A Qualitative Generalization of Birnbaum's Theorem

Conor Mayo-Wilson

Abstract

We prove a generalization of Birnbaum's theorem, which states that the sufficiency and conditionality principles together entail the likelihood principle. Birnbaum's theorem poses a dilemma for frequentists, who typically accept versions of the former two principles but reject the third. Our generalization of Birnbaum's theorem relies only on axioms for qualitative/comparative, conditional probability.

Statistical reasoning is just one among many types of reasoning, and so norms for statistical reasoning ought to be special cases of norms for reasoning more broadly. Thus, one way of assessing the plausibility of the three major approaches to statistical inference – frequentistism, likelihoodism, and Bayesianism – is to assess to what extent their underlying principles generalize to everyday reasoning. And to model everyday reasoning, it might be impractical, unfruitful, or misleading to model agents as if they (i) know the precise numerical likelihood functions associated with various experimental outcomes and then (ii) update (perhaps unconsciously) their precise numerical degrees of belief upon making such observations.

In this paper, we argue that a central argument for likelihoodist and Bayesian approaches to statistical inference generalizes naturally to evidential reasoning that involves only qualitative comparisons of the form "A given B is more likely than C given D." Specifically, we generalize Birnbaum's theorem that the sufficiency and conditionality principles – which frequentists often endorse – together entail the likelihood principle, which frequentists universally reject.

In §1, we begin with a few philosophical remarks about modeling belief using comparative, conditional probability; this section motivates the specific framework we have chosen for generalizing Birnbaum's theorem. In §2, we review the sufficiency, conditionality, and likelihood principles. We motivate the three principles, explore their relationship to Bayesian and frequentist methods, and finally, review Birnbaum's theorem and its proof. This section is extensive to make the paper self-contained, but all but §2.3 can be skipped by readers familiar with the likelihood principle and Birnbaum's theorem. In §3, we generalize the three statistical principles and prove a generalization of Birnbaum's theorem. In §4, we conclude by discussing further philosophical implications of the theorem and avenues for future research.

1 Incorporating Probabilistic Data into Non-Probabilistic Beliefs

Everyone agrees that data should sometimes change our beliefs. Because experimental data are often finite frequencies, which obey the probability axioms, it is only natural that many philosophers and statisticians have have argued (via Dutch Books, accuracy-dominance considerations, etc.) for **probabilism**, i.e., the thesis that our degrees of belief should obey the probability axioms. Equipped with probabilism, one can then easily explain how the probabilistic structure of data should incorporated into a rational agent's beliefs, as the latter also have probabilistic structure.

Yet everyone recognizes that, at best, probabilism is a norm for an *ideal* agent. It is well-known that if our degrees of belief should obey the probability axioms, then there's a sense in which we should also be logically omniscient [Savage, 1972, p. 7, especially footnote †]. And it seems unreasonable to require logical omniscience from fallible reasoners with limited computational resources. Moreover, we risk introducing modeling artifacts when we represent degrees of belief using precise probabilities, when in fact many arguments for probabilism (e.g., representation theorems) begin with primitive *comparative* judgments of likelihood.

So here is the dilemma. On one hand, it seems far too demanding to require that degrees of belief obey the probability axioms. On the other hand, it is difficult to specify how probabilistic data ought to influence our beliefs unless we model belief itself using probability theory.

The framework for qualitative/comparative, conditional probability below provides one way to address that dilemma. We axiomatize expressions of the form "$A|B \succeq C|D$", which can be roughly understood as the representing the claim "A given B is more likely than C given D." The axioms are *necessary* for the binary relation $\succeq$ to be representable using (conditional) probabilities, and so any data obeying the probability axioms will satisfy our qualitative axioms. However, our axioms are not sufficient for a probabilistic representation, and so an agent's beliefs might satisfy our axioms without being probabilistically "coherent."

Perhaps most importantly, our framework addresses some of the motivations for thinking rational degrees of belief need not satisfy the probability axioms. Instead of requiring $P(\varphi) = 1$ for any tautology φ, our axioms more-or-less require a qualitative analog of the claim that $P(\varphi|\varphi) = 1$ for any formula φ. Famously, there is no algorithm for determining whether a formula of predicate logic is a tautology (and so there is no program for deciding whether the standard axiom entails that $P(\varphi)$ ought to equal one), but it is computationally trivial to check whether two formula tokens are syntactically identical (and hence, it is trivial to determine if our axiom requires that $P(\varphi|\varphi) = 1$). Further, because our axioms concern a binary relation, we avoid worries about introducing modeling artifacts by representing comparative judgments using precise, numerical probabilities.

To be clear, the likelihood principle – and the related "law" of likelihood –

say nothing about rational belief. Yet likelihoodists often motivate the law of likelihood and likelihood principle by arguing the two theses dovetail with the Bayesian theory of rational belief [Edwards, 1984, p. 28], [Royall, 1997, p. 87]. It is only natural, therefore, to ask whether there are qualitative/comparative likelihoodist principles that dovetail nicely with a generalization of Bayesianism, in which rational degrees of belief are not required to satisfy the probability axioms. The generalization of Birnbaum's proof (of the likelihood principle) below suggests an affirmative answer to that question.

2 Frequentism vs. Bayesianism

In the foundations of mathematics, rival philosophies (e.g., intuitionist vs. classical) can be characterized by which axioms and rules of inference they accept and which they reject. In the foundations of statistics, things are messier. Although likelihoodism and Bayesianism each might be identified with axiomatic theories, frequentism cannot.[1] Instead, frequentist methods – including Neyman-Pearsonian tests, confidence intervals, and particular approaches to point estimation – are typically motivated by a hodgepodge of theses about minimizing error, maximizing "coverage" probabilities, and avoiding inadmissibility.

Despite lacking an axiomatic basis, frequentists often endorse variants of the sufficiency and conditionality principles discussed below. Like the likelihood principle, which frequentists reject, sufficiency and conditionality are theses about *evidential equivalence.* Here, "evidential equivalence" is intended to denote an explication of an informal, normative concept. Intuitively, two samples and might convey the exact same information about an unknown quantity. In such cases, statisticians would call the two samples "evidentially equivalent", and the three statistical principles discussed below aim to provide precise conditions for evidential equivalence.

To state the three principles, we must introduce some notation. Let Θ represent the **simple hypotheses** under investigation. Each experiment $\mathbb{E}$ has a set of possible outcomes $\Omega_{\mathbb{E}}$, which represents the data one might obtain. Statisti-

[1] Calling likelihoodism and Bayesianism "axiomatic" theories for the "foundations of statistics" might be misleading for two reasons. First, it's not clear the two schools' axioms concern the same subjects. Likelihoodists (e.g. Royall [1997]) argue that their "axioms" – the law of likeliihood and likelihood principle – answer the question, "Which hypotheses does the evidence support?", not "What should one believe in light of the evidence?" or "What should one do?" Arguably, Bayesianism is first and foremost a theory answering the latter two questions. Comparing Bayesianism and likelihoodism, therefore, might seem like comparing group theory and plane geometry. Second, contrary to what some Bayesians claim [Bernardo, 2011], Bayesianism is rarely presented as an axiomatic theory like most mathematical subjects, and self-identified Bayesians might disagree about the axioms. One reason for the disagreement is that it's not clear what is being axiomatized. In geometry, one axiomatizes the relationship between points and lines; what exactly is axiomatized by Bayesianism? The obvious choices are rational belief and action, but should the axioms for belief concern pairwise confidence judgments (i.e., Is A or B more likely?) or functions from a set of propositions to the real number line (e.g., what properties should a rational belief function P from a set of propositions to $\mathbb{R}$ have?). Are preferences manifested in binary choices, or better modeled by choice functions? And so on.

cians assume that, for each experiment $\mathbb{E}$ and each hypothesis $\theta \in \Theta$, there is a probability measure $P_\theta^{\mathbb{E}}$ over $\Omega_{\mathbb{E}}$. To avoid unnecessary measure-theoretic complications, we assume that the set of hypotheses Θ and the set of experimental outcomes $\Omega_{\mathbb{E}}$ are finite. We also assume all probability measures are defined on the appropriate power set algebra.

Example 1: Imagine you want to test the success rate of a new drug in treating a serious cancer; you are interested in how many people survive three months after treatment. One experiment $\mathbb{E}_1$ consists in treating 100 patients. The set of hypotheses Θ is best represented by some subset of the unit interval $[0, 1]$, where $\theta \in \Theta$ represents the hypothesis that, any given patient's probability of surviving three months is precisely θ. The set of possible outcomes $\Omega_{\mathbb{E}_1}$ contains all binary sequences of length 100, representing which of the 100 patients survive three months. For each θ, the probability measure $P_\theta^{\mathbb{E}_1}$ specifies how likely various outcomes are, if the drug's success rate equals θ.

