Significance Testing in Theory and Practice

Daniel Greco

dlgreco@mit.edu

Formal Epistemology Workshop '09

Significance testing, a commonly used method for testing statistical hypotheses, has been harshly criticized by philosophers. Howson and Urbach (1993, p. 208) claim that "the results of a significance test, either of the Fisher or Neyman-Pearson variety, are often in flat contradiction to the conclusions which an impartial scientist or ordinary observer would draw" and that "although they are thoroughly fallacious, the methods of significance testing and classical estimation are still being advocated in hundreds of books, required texts in thousands of institutions of higher education, where hundreds of thousands of students are obliged to learn them." (1993, p. 252) While they make their criticisms of significance testing in the context of defending an alternative, Bayesian approach to statistical inference, you don't need to be a Bayesian true believer to think that the reasoning used in significance tests is fallacious in at least two ways.[1]

Nevertheless, in practice, most of us treat studies that use significance tests as reliable, and it seems as if we're reasonable in doing so. We're more likely to take a new drug if we read that medical studies (which typically use significance tests) support the claim that it's effective, we're more likely to vote for a new educational reform if statisticians tell us that it improves literacy, etc. While Bayesians like Howson and Urbach have alternative methods that they'd like to see used to evaluate statistical hypotheses, in the status quo these methods generally aren't used. In light of this, it looks as if those who are sympathetic to the criticisms of significance

[1] As I'll discuss later in the paper, Sober (2008) criticizes significance testing, though he is not a Bayesian, and accepts some of the standard criticisms of Bayesianism made by defenders of orthodox statistical methods like Mayo (1996).

testing made by Howson, Urbach, and others face an unpleasant choice: find fault with the criticisms, or become skeptical about a wide range of empirical scientific conclusions. And even if we're not worried about *whether* significance testing is reliable in practice (maybe a simple inductive argument is enough to support this conclusion), insofar as we accept the standard criticisms we might still be uncertain as to *why*. In this paper I'll offer an explanation; while significance testing seems to allow for hypotheses to receive spurious support, there are certain conditions under which the fact that a significance test would have us accept some hypothesis does constitute good grounds—both from a common-sense point of view, and a Bayesian one—for believing it. Furthermore, I'll argue that we have good reason to think that these conditions are generally met when significance tests are actually used. If I'm right, we can agree with the Bayesian about there being theoretical/foundational problems with significance tests, without being skeptical about empirical conclusions drawn on the basis of them.

The structure of this paper will be as follows. First, I'll explain some of the foundational views about probability that inspired the development of significance tests. I'll then explain how significance tests work. Next, I'll introduce two prima facie problems with significance testing—respects in which they apparently allow for hypotheses to receive spurious support. I'll consider each problem in turn, arguing that we have good reason to think that cases in which these features of significance tests lead to flawed evaluations of our evidence will be quite atypical.

1. Frequentism and Bayesianism

The debate between defenders of classical statistical methods and their opponents is in part a debate about in which contexts we can fruitfully use the notion of probability. To the Bayesian, probabilities can represent degrees of belief, and the problem of testing statistical

hypotheses is understood as a problem of how to move from a prior state of belief with respect to a hypothesis to a post-evidence state of belief. Defenders of significance testing tend to favor an alternative view known as frequentism, according to which we can only fruitfully talk about probabilities in the context of repeatable types of events; according to frequentists, the probability of a token event of a repeatable event type should be identified with the relative frequency with which that event would occur were the event repeated many times.[2] Frequentists understand the problem of hypothesis testing quite differently from Bayesians, since except in rare cases,[3] hypotheses aren't the sort of things that should be assigned probabilities according to a frequentist—because of this, the problem isn't conceived in terms of moving from a pre-evidence probability for a hypothesis to a post-evidence probability. Deborah Mayo explains this distinction and endorses the frequentist approach:

> As C.S. Peirce urged in anticipation of modern frequentists, what we really want to know is not the probability of hypotheses, but the probability with which certain outcomes would occur given that a specified experiment is performed. It was the genius of classical statisticians, R.A. Fisher, Jerzy Neyman, Egon Pearson, and others, to have developed approaches to experimental learning that did not depend on prior probabilities and where probability refers only to relative frequencies of types of outcomes or events. (1996, p. 10)

The statisticians Mayo refers to were aware of the possibility of allowing probabilities to represent degrees of belief. However, they tended to regard methods of statistical hypothesis testing based on this approach as too subjective. According to R.A. Fisher:

> Advocates of inverse probability [this was the traditional name for the use of Bayes theorem to generate posterior probabilities] seem to regard mathematical probability…as

[2] There are many difficulties associated with expressing this idea precisely. For a discussion, see Mellor (2005, chapter 3)

[3] When hypotheses themselves are the outcomes of repeatable event types, (for instance, if our hypothesis is that a given coin is fair, we may assign that hypothesis a probability if the coin was randomly drawn from a barrel of coins with some known proportion of fair coins) frequentists are fine with assigning them probabilities, and they agree with Bayesians about what should be said about such cases.

> measuring merely psychological tendencies, theorems respecting which are useless for scientific purposes. (Fisher, 1947, pp. 6-7, quoted in Howson and Urbach, 1993, p. 72)

Defenders of the Bayesian approach are often quick to point out that there are many respects in which frequentist statistical methods call for subjective input on the part of the researcher, and to argue that this vitiates any supposed objectivity advantage held by classical methods:

> In order to derive a unique conclusion from a Fisherian test of significance, arbitrary decisions need to be taken…As a leading advocate of Fisherian methods admitted, "[t]here is no answer to [the question 'Which significance test should one use?']…except a subjective one", adding in parentheses that it was "curious that personal views intrude always" (Kempthorne, 1971, p. 480). Indeed, it *is* curious, when one considers that Fisherian methods arose from a dissatisfaction with the Bayesian approach on account of its supposed subjectivity! (Howson and Urbach, 1993, pp. 192-3).

While I agree with Howson and Urbach, ultimately I'll argue that in a sense, it's the subjectivity of the classical methods that saves them from the putative problems I'll consider. That is, I'll argue that it's only because of the discretion researchers have in deciding when to employ significance tests, and the motives they have to only employ them when they're likely to lead to favorable results, that the conditions under which they actually are used are also conditions under which they are reliable. But I'm getting ahead of myself. Now that I've explained some of the background views that motivate significance tests, I can explain how they work.

2. *Significance Testing: Three Steps*

Using significance testing to analyze the results of an experiment is a three step process.[4] The first step involves formulating the null hypothesis. How do we decide which hypothesis is the null hypothesis? Actually, this is sometimes a difficult question, and one strand of criticism

[4] In this paper I'll discuss Fisherian significance testing. While I think similar remarks would apply to Neyman-Pearson methods, I don't explicitly argue for this in the body of the paper. Readers suspicious that my arguments are only of limited interest as a result should wait until they get to note 18 before making up their minds—I'll give some reasons to think that while discussing Neyman-Pearson methods would complicate things, it wouldn't change the main thrust of the argument.

of significance testing focuses on the arbitrariness involved in designating a particular hypothesis as the null, and the ways in which our conclusions can depend on this seemingly arbitrary choice.[5] I won't address this worry about significance testing here, except to mention that in many (perhaps most) cases, there's a natural choice for the null hypothesis. Intuitively, the null hypothesis is the boring hypothesis—the one according to which any apparently interesting results in the experiment are just due to chance. If we're interested in testing to see whether a treatment for a disease is effective, and we have a control group and a treatment group, the null hypothesis is the hypothesis according to which the treatment has no effect. What does it mean to say that results are just due to chance? I can't fully address this question here, but it should be taken at least to imply that in cases involving a control group and a treatment group, any discrepancies between the control group and the treatment group don't reflect a causal impact of the treatment.

The second step involves asking the following question:

> *Assuming that the null hypothesis is true, what is the probability of observing a value for the test statistic that is at least as extreme as the value that was actually observed?*"[6]

The test statistic is just the quantity measured in the experiment (or a quantity obtained from calculation using quantities directly measured in the experiment). In a study about whether the children in Horseshoe Creek have stunted growth, the test statistic might be the difference between the average height of children in that town and the national average. In some cases, completing the second step is relatively simple. For instance, in the above case, suppose the null

---

[5] (Sober 2008, p. 62) However, I do think some of my arguments later in the paper will suggest that the choice of the null hypothesis, while important, isn't completely arbitrary. Also, later in the paper I discuss a case in which Howson and Urbach argue that arbitrariness in the selection of the null leads to problems later in the paper, and I argue that while they're right about that case, there are important respects in which it is atypical--cases like it are not likely to arise in practice.

[6] "Statistical Hypothesis Testing," in Wikipedia, summarizing description in *Sage Dictionary of Statistics* (2004, p. 76)

hypothesis is that any differences in height between the children in Horseshoe Creek and the national average are due to chance variation (and not, say, due to pollution in the creek). If we know that that height in the national juvenile population follows a normal distribution, *and* we know the standard deviation from the national mean, and sample sizes are sufficiently large, then it's not hard to complete step two.[7] When we don't know these things (that is, when sample sizes are too small for the sample mean to be distributed normally, or when we don't know that the distribution is normal, or when we know it's normal, but don't know the standard deviation) things get more complex, and more sophisticated mathematical methods are used. But no matter how sophisticated the mathematical methods used to complete step two, the basic three-step structure of significance testing is the same.

The third step involves the decision of whether or not to reject the null hypothesis. This decision depends on the answer to step two. If the probability obtained in step two is below some predetermined critical threshold (often .05, or .01), the experimenter rejects the null hypothesis (i.e. accepts that the null hypothesis is false). If it is above the critical threshold, the experimenter does not reject the null hypothesis (which is not the same thing as accepting the null hypothesis).

