philosophies

MDPI

Article

On Falsifiable Statistical Hypotheses

Konstantin Genin

Cluster of Excellence—Machine Learning for Science, University of Tübingen, 72076 Tübingen, Germany; konstantin.genin@uni-tuebingen.de

Abstract: Popper argued that a statistical falsification required a prior methodological decision to regard sufficiently improbable events as ruled out. That suggestion has generated a number of fruitful approaches, but also a number of apparent paradoxes and ultimately, no clear consensus. It is still commonly claimed that, since random samples are logically consistent with all the statistical hypotheses on the table, falsification simply does not apply in realistic statistical settings. We claim that the situation is considerably improved if we ask a conceptually prior question: when should a statistical hypothesis be regarded as falsifi*able*. To that end we propose several different notions of statistical falsifiability and prove that, whichever definition we prefer, the same hypotheses turn out to be falsifiable. That shows that statistical falsifiability enjoys a kind of conceptual robustness. These notions of statistical falsifiability are arrived at by proposing statistical analogues to intuitive properties enjoyed by exemplary falsifiable hypotheses familiar from classical philosophy of science. That demonstrates that, to a large extent, this philosophical tradition was on the right conceptual track. Finally, we demonstrate that, under weak assumptions, the statistically falsifiable hypotheses correspond precisely to the closed sets in a standard topology on probability measures. That means that standard techniques from statistics and measure theory can be used to determine exactly which hypotheses are statistically falsifiable. In other words: the proposed notion of statistical falsifiability both answers to our conceptual demands and submits to standard mathematical techniques.

Keywords: falsifiability; statistical hypotheses; induction

Citation: Genin, K. On Falsifiable Statistical Hypotheses. *Philosophies* **2022**, *7*, 40. https://doi.org/10.3390/philosophies7020040

Academic Editor: Benjamin Jantzen

Received: 1 October 2021
Accepted: 6 March 2022
Published: 2 April 2022

1. Introduction

Probabilistic theories posed a challenge for Popper's falsificationism. But, on first glance, there are strong analogies between "naive" falsificationism, and standard practice in frequentist hypothesis testing. For example, a standard frequentist test of a sharp null hypothesis rejects upon observing an event that would be highly improbable if the null hypothesis were true. Such a test has a very low chance of erroneously rejecting the null hypothesis. That falls short of, but is closely analogous to, the infallibility of rejecting a universal law upon observing a countervailing instance. Alternatively, a standard frequentist test of a sharp null hypothesis has a high chance of erroneously *failing* to reject if the null hypothesis is subtly false. That is similar to the fallibility of inferring a universal hypothesis from finitely many instances. Such analogies are natural and sometimes made explicitly by statisticians. Here are Gelman and Shalizi [1]:

> Extreme *p*-values indicate that the data violate regularities implied by the model, or approach doing so. If these were strict violations of deterministic implications, we could just apply *modus tollens* . . . as it is, we nonetheless have evidence and probabilities. Our view of model checking, then, is firmly in the long hypothetico-deductive tradition, running from Popper (1934/1959) back through Bernard (1865/1927) and beyond (Laudan, 1981).

Statistical falsification, Gelman and Shalizi suggest, is *all but deductive*.[1] But how extremal, exactly, does a *p*-value have to be for a test to count as a falsification? Popper was loathe to draw the line at any particular value: "It is fairly clear that this 'practical falsification' can

be obtained only through a methodological decision to regard highly improbable events as ruled out . . . But with what right can they be so regarded? Where are we to draw the line? Where does this 'high improbability' begin?" ([3], p. 182) The problem of when to consider a statistical hypothesis to be falsified has engendered a significant literature [4–6] but no universally accepted answer.

A natural idea, due to Fisher [7] and Gillies [4], is to pick some canonical event that is highly improbable if the hypothesis is true. The problem with this idea, pointed out by Neyman [8] and given dramatic expression by Redhead [9], is that there is no unique methodological convention satisfying this property; indeed there may be competing conventions giving conflicting verdicts on every sample.[2] Neyman and Pearson [11] attempt to solve this uniqueness problem by requiring the falsifying event to have maximum probability if the hypothesis is false. Falsificationists like Gillies [4] object that this strategy involves restricting the set of alternatives in artifical ways. But even if we grant the basics of the Neyman–Pearson framework, there is typically no way to choose a single falsifying event with "uniformly" maximum probability in all the alternative possibilities in which the hypothesis is false: one must always favor some alternative over others (See Casella and Berger [12], p. 393). In short, the question of when to consider an event falsifying seems to have no answer entirely free of arbitrariness.

Our response to this situation is two-fold. The first response is nicely summarized by Redhead [9]: "Popper demands in science refutability, not refutation". The question of which hypotheses are refutable is at least as important as the question of when data should be taken to have refuted a hypothesis.[3] The existence of a univocal methodological convention is not necessary for refutability; indeed, if there were a univocal answer, the problem would no longer have any conventional or methodological aspect. The results of this work allow us to be fairly ecumenical about which methodology of falsification should be adopted without losing any clarity about which hypotheses are falsifiable: Theorem 2 says that a variety of different methodologies of falsification—ranging from the permissive to the stringent—give rise to exactly the same collection of falsifiable hypotheses. Moreover, it provides a relatively simple mathematical criterion for identifying the falsifiable hypotheses. That means scientific controversies about the testability of hypotheses can be adjudicated without awaiting a precise resolution of the problem of statistical falsification.

Although the existence of a univocal methodological convention is not necessary for refutability, what is important is the existence of a coherent set of methodological decisions exhibiting some desirable properties. Statistical tests are implicit proposals for such a convention, proposing a falsifying event (rejection region) for every sample size. To the extent possible, these tests should exhibit the synchronic virtues sought after by Fisher, Neyman, Pearson and others: at every sample size, if the hypothesis is true, the chance that it is rejected should be small; and if it is false, the chance that it is rejected should be large. But excessive focus on these synchronic virtues tends to obscure the social dimensions of these decisions. Our second response is a plea to consider another axis on which competing methodological proposals can be compared. If it were adopted by researchers conducting independent investigations of the same hypothesis, the method should also exhibit certain diachronic virtues: if the hypothesis is true, the chance that it is falsely rejected should shrink as monotonically as possible as the investigation is replicated at larger sample sizes; and similarly, if the hypothesis is false, the chance that it is correctly rejected should grow as monotonically as possible. In other words: a good methodological convention would, if adopted, support a pattern of successful replication by independent investigators. Crucially, possessing the synchronic virtues at every sample size is not sufficient to ensure the diachronic virtues: care must be taken such that the methodological decisions for different sample sizes cohere in a particular way.[4] Finally, falsifiable hypotheses would be those for which a methodological convention exhibiting all of these virtues exists.

We end this introductory section with a few comments on routes not taken in the following. The work of Mayo and Spanos [6] is in its own way a mathematical elaboration

of Popperian falsifiability. It is natural to equate falsifiability with Mayo and Spanos' severe testability. Although severe testability is an important notion of independent interest, we do not adopt this identification—it would be too radical a revision of falsifiability. Roughly, a hypothesis is severely tested by some evidence if, were the hypothesis false, then with high probability evidence less favorable to the hypothesis would have been observed. On that stringent notion, hypotheses like *the coin is precisely fair* are not severely testable, since whatever sequence of flips has been observed, a sequence equally favorable to the *precisely fair* hypothesis could have been generated by a subtly biased coin. But we take sharp null hypotheses like *the coin is precisely fair* as archetypical falsifiable hypotheses. Finally, it might be justly said that any contemporary philosophical discussion of statistics should say something about the Bayesian point of view. We do not attempt to fully satisfy a philosophical Bayesian. But although this work is inspired by a rival philosophical tradition, it should not be understood as incompatible with Bayesianism. On any of a number of reasonable correspondence principles, any hypothesis falsifiable by a Bayesian method will be falsifiable in our sense. Although the question of whether the converse is true is an interesting one, we do not pursue it here.

2. In Search of Statistical Falsifiability

Accordingly, instead of asking 'when should a statistical hypothesis be regarded as falsified?', we are guided in the following by the question: 'when should a statistical hypothesis be regarded falsifi*able*?'. We suggest that statistical falsifiability will be found by analogy with exemplars from philosophy of science and the theory of computation. We have in mind universal hypotheses like 'all ravens are black', or co-semidecidable formal propositions like 'this program will not halt'. Although there is no *a priori* bound on the amount of observation, computation, or proof search required, these hypotheses may be falsified by suspending judgement until the hypothesis is decisively refuted by the provision of a non-black raven or a halting event. We wish to call the reader's attention to several paradigmatic properties of these falsification methods. We do not claim that these properties are typically achievable in empirical inquiry; rather, they should be seen as regulative ideals that falsification methods ought to approximate.

Error Avoidance: Output conclusions are true.

By suspending judgement until the hypothesis is logically incompatible with the evidence, falsification methods never have to "stick their neck out" by making a conjecture that might be false.[5]

Monotonicity: Logically stronger inputs yield logically stronger conclusions.

Exemplary falsification methods never have to retract their previous conclusions; their conjecture at any later time always entails their conjecture at any previous time. In the ornithological context, conjectures made on the basis of more observations always entail conjectures made on the basis of fewer; once a non-black raven has been observed, the hypothesis is decisively falsified. (We appeal here to an idealization that we later discharge in the statistical setting: an exemplary method never mistakes a black raven for a non-black raven). In the computational context, conjectures made on the basis of more computation always entail conjectures made on the basis of less; once the program has entered a halting state, it will never exit again.

Limiting Convergence: The method converges to $\neg H$ iff H is false.

If all ravens are black, the falsification method may suspend judgement forever; but if there is some non-black raven, diligent observation will turn up a falsifying instance eventually. Similarly, if the program eventually halts, the patient observer will notice.[6]

We propose that statistical falsifiability will be found if we look for the minimal weakening of these paradigmatic properties that is feasible in statistical contexts. First, some definitions.[7] *Inference methods* output conjectures on the basis of input information. Here "information" is understood broadly. One conception of information, articulated explicitly by Bar-Hillel and Carnap [19] and championed by Floridi [20,21], is true, propositional semantic content, logically entailing certain relevant possibilities, and logically refuting others. This notion is prevalent in epistemic logic and many related formal fields—call it the *propositional* notion of information. A second conception, ubiquitous in the natural sciences, is random samples, typically independent and identically distributed, and logically consistent with all relevant possibilities, although more probable under some, and less probable under others. Call that the *statistical* notion of information. Of course, statistical information can be expressed propositionally. The trouble is that—at least so far as the relevant possibilities are concerned—the proposition is always the same, since every proposition about the observed data is logically consistent with every probabilistic proposition.

Statistical methods cannot be expected to be infallible. We liberalize that requirement as follows:

> α-*Error Avoidance*: For every sample size, the objective chance that the output conclusion is false is not higher than α.

The α-error avoidance property is closely related to frequentist statistical inference. A confidence interval with coverage probability $1 - \alpha$ is straightforwardly α-error avoiding: the chance that the interval excludes the true parameter is bounded by α. A hypothesis test with significance level α is also α-error avoiding, so long as one understands failure to reject the null hypothesis as recommending suspension of judgment, rather than concluding that the null hypothesis is true. The chance of falsely rejecting the null is bounded by α, and failing to reject outputs only the trivially true, or tautological, hypothesis.

We first define a weaker notion. A *method* is a function from information to conjectures. A method *falsifies hypothesis H in the limit* by converging, on increasing information, to not-H iff H is false. In statistical settings, that means that in all and only the worlds in which H is false, the chance that $L(\cdot)$ outputs not-H converges to 1 as sample size increases.[8] A method *verifies H in the limit* if it falsifies not-H in the limit. A method *decides H in the limit* if it verifies H and not-H in the limit. Hypothesis H is verifiable, refutable, or decidable in the limit iff there exists a method that verifies, refutes or decides it in the limit.

Falsification *in the limit* is a relatively undemanding concept of success — it is consistent with any finite number of errors and volte-faces prior to convergence. Falsification is a stronger success notion. Consider the following cycle of definitions.[9]

F1. Hypothesis H is falsifiable iff there is a monotonic, error avoiding method M that falsifies H in the limit;

F1.5. Hypothesis H is falsifiable iff there is a method M that falsifies H in the limit, and for every $\alpha > 0$, M is α-error avoiding;

F2. Hypothesis H is falsifiable iff for every $\alpha > 0$ there is an α-error evoiding method that falsifies H in the limit.

