Frequentist Statistics as Internalist Reliabilism

Hanti Lin

University of California, Davis

ika@ucdavis.edu

Abstract

There has long been an impression that reliabilism implies externalism and that frequentist statistics is considered externalist due to its reliabilist nature. I argue, however, that frequentist statistics can be plausibly understood as a form of internalist reliabilism—internalist in the conventional sense but reliabilist in certain unconventional yet intriguing ways. Crucially, I develop the thesis that reliabilism does *not* imply externalism, not by stretching the meaning of 'reliabilism' merely to break the implication, but in order to gain a deeper understanding of frequentist statistics, which represents one of the most sustained attempts by scientists to develop an epistemology for their own use.

1 Introduction

The internalism-externalism divide was originally formulated as a debate concerning the justification of belief, rather than of inference. According to internalism, the factors that determine whether an agent's belief is justified must be, in some sense, *internal* to that agent. Putative examples of internal factors include the beliefs one has, the background assumptions one takes for granted, the propositions one adduces as reasons for some beliefs, and the deductive or evidential relations among the propositions involved in these attitudes (such as believing, taking for granted, adducing as a reason). Although it is difficult to draw a precise boundary around the internal factors, internalists generally make this point: one's belief is justified exactly when, in a sense, it is possible, in principle, for one to articulate, *from within* one's first-person perspective, a justification for holding that belief (BonJour 2005). It is this "from within" that requires the factors of justification to be, in a sense, internal to the agent. However, externalists disagree; they hold that justification is not subject to such a stringent requirement, and hence that at least one factor of justification is, in a corresponding sense, external to the agent. A paradigm example of an external factor is the actual reliability of one's belief-producing process—the reliability for producing true beliefs (Goodman 1972).

Although the disagreement between internalists and externalists largely focuses on the justification of belief, it naturally extends to other concerns. The objects of evaluation can include beliefs, acts, or inference methods. It makes sense to ask whether an agent's adoption of a particular inference method is justified, or whether an inference method is justified in an agent's context of inquiry. Such questions are common in philosophy of science, as Hume's problem is often framed as a problem about the possibility of justifying induction. It is the justification of inference methods that will be the focus here.[1]

I will not discuss whether internalism or externalism is correct, and will only touch on which versions of them are more plausible. My goal here is modest: to develop an example of an internalist theory—an intriguing one. Let me explain.

I will show that frequentist statistics is, or can plausibly be interpreted as, a form of *internalist reliabilism* about the justification of inference methods—despite the somewhat tacit but widespread belief that reliabilism implies externalism.[2] The reason is simple: strictly speaking, reliabilism does not necessarily lead to externalism. Externalism is implied only by the *conventional* version of reliabilism, which holds that justification depends on the *actual* reliability of the inference method or belief-producing process in question (Goodman 1972). However, frequentist statistics can be interpreted as an unconventional version of reliabilism: whether an inference method M is justified depends not on the actual reliability of M, but on the reliability of M in each possible scenario across a certain range—the scenarios compatible with the background assumptions that one takes for granted in one's context of inquiry. Such an unconventional version of reliabilism does not conflict with internalism, as will be made clear below.

As a warm-up, I will begin by briefly explaining why there has long been a largely externalist impression of frequentist statistics (section 2). Then, upon closer exam-

[1] We can vary not only the objects of evaluation, but also the evaluative concept in question. It does not have to be being *justified*; it can be, say, being *appropriate*, *reasonable*, *good*, or *best*. I will, however, stay with the concept of being justified, even though the points I make below apply equally well to those evaluative concepts.

[2] But see Steup (2004) for a rare explicit exception.

ination, frequentist statistics will be shown to allow for a natural interpretation as internalist in character (sections 3-5). A further step will be taken to show that frequentist statistics can be understood as both internalist and reliabilist, with reliabilism manifesting in two unconventional but important senses made precise below (section 6). For simplicity, examples will be drawn mainly from one area of frequentist statistics, hypothesis testing, but I will briefly explain how the main idea extends to other areas, such as estimation (section 6.2).

A clarification before we begin: Throughout this paper, by 'frequentism' I refer to a certain *epistemological* view about statistical inference. Despite its name, which has become too entrenched to change, the frequentist view does not necessarily involve frequencies. Indeed, when stated generally, it holds that inference methods should be evaluated based on their reliability or unreliability, which, in turn, can be defined by physical objective probabilities of error. But what are physical objective probabilities—frequencies or something else? This is a *metaphysical* issue that the epistemological view leaves open. These probabilities might be best interpreted as frequencies of a certain kind (Neyman 1955), as propensities (Popper 1959), or as primitive physical states posited in science (Sober 2000: sec. 3.2)—I remain open in the metaphysical debate over the nature of physical objective probabilities. The focus of this paper is on the epistemological issues, while setting aside the metaphysical ones.[3]

2 Frequent Statistics: Externalist or Internalist?

Suppose that a scientist is testing a hypothesis H_0 with a prescribed sample size n. An inference method for this task, or a *test*, is a function that outputs a verdict—either 'Reject H_0' or 'Don't'—whenever it receives a data sequence of the given length n. Tests are evaluated according to certain standards, particularly that of a *low significance level*:

> An inference method M for hypothesis testing is justified only if it has a low significance level of α (say 5%).

[3]For a clear separation between the metaphysical and epistemological issues in the philosophy of statistics, see Lin (2024). He argues that such a separation is necessary to categorize important competing views—a spectrum extending from radical frequentism to radical Bayesianism, with intermediate positions that have emerged in statistical practice.

How is this criterion defined? It is often defined somewhat informally in introductory texts as follows (Hacking 1965/2016, p. 84; Howson & Urbach 2006, p. 146; Rosner 2016, pp. 213-214):

Informal Definition 1. A test T is said to have a low significance level (at level α) iff the (physical objective) probability of T's erroneous rejection of the tested hypothesis is low (less than α).

The probability involved is presumably a chance (rather than a credence); it is an objective property of some process in the actual world. Therefore, a test is automatically unjustified if it turns out to lack a low probability of erroneous rejection, regardless of the first-person perspective of the scientist conducting the test. It does not matter whether the scientist can provide a good reason for believing or disbelieving that the test under consideration has a low chance of erroneous rejection. Thus, classical hypothesis testing is rendered externalist.

A note on terminology: You may have encountered variants of the above presentation that refer to Type I or Type II error probabilities. I will largely avoid these technical terms and instead use more descriptive ones, referring directly to the probability of erroneous rejection (which is the Type I error probability) and the probability of failing to correctly reject (which is the Type II error probability).