To someone with broadly Bayesian sympathies, this talk of accepting and rejecting in binary terms can seem strange—we might instead be interested in a method that lets us assign a probability to the null hypothesis, rather than one that simply gives us a binary reject or don't reject answer. How to interpret the decision to reject the null hypothesis is actually a matter of

[7] Basically, we'd see how far the sample mean was from the national mean. We'd then convert the difference between the sample mean and the national mean into standard units (i.e., we'd see how many standard deviations the sample mean was from the national mean). Suppose the difference was *x* standard units. We'd then check to see what proportion of the area under the normal curve is more than *x* standard units away from the mean. That proportion is the probability that we should observe a value for the test statistic at least as extreme as the one we did, on the null hypothesis.

some controversy. Howson and Urbach bring a number of quotes from classical statisticians to bear in defense of the interpretation according to which rejecting a hypothesis amounts to resolving to behave as if it were definitely false, while not necessarily believing this. They go on to point out that significance tests don't rationalize such behavior—a hypothesis' being rejected in a significance test rarely warrants a scientist in betting his life on against a penny that it is false, though such behavior would be appropriate were the scientist certain that the hypothesis were false. (1993, pp. 203-6) However, Mayo is quick to argue against this interpretation:

> The misunderstanding concerns the construal of "accept" and "reject" on the behavioristic model. Actually, Neyman is quite clear on what he intends. Accept *H*, Neyman says, means to *take action A* rather than *B*. Accept *H* does not mean believe *H* is true. Accept *H* does not mean act as if you knew *H* was true, in the sense of behaving in any and all of the ways you would if you knew that *H* was true…Neyman's behavioristic model literally identifies the acceptance of *H* with the adoption of a decision to take some specific action *A* rather than *B* where *A* is set out at the start. (1993, pp. 369-70)

This interpretation of significance tests is often associated with their use in industry. A beer brewer might be interested in ensuring that his beer is of sufficiently high quality before shipping it to market. The brewer might designate the hypothesis that the beer is of acceptable quality the null hypothesis, and might associate this hypothesis with the action of shipping the beer to market. Rejecting the null would correspond to not shipping the beer and instead brewing a new batch. The brewer might select a sample of the beer, perform a significance test based on an analysis of its quality, and use this test to decide whether or not to accept the hypothesis that the beer is ready to ship--i.e., whether or not to sell the beer.[8]

[8] William Saley Gossett devised the t-test (a significance test designed for analysis of small sample sizes) based on his work for the Guiness brewery in circumstances like the ones described above. The Student's t-test (so-called because Gossett published anonymously as "Student" in order to protect Guiness' trade secrets) was used as a method for cheaply monitoring beer quality. (Mankiewicz 2000, p. 158)

While this interpretation avoids the problems discussed by Howson and Urbach, it's not clear how well it justifies the uses to which we'd like to put significance tests. In many cases, significance tests are used to justify an action *A* even when the decision to reject the null hypothesis wasn't associated with taking action *A* by the researchers who conducted the test. Suppose a school board is interested in deciding whether or not to implement some new method of instruction—call the action of implementing this method on a wide scale *A*. They find a study in which a significance test was used to analyze some experimental results in which the null hypothesis that the method of instruction didn't improve literacy was rejected. They decide to take action *A*, and the study is part of their grounds for doing so. This sounds like a paradigm case of the use of statistical studies to guide action, and (hopefully) it can be reasonable even if the researchers who performed the study had no idea that the school board was considering action *A*, and wouldn't have taken the study to justify *A*. It could be that the researchers thought that even if they rejected the null hypothesis in their study, taking action *A* would cost too much money to be worthwhile. That the researchers were of this opinion shouldn't undermine the school board's decision to use the study to support taking action *A*, at least not if studies that use significance tests are to be of much help to people other than the researchers who perform them. If significance tests are to play the roles we'd like them to, they must be allowed to guide action in a more general way than the one Mayo suggests; it's not enough to only allow them to guide us with respect to specific actions pre-set by the researchers.

Ultimately, Mayo rejects what she calls Neyman's behavioristic construal of acceptance and rejection spelled out above (though not for the reasons I've urged), in favor of an alternative, more epistemological understanding.

> The present account of testing licenses claims about hypotheses that are and are not indicated by tests without assigning quantitative measures of support or probability to those hypotheses…The Bayesian critic may persist that if I do not secretly mean to assign some number to the inferences licensed by my tests, then what do I mean by evidence indicating hypotheses? My answer is the one I have been giving throughout this book. That data indicate hypothesis *H* means that the data indicate or signal that *H* is correct, much as I might say that a scale reading indicates my weight…What does it mean to infer that *H* is indicated by the data? It means that the data provide good grounds for the correctness of *H*. (1996, pp. 409-10)

Perhaps uncharitably, I'll interpret Mayo as arguing that when significance tests would have us reject a hypothesis, we ought be relatively confident that the hypothesis is false (though the tests don't warrant any specific numerical degree of confidence). That is, on this interpretation significance tests provide qualitative constraints on our levels of confidence in the hypotheses they are applied to. I say that this is perhaps uncharitable because it may look too close to taking the goal of hypothesis testing to be finding out which hypotheses are probably true/false, rather than some other goal more congenial to the frequentist. However, without interpreting the upshot of significance tests along these lines, I don't see how they can be taken to guide action in the ways they are typically thought to be able to. That is, an adequate interpretation of what it is to accept that an educational method is effective should explain why taking such an attitude might justify implementing the method on a wide scale, but not betting one's life against a penny that doing so will improve literacy. If accepting that an educational method is effective means becoming pretty confident that it is effective while not being certain, then we have a straightforward explanation. Furthermore, it should do these things without requiring that accepting the hypothesis involves assigning it some precise probability. Perhaps it's lack of imagination, but other than my "qualitative constraints on levels of confidence," interpretation, I'm not sure what will do the job. In my defense, it's possible to find similar interpretations urged in classical statistics textbooks. For instance, in *Principles of Statistics*, M.G. Bulmer

writes that "the rejection of a hypothesis…provides good reason, in the sense of rational degree of belief, for supposing the hypothesis to be false, but no numerical value can be placed upon this degree of belief." (1979, p. 165)

So how would we use significance testing in the case of the children of Horseshoe Creek? We'd formulate our null hypothesis—in this case, it would be the hypothesis that any differences in height between the children of Horseshoe Creek and the national average were due to chance variation. We'd then collect our results—we'd measure the children, or look at the results of previously obtained measurements. Suppose we found that the children of Horseshoe Creek were, on average, 2.7 inches shorter than the national average. We'd then compute the probability that the difference between the average height of the children of Horseshoe Creek and the national average height should be *at least* 2.7 inches, assuming that any differences in height between these two populations are just due to chance variation. If this probability were below the critical threshold we'd reject the null hypothesis, and conclude that probably, the difference in height between these two populations is at least in part due to some factor(s) other than chance variation. In the next section, I'll consider some prima facie problems for this method.

3. *Two Apparent Fallacies in Significance Testing*

Consider the following, deductively valid instance of modus tollens:

P1. If the null hypothesis is true, then the value for the test statistic will not be at least as extreme as *x*.

P2. The value for the test statistic is at least as extreme as *x*. Therefore:

C. The null hypothesis is false.

If we're warranted in being certain in P1 and P2, then we're warranted in being certain in C. We can see significance testing as moving from this uncontroversial (at least if appropriately hedged) idea to the incorrect one that mere high probability in P1, combined with certainty in P2, warrants high probability in C. That is, the argument in a significance test might be represented as follows:

P1*. If the null hypothesis is true, then the value for the test statistic will *probably* not be at least as extreme as *x*.

P2. The value for the test statistic is at least as extreme as x. Therefore:

C*. *Probably*, the null hypothesis is false.

Elliott Sober calls this form of reasoning *Probabilistic Modus Tolens*, (hereafter *PMT*) and he points out that it is invalid. (2008, pp. 48-51) I'll go on to explain why, but first it's worth noting that what's important isn't how we represent the argument used in significance testing. A significance tester might agree that the argument from P1* and P2 to C* is invalid, and insist that his argument runs from P1 and P2 to C; he could do this if he allowed premises to be used in arguments not just when they were certain, but also when they had high but sub-maximal probability. His reasoning would be just as fallacious. The fallacy is committed whenever someone takes an occurrence that was unlikely on the assumption that some hypothesis is true to be enough to establish that the hypothesis is unlikely to be true—someone who makes this mistake might represent their inference as moving from P1* and P2 to C*, or as moving from P1 and P2 to C.

But what's wrong with *PMT*? I roll a die ten times. The sequence of numbers showing on the face of the die is as follows: 4, 4, 1, 3, 1, 3, 6, 3, 4, 3. Call this sequence *S*. Now, consider the hypothesis that the die is fair—each face is equally likely to come up, and the each roll is

independent from the rest. The probability that I should obtain sequence *S* upon rolling the die 10 times, on the hypothesis that the die is fair, is quite low (in particular, it's the same as the probability for any other particular sequence, 1/6 to the tenth power). But I did obtain sequence *S*. *PMT* would tell us to conclude that probably, the die isn't fair. But this would be silly (full disclosure: I actually obtained *S* using a random number generator). Given typical background assumptions, if you saw a die produce the sequence *S*, it would be unreasonable for you to think that it probably wasn't fair.[9] Significance testing seems to rely on *PMT*, and *PMT* is a fallacious form of argument, so we might worry that significance testing is unreliable.

One way of fleshing out this worry is that whether the occurrence of an event that was unlikely according to some hypothesis *H* counts as strong evidence against *H* depends on things that significance testing doesn't seem to take into account, such as whether there are other hypotheses that might explain the event, and how probable these hypotheses are relative to *H*. For instance, if there are no alternative hypotheses that provide a better explanation of the event than *H*, then the event may count as evidence *in favor* of *H*, rather than a reason to reject it, even though on the supposition that *H* is true, the event was very improbable. In such circumstances, it would seem that *PMT*-based significance testing would support spuriously rejecting the null hypothesis. Before responding to this worry, I'd like to introduce the second main (apparent) problem for significance testing I'll consider in this paper.

When carrying out step two of a significance test, we don't compute the probability that we should observe the results we actually did observe on the assumption that the null hypothesis

[9] Sober considers other cases of probabilistic modus tollens leading to rejection of warranted hypotheses. In particular, he points out that any probabilistic theory, given enough data, will assign a low probability to the total dataset, and so would be rejected under *PMT*. (2008, p. 51)

is true. Rather, we compute the probability that we should observe results *at least as extreme* as the ones we actually did observe, on the assumption that the null hypothesis is true. That is, rather than making our decision about whether or not to reject the null based on our actual evidence, we use a strictly logically weaker description of that evidence (that the observed value for the test statistic was x entails that it was at least as extreme as x, but the reverse entailment does not hold).