□

Example 2: Imagine you are interested in the same drug as in the previous example, but you are worried that it might be less effective than the conventional treatment, which has a 94% survival rate. So you design an experiment $\mathbb{E}_2$ in which you treat a new patient every three months until two patients die in total. The set of hypotheses Θ is as before, but the set $\Omega_{\mathbb{E}_2}$ of outcomes differs. Here, $\Omega_{\mathbb{E}_2}$ contains (i) all finite binary sequences containing exactly two zeroes (representing the two deaths), where the last digit is zero (because the trial stops with the fourth death) and (ii) all infinite binary sequences containing one or fewer zeroes.

□

Informally, evidential equivalence is a relation between two outcomes of (i) the same experiment or (ii) two different experiments. For instance, consider the above experiment $\mathbb{E}_1$. One might think that any two outcomes of $\mathbb{E}_1$ containing exactly the same number of deaths are evidentially equivalent (assuming one knows nothing else about the patients); that turns out to be a consequence of the sufficiency principle discussed below. More controversially, let ω_1 be a sequence containing two deaths in experiment $\mathbb{E}_1$, and let $\omega_2 = \omega_1$ be *exactly the same* sequence of 100 outcomes obtained in experiment $\mathbb{E}_2$ (so exactly 100 patients are treated). One might claim that, in virtue of being exactly the same sequence of outcomes, ω_1 and ω_2 ought to demand the same inferences, even though in experiment $\mathbb{E}_1$ the number of patients was fixed in advance whereas in $\mathbb{E}_2$, it was not. That claim is a consequence of the likelihood principle.

In the ensuing sections, we summarize three common principles about evidential equivalence. For now, we flag two important features of the discussion. First, evidential equivalence is a *relation between experimental outcomes*. Some frequentists argue that statisticians can only characterize the reliability of *repeatable procedures*, like experiments, estimators, or tests. According to such

frequentists, statisticians ought to remain silent about the "evidence" or "information" contained in a particular experimental outcome (i.e., a particular data set or sample).[2] Such frequentists will reject all of the principles below. We emphasize, however, that we believe most frequentists do not take that extreme position.

Second, evidential equivalence is really a three-place predicate that relates two experimental outcomes *and* a set of hypotheses Θ. That means that two outcomes might be evidentially equivalent with respect to one set of hypotheses Θ but not with respect to another Θ'. As a trivial case, two distinct observations (e.g., heads and tails of a coin flip) might be equivalent in one sense because they tell one nothing about the hypotheses under investigation (e.g., the success rate of a particular treatment), but if the hypotheses are identical with the experimental outcomes, then clearly the two are not evidentially equivalent. Nonetheless, we will omit discussing the dependence of the evidential equivalence relation on Θ, assuming the parameter space is clear from context.

2.1 Conditionality and Sufficiency

Conditionality

Suppose a researcher is unsure which of the two experiments above (i.e., $\mathbb{E}_1$ or $\mathbb{E}_2$) to conduct. She decides to flip a coin to pick an experiment, and on the basis of flip, she chooses to conduct $\mathbb{E}_1$. Using a randomizing device like a coin or die to select from several possible experiments is called performing a **mixed experiment.** In the case at hand, because a coin flip tells one nothing about the efficacy of cancer drugs, intuitively, any sample the researcher acquires in her mixed experiment seems evidentially equivalent to the same sample she would have obtained had she just decided, without flipping a coin, to conduct $\mathbb{E}_1$. That is what the conditionality principle says.

Formally, given some probability $r \in (0, 1)$ representing the known bias of a coin and two experiments $\mathbb{E}$ and $\mathbb{F}$, let $\mathbb{M}(r, \mathbb{E}, \mathbb{F})$ represent the mixed experiment in which a coin of bias $1 - r$ is flipped, $\mathbb{E}$ is conducted if the coin lands tails, and $\mathbb{F}$ is conducted otherwise. Clearly, $\mathbb{M}(r, \mathbb{E}, \mathbb{F})$ is representable by a pair $\langle \Omega_{\mathbb{M}(r,\mathbb{E},\mathbb{F})}, \{P_\theta^{\mathbb{M}(r,\mathbb{E},\mathbb{F})}\}_{\theta\in\Theta}\rangle$ with the following properties:

- The outcome space $\Omega_{\mathbb{M}(r,\mathbb{E},\mathbb{F})}$ equals $(\{0\} \times \Omega_{\mathbb{E}}) \cup (\{1\} \times \Omega_{\mathbb{F}})$. For any $\omega \in \Omega_{\mathbb{E}}$, the pair $\langle 0, \omega\rangle \in \Omega_{\mathbb{M}(r,\mathbb{E},\mathbb{F})}$ represents the outcome in which the coin lands tails, one conducts $\mathbb{E}$, and one obtains the outcome ω.
- $P_\theta^{\mathbb{M}(r,\mathbb{E},\mathbb{F})}(0, \omega) = r \cdot P_\theta^{\mathbb{E}}(\omega)$ for every $\theta \in \Theta$ and any $\omega \in \Omega_{\mathbb{E}}$. Similarly, $P_\theta^{\mathbb{M}(r,\mathbb{E},\mathbb{F})}(1, \omega) = (1 - r) \cdot P_\theta^{\mathbb{F}}(\omega)$ for every $\omega \in \Omega_{\mathbb{F}}$.

It is assumed that one knows the value of r, and so the coin flip is uninformative with respect to Θ.

[2] Arguably, [Neyman and Pearson, 1933] endorse this position. See [Mayo, 2018] for an alternative interpretation of the famous passage in Neyman and Pearson's landmark paper.

Conditionality Principle (C): Let $\mathbb{E}$ and $\mathbb{F}$ be any two experiments, and $\mathbb{M}$ a mixture of the two. Then the outcome $\langle 0, \omega \rangle$ of $\mathbb{M}$ is evidentially equivalent to the outcome ω of $\mathbb{E}$.

Sufficiency

Informally, a statistic is a way of summarizing a data set. For example, instead of reporting the exact sequence of fifty coin tosses to you, we might report the number of heads. Just as you might learn all the important events in a movie from a comprehensive preview, so some statistics summarize all the important information in a data set. For both Bayesians and many frequentists, a sufficient statistic is such a comprehensive summary. Thus, the sufficiency principle says that learning the value of a sufficient statistic is evidentially equivalent to seeing the entire sample.

Formally, a *statistic* is a function $T : \Omega_{\mathbb{E}} \to \mathcal{R}$ of the outcomes of some experiment $\mathbb{E}$, where elements of the range $\mathcal{R}$ are typically real numbers or vector of numbers. In some cases, a statistic compresses the data substantially. For instance, if $\Omega_{\mathbb{E}}$ is the set of all binary sequences of length 50 (representing 50 coin tosses), then one succinctly summarizes the data using the statistic T that maps every sequence ω to the number of heads it contains $T(\omega) \in \mathcal{R} = \{0, 1, 2, \ldots 50\}$. But a degenerate example of a statistic is the identity map, which maps every sample ω to itself $T(\omega) = \omega \in \mathcal{R} = \Omega_{\mathbb{E}}$.

Given an experiment $\mathbb{E}$, a statistic is called *sufficient* if, $P_\theta(\omega | T = t) = P_\upsilon(\omega | T = t)$ for any value of the statistic $t \in \mathcal{R}$, any data set $\omega \in \Omega_{\mathbb{E}}$ and for all $\theta, \upsilon \in \Theta$. Informally, the value of a sufficient statistic "screens off" the parameter from any data set.

Example 1 (continued): The number of successes/survivals in $\mathbb{E}_1$ is a sufficient statistic. See, [Casella and Berger, 2002, p. 274] for a proof.

Example 2 (continued): The number of survivals (which determines the number of treated patients) is again a sufficient statistic for $\mathbb{E}_2$. Why? Let t_ω represent the number of survivals in a sample ω. If $t \neq t_\omega$, then clearly $P_\theta(\omega | T = t) = 0$ for all θ. And if $t = t_\omega$, then:

$$\begin{aligned}
P_\theta(\omega | T = t_\omega) &= \frac{P_\theta(\{\omega\} \cap \{T = t_\omega\})}{P_\theta(T = t_\omega)} \text{ by definition of conditional probability} \\
&= \frac{P_\theta(\omega)}{P_\theta(t_\omega)} \text{ because } \omega \in \{T = t_\omega\} := \{\omega' \in \Omega_{\mathbb{E}_2} : t_{\omega'} = t\} \text{ since } t = t_\omega \\
&= \frac{\theta^2(1-\theta)^{t_\omega}}{\binom{t_\omega + 1}{1} \cdot \theta^2(1-\theta)^{t_\omega}} \\
&= \frac{1}{t_\omega + 1}
\end{aligned}$$

Since the final line does not depend upon θ, it follows that $P_\theta(\omega | T = t_\omega) = \frac{1}{t_\omega + 1} = P_\upsilon(\omega | T = t_\omega)$ for all $\theta, \upsilon \in \Theta$.

