev}(E_1^{CE}, (d, z_0))$ and $\text{Ev}(E_1^{CE}, (d, z_0)) = \text{Ev}(E^B, g)$, it follows that $\text{Ev}(E_1, x_0) = \text{Ev}(E^B, g)$. An analogous construction establishes $\text{Ev}(E_2, y_0) = \text{Ev}(E^B, g)$, and thus $\text{Ev}(E_1, x_0) = \text{Ev}(E_2, y_0)$. (See the appendix for the details of this construction.) The only restriction we

placed on (E_1, x_0) and (E_2, y_0) in establishing this result is that they have proportional likelihood functions, so the Likelihood Principle follows by universal generalisation. □

Figure 4 provides a graphical depiction of this proof.

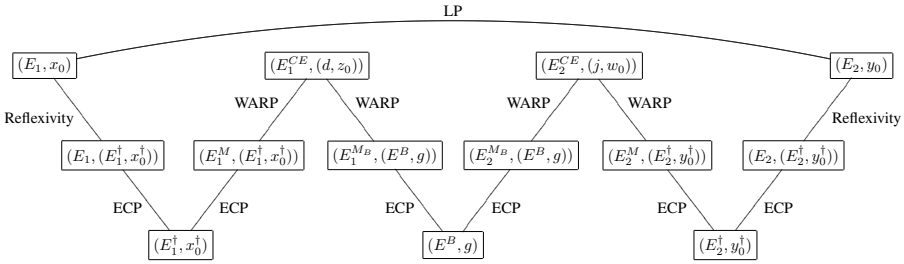

Figure 4: Graphical depiction of the series of equivalences used to establish Theorem 1. Boxes refer to experimental outcomes, edges between boxes indicate evidential equivalence, and labels on edges indicate the principle used to establish evidential equivalence. (ECP = Experimental Conditionality Principle, WARP = Weak Ancillary Realisability Principle, LP = Likelihood Principle)..

3 How the New Proof Addresses Proposals to Restrict Birnbaum's Premises

Birnbaum ([1962]) shows that the Likelihood Principle follows from the conjunction of the Sufficiency Principle and the Weak Conditionality Principle. Durbin ([1970]) responds to Birnbaum's proof by restricting the Weak Conditionality Principle, while Kalbfleisch ([1975]) responds by restricting the the Sufficiency Principle.[22] Analogous restrictions on the premises of the proof given in the previous section do not suffice to save evidential frequentism.

I will briefly present Birnbaum's proof, and then explain how the new proof addresses Durbin's and Kalbfleisch's responses to it. The Weak Conditionality Principle

[22]Cox and Hinkley suggest a response similar to Kalbfleisch's in their earlier ([1974]), but they do not develop the idea as fully as Kalbfleisch.

is the Experimental Conditionality Principle restricted to simple mixture experiments. It can be stated as follows:

> **The Weak Conditionality Principle**: For any outcome x of any component E' of any simple mixture experiment E, $\text{Ev}(E, (E', x)) = \text{Ev}(E', x)$.

This principle is logically weaker than the Experimental Conditionality Principle, but there does not seem to be any reason to accept the former but not the latter.

The Sufficiency Principle says that two experimental outcomes that give the same value of a sufficient statistic are evidentially equivalent. The notion of a *sufficient statistic* is intended to explicate the informal idea of a statistic that simplifies the full data without losing any of the information about the model it contains. Formally, a statistic is called sufficient with respect to the hypotheses of interest if and only if it takes the same value for a set of outcomes only if the probability distribution over those outcomes given that one of them occurs does not depend on which of those hypotheses is true. The Sufficiency Principle seems eminently plausible: if the probability distribution over a set of outcomes given that one of them occurs does not depend on which hypothesis is true, then it is not clear how those outcomes could support different conclusions about those hypotheses. The Weak Sufficiency Principle can be stated formally as follows:

> **The Sufficiency Principle** (S): Consider an experiment $E = (\mathcal{X}, \{\theta\}, \mathbf{P})$ where T(X) is a sufficient statistic for θ. For any $x_1, x_2 \in \mathcal{X}$, if $T(x_1) = T(x_2)$ then $\text{Ev}(E, x_1) = \text{Ev}(E, x_2)$.