In many cases, significance testing couldn't get off the ground without this logically weakened description of the evidence. This is because sometimes the null hypothesis entails that each possible maximally specific result of the experiment is equally likely (for instance, the hypothesis that the die is fair entails that each possible specific sequence of rolls is equally likely). If we must make our decision about whether or not to reject the null based only on how likely the exact result is on the assumption that the null is true, then in cases like these, depending on the chosen threshold either every result would force us to reject the null, or none would. For example, in a case where we rolled the die 10 times, and used significance testing applied to the specific sequence of rolls, if our threshold was $> (1/6)^{10}$ we'd always reject the null. If the inequality went the other way, we'd never reject it. By using some logically weaker description of the evidence (e.g., that there were four 3's, three 4's, one 6, and two 1's) we can ensure that some results will be more likely, given the null, than others; while each specific sequence of rolls is equally likely, for combinatorial reasons, each distribution of numbers is not (e.g. there are more ways to get a sequence with three 1's, two 3's, and five 5's than there are to get a sequence with ten 1's). This way, whether we reject the null can depend what the outcome of the experiment turns out to be, rather than being mandated by the fact that each possible outcome is equally likely.

However, the fact that significance testing couldn't be fruitfully applied in certain cases without using logically weakened descriptions of the results of an experiment doesn't show that reasoning with logically weakened versions of one's evidence is legitimate. Two wrongs don't make a right—if significance testing didn't use *PMT*, weakening the evidence wouldn't be necessary. In general, weakening the evidence isn't ok.[10] Consider the following case. I'm at a party, and I can't remember who's hosting it, though it might be Sam. I pick a guest at random to introduce myself to, and lo and behold, it's Joe the plumber. I know that Sam hates plumbers and tends not to invite them to his parties (at least, he invites them at far lower rates than do other people who might be hosting the party). However, Sam loves Joe the plumber, and always invites him to the parties he hosts. The natural thing to say about this case is that when I run into Joe the plumber, I get evidence that the party is hosted by Sam (to make the case rock solid, we may assume that the other people who might be hosting the party hate Joe the plumber).

However, if instead of reasoning with my actual evidence, I use a logically weaker version of it, perhaps the proposition that I ran into some plumber or other, then my evidence may seem to support rejecting the hypothesis that Sam is hosting the party. After all, if Sam were hosting the party, the chance that a randomly selected guest would be a plumber would be quite a good deal lower than if someone else were (Sam generally doesn't like plumbers—other hosts tend to invite more of them). In general, there's no guarantee that in reasoning with weakened versions of one's evidence, one won't find oneself rejecting hypotheses spuriously—i.e., rejecting them only because one isn't taking into account one's total evidence. Insofar as one ought to accept or reject hypotheses based on what one's total evidence supports, significance

[10] This has been pointed out by a number of authors. Sober (2008, p. 53) discusses weakening the evidence in the context of Fisherian significance testing. White (2000) does so in the context of fine tuning arguments for multiple universes, and Kotzen (Draft) does so in his paper "Multiple Studies and Evidential Defeat."

testing seems misguided. It looks as if it commits the epistemic sin of *weakening the evidence*. Reasoning from the fact that some logically weakened version of one's evidence supports rejecting a hypothesis *H* to the conclusion that one's total evidence supports rejecting *H* amounts to committing the *Fallacy of Weakening the Evidence*.[11]

We've found two apparent fallacies that significance testing seems to commit: the *PMT Fallacy*, and the *Fallacy of Weakening the Evidence*. Both of these fallacies seem to allow that a significance test could recommend rejecting the null hypothesis—concluding that it is unlikely to be true—when a reasonable evaluation of one's evidence would point the other way. In what follows, I'll address these issues in turn. My response in both cases will be similar. First, I'll identify some conditions under which these forms of reasoning are valid—i.e., conditions such that, when they hold, reasoning with *PMT* and weakening the evidence won't lead you from true premises to false conclusions. I'll then argue that we have good reason to think that these conditions tend to hold when significance testing is actually used.

Before moving on, however, I'd like to say a bit about what I won't do. In this paper I'm mainly interested in giving reasons to think that in practice, researchers who use null hypothesis won't reject the null when the evidence doesn't warrant assigning it a low probability, i.e., won't commit are called type one errors. Statisticians are also concerned to design procedures that are unlikely to fail to reject the null when it is in fact false—failing to reject the null when it's false is known as a type two error. [12] In this paper, ultimately I'm interested in arguing that when significance tests tell us to reject the null, we should agree that the null is probably false—I

[11] Sober (2008, p. 53) lodges this complaint against significance testing, as do Howson and Urbach (1993, p. 176)
[12] Mayo explains the distinction in her (1996, pp. 159-60)

won't take up the question of whether significance testing too often fails to recognize evidence against the null.[13]

4. PMT and Alternative Hypotheses

Let's return to the Horseshoe Creek case. Suppose some parents, concerned about chemicals in the creek (suppose that children in the town tend to swim in the creek), have asked that studies be done to determine whether pollution is stunting the growth of their children. As before, the difference between the mean height of children in Horseshoe Creek and the national mean height is 2.7 inches. Now, suppose that given the number of children in Horseshoe Creek, and given how small the standard deviation in national height is, the probability that the difference between these means should be 2.7 inches, on the hypothesis that the difference is just due to chance variation, is vanishingly small, say, .003. In this case, it seems quite natural to think that good evidence has been obtained that some factor is stunting the children's growth. Why is this? Is there a way of explaining this that doesn't just rely on *PMT*?

In this case, collecting some data produces some very improbable results, and it seems as if we get strong evidence that the hypothesis that the results are just due to chance is false. In the case where a die is rolled 10 times, we get a very improbable result, but we don't get strong evidence that the result isn't due to chance. What's the difference? I'll first give an informal explanation of the difference, which I'll then put a Bayesian gloss on. In the Horseshoe Creek case, there's an alternative hypothesis—the hypothesis that something in the Creek is stunting the children's growth—which, if true, would make it quite likely that the results would be as they

[13] Ultimately, I suspect that the most troubling criticisms Howson and Urbach have to offer of classical statistical are the ones that suggest that significance testing is often too slow to recognize good evidence against the null hypothesis, rather than ones that suggest that significance testing makes it too easy to reject the null. See especially (Howson and Urbach 1993, sections 9.c6, 10.c4, 11.e-f)

actually are (i.e., that there would be a big difference between the Horseshoe Creek mean height and the national mean height). Furthermore, this hypothesis, even before seeing the data, isn't antecedently extremely implausible. In the die case, this doesn't hold. While there are alternative hypotheses on which the sequence *S* is very likely to arise, (for instance, the hypothesis according to which a genie is influencing the rolls, and he really loves that particular sequence) these alternative hypotheses are extremely implausible.[14] In general, when there's an alternative hypothesis according to which the evidence was quite likely, and that alternative hypothesis isn't itself too implausible, *PMT* is a good heuristic.

In Bayesian terms, when some hypothesis *H* assigns the evidence *E* a low probability (i.e., P(*E*|*H*) is low), and there's some alternative hypothesis *H'* such that P(*E*|*H'*) is high, then if the prior probability of *H'* is high enough, then the posterior probability of *H* given *E* must be low. To see this, take a simple case where the only two hypotheses assigned positive probability are *H* and *H'* (i.e., these are mutually exclusive and exhaustive hypotheses)[15]. Let's suppose that P(*E*|*H*) = .05 (typically P(*E*|*H*) must be $\leq$ .05 for results to be considered statistically significant, which is usually a necessary condition for their being considered publishable). Furthermore, let's

[14] However plausible the genie hypothesis is, there are other genie hypotheses that are equally plausible that apply to each specific sequence other than *S*. Because there are so many possible sequences, and none of the genie hypotheses is much more probable than any other, none of the genie hypotheses can get assigned a non-negligible probability.

You may challenge the claim that alternative hypotheses according to which *S* is likely to arise are implausible. While I think that such hypotheses are very implausible relative to typical background assumptions, I grant that one could have background information relative to which these alternative hypotheses would in fact not be very implausible. But in cases like these, I submit that it will not be counterintuitive to say that given that the sequence of rolls was *S*, one's evidence supports rejecting the hypothesis that the die is fair.

[15] This assumption may seem unrealistic—why shouldn't there by other hypotheses, besides the null and the alternative hypothesis, that get positive probability? If we set up our null hypothesis as "any differences between the control group and the experimental group are just due to chance variation" and our alternative hypothesis as "there are differences between the control group and the experimental group that are due in part to some factor other than chance variation", it's clear that they'll be exclusive and exhaustive. Even if we don't set up our alternative hypothesis this way—we might set it up as claiming that differences are due to some particular factor—it needn't be unrealistic to set up a case in which any other hypotheses are improbable enough so as to be such that their inclusion wouldn't significantly change the analysis.

assume that that P($E|H'$) = .95, i.e., that the evidence is quite likely on the assumption of the alternative hypothesis. The graph below plots the posterior probability for the null hypothesis *H* against the prior probability for alternative hypothesis *H'*. The *x* axis represents the prior probability of the alternative hypothesis *H'*, and the *y* axis represents the posterior probability P($H|E$).