□

It turns out that, no matter the experiment, there are always at least two sufficient statistics. First, note the identity map $T(\omega) = \omega$ is sufficient because $P_\theta(\omega|T = t) = 1$ if $t = \omega$ and is zero otherwise, no mater the parameter. Second, for any experiment and any outcome $\omega \in \Omega_{\mathbb{E}}$, the function $L_\omega : \theta \mapsto P_\theta^{\mathbb{E}}(\omega)$ is called the **likelihood function**. The next lemma, which is crucial for Birnbaum's proof, shows the likelihood function is a sufficient statistic.[3]

Lemma 1 *In every experiment, the statistic $T : \omega \mapsto L_\omega$ is sufficient.*

Proof: Like Example 2, let L_ω be the likelihood function of ω. If $t \neq L_\omega$, then clearly $P_\theta(\omega|T = t) = 0$ for all θ. And if $t = L_\omega$, then

$$\begin{aligned}
P_\theta(\omega|T = t = L_\omega) &= \frac{P_\theta(\{\omega\} \cap \{T = L_\omega\})}{P_\theta(T = L_\omega)} \text{ by definition of conditional probability} \\
&= \frac{P_\theta(\omega)}{P_\theta(T = L_\omega)} \text{ because } \omega \in \{T = L_\omega\} := \{\omega' \in \Omega_{\mathbb{E}} : L_{\omega'} = L_\omega\} \\
&= \frac{P_\theta(\omega)}{\sum_{\omega' \in \{T = L_\omega\}} P_\theta(\omega')} \\
&= \frac{P_\theta(\omega)}{P_\theta(\omega) \cdot |\{T = L_\omega\}|} \text{ because } P_\theta(\omega') = P_\theta(\omega) \text{ if } L_\omega = L_{\omega'} \\
&= \frac{1}{|\{T = L_\omega\}|}
\end{aligned}$$

Again, the final line does not depend on θ, and so we're done.

□

With this background, we can state the sufficiency principle. Given any experiment $\mathbb{E} = \langle \Omega_{\mathbb{E}}, \{P_\theta^{\mathbb{E}}\}_{\theta \in \Theta} \rangle$ and any sufficient statistic $T : \Omega_{\mathbb{E}} \to \mathcal{R}$, let $\mathbb{E}^T$ denote the experiment in which one performs $\mathbb{E}$ but learns only the value of T. Clearly, $\mathbb{E}^T$ is representable by a pair $\langle \Omega^{\mathbb{E}^T}, \{P_\theta^{\mathbb{E}^T}\}_{\theta \in \Theta} \rangle$ such that:

- The outcome space $\Omega^{\mathbb{E}^T}$ equals $T(\Omega_{\mathbb{E}}) = \{T(\omega) : \omega \in \Omega_{\mathbb{E}}\} \subseteq \mathcal{R}$. Informally, instead of seeing the outcome ω of $\mathbb{E}$, one sees the value $T(\omega)$ of the sufficient statistic.

[3] Technical aside: The lemma above differs from the standard lemma used in Birnbaum's theorem because, as I have defined matters, the statistic $T : \omega \mapsto L_\omega$ is sufficient but not *minimal*. According to my definition, $T(\omega) = T(\omega')$ only if the likelihood functions of ω and ω' are *precisely equal*, i.e., it is not enough for them to be related by a positive multiplicative constant. Typically, in proving Birnbaum's theorem, one defines T^* to be the statistic such that $T(\omega) = T(\omega')$ if there is some $c > 0$ such that $P_\theta^{\mathbb{E}}(\omega) = c \cdot P_\theta^{\mathbb{E}}(\omega')$ for all θ, and then one proves T^* is a minimal sufficient statistic in a way analogous to above. Some frequentist statisticians do think that minimality is a crucial epistemic notion, but I lack the space to enter that debate. For my purposes, I have chosen to use T in my reconstruction of Birnbaum's proof as the strategy generalizes naturally to the qualitative setting; there is no direct qualitative analog of T^*.

- For each $\theta \in \Theta$ and each $t \in \Omega^{\mathbb{E}^T}$, one has $P_\theta^{\mathbb{E}^T}(t) = P_\theta^{\mathbb{E}}(T^{-1}(t)) := P_\theta^{\mathbb{E}}(\{\omega \in \Omega_{\mathbb{E}} : T(\omega) = t)$.

Sufficiency Principle (S): Let $\mathbb{E}$ be any experiment and T a sufficient statistic. Then $T(\omega)$ in $\mathbb{E}^T$ is evidentially equivalent to the outcome ω of $\mathbb{E}$.

An immediate corollary of the sufficiency principle (assuming evidential equivalence is transitive) is the following:

Weak Sufficiency Principle: Let $\mathbb{E}$ be any experiment. If T is a sufficient statistic and $\omega, \omega \in \Omega_{\mathbb{E}}$ are two outcomes of $\mathbb{E}$ such that $T(\omega) = T(\omega')$, then ω and ω' are evidentially equivalent.

2.2 The Likelihood Principle and its consequences

The most controversial statistical principle is the likelihood principle.

Likelihood Principle (LP): Let $\mathbb{E}$ and $\mathbb{F}$ be any two experiments, and let $E \subseteq \Omega_{\mathbb{E}}$ and $F \subseteq \Omega_{\mathbb{F}}$. If there is some $c > 0$ such that $P_\theta^{\mathbb{E}}(E) = c \cdot P_\theta^{\mathbb{F}}(F)$ for all $\theta \in \Theta$, then E and F are evidentially equivalent.

Frequentists reject LP for at least three related reasons: (1) LP has a number of consequences (e.g., the stopping rule principle) that some find unintuitive, (2) LP seems to contraindicate the use of p-values, confidence intervals, and other classical methods, and (3) LP is intricately connected with Bayesian methodology. We discuss those three issues in order.

Controversial Consequences: Irrelevance of Stopping Rules

Perhaps the most controversial consequence of LP is the stopping rule principle, which asserts that, under certain precisely circumscribed conditions, statisticians may safely ignore the reason an experiment was halted. Rather than precisely defining which types of stopping rules can be ignored according to that principle, we give a simple example.

Examples 1 and 2 continued: Experiment $\mathbb{E}_1$ stops, no matter what, when 100 patients are treated; experiment $\mathbb{E}_2$ is stops when two patients have died. So the rules for terminating the experiments differ. Nonetheless, certain pairs of outcomes from the two experiments are, according to LP, evidentially equivalent.

For example, suppose $\mathbb{E}_1$ is conducted and two patients die. Let ω be the binary sequence representing which patients died, and suppose the second death is the 100th treated patient. Next, suppose the same data sequence ω is obtained in experiment $\mathbb{E}_2$. Then in $\mathbb{E}_1$, we have $P_\theta^{\mathbb{E}_1}(\omega) = P_\theta^{\mathbb{E}_2}(\omega) = \theta^2(1-\theta)^{98}$. Thus, by LP (here, the constant c is 1), the outcome ω in $\mathbb{E}_1$ is equivalent to the outcome ω in $\mathbb{E}_2$.

In the literature on stopping rules, it is typically shown that the same argument works if one learns only the values of a sufficient statistic in each of the

two experiments. That is, suppose $\mathbb{E}_1$ is conducted and one learns only E that two patients die; formally, E is the set of *all* binary sequences of length 100 containing two zeroes (for each death). Similarly, suppose $\mathbb{E}_2$ is conducted and one learns only F that 100 patients are treated; formally, F is the set of binary sequences of length 100 containing two zeroes, one of which occurs in the 100th spot. Then $P_\theta^{\mathbb{E}_1}(E) = \binom{100}{2}\theta^2(1-\theta)^{98}$ and $P_\theta^{\mathbb{E}_2}(F) = \binom{99}{1}\theta^2(1-\theta)^{98}$, and so letting $c = \frac{99}{\binom{100}{2}}$, it follows that $P_\theta^{\mathbb{E}_1}(E) = c \cdot P_\theta^{\mathbb{E}_2}(F)$ for all $\theta \in \Theta$. Thus, by LP, the outcomes E and F are evidentially equivalently.