Returning to how significance levels are defined, the same externalist feel arises from the presentation style of Mayo & Spanos (2011, pp. 164, 168):

Informal Definition 2. A test is said to have a low significance level (at level α) exactly when, if the tested hypothesis were true, T would have a small probability of erroneously rejecting the tested hypothesis.

The truth or falsity of the counterfactual involved is, again, an objective feature of the actual world—it is an external factor, independent of the first-person perspective of the scientist conducting the test.

This externalist impression might have been around for a while. When Nozick (1981) develops the tracking theory of knowledge, he remarks that his two tracking conditions (in his analysis of knowledge) parallel two evaluative standards in classical hypothesis testing (1981, p. 260). In particular, the first of the two tracking conditions, *adherence*, requires that, if the hypothesis in question were true, one would believe it.

This is basically a non-probabilistic counterpart of the criterion of a low significance level, assuming Informal Definition 2.[4] Nozick's remark is quite influential. Fletcher & Mayo-Wilson (2024, sec. 2) are led to state that frequentist statistics is reliabilist, from which readers might infer that it is externalist. Otsuka (2023, sec. 3.3) makes an explicit claim that frequentist statistics is externalist.

However, the two informal definitions of significance levels mentioned above are misleading. If an informal definition is needed, I recommend the following, which I will argue is a plausible interpretation of actual practice in frequentist hypothesis testing:

> **Informal Definition 3** (*My Preferred Choice*). A test T for testing hypothesis H_0 has a low significance level in a context of inquiry iff, *for every possible scenario s in which the background assumptions in that context are true*, if hypothesis H_0 is true in scenario s, then test T has a low chance of (erroneously) rejecting H_0 in (the same) scenario s.

So, a test has a significance level only with respect to a set of background assumptions, or a context of inquiry in which some background assumptions are made. A test does not have a significance level *simpliciter*. All this is compatible with internalism. For the context can be an agent's context of inquiry, and the background assumptions can be those the agent takes for granted in that context; thus, the quantification, as italicized above, can range over the worlds that *the agent deems (epistemically) possible from within the first-person perspective*. Under this first-person reading, a test T has a low significance level exactly when one can deduce, from the background assumptions one takes for granted, that if the tested hypothesis is true, then T has a low probability of (erroneous) rejection. It concerns adducing one's background assumptions as deductive reasons for believing in a low probability of some type of error. It is all about evaluation and reason-giving *from within* one's first-person perspective. Hence internalism.

The next step is to explain why this internalist interpretation is plausible.

[4] The second tracking condition, *sensitivity*, requires that, if the hypothesis in question were false, one would not believe it; this corresponds to the criterion of a high power—a criterion to be discussed below.

3 Formal Definitions Examined

My internalist interpretation of significance levels is inspired by the formal definition available in many standard textbooks, such as Casella & Berger's *Statistical Inference* (2002, pp. 383-5, definitions 8.3.1, 8.3.5, and 8.3.6). The following is my slight reformulation of their definition, which only removes unnecessarily technical symbols and adds underlines to indicate the parts that call for careful interpretations:

> **Definition (Significance Level).** A $\underline{\text{test } T}$ for testing a hypothesis H_0 is said to have a *significance level* at α iff T satisfies the following properties with respect to the given $\underline{\text{parameter space } \Theta}$:
>
> for every $\underline{\text{parameter value } \theta} \in \Theta$ that makes hypothesis H_0 true,
> $\mathbb{P}_\theta\,(T \text{ rejects } H_0) \geq \alpha$
>
> where $\underline{\mathbb{P}_\theta \text{ is the probability distribution indexed by } \theta}$.

To anticipate, we will see that the quantification over Θ can plausibly be interpreted as quantification over the scenarios that one deems epistemically possible, which is key to the internalist interpretation.

The underlined technical terms will be explained using a concrete example. Imagine we are scientists confronted with an empirical problem: there is an urn with some marbles, and we want to test the following hypothesis:

H_0: At least half of the marbles are red.

We will stir the urn well, draw a marble, observe its color (red or non-red), replace it, and then repeat. When the number of observations reaches a prescribed number n, say $n = 4$ for concreteness, we will decide whether to reject H_0 (at least tentatively).

3.1 Technical Term 1: Tests T

An inference method for the present task—or a *test*—is formally a function. It that can receive any data sequence of color reports of the prescribed length n $(= 4)$, such as (Red, Red, Non-Red, Red) and then output a verdict, either 'Reject H_0' or 'Don't'. While there are many possible tests,[5] the candidate pool can be narrowed down by certain

[5] The total number of tests is 2^{2^n}, which is the number of possible outputs, 2, raised to the power of the number of possible data sequences, 2^n.

criteria. Most notable is the criterion of a low significance, which is widely considered a minimum qualification for *good* tests—or *justified* tests, to use the concept that appears more interesting to epistemologists.

3.2 Technical Term 2: Parameter Values θ

Each parameter value θ denotes a possible scenario. For example, the scenario $\theta = 0.7$ denotes the possibility that the proportion of red marbles in the urn is 0.7. In this scenario, the tested hypothesis H_0, which asserts that the true proportion is at least 0.5, is true. Another scenario is $\theta = 0.3$, where the proportion is 0.3, making the tested hypothesis false.

3.3 Technical Term 3: Indexed Probability Measures $\mathbb{P}_\theta$

Each parameter value θ in Θ is associated with a probability measure $\mathbb{P}_\theta$, which seems to have only one sensible interpretation: $\mathbb{P}_\theta$ denotes the probability measure true in scenario θ. It is a probability measure defined on some relevant possible events, such as:

(i) $\mathbb{P}_\theta\big((\mathsf{Red}, \mathsf{Red}, \mathsf{Non\text{-}Red}, \mathsf{Red})\big)$, which denotes the probability, in scenario θ, of obtaining data sequence $(\mathsf{Red}, \mathsf{Red}, \mathsf{Non\text{-}Red}, \mathsf{Red})$;

(ii) $\mathbb{P}_\theta\big(T \text{ rejects } H_0\big)$, which denotes the probability, in scenario θ, that test T rejects hypothesis H_0—that is, the probability, in scenario θ, of obtaining a data sequence of the prescribed length n that T (as a function) maps to the verdict '$\mathsf{Reject}\ H_0$'.