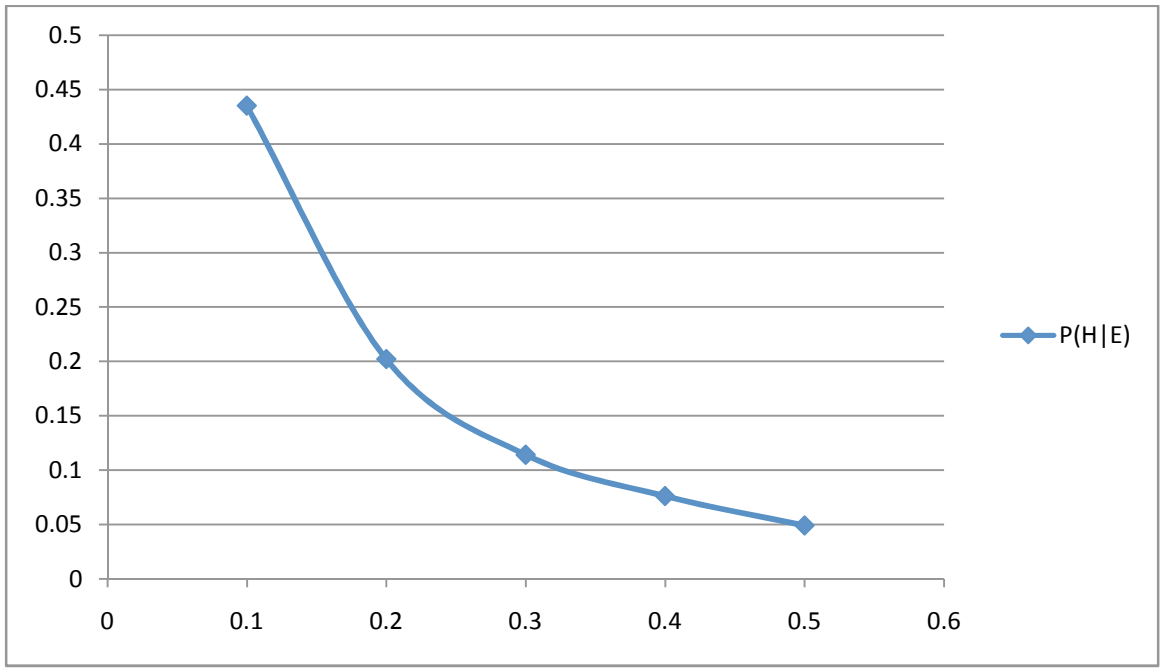

Figure 1

As the graph shows, even when the alternative hypothesis *H'* has a relatively low prior probability, if it makes the evidence *E* much more likely than does the null hypothesis *H*, the null hypothesis ends up strongly disconfirmed by *E*. As the prior probability of *H'* increases, the posterior probability of the null hypothesis given the evidence falls lower and lower.

This observation isn't new; many writers have noticed that the occurrence of an event that is regarded as unlikely according to a hypothesis *H* counts as good evidence against *H* when

there are plausible alternative hypotheses according to which the event is much more likely. Under the conditions that there are alternative hypotheses that predict the evidence with much higher probability, and these hypotheses aren't themselves too implausible, *PMT* is a reliable heuristic. Sober (2008, p. 57) makes essentially this point in his discussion of significance tests, and he found the point in Hacking (who was himself quoting Gossett). However, this observation isn't yet enough to comfort critics of significance testing—while it should be uncontroversial that under these conditions, a hypothesis' being rejected in a significance test really does give us good reason to regard it as unlikely to be true, we haven't yet been given reason to think that these conditions generally hold when significance tests are used. In the next section, I'll try to provide such reasons.

4.1 PMT, Predesignation, and the Decision to Perform a Significance Test

Before arguing directly that there's generally a plausible alternative hypothesis that assigns the results a high probability in cases where significance testing recommends rejecting the null, I'll need to take a brief detour. Results of experiments aren't extreme simpliciter. They are extreme in certain respects. Recall the sequence of die rolls *S*: 4, 4, 1, 3, 1, 3, 6, 3, 4, 3. Is this sequence extreme? The question is poorly formed. It is extreme with respect to how few 2's and 5's it contains—it couldn't contain any fewer, and the probability that a sequence of 10 rolls of a fair die should contain so few is quite low: $(2/3)^{10}$. However, it is not extreme at all with respect to the sum of the members of the sequence.

That experimental results are not extreme simpiciter, but are only extreme in certain respects might seem to raise a worry about significance testing. Almost all realistic sets of data will be extreme in *some* respect. Even for data that are intuitively "typical", there will usually,

nevertheless, be respects (perhaps unnatural, gerrymandered respects) in which they are quite extreme, and in which it was improbable that they should've turned out to be so extreme. In the coin case, this is obvious. We might worry that this means it's too easy to reject the null when it's true: just run an experiment, look for a respect in which the results are improbably extreme (there will almost always be at least one, even if the null is true), and then use a significance test to conclude that the null ought to be rejected. If significance testing really were so easy to exploit, we'd have reason to be worried. However, things are not so dire, and understanding why they are not so dire will help us see why the conditions under which *PMT* is reliable are generally met when significance testing is used.

In fact, statisticians are well aware that were it possible to inspect data after they had been collected and only then decide which hypothesis to subject to a significance test, it would be quite easy to reject the null; under such a testing regime, knowledge that a researcher had rejected a hypothesis in a significance test would be poor evidence that it was false. Egon Pearson (quoted by Mayo) writes:

> To base the choice of the test of a statistical hypothesis upon an inspection of the observations is a dangerous practice; a study of the configuration of a sample is almost certain to reveal some feature, or features, which are exceptional if the [chance] hypothesis is true. (Mayo 1996, p. 194)

This danger is typically guarded against by norms of predesignation, which require that researchers decide which hypothesis to test before inspecting the data. While such norms don't forbid forming new hypotheses after inspecting the data, if such hypothesis are to be subjected to

a significance test, they must be tested using a different set of data than the one that inspired them.[16]

One way of understanding the function of norms of predesignation is as requiring the following. Researchers must decide in advance which respect of extremity is such that, if and only if the results are extreme *in that respect*, and it was improbable on the assumption of the null that they should be so extreme in that respect, they'll reject the null. This requirement rules out using the process for automatically rejecting the null discussed above. Because researchers must decide in advance how they'll analyze their data, they can't collect data and then look for (possibly spurious) respects in which the results are improbably extreme.

Why does the fact that researchers must decide in advance what statistical test they'll apply to their data make it likely that the conditions under which *PMT* is reliable are generally satisfied when significance testing is used? I submit that being required to decide in advance that they'll reject the null if and only if results are (with respect to the null) improbably extreme in some particular respect *R* makes it probable that significance testers have in mind an alternative hypothesis *H'* that satisfies two conditions:

1. On the assumption of *H'*, it is highly probable that results should be improbably extreme in respect *R*. (that is, improbably extreme with respect to the null)
2. *H'* itself is relatively plausible.

Why is it natural to think that significance testers typically have in mind an alternative hypothesis *H'* that satisfies the two conditions above in mind? Just because if these conditions hold, then analyzing your data with the policy of rejecting the null in case results are improbably

[16] Mayo (1996, chapter nine) discusses such norms at length. Bulmer also says that a hypothetical violation of predesignation which he discusses "would clearly be cheating." (1979, p. 143)

extreme in respect *R* is a strategy that's likely to lead to interesting, publishable results (i.e., results on which you reject the null). If either of these conditions fails to hold—if there are no plausible alternative hypothesis according to which the results are likely to be improbably extreme in respect *R*—then a policy of rejecting the null just in case results are improbably extreme in respect *R* will look like a waste of time. That is, if the null hypothesis says that results are unlikely to be extreme in respect *R*, and no other plausible hypotheses say that results are likely to be extreme in respect *R*, then results are unlikely to be extreme in respect *R*, and running a significance test that only rejects the null under these conditions is unlikely to lead to rejecting the null.

The following example will help make this clear. Suppose there are two groups of people who have some disease, the control group and a group that's been given a treatment. The null hypothesis is that the treatment is ineffective. Furthermore, suppose we have some quantitative measure of how bad a given case of the disease is—days to recovery, perhaps. Assuming that the null hypothesis is true, the difference between the average number of days to recovery among the control group and the treatment group will probably be small. It's quite unlikely (suppose the probability is 5%) that it should be greater than two days (i.e., it's unlikely that the treatment group should recover on average more than two days faster than the control group, if the null hypothesis is true). On the alternative hypothesis, the treatment is effective in shortening recovery times, and it's quite likely that the difference between the average time to recovery in the control group and the treatment group will be at least two days. This situation is represented in figure 2. The dark and light curves represent, respectively, probability distributions over the difference between average recovery times in the two groups on the null hypothesis, and on the alternative hypothesis. That the dark curve peaks at zero means that, according to the null

hypothesis, the most likely difference between the two averages is zero days. That the light curve peaks at three means that, on the alternative hypothesis, the most likely difference between the two averages is three days.

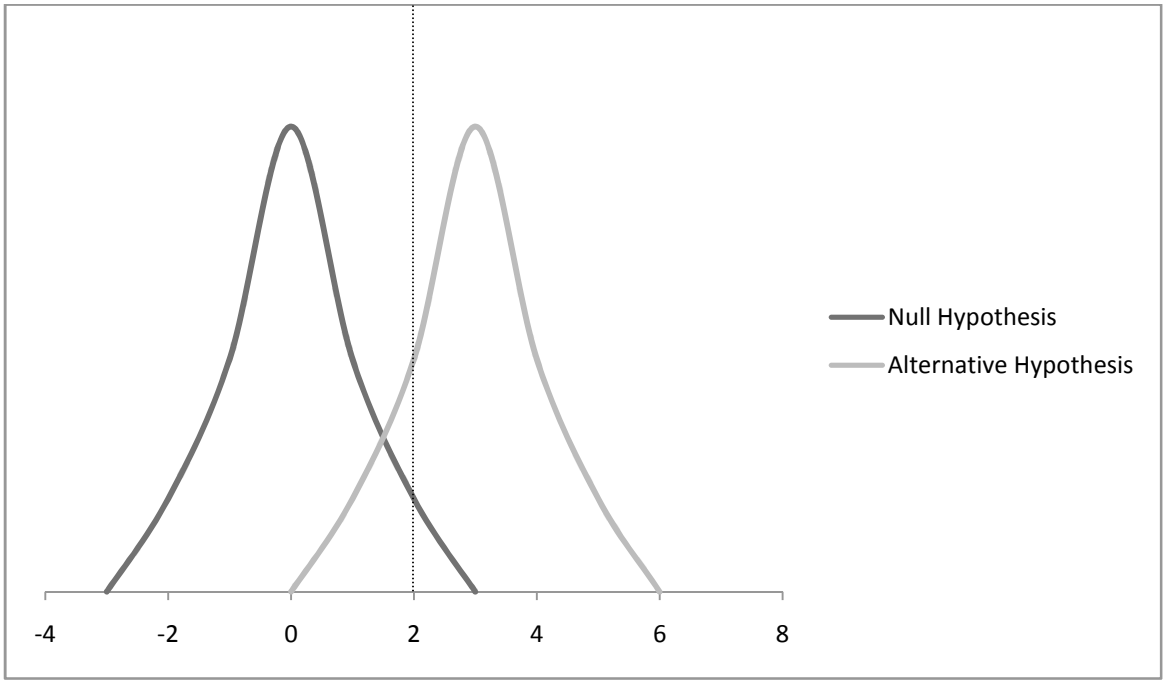

Figure 2

Now, suppose a researcher is interested in using a significance test with critical threshold .05 to evaluate the effectiveness of the drug. That is, the null hypothesis will be rejected if the treatment group recovers at least two days faster, on average, than the control group (remember, there was only a 5% chance on the null hypothesis that this would happen). The area to the right of the vertical line represents this "rejection region" i.e., the set of values for the difference between the average times to recovery in the two groups such that, if the actual value is in that set, the null hypothesis will be rejected.[17]