□

Proponents of LP think the stopping rule principle is a victory for their methodology: if an experiment is stopped early because of lack of funding or because the principal investigator passes away, a proponent of LP need not modify the way she analyzes the data. Frequentists, in contrast, argue this convenience comes at too great an epistemic cost. Namely, frequentists argue that *modal* properties of the two experiments differ, and those modal properties are epistemically important. That, even if the *actual* outcomes of $\mathbb{E}_1$ and $\mathbb{E}_2$ are identical, in experiment $\mathbb{E}_1$, more than two patients *could have* died, even if 100 must be treated. Similarly, in $\mathbb{E}_2$, more than 100 patients could have been treated, even if two deaths are more-or-less guaranteed. For this reason, some statisticians say that frequentist inferences, unlike Bayesian ones, *depend upon the sample space* [Lindley, 2006].

Incompatibility with Classical Techniques

The differences between the modal properties of the two properties of the two experiments is critical for certain classical methods. Here, we provide a typical example that illustrates how the use of p-values in hypothesis testing can render different verdicts in situations that are regarded as evidentially equivalent by LP. Similar conflicts between LP and other frequentist devices (e.g., confidence intervals) can be found in the opening chapters of [Berger and Wolpert, 1988].

Example 1 continued: In experiment $\mathbb{E}_1$, imagine the new drug is compared with a conventional treatment that has a survival rate of 94%. Again, imagine that two patients die in $\mathbb{E}_1$, and so the experimenter wants to know if there is a statistically significant difference between the efficacy of the new treatment (which produced a 98% survival rate in the sample) and that of the conventional treatment. To do so, the experimenter decides to reject the null hypothesis – that the new treatment is no better than the conventional one – if, under the assumption the null hypothesis is true, the probability of two or fewer deaths in a sample of 100 patients is below .05. However, given $\theta = .94$, it is easy to see that, if T is the number of deaths in the sample:

$$P_\theta^{\mathbb{E}_1}(T \leq 2) = \sum_{n=0}^{2} \binom{100}{n} .94^{100-n} \cdot (.06)^n \approx .0566$$

which is *not* significant at the .05 level.

□

Example 2 continued: Suppose, in $\mathbb{E}_2$, 100 patients are treated before the second death, and as above, imagine the new drug is compared with the conventional treatment. Again, the experimenter wants to know if there is a statistically significant difference between the new and conventional treatment. To do so, the experimenter decides to reject the null hypothesis if, under its supposition, there is less than a 5% chance of needing to treat at least 100 patients. Given $\theta = .94$, it is easy to see that, if T is the number of patients treated in $\mathbb{E}_2$

$$P_\theta^{\mathbb{E}_2}(T \geq 100) = \sum_{n=100}^{\infty} \binom{n}{1} .94^n \cdot (.06)^2 \approx .014$$

which *is* significant at the .05 level.

□

2.3 Bayesianism and the three statistical principles

Bayesians endorse C, S, and LP: if two experimental outcomes are entailed to be evidentially equivalent by atleast one of the principles, then *every* Bayesian will treat the two outcomes identically when updating his or her beliefs.

In greater detail, Bayesianism is often defined to be the conjunction of two theses. First, for any experiment $\mathbb{E}$, one's degrees of belief $Q^{\mathbb{E}}$ (over the space $\Theta \times \Omega_{\mathbb{E}}$ consisting both of hypotheses and experimental outcomes) should satisfy the probability axioms. Given $Q^{\mathbb{E}}$, one can define a measure $\pi_Q^{\mathbb{E}}$ on Θ by $\pi_Q^{\mathbb{E}}(\Theta_0) = Q^{\mathbb{E}}(\Theta_0 \times \Omega_{\mathbb{E}})$ where $\Theta_0 \subseteq \Theta$; the function $\pi_Q^{\mathbb{E}}$ is called the experimenter's *prior*.

Second, for any outcome $\omega \in \Omega_{\mathbb{E}}$, one's degrees of belief upon learning ω should be updated by conditionalization, i.e, one's degree of beliefs over Θ ought to be represented by the *posterior* distribution:

$$Q^{\mathbb{E}}(\theta|\omega) = \frac{Q^{\mathbb{E}}(\theta) \cdot P_\theta(\omega)}{\sum_{\upsilon \in \Theta} P_\upsilon(\omega) \cdot Q^{\mathbb{E}}(\upsilon)}$$

In philosophy and decision theory, Bayesians typically also tacitly assume that one's degrees of belief ought not vary with one's choice of experiment, i.e., that $\pi_Q^{\mathbb{E}} = \pi_Q^{\mathbb{F}}$ for any two experiments $\mathbb{E}$ and $\mathbb{F}$. This seems like a reasonable assumption if the measures in $\{Q^{\mathbb{E}}\}_{\mathbb{E} \in \mathcal{E}}$ (where $\mathcal{E}$ represents all experiments) do in fact represent some particular experimenter's degrees of belief. For that reason, we adopt the assumption for the remainder of the paper and we write π_Q for the agent's prior.[4]

Define two experimental outcomes $\langle \mathbb{E}, \omega_{\mathbb{E}} \rangle$ and $\langle \mathbb{F}, \omega_{\mathbb{F}} \rangle$ to be **Bayesian posterior equivalent** if $Q^{\mathbb{E}}(\theta|\omega_{\mathbb{E}}) = Q^{\mathbb{F}}(\theta|\omega_{\mathbb{F}})$ for all priors π_Q and all $\theta \in \Theta$. In

[4]In practice, many Bayesian statisticians use particular priors for computational ease (e.g., by fitting a conjugate prior to one's beliefs) and still others (typically, "objective" Bayesians) think certain priors are rationally required in a way that might vary with the experiment.

other words, two outcomes are Bayesian posterior equivalent if, no matter one's prior, one would update one's degrees of belief (in the hypotheses) in identical ways no matter which observation was made. It's well-known that:

Proposition 1 *If* C, S, *or* LP *entails two experimental outcomes are evidentially equivalent, then the two outcomes are Bayesian posterior equivalent.*

For brevity, we omit the proof. Further, Mayo-Wilson and Saraf [2020] prove that LP completely characterizes Bayesian posterior equivalence.

Proposition 2 *If two experimental outcomes are Bayesian posterior equivalent, then* LP *entails the two are evidentially equivalent.*[5]

Together, the two propositions may explain yet another reason frequentists reject LP. The two propositions show not only do Bayesians endorse LP, but if one endorses LP, then one must agree with Bayesians about which outcomes are evidentially equivalent. And doing that seemingly requires rejecting many classical methods.

2.4 Birnbaum's Theorem

Birnbaum's theorem is easy to state and prove: C and S entail LP. See below. Since its publication, frequentists have argued that the Birnbaum's formalizations of the conditionality and sufficiency principles are improper, i.e., that C and S do not capture the intended, informal methodological precepts.[6] That allows frequentists to endorse other versions of the conditionality and sufficiency principles but continue to reject LP. We will say a bit more about these maneuvers, but for now, it suffices to say that a primary motivation of our qualitative generalization of Birnbaum's theorem is that it shows Birnbaum's result is robust under different formalizations of sufficiency and conditionality.[7]

Theorem 1 (Birnbaum [1962]) C *and* S *entail* LP.

Proof: Let $\omega_{\mathbb{E}} \in \Omega_{\mathbb{E}}$ and $\omega_{\mathbb{F}} \in \Omega_{\mathbb{F}}$ be outcomes of experiments $\mathbb{E}$ and $\mathbb{F}$ respectively. Suppose there is some $c > 0$ such that $P_\theta^{\mathbb{E}}(\omega_{\mathbb{E}}) = c \cdot P_\theta^{\mathbb{F}}(\omega_{\mathbb{F}})$ for all $\theta \in \Theta$. We must show that $\omega_{\mathbb{E}}$ and $\omega_{\mathbb{F}}$ are evidentially equivalent using C and S.

[5] This theorem is valid only under the assumption that the two outcomes are drawn from experiments with the same parameter space. In cases in which there are "non-informative nuisance parameters" in one experiment but not the other, there might be pairs of outcomes that are Bayesian posterior equivalent but which are unrelated via LP, as LP concerns only outcomes drawn from experiments with the same parameter space. See [Berger and Wolpert, 1988, p. 41.5] for a generalization of LP that, we conjecture, addresses this problem.

[6] For instance, both Durbin [1970] and Kalbfleisch [1975] argues that Birnbaum's version of C is too strong. Some authors, like [Evans et al., 1986], do not pinpoint a problem with one principle but rather argue that the way Birnbaum combines them is problematic. [Mayo, 2014] alleges Birnbaum's proof is "invalid", but this claim, we believe, is false, as Mayo's analysis rests on a mathematical framework that differs from Birnbaum's.