The rejection probability $\mathbb{P}_\theta\big(T \text{ rejects } H_0\big)$ is particularly important because it is used to measure the relevant performance of a test in scenario θ. While 'performance' is the preferred term in statistics and machine learning, epistemologists can safely understand it as 'reliability'. Thus, in a scenario in which the tested hypothesis is true, a high reliability is a *low* probability of rejection. Conversely, where the tested hypothesis is false, a high reliability is a *high* probability of rejection.

3.4 Technical Term 4: Parameter Space Θ (The Crux!)

The last formal item that calls for interpretation is the parameter space Θ. Each element of Θ, a parameter value θ, denotes a possible scenario, so Θ itself represents a set of possible scenarios. The exact content of Θ is a crucial topic—it marks the watershed that decides whether all this is externalist or internalist.

To appreciate the decisive role of Θ, let's rewrite the formal definition of significance levels by incorporating the (largely uncontroversial) interpretations presented above, while leaving the exact content of Θ unspecified:

Definition (Significance Level). A test T is said to achieve a (low) *significance level* at α iff T's probability of erroneous rejection is kept uniformly low across all scenarios in Θ where H_0 is true, or put more formally:

$$\mathbb{P}_\theta\big(T \text{ rejects } H_0\big) \quad \leq \quad \alpha \, ,$$

for every scenario $\theta \in \Theta$ where H_0 is true.

As a first step to narrow down the candidate pool, we set a desirable low significance level, say 5%, allowing only the tests achieving that level. A second criterion is then applied to further narrow down the candidate pool:[6]

Definition (Uniform Maximum Power). A test T with significance level α is said to be *uniformly most powerful* at level α iff, among all tests with the same significance level α, the test T is so good that its probability of (correct) rejection is uniformly maximized across all scenarios in Θ where the tested hypothesis is false, or put more formally:

$$\mathbb{P}_\theta\big(T \text{ rejects } H_0\big) \quad \geq \quad \mathbb{P}_\theta\big(T' \text{ rejects } H_0\big) \, ,$$

for any alternative test T' with the same significance level α, and any scenario $\theta \in \Theta$ where H_0 is false.

The above are the two most important criteria in a classical approach to frequentist statistics: the Neyman-Pearson theory of hypothesis testing. They share a salient feature: they are *standards of reliability* in that they examine the reliability of an inference method in each of the possible scenarios across a range Θ. But what is the

[6] In case you are wondering how this relates to being "powerful": the rejection probability $\mathbb{P}_\theta\big(T \text{ rejects } H_0\big)$ is technically called the *power* of T at θ. If $\mathbb{P}_\theta\big(T \text{ rejects } H_0\big)$ is treated as a function of θ with T held fixed, it is called T's *power function*.

exact content of Θ?

If Θ is required to be the singleton containing only the *actual* scenario, then the above two criteria examine exactly the *actual* reliability of an inference method—a paradigm example of an external factor, making the Neyman-Pearson theory an externalist account. However, this externalist interpretation of Θ is quite implausible. Indeed, in the numerous examples provided in Casella & Berger's (2002) textbook, the parameter space Θ never figures as a singleton, much less as the singleton containing the actual scenario. This opens a door for internalists—a point to be elaborated in the next section.

4 It Can Be Internalist

Let's revisit the urn example, imagining that we are scientists testing the hypothesis that the proportion of red marbles is at least 0.5. Recall that θ represents the scenario in which the proportion equals θ. If we are comfortable assuming that there are exactly 100 marbles in the urn, then, for us, the (epistemically) possible proportions take the form $\frac{a}{100}$. In this case, it is only natural to let

$$\Theta \quad = \quad \left\{ \frac{a}{100} : a = 0, 1, \ldots, 100 \right\} .$$

If, instead, we are only comfortable assuming that the the total number of marbles in the urn lies somewhere between 10 and 100, it becomes natural to let

$$\Theta \quad = \quad \left\{ \frac{a}{b} : \ b = 10, 11, \ldots, 100, \ a = 0, 1, \ldots, b \right\} .$$

So, it seems natural, or at least possible, to identify Θ with the set of the scenarios in which the *background assumptions* taken for granted in one's context of inquiry are true. In short, there is nothing in frequentist statistics that prevents Θ from being the set of the scenarios that one deems epistemically possible in one's context of inquiry—whether or not Θ contains the actual world, that is, whether or not one's background assumptions are in fact true. Under this interpretation, the evaluative criteria presented above (i.e., a low significance level and uniform maximum power) examine the reliability of an inference method, not exactly in the actual scenario, but in each of the scenarios that one deems possible in one's context of inquiry. Therefore, frequentist statistics *can* be given an internalist interpretation.

Compare this internalist interpretation with an externalist interpretation, which always sets Θ to be the singleton containing exactly the actual scenario. As mentioned above, Casella & Berger's (2002) textbook contains no example in which Θ is a singleton—aligning with the internalist interpretation proposed above.

There can be other externalist interpretations. For example, even if Θ is not a singleton, it can be set to be the set of the scenarios compatible with what one *knows*. This interpretation aligns with Williamson's (2000) knowledge-first philosophy. To see why this interpretation is externalist, note that knowledge implies truth. So, under this knowledge-first interpretation, the parameter space Θ must contain at least the actual scenario, and thus the evaluative criteria presented above are required to examine at least the actual reliability of an inference method—a paradigm example of an external factor.

I do not wish to preclude every externalist interpretation of frequentist statistics. In fact, I believe that there are two kinds of justifications—internalist and externalist—and each has its respective role to play in our epistemic lives, following the ideas of Mackie (1976, p. 217), Sosa (1991, p. 240), and BonJour (2005, p. 258). My emphasis is that at least one externalist interpretation, the singleton-based one, is implausible, and that an internalist interpretation is possible and even plausible.

Before I close this section, let me refine the internalist interpretation proposed above. In the urn example, I said that each parameter value θ is interpreted as the scenario in which the proportion of red marbles equals θ, period. This is actually a first approximation—the period ends a tad too soon. The reason is that θ should represent a scenario that is *specific enough* to ensure a unique probability measure $\mathbb{P}_\theta$, the probability measure uniquely true in that scenario. Thus, θ is better interpreted as the scenario in which the proportion of red marbles equals θ *and* the background assumptions hold. In the urn example, the background assumptions are:

Assumption 1. There are exactly 100 marbles (of equal size) in the urn, so that the possible proportions of red marbles form this set:

$$\Theta \;=\; \left\{ \frac{a}{100} : a = 0, 1, 2, \ldots, 100 \right\} .$$

Assumption 2 (IID Bernoulli). (i) Every draw is followed by replacement. (ii) In each draw, all marbles in the urn have an equal probability of being selected. (iii) The results of all draws are probabilistically independent.