[17] In my example, the researcher would be using a one-tailed test—that is, the rejection region would comprise only one tail of the null hypothesis. Fisher favored two-tailed tests. Were a two-tailed test applied to the above case, the

Suppose the researcher thinks that the alternative hypothesis isn't all that implausible. In this case, performing a significance test will look like an attractive option. After all, on the assumption of the alternative hypothesis, it's quite probable that the result of the experiment will lead to rejecting the null hypothesis—most of the area under the curve representing the probability distribution on the alternative hypothesis is in the rejection region for the null hypothesis. This fact wouldn't be of interest if the alternative hypothesis were extremely implausible, but if the researcher thinks that there's a decent chance that the alternative hypothesis is true, then performing a significance test will look like a promising strategy for finding some interesting, publishable results.

But conditions like the ones described above are exactly the sort of conditions under which, as argued earlier, *PMT* is reliable. That is, when there's an alternative hypothesis according to which the evidence (in this case, an average recovery time in the treatment group at least two days faster than that in the control group) is quite likely, and that alternative hypothesis isn't too implausible, then it's a good inference to move from the premise that the evidence is improbable on the assumption of the null hypothesis to the conclusion that the null hypothesis

---

null would be rejected if the treatment group recovered on average at least two days faster, *or* at least two days slower, than the control group. (Actually, it'd have to be more than 2 days to maintain the .05 significance level—if a test were run that would reject the null if the average difference were more than 2 days in either direction, it would only be significant at the .1 level) However, modern statistical practice recognizes both one-tailed and two-tailed tests. If I'm right, whether researchers opt for one-tailed or two-tailed tests will depend on what they consider likely results of their experiments. A researcher who thinks it likely that a treatment will shorten recovery times, and very unlikely that it will hasten them, will opt for a one-tailed test; doing so will maximize his probability of rejecting the null. A researcher who thinks it both reasonably likely that a treatment will shorten recovery times, and reasonably likely that treated patients will react badly and will be sick for longer, but unlikely that it will have little or no effect, will opt for a two-tailed test. In *Principles of Statistics*, Bulmer offers some remarks along these lines: "A one sided range of alternative hypotheses gives rise naturally to a one-tailed significance test and a two-sided range of alternative hypotheses to a two-tailed test." (1979, p. 143). While Bulmer doesn't offer any explanation of these remarks, I think my explanation is the right one—we should understand the practice Bulmer refers to as involving researchers doing their best to design tests that will lead to significant, publishable results, in light of their judgments about the plausibility of various alternative hypotheses and how the probability distributions on these hypotheses differ from the probability distribution on the null. The decision of whether to use a one-tailed or two-tailed test is a nice illustration of a case where something like researchers' prior probabilities play an important (and salutary) role in classical statistical practice, even if they're not explicitly acknowledged.

probably isn't true. Because researchers must decide in advance what sort of extreme results they're looking for, and because they want to be able to publish interesting findings, (i.e., they want to be able to reject the null) they have incentives to only do significance tests when there are plausible alternative hypotheses according to which results are quite likely to be extreme in the relevant respects. These are the conditions under which *PMT* is not a fallacious and confused form of reasoning, but rather a reliable simplifying heuristic.[18,19]

---

[18] Statistically sophisticated readers might worry that my explanation of why Fisherian significance testing is epistemologically kosher is uninteresting, because Neyman-Pearson testing explicitly involves alternative hypotheses. Haven't I just said that Fisher tests implicitly involve what Neyman-Pearson tests explicitly do? No. While Neyman-Pearson testing does involve bringing in an alternative hypothesis, it doesn't bring in the prior probability of the alternative hypothesis—in particular, it doesn't require that the alternative hypothesis have a sufficiently high probability—nor does it require that the probability that the test statistic should be in the rejection region for the null, conditional on the alternative hypothesis, be high in absolute terms. My defense of Fisher testing, however, involved claiming that it's precisely because these conditions are met (even though they're not explicitly required) that Fisher testing doesn't lead us astray in rejecting the null. Because Neyman-Pearson testing doesn't involve prior probabilities, criticisms along the lines of *PMT* can be mounted, and a defense along much the same lines as the one I've given for Fisherian testing could be given. Similarly, Neyman-Pearson testing involves weakening the evidence in much the same manner as Fisherian testing, and I believe that my discussion of weakening the evidence in later sections of this paper would apply in much the same manner to Neyman-Pearson testing.

[19] It's also worth noting a point of contact between my argument above, and the debate about the epistemic significance of predicting experimental results versus merely accommodating them. A number of authors have been attracted to the idea that when a theory entails some experimental result, it counts more in favor of that theory if it predicted the result, rather than merely accommodated it (i.e., if the theory was designed so that it would be consistent with the result). See e.g., White (2003) and Lipton (2004). For authors on the other side, see e.g. Collins (1994) and Achinstein (1994).

A familiar point from this debate is that theories that predict tend to be more elegant and less ad hoc than theories that accommodate. After all, it's easy to get an empirically adequate theory by adding epicycles, but one can only do this when one is accommodating data—when predicting, one doesn't know which ad hoc epicycles one must add to end up with a theory that will entail the data.

While this observation doesn't settle the prediction/accommodation debate (authors are typically interested in whether prediction is better evidence for a theory than accommodation *holding fixed* things like how simple and elegant the theory is), it is relevant to my point. The illegitimate applications of significance testing discussed above, in which researchers first collect data and then look for respects in which they are extreme, are in a sense very much like the practice of accommodating rather than predicting data. Take a case where researchers first collect their data and then find only gerrymandered, unnatural respects in which the data are extreme with respect to the predictions of the null. In such a case, running a significance test and claiming that we should reject the null because the results are so improbably extreme is highly reminiscent of claiming support for a theory that accommodates a datum by incorporating ad hoc, baroque epicycles. My arguments in the text suggest a reason why—if rejecting the null is justified only when there's an alternative hypothesis that is itself relatively plausible that better explains the results, and hypotheses that predict that the results should be extreme in ad hoc, baroque respects tend to be implausible, then the reasons why snooping for significance can lead to spurious rejections of the null are much the same as the reasons why accommodating data by adding complex epicycles to theories can lead to spurious confirmation.

The above argument depends on taking the plausibility judgments of researchers to be generally reliable. In cases where this assumption fails, it's possible that a researcher might run a significance test because he has in mind an alternative hypothesis *H'* that he takes to be relatively plausible (but which in fact is not), and on which it is quite probable that the results of the test should be in the rejection region for the null. In such a case, the test might lead to rejecting the null (after all, flukes happen), even though there is no alternative hypothesis satisfying the two conditions mentioned earlier. In this case, the null would be rejected, even though a Bayesian would deny that good evidence against the null had been obtained. So the above considerations should comfort us only if we take researchers' judgments of the prima facie plausibility of the hypotheses they're interested in testing to be reliable. But I take it that this assumption is reasonable.

However, that the response to the *PMT* worry depends on such an assumption is significant for the foundational debates I mentioned earlier in the paper. As I noted in section 1, frequentists sometimes criticize Bayesianism for its supposed subjectivity. However, this criticism is dialectically ineffective if, in order to respond to the worry about probabilistic modus tolens, frequentists themselves need to appeal to considerations that are intuitively just as subjective as anything Bayesians require. And the assumption that researchers' prior probability judgments about the hypotheses they subject to significance tests are typically reliable seems to be just that. While it perhaps isn't subjective in exactly the same way that assigning prior probabilities to hypotheses is thought to be—it seems to amount to much the same thing; it amounts to relying on the reliability of the prior probability assignments of researchers.

The preceding considerations should allay the worry that significance testing leads to spuriously rejecting the null because it uses *PMT*. In the next section I'll take up the worry that significance testing may have us spuriously reject the null by weakening the evidence.

5. *Weakening the Evidence*

As the example of Joe the Plumber showed, sometimes one's total evidence can fail to support rejecting a hypothesis, even though a weakened version of one's evidence would support rejecting that hypothesis. Because significance testing seems to involve weakening the evidence, we might worry that it often leads researchers to reject the null hypothesis in situations when, were they to take their total evidence into account, they would not be reasonable in rejecting the null. In this section, I'll argue that this worry is misplaced—the cases where significance testing is typically used are importantly unlike the case of Joe the Plumber, where weakening the evidence significantly changes the direction in which the evidence points.