The Sufficiency Principle would underwrite the practice, common among frequentists as well as advocates of other statistical paradigms, of 'reducing to a sufficient statistic,' that is, reporting only the value of a sufficient statistic rather than reporting the full data. For instance, from a sequence of a fixed number of coin tosses that are assumed to be independent and identically distributed with probability p of heads, a fre-

quentist would typically report only the number of heads in the sequence (a sufficient statistic for p) rather than the sequence itself.

One can also formalise the notion of a *minimal sufficient* statistic, which retains all of the information about the model that is in the full data but cannot be simplified further without discarding some such information. Formally, a statistic is minimal sufficient for θ in a given experiment if and only if it is sufficient for θ and is a function of every sufficient statistic for θ in that experiment. A minimal sufficient statistic is more coarse-grained than any non-minimal sufficient statistic for the same parameter and experiment, so it provides the greatest simplification of the data that the Sufficiency Principle warrants.

Minimal sufficient statistics are unique up to one-to-one transformation, and a statistic that assigns the same value to a pair of outcomes if and only if they have the same likelihood function is minimal sufficient (Cox and Hinkley [1974], p. 24). Thus, the Sufficiency Principle implies that two outcomes of the same experiment are evidentially equivalent if they have the same likelihood function. This consequence of the Sufficiency Principle is sometimes called the Weak Likelihood Principle (Cox and Hinkley [1974], p. 24); it differs from the Likelihood Principle only in that the latter applies also to outcomes from different experiments. This difference may seem slight, but the Likelihood Principle has implications that are radical from a frequentist perspective, such as the evidential irrelevance of stopping rules, that the Weak Likelihood Principle lacks.

Proving the Likelihood Principle from the Sufficiency Principle requires an additional principle that allows one to 'bridge' different experiments. The Weak Conditionality Principle plays this role in Birnbaum's proof: one simply constructs a hypothetical mixture of the two experiments in question. The proof proceeds as follows. Take an arbitrary pair of experimental outcomes (E_1, x) and (E_2, y) that have the same likelihood function for the same parameter θ. Construct a simple mixture

E^M of E_1 and E_2. $(E^M, (E_1, x_0))$ and $(E^M, (E_2, y_0))$ are two outcomes of the same experiment that have the same likelihood function, so a minimal sufficient statistic for E^M has the same value for those two outcomes. By the Sufficiency Principle, then, $\text{Ev}(E^M, (E_1, x_0)) = \text{Ev}(E^M, (E_2, y_0))$. By the Conditionality Principle, $\text{Ev}(E^M, (E_1, x_0)) = \text{Ev}(E_1, x_0)$ and $\text{Ev}(E^M, (E_2, y_0)) = \text{Ev}(E_2, y_0)$. It follows that $\text{Ev}(E_1, x_0) = \text{Ev}(E_2, y_0)$. (E_1, x_0) and (E_2, y_0) are arbitrary except for the fact that they have the same likelihood function, so the Likelihood Principle follows by a universal generalisation. Figure 5 displays the steps of this proof in a graphical format.

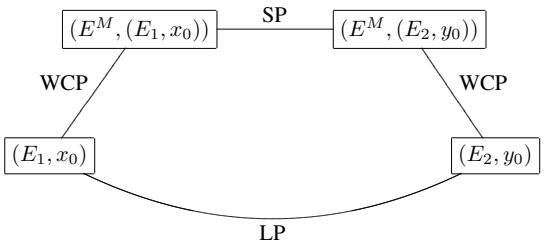

Figure 5: Birnbaum's proof of the Likelihood Principle. Boxes refer to experimental outcomes, edges between boxes indicate evidential equivalence, and labels on edges indicate the principle used to establish evidential equivalence. (WCP = Weak Conditionality Principle, SP = Sufficiency Principle, LP = Likelihood Principle).

Having presented Birnbaum's proof, I will now discuss Durbin's and Kalbfleisch's responses to it in turn. Durbin ([1970]) proposes restrict the Weak Conditionality Principle to experimental ancillaries that are functions of a minimal sufficient statistic. This restriction suffices to block Birnbaum's proof because the outcomes (E_1, x_0) and (E_2, y_0) of Birnbaum's mixture experiment E^M share the same value of a minimal sufficient statistic. Thus, one value of a minimal sufficient statistic for E^M corresponds to two different values of the experimental ancillary that indexes the outcomes of the process used to select which of E_1 and E_2 to perform. It follows that the ancillary statistic that indexes the outcome of the coin flip used to decide which component experiment to perform is not a function of a minimal sufficient statistic, and thus that Durbin's restricted Weak Conditionality Principle does not warrant conditioning on it.