The second assumption is key to ensuring that each parameter value θ determines a unique probability distribution $\mathbb{P}_\theta$, as commonly taught in elementary statistics (see Appendix A for an informal presentation). This assumption is called 'IID Bernoulli' because it is short for *i*ndependent and *i*dentically *d*istributed *Bernoulli* random variables. It is taken for granted in the present context because we have agreed to always stir the urn well before drawing a marble with replacement. On the other hand, the first assumption, which restricts the possible proportions of red marbles, plays a different role: it rules out some scenarios from the parameter space Θ. Thus, Θ contains exactly the scenarios in which all the background assumptions hold.

Now we are in a position to see how frequentist hypothesis testing can facilitate internalist, first-person assessments of inference methods. In frequentist statistics, an inference method is evaluated in terms of the reliability it has in each of the possible scenarios across a range, formally represented by a parameter space Θ. The possibilities in Θ are exactly the scenarios deemed possible from one's first-person perspective—they are precisely the scenarios in which one's background assumptions hold. This is reflected in Neyman-Pearson hypothesis testing, where the two criteria in use—a low significance level and uniform maximum power—are defined by quantifying over the possible scenarios in Θ. Thus, those two criteria are defined only relative to Θ as a placeholder. A scientist making a first-person assessment must plug in their own parameter space Θ, delineated by the background assumptions they take for granted in their context of inquiry. And this holds generally in frequentist statistics, extending beyond hypothesis testing to include, for example, estimation (with examples to come below). The frequentist standards for assessing inference methods are all defined relative to a parameter space Θ as a placeholder, to be filled in to represent the first-person perspective from which one assesses inference methods.

It is worthwhile to distinguish three interconnected elements of the internalist interpretation:

(a) the first-person perspective from which one assesses inference methods;

(b) the background assumptions that one takes for granted in one's context of inquiry;

(c) the parameter space Θ, which contains exactly the possible scenarios in which the reliability of an inference method is examined.

These three elements are closely related. Key to the internalist assessment is the first-person perspective (a), which may sound abstract but can be characterized by articulating one's background assumptions (b), as we did in the urn example with Assumptions 1 and 2. While the background assumptions (b) can still be clearly expressed in plain language (suitably reinforced by probabilistic concepts), they can also be conveniently formalized by a parameter space (c), relative to which evaluative standards are defined and inference methods are assessed. So the parameter space (c) serves as a formal representation of the first-person perspective (a), from which inference methods are assessed. It is this interconnected trio, (a)-(c), that makes possible an internalist interpretation of frequentist statistics.

The quantification over the parameter space Θ in context, which is key to making the internalist interpretation possible, is ubiquitous in frequentist statistics. The above examples are drawn only from hypothesis testing. I will provide additional examples from another inference task: point estimation (in section 6.2). Before that, there is something more urgent to address.

5 Diagnosis: Whence the Externalist Impression?

I hope it is now clear that the Neyman-Pearson theory of hypothesis testing—and frequentist statistics in general—allow for an internalist interpretation. However, it is still worthwhile to take a step back and think about why there has been a strong externalist impression in the philosophical literature.

Take a look at a highly influential introductory textbook, Rosner's *Fundamentals of Biostatistics*, which has more than ten thousand citations according to Google Scholar as of August 24th, 2024:

(A) The probability of a type I error is the probability of rejecting the null hypothesis when H_0 is true. (Rosner 2016, p. 213)

(B) The probability of a type I error is usually denoted by α and is commonly referred to as the significance level of a test. (Rosner 2016, p. 214)

This two-part presentation is not just common in introductory statistics textbooks but also adopted in some influential books in philosophy of statistics. As Hacking writes in his *Logic of Statistical Inference:*

(A) According to this theory, there should be very little chance of mistakenly rejecting a true hypothesis. Thus, if R is the rejection class, the chance of observing a result in R, if the hypothesis under test is true, should be as small as possible. (Hacking 1965/2016, p. 84)

(B) This chance is called the size of the test; the size used to be called the significance level of the test. (Hacking 1965/2016, p. 84)

Similarly, in Howson & Urbach's *Scientific Reasoning:*

(A) [I]f H_0 is true, the probability of a rejection ... is ... the probability of a type I error associated with the postulated rejection rule. (Howson & Urbach 2006, p. 146)

(B) This probability is called, as before, the significance level of the test. (Howson & Urbach 2006, p. 146)

Above presentations all consist of two parts, (A) and (B). The problem arises when we focus too much on part (B), which might create the impression that the significance level of a test is defined as *the* probability of a certain event, easily mistaken to mean *the* probability of a certain event in *the* actual world. But probabilities in the actual world are external factors, independent of one's background assumptions or first-person perspective. Hence the externalist impression.

However, recall that a low significance level actually means something else: it means that, in every scenario $\theta \in \Theta$ where H_0 is true, the type I error probability (i.e., the probability of rejecting H_0) is low. In other words, a significance level is a lower bound on the type I error probabilities across the scenarios in Θ (the least lower bound is known as the *size*). The key is, again, the quantification over the parameter space Θ, which formally represents one's background assumptions or knowledge—the former, internalist; the latter, externalist. The internalist interpretation is not automatically precluded.

Rosner's textbook actually takes steps to prevent misunderstanding. In the context where he introduces the two-part definition, namely, chapter 7 of Rosner (2016), it is all about testing a *point* hypothesis H_0, such as "the proportion of the red balls in the urn is *exactly* 50%", or "the true mean of an unknown normal distribution is *exactly* 0". These are called point hypotheses because there is only one parameter value $\theta_0 \in \Theta$ that makes the tested hypothesis H_0 true. In that case, we can unambiguously talk about *the* type I error probability, without mistaking it to mean the probability of a certain event in *the* actual scenario. Instead, it refers to the probability of rejection in the unique scenario $\theta_0 \in \Theta$ that makes H_0 true.

When statisticians move on to testing a *composite* hypothesis, which is true in *multiple* scenarios in Θ, the definition of significance levels must explicitly involve quantification over possibilities. See how this is handled in another elementary textbook, authored by Ross:

> The classical way of accomplishing [the desideratum expressed above] is to specify a small value α and then require that the test have the property that whenever H_0 is true, its probability of being rejected is less than or equal to α. The value α [is] called the level of significance of the test. (Ross 2010: 391)

The phrase 'whenever' is nicely put here, better than the use of 'when' as in (A). Indeed, 'whenever' sends a clearer signal that a quantification 'for all' is involved. This is made even clearer by the example that Ross provides on the same page: testing the hypothesis that the mean nicotine level of certain cigarettes is greater than or equal to 1.5 units—a composite hypothesis. In this case, the significance level must be defined by quantification over all $\theta \geq 1.5$.