Concept F1 is the familiar one from epistemology, the philosophy of science, and the theory of computation. Our motivating examples are all of this type. These may be falsified by suspending judgement until the relevant hypothesis is logically refuted by information. Although there is no *a priori* bound on the amount of information (or computation) required, the outputs of a falsifier are guaranteed to be true, without qualification.

Concept F1.5 weakens concept F1 by requiring only that there exist a method that avoids error with probability one. Hypotheses of type F1.5 are less frequently encountered in the wild.[10] Concept F1.5 is introduced here to smooth the transition to F2.

F2 weakens F1.5 by requiring only that for every bound on the chance of error, there exists a method that achieves the bound. Hypotheses of this type are ubiquitous in statistical settings. The problem of falsifying that a coin is fair by flipping it indefinitely is an archetypical problem of the third kind. For any $\alpha > 0$, there is a consistent hypothesis test with significance level α that falsifies, in the third sense, that the coin is fair. Moreover, it is hard to imagine a more stringent notion of falsification that could actually be implemented in digital circuitry. Electronics operating outside the protective cover of Earth's atmosphere are often disturbed by space radiation—energetic ions can flip bits or change the state of memory cells and registers [22]. Even routine computations performed in space are subject to non-trivial probabilities of error, although the error rate can be made arbitrarily small by redundant circuitry, error-correcting codes, or simply by repeating the calculation many times and taking the modal result. Electronics operating on Earth are less vulnerable, but are still not immune to these effects.

Since issues of monotonicity are ignored, F2 provides only a partial statistical analogue for F1. A statistical analogue of monotonicity is suggested by considerations of replication. Suppose that researchers write a grant to study whether NewDrug is better at treating migraine than OldDrug. They perform a pre-trial analysis of their methodology and conclude that if NewDrug is indeed better OldDrug, the objective chance that they will reject the null hypothesis of "no improvement" is greater than 50%. The funding agency is satisfied, providing enough funding to perform a pilot study with $N = 100$. Elated, the researchers perform the study, and correctly reject the null hypothesis. Now suppose that a replication study is proposed at sample size 150, but the chance of rejecting has decreased to 40%. The chance of rejecting correctly, and thereby replicating successfully, has gone down, even though the first study was correct. Nevertheless, investigators propose going to the trouble and expense of collecting a larger sample! Such methods are epistemically defective. They may even be immoral: why expose more patients to potential side effects for no epistemic gain? Accordingly, consider the following statistical norm:

> *Monotonicity in chance*: If H is false, then the objective chance of rejecting H is strictly increasing with sample size.

Unfortunately, strict monotonicity is often infeasible (see Lemma 1). Nevertheless, it should be a regulative ideal that we strive to approximate. The following principle expresses that aspiration:

> *α-Monotonicity in chance*: If H is false, then for any sample sizes $n_1 < n_2$, the objective chance of rejecting H decreases by no more than α.

That property ensures that collecting a larger sample is never a disastrously bad idea. Equipped with a notion of statistical monotonicity, we state the final definition in our cycle:

F3. Hypothesis H is falsifiable iff for every $\alpha > 0$ there is an α-error avoiding and α-monotonic method that falsifies H in the limit.

F3 seems like a rather modest strengthening of F2. Surprisingly, many standard frequentist methods satisfy F2, but not F3. Chernick and Liu [23] noticed non-monotonic behavior in the power function of standard hypothesis tests of the binomial proportion, and proposed heuristic software solutions. That defect would have precisely the bad consequences that inspired our statistical notion of monotonicity: attempting replication with a larger sample might actually be a bad idea. That issue has been raised in consumer safety regulation, vaccine studies, and agronomy [24–26]. But Chernick and Liu [23] have noticed only the tip of the iceberg—similar considerations attend all statistical inference methods. One of the results of this paper (Theorem A1) is that F3 is feasible whenever F2 is. That suggests that stasticial feasibility is a robust notion; many different formulations yield the same collection of falsifiable hypotheses.

It is sometimes more natural to speak not in terms of falsifiability but its dual: verifiability. The falsifiability notions F1, F2, F3 give rise to corresponding notions of verifiability V1, V2, V3 by letting H be verifiable (in the relevant sense) iff its logical complement $\neg H$ is falsifiable (in the relevant sense). Then, such archetypical hypotheses as, 'not all ravens are black', or semidecidable formal proposition like 'this program will halt' are verifiable in the sense of V1. The statistical hypotheses that typically serve as composite alternatives to sharp null hypotheses turn out to be verifiable in the sense of V2 and V3.

Both verifiability and falsifiability are one-sided notions. They can be symmetrized by defining H to be decidable (in the relevant sense) iff H is both verifiable and falsifiable (in the relevant sense). That gives rise to the three corresponding decidability concepts D1, D2 and D3. Finite-horizon empirical hypotheses such as 'the first hundred observed ravens will be black' and formal propositions like 'this program will halt in under a hundred steps' are decidable in the sense of D1. Typically, simple vs. simple hypothesis testing problems turn out to be decidable in the sense of D3. However, some more interesting problems also turn out to be statistically decidable (see Genin and Mayo-Wilson [27]).

3. The Topology of Inquiry

A central insight of Abramsky [28], Vickers [29], Kelly [30] is that falsifiable propositions of type F1 enjoy the following properties:

C1. If H_1, H_2 are falsifiable, then so is their disjunction, $H_1 \cup H_2$;
C2. If $\mathcal{H}$ is a (potentially infinite) collection of falsifiable propositions, then their conjunction, $\cap\mathcal{H}$, is also falsifiable.

Together, C1 and C2 say that falsifiable propositions of type F1 are closed under conjunction, and *finite* disjunction. Why the asymmetry? For the same reason that it is possible to falsify that bread will cease to nourish sometime next week, but not possible to falsify that it will cease to nourish on some day in the future. It is also important to notice what C1 and C2 *do not* say: if H is falsifiable, its negation may not be. To convince yourself of this it suffices to notice that it is possible to falsify that bread will always nourish, but not that it will cease to nourish on some day in the future.

For their part, verifiable propositions of type V1 satisfy the dual properties:

O1. If H_1 and H_2 are verifiable, then so is their conjunction, $H_1 \cap H_2$;
O2. If $\mathcal{H}$ is a (potentially infinite) collection of verifiable propositions, then their disjunction, $\cup\mathcal{H}$, is also verifiable.

O1 and O2 express the characteristic asymmetries of verifiability. Although it is possible to verify that bread nourishes for any finite number of days, it is not possible to verify that bread will continue to nourish into the indefinite future. Moreover, if H is verifiable, its negation may not be: if bread will cease to nourish, we will read about it in the newspaper; but nothing we can observe about bread will ever rule out the possibility of a future dereliction of duty.

Jointly, C1 and C2 ensure that the collection of all propositions of type F1 constitute the *closed* sets of a *topological space*.[11] The collection of all propositions of type V1 constitute the *open* sets of the topology; and the propositions of type D1 constitute the *clopen* sets of the topology. Sets of greater topological complexity are formed by set-theoretic operations on open and closed sets. The central point of Kelly [30] is that degrees of methodological accessibility correspond exactly to increasingly ramified levels of topological complexity, corresponding to elements of the *Borel hierarchy.* Roughly speaking, the Borel complexity of a hypothesis measures how complex it is to construct the hypothesis out of logical combinations of verifiable and falsifiable propositions. Higher levels of Borel complexity correspond to *inductive* notions of methodological success, where by inductive we mean any success notion where the chance of error is unbounded in the short run (see Figure 1).

Taking falsifiability of type F1 as the fundamental notion, the view sketched above was worked out in its essentials by Kelly [30] and further generalized by de Brecht and Yamamoto [31], Genin and Kelly [32] and Baltag et al. [33]. Genin and Kelly [34] exhibit a topology on probability measures in which the closed sets are exactly the propositions falsifiable in the sense of F2. That shows that the structure of statistical verifiability is *also* topological, at least when issues of monotonicity are ignored. The characteristic asymmetries are all present: while it is possible to falsify that the coin is fair, it is not possible to falsify that it is biased.

In this work, we show that hypotheses of type F3 also enjoy a topological structure; in fact, under a weak assumption, every hypothesis falsifiable in the sense of F2 is also falsifiable in the sense of F3 (Theorem A1). Imposing the demands of monotonicity does not make any fewer hypotheses falsifiable. That result provides a kind of cross-check on our notion of statistical falsifiability: it enjoys the same algebraic closure properties as the traditional notion F1, familiar from classical philosophy of science.

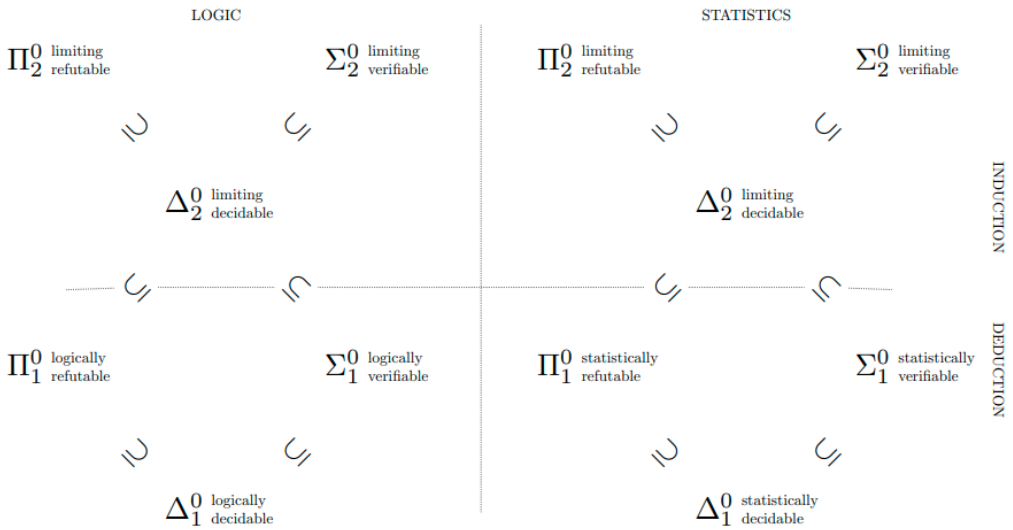

Figure 1. Pictured is a hierarchy of topological complexity and corresponding notions of methodological success. The set of all open sets is referred to as Σ^0_1; the set of all closed sets as Π^0_1; and the set of all clopen sets as $\Delta^0_1 = \Sigma^0_1 \cap \Pi^0_1$. Depending on whether the Σ^0_1 sets are propositions of type V1 or V2, we get the logical (**left**) and statistical (**right**) hierarchies. Sets of greater complexity are built out of Σ^0_1 sets by logical operations, e.g., Σ^0_2 sets are countable unions of locally closed sets. Inclusion relations between notions of complexity are also indicated. For more on the notions of methodological success characterized by higher levels of Borel complexity, see Genin and Kelly [34].

4. The Statistical Setting

4.1. Models, Measures and Samples

A *model* characterizes the contextually relevant features of a data-generating process. Models may be quite simple, as when they specify the bias of a coin. Models may also be elaborate structural hypotheses, as when they specify a set of structural equations expressing the causal processes by which data are generated. Inquiry typically begins with the specification of a set of possible models $\mathcal{M}$, any one of which, for all the inquirers know, may characterize the true data-generating process. Each model $\theta \in \mathcal{M}$ determines a *probability measure*, μ_θ over a common *sample space* $\mathfrak{S} = (\Omega, \mathcal{F})$. A sample space is a set of possible random samples Ω, together with a σ-algebra of events $\mathcal{F} \subseteq \mathcal{P}(\Omega)$ over the set of samples. Each probability measure μ_θ is a possible assignments of probabilities to events in $\mathcal{F}$ that is consistent with the axioms of probability and with the constraints specified by the model θ. We let W be the set $\{\mu_\theta : \theta \in \mathcal{M}\}$, i.e., the set of all probability measures on the sample space $\mathfrak{S}$ generated by possibilities in $\mathcal{M}$.

A set of models $\mathcal{M}$ is said to be *identified* if the map $\theta \mapsto \mu_\theta$ is one-to-one, i.e., if no two models generate the same probability measure over the space of observable outcomes $\mathfrak{S}$. If two models θ, θ' generate the same probability measure over the observables, then no amount of random sampling can possibly distinguish them and the question of whether θ or θ' is the generating mechanism is, in a sense, hopeless from a statistical perspective.[12] In the following, we will assume that the set of models is identified. Since each probability measure uniquely determines a model, we can identify the model with its probabilistic consequences and drop the subscript θ from our notation.