Durbin's proposal faces many strong objections (see e.g. Birnbaum [1970a]; Savage [1970]; and Berger and Wolpert [1988], pp. 45–6). However, influential authorities continue to cite it as a reason to reject Birnbaum's proof (e.g. Casella and Berger [2002], p. 296). Regardless of how those objections to Durbin's approach fare, the proof presented in the previous section makes it effectively moot because the analogue of Durbin's response—restricting the Experimental Conditionality Principle to experimental ancillaries that are functions of a minimal sufficient statistic—allows the proof to go through in all but a restricted class of cases. The view that the Likelihood Principle holds only outside that class of cases is implausible, unattractive, and insufficient to satisfy an evidential frequentist, so for present purposes a proof of the Likelihood Principle outside that class is as good as a proof of the Likelihood Principle in general.

That the proof in the previous section goes through in all but a restricted class of cases when one restricts the Experimental Conditionality Principle as Durbin proposes to restrict the Weak Conditionality Principle can be seen as follows. It suffices to consider the applications of the Experimental Conditionality Principle to $(E_1, (E_1^\dagger, x_0^\dagger))$, $(E_1^M, (E_1^\dagger, x_0^\dagger))$, and $(E_1^{M_B}, (E^B, d))$ because its applications to $(E_2, (E_2^\dagger, y_0^\dagger))$, $(E_2^M, (E_2^\dagger, y_0^\dagger))$, and $(E_2^{M_B}, (E^B, d))$, respectively, are analogous. With Durbin's restriction the Experimental Conditionality Principle can be applied to $(E_1, (E_1^\dagger, x_0^\dagger))$ unless at some stage in the possibly nested mixture experiment E_1, two experimental outcomes from different component experiments have the same likelihood function. E_1 may have this property, but actual experiments typically do not, and the view that the Likelihood Principle holds only for those that do not is implausible, unattractive, and insufficient to satisfy an evidential frequentist. The principle can be applied to $(E_1^M, (E_1^\dagger, x_0^\dagger))$ and $(E_1^{M_B}, (E^B, d))$ unless outcomes in the two components of one of E_1^M or $E_1^{M_B}$, respectively, have the same likelihood function. The sampling distribution of these experiments, given by table 2, has been chosen so that it is not the case that outcomes in the two components of either of those experiments have the same likelihood function

by construction. Some such pair of outcomes may have the same likelihood function because of incidental features of the sampling distribution of E_1, but such cases are of no special interest. Again, the claim that the Likelihood Principle applies only outside such cases is implausible, unattractive, and insufficient to satisfy an evidential frequentist. Thus, the analogue of Durbin's restriction allows the proof given in the previous section to go through for a large class of cases that is sufficient for present purposes.

A second proposal to weaken the premises of Birnbaum's proof is Kalbfleisch's proposal ([1975]) to restrict the Sufficiency Principle to outcomes of minimal experiments.[23] Kalbfleisch's proposal, like Durbin's, faces many serious objections but continues to be cited influential authorities such as Casella and Berger ([2002], p. 296).[24] The proof given in the previous section neatly sidesteps the analogue of Kalbfleisch's proposal because the Weak Ancillary Realisability already applies only to minimal experiments.

Kalbfleisch's proposal is a bit stronger than it needs to be to block Birnbaum's proof: Kalbfleisch prohibits applying the Sufficiency Principle to any mixture experiment, but Birnbaum's proof involves applying it only to a simple mixture experiment. However, a weakened version of Kalbfleisch's proposal that applied only to simple mixtures, call it 'Kalbfleisch*,' would be inadequate because one could easily avoid it by modifying Birnbaum's proof slightly, for instance by adding a third component to E^M or by giving E_1 and E_2 unequal probabilities within E^M.

One might wonder if there is a set of restrictions on the Weak Conditionality Principle and the Sufficiency Principle not stronger than Durbin's or Kalbfleisch* that would suffice to block Birnbaum's proof and the analogue of which would suffice to block

[23] Kalbfleisch also proposes to restrict the Strong Conditionality Principle to allow conditioning on mathematical ancillaries only after reducing by minimal sufficiency, but this change irrelevant to Birnbaum's proof.