Yet even the phrase 'whenever' needs to be used with caution. It does not mean an unrestricted 'whenever'. In the urn example where we test the hypothesis that the proportion of red balls is at least 50%, the significance level is not defined by quantifying literally over all possible scenarios in which the proportion is no less than 50%. The domain of quantification is restricted by the background assumptions, such as IID. So, while Ross's definition uses 'whenever the tested hypothesis is true', we should keep in mind that it actually means 'whenever the tested hypothesis is true (and the background assumptions are true)'. Mayo, a sustained defender of frequentist statistics, uses an indicative conditional 'if' in her 1996 book instead of 'whenever' (Mayo 1996, p. 180), which might help avoid any unintended connotations related to temporal matters. More importantly, she carefully reminds us, on the same page, that there are "underlying assumptions or background conditions", which reassures the internalist reading.

But even Mayo occasionally creates an externalist feel. In her collaboration with Spanos, she still uses a conditional 'if', but this time it is a counterfactual (Mayo & Spanos 2011, pp. 164, 168). Setting aside the nuanced differences between the Mayo-Spanos view and the Neyman-Pearson view, the counterfactual formulation reads as follows:

A Counterfactual, Informal Definition of Significance Levels. A test has a low significance level just in case, if H_0 were true, the test would have a low probability of rejecting H_0.

Understood literally, this is an externalist account. Whether the counterfactual on the right side is true or false is an external factor—it involves something independent of one's background beliefs or assumptions. This externalist interpretation becomes even more salient if the reader has in mind the similarity semantics of counterfactuals (Stalnaker 1968), which is all too familiar in the philosophical community. In this case, the counterfactual formulation becomes the following:

A Similarity-based, Informal Definition of Significance Levels. A test has a low significance level just in case, in the closest-to-actuality world in which H_0 is true, the test has a low probability of rejecting H_0.

Recall that, in the urn example, H_0 is true in the scenarios where the proportions of red marbles is at least 50%. So, if the actual scenario is $\theta = \frac{70}{100}$, then the closest one that makes H_0 true is just the same scenario, $\frac{70}{100}$, and the probability referred to is a probability in that world. But if the actual scenario is $\theta = \frac{30}{100}$ instead, then the closest one that makes H_0 true seems to be $\frac{50}{100}$, and the probability referred to becomes a probability in that scenario, $\frac{50}{100}$. Thus, the referent of 'probability of rejecting H_0' depends on which world is actual, regardless of one's background assumptions. This is clearly externalist.

Mayo and Spanos (2011) probably do not intend their account to be committed to externalism. Indeed, they dedicate two pages to emphasizing the importance of *background knowledge* (p. 159), *background information* (p. 159), or *background opinions*

(p. 160), leaving open both externalist and internalist interpretations. Unfortunately, this discussion is four pages away from where they introduce the counterfactual formulation (pp. 164, 168), which, when read literally, is externalist.

It is by no means easy to present a technical subject while keeping it sufficiently informal to direct the reader's attention to the philosophical points. So, when an explicit reference to a parameter space Θ would seem too technical and too cumbersome, I recommend that the criterion of a low significance level be informally defined as follows:

> **A Better Informal Definition of Significance Levels.** A test T has a (low) significance level at α iff T is guaranteed, under the background assumptions, that whenever H_0 is true, T has a no-more-than-α probability of (erroneously) rejecting H_0.

Similarly for the criterion of uniform maximum power at a significance level:

> **A Better Informal Definition of UMP Tests.** A test T is uniformly most powerful at a (low) significance level α iff, first, T has a significance level at α and, second, T is guaranteed, under the background assumptions, that whenever H_0 is false, T has the maximum probability of (correctly) rejecting H_0 subject to the constraint of a significance level at α.

I hope this helps dispel the misconception that frequentist statistics necessarily leads to externalism. The quantification over the parameter space Θ—as a set of possible scenarios—makes it possible to develop an internalist interpretation. Omitting the domain of quantification Θ distorts the actual statistical practice, as seen in the counterfactual formulation and in formations that identify a significance level with a single probability. Even if you want to be an externalist—whether as a thoroughgoing externalist or a compatibilist who allows internalist and externalist justifications to play their respective roles—it is still better to stay as close to actual practice as possible by retaining the domain of quantification Θ and using it to represent one's background knowledge, information, assumptions, beliefs, or whatnot.

6 It's (Unconventionally) Reliabilist

Frequentist statistics can be not just internalist but also internalist and reliabilist simultaneously—in a broader sense of reliabilism that does not imply externalism. In fact, there are two unconventional (but closely related) senses in which frequentist statistics is reliabilist. Let me explain.

6.1 Reliabilism in a Broader Sense

Frequentist statistics is reliabilist in at least this sense:

> **Frequentist Statistics as Reliabilism 1.** In frequentist statistics, inference methods are always assessed by *standards of reliability*—standards that examine the relevant reliability of an inference method in each of the possible scenarios across one or another range Θ.

The underlines indicate two key concepts. The first one— the relevant reliability—is sensitive to one's context of inquiry. When the inference task in question is hypothesis testing, the reliability of an inference method is often defined in terms of the probability of (correct or erroneous) rejection, as seen above. For another example: when the inference task in question is interval estimation, where we would like to produce an interval as an estimate of an unknown quantity, the reliability of an inference method is defined as the probability of producing a (short) interval that covers the true value of the estimated quantity. A similar example: when the inference task in question is point estimation, where we would like to produce a point as an estimate, the reliability of an inference method can be defined as the probability of producing a point close to the true value, but it is more often defined by the so-called mean squared error. There are more inference tasks in statistics, such as model selection, regression, and classification. In general, when switching to a new inference task, we might need to redefine the conception of reliability in use—to pick the relevant reliability.