[7] Gandenberger [2014] likewise shows Birnbaum's proof is robust under weaker versions of S and C. What distinguishes the present paper from [Gandenberger, 2014]'s is that we work within a qualitative framework.

To do so, define $\mathbb{M}$ to be the mixed experiment in which a coin of bias $r = \frac{c}{1+c}$ is flipped, and then $\mathbb{F}$ is conducted if heads is obtained and $\mathbb{E}$ otherwise. Notice that $r \in (0,1)$ because $c > 0$.

By the conditionality principle C, it follows that the outcome $\omega_{\mathbb{E}}$ of $\mathbb{E}$ is evidentially equivalent to the outcome $\langle 0, \omega_{\mathbb{E}} \rangle$ of $\mathbb{M}$. Similarly, $\omega_{\mathbb{F}}$ of $\mathbb{F}$ is evidentially equivalent to the outcome $\langle 1, \omega_{\mathbb{F}} \rangle$ of $\mathbb{M}$. Thus, it suffices to show that the outcomes $\langle 0, \omega_{\mathbb{E}} \rangle$ and $\langle 1, \omega_{\mathbb{F}} \rangle$ of $\mathbb{M}$ are evidentially equivalent.

To do so, note that in the mixed experiment $\mathbb{M}$, the observations $\langle 0, \omega_{\mathbb{E}} \rangle$ and $\langle 1, \omega_{\mathbb{F}} \rangle$ have the same likelihood function because for all θ:

$$P_\theta^{\mathbb{M}}(\langle 1, F \rangle) = \frac{c}{1+c} \cdot P_\theta^{\mathbb{F}}(F) = \frac{c}{1+c} \cdot \left(\frac{1}{c} P_\theta^{\mathbb{E}}(E) \right) = \left(1 - \frac{c}{1+c} \right) \cdot P_\theta^{\mathbb{E}}(E) = P_\theta^{\mathbb{M}}(\langle 0, E \rangle)$$

By lemma 1, the likelihood function is a sufficient statistic. Hence, because $\langle 0, \omega_{\mathbb{E}} \rangle$ and $\langle 1, \omega_{\mathbb{F}} \rangle$ have the same likelihood function, S entails that $\langle 0, \omega_{\mathbb{E}} \rangle$ and $\langle 1, \omega_{\mathbb{F}} \rangle$ are evidentially equivalent, as desired.

□

Evans et al. [1986] allege that LP follows from C alone, but their proof relies on a considerably stronger version of C. Specifically, Evans et al. [1986]'s statement of C, unlike Birnbaum's, entails that ancillaries statistics (whether they are the value of some "mixing" device) can often be ignored in the analysis of an experiment. Because the technical details are subtle and irrelevant to the remainder of the paper, we omit them.

3 Three Qualitative Evidential Principles

3.1 Axioms for Qualitative Conditional Probability

Given a set of hypotheses Θ, every **qualitative experiment** $\mathbb{E}$ will be representable by a pair $\langle \Omega_{\mathbb{E}}, \preceq_{\mathbb{E}} \rangle$ such that $\preceq_{\mathbb{E}}$ is binary relation satisfying the axioms below.[8] The relation $\preceq_{\mathbb{E}}$ is the qualitative analog of the set of likelihood functions in an experiment. The idea is that $A|\theta \preceq_{\mathbb{E}} B|\eta$ represents the claim that "experimental outcome B is at least as likely under supposition η as outcome A is under supposition θ." Notice that here, as in the remainder of the document, we will typically use commas and write $A|\theta$ instead of $A|\{\theta\}$ when a singleton $\{\theta\}$ appears to the right of the conditioning bar. Define $A|\theta \approx_{\mathbb{E}} B|\theta$ to hold if $A|\theta \preceq_{\mathbb{E}} B|\theta$ and vice versa.

Likelihoodists, like frequentists, do not think of $P_\theta^{\mathbb{E}}(E)$ as a conditional probability, but nonetheless, all statisticians agree that, in quantitative setting, the measure $P_\theta^{\mathbb{E}}$ allows one to compare particular types of conditional probabilities. For instance, if E, F, and G respectively represent the events that at

[8] For reasons discussed below, a qualitative experiment ought not be *defined* to be a pair, just as in the quantitative setting an experiment is *not* to be defined to be a pair $\langle \Omega, \{P_\theta\}_{\theta \in \Theta} \rangle$: experiments are procedures that can be conducted, and not every mathematical possibility need be physically realizable.

least six, seven, and eight heads are observed in a sequence of ten tosses, then $P_\theta(E|F) = 1 > P_\eta(G|F)$ for any $\theta, \eta \in (0,1)$. Although we list parameters to the right of a conditioning bar below, note the qualitative proof of Birnbaum's theorem never treats $\{\theta\}$ as an event in a way that frequentists or likelihoodists would criticize. But just like in the quantitative setting, we define the relation $\preceq_{\mathbb{E}}$ so that one can make comparisons of the form $E|\theta, F \preceq_{\mathbb{E}} G|\eta, F$, where $E, F, G \subseteq \Omega_{\mathbb{E}}$ are observable events. Notice here we write $A|E, \theta$ instead of $A|E \cap \{\theta\}$.

For all A, B, C for which the following events are well-defined, we assume $\preceq$ satisfies the following properties.

> Axiom 1: $\preceq_{\mathbb{E}}$ is total, transitive, and reflexive.
>
> Axiom 3: $A|A \approx_{\mathbb{E}} B|B$.
>
> Axiom 4: $A \cap B|B \approx_{\mathbb{E}} A|B$,
>
> Axiom 5: Suppose $A \cap B = A' \cap B' = \emptyset$. If $A|C \preceq_{\mathbb{E}} A'|C'$ and $B|C \preceq_{\mathbb{E}} B'|C'$, then $A \cup B|C \preceq_{\mathbb{E}} A' \cup B'|C'$; moreover, if either hypothesis is $\prec_{\mathbb{E}}$, then the conclusion is $\prec_{\mathbb{E}}$.
>
> Axiom 6: Suppose $C \subseteq B \subseteq A$ and $C' \subseteq B' \subseteq A'$. If $B|A \preceq_{\mathbb{E}} C'|B'$ and $C|B \preceq_{\mathbb{E}} B'|A'$, then $C|A \preceq_{\mathbb{E}} C'|A'$; moreover, if either hypothesis is $\succ$, the conclusion is $\succ$.

Axioms 1-6 are a subset of [Krantz et al., 2006b]'s axioms for qualitative conditional probability. Importantly, the above axioms are not sufficient for for $\preceq_{\mathbb{E}}$ to be representable as a probability measure without some sort of richness assumption.[9] Alternatively, other representation theorems (e.g., Alon and Lehrer [2014]) assume a stronger form of additivity, which generalizes Scott's axiom [Fishburn, 1986]).

We briefly discuss the axioms and the ways in which we depart from [Krantz et al., 2006a]'s presentation. We have omitted [Krantz et al., 2006a]'s second axiom, which more or less prohibits conditioning on events that are equiprobable with the empty set. That axiom is not necessary for our proofs, and since it's controversial whether one ought to be able to condition on events of probability zero, we omit it. Axiom 3 is only one conjunct of [Krantz et al., 2006b]'s third axiom, and their stronger axiom raises the same logical omniscience concerns that the normality axiom (i.e., that the whole space is assigned probability one) does in the quantitative setting. Our Axiom 3 requires only one to be able to recognize that an event is self-identical.

Axiom 6 is primarily useful because it allows us to "multiply" in a qualitative setting. To see it's motivation in the quantitative setting, notice that if $C \subseteq B \subseteq A$ and $C' \subseteq B' \subseteq A'$, then

$$P(C|B) = \frac{P(C)}{P(B)} \text{ and } P(B|A) = \frac{P(B)}{P(A)},$$

[9] See [Krantz et al., 2006b, p. 205], which draws on [Kraft et al., 1959].

and similarly for the A', B', and C'. So if $P(B|A) \geq P(C'|B')$ and $P(C|B) \geq P(A'|B')$, then

$$\frac{P(B)}{P(A)} \geq \frac{P(C')}{P(B')} \text{ and } \frac{P(C)}{P(B)} \geq \frac{P(A')}{P(C')}.$$

When we multiply the left and right-hand sides of those inequalties, we obtain $P(C)/P(A) \geq P(C')/P(A')$, which is equivalent to $P(C|A) \geq P(C'|A')$ given our assumption about the nesting of the sets. For readers who find the criss-crossed terms of Axiom 6 difficult to follow, we introduce the following alternative axiom, which can be proven from the above axioms (see [Krantz et al., 2006b, Lemma 11, p. 231])

> Axiom 6': Suppose $C \subseteq B \subseteq A$ and $C' \subseteq B' \subseteq A'$. If $B|A \preceq_{\mathbb{E}} B'|A'$ and $C|B \preceq_{\mathbb{E}} C'|B'$, then $C|A \preceq_{\mathbb{E}} C'|A'$; moreover, if either hypothesis is $\succ$, the conclusion is $\succ$.