Let's consider a case in which weakening the evidence is uncontroversially harmless. Suppose one has a coin of constant but unknown bias—the probability of heads is unknown, but it is known that the probability doesn't vary from toss to toss. We toss the coin ten times in order to get some evidence about the bias of the coin. In this case, if we describe our evidence just in terms of how many heads and tails there were (rather than specifying the exact sequence obtained) and reason about how well our evidence supports various hypotheses about the coin's bias, we won't be making any mistake. The degree to which various hypothesis about the coin's

bias are supported or disconfirmed by evidence about how the coin landed depends only on how many heads and tails there were—not on the order in which they arose.[20]

There's a nice explanation of why this holds in the coin case. Take a logically weakened description of a sequence of tosses—that there were four heads and six tails, for instance. Each of the hypotheses under consideration (hypotheses about the direction and strength of the coin's bias) will agree that each possible specific sequence consistent with there having been four heads and six tails is equally likely. While they'll disagree on just how likely each of these sequences are, (e.g., the hypothesis that the coin is biased 0.6 in favor of tails will say that they're more likely than will the hypothesis that the coin is biased 0.9 in favor of heads) none of them will say that some of these sequences are more probable than others. When all the hypotheses under consideration agree that each possible specific body of evidence consistent with some logically weakened description of a body evidence is equally likely, then it doesn't make a difference whether we reason with the weakened description of our evidence or with the more specific version—the same hypotheses will be supported to the same degree either way.[21]

---

[20] Sober has a helpful explanation of this case. (2008, pp. 46-8) Classical statisticians would put this by saying that subject to the assumption that the probability of heads is constant from toss to toss, the proportion of heads is a sufficient statistic for estimating the bias—once the proportion of heads in a sample of tosses is known, no more information can be obtained from the sample that would help better estimate the bias of the coin. See Bulmer (1979, pp. 196-7) for a discussion of sufficient statistics. Howson and Urbach (1993) discuss sufficient statistics in a number of places, both to criticize the classical explanations of why they should be used when possible, and to offer what they take to be a more illuminating Bayesian alternative.

[21] PROOF:

Assume E is our weakened evidence, and there are n possible pairwise incompatible strengthenings of E: E_1, E_2… E_n. Furthermore, for each of the m hypothesis under consideration (i.e., each hypothesis that gets positive probability) H_i, $P(E_1|H_i) = P(E_2|H_i)$… $=P(E_n|H_i)$, though they needn't all agree on what the value for this probability is. We want to show that for each hypothesis H_i, $P(H_i|E) = P(E_i)$ for each E_i—that is, whether we update on the weakened evidence, or on some particular strengthening of it, the posterior probabilities of the hypotheses under consideration will be unchanged.

First, we'll prove that for each H_i, $P(H_i|E_1) = P(H_i|E_2)$…$= P(H_i|E_n)$. For any given H_k, for each E_i, by Bayes theorem $P(H_k|E_i) = P(E_i|H_k)P(H_k)/P(E_i)$. By the setup of the problem, $P(E_i|H_k)$ is the same for each E_i and $P(H_k)$ is a constant, so if we can show that $P(E_i)$ is the same for each E_i, then we'll have shown that $P(E_i|H_k)P(H_k)/P(E_i)$ is the same for each E_i. For any given E_k, $P(E_k) = P(E_k|H_1) + P(E_k|H_2)$…$+ P(E_k|H_n)$. But by the setup of the problem,

In the Joe the plumber case, it is not the case that all the hypotheses under consideration (that it's Sam hosting the party, and that it's someone else) regard each possible specific body of evidence consistent with the evidence that I ran into a plumber as equally likely. On the contrary, assuming that Sam is throwing the party, some propositions you get by strengthening the claim that I ran into a plumber are much more probable than others. (That is, it's much more probable, assuming that Sam is throwing the party, that I should run into Joe the plumber than it is that I should run into Suzy the plumber, or Steve the plumber, etc.) When some of the hypotheses under consideration regard some possible strengthenings of a weakened description of one's evidence as more likely than others, there's no guarantee that reasoning with the weakened version of one's evidence is legitimate—some hypotheses might be confirmed, even though they'd be disconfirmed (or confirmed to a significantly smaller degree) if you reasoned with your total evidence.

It would be nice to be able to apply this lesson straightforwardly to the case significance testing. Sadly, things aren't so easy. In a significance test, it's almost never the case that all hypotheses under consideration agree that each way of strengthening the claim that the test statistic was at least as extreme as x is equally likely. Nevertheless, I'll argue that something almost as good holds.

Let's return to the case represented in figure 2. Suppose the treatment group actually recovered 2.5 days faster, on average, than the control group. In this case, a significance tester

---

conditional on each H_i, $P(E_1) = P(E_2)\ldots=P(E_n)$. So the values of the terms in the series $P(E_i|H_1) + P(E_i|H_2)\ldots+ P(E_i|H_n)$ must be the same for each E_i. So $P(E_i|H_k)P(H_k)/P(E_i)$ is the same for each E_i, so $P(H_i|E_1) = P(H_i|E_2)\ldots= P(H_i|E_n)$.

Now, For each hypothesis under consideration H_i, $P(H_i|E) = \Sigma_1^n P(H_i|E_i)$. By the result of the previous paragraph, this implies that $P(H_i|E) = P(H_i|E_i)n$. Also, $P(E) = P(E_i)n$ for each E_i. This is because each E_i has the same probability, they are pairwise incompatible, and their disjunction is equivalent to E. So $P(H_i|E) = P(E|H_i)P(H_i)/P(E) = P(E|H_i)P(H_i)/P(E_i)n = P(E_i|H_i)(n)P(H_i)/P(E_i)n = P(E_i|H_i)P(H_i)/P(E_i) = P(H_i|E_i)$. QED

will compute the probability that the difference was *at least as extreme as* 2.5 days. Figure 3 is just like figure 2, but with the shaded region representing the set of values for the difference that are at least as extreme as 2.5 days. All the points in the shaded region are to the right of the vertical line—this is because all the points in the shaded region are in the rejection region for the null—if the value for the difference lies in the shaded region, the null will be rejected. If the shaded region represents the weakened version of the evidence, particular vertical lines in the shaded region would represent the possible strengthenings of this evidence—e.g., that the difference between the average times to recovery was 3 days, 3.5 days, 4 days, etc.

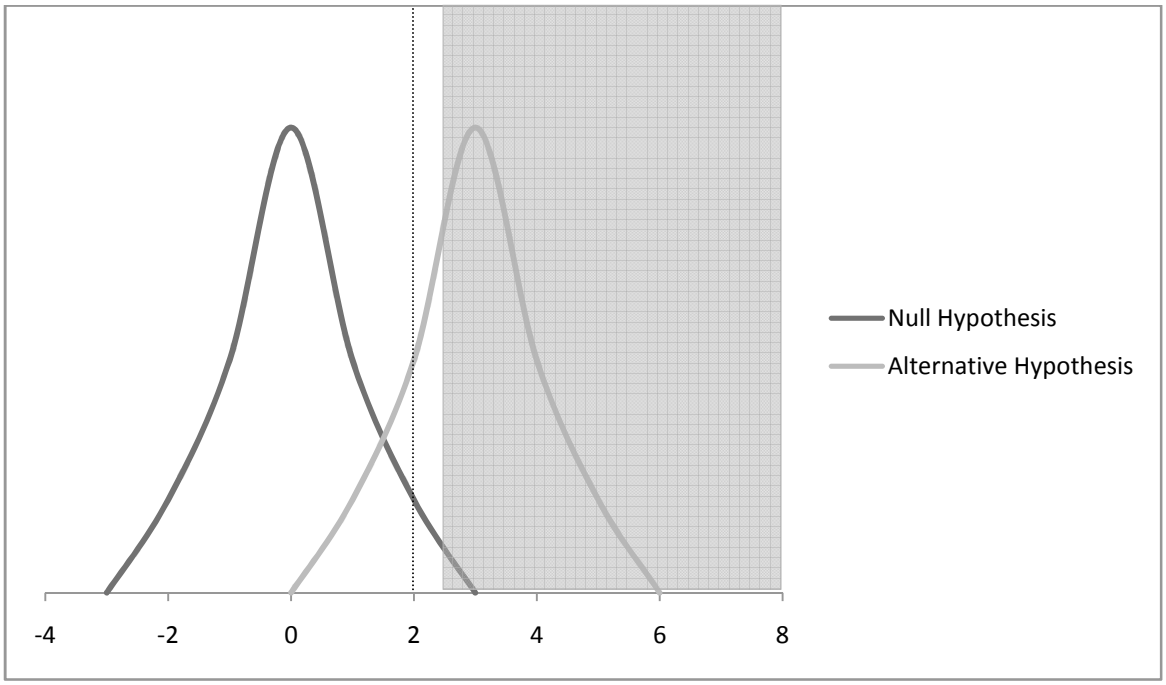

Figure 3

It's clear that, unlike in the coin flipping case, the hypotheses under consideration don't regard the various possible strengthenings of the weakened version of the evidence as equally probable. For instance, on the alternative hypotheses it's more probable that the difference should be 4 days than it is that the difference should be 6 days, and we can see this by noting that

the curve representing the probability distribution on the alternative hypothesis is higher when $x = 4$ than when $x = 6$. Similarly, on the null hypothesis, it's more likely that the treatment group should recover on average 2.5 days faster than the control group than it is that they should recover 4 days faster.

However, if we just focus on the shaded region, we notice an interesting property. For each value in the shaded region, the probability that the difference in average recovery times should take that value is significantly higher on the alternative hypothesis than on the null hypothesis. If we assume that the actual value is in the shaded region, then how much the alternative hypothesis is supported over the null won't change much depending on where in that region it is. When only two hypotheses *H* and *H'* are in play, what matters (in addition to their prior probabilities) for determining how likely they are in light of some body of evidence *E* isn't the absolute probabilities they assigned to that body of evidence, but instead the likelihood ratio $P(E|H)/P(E|H')$. In particular, if all the specific values *x* for the difference in average recovery time compatible with the claim that the value is in the shaded region are such that the probability of *x* given the alternative hypothesis is much higher than the probability of *x* given the null hypothesis, then it won't matter much whether we update with the evidence about the specific value, or just the weaker evidence that it was somewhere in the shaded region.

I take it that in this case, it's clear that rejecting the null based on weakened evidence isn't problematic; any strengthening of the evidence would also lead to rejecting the null because of the points made about the likelihood ratio above. But why should we think that this case is typical in that respect? This is the question I'll be concerned with in the remainder of the paper.