Call the triple $(W, \Omega, \mathcal{F})$, consisting of a set of probability measures W on a sample space $(\Omega, \mathcal{F})$, a *statistical setup.* The inquirer observes events in $\mathcal{F}$, but the hypotheses she is interested in are typically propositions over W. In paradigmatic cases, an event A in $\mathcal{F}$ is logically consistent with all probability measures in W.

Although measures in W assign probabilities to *all* events in $\mathcal{F}$, not all these events are the kinds of events than an agent can observe. No one can tell whether a real-valued sample point is rational or irrational, or if it is exactly π. For another example, suppose that region A is the closed interval $[1/2, \infty]$, and that the sample ω happens to land right on the boundry-point $1/2$ of A. Suppose, furthermore, that given enough time and computational power, the sample ω can be specified to arbitrary, finite precision. No finite degree of precision: $\omega \approx 0.50$; $\omega \approx 0.500$; $\omega \approx 0.5000$; ... suffices to determine that ω is truly in A. But the mere possibility of a sample hitting the boundary of A does not matter statistically, if the chance of obtaining such a sample is zero, as it typically is, unless there is discrete probability mass on the point $\frac{1}{2}$. Both these examples hinge on an implicit underlying topology $\mathcal{T}$ on the sample space Ω.

The topology $\mathcal{T}$ the sample space reflects what is formally verifiable about the sample itself. For example, if Ω is the real line, then it is typically verifiable whether a sample point ω is greater or less than $r \in \mathbb{R}$; and whether it lies in the interval (r, s). It is typically not verifiable whether the sample ω is rational or irrational, or whether it is exactly π. We assume that these formally verifiable propositions over Ω are closed under finite conjunctions and arbitrary disjunctions and therefore, form the open sets of a topological space. Then, the formally decidable proposition in $\mathcal{F}$ form the clopen sets of the same topological space. Furthermore, we assume that the σ-algebra $\mathcal{F}$ contains all the verifiable proposition $\mathcal{T}$, and in fact that it is the least such σ-algebra, i.e., that it is generated by closing the verifiable propositions under countable unions and negations. A σ-algebra that arises from a topology in this way is called a *Borel algebra* and its members are called *Borel sets*.

A Borel set A for which $\mu(\mathsf{bdry}A) = 0$ is said to be *almost surely clopen (decidable) in* μ.[13] Say that a collection of Borel sets $\mathcal{S}$ is almost surely clopen in μ iff every element of $\mathcal{S}$ is almost surely clopen in μ. We say that a Borel set A is almost surely decidable iff it is almost surely decidable in every μ in W. Similarly, we say that a collection of Borel sets $\mathcal{S}$ is almost surely clopen iff every element of $\mathcal{S}$ is almost surely clopen.

A *topological basis* $\mathcal{I}$ on Ω is a collection of subsets of Ω such that (1) the elements of $\mathcal{I}$ cover Ω; and (2) if E, F are elements of $\mathcal{I}$, then for each $\omega \in E \cap F$, there is $G \in \mathcal{I}$ containing ω and contained in $E \cap F$. Closing a topological basis under unions generates a topology, and every topology is generated by some basis. In the following, we will assume that the topology $\mathcal{T}$ on Ω is generated by a basis $\mathcal{I}$ that is at most countably infinite and almost surely clopen. Say that a statistical setup $(W, \Omega, \mathcal{F})$ is *feasibly based* if $\mathcal{F}$ is a Borel σ-algebra arising from a countable, almost surely clopen basis. That assumption is satisfied, for example, in the standard case in which the worlds in W are Borel measures on $\mathbb{R}^n$, and all measures are absolutely continuous with respect to Lebesgue measure, i.e., when all measures have probability density functions, which includes normal, chi-square, exponential, Poisson, and beta distributions. It is also satisfied for discrete distributions like the binomial, for which the topology on the sample space is the discrete (power set) topology, so every region in the sample space is clopen. It is satisfied in the particular cases of Examples 1 and 2.

Example 1. *Consider the outcome of a single coin flip. The set Ω of possible outcomes is $\{H, T\}$. Since every outcome is decidable, the appropriate topology on the sample space is $\mathcal{T} = \{\varnothing, \{H\}, \{T\}, \{H, T\}\}$, the discrete topology on Ω. Let W be the set of all probability measures assigning a bias to the coin. Since every element of $\mathcal{T}$ is clopen, every element is also almost surely clopen.*

Example 2. *Consider the outcome of a continuous measurement. Then, the sample space Ω is the set of real numbers. Let the basis $\mathcal{I}$ of the sample space topology be the usual interval basis on the real numbers. That captures the intuition that it is verifiable that the sample landed in some open interval, but it is not verifiable that it landed exactly on the boundary of an open interval. There are no non-trivial decidable (clopen) propositions in that topology. However, in typical statistical applications, W contains only probability measures μ that assign zero probability to the boundary of an arbitrary open interval. Therefore, every open interval E is almost surely decidable, i.e., $\mu(\mathsf{bdry}(E)) = 0$.*

Product spaces represent the outcomes of repeated sampling. Let I be an index set, possibly infinite. Let $(\Omega_i, \mathcal{T}_i)_{i \in I}$ be sample spaces, each with basis $\mathcal{I}_i$. Define the *product* $(\Omega, \mathcal{T})$ of the $(\Omega_i, \mathcal{T}_i)$ as follows: let Ω be the Cartesian product of the Ω_i; let $\mathcal{T}$ be the product topology, i.e., the topology in which the open sets are unions of Cartesian products $\times_{i \in I} O_i$, where each O_i is an element of $\mathcal{T}_i$, and all but finitely many O_i are equal to Ω_i. When I is finite, the Cartesian products of basis elements in $\mathcal{I}_i$ are the intended basis for $\mathcal{T}$. Let $\mathcal{B}$ be the σ-algebra generated by $\mathcal{T}$. Let μ_i be a probability measure on $\mathcal{B}_i$, the Borel σ-algebra generated by the $\mathcal{T}_i$. The *product measure* $\mu = \times_{i \in I} \mu_i$ is the unique measure on $\mathcal{B}$ such that, for each $B \in \mathcal{B}$ expressible as a Cartesian product of $B_i \in \mathcal{B}_i$, where all but finitely many of the B_i are equal to Ω_i, $\mu(B) = \prod \mu_i(B_i)$. (For a simple proof of the existence of the infinite product measure, see Saeki [37].) Let $\mu^{|I|}$ denote the $|I|$-fold product of μ with itself. If W is a set of measures, let $W^{|I|}$ denote the set $\{\mu^{|I|} : \mu \in W\}$. If $\mathcal{B}$ is a Borel σ-algebra generated by $\mathcal{T}$, let $\mathcal{B}^{\otimes n}$ be the Borel σ-algebra generated by n-fold product of $\mathcal{T}$ with itself.

4.2. Statistical Tests

A statistical *method* is a measurable function ϕ from random samples to propositions over W, i.e., for every A in the range of ϕ, its preimage $\phi^{-1}(A)$ is an element of $\mathcal{F}$. A *test* of a statistical hypothesis $H \subseteq W$ is a statistical method $\psi : \Omega \to \{W, H^{\mathsf{c}}\}$. Call $\psi^{-1}(W)$ the *acceptance region*, and $\psi^{-1}(H^{\mathsf{c}})$ the *rejection region* of the test.[14] The *power* of test $\psi(\cdot)$ is the worst-case probability that it rejects truly, i.e., $\inf_{\mu \in H^{\mathsf{c}}} \mu[\psi^{-1}(H^{\mathsf{c}})]$. The *significance level* of a test is the worst-case probability that it rejects falsely, i.e., $\sup_{\mu \in H} \mu[\psi^{-1}(H^{\mathsf{c}})]$.

A test is *feasible in* μ iff its acceptance region is almost surely decidable in μ. Say that a test is *feasible* iff it is feasible in every world in W. More generally, say that a method is *feasible* iff the preimage of every element of its range is almost surely decidable in every world in W. Tests that are not feasible in μ are impossible to implement—as described above, if the acceptance region is not almost surely clopen in μ, then with non-zero probability, the sample lands on the boundary of the acceptance region, where one cannot decide whether to accept or reject. If one were to draw a conclusion at some finite stage, that conclusion might be reversed in light of further computation. Tests are supposed to *solve* inductive problems, not to generate new ones.

Considerations of feasibility provide a new perspective on the assumption that appears throughout this work: that the basis $\mathcal{I}$ is almost surely clopen. If that assumption fails, then it is not an *a priori* matter whether geometrically simple zones are suitable acceptance zones for statistical methods. But if that is not determined *a priori*, then presumably it must be investigated by statistical means. That suggests a methodological regress in which we must use statistical methods to decide which statistical methods are feasible to use. Therefore, we consider only feasible methods in the following development.

4.3. The Weak Topology

A sequence of measures (μ_n) *converges weakly* to μ, written $\mu_n \Rightarrow \mu$, iff $\mu_n(A) \to \mu(A)$, for every A almost surely clopen in μ. It is immediate that $\mu_n \Rightarrow \mu$ iff for every μ-feasible test $\psi(\cdot)$, $\mu_n(\psi \text{ rejects}) \to \mu(\psi \text{ rejects})$. It follows that no feasible test of $H = \{\mu\}$ achieves power strictly greater than its significance level. Furthermore, every feasible method that correctly infers H with high chance in μ, exposes itself to a high chance of error in "nearby" μ_n.

It is a standard fact that one can topologize W in such a way that weak convergence is exactly convergence in the topology.[15] That topology is called the *weak topology*, n.b.: the weak topology is a topology on *probability measures*, whereas all previously mentioned topologies were topologies on *random samples*. In other words, the open sets of the weak topology are propositions over W, not over Ω. If we interpret the closure operator in terms of the weak topology, $\mu \in \mathsf{cl}A$ iff there is a sequence (μ_n) lying in A such that $\mu_n \Rightarrow \mu$.

If A is an almost surely decidable event in the appropriate σ-algebra then it is immediate from the definition of weak convergence that $\{\mu^n : \mu^n(A) > r\}$ and $\{\mu^n : \mu^n(A) < r\}$ are open in the weak topology over W^n. In this case, it is also true that $\{\mu : \mu^n(A) > r\}$ and $\{\mu : \mu^n(A) < r\}$ are open in the weak topology over W. That observation will often be appealed to in the following. For a proof, see Theorem 2.8 in Billingsley [38].

5. Statistical Falsifiability

Hypothesis H was said to be falsifiable (in the sense of F1) if there is an error avoiding method that converges on increasing information to not-H iff H is false. That condition implies that there is a method that achieves *every* bound on *chance* of error, and converges to not-H iff H is false. In statistical settings, one cannot insist on such a high standard of infallibility. Instead, we say that H is *falsifiable in chance* iff for every bound on error, there is a method that achieves it, and that converges in probability to not-H iff H is false. The reversal of quantifiers expresses the fundamental difference between statistical and propositional falsifiability. The central point in this section is encapsulated in Theorem 1: for feasibly based statistical setups, the statistically falsifiable propositions (in the sense of F2) are exactly the closed sets in the weak topology.

Say that a family $(\lambda_n)_{n\in\mathbb{N}}$ of feasible tests of H is an *α-falsifier in chance* of $H \subseteq W$ iff for all $n \in \mathbb{N}$:

BndErr. $\mu^n[\lambda_n^{-1}(W)] \geq 1 - \alpha$, if $\mu \in H$;
LimCon. $\mu^n[\lambda_n^{-1}(H^{\mathsf{c}})] \xrightarrow{n} 1$, if $\mu \in H^{\mathsf{c}}$.

Say that H is *α-falsifiable in chance* iff there is an α-falsifier in chance of H. Say that H is *falsifiable in chance* iff H is α-falsifiable in chance for every $\alpha > 0$.

Several strengthenings of falsifiability in chance immediately suggest themselves. One could demand that, in addition to BndErr, the chance of error vanishes to zero:

VanErr. $\mu^n[\lambda_n^{-1}(W)] \xrightarrow{n} 1$, if $\mu \in H$.

A strengthening of this requirement will be taken up in the following section.

Defining statistical verifiability requires no new ideas. Say that H is *α-verifiable in chance* iff there is an α-falsifier in chance of H^{c}. Say that H is *verifiable in chance* iff H^{c} is α-falsifiable in chance for every $\alpha > 0$.

The central theorem of Genin and Kelly [34] states that, for feasibly based statistical setups, falsifiability in chance is equivalent to being closed in the weak topology.