3.2 Real vs. Conceptual Experiments

Before stating qualitative analogs of the three above statistical principles, we first discuss one perhaps subtle philosophical issue between experiments and their mathematical representations.

In the quantitative setting, two *physical* experiments $\mathbb{E}$ and $\mathbb{F}$ might be represented by exactly the same *mathematical pair* $\langle \Omega, \{P_\theta\}_{\theta \in \Theta} \rangle$. For example, one might randomly assign patients a new treatment or placebo on the basis of a coin flip. Alternatively, one could roll a fair die and assign treatment based upon whether the roll was even or odd. Because each patient is treated with the same probability in the two experiments, the probabilities of various experimental outcomes are also the same. Of course, the difference between the two experiments might not be epistemically relevant. Why? One might think evidence is a function of exclusively the *probabilities* of various outcomes, not how those outcomes are actually produced.[10]

Regardless, in the qualitative setting, that distinction between experiments and their mathematical representation is obviously epistemically relevant. For example, imagine you are interested in the unknown bias $\theta \in (0, 1)$ of a coin. In experiment $\mathbb{E}$ the coin will be flipped 10 times and you learn only whether the number of heads is greater than (G), equal to (E), or less than (L) the number of tails. $\mathbb{F}$ has an identical setup, but the coin will be flipped 20 times. The experiments $\mathbb{E}$ and $\mathbb{F}$ have the same set of experimental outcomes $\{G, E, L\}$, and no matter the bias of the coin θ, the qualitative likelihood ordering of the outcomes is the same: $G|\theta \succ L|\theta \succ E|\theta$ if $\theta > 1/2$; $L|\theta \succ G|\theta \succ E|\theta$ if $\theta < 1/2$, and $G|\theta \sim L|\theta \succ E|\theta$ if $\theta = 1/2$. So $\mathbb{E}$ and $\mathbb{F}$ are represented by exactly the same mathematical pair $\langle \Omega, \preceq \rangle$, despite conveying different amounts of information.

To state plausible, qualitative analogs of C, S, and LP, therefore, one must appeal to more than the qualitative orderings representing various experiments. Instead, one must assume there is an extra-mathematical, primitive notion of

[10] For a dissenting opinion, see [Kalbfleisch, 1975], who draws on [Basu, 1964]'s distinction between an experiment that can be *performed* and one that is a mere *mathematical* possibility.

experimental identity that can be used to distinguish, for example, (1) an experiment in which $\mathbb{E}$ is conducted and then a particular sufficient statistic is reported, from (2) an experiment that is represented by the same qualitative structure as $\mathbb{E}$ and then only a sufficient statistic is reported. Similarly, to state a plausible version of C, we will need to be able to distinguish a *genuine* mixture $\mathbb{M}$ of two experiments $\mathbb{E}$ and $\mathbb{F}$ from some other experiment with the same qualitative structure.[11]

3.3 Qualitative Sufficiency

A sufficient statistic, recall, is intended to be one the summarizes all relevant information in the data. Formally, a sufficient statistic T for an experiment $\mathbb{E}$ was defined to satisfy $P^{\mathbb{E}}_{\theta}(\omega|T=t) = P^{\mathbb{E}}_{\upsilon}(\omega|T=t)$ for all ω and all $\theta, \upsilon \in \Theta$. That definition generalizes naturally:

Definition 1 (Sufficient Statistic) *Let $\mathbb{E}$ be a qualitative experiment. A statistic $T : \Omega_{\mathbb{E}} \to \mathcal{R}$ is **sufficient** (for $\mathbb{E}$) if $\omega|T=t, \theta \approx_{\mathbb{E}} \omega|T=t, \upsilon$ for all $\omega \in \Omega_{\mathbb{E}}$ and for all $\theta, \upsilon \in \Theta$.*

Given that definition, we can now state a **qualitative sufficiency principle** (QS): its statement is exactly the same as the quantitative version, except one uses the qualitative definition of "sufficient." Importantly, in light of the above discussion of the distinction between real and conceptual experiments, we emphasize that statement of QS involves a relation between $\mathbb{E}$ and $\mathbb{E}^T$, not $\mathbb{E}$ and some experiment that is represented by the same mathematical pair as $\mathbb{E}^T$.

As in the quantitative case (see lemma 1), it turns out the likelihood function is again a sufficient statistic.

Lemma 2 *Let $T : \omega \mapsto L_{\omega} = \{\omega' \in \Omega_{\mathbb{E}} : \omega'|\theta \approx_{\mathbb{E}} \omega|\theta \text{ for all } \theta \in \Theta\}$. Then T is sufficient.*

To prove lemma 2, we need a preliminary lemma.

Lemma 3 *If $L_{\omega} = L_{\omega'}$, then $\omega'|L_{\omega}, \theta \approx_{\mathbb{E}} \omega|L_{\omega}, \theta$ for all θ.*

Proof: Suppose for the sake of contradiction there is some θ such that $\omega'|L_{\omega}, \theta \not\approx_{\mathbb{E}} \omega|L_{\omega}, \theta$. Then by Axiom 1 (specifically, totality of the ordering), either $\omega|L_{\omega}, \theta \prec \omega'|L_{\omega}, \theta$ or vice versa. Without loss of generality, assume $\omega|L_{\omega}, \theta \prec_{\mathbb{E}} \omega'|L_{\omega}, \theta$. Define:

$$\begin{array}{ll} A = \{\theta\} & A' = A = \{\theta\} \\ B = L_{\omega} \cap \{\theta\} & B' = B = L_{\omega} \cap \{\theta\} \\ C = \{\omega'\} \cap L_{\omega} \cap \{\theta\} & C' = \{\omega\} \cap L_{\omega} \cap \{\theta\} \end{array}$$

[11] To avoid objections to Birnbaum's original argument for LP, Gandenberger [2014] likewise draws extra-mathematical distinctions between different types of experiments. Because [Gandenberger, 2014]'s proof establishes LP in full generality, however, those distinctions turn out not to matter. In contrast, the above examples show that the qualitative likelihood ordering does not encode all of information in a sample, and so the distinction between two experiments with identical qualitative structure is often epistemically significant.

Clearly $B|A \approx_{\mathbb{E}} B'|A'$ as $B|A = B'|A'$. Further:

$$\begin{aligned} C|B &= \{\omega'\} \cap L_\omega \cap \{\theta\}|L_\omega \cap \{\theta\} \\ &\approx_{\mathbb{E}} \{\omega'\}|L_\omega \cap \{\theta\} \text{ by Axiom 4} \\ &\succ_{\mathbb{E}} \{\omega\}|L_\omega \cap \{\theta\} \text{ by assumption} \\ &\approx_{\mathbb{E}} C'|B' \text{ by Axiom 4} \end{aligned}$$

So by Axiom 6', it follows that $C|A \succ C'|A'$, i.e., that

$$\dagger\ \omega' \cap L_\omega \cap \{\theta\}|\theta \succ \omega \cap L_\omega \cap \{\theta\}|\theta.$$

Now because $\omega, \omega' \in L_\omega = L_{\omega'}$, it follows that $\{\omega\} \cap L_\omega = \{\omega\}$ and similarly $\{\omega'\} \cap L_\omega = \{\omega'\}$. Thus, $\dagger$ entails $\{\omega'\} \cap \{\theta\}|\theta \succ \{\omega\} \cap \{\theta\}|\theta$. By Axiom 4, it follows that $\omega'|\theta \succ \omega|\theta$. But $L_\omega = L_{\omega'}$ by assumption, and hence, $\omega'|\theta \approx_{\mathbb{E}} \omega|\theta$, contradiction.

□

We can now prove lemma 2.

Proof: We must show that $\omega|T = t, \theta \approx_{\mathbb{E}} \omega|T = t, \upsilon$ for all ω and for all $\theta, \upsilon \in \Theta$, where $T : \omega \mapsto L_\omega$. So let θ, υ be arbitrary.

Consider first the case in which $t \neq L_\omega$, and so $\omega \cap \{T = t\} = \emptyset$. Then $\omega|T = t, \theta \approx_{\mathbb{E}} \omega \cap \{T = t\} \cap \{\theta\}|T = t, \theta$ by Axiom 4, and so $\omega|T = t, \theta \approx_{\mathbb{E}} \emptyset|T = t, \theta$. Similarly, $\omega|T = t, \upsilon \approx_{\mathbb{E}} \emptyset|T = t, \upsilon$. It suffices to show only that $\emptyset|A \approx_{\mathbb{E}} \emptyset|B$ for all A, B (when defined), and this is a fairly routine exercise (see [Krantz et al., 2006a, p.229] for a proof).