First, most of the area under the curve representing the probability distribution on the alternative hypothesis is to the right of the vertical line—it's in the rejection region for the null hypothesis. This just means that it's quite probable, on the alternative hypothesis, that the result of the experiment will be (relative to the null hypothesis) improbably extreme. For the reasons discussed in the previous section, there's good reason to think that this feature of the example will be typical—normally, most of the area under the curve representing the probability distribution on the alternative hypothesis will be in one of the tails of the curve representing the distribution on the null.

Another feature of this example is that the alternative hypothesis assigns a significantly higher probability than the null hypothesis to all the values in the right-hand portion of the shaded region. This feature of the example is critical—in cases where it fails, it can be the case that the weakened version of the evidence (i.e., the fact that the value for the test statistic was in the shaded region) supports rejecting the null, while the actual evidence would not. Consider a different example represented by the following chart:

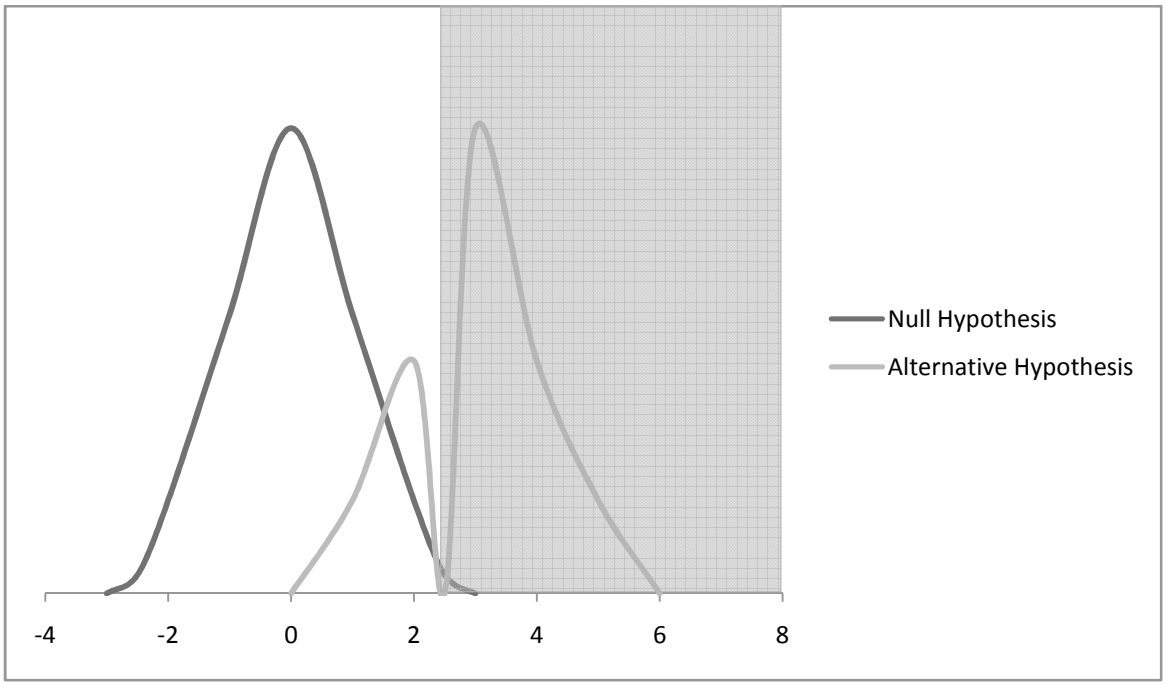

Figure 4

In this case, the actual value for the test statistic was 2.5. The weakened version of the evidence—that the value for the test statistic was at least as extreme as 2.5—supports the alternative hypothesis. The region in which the value for the test statistic is at least as extreme as 2.5 is shaded, and it's clear that there's far more area in that region under the curve representing the probability distribution on the alternative hypothesis than there is under the curve representing the probability distribution on the null hypothesis. However, the actual evidence—that the value for the test statistic was 2.5—does not support the alternative hypothesis. The probability distribution for the alternative hypothesis takes a nose dive at 2.5. In fact, while it's quite unlikely on the null hypothesis that the test statistic should take the value of 2.5, it's even more unlikely on the alternative hypothesis. This case is like the example of Joe the plumber—the weakened version of the evidence supports rejecting the null, but the actual evidence does not—in fact, it supports rejecting the alternative hypothesis in favor of the null. We should hope that cases like these aren't typical—if they are, then significance testing will lead us astray when

the test statistic takes a value *x* such that values at least as extreme as *x* are more likely on the alternative hypothesis, but *x* itself is more likely on the null. In the remainder of this section, I'd like to consider some examples that will hopefully make it plausible that cases like the one represented in figure 4 are unusual—that is, realistic cases will be ones where the probability distribution on the alternative hypothesis doesn't have any steep valleys.

Suppose we took figure 4 to represent a case like the one in figure 3—one involving a treatment group and a control group, where the test statistic is the difference in average recovery times. In this case, the valley in the curve representing the probability distribution on the alternative hypothesis would mean that according to the alternative hypothesis, the following is the case: it's not that unlikely that people should recover about two days faster if they take the treatment, and it's quite likely that they should recover three or four days faster if they take the treatment, but it's extremely unlikely that people should recover in the neighborhood of two and a half days faster if they take the treatment. I take it that this is a bizarre hypothesis—normal drugs don't work this way. Let's consider some other examples.

Suppose figure 4 represented a case like the Horsheshoe Creek one. In this case, according to the alternative hypothesis, it would be quite likely that children from Horseshoe Creek should be on average three or four inches shorter than the national average, and not that unlikely that they should be two inches shorter, but extremely unlikely that they should be two and a half inches shorter. Again, this would be a bizarre hypothesis. Pollutants don't affect people's heights in such an irregular manner. Lastly, suppose figure 4 represented a case involving two groups of students, each of which was given a test, one of which was taught using an experimental educational method, and the other of which was a control group. The null

hypothesis would be that the method doesn't improve test scores. The test statistic would be the difference between the average score in the control group and the average score in the experimental group. Here, the alternative hypothesis would be that the educational method was pretty likely to improve scores by three or four points on average, and not unlikely to improve scores by two points, but extremely unlikely to improve scores by two and a half points. It's hard to imagine a realistic case where this could be a reasonable hypothesis about the likely effects an educational method might have.

Why is it that typical cases don't involve alternative hypothesis on which there are large regions of values for the test statistic that are very favorable to the alternative hypothesis, but smaller regions within those large regions that are much less favorable to the alternative hypothesis (or even favorable to the null)? Why aren't narrow valleys more common? An unsatisfying answer would involve pointing out that many of the probability distributions of interest to empirical researchers are (approximately) normal, and normal distributions never have valleys.[22] Why is this answer unsatisfying? For one, in some cases on some hypotheses under consideration the probability distribution for the test statistic will not be approximately normal. But even typical non-normal distributions studied by statisticians (such as the Cauchy distribution) don't have steep valleys of the sort represented by the probability distribution on the alternative hypothesis in figure 4. What we'd like would be a general explanation of why we should expect narrow valleys to be rare—one whose assumptions about the character of the hypotheses under consideration are as minimal as possible. I believe such an explanation is available, and I'll introduce the machinery necessary to give it in the next section.

[22] Any discussion of the Central Limit Theorem, such as that in Bulmer (1979, pp. 115-20) will shed some light on why so many distributions are normal.

5.1 Natural Properties and Natural Quantities

The idea that certain ways of grouping objects are more natural than others—that some but not other classifications "cut nature at its joints"—has an old pedigree in philosophy.[23] Another way of putting the idea that some classifications are more natural than others is by saying that certain respects of similarity are distinguished compared to others. For instance, we might think that two objects that are both electrons, or both horses, are truly similar to one another, but two objects that were both once within seven yards of someone with a moustache are not (or at least, they aren't similar just by virtue of sharing *that* property). I think that something along the lines of this idea can help answer the question of why cases like the one represented in figure 4 are atypical—that it can explain why we shouldn't worry that weakening the evidence in the context of significance testing will lead to spuriously rejecting the null.

What's supposed to follow from the thought that certain respects of similarity are distinguished? Some philosophers have put this idea to quite controversial uses.[24] But many philosophers—even some who would probably balk at the more metaphysically loaded applications of this notion—have thought that this idea is helpful in thinking about induction.[25] Consider the following argument schema:

P1. Objects $o_{1,}\, o_2 \ldots o_n$ are all both F and G

P2. Object o_{n+1} is F

C. Object o_{n+1} is G

---

[23] It goes back to Plato's Phaedrus:

> And what is the other principle, Socrates?
> That of dividing things again by classes, where the natural joints are, and not trying to break any part, after the manner of a bad carver. (Plato 1925, 265d-e)

[24] E.g., Lewis (1984), Sider (Forthcoming)

[25] E.g, Quine (1969)

It’s plausible that P1 and P2 typically provide good inductive support for C when F and G are natural properties—properties the sharing of which makes for real similarity—but not when one of F or G is not a natural property. For instance, suppose we think that being an emerald and being green are natural properties, but that being grue—green if observed before January 1, 2050 and blue if observed afterwards—is not. Furthermore, suppose it’s January 1, 2050, and we’re about to check the color of a previously unobserved emerald (all previously observed emeralds have been green). If we plug in “emerald” for “F” and “green” for G, we get an inductive argument that looks like a good one; if the case is typical, it is reasonable for us to expect the new emerald to be green. But if we plug in “grue” for G, we get what looks like a bad argument; we shouldn’t expect the new emerald to be blue.

I suggest that it may be fruitful to apply something like this idea not just to binary properties that objects either have or don't, but also to quantities such as mass or volume. How would this work? Well, suppose some object o_1 has n units of some quantity, o_2 has $n + 1$ units of this quantity, and o_3 has $n + 10$ units of this quantity. If the quantity is a natural one, then o_1 is more similar to o_2 in a natural respect than it is to o_3. But if the quantity is not a natural one, then we won't think that o_1's differing from o_2 only by one unit of the quantity really makes it any more similar to o_2 than to o_3. So far this is very vague—what's the cash value of such intuitions about similarity? The following example will help.