[24] For objections to Kalbfleisch's proposal, see (Birnbaum [1975]) and (Berger and Wolpert [1988], pp. 46–67). Savage's objection to Durbin's approach ([1970]) also applies to Kalbfleisch's approach with slight modifications. Savage's objection to Durbin's approach is that it could lead statisticians to draw very different conclusions from experiments that differ 'only microscopically' when the microscopic difference makes a minimal sufficient statistic no longer quite sufficient. The same point applies to Kalbfleisch's approach within a minimal component experiment.

the new proof. In fact, any such set of restrictions would have to appeal to possible incidental features of the arbitrary experiments E_1 and E_2 between outcomes of which evidential equivalence is to be established, rather than to features of other experiments in the proofs that are present by construction. Thus, it would only suffice to restrict the scope of the new proof in a way that one suspects would be implausible, unattractive, and insufficient to satisfy an evidential frequentist. The only experiment Birnbaum constructs in his proof is the simple mixture E^M of E_1 and E_2, which are arbitrary except for the fact that they have a pair of respective outcomes with proportional likelihood functions. Thus, the weakest restriction on the Weak Conditionality Principle that appeals only to features of the proof that are present by construction which suffices to block Birnbaum's proof is to prohibit applying it to mixture experiments outcomes from different components of which have proportional likelihood functions. This restriction is equivalent to Durbin's. And the weakest restriction on the Sufficiency Principle that appeals only to features of experiments in the proof that are present by construction which suffices to block the proof is to prohibit applying it to simple mixture experiments, which is exactly Kalbfleisch*.

Thus, Durbin and Kalbfleisch's responses are in a sense the only options for those who would like to block Birnbaum's proof by restricting its premises. Of course, there are stronger responses that would suffice to block both Birnbaum's proof and the new proof given here, but such a strong response would require a proportionally strong argument.

4 A Response to Arguments that the Proofs are Fallacious

In addition to Durbin's and Kalbfleisch's proposals to block Birnbaum's proof by restricting its premises, there are also arguments due to Joshi ([1990]) and Mayo ([2009],

[2011], [2012]) that Birnbaum's proof is fallacious. The objections Joshi and Mayo present would also apply (*mutatis mutandis*) to the proof presented here, but I will argue that they are mistaken.

The arguments that Birnbaum's proof is fallacious mistake Birnbaum's premises for what we might call their operational counterparts. The Weak Conditionality Principle and the Sufficiency Principle each posit sufficient conditions for experimental outcomes to be evidentially equivalent. They are different, respectively, from what we might call the Operational Weak Conditionality Principle and the Operational Sufficiency Principle. The Operational Weak Conditionality Principle says that in drawing inferences from the outcome of a simple mixture experiment, one ought to use the sampling distribution of the component experiment actually performed, rather than the sampling distribution of the mixture experiment as a whole. The Operational Sufficiency Principle says that one ought to 'reduce by sufficiency' as far as possible—that is, to use for inference the sampling distribution of a minimal sufficient statistic rather than the sampling distribution of the original sample space. Even on the assumption that one ought to use methods that conform to the Weak Conditionality Principle and Sufficiency Principle if they are true, it does not follow that one ought to change the sample space as their operational counterparts prescribe: Bayesian conditioning and likelihoodist methods are insensitive to sample spaces, so they conform to the Weak Conditionality Principle and the Sufficiency Principle regardless of whether one conditions on ancillaries and/or reduces by sufficiency or not.

The Weak Conditionality Principle and the Sufficiency Principle are logically consistent: each of those principles merely asserts that certain sets of experimental outcomes are evidentially equivalent, and the assertion of any set of equivalences is logically consistent with the assertion of any other set of equivalences. However, their operational counterparts can conflict: conditioning on an ancillary statistic can preclude reducing by a particular sufficient statistic and vice versa. A conflict of this kind

does arise in Birnbaum's proof: a minimal sufficient statistic of the mixture experiment E^M assigns the same value to the outcomes (E_1, x_0) and (E_2, y_0), so conditioning on which component experiment is performed precludes reducing to that statistic because only one of (E_1, x_0) and (E_2, y_0) is in the resulting sample space, and reducing to that statistic precludes conditioning on which component experiment is performed when either (E_1, x_0) or (E_2, y_0) occurs because those outcomes come from different component experiments but are indistinguishable in the reduced sample space.