Next, consider the case in which $t = L_\omega$, and so $\{T = t\} = L_\omega$. Then we must show that $\omega|L_\omega, \theta \approx_{\mathbb{E}} \omega|L_\omega, \upsilon$. Suppose for the sake of contradiction not. Then by Axiom 1 (specifically, totality of $\preceq$), it follows that $\omega|L_\omega, \theta \prec \omega|L_\omega, \upsilon$ or vice versa. Without loss of generality, assume $\omega|L_\omega, \theta \prec \omega|L_\omega, \upsilon$.

Next, define $R_\omega = L_\omega \setminus \{\omega\}$. We claim there is some $\omega' \in R_\omega$ such that either (1) $\omega'|L_\omega, \theta \not\approx_{\mathbb{E}} \omega|L_\omega, \theta$ or (2) $\omega'|L_\omega, \upsilon \not\approx_{\mathbb{E}} \omega|L_\omega, \upsilon$. If neither (1) nor (2) holds, then we have $\omega'|L_\omega, \theta \approx_{\mathbb{E}} \omega|L_\omega, \theta$ AND $\omega'|L_\omega, \upsilon \approx_{\mathbb{E}} \omega|L_\omega, \upsilon$ for all $\omega' \in R_\omega$. So by repeatedly applying Axiom 5 (recall, we've assumed $\Omega_{\mathbb{E}}$ is finite) to the assumption that $\omega|L_\omega, \theta \prec \omega|L_\omega, \upsilon$, we obtain that $R_\omega|L_\omega, \theta \prec R_\omega|L_\omega, \upsilon$. But then because $R_\omega|L_\omega, \theta \prec R_\omega|L_\omega, \upsilon$ and $\omega|L_\omega, \theta \prec \omega|L_\omega, \upsilon$, one last application of Axiom 5 yields that $L_\omega|L_\omega, \theta \prec L_\omega|L_\omega, \upsilon$. By Axiom 4, it follows that $L_\omega, \theta|L_\omega, \theta \prec L_\omega, \upsilon|L_\omega, \upsilon$. But that contradicts Axiom 3.

So either (1) or (2) obtains. But notice that both contradict lemma 3 as $\omega' \in R_\omega \subseteq L_\omega$.

□

3.4 Qualitative Conditionality

As in the case of sufficiency, to state a plausible, qualitative analog of C, one cannot appeal to strictly mathematical facts about the ordering. Instead, one must

appeal to the extra-mathematical notion of experimental identity to distinguish a *genuine* mixture $\mathbb{M}$ of two experiments $\mathbb{E}$ and $\mathbb{F}$ from some other experiment with the same (or worse yet, isomorphic) qualitative structure. But if one can identify a "genuine" mixture $\mathbb{M}$, the statement of C remains unchanged.

3.5 Qualitative LP?

Unlike S and C, there is no *direct* qualitative analog of LP. Why? The statement of LP involves both multiplication and numerical constants. We will not define "direct", but we hope the idea is clear: a quantitative statement has a direct qualitative analog if one can simply erase some "P"s, "Q"s , and parentheses to obtain a well-formed expression in our qualitative framework.

Given there is no *direct* qualitative analog of LP, what properties must an *indirect* analog φ have? We propose the following: (1) there a quantitative statement ψ that is mathematically equivalent to LP and (2) φ is the direct qualitative analog of ψ. Here is the quantitative principle ψ that is mathematically equivalent to LP.

Mixed Experiment Principle (MEP): Let $\mathbb{E}$ and $\mathbb{F}$ be two experiments over the same parameter space Θ. Two pieces of evidence E from $\omega_{\mathbb{E}}$ and F from $\omega_{\mathbb{F}}$ are evidentially equivalent if there is a mixture $\mathbb{M}$ of $\mathbb{E}$ and $\mathbb{F}$ such that $P_\theta^{\mathbb{M}}(\langle 0, \omega_{\mathbb{E}}\rangle) = P_\theta^{\mathbb{M}}(\langle 1, \omega_{\mathbb{F}}\rangle)$ for all $\theta \in \Theta$.

It is easy to see that LP and MEP are equivalent.

Proposition 3 LP *and* MEP *are equivalent.*

Proof: To show MEP entails LP, suppose there is a mixture $P_\theta^{\mathbb{M}}(\langle 0, \omega_{\mathbb{E}}\rangle) = P_\theta^{\mathbb{M}}(\langle 1, \omega_{\mathbb{F}}\rangle)$ for all $\theta \in \Theta$. By definition of mixed experiment, this means that $c \cdot P_\theta^{\mathbb{E}}(\omega_{\mathbb{E}}) = (1-c) \cdot P_\theta^{\mathbb{F}}(\omega_{\mathbb{F}})$ for all $\theta \in \Theta$. Thus, $P_\theta^{\mathbb{E}}(\omega_{\mathbb{E}}) = \frac{1-c}{c} \cdot P_\theta^{\mathbb{F}}(\omega_{\mathbb{F}})$ for all $\theta \in \Theta$. Since $0 < c < 1$, we know $\frac{1-c}{c} > 0$, and so LP entails E and F are evidentially equivalent.

The proof of the other direction is omitted because it's similar to the proof of Birnbaum's theorem.

□

Importantly, if LP entails two outcomes $\omega_{\mathbb{E}}$ and $\omega_{\mathbb{F}}$ to be evidentially equivalent, then the mixed experiment constructed in the above proposition is not some purely theoretical or "conceptual" experiment: it really can be performed by constructing a randomizing device that yields one outcome with probability $c/(1+c)$ and a second outcome otherwise, and to construct such a randomizing device, one needs only a random number generator with sufficiently many outcomes, or an urn with sufficiently many balls, etc. So those who wish to reject LP must also reject MEP. Notice that MEP generalizes straightforwardly to the qualitative setting, and so we take the following principle to be the qualitative analog of LP.

Qualitative Mixed Experiment Principle (QMEP): Let $\mathbb{E}$ and $\mathbb{F}$ be two experiments over the same parameter space Θ. Two pieces of evidence $\omega_{\mathbb{E}}$ from $\mathbb{E}$ and $\omega_{\mathbb{F}}$ from $\mathbb{F}$ are evidentially equivalent if there is a mixture $\mathbb{M}$ of $\mathbb{E}$ and $\mathbb{F}$ such that $\langle 0, \omega_{\mathbb{E}} \rangle | \theta \approx_{\mathbb{M}} \langle 1, \omega_{\mathbb{F}} \rangle | \theta$ for all $\theta \in \Theta$.

3.6 Generalizing Birnbaum's Theorem

We now prove a qualitative analog of Birnbaum's theorem, namely, we show that the qualitative versions of sufficiency and conditionality (which have the same statements of the quantitative ones for all intents and purposes) together entail QMEP. The proof is essentially the same as in the quantitative case.

Theorem 2 QS *and* QC *entail* QMEP.

Proof: Let $\mathbb{M}$ be a mixture of two experiments $\mathbb{E}$ and $\mathbb{F}$ such that $\langle 0, \omega_{\mathbb{E}} \rangle | \theta \approx_{\mathbb{M}} \langle 1, \omega_{\mathbb{F}} \rangle | \theta$ for all $\theta \in \Theta$. We must show that $\omega_{\mathbb{E}}$ from $\mathbb{E}$ and $\omega_{\mathbb{F}}$ from $\mathbb{F}$ are evidentially equivalent. To do so, note that by conditionality, $\omega_{\mathbb{E}}$ from $\mathbb{E}$ is evidentially equivalent to $\langle 0, \omega_{\mathbb{E}} \rangle$ from $\mathbb{M}$. Similarly, again by conditionality, $\omega_{\mathbb{F}}$ from $\mathbb{F}$ is evidentially equivalent to $\langle 1, \omega_{\mathbb{F}} \rangle$ from $\mathbb{M}$. So by transitivity of evidential equivalence, it suffices to show that the two outcomes $\langle 0, \omega_{\mathbb{E}} \rangle$ and $\langle 1, \omega_{\mathbb{F}} \rangle$ of $\mathbb{M}$ are evidentially equivalent. By assumption, $\langle 0, \omega_{\mathbb{E}} \rangle | \theta \approx_{\mathbb{M}} \langle 1, \omega_{\mathbb{F}} \rangle | \theta$ for all $\theta \in \Theta$. In other words, the outcomes $\langle 0, \omega_{\mathbb{E}} \rangle$ and $\langle 1, \omega_{\mathbb{F}} \rangle$ have the same likelihood functions in $\mathbb{M}$, or in symbols, $L_{\langle 0, \omega_{\mathbb{E}} \rangle} = L_{\langle 1, \omega_{\mathbb{F}} \rangle}$. By lemma 2, the likelihood function is a qualitative sufficient statistic. Hence, by the Weak Sufficiency Principle, $\langle 0, \omega_{\mathbb{E}} \rangle$ and $\langle 1, \omega_{\mathbb{F}} \rangle$ are evidentially equivalent, as desired.