Theorem 1 (Fundamental Characterization Theorem). *Suppose that the statistical setup $(W, \Omega, \mathcal{F})$ is feasibly based. Then, for $H \subseteq W$, the following are equivalent:*

1. *H is α-falsifiable in chance for some $\alpha > 0$;*
2. *H is falsifiable in chance;*
3. *H is closed in the weak topology on W.*

Since a hypothesis is verifiable iff its complement is refutable, the characterization of statistical verifiability follows immediately.

Since a topological space is determined uniquely by its closed sets, Theorem 1 implies that the weak topology is the *unique* topology that characterizes statistical falsifiability (in the sense of F2), at least under the weak conditions stated in the antecedent of the theorem. Thus, under those conditions, the weak topology is not merely a convenient formal tool; it is *the* fundamental topology for statistical inquiry.

6. Monotonic Falsifiability

In Section 5, we said that (λ_n) is a statistical falsifier of H if it converges to not-H if H is false, and otherwise has a small chance of drawing an erroneous conclusion. But that standard is consistent with a wild see-sawing in the chance of producing the informative conclusion not-H as sample sizes increase, even if H is false. Of course, it is desirable that the chance of correctly rejecting H increases with the sample size, i.e., that for all $\mu \in H^{\mathsf{c}}$ and $n_1 < n_2$,

MON. $\mu^{n_2}[\lambda_{n_2}^{-1}(H^{\mathsf{c}})] > \mu^{n_1}[\lambda_{n_1}^{-1}(H^{\mathsf{c}})]$.

Failing to satisfy MON has the perverse consequence that collecting a larger sample might be a bad idea. Researchers would have to worry whether a failure of replication was due merely to a clumsily designed statistical method that converges to the truth along a needlessly circuitous route. Unfortunately, MON is infeasible in typical cases, so long as we demand that verifiers satisfy VANERR. Lemma 1 expresses that misfortune.

Say that a statistical setup $(W, \Omega, \mathcal{I})$ is *purely statistical* iff for all $\mu, \nu \in W$ and all events $B \in \mathcal{B}^{\otimes n}$ such that $\mu^n(\mathsf{bdry}B) = 0$, $\mu^n(B) > 0$ iff $\nu^n(B) > 0$. This is the formal expression of the idea that almost surely clopen sample events have no logical bearing on statistical hypotheses and is easily seen to be a weakening of mutual absolute continuity.

Lemma 1. *Suppose that the statistical setup $(W, \Omega, \mathcal{F})$ is feasibly based and purely statistical. Let H be closed, but not open in the weak topology. If (λ_n) is an α-falsifier in chance of H, then, (λ_n) satisfies* VANERR *only if it does not satisfy* MON.

Proof of Lemma 1. Suppose for a contradiction that (λ_n) is an α-falsifier in chance of H satisfying VANERR. Since H is not open, there is $\nu \in H \cap \mathsf{cl}(H^{\mathsf{c}})$. Since the statistical setup is purely statistical and (λ_n) satisfies LIMCON, there must be a sample size n_1 such that $0 < \alpha_{n_1} := \nu^{n_1}[\lambda_{n_1}^{-1}(H^{\mathsf{c}})]$. Let $O' = \{\mu : \alpha_{n_1} - \epsilon < \mu^{n_1}[\lambda_{n_1}^{-1}(H^{\mathsf{c}})]\}$. By construction, $\nu \in O'$. Since (λ_n) satisfies VANERR, there is $n_2 > n_1$ such that $\nu^{n_2}[\lambda_{n_2}^{-1}(H^{\mathsf{c}})] < \alpha_{n_1} - \epsilon$. Let $O'' = \{\mu : \alpha_{n_1} - \epsilon > \mu^{n_2}[\lambda_{n_2}^{-1}(H^{\mathsf{c}})]\}$. By construction, $\nu \in O := O' \cap O''$. Since (λ_n) is feasible, both O', O'' are open in the weak topology (see the observation at the end of Section 4.3). Since open sets are closed under finite conjunction, O is open in the weak topology. But since $\nu \in \mathsf{cl}(H^{\mathsf{c}})$, there is $\mu \in H^{\mathsf{c}} \cap O$. But then, $\mu^{n_1}[\lambda_{n_1}^{-1}(H^{\mathsf{c}})] > \alpha_{n_1} - \epsilon$, whereas $\mu^{n_2}[\lambda_{n_2}^{-1}(H^{\mathsf{c}})] < \alpha_{n_1} - \epsilon$. Therefore, (λ_n) does not satisfy MON. □

But even if strict monotonicity of power is infeasible, it ought to be our regulative ideal. Say that an α-falsifier $(\lambda_n)_{n\in\mathbb{N}}$ of H, whether in chance, or almost sure, is *α-monotonic* iff for all $\mu \in H^{\mathsf{c}}$ and $n_1 < n_2$:

α-MON $\mu^{n_2}[\lambda_{n_2}^{-1}(H^{\mathsf{c}})] + \alpha > \mu^{n_1}[\lambda_{n_1}^{-1}(H^{\mathsf{c}})]$.

Satisfying α-MON ensures that collecting a larger sample is not a disastrously bad idea. Surprisingly, some standard hypothesis tests fail to satisfy even this weak requirement. Chernick and Liu [23] noticed non-monotonic behavior in the power function of textbook tests of the binomial proportion, and proposed heuristic software solutions. The test exhibited in the Genin and Kelly's proof of Theorem 1 also displays dramatic non-monotonicity (Figure 2). Others have raised worries of non-monotonicity in consumer safety regulation, vaccine studies, and agronomy [24–26].

We now articulate a notion of statistical falsifiability that requires α-monotonicity. Write $a_n \downarrow 0$ if the sequence (a_n) converges monotonically to zero. Say that a family $(\lambda_n)_{n\in\mathbb{N}}$ of feasible tests of $H \subseteq W$ is an *α-monotonic falsifier* of H iff

MVanErr. For all $\mu \in H$, there exists a sequence (α_n) such that each $\alpha_n \leq \alpha$, $\alpha_n \downarrow 0$, and $\mu^n[\lambda_n^{-1}(H^c)] \leq \alpha_n$;
LimCon. For all $\mu \in H^c$, $\mu^n[\lambda_n^{-1}(H^c)] \xrightarrow{n} 1$;
α-Mon. For all $\mu \in W$, $\mu^{n_2}[\lambda_{n_2}^{-1}(H^c)] + \alpha > \mu^{n_1}[\lambda_{n_1}^{-1}(H^c)]$, if $n_1 < n_2$.

Say that H is *α-monotonically falsifiable* iff there is an α-monotonic falsifier of H. Say that H is *monotonically falsifiable* iff H is α-monotonically falsifiable, for every $\alpha > 0$. It is clear that every α-monotonic falsifier of H is also an α-falsifier in chance. However, not every α-monotonic falsifier of H is an almost sure α-falsifier. The converse also does not hold.

Defining monotonic verifiability requires no new ideas. Say that H is *α-monotonically verifiable in chance* iff there is an α-monotonic falsifier of H^c. Say that H is *montonically verifiable* iff H^c is α-monotonically refutable for every $\alpha > 0$.

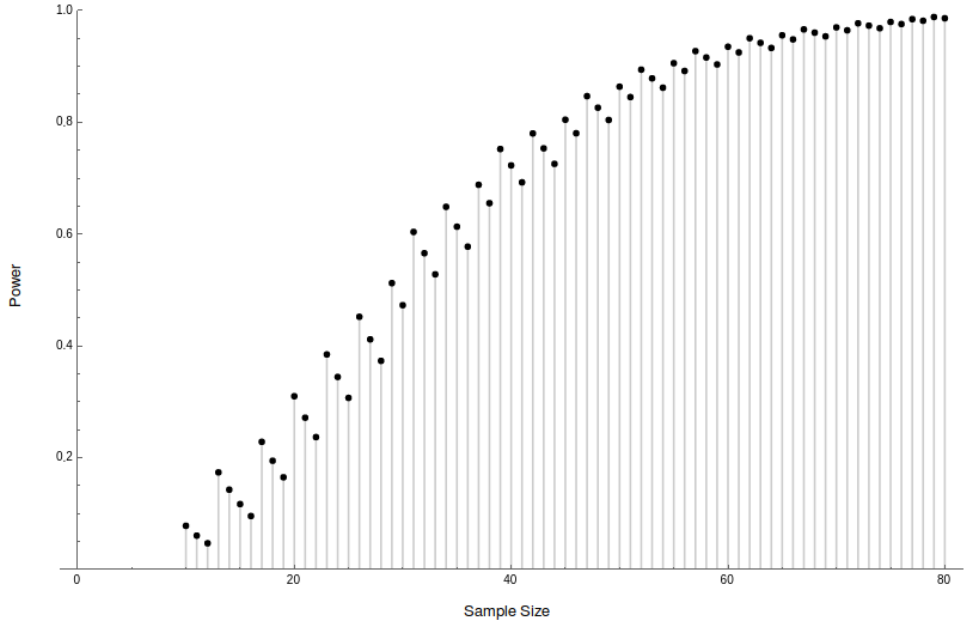

Figure 2. Diachronic plot of power of typical test of the null hypothesis that a coin is not head-biased, when $p(H) = 0.775$. The plot exhibits the characteristic "saw-tooth" shape identified by Chernick and Liu [23]. Note that the drops in power are significant e.g., >0.07 between sample sizes 31 and 33.

The central theorem of this section states that every statistically falsifiable hypothesis is also monotonically falsifiable. The proof is provided in the Appendix **??**.

Theorem 2. *Suppose that the statistical setup $(W, \Omega, \mathcal{F})$ is feasibly based. Then, for $H \subseteq W$, the following are equivalent:*

1. *H is α-falsifiable in chance for some $\alpha > 0$;*
2. *H is statistically falsifiable;*
3. *H is monotonically falsifiable;*
4. *H is closed in the weak topology on W.*

The characterization of monotonic verifiability follows immediately.

7. Conclusions: Falsifiability and Induction

This work is inspired by Popper's falsificationism and its difficulties with statistical hypotheses. We have proposed several different notions of statistical falsifiability and proven that, whichever definition we prefer, the same hypotheses turn out to be falsifiable. That shows that the notion enjoys a kind of conceptual robustness. Finally, we have demonstrated that, under weak assumptions, the statistically falsifiable hypotheses correspond precisely to the closed sets in a standard topology on probability measures. That means that standard techniques from statistics and measure theory can be used to determine exactly which hypotheses are statistically falsifiable. Hopefully, this result will be a boon to statistical practice by providing a simple diagnostic for statistical falsifiability—controversies about testability may be resolved while maintaining an ecumenical attitude about what, precisely, statistical falsification consists in.

However, this work should not be taken as a wholesale endorsement of Popperian falsificationism. It is easy to generate respectable statistical hypotheses that are not falsifiable or, for that matter, verifiable. Indeed, many pressing scientific hypotheses are of this nature, especially when researchers are interested in answering causal questions (see Genin [39]). These should by no means be regarded as unscientific. The hallmark of these kinds of hypotheses is that, although it is possible to converge to the truth as sample sizes increase, it is not possible to do so with finite sample bounds on the probability of error. This fact should be faced squarely—and not necessarily by searching for stronger assumptions that make it possible to provide finite sample guarantees.

A large class of scientific problems can be reconstructed in the following way: we enumerate a collection of disjoint hypothesis $H_1, H_2, \ldots$ in such a way that that $\cup_{i=1}^{n} H_i$ is falsifiable no matter how large n is.[16] Then, we test longer and longer initial segments, conjecturing the first hypothesis that we fail to reject. Since the procedure involves nesting tests, we cannot expect finite sample bounds on the probability of conjecturing a false hypothesis. Nevertheless, this procedure answers pretty closely to a Popperian methodology of conjectures and refutations. Unlike Popper, we have no problem calling the outcome of such a procedure—belief in, or acceptance of, the first unrejected hypothesis in the enumeration—an induction. But is there anything to be said in favor of this natural and commonplace scientific procedure? Popper hoped that it would produce theories of greater and greater *truthlikeness* or *verisimilitude*. That notion has had a troubled and fascinating career, which we do not review here.[17] There is, so far as we know, no demonstration that this procedure must produce theories of increasing truthlikeness.[18]

Let us briefly consider a different idea. Call such a procedure *progressive* if (1) it converges in chance to the true H_i (if any is true) and (2) if H_i is true, then the objective chance of conjecturing H_i increases monotonically with sample size. It should come as no surprise that this kind of strict monotonicity is not usually feasible. However, it should be a regulative ideal: say that such a procedure is *α-progressive* if (1′) it converges in chance to the true H_i (if any is true) and (2′) if H_i is true, then the objective chance of conjecturing H_i never decreases by more than α as sample sizes increase. The latter is a natural generalization of α-MON to problems of theory choice and precludes egregious backsliding. It should be intuitive that such a procedure can only be α-progressive if the constituent test are themselves α-monotonic in chance. Genin [45] Theorem 3.6.3, shows that, for all $\alpha > 0$ it is possible to produce an α-progressive solution to this problem, so long as the consitutent tests are chosen to be sufficiently monotonic in chance. The existence of such tests is guaranteed by Theorem 2. That means that a carefully calibrated methodology of conjectures and refutations is conducive to progress—it converges to the right answer (if any candidate is right) with an arbitrarily low degree of backsliding. We hope that these remarks suggest to some degree the relevance of this work to a positive theory of induction.