It is easy to see how this conflict between the Operational Weak Conditionality Principle and the Operational Sufficiency Principle could give rise to the claim that Birnbaum's premises are inconsistent. For a frequentist, the only way to conform to each of Birnbaum's premises is to follow the corresponding operational principle. The two operational principles come into conflict, so from a frequentist perspective the original premises seem inconsistent. But this apparent inconsistency is not a legitimate objection to Birnbaum's proof because it presupposes a frequentist use of sampling distributions. Whether or not sampling distributions are relevant to evidential meaning is exactly what is at issue in the debate about the Likelihood Principle, so this presupposition begs the question.

The following passage shows that Joshi does in fact mistake Birnbaum's premises for their operational counterparts:

> For the assumed set-up [in Birnbaum's proof], the conditionality principle essentially means that only the experiment actually performed (E_1 or E_2) is relevant[...] But the same relevancy must hold good when applying the sufficiency principle[...] [A minimal sufficient statistic for E^M] is not a statistic—and hence not a sufficient statistic—under the probability distribution relevant for the inference in the assumed set-up. Hence the sufficiency principle cannot yield [Ev$(E^M, (E_1, x_0))$ = Ev$(E^M, (E_2, y_0))$] [...]so the proof fails. ([1990], pp. 111–2)

Joshi assumes that the Weak Conditionality Principle implies that one must condition on which component of E^M is actually performed before applying the sufficiency principle. But the Weak Conditionality Principle is not a directive to condition at all. Joshi appears to have in mind the Operational Weak Conditionality Principle and a version of the Operational Sufficiency Principle that is restricted along the lines of Kalbfleisch's approach to require conditioning on an experimental ancillary before reducing by sufficiency. Conforming to Birnbaum's premises requires following their operational counterparts only within a frequentist approach, so Joshi's objection begs the question.

Birnbaum himself insisted that his premises were to be understood as 'equivalence relations' rather than as 'substitution rules' and recognised that his proof is valid only when they are understood in this way. As he put, 'It was the adoption of an unqualified equivalence formulation of conditionality, and related concepts, which led, in my 1972 paper, to the monster of the [Likelihood Principle]' ([1975], 263).

Mayo's objection to Birnbaum's proof is more elaborate than Joshi's but rests on the same error.[25] Mayo reconstructs Birnbaum's argument as having two premises. She writes the following about the first of those premises ([2012], p. 19, notation changed for consistency):

> Suppose we have observed (E_1, x_0) [such that some (E_2, y_0) has the same likelihood function]. Then we are to view (E_1, x_0) as having resulted from getting heads on the toss of a fair coin, where tails would have meant performing E_2[...] Inference based on [a minimal sufficient statistic for E^M] is to be computed averaging over the performed and unperformed experiments E_1 and E_2. This is the *unconditional formulation* of [E^M].

Mayo's comments here are true of the Operational Sufficiency Principle, but not of Birnbaum's Sufficiency Principle. Birnbaum's Sufficiency Principle does not say anything about how inference is to be performed—in particular, it does not say that in-

[25] I consider Mayo's most recent presentation of the objection here ([2012]). See also Cox and Mayo ([2011]) and Mayo ([2009]).

ference is to be computed by averaging over E_1 and E_2. It is compatible with the possibility that inference is to be performed in that way, but it is also compatible with the possibility that inference is to be performed in a way that does not take sampling distributions into account at all, such as by a likelihoodist method or Bayesian conditioning.

Mayo then states her second premise as follows:

> Once it is known that E_1 produced the outcome x_0, the inference should be computed just as if it were known all along that E_1 was going to be performed, i.e. one should use the conditional formulation, ignoring any mixture structure.

She then claims that Birnbaum's argument is unsound because her two premises are incompatible: premise one says that one should use the unconditional formulation, while premise two says that one should use the conditional formulation. It is true that the argument Mayo has constructed is unsound, but that argument is not Birnbaum's. In reconstructing Birnbaum's argument Mayo has assumed that one must choose between a 'conditional' and an 'unconditional' formulation of the mixture experiment E^M. Frequentists need to make that choice because they use sampling distributions for inference, but likelihoodists and Bayesians do not. Thus, Mayo's response to Birnbaum's proof begs the question by presupposing a frequentist approach.