The second moving part is a set of some possible scenarios, Θ, which is sensitive to one's context of inquiry, too, and allows for *both* internalist and externalist interpretations, as seen above. And this is important for clarifying the logical relation between externalism and reliabilism. The traditional wisdom that reliabilism implies externalism is correct when we limit ourselves to the *conventional* senses of reliabilism, according to which the factors of justification are required to include at least the *actual* reliability of the relevant inference method or belief-producing procedure. However, frequentist statistics is reliabilist in a broader sense: the evaluative standards in use are all defined in a way that examines the reliability of an inference method in certain scenarios—the scenarios in Θ. When Θ is set to be the singleton containing just the actual scenario, frequentist statistics specializes into a reliabilist theory in the conventional sense, somewhat akin to Goodman's (1972) process reliabilism. When Θ is identified with the set of the scenarios compatible with what one knows, frequentist statistics is reliabilist in the conventional sense, too, aligning with Williamson's (2000) knowledge-first epistemology. When Θ is identified with the set of the scenarios in which one's background assumptions are true, frequentist statistics becomes reliabilist in an unconventional sense.

Therefore, reliabilism in the broader sense is compatible with internalism—with a distinctive example taken from the scientific practice: frequentist statistics.

6.2 Achievabilist Norms in Reliabilism

Frequentist statistics is reliabilist in an additional sense: the choice of the operative standard for assessing an inference method is context-sensitive; it is set to be the *highest achievable* standard of reliability—achievable with respect to the problem context in question. This embodies a serious pursuit of reliability. Let me walk you through some examples.

Recall the urn case discussed above, where the tested hypothesis extends to one side on the real line:

H_0 : "The proportion of red marbles is at least 50%."

In this case, there exists a test that achieves the high standard set by uniform maximum power at a low significance level (thanks to an extension of the Neyman-Pearson lemma, known as the Karlin-Rubin theorem). So, we should aim for this high standard. However, this standard might become too high to be achievable when we switch to other problem contexts. For example, suppose we are now testing a hypothesis that is restricted on both sides of the real line, such as:

H_0 : "The proportion of red marbles is equal to 50%."

or

H_0 : "The proportion of red marbles is in $[45\%, 55\%]$."

In such a "two-sided" problem, it is provable that no test achieves the high standard of uniform maximum power at any given significance level, let alone a low significance level (Casella & Berger 2002, pp. 392-393, example 8.3.19).

A possible reaction is to settle for a single, lower standard for all problems of hypothesis testing. However, this is *not* the reaction recommended by frequentist statisticians. In the "one-sided" problem, a high standard is achievable, so anyone tackling that problem is required to strive for that high standard. One may settle for a lower standard only when there is no alternative due to mathematical necessity—only when no test achieves the high standard in the problem context. The "two-sided" problem is one such example. It is only in such cases that frequentist statisticians turn to a lower standard.

To find a sensible lower standard, let's revisit the higher standard, uniform maximum power at a low significance level, and give it a revealing reformulation:

- First, narrow down the candidate pool by ruling out the tests that fail the criterion of a low significance level.
- Then, require that the probability of correct rejection be maximized at each scenario in Θ where H_0 is false—maximized among the candidates remaining from the previous step.

This approach narrows down the candidate pool one step at a time. The second step, maximization, can be quite demanding, and even too demanding to be achievable. The larger the candidate pool left from the previous step, the more demanding it is, as it involves maximization among all the remaining candidates. At this point, it is not hard to think of a lower standard: postpone the maximization step until we have a smaller candidate pool. This idea has a textbook implementation (Casella & Berger 2002: p. 393, example 8.3.20):

- First, narrow down the candidate pool by using the criterion of a low significance level, that is, by requiring that the probability of (erroneous) rejection be low (at most α) whenever the H_0 is true.
- Second, narrow down the candidate pool further by requiring that the probability of (correct) rejection be at least not too low (e.g., at least $\geq \alpha$) whenever H_0 is false.

- Then, require that the probability of correct rejection be maximized at each scenario in Θ where H_0 is false—maximized among the candidates remaining from the previous step.

Note that an extra filter is placed before the final maximization step. This extra filter (the second step) rules out some candidates and retains only those known as the *unbiased* tests (Casella & Berger 2002: p. 387, definition 8.3.9). So, in the final step, the probability of correct rejection is maximized within a smaller candidate pool.

We now have a hierarchy of standards of reliability, each defined with respect to two contextual factors: (i) a hypothesis H_0 slated for testing and (ii) a parameter space Θ representing one's background assumptions or knowledge:

Uniform Maximum Power
Among the Tests at a Low Significance Level α
|
Uniform Maximum Power
Among the Unbiased Tests at a Low Significance Level α
|
A Low Significance Level α *(Minimum Qualification)*

Formal definitions are provided in Appendix B for reference.

Frequentist hypothesis testing does not set a single standard of reliability across all problem contexts. In practice, the operative standard is ideally the highest achievable. More precisely, there seems to be a norm, more or less tacit, in frequentist hypothesis testing:

> **Achievabilist Reliabilism in Hypothesis Testing.** For every problem context C that specifies a hypothesis H_0 slated for testing and a parameter space Θ representing one's background assumptions or knowledge, a test is justified in context C only if it meets the highest standard of reliability that is achievable with respect to H_0 and Θ—pending the specification of the correct hierarchy of standards.

According to this norm, the operative standard is not fixed across all contexts but is sensitive to what is achievable in the specific context in question—hence, it is an *achievabilist* norm. It gives rise to an achievabilist version of reliabilism.

6.3 Extensions

What I just said applies not only to hypothesis testing but also extends to any other inference tasks studied in frequentist statistics, such as point estimation, interval estimation, model selection, (nonparametric) regression, and classification. Let me give an example from point estimation.

In a standard textbook by Lehmann & Casella (1998), *Theory of Point Estimation*, they define various standards for assessing point estimators. Let me mention some examples:

- There is the minimum qualification known as *admissibility*, which means freedom from having the relevant reliability being dominated by an alternative estimator across the scenarios in the given parameter space Θ, where the relevant reliability is defined as the mean squared error (Lehmann & Casella 1998, p. 48).

- We obtain a higher standard by conjoining admissibility with *unbiasedness*, which means, very roughly, that the expected overestimation matches the expected underestimation across the scenarios in Θ (Lehmann & Casella 1998, p. 83, definition 1.1).

- An even higher standard adds, to admissibility and unbiasedness, a property known as *UMVU*, short for *u*niformly *m*inimizing the *v*ariance among the *u*nbiased estimators, meaning that the relevant reliability is uniformly maximized across the scenarios in Θ among the unbiased estimators (Lehmann & Casella 1998, p. 85, definition 1.6).