Funding: This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC number 2064/1—Project number 390727645.

Acknowledgments: This work owes a great deal to Kevin Kelly, with whom I had the good fortune of completing my philosophical apprenticeship. I am also grateful to Franz Huber, Conor Mayo-Wilson, Samuel Fletcher, Jonathan Livengood, Kino Zhao and several anonymous reviewers for their generous attention and invaluable input.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A. Proof of Theorem 2

In order to prove Theorem 2, we introduce yet another strengthening of statistical falsifiability. The following strengthens BNDERR to the requirement that the total chance of error is bounded:

σ-BNDERR. $\sum_{n=1}^{\infty} \mu^n[\lambda_n^{-1}(W)] \geq 1 - \alpha$, if $\mu \in H$.

That implies both BNDERR and VANERR, but also, by the Borel-Cantelli lemma, that with probability one, a verifier makes only finitely many errors on an infinite sample path, whenever H is true:

SVANERR. $\mu^{\infty}[\liminf \lambda_n^{-1}(W)] = 1$, if $\mu \in H$.

We might demand similar asymptotic behavior when H is false. Say that a family $(\lambda_n)_{n\in\mathbb{N}}$ of feasible tests of $H \subseteq W$ is an *almost sure α-falsifer* of H iff

σ-BNDERR. $\sum_{n=1}^{\infty} \mu^n[\lambda_n^{-1}(W)] \geq 1 - \alpha$, if $\mu \in H$, and
SLIMCON. $\mu^{\infty}\left[\liminf \lambda_n^{-1}(H^{\mathsf{c}})\right] = 1$, if $\mu \in H^{\mathsf{c}}$.

Say that H is *almost surely α-falsifiable* iff there is an almost sure α-falsifier of H. Say that H is *almost surely falsifiable* iff H is almost surely α-falsifiable, for every $\alpha > 0$. As usual, we say that H is almost surely verifiable iff its complement is almost surely falsifiable.
Since almost sure convergence entails convergence in chance, almost sure falsifiable entails falsifiability in chance.[19]

In this section, we prove the following, from which Theorem 2 follows immediately.

Theorem A1. *Suppose that the statistical setup $(W, \Omega, \mathcal{F})$ is feasibly based. Then, for $H \subseteq W$, the following are equivalent:*

1. *H is α-verifiable in chance for some $\alpha > 0$;*
2. *H is monotonically verifiable;*
3. *H is almost surely verifiable;*
4. *H is open in the weak topology on W.*

In Genin and Kelly [34], it is proven that 1, 3 and 4 are equivalent (see their Theorem 4.1). The fact that 2 implies 1 is immediate from the definitions. To prove that 4 implies 2, we proceed largely as in Genin and Kelly [34], but with greater attention to details.

Recall that a statistical setup $(W, \Omega, \mathcal{F})$ is *feasibly based* if $\mathcal{F}$ is a Borel σ-algebra arising from $\mathcal{I}$, a countable, almost surely clopen basis. Let $\mathcal{A}$ be the algebra generated by $\mathcal{I}$. Genin and Kelly [34] show that when the statistical setup is feasibly based, the collection of sets of the form $\{\mu : \mu(A) > a\}$, where $a \in \mathbb{Q}$ and $A \in \mathcal{A}$, is a sub-basis for the weak topology. That means that every open set in the weak topology over W can be expressed as a union of finite intersections of elements of the collection. Therefore, we proceed by showing that every element of that collection is monotonically verifiable (Lemma A2). We conclude by showing that the monotonically verifiable propositions are closed under finite intersections (Lemma A3), and countable unions (Lemma A4), which completes the proof. But first, we prove the somewhat technical Lemma A1, which says that if you have a countable collection of α_i-monotonic verifiers $(\lambda_n^1), (\lambda_n^2), \ldots$, of hypotheses $A_1, A_2, \ldots$, then

it is possible to construct a countable collection of *mutually independent* monotonic verifiers for the A_i. The idea behind the construction is very rudimentary: you can always render methods independent by splitting up the sample in the appropriate way.

Lemma A1. *Suppose that I is a countable index set and that for $i \in I$, $(\psi_n^i)_{n\in\mathbb{N}}$ is an α_i-monotonic verifier of A_i. Then, there exist $(\lambda_n^1), (\lambda_n^2), \ldots$ where each $(\lambda_n^i)_{n\in\mathbb{N}}$ is an α_i-monotonic verifier of A_i, and for each n, $(\sigma(\lambda_n^i), i \in I)$ are mutually independent.*

Proof of Lemma A1. The basic idea is to render the (ψ_n^i) mutually independent by splitting the sample and feeding it to the individual verifiers according to a triangular dovetailing scheme. Represented in tabular form:

$$\begin{array}{ccccc} (\psi_n^1) & (\psi_n^2) & (\psi_n^3) & (\psi_n^4) & \cdots \\ \\ \omega_1 & & & & \\ \omega_2 & \omega_3 & & & \\ \omega_4 & \omega_5 & \omega_6 & & \\ \omega_7 & \omega_8 & \omega_9 & \omega_{10} & \\ \omega_{11} & \omega_{12} & \omega_{13} & \omega_{14} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \cdots \end{array}$$

The samples in the ith column of the table are the samples that are fed to the ith verifier. Since the verifiers are essentially functions of disjoint samples, they are mutually independent. Since the samples are i.i.d. all of the desirable statistical properties of the verifiers are preserved. For a detailed proof, including all the measure-theoretic niceties, see Lemma 3.3.2 in Genin [45]. □

Lemma A2. *Suppose that B is almost surely decidable for every $\mu \in W$. Then, for all real b, the hypothesis $H = \{\mu : \mu(B) > b\}$ is monotonically verifiable.*

Proof of Lemma A2. We restrict attention to the non-trivial cases where $b \in (0,1)$. The idea is to take an almost sure α-verifier of H and modify it slightly to ensure α-monotonicity. See Figure A1 for an overview of the proof strategy. Let $t_n = \sqrt{\frac{1}{2n}\ln(\pi^2 n^2/6\alpha)}$ and

$$\lambda_n(\vec{\omega}) := \begin{cases} H, & \text{if } \sum_{i=1}^n \mathbb{1}_B(\omega_i) \geq \lceil n(b+t_n)\rceil, \\ W, & \text{otherwise.} \end{cases}$$

Genin and Kelly [34] Theorem 4.1, show that $(\lambda_n)_{n\in\mathbb{N}}$ is an a.s. α-verifier. Let

$$\beta_n(\theta) = \sum_{\lceil n(b+t_n)\rceil}^{n} \binom{n}{k} \theta^k (1-\theta)^{n-k}.$$

Then, $\beta_n(\mu B) = \mu^n[\lambda_n^{-1}(H)]$. It is something of a nuisance that $\beta_n(\theta)$ is exactly zero for n such that $\lceil n(b+t_n)\rceil > n$. Since the t_n converge monotonically to 0, this is the case for only finitely many initial sample sizes n. For example, for $b = 0.5, \alpha = 0.05$, $\beta_n(\theta)$ is non-trivial for samples sizes larger than 20. Let $n_0 = \min\{n : \lceil n(b+t_n)\rceil \leq n\}$.

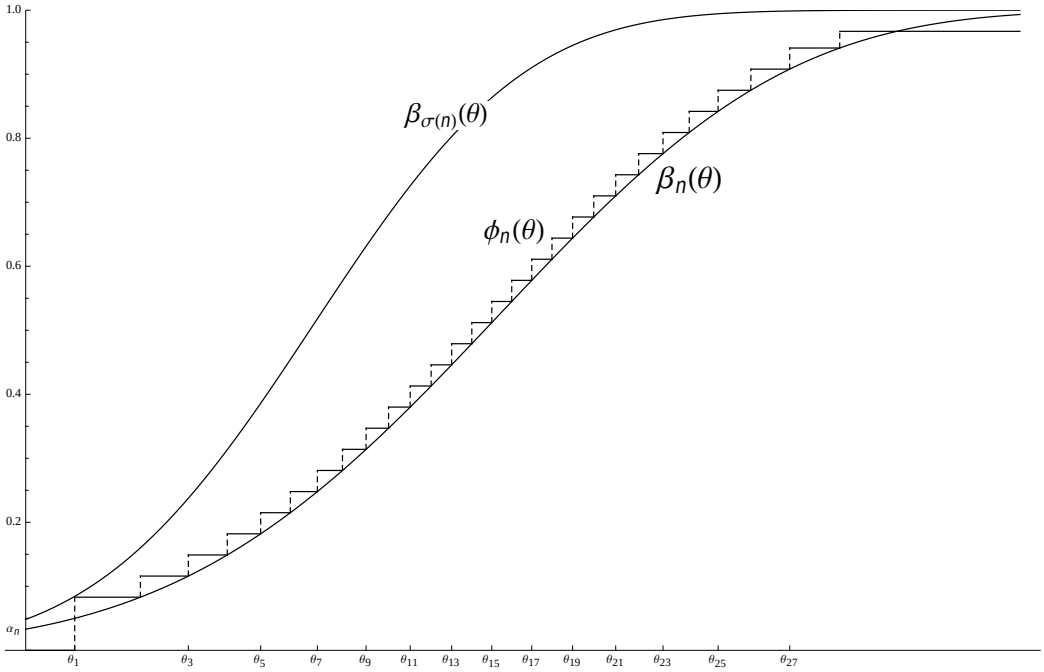

Figure A1. The basic idea of the proof is encapsulated in the figure. For any sample size n, we construct a step function ϕ_n that "almost dominates" the power function β_n. The step function dominates β_n for all θ except those for which $\beta_n(\theta)$ is less than α, or close to 1. Then, since the set of steps is finite, and the original test satisfies LIMCON, there must be a sample size $\sigma(n)$ such that the power function $\beta_{\sigma(n)}$ strictly dominates the step function. Since $\beta_{\sigma(n)}$ dominates the step function, $\beta_{\sigma(n)}(\theta)$ can only be less than $\beta_n(\theta)$ if $\beta_n(\theta)$ is less than α, or they are both close to 1. The loss of power from n to $\sigma(n)$ is thereby bounded by α. Iterating this process we get a sequence of "good" sample sizes $n, \sigma(n), \sigma^2(n), \ldots$ such that the power is "almost increasing". It remains only to interpolate the intermediate sample sizes with tests that throw out samples until they arrive at the nearest "good" sample size.

It is worth pointing out some additional features of the $\beta_n(\theta)$ that we will be appealing to in the following. It is evident that, since $\beta_n(\theta)$ is a polynomial, it is a continuous function of the parameter θ. Although it is "obviously" true that, for $n \geq n_0$, $\beta_n(\theta)$ is strictly increasing in θ, it is surprisingly non-trivial to demonstrate. For an elegant proof of this fact, see Gilat [46]. It is a standard fact of analysis that that if $[a, b]$, $[c, d]$ are closed real intervals and $f : [a, b] \to [c, d]$ is a continuous real function, then f is bijective iff it is strictly increasing. Therefore, for $n \geq n_0$, $\beta_n : [0, 1] \to [0, 1]$ is bijective.