In personal communication, Mayo has responded that Birnbaum's proof is irrelevant to a sampling theorist if it requires assuming that sampling distributions are irrelevant to evidential meaning. But his proof does not require that assumption. Each of Birnbaum's premises is compatible with the claim that sampling distributions are relevant to evidential import. The fact that they are not *jointly* compatible with that assumption is not an *objection* to Birnbaum's proof—it is the whole *point* of Birnbaum's proof!

5 Conclusion

I have shown that the Likelihood Principle follows from the conjunction of the Experimental Conditionality Principle and the Weak Ancillary Realisability Principle. My proof of this result addresses responses to Birnbaum's proof that involve restricting its premises. Joshi's and Mayo's arguments for the claim that Birnbaum's proof is logically flawed would with appropriate modifications apply to the new proof given here as well, but those arguments are in error.

The case for the Likelihood Principle seems quite strong. However, the Likelihood Principle as formulated here only implies that one ought to use methods that conform to the Likelihood Principle on the assumption that one ought to use methods that track evidential meaning. This assumption is not mandatory. Frequentists claim that their methods have many virtues, including objectivity and good long-run operating characteristics, that are not characterised in terms of evidential meaning. Tracking evidential meaning is intuitively desirable, but one could maintain that it is less important than securing one or more of those putative virtues.

Greg Gandenberger
1017 Cathedral of Learning
Pittsburgh, PA 15260
greg@gandenberger.org

6 Appendix: Proof that $\mathrm{Ev}(E_2, y_0) = \mathrm{Ev}(E^B, g)$

What follows is completely analogous to the proof in the main text that $\mathrm{Ev}(E_1, x_0) = \mathrm{Ev}(E^B, g)$. Some expository comments and footnotes given there are not repeated here.

Proof. We have assumed that $P^1_\theta(x_0) = cP^2_\theta(y_0)$ for all $\theta \in \Theta$ and

some $c \geq 1$ that is constant in θ. y_0 is the outcome $y_0^\dagger$ of some unique minimal experiment $E_2^\dagger$ with sample space $\mathcal{Y}^\dagger$ that is performed with some known probability r when E_2 is performed. $E_2^\dagger$ is either E_2 itself or a proper component of E_2. y_0 just is $(E_2^\dagger, y_0^\dagger)$, so by the reflexivity of the evidential equivalence relation $\text{Ev}(E_2, y_0) = \text{Ev}(E_2, (E_2^\dagger, y_0^\dagger))$. By the Experimental Conditionality Principle, $\text{Ev}(E_2, (E_2^\dagger, y_0^\dagger)) = \text{Ev}(E_2^\dagger, y_0^\dagger)$.

Construct a hypothetical minimal experiment E_2^{CE} with sample space and sampling distribution given by table 3. The outcomes in the first row that correspond to outcomes of $E_2^\dagger$ (that is, $\{(j, w_i) : \exists(y^\dagger \in \mathcal{Y}^\dagger)(y_i = (E_2^\dagger, y^\dagger)\})$ constitute a set $A \subset \mathcal{Y}$ such that $\Pr(Y \in A) = p$ for some known $0 \leq p \leq 1$, namely rc^*c where $c^* = \frac{1}{1+c}$. Likewise, the outcomes in the first column (that is, $\{(j, w_0), (k, w_0)\}$ constitute a set $A \subset \mathcal{Y}$ such that $\Pr(Y \in A) = p$ for some known $0 \leq p \leq 1$, namely c^*.

	w_0	w_1	w_2	w_3	$\ldots$	w_n
j	$c^*cP_\theta^2(y_0)$	$c^*cP_\theta^2(y_1)$	$c^*cP_\theta^2(y_2)$	$c^*cP_\theta^2(y_3)$	$\ldots$	$c^*cP_\theta^2(y_n)$
k	$c^* - c^*cP_\theta^2(y_0)$	$c^*cP_\theta^2(y_0) - c^*\min_\theta P_\theta^2(y_0)$	$c^*\min_\theta P_\theta^2(y_0)$	0	$\ldots$	0

Table 3: Sampling Distribution of E_2^{CE} ($c^* = \frac{1}{1+c}$)