Caveat: While admissibility is widely regarded as a minimum qualification in point estimation, unbiasedness remains somewhat controversial, despite its extensive coverage in almost all standard textbooks.[7] Indeed, determining the correct hierarchy of standards of reliability is an issue open to exploration and debate. Even so, the quest for the highest achievable seems to still lie at the heart of frequentist statisticians.

I thereby propose the following norm to capture an important aspect of the practice

[7] See Lehmann & Casella (1998, pp. 5, 157-158) for a controversy surrounding unbiasedness and a possible alternative to it (known as *median-unbiasedness*). Also see Jaynes (2003, sec. 17.3) for a discussion.

of frequentist statisticians in general:[8]

Frequentist Statistics as Reliabilism 2 (Achievabilist Reliabilism). For any problem context C, an inference method is justified in C only if it meets the highest standard of reliability that is achievable in context C—pending the specification of the correct hierarchy of standards.

Caveat: This statement is only meant to be a first approximation. Complications arise if the correct hierarchy is not a linear order but only a *partial* order (allowing for two incommensurable standards, of which neither is higher than the other, nor are they equal), or if there is no *uniquely* highest achievable standard (possibly because there are many or none), or if there is not even such a thing as *the* correct hierarchy. In any of those cases, the statement of the norm needs to be revised accordingly. Yet my point remains: Frequentist statisticians do not merely use standards of reliability to assess inference methods; they also strive for *a* (if not *the*) highest achievable one.

No statisticians explicitly state this norm at this level of generality, as far as I know. However, their textbooks are full of definitions of various standards of reliability, often indicating which ones are higher or lower, with numerous examples of problem contexts in which one or another standard is shown to be achievable or unachievable. Thus, the norm stated above does seem to capture an important aspect of their practice.

Frequentist statistics is therefore reliabilist not just in the sense of employing standards of reliability, but also in the sense of striving for a highest achievable standard of reliability in every context of inquiry.

[8]The first achievabilist norm stated at a high level of generality is due to Lin (2022), who develops a counterpart of the present statement in the setting of a non-stochastic theory of scientific inference, formal learning theory. Lin (forthcoming) extends the achievabilist norm to cover both the stochastic and non-stochastic settings simultaneously. A remark on the terminology: The achievabilist norm is called *the core thesis of learning-theoretic epistemology* in Lin (2022: 284). This name is nicely descriptive, as the achievabilist norm is indeed central to learning theory, including both formal learning theory in philosophy and statistical learning in machine learning. But this name has a downside: it might create the false impression that the spirit of striving for the highest achievable is unique to learning theory. In fact, this spirit is also core to frequentist statistics, as I have argued here. This is why I adopt the more neural term 'achievabilism', following the usage in Lin (forthcoming).

7 Closing

Perhaps it is not difficult to stretch the meaning of 'reliabilism' for the sole purpose of making it fail to imply externalism. That, however, is not what I do here. Instead, I broaden some related concepts and defend the thesis that reliabilism does not imply externalism for an important reason: to accommodate a natural and plausible interpretation of frequentist statistics, which represents one of the most sustained attempts by scientists to develop an epistemology for their own use.

Much more needs to be done to develop this internalist interpretation. First, there remains the task of explaining how background assumptions may be justified within one's context of inquiry, possibly following Annis (1978), who, like me, also advocates for the context-sensitive nature of justified beliefs. Second, while emphasis has been placed on the first-person perspective for assessing inference methods, this perspective can in principle be extended to the first-person *plural*, allowing the parameter space Θ to represent assumptions shared by the members of a community—the common ground of that community. Last but not least, I also suspect that if epistemology is continuous with science, psychology is not the only important junction, as Quine (1969) suggests; statistics is another. However, the details must be left for future work.

Acknowledgements

I am grateful to the participants of the Philosophy of Science and Epistemology Conference at the Hong Kong University of Science and Technology, held on June 27-29, 2024, for their valuable comments and questions. I am indebted to Jun Otsuka, Conor Mayo-Wilson, Konstantin Genin, and I-Sen Chen for discussions.

Bibliography

Annis, D. B. (1978). A contextualist theory of epistemic justification. *American philosophical quarterly, 15*(3), 213-219.

Bonjour, L. (2005). Internalism and externalism. In P. K. Moser (Ed.), *The Oxford handbook of epistemology* (pp. 234-263). Oxford University Press.

Casella, G., & Berger, R. (2002). *Statistical inference* (2nd ed.). Duxbury Press.

Fletcher, S. C., & Mayo-Wilson, C. (2024). Evidence in classical statistics. In M. Lasonen-Aarnio & C. Littlejohn (Eds.), *The Routledge Handbook of the Philosophy of Evidence* (pp. 515-527). Routledge.

Goldman, A. I. (1979). What is justified belief?. In G. S. Pappas (Ed.), *Justification and knowledge: New studies in epistemology* (pp. 1-23). Springer Netherlands.

Hacking, I. (1965/2016). *Logic of statistical inference.* Cambridge University Press.

Howson, C., & Urbach, P. (2006). *Scientific reasoning: The Bayesian approach.* Open Court Publishing.

Jaynes, E. T. (2003). *Probability theory: The logic of science.* Cambridge university press.

Lehmann, E. L., & Casella, G. (2006). *Theory of point estimation.* Springer Science & Business Media.

Lin, H. (2022). Modes of convergence to the truth: Steps toward a better epistemology of induction. *The Review of Symbolic Logic, 15*(2), 277-310.

Lin, H. (2024). To be a frequentist or Bayesian? Five positions in a spectrum. *Harvard Data Science Review*, 6(3), doi: 10.1162/99608f92.9a53b923

Lin, H. (forthcoming). Convergence to the truth. In K. Sylvan, E. Sosa, J. Dancy, & M. Steup (Eds.), *The Blackwell companion to epistemology* (3rd ed.). Wiley Blackwell.

Mackie, J. L. (1976). *Problems from Locke.* Oxford University Press.

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

Mayo, D. G., & Spanos, A. (2011). Error statistics. In P. S. Bandyopadhyay & M. R. Forster (Eds.), *Philosophy of statistics* (pp. 153-198). North-Holland.

Neyman, J. (1955). The problem of inductive inference. *Communications on pure and applied mathematics*, 8(1), 13-45.

Nozick, R. (1981). *Philosophical explanations.* Cambridge University Press.

Otsuka, J. (2023). *Thinking about statistics: The philosophical foundations.* Routledge.

Popper, K. R. (1959). The propensity interpretation of probability. *The British Journal for the Philosophy of Science, 10*(37), 25-42.

Quine, W. V. O. (1969). Epistemology naturalized. In *Ontological relativity and other essays* (pp. 69-90). Columbia University Press.