There are two important properties of the collection $(\beta_n(\theta))_{n\in\mathbb{N}}$ that follow from the fact that (λ_n) is an almost sure verifier: (1) for $\theta > b$, $\lim_n \beta_n(\theta) = 1$; (2) $\sum_n \sup_{\theta\leq b} \beta_n(\theta) = \sum_n \beta_n(b) < \alpha$. The first property follows from LIMCON. The second property follows from σ-BNDERR and the fact that $\beta_n(\theta)$ is increasing. Define $\alpha_n = \beta_n(b)$. Note that for $n \geq n_0$, $\alpha_n > 0$. Let $\alpha_n^* = \min\{\alpha_m : n_0 \leq m \leq n\}$. For $n \geq n_0$, let

$$K_n = \max\{k \in \mathbb{N} : \alpha + k\alpha_n < 1 - \alpha_n^*\}.$$

Now, define the increasing step function:

$$\phi_n(\theta) = \begin{cases} 0, & \beta_n(\theta) < \alpha, \\ \alpha + k\alpha_n, & \alpha + (k-1)\alpha_n \leq \beta_n(\theta) < \alpha + k\alpha_n, \quad k \in \{1, \ldots, K_n\}, \\ 1 - \alpha_n^*, & \beta_n(\theta) \geq \alpha + K_n\alpha_n. \end{cases}$$

We first show that $\beta_n(\theta) - \phi_n(\theta) < \alpha$. If $\beta_n(\theta) < \alpha$, then $\beta_n(\theta) - \phi_n(\theta) \leq \beta_n(\theta) < \alpha$. Furthermore, $\phi_n(\theta) > \beta_n(\theta)$ for θ such that $\alpha \leq \beta_n(\theta) < \alpha + K_n\alpha_n$. Finally, for θ such that $\beta_n(\theta) \geq \alpha + K_n\alpha_n$, $\beta_n(\theta) - \phi_n(\theta) \leq 1 - \phi_n(\theta) = \alpha_n^* < \alpha$.

For $n \geq n_0$, let $\theta_{n,i} = \beta_n^{-1}(\alpha + i\alpha_n)$, for $i \in \{0, \ldots, K_n\}$. (The $\theta_{n,i}$ are well defined because $\beta_n(\theta)$ is surjective whenever $n \geq n_0$.) Note that since for each i, $\beta_n(\theta_{n,i}) \geq \alpha$, each $\theta_{n,i} > b$. Since $\lim_n \beta_n(\theta) \to 1$ for all $\theta > b$, there is a least $N > n$, such that for all $i \in \{0, \ldots, K_n\}$, $\beta_N(\theta_{n,i}) > \phi_n(\theta_{n,i})$. For all n, define $\sigma(n)$ to be the least such N. Define $\sigma^0(n) := n, \sigma^1(n) := \sigma(n)$, and $\sigma^m(n) := \sigma(\sigma^{m-1}(n))$.

We show that $\beta_{\sigma(n)}(\theta) \geq \phi_n(\theta)$. Suppose that $\theta < \theta_{n,0}$. Then $\beta_n(\theta) < \alpha$, and $\phi_n(\theta) = 0 \leq \beta_{\sigma(n)}(\theta)$. Suppose that $\theta_{n,i} \leq \theta < \theta_{n,i+1}$ for $i \in \{0, \ldots, K_n - 1\}$. Then, $\beta_{\sigma(n)}(\theta) \geq \beta_{\sigma(n)}(\theta_{n,i}) > \phi_n(\theta_{n,i}) = \phi_n(\theta)$. Finally, suppose that $\theta \geq \theta_{n,K_n}$. Then, similarly, $\beta_{\sigma(n)}(\theta) \geq \beta_{\sigma(n)}(\theta_{n,K_n}) > \phi_n(\theta_{n,K_n}) = \phi_n(\theta)$.

Now, we show that $\phi_{\sigma(n)}(\theta) \geq \phi_n(\theta)$. Suppose that $\beta_{\sigma(n)}(\theta) < \alpha$. Then, since, $\phi_n(\theta) < \beta_{\sigma(n)}(\theta)$, it follows that $\phi_n(\theta) < \alpha$, and therefore, $\phi_n(\theta) = 0 \leq \phi_{\sigma(n)}(\theta)$. Suppose that $\alpha \leq \beta_{\sigma(n)} < \alpha + K_{\sigma(n)}\alpha_{\sigma(n)}$. Then, $\phi_n(\theta) \leq \beta_{\sigma(n)}(\theta) < \phi_{\sigma(n)}(\theta)$. Finally, if $\beta_{\sigma(n)} \geq \alpha + K_{\sigma(n)}\alpha_{\sigma(n)}$, then $\phi_\sigma(n) = 1 - \alpha^*_{\sigma(n)} \geq 1 - \alpha_n^* \geq \phi_n(\theta)$.

It is now easy to show that $\beta_n(\theta) - \beta_{\sigma^m(n)}(\theta) < \alpha$, since $\beta_n(\theta) - \beta_{\sigma^m(n)}(\theta) \leq \beta_n(\theta) - \phi_{\sigma^{m-1}(n)}(\theta) \leq \beta_n(\theta) - \phi_n(\theta) < \alpha$. Therefore, for an increasing sequence of "good" sample sizes, $n, \sigma(n), \sigma^2(n), \ldots$, the verifier $\{\lambda_n\}$ is α-monotonic. Furthermore, since the sequence $(\beta_{\sigma^m(n)}(b))_{m\in\mathbb{N}}$ converges to zero, there exists a subsequence $(\beta_{n_i^*}(b))_{i\in\mathbb{N}}$ that converges monotonically to zero. We use these facts to construct a verifier that is α-monotonic, by patching over the "bad" sample sizes.

Let $\pi(n) = \max\{n_i^* : n_i^* \leq n\}$. Let $\lambda_n^*(\omega_1, \ldots, \omega_n) := \lambda_{\pi(n)}(\omega_1, \ldots, \omega_{\pi(n)})$. We have taken pains to ensure that (λ_n^*) satisfies α-MON. Since

$$\sup_{\mu\in H^c} \mu^n[\lambda_n^{-1}(H)] \leq \beta_{\pi(n)}(b) \downarrow_n 0,$$

(λ_n) satisfies MVANERR. Since (λ_n) is an almost sure verifier, the set $C = \{\omega \in \Omega^\infty : \lambda_n(\omega|_n) \to H\}$ has $\mu^\infty(C) = 1$ for every $\mu \in H$. But if the sequence $(\lambda_n(\omega|_n))$ converges to H then so does the subsequence $(\lambda_{\pi(n)}(\omega|_{\pi(n)}))$. Therefore, $C \subseteq C^* = \{\omega \in \Omega^\infty : \lambda_{\pi(n)}(\omega|\pi(n)) \to H\}$. Therefore, $\mu^\infty[\liminf(\lambda_n^*)^{-1}(H)] = 1$, if $\mu \in H$, and (λ_n^*) satisfies SLIMCON. A fortiori, it also satisfies LIMCON. Since $\alpha > 0$ was arbitrary, we are done. □

Lemma A3. *The monotonically verifiable propositions are closed under finite conjunctions.*

Proof of Lemma A3. By Lemma A1, it suffices to show that if $(\lambda_n^1), \ldots, (\lambda_n^k)$ are mutually independent α_i-monotonic verifiers of $A_1, \ldots, A_k$, with $\alpha > \sum_i \alpha_i$, then

$$\lambda_n(\vec{\omega}) = \begin{cases} \cap_{i=1}^k A_i, & \text{if } \lambda_n^i(\vec{\omega}) = A_i \text{ for all } i \in \{1, \ldots, k\}, \\ W, & \text{otherwise,} \end{cases}$$

is an α-monotonic verifier of $\cap_{i=1}^k A_i$.

Consider the case where $k = 2$. We first demonstrate that (λ_n) satisfies MVANERR. Suppose $\mu \notin A_1 \cap A_2$. Without loss of generality, suppose that $\mu \notin A_1$. By assumption there exists a sequence (α_n^1) such that $\alpha_n^1 < \alpha_1$, $\alpha_n \downarrow 0$ and $\mu^n[(\lambda_n^1)^{-1}(A_1)] < \alpha_n^1$. Noticing that

$$\begin{aligned} \mu^n[\lambda_n^{-1}(A_1 \cap A_2)] &= \mu^n[(\lambda_n^1)^{-1}(A_1) \cap (\lambda_n^2)^{-1}(A_2)] \\ &\leq \mu^n[(\lambda_n^1)^{-1}(A_1)] \\ &\leq \alpha_n^1, \end{aligned}$$

we see that (λ_n) also satisfies MVANERR.

Suppose that $\mu \in A_1 \cap A_2$. We demonstrate that (λ_n) satisfies LIMCON.

$$\begin{aligned}\mu^n[\lambda_n^{-1}(A_1 \cap A_2)] &= \mu^n[(\lambda_n^1)^{-1}(A_1) \cap (\lambda_n^2)^{-1}(A_2)] \\ &= 1 - \mu^n[(\lambda_n^1)^{-1}(W) \cup (\lambda_n^2)^{-1}(W)] \\ &\geq 1 - \mu^n[(\lambda_n^1)^{-1}(W)] + \mu^n[(\lambda_n^2)^{-1}(W)].\end{aligned}$$

But since $\mu^n[(\lambda_n^1)^{-1}(W) + \mu^n[(\lambda_n^2)^{-1}(W)] \to 0$, it follows that

$$\mu^n[\lambda_n^{-1}(A_1 \cap A_2)] \to 1.$$

It remains to show that $(\lambda_n)_{n\in\mathbb{N}}$ satisfies α-MON. To make the expressions more manageable, we make the following substitutions:

$$a_i^{n_j} = \mu^{n_j}[(\lambda_{n_j}^i)^{-1}(A_i)].$$

Under the assumption of independence:

$$\mu^{n_1}[\psi_{n_1}^{-1}(A_1 \cap A_2)] - \mu^{n_2}[\psi_{n_2}^{-1}(A_1 \cap A_2)] = a_1^{n_1}a_2^{n_1} - a_1^{n_2}a_2^{n_2}.$$

By assumption, either (i) $a_1^{n_1} - a_1^{n_2} < \alpha/2$ or (ii) $a_2^{n_1} - a_2^{n_2} < \alpha/2$. Assume, without loss of generality, that (ii). Then,

$$\begin{aligned}a_1^{n_1}a_2^{n_1} - a_1^{n_2}a_2^{n_2} &< a_1^{n_1}a_2^{n_1} - a_1^{n_2}(a_2^{n_1} - \alpha/2) \\ &= a_1^{n_1}a_2^{n_1} - a_1^{n_2}a_2^{n_1} + a_1^{n_2} \cdot \alpha/2 \\ &= a_2^{n_1}(a_1^{n_1} - a_1^{n_2}) + a_1^{n_2} \cdot \alpha/2 \\ &< a_2^{n_1} \cdot \alpha/2 + a_1^{n_2} \cdot \alpha/2 \\ &\leq \alpha.\end{aligned}$$

The case when $k > 2$ follows immediately by induction. □

Lemma A4. *The monotonically verifiable propostions are closed under countable disjunction.*

Proof of Lemma A4. By Lemma A1, it suffices to show that if $(\lambda_n^1), \ldots, (\lambda_n^k), \ldots$, are mutually independent α_i-monotonic verifiers of $A_1, \ldots, A_k, \ldots$, such that $\sum_{i=1}^{\infty} \alpha_i$ converges to α, then,

$$\lambda_n(\vec{\omega}) = \begin{cases} \cup_{i=1}^{\infty} A_i, & \text{if } \lambda_n^i(\vec{\omega}) = A_i \text{ for some } i \in \{1, \ldots, n\}, \\ W, & \text{otherwise,} \end{cases}$$

is an α-monotonic verifier of $\cup_{i=1}^{\infty} A_i$.

We first demonstrate that (λ_n) satisfies MVANERR. Suppose that $\mu \notin \cup_{i=1}^{\infty} A_i$. Then, for each i there is (α_n^i) such that $\alpha_n^i < \alpha_i$, $\alpha_n^i \downarrow 0$, and $\mu^n[(\lambda_n^i)^{-1}(A_i)] \leq \alpha_n^i$. Therefore,

$$\begin{aligned}\mu^n[\lambda_n^{-1}(\cup_{i=1}^{\infty} A_i)] &= \mu^n[\cup_{i=1}^{n}(\lambda_n^i)^{-1}(A_i)] \\ &\leq \sum_{i=1}^{n} \mu^n[(\lambda_n^i)^{-1}(A_i)] \\ &\leq \sum_{i=1}^{\infty} \alpha_n^i \leq \alpha.\end{aligned}$$

It remains to show that $S_n = \sum_{i=1}^{\infty} \alpha_n^i$ converges monotonically to zero as $n \to \infty$. Since $\alpha_n^i \geq \alpha_{n+1}^i$ for each i, we have that $S_n \geq S_{n+1}$. Therefore, the sequence (S_n) is decreasing and bounded below by zero. By the monotone convergence theorem, the sequence (S_n) converges to its infimum. We show that the infimum is zero. Let $\epsilon > 0$. Since the tail of a convergent series tends to zero, there is K such that $\sum_{i=K}^{\infty} \alpha_i < \epsilon/2$. Therefore,

$T_n = \sum_{i=K}^{\infty} \alpha_n^i < \epsilon/2$. Since $S_n = T_n + \sum_{i=1}^{K-1} \alpha_n^i$, and $\sum_{i=1}^{K-1} \alpha_n^i \to 0$ as $n \to \infty$, there is N such that $S_n < \epsilon$ for $n \geq N$. Since ϵ was arbitrary, we are done.