Let E_2^M be an experiment that consists of flipping a coin with bias rc^*c for heads to choose between performing $E_2^\dagger$ if the coin lands heads and performing a minimal experiment with sampling distribution given by the distribution of E_2^{CE} conditional on the complement of $\{(j, w_i) : \exists(y^\dagger \in \mathcal{Y}^\dagger)(y_i = (E_2^\dagger, y^\dagger))\}$ if the coin lands tails. By the Weak Ancillary Realisability Principle, $\text{Ev}(E_2^{CE}, (j, w_0)) = \text{Ev}(E_2^M, (E_2^\dagger, y_0^\dagger))$. By the Experimental Conditionality Principle, $\text{Ev}(E_2^M, (E_2^\dagger, y_0^\dagger)) = \text{Ev}(E_2^\dagger, y_0^\dagger)$. It follows that $\text{Ev}(E_2, y_0) = \text{Ev}(E_2^{CE}, (j, w_0))$.

Next take the hypothetical Bernoulli experiment E^B constructed in the proof given in the main text, which has sample space (g, h) and sam-

pling distribution given by $P_\theta^B(g) = P_\theta^1(x_0) = cP_\theta^2(y_0)$. Finally, construct a mixture experiment $E_2^{M_B}$ that consists of first flipping a coin with bias c^* for heads to decide between performing E^B and performing a minimal experiment with the known sampling distribution given by the distribution of E_2^{CE} conditional on the complement of the first-column outcomes $\{(j, w_0), (k, w_0)\}$. By the Weak Ancillary Realisability Principle, $\text{Ev}(E_2^{CE}, (j, w_0)) = \text{Ev}(E_2^{M_B}, (E^B, g))$. By the Experimental Conditionality Principle, $\text{Ev}(E_2^{M_B}, (E^B, g))= \text{Ev}(E^B, g)$. It follows that $\text{Ev}(E_2^{CE}, (j, w_0))= \text{Ev}(E^B, g)$. From $\text{Ev}(E_2, y_0) = \text{Ev}(E_2^{CE}, (j, w_0))$ and $\text{Ev}(E_2^{CE}, (j, w_0)) = \text{Ev}(E^B, g)$, it follows that $\text{Ev}(E_2, y_0)= \text{Ev}(E^B, g)$.

□

In the main text it was shown that $\text{Ev}(E_1, x_0)= \text{Ev}(E^B, g)$, from which it now follows that $\text{Ev}(E_1, x_0)= \text{Ev}(E_2, y_0)$.

Acknowledgements

Thanks to James Berger, Andrew Gelman, Leon Gleser, Jason Grossman, Nicole Jinn, James Joyce, Kevin Kelly, Jonathan Livengood, Edouard Machery, Deborah Mayo, Conor Mayo-Wilson, John Norton, Jonah Schupbach, Teddy Seidenfeld, Elizabeth Silver, Jan Sprenger, Paul Weirich, James Woodward, John Worrall, and two anonymous referees for helpful discussions about the topic of this paper and/or feedback on previous drafts.

References

Basu, D. 1964. "Recovery of Ancillary Information." *Sankhya: The Indian Journal of Statistics, Series A (1961-2002)* 26:3–16.

—. 1975. “Statistical Information and Likelihood [with Discussion].” *Sankhya: The Indian Journal of Statistics, Series A (1961-2002)* 37:1–71.

Berger, James. 2006. “The Case for Objective Bayesian Analysis.” *Bayesian Analysis* 1:385–402.

Berger, James and Wolpert, Robert. 1988. *The Likelihood Principle*, volume 6 of *Lecture Notes—Monograph Series*. Beachwood, OH: Institute of Mathematical Statistics, 2nd edition.

Birnbaum, Allan. 1962. “On the Foundations of Statistical Inference.” *Journal of the American Statistical Association* 57:269–306.

—. 1964. “The Anomalous Concept of Statistical Evidence: Axioms, Interpretations, and Elementary Exposition.” Technical Report IMM-NYU 332, New York University Courant Institute of Mathematical Sciences.

—. 1970a. “On Durbin’s Modified Principle of Conditionality.” *Journal of the American Statistical Association* 65:402–3.

—. 1970b. “Statistical Methods in Scientific Inference.” *Nature* 225:1033.

—. 1972. “More on Concepts of Statistical Evidence.” *Journal of the American Statistical Association* 67:858–61.

—. 1975. “Comments on Paper by J. D. Kalbfleisch.” *Biometrika* 62:262–4.