□

4 Future work: Relation to Qualitative Bayesianism

S, C, and LP are all intricately connected to Bayesianism, as discussed in §2.3. Is there any relationship between these principles and some sort of qualitative version of Bayesianism?

As discussed above, Bayesianism is typically understood as the conjunction of (1) probabilism and (2) the thesis that one's degrees of belief ought to be updated by conditionalization. To obtain a weaker (perhaps more plausible) theory of rationality, one might instead pick some subset $\mathcal{A}$ of the above axioms (or the necessary axioms of some other representation theorem) and endorse the following conjunction: (1') rational comparative beliefs should satisfy $\mathcal{A}$ (which are necessary but insufficient for a probabilistic representation) and (2') after learning E, a rational agent's degree of belief in H should be $H|E$. For each set $\mathcal{A}$, we can obtain in this way a theory that might be called **$\mathcal{A}$-qualitative Bayesianism**. Here, we will investigate the case in which $\mathcal{A}$ is all the axioms

above, but we do not claim those axioms $\mathcal{A}$ are sufficient for rational (conditional) belief.

Recall, in the quantitative setting, we defined two experimental outcomes $\omega_{\mathbb{E}}$ and $\omega_{\mathbb{F}}$ to be Bayesian posterior equivalent if $Q^{\mathbb{E}}(\cdot|\omega_{\mathbb{E}}) = Q^{\mathbb{F}}(\cdot|\omega_{\mathbb{F}})$ for all priors π_Q, i.e., *every* Bayesian would update her degrees of belief upon learning $\omega_{\mathbb{E}}$ in the same way she would upon learning $\omega_{\mathbb{F}}$. Although our qualitative framework does not allow us to compare posterior probabilities of outcomes drawn from different experiments (because the relation $\approx_{\mathbb{E}}$, unlike numerical equality, is indexed to a specific experiment $\mathbb{E}$), we can analogously define two outcomes $\omega, \omega' \in \Omega_{\mathbb{E}}$ of the same experiment $\mathbb{E}$ to be **$\mathcal{A}$-posterior equivalent$_{\mathbb{E}}$** if $\theta|\omega \approx_{\mathbb{E}} \theta|\omega'$ for all θ and all orderings $\preceq_{\mathbb{E}}$ satisfying the axioms of $\mathcal{A}$, i.e., *every* agent whose degrees of belief satisfy the axioms of $\mathcal{A}$ would update her degrees of belief upon learning $\omega_{\mathbb{E}}$ in the same way she would upon learning $\omega_{\mathbb{F}}$.[12] Then we can prove a qualitative analog of (part of) proposition 1.

Theorem 3 *If the Weak Sufficiency Principle entails two outcomes are evidentially equivalent, then they are $\mathcal{A}$-posterior equivalent$_{\mathbb{E}}$.*

The proof of this theorem is long, and so it is omitted. We state the theorem only to motivate several important open technical and philosophical questions raised by our work.

From a technical perspective, here are several central issues. In the quantitative setting, it is well-known that if any of S, C, or LP/MEP entail that two outcomes are evidentially equivalent, then they are Bayesian posterior equivalent (again, see proposition 1). Theorem 3 is analogous to the result for the Weak Sufficiency Principle, but it is not clear how to *formulate* analogous results for QS, QC and QMEP in our framework. Why? As noted above, the relation $\approx_{\mathbb{E}}$, unlike numerical equality, is indexed to a specific experiment $\mathbb{E}$. So in our current framework, there is no mechanism for comparing an agent's posterior degree of belief $H|\omega_{\mathbb{E}}$ after an experiment $\mathbb{E}$ to her posterior $H|\omega_{\mathbb{F}}$ after an experiment $\mathbb{F}$. Further research is necessary.

From a philosophical perspective, those technical questions matter. Recall, the in quantitative setting, LP completely characterizes Bayesian posterior equivalence (i.e., two outcomes are Bayesian posterior equivalent if and only if LP entails they are evidentially equivalent). See proposition 2. In the quantitative setting, Birnbaum's theorem, therefore, amounts to a proof that frequentist assumptions entail that one's beliefs ought to abide by Bayesian norms. So an analogous proof, showing that two outcomes are $\mathcal{A}$ posterior equivalent if and only if QMEP entails them to be, would show that norms analogous to those adopted by the Bayesian persist even if one substantially weakens the axioms for rational belief. And our qualitative generalization of Birnbaum's theorem, therefore, would amount to a proof that very weak frequentist assumptions entail that one's beliefs ought to abide by qualitatively-Bayesian norms.

[12] Of course, one must assume that such agents agree upon the qualitative likelihood orderings just as, in the quantitative setting, Bayesians agree upon the likelihood functions $\{P_\theta(\omega)\}_{\theta\in\Theta}$.

References

Shiri Alon and Ehud Lehrer. Subjective multi-prior probability: A representation of a partial likelihood relation. *Journal of Economic Theory*, 151: 476–492, 2014.

D. Basu. Recovery of ancillary information. *Sankhyā: The Indian Journal of Statistics*, pages 91–104, 1964.

James Orvis Berger and Robert L. Wolpert. The likelihood principle. In *Ims Lecture Notes-Monograph*, volume 6. Institute of Mathematical Statistics, 1988.

Jose M. Bernardo. Modern Bayesian inference: Foundations and objective methods. In M. Forster and Prasanta S Bandyopadhyay, editors, *Philosophy of Statistics*, number 7 in Handbook of the Philosophy of Science, pages 263–306. 2011.

Allan Birnbaum. On the foundations of statistical inference. *Journal of the American Statistical Association*, 57(298):269–306, 1962.

George Casella and Roger L. Berger. *Statistical inference*, volume 70. Duxbury Press Belmont, CA, second edition, 2002.

James Durbin. On Birnbaum's theorem on the relation between sufficiency, conditionality and likelihood. *Journal of the American Statistical Association*, 65(329):395–398, 1970.

Anthony William Fairbank Edwards. *Likelihood*. CUP Archive, 1984.

Michael J. Evans, Donald AS Fraser, and Georges Monette. On principles and arguments to likelihood. *Canadian Journal of Statistics*, 14(3):181–194, 1986.

Peter C. Fishburn. The axioms of subjective probability. *Statistical Science*, 1(3):335–358, 1986. URL `http://projecteuclid.org/euclid.ss/1177013611`.

Greg Gandenberger. New responses to Three counterexamples to the Likelihood Principle. *Unpublished manuscript*, 2014.

John D. Kalbfleisch. Sufficiency and conditionality. *Biometrika*, 62(2):251–259, 1975.

Charles H. Kraft, John W. Pratt, and Abraham Seidenberg. Intuitive probability on finite sets. *The Annals of Mathematical Statistics*, pages 408–419, 1959.

David H. Krantz, R. Duncan Luce, Patrick Suppes, and Amos Tversky. *Foundations of Measurement Volume I: Additive and Polynomial Representations*. Dover Publications, Mineola, NY, December 2006a. ISBN 978-0-486-45314-9.

David H. Krantz, R. Duncan Luce, Patrick Suppes, and Amos Tversky. *Foundations of Measurement Volume II: Geometrical, Threshold, and Probabilistic Representations*. Dover Publications, Mineola, N.Y, December 2006b. ISBN 978-0-486-45315-6.

Dennis V. Lindley. *Understanding uncertainty*. Wiley-Interscience, 2006.

Deborah G. Mayo. On the Birnbaum argument for the strong likelihood principle. *Statistical Science*, pages 227–239, 2014.

Deborah G. Mayo. *Statistical inference as severe testing*. Cambridge: Cambridge University Press, 2018.

Conor Mayo-Wilson and Aditya Saraf. Qualitative Robust Bayesianism and the Likelihood Principle. 2020.

Jerzy Neyman and Egon S. Pearson. 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, 1933.

Richard Royall. *Statistical evidence: a likelihood paradigm*, volume 71. Chapman & Hall/CRC, 1997.

L. J. Savage. *The foundation of statistics*. Dover publications, 1972.