Suppose that $\mu \in \cup_{i=1}^{\infty} A_i$. We show that (λ_n) satisfies LIMCON. Since $\mu \in \cup_{i=1}^{\infty} A_i$, there is k such that $\mu \in A_k$. For $n \geq k$,

$$\begin{aligned}\mu^n[\lambda_n^{-1}(\cup_{i=1}^{\infty} A_i)] &= \mu^n[\cup_{i=1}^{n}(\lambda_n^i)^{-1}(A_i)] \\ &\geq \mu^n[(\lambda_n^k)^{-1}(A_k)].\end{aligned}$$

But since $\mu^n[(\lambda_n^k)^{-1}(A_k)] \to 1$, we have that $\mu^n[\lambda_n^{-1}(\cup_{i=1}^{\infty} A_i)] \to 1$.

It remains to show that (λ_n) satisfies α-MON. By the inclusion-exclusion formula:

$$\begin{aligned}&\mu^n[\lambda_n^{-1}(\cup_{i=1}^{\infty} A_i)] = \mu^n[\cup_{i=1}^{n}(\lambda_n^i)^{-1}(A_i)] = \\ &= \sum_{i=1}^{n} \mu^n[(\lambda_n^i)^{-1}(A_i)] - \sum_{i<j<n} \mu^n[(\lambda_n^i)^{-1}(A_i) \cap (\lambda_n^j)^{-1}(A_j)] \\ &+ \sum_{i<j<k<n} \mu^n[(\lambda_n^i)^{-1}(A_i) \cap (\lambda_n^j)^{-1}(A_j) \cap (\lambda_n^k)^{-1}(A_k)] + \cdots \\ &+ (-1)^{n-1} \mu^n[\cap_{i=1}^{n}(\lambda_n^i)^{-1}(A_i)].\end{aligned}$$

To make the expressions more manageable, we make the following substitutions:

$$a_i^n = \mu^n[(\lambda_n^i)^{-1}(A_i)].$$

Since the verifiers are mutually independent:

$$\mu^n[\lambda_n^{-1}(\cup_i A_i)] = \sum_{i=1}^{n} a_i^n - \sum_{i<j<n} a_i^n a_j^n + \cdots + (-1)^{n-1} a_1^n a_2^n \cdots a_n^n.$$

Or, in closed form:

$$\mu^n[\lambda_n^{-1}(\cup_i A_i)] = \sum_{j=1}^{n} \left((-1)^{j-1} \sum_{\substack{I \subset \{1,\ldots,n\} \\ |I|=j}} \prod_{i \in I} a_i^n \right).$$

Furthermore, for $n_1 < n_2$,

$$\begin{aligned}\mu^{n_2}[\lambda_{n_2}^{-1}(\cup_i A_i)] &= \mu^{n_2}[\cup_{i=1}^{n_2}(\lambda_{n_2}^i)^{-1}(A_i)] \\ &\geq \mu^{n_2}[\cup_{i=1}^{n_1}(\lambda_{n_2}^i)^{-1}(A_i)] \\ &= \sum_{j=1}^{n_1} \left((-1)^{j-1} \sum_{\substack{I \subset \{1,\ldots,n_1\} \\ |I|=j}} \prod_{i \in I} a_i^{n_2} \right).\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mu^{n_1}[\lambda_{n_1}^{-1}(\cup_i A_i)] - \mu^{n_2}[\lambda_{n_2}^{-1}(\cup_i A_i)] = \\
&= \sum_{j=1}^{n_1}\left((-1)^{j-1}\sum_{\substack{I\subset\{1,\ldots,n_1\}\\|I|=j}}\prod_{i\in I}a_i^{n_1}\right) - \sum_{j=1}^{n_2}\left((-1)^{j-1}\sum_{\substack{I\subset\{1,\ldots,n_2\}\\|I|=j}}\prod_{i\in I}a_i^{n_2}\right) \\
&\leq \sum_{j=1}^{n_1}\left((-1)^{j-1}\sum_{\substack{I\subset\{1,\ldots,n_1\}\\|I|=j}}\prod_{i\in I}a_i^{n_1}\right) - \sum_{j=1}^{n_1}\left((-1)^{j-1}\sum_{\substack{I\subset\{1,\ldots,n_1\}\\|I|=j}}\prod_{i\in I}a_i^{n_2}\right) \\
&= \sum_{j=1}^{n_1}\left((-1)^{j-1}\sum_{\substack{I\subset\{1,\ldots,n_1\}\\|I|=j}}\left(\prod_{i\in I}a_i^{n_1} - \prod_{i\in I}a_i^{n_2}\right)\right).
\end{aligned}$$

Next, we demonstrate that

$$\prod_{i\in I}a_i^{n_1} - \prod_{i\in I}a_i^{n_2} = \sum_{i\in I}\left((a_i^{n_1} - a_i^{n_2})\prod_{j\in I, j<i}a_j^{n_2}\prod_{j\in I, j>i}a_j^{n_1}\right).$$

By induction on $|I|$. Let $k = \max i \in I$.

$$\begin{aligned}
&\prod_{i\in I}a_i^{n_1} - \prod_{i\in I}a_i^{n_2} = a_k^{n_1}\prod_{i\in I\setminus\{k\}}a_i^{n_1} - a_k^{n_1}\prod_{i\in I\setminus\{k\}}a_i^{n_2} + (a_k^{n_1} - a_k^{n_2})\prod_{i\in I\setminus\{k\}}a_i^{n_2} \\
&= a_k^{n_1}\left(\prod_{i\in I\setminus\{k\}}a_i^{n_1} - \prod_{i\in I\setminus\{k\}}a_i^{n_2}\right) + (a_k^{n_1} - a_k^{n_2})\prod_{i\in I\setminus\{k\}}a_i^{n_2} \\
&= a_k^{n_1}\left(\sum_{i\in I\setminus\{k\}}\left((a_i^{n_1} - a_i^{n_2})\prod_{j\in I\setminus\{k\}, j<i}a_j^{n_2}\prod_{j\in I\setminus\{k\}, j>i}a_j^{n_1}\right)\right) + (a_k^{n_1} - a_k^{n_2})\prod_{i\in I\setminus\{k\}}a_i^{n_2} \\
&= \sum_{i\in I\setminus\{k\}}\left((a_i^{n_1} - a_i^{n_2})\prod_{j\in I, j<i}a_j^{n_2}\prod_{j\in I, j>i}a_j^{n_1}\right) + (a_k^{n_1} - a_k^{n_2})\prod_{i\in I\setminus\{k\}}a_i^{n_2} \\
&= \sum_{i\in I}\left((a_i^{n_1} - a_i^{n_2})\prod_{j\in I, j<i}a_j^{n_2}\prod_{j\in I, j>i}a_j^{n_1}\right).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mu^{n_1}[\lambda_{n_1}^{-1}(\cup_i A_i)] - \mu^{n_2}[\lambda_{n_2}^{-1}(\cup_i A_i)] \\
&\leq \sum_{j=1}^{n_1}(-1)^{j-1} \sum_{\substack{I\subset\{1,\ldots,n_1\}\\|I|=j}} \left(\prod_{i\in I} a_i^{n_1} - \prod_{i\in I} a_i^{n_2}\right) \\
&= \sum_{j=1}^{n_1}(-1)^{j-1} \sum_{\substack{I\subset\{1,\ldots,n_1\}\\|I|=j}} \sum_{i\in I}\left((a_i^{n_1} - a_i^{n_2}) \prod_{k\in I,k<i} a_k^{n_2} \prod_{k\in I,k>i} a_k^{n_1}\right) \\
&= \sum_{j=1}^{n_1}(-1)^{j-1} \sum_{i=1}^{n_1}(a_i^{n_1} - a_i^{n_2}) \sum_{\substack{I\subset\{1,\ldots,n_1\}\\|I|=j,i\in I}} \prod_{k\in I,k<i} a_k^{n_2} \prod_{k\in I,k>i} a_k^{n_1} \\
&= \sum_{i=1}^{n_1}(a_i^{n_1} - a_i^{n_2}) \sum_{j=1}^{n_1}(-1)^{j-1} \sum_{\substack{I\subset\{1,\ldots,n_1\}\\|I|=j,i\in I}} \prod_{k\in I,k<i} a_k^{n_2} \prod_{k\in I,k>i} a_k^{n_1}
\end{aligned}$$

Noticing that

$$\begin{aligned}
&\sum_{j=1}^{n_1}(-1)^{j-1} \sum_{\substack{I\subset\{1,\ldots,n_1\}\\|I|=j,i\in I}} \prod_{k\in I,k<i} a_k^{n_2} \prod_{k\in I,k>i} a_k^{n_1} = \\
&= \sum_{j=1}^{n_1-1}(-1)^{j-1} \sum_{\substack{I\subset\{1,\ldots,n_1\}\setminus\{i\}\\|I|=j}} \prod_{k\in I,k<i} a_k^{n_2} \prod_{k\in I,k>i} a_k^{n_1} \\
&= \sum_{j=1}^{n_1-1}(-1)^{j-1} \sum_{\substack{I\subset\{1,\ldots,n_1\}\setminus\{i\}\\|I|=j}} \mu^{n_2}\left[\bigcap_{k\in I,k<i} (\lambda_{n_2}^k)^{-1}(A_k)\right] \mu^{n_1}\left[\bigcap_{k\in I,k>i} (\lambda_{n_1}^k)^{-1}(A_k)\right] \\
&= \sum_{j=1}^{n_1-1}(-1)^{j-1} \sum_{\substack{I\subset\{1,\ldots,n_1\}\setminus\{i\}\\|I|=j}} \mu^{n_1}\times\mu^{n_2}\left[\bigcap_{k\in I,k>i} (\lambda_{n_1}^k)^{-1}(A_k) \times \bigcap_{k\in I,k<i} (\lambda_{n_2}^k)^{-1}(A_k)\right] \\
&= \mu^{n_1}\times\mu^{n_2}\left[\bigcup_{k>i} (\lambda_{n_1}^k)^{-1}(A_k)\times\Omega^{n_2} \cup \bigcup_{k<i} \Omega^{n_1}\times(\lambda_{n_2}^k)^{-1}(A_k)\right] \leq 1,
\end{aligned}$$

where the last equality follows from the inclusion-exclusion principle. It follows that

$$\begin{aligned}
\mu^{n_1}[\lambda_{n_1}^{-1}(\cup_i A_i)] - \mu^{n_2}[\lambda_{n_2}^{-1}(\cup_i A_i)] &\leq \sum_{i=1}^{n_1}(a_i^{n_1} - a_i^{n_2}) \\
&\leq \sum_{i=1}^{n_1}\alpha_i < \alpha,
\end{aligned}$$

as required. □

Notes

1 Some frequentists go even further. In their response to the American Statistical Association's controversial statement on *p*-values, Ionides et al. [2] identify frequentist method with deduction, and Bayesian method with induction.

2 Albert [10] answers Redhead's challenge by suggesting that we drop the fiction that observed variables are continuous. Since we would like our results to apply directly to problems as formulated in the sciences, where continuous variables are commonplace, we do not adopt Albert's solution. However, we assimilate this insight in Sections 4.1 and 4.2 by insisting on "feasible" test methods, i.e., methods whose verdicts depend only on a discretization of the data, even if the underlying variables are continuous.

[3] For example, the testability of conditional independence hypotheses is an area of active research [13,14] with far-reaching consequences.

[4] Kvanvig [15] makes a parallel point for epistemology: "...it is far from obvious that ...the best way of structuring a fruitful cognitive life is to concentrate on individual time-slices, and whether knowledge or justification is posessed at each time-slice, and let the totality ...get generated by cementing together these time-slices". Laudan [16] makes a similar point in philosophy of science: "Progress necessarily involves a ...process through time. Rationality ...has tended to be viewed as an atemporal concept ...most writers see progress as nothing more than the temporal projection of individual rational choices ...we may be able to learn something by inverting the presumed dependence of progress on rationality".