Casella, George and Berger, Roger. 2002. *Statistical Inference*. Duxbury Advanced Series in Statistics and Decision Sciences. Pacific Grove, CA: Thomson Learning, 2nd edition.

Cox, D. and Mayo, D. 2011. “Statistical Scientist Meets a Philosopher of Science: A Conversation.” *Rationality, Markets and Morals* 2:103–14.

Cox, David. 1958. "Some Problems Connected with Statistical Inference." *The Annals of Mathematical Statistics* 29:357–72.

Cox, David and Hinkley, David. 1974. *Theoretical Statistics*. London: Chapman and Hall.

Dawid, A. P. 1977. "Conformity of Inference Patterns." In J.R. Barra and the European Meeting of Statisticians (eds.), *Recent Developments in Statistics: Proceedings of the European Meeting of Statisticians, Grenoble, 6-11 Sept., 1976*, 245–67. Amsterdam: North-Holland.

Durbin, James. 1970. "On Birnbaum's Theorem on the Relation Between Sufficiency, Conditionality and Likelihood." *Journal of the American Statistical Association* 65:395–8.

Evans, Michael, Fraser, Donald, and Monette, Georges. 1986. "On Principles and Arguments to Likelihood." *Canadian Journal of Statistics* 14:181–94.

Fisher, Ronald. 1955. "Statistical Methods and Scientific Induction." *Journal of the Royal Statistical Society. Series B (Methodological)* 17:69–78.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. 2003. *Bayesian Data Analysis, Second Edition*. Chapman & Hall/CRC Texts in Statistical Science. London: Taylor & Francis.

Gelman, Andrew. 2012. "It is not necessary that Bayesian methods conform to the likelihood principle." *Statistical Modeling, Causal Inference, and Social Science* (Blog) . <andrewgelman.com/2012/10/it-not-necessary-that-bayesian-methods-conform-to-the-likelihood-principle/>. November 1 comment.

Grossman, Jason. 2011a. "The Likelihood Principle." In D.M. Gabbay, P.S. Bandyopadhyay, P. Thagard, and M.R. Forster (eds.), *Philosophy of Statistics*, Handbook of the Philosophy of Science, 553–80. Amsterdam: Elsevier.

—. 2011b. "Statistical Inference: From Data to Simple Hypotheses." Unpublished manuscript. Available at <bunny.xeny.net/linked/Grossman-statistical-inference.pdf>.

Joshi, V. M. 1990. "Fallacy in the proof of Birnbaum's Theorem." *Journal of Statistical Planning and Inference* 26:111–2.

Kalbfleisch, John D. 1975. "Sufficiency and Conditionality." *Biometrika* 62:251–9.

Kiefer, J. 1977. "Conditional Confidence Statements and Confidence Estimators." *Journal of the American Statistical Association* 72:789–808.

Mayo, Deborah. 1985. "Behavioristic, Evidentialist, and Learning Models of Statistical Testing." *Philosophy of Science* 52:493–516.

—. 1992. "Did Pearson Reject the Neyman-Pearson Philosophy of Statistics?" *Synthese* 90:233–62.

—. 1996. *Error and the Growth of Experimental Knowledge*. Science and Its Conceptual Foundations. University of Chicago Press.

—. 2009. "An Error in the Argument from Conditionality and Sufficiency to the Likelihood Principle." In Deborah Mayo and Aris Spanos (eds.), *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science*, chapter 7(III), 305–14. Cambridge University Press.

—. 2012. "On the Birnbaum Argument for the Strong Likelihood Principle." Unpublished. Available at <www.phil.vt.edu/dmayo/conference_2010/9-18-12MayoBirnbaum.pdf>.

Neyman, J. and Pearson, E. S. 1933. "On the Problem of the Most Efficient Tests of Statistical Hypotheses." *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 231:289–337.

Savage, L. J. 1962. *The Foundations of Statistical Inference: A Discussion*. Methuen's Monographs on Applied Probability and Statistics. London: Methuen.

Savage, Leonard James. 1970. "Comments on a Weakened Principle of Conditionality." *Journal of the American Statistical Association* 65:399–401.

Wasserman, Larry. 2012. "Statistical Principles?" *Normal Deviate* (Blog) . <normaldeviate.wordpress.com/2012/07/28/statistical-principles>. July 28 blog post.