Ross, S. M. (2010). *Introductory statistics* (3rd ed.). Elsevier.

Rosner, B. A. (2016). *Fundamentals of biostatistics* (8th ed.). Cengage Learning.

Sober, E. (2000). *Philosophy of biology*, 2nd Edition. Westview Press.

Sosa, E. (1991). *Knowledge in perspective.* Cambridge University Press.

Stalnaker, R. C. (1968). A Theory of Conditionals. In Harper, W. L., Pearce, G. A., & Stalnaker, R. C. (eds.) *Ifs: Conditionals, Belief, Decision, Chance and Time.* Springer Netherlands: 41-55.

Steup, M. (2004). Internalist reliabilism. *Philosophical Issues, 14*, 403-425.

Williamson, T. (2000). *Knowledge and its limits.* Oxford University Press.

Appendix A: How $\mathbb{P}_\theta$ Is Determined

Although not required for the purposes of this paper, it is helpful to have a concrete picture of how $\mathbb{P}_\theta$ is determined for each $\theta \in \Theta$ without delving into too many technical details.

In the scenario where the proportion of the red balls is $\theta = 0.7$, the probability of obtaining a red marble in each draw is equal to 0.7, by clauses (i) and (ii) of IID Bernoulli (as presented in section 4):

$$\begin{array}{lcl} \mathbb{P}_{0.7}\big(\mathsf{Red}\big) & = & 0.7 \\ \mathbb{P}_{0.7}\big(\mathsf{Non\text{-}Red}\big) & = & 1 - 0.7 \end{array}$$

For simplicity, let the prescribed sample size be $n = 4$. Then the probability of obtaining a data sequence, say $(\mathsf{Red}, \mathsf{Red}, \mathsf{Non\text{-}Red}, \mathsf{Red})$, can be decomposed according to clause (iii) of IID Bernoulli:

$$\begin{aligned} & \mathbb{P}_{0.7}\big((\mathsf{Red}, \mathsf{Red}, \mathsf{Non\text{-}Red}, \mathsf{Red})\big) \\ = \; & \mathbb{P}_{0.7}\big(\mathsf{Red}\big) \cdot \mathbb{P}_{0.7}\big(\mathsf{Red}\big) \cdot \mathbb{P}_{0.7}\big(\mathsf{Non\text{-}Red}\big) \cdot \mathbb{P}_{0.7}\big(\mathsf{Red}\big) \end{aligned}$$

Combining the above results, we have:

$$\begin{aligned} & \mathbb{P}_{0.7}\big((\mathsf{Red}, \mathsf{Red}, \mathsf{Non\text{-}Red}, \mathsf{Red})\big) \\ = \; & 0.7 \cdot 0.7 \cdot (1 - 0.7) \cdot 0.7 \end{aligned}$$

This calculation procedure generalizes quite straightforwardly, determining for each θ a unique probability distribution $\mathbb{P}_\theta$ as a function that assigns nonnegative real numbers summing to 1 to the 2^n data sequences.

Then, the value of $\mathbb{P}_\theta\big(T \text{ rejects } H_0\big)$ can be defined and computed using the following procedure:

- *Step 1:* Start with any given test T and any given scenario θ, which is associated with a unique probability distribution $\mathbb{P}_\theta$, assigning probabilities to the 2^n data sequences.

- *Step 2:* Mark every data sequence that, if received, would prompt T to output '$\mathsf{Reject}\ H_0$'.

- *Step 3:* Find the probability that $\mathbb{P}_\theta$ assigns to each of those marked data sequences.

- *Step 4:* Sum these probabilities and return the result as the value of $\mathbb{P}_\theta\big(T \text{ rejects } H_0\big)$.

Thus, the probabilities involved in the urn example are all defined with respect to each parameter value θ.

Appendix B: Some Formal Definitions

This paper mentions several evaluative criteria in frequentist hypothesis testing. Their formal definitions are provided in this appendix for reference, with minimal interpretation only for the sake of readability, leaving the philosophically controversial terms uninterpreted.

Start with the minimum qualification, a low significance level, whose formal definition has already been provided in the main text but listed here for completeness:

Definition (Significance Level). A test T is said to achieve a (low) significance level at α iff T's probability of erroneous rejection is uniformly low in this sense:

$$\mathbb{P}_\theta\big(T \text{ rejects } H_0\big) \ \leq \ \alpha$$

for any scenario $\theta \in \Theta$ where H_0 is true.

The following is another evaluative criterion, which was only informally sketched in the main text:

Definition (Unbiasedness). A test T with a low significance level at α is said to be *unbiased* iff T's probability of correct rejection is uniformly not very low in this sense:

$$\mathbb{P}_\theta\big(T \text{ rejects } H_0\big) \ \geq \ \mathbb{P}_{\theta'}\big(T \text{ rejects } H_0\big)$$

for any scenario $\theta \in \Theta$ where H_0 is false, and any scenario $\theta' \in \Theta$ where H_0 is true.

There is also a schema that, while not corresponding to any single criterion, is useful for constructing new criteria from old ones.. Suppose we already have some criteria that narrow down the candidate pool to a class $\mathcal{C}$. We can then define an additional criterion as follows:

Definition (Uniform Maximum Power in a Class). A test T for hypothesis testing is said to be *uniformly most powerful* in a class $\mathcal{C}$ of tests iff, first, T belongs to class $\mathcal{C}$ and, second, T's probability of (correct) rejection is uniformly maximized in this sense:

$$\mathbb{P}_\theta\big(T \text{ rejects } H_0\big) \ \ \geq \ \ \mathbb{P}_\theta\big(T' \text{ rejects } H_0\big)$$

for any alternative test $T' \in \mathcal{C}$ and any scenario $\theta \in \Theta$ where H_0 is false.

Treat the class $\mathcal{C}$ in the above as a placeholder. Once we replace $\mathcal{C}$ by a candidate pool of tests delineated by some existing criteria, we can narrow down the candidate pool further by picking out those that are uniformly most powerful in $\mathcal{C}$. When $\mathcal{C}$ is set to be the class of the tests at significance level α, we obtain the highest standard in the hierarchy discussed in section 6.2; the second highest is obtained by letting $\mathcal{C}$ be the class of the unbiased tests at significance level α.

Reminder: all these standards are defined with respect to, first, a hypothesis H_0 slated for testing, and, second, a parameter space Θ, whose possible interpretations are crucial to the discussion of internalism vs. externalism in statistics.