[5] We adopt here the idiom of "negative" falsificationism, according to which one should suspend judgement on a hypothesis unless it is falsified. "Positive" falsificationism, on the other hand, endorses belief in hypotheses which have passed an appropriate test (see Musgrave [17], Section 6). If we adopted the positive formulation, we could no longer speak of error avoidance *tout court*, but would have to rephrase the norm in terms of avoidance of errors of Type I (false rejection). Not much hinges on the choice, since the set of falsifiable hypotheses remains the same.

[6] Here an astute reader may object: how do we know that observation must turn up an elusive non-black raven if one exists? We might rephrase the hypothesis as 'all ravens that will ever be observed are black.' We might simply appeal to the background assumptions of inquiry: the method must converge to not-H not in all those possibilities in which H is false, but in all possibilities in which H is false and the background assumptions of inquiry are true. A more careful answer might require a falsification method to converge to not-H in a "maximal set" of possibilities in which H is false, e.g., in all those possibilities in which convergence is compatible with error avoidance. See Lin [18] for a rigorous development of this idea.

[7] The definitions in this section are intentionally rather schematic. Hopefully this will aid, rather than hinder, comprehension. All definitions are formalized in the following.

[8] The reader may object: surely no method can be expected to converge to not-H in *all* the possibilities in which H is false. For example, what if H is false because the assumption of i.i.d sampling is violated? A more careful formulation requires that the method converge to not-H in all those worlds in which H is false but the background assumptions of inquiry are true—if it is *statistical* falsification which is at issue, then some kind of statistical regularity must be taken for granted.

[9] The following are only proto-definitions, leaving many things unspecified. The notion of verification in the limit is here a free parameter: compatible notions include convergence in propositional information, convergence in probability and almost sure convergence. The notion of α-error avoidance is also parametric: it can mean that the chance of error at any sample size is bounded by α, or that the sum of the chances of error over all sample sizes is bounded by α. Each of these concepts will be developed in detail in the following. These parametric details are omitted here to expose the essential differences between the three concepts.

[10] For a contrived example, suppose it is known that random samples are distributed uniformly on the interval $(\mu - 1/2, \mu + 1/2)$, for some unknown parameter μ. Although samples may land outside the interval, they only do so with probability zero. Let H be the hypothesis that the true parameter is μ. Let M be the method that concludes not-H if some sample lands outside of the interval $(\mu - 1/2, \mu + 1/2)$, and draws no non-trivial conclusion otherwise. Then, M is a deductive falsifier of the second type, although not of the first. Clearly, every falsifier of the first type is a falsifier of the second type.

[11] A topological space $\mathcal{T}$ is a structure $\langle W, \mathcal{V} \rangle$ where W is a set, and $\mathcal{F}$ is a collection of subsets of W closed under conjunction and finite disjunction. The elements of $\mathcal{V}$ are called the closed sets of $\mathcal{T}$. In our case W is a set of epistemic possibilities, or possible worlds, and $\mathcal{F}$ is the set of falsifiable propositions over W. Although we define a topology here in terms of closed sets, we could just as easily have done it in terms of open sets since the complement of every closed set is open.

[12] See, however, Lin [35], Lin and Zhang [36] for how to proceed when identifiability fails.

[13] We use the notation $\mathsf{bdry}A$ to denote the set of boundary points of A.

[14] The acceptance region is $\psi^{-1}(W)$, rather than $\psi^{-1}(H)$, because failing to reject H licenses only suspension of belief i.e., the trivial inference W.

[15] Every topology gives rise to a kind of convergence by setting $\mu_n \Rightarrow \mu$ iff for every open E containing μ there is N such that $\mu_n \in E$ for $n \geq N$. Every kind of convergence gives rise to a topology by letting $E \subseteq W$ be open iff for each sequence $\mu_n \Rightarrow \mu \in E$ there is N such that $\mu_n \in E$ for $n \geq N$. The notion of convergence arising from a topology defined in this way will agree with the original notion of convergence.

[16] For example, if we are interpolating points with polynomials, *linear, quadratic, cubic, ...* is such a collection, when the hypotheses are understood in the strict sense. Although the individual hypotheses are not falsifiable—if the truth is linear, quadratic will never be refuted—unions of inital segments are—if the generating polynomial is of degree greater than 2, we will get a sign. Many statistical model selection problems also fit this description.

[17] See Miller [40], Tichý [41], Oddie [42], Niiniluoto [43,44].

[18] Niiniluoto [43] admits: "the problem of estimating verisimilitude is neither more nor less difficult than the traditional problem of induction".

[19] This entailment holds only for countably additive measures, to which we restrict attention.

References

1. Gelman, A.; Shalizi, C.R. Philosophy and the practice of Bayesian statistics. *Br. J. Math. Stat. Psychol.* **2013**, *66*, 8–38. [CrossRef] [PubMed]
2. Ionides, E.L.; Giessing, A.; Ritov, Y.; Page, S.E. Response to the ASA's Statement on *p*-Values: Context, Process, and Purpose. *Am. Stat.* **2017**, *71*, 88–89. [CrossRef]
3. Popper, K.R. *The Logic of Scientific Discovery*, 1st ed.; Hutchinson: London, UK, 1959.
4. Gillies, D.A. A Falsifying Rule for Probability Statements. *Br. J. Philos. Sci.* **1971**, *22*, 231–261. [CrossRef]
5. Albert, M. Die Falsifikation Statistischer Hypothesen. *J. Gen. Philos. Sci.* **1992**, *23*, 1–32. [CrossRef]
6. Mayo, D.G.; Spanos, A. Severe testing as a basic concept in a Neyman–Pearson philosophy of induction. *Br. J. Philos. Sci.* **2006**, *57*, 323–357.
7. Fisher, R.A. *Statistical Methods and Scientific Inference*; Oliver and Boyd: Edinburgh, UK, 1959.
8. Neyman, J. *Lectures and Conferences on Mathematical Statistics and Probability*; Graduate School, US Department of Agriculture: Washington, DC, USA, 1952.
9. Redhead, M. On Neyman's paradox and the theory of statistical tests. *Br. J. Philos. Sci.* **1974**, *25*, 265–271.
10. Albert, M. Resolving Neyman's Paradox. *Br. J. Philos. Sci.* **2020**, *53*, 69–76. [CrossRef]
11. Neyman, J.; Pearson, E.S. IX. On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. R. Soc. Lond. Ser. A Contain. Pap. Math. Phys. Character* **1933**, *231*, 289–337.
12. Casella, G.; Berger, R.L. *Statistical Inference*, 2nd ed.; Duxbury: Pacific Grove, CA, USA, 2002.
13. Shah, R.D.; Peters, J. The hardness of conditional independence testing and the generalised covariance measure. *Ann. Stat.* **2020**, *48*, 1514–1538.
14. Neykov, M.; Balakrishnan, S.; Wasserman, L. Minimax optimal conditional independence testing. *Ann. Stat.* **2021**, *49*, 2151–2177.
15. Kvanvig, J.L. *The Intellectual Virtues and the Life of the Mind: On the Place of Virtues in Contemporary Epistemology*; Rowman and Littlefield: Savage, MD, USA, 1992.
16. Laudan, L. *Progress and its Problems: Towards a Theory of Scientific Growth*; University of California Press: Berkeley, CA, USA, 1978.
17. Musgrave, A. Critical Rationalism. In *The Power of Argumentation*; Poznań Studies in the Philosophy of the Sciences and the Humanities; Suárez-Iñiguez, E., Ed.; Brill: Leiden, The Netherlands, 2007; pp. 171–211.
18. Lin, H. Modes of convergence to the truth: Steps toward a better epistemology of induction. *Rev. Symb. Log.* **2022**, *forthcoming*.
19. Bar-Hillel, Y.; Carnap, R. Semantic Information. *Br. J. Philos. Sci.* **1953**, *4*, 147–157.
20. Floridi, L. Is semantic information meaningful data? *Philos. Phenomenol. Res.* **2005**, *70*, 351–370. [CrossRef]
21. Floridi, L. *The Philosophy of Information*; Oxford University Press: Oxford, UK, 2011.
22. Niranjan, S.; Frenzel, J.F. A comparison of fault-tolerant state machine architectures for space-borne electronics. *IEEE Trans. Reliab.* **1996**, *45*, 109–113. [CrossRef]
23. Chernick, M.R.; Liu, C.Y. The saw-toothed behavior of power versus sample size and software solutions: Single binomial proportion using exact methods. *Am. Stat.* **2002**, *56*, 149–155. [CrossRef]
24. Schuette, P.; Rochester, C.G.; Jackson, M. Power and sample size for safety registries: New methods using confidence intervals and saw-tooth power curves. In Proceedings of the 8th International R User Conference, Nashville, TN, USA, 12–15 June 2012.
25. Musonda, P. The Self-Controlled Case Series Method: Performance and Design in Studies of Vaccine Safety. Ph.D. Thesis, Open University, Milton Keynes, UK, 2006.
26. Schaarschmidt, F. Experimental design for one-sided confidence intervals or hypothesis tests in binomial group testing. *Commun. Biometry Crop. Sci.* **2007**, *2*, 32–40.
27. Genin, K.; Mayo-Wilson, C. Statistical Decidability in Linear, Non-Gaussian Causal Models. In Proceedings of the Causal Discovery & Causality-Inspired Machine Learning Workshop, NeurIPS, Virtual, 11–12 December 2020.
28. Abramsky, S. Domain Theory and the Logic of Observable Properties. Ph.D. Thesis, University of London, London, UK, 1987.
29. Vickers, S. *Topology Via Logic*; Cambridge University Press: Cambridge, UK, 1996.
30. Kelly, K.T. *The Logic of Reliable Inquiry*; Oxford University Press: Oxford, UK, 1996.
31. de Brecht, M.; Yamamoto, A. Interpreting learners as realizers for Σ^0_2-measurable functions. *Spec. Interes. Group Fundam. Probl. Artif. Intell. (SIG-FPAI)* **2009**, *74*, 39–44.
32. Genin, K.; Kelly, K.T. Theory Choice, Theory Change, and Inductive Truth-Conduciveness. In Proceedings of the Fifteenth Conference on Theoretical Aspects of Rationality and Knowledge (TARK), Pittsburgh, PA, USA, 4–6 June 2015; pp. 111–121.
33. Baltag, A.; Gierasimczuk, N.; Smets, S. On the Solvability of Inductive Problems: A Study in Epistemic Topology. *Electron. Proc. Theor. Comput. Sci.* **2016**, *215*, 81–98. [CrossRef]
34. Genin, K.; Kelly, K.T. The Topology of Statistical Verifiability. In Proceedings of the Sixteenth Conference on Theoretical Aspects of Rationality and Knowledge (TARK), Liverpool, UK, 24–26 July 2017; pp. 236–250. [CrossRef]
35. Lin, H. The Hard Problem of Theory Choice: A Case Study on Causal Inference and Its Faithfulness Assumption. *Philos. Sci.* **2019**, *86*, 967–980.
36. Lin, H.; Zhang, J. On Learning Causal Structures from Non-Experimental Data without Any Faithfulness Assumption. In Proceedings of the Algorithmic Learning Theory, PMLR, San Diego, CA, USA, 8–11 February 2020; pp. 554–582.
37. Saeki, S. A proof of the existence of infinite product probability measures. *Am. Math. Mon.* **1996**, *103*, 682–683. [CrossRef]
38. Billingsley, P. *Convergence of Probability Measures*, 2nd ed.; John Wiley & Sons: Hoboken, NJ, USA, 1999. [CrossRef]

39. Genin, K. Statistical Undecidability in Linear, Non-Gaussian Models in the Presence of Latent Confounders. In Proceedings of Advances in Neural Information Processing Systems 34, Virtual, 6–12 December 2020.
40. Miller, D. Popper's qualitative theory of verisimilitude. *Br. J. Philos. Sci.* **1974**, *25*, 166–177. [CrossRef]
41. Tichý, P. On Popper's definitions of verisimilitude. *Br. J. Philos. Sci.* **1974**, *25*, 155–160. [CrossRef]
42. Oddie, G. *Likeness to Truth*; The University of Western Ontario Series in Philosoph of Science; Reidel: Dordrecht, The Netherlands, 1986; Volume 30.
43. Niiniluoto, I. *Truthlikeness*; The Synthese Library; Reidel: Dordrecht, The Netherlands, 1987; Volume 185.
44. Niiniluoto, I. *Critical Scientific Realism*; Oxford University Press: Oxford, UK, 1999.
45. Genin, K. The Topology of Statistical Inquiry. Ph.D. Thesis, Carnegie Mellon University, Pittsburgh, PA, USA, 2018.
46. Gilat, D. Monotonicity of a power function: An elementary probabilistic proof. *Am. Stat.* **1977**, *31*, 91–93. [CrossRef]