Suppose we've taken a number of measurements of a class of objects, and the following graph represents the data compiled so far.

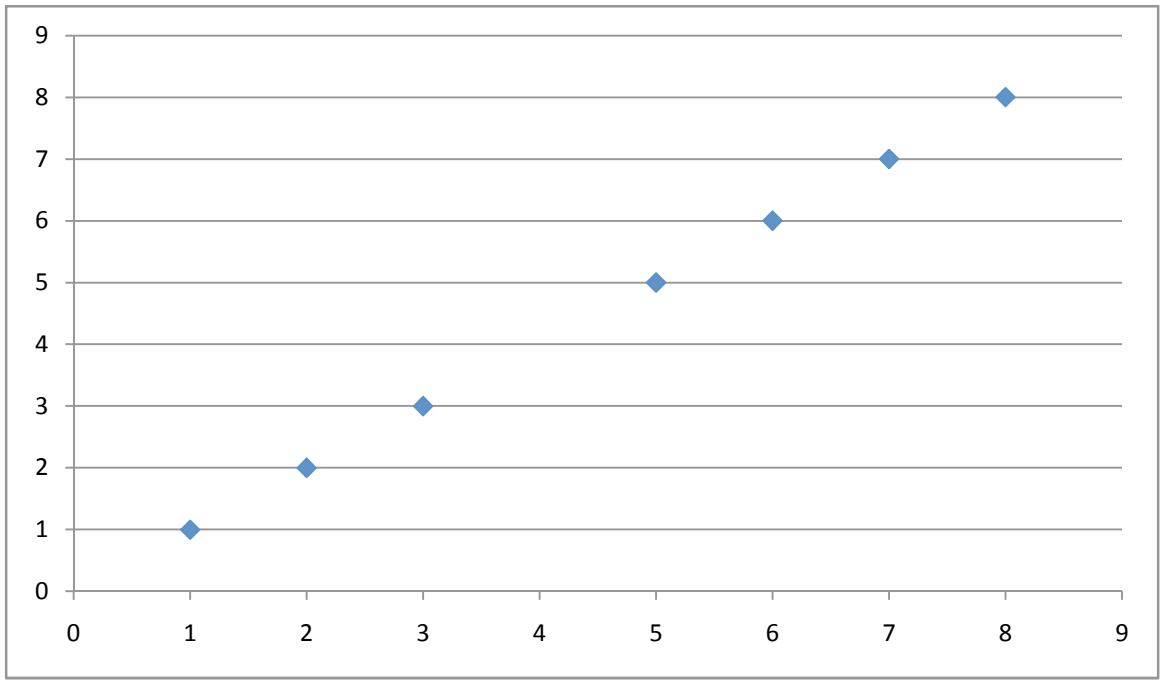

Figure 5

In this case, it seems natural to predict new data points by fitting a linear curve to the data—we might expect that if we measure an object in this class whose *x* value is 4 units, the *y* value will also be 4 units. In fact, there are sophisticated methods for fitting curves to sets of data points like the one above to make predictions about new data points, and such methods often favor curves that require fewer parameters to specify—if they fit the data equally well, a line will be preferred over a parabola, which will be preferred over yet more complex curves.[26] But judgments based on such methods are only credible—even with appropriate *ceteris paribus* hedges—when the quantities represented on the *x* and *y* axes are natural quantities. For instance, suppose the data comes from an experiment in which hamsters were fed various amounts of food, and then weighed. Suppose the *x* axis represents the amount of food the hamsters were fed each day (in ounces), and the *y* axis represents weight in ounces after one year. In this case,

[26] Elliott Sober offers a sympathetic discussion of one such method—the Akaike Information Criterion. (Sober 2008, pp. 82-96)

ceteris paribus, it seems quite reasonable to infer that a hamster fed 4 units of food per day will weigh 4 ounces after one year. But suppose the *y* axis doesn't represent weight, but instead represents queight. An object's queight is defined as follows:

Queight(x) = 100 quounds if weight(x) = 4 ounces,
4 quounces if weight(x) = 100 pounds
Weight(x) otherwise

That is, most objects weigh exactly as much in pounds as they queigh in quounds, but objects that weigh 4 ounces queigh 100 quounds and vice versa. If the *x* axis represents ounces of food per day, and the *y* axis represents values for queight, it would be quite unreasonable to fit the data with a straight line running through all the points. That is, if the hypothetical data set above would make it reasonable to expect that a hamster fed 4 ounces of food a day will weigh 4 ounces, then if our graph plotting units of food per day against queight, we won't want to fit our data with a line, even though doing so would hit all the data points exactly. Except in very atypical circumstances, we shouldn't expect that feeding a hamster 4 units of food a day will produce a behemoth, while feeding it slightly more or slightly less food will not. I hope it will also seem plausible that queight is not a natural quantity. Take three objects, *x*, *y*, and *z*, which weigh 3.99 ounces, 100 pounds, and 4.01 ounces respectively. Intuitively, *x* and *z* are quite similar to one another, and both of them are quite dissimilar to *y*. If we measure similarity by queight, the three objects are practically peas in a pod, and *y* counts as more similar to *z* than *x* does.

As outlined above, one potential application of the idea that some quantities are more natural than others is that it's only when our axes represent relatively natural quantities that we should try to predict future data points by fitting simple curves to already collected data. This

isn't the application, however, that I'll suggest can aid in explaining why we needn't worry about weakening the evidence in the context of significance testing; I introduced the above application to make it plausible that using the notion of natural quantities in thinking about inductive practices isn't an unmotivated, ad hoc move.

5.2 Natural Quantities and Weakening the Evidence

In his paper "Inferring Probabilities from Symmetries", (1998) Michael Strevens addresses questions such as how it is that we manage to correctly infer that a die is as likely to land on any one of its faces as on any other when tossed (a claim about probabilities) from the fact that it is perfectly symmetrical (a claim about physical symmetries). In the context of discussing such questions, he introduces the notion of a probability distribution being "smooth, in the sense that it does not fluctuate rapidly." (1998, p. 237) In particular, a smooth probability distribution wouldn't display any narrow valleys like the probability distribution for the test statistic on the alternative hypothesis in figure 4, nor would it display steep climbs to pointy peaks. He goes on to make some general remarks that I believe are directly applicable to the issue discussed in this paper:

> Call the kinds of variables in terms of which we usually work our "standard" variables. It seems to be the case that, for whatever reason, our standard variables are usually smoothly distributed. If we go ahead and generalize from this observation (by enumerative induction), we arrive at the conclusion that most standard variable distributions are smooth. We may consequently take ourselves to have empirical grounds for adopting a revised and differently deployed "Principle of Insufficient Reason" of the following form:
>
> > *In the absence of any reason to think otherwise, assume that any standard variable is fairly smoothly distributed*. (Strevens 1998, p. 241)

I take it that Strevens is making claims both about standard variables being smoothly objectively distributed, and smoothly subjectively distributed. That is, Strevens is claiming that typically,

physical systems produce outcomes such that the objective probability distributions over those outcomes (when the outcomes are described using standard variables) are smooth, and that we know this by induction. Furthermore, because we know this, and because we accept something like Lewis' Principal Principle (Lewis 1987),[27] our subjective probability distributions over the outcomes of chance setups should also typically be smooth when those outcomes are described in standard variables and when we don't have any special information about the outcomes (e.g., we haven't yet observed them).

While I take this to be the correct interpretation of Strevens, and by and large I agree with the claims I take Strevens to be making, it does raise some questions—for instance, we might have a notion of objective chance such that there can be no non-trivial objective chances in a deterministic universe. If this is our notion of objective chance, then Strevens' claim that standard variables are typically smoothly objectively distributed will look highly contentious—it will imply indeterminism. For this reason, I suspect Strevens is working with a notion of objective chance such that non-trivial objective chances are compatible with determinism.[28] Luckily, details like this won't affect my project—for my purposes, all that's important is that it typically be the case that our subjective probabilities over standard variables should be smoothly distributed. I take it that this claim is plausible independently of whether or not it can be supported by claims about objective probabilities being smoothly distributed together with the Principal Principle.

[27] Roughly, the Principal Principle requires that in the absence of special information, our beliefs about the objective probabilities of various outcomes should line up with the subjective probabilities we assign those outcomes

[28] Roger White suggests that we have a notion of objective chance that's compatible with determinism, (2007, p.4) and I'm inclined to agree, but I don't think I need to commit myself one way or the other for my purposes here.

Using the terminology already developed, I'd say that our standard variables typically represent natural quantities, and that best hypotheses typically induce smooth subjective probability distributions over natural quantities (whether this is because the Principal Principle holds, and physical systems typically produce objective probability distributions that are smooth over natural quantities, is not a position I intend to take a stand on). If this idea is right, then the application to the case of weakening the evidence is relatively straightforward. If the test statistics used in significance tests represent natural quantities, and natural quantities are typically subjectively smoothly distributed, then cases in which some of the hypotheses under consideration induce subjective probability distributions on the test statistic that display narrow valleys—cases like those represented by figure 4—will be atypical.

None of what I've written so far commits me to any particular view about the metaphysical status of natural quantities. Strevens writes: "let me stress that I am not proposing that our "standard" variables have any special logical status. They are simply the variables with which we prefer to operate, and which are, conveniently for us, for the most part smoothly distributed." (1998, p. 241) I agree—it doesn't matter for my purposes whether we think that natural quantities represent objective metaphysical joints in nature. We might instead think that all it is for some quantity to be a natural quantity is for it to play a certain role in our best scientific theories. Alexander Bird and Emma Tobin (2008) suggest that Quine's view about natural properties was something like this:

> Quine thinks that in due course science will obviate the need for a general notion of similarity or kind: in each area of science more specific notions will take the place of the generic notion; this is a sign of the maturity of a branch of science. For example, in zoology we may replace talk of the similarity between two animals by discussion of the historical proximity of their closest common ancestor.

One could take a similar line about natural quantities—that is, one might think that whether a quantity is a natural one is a domain-relative question. Mass might be a natural quantity in the context of physics, but not in the context of psychology, where knowing that two people are roughly equally massive sheds little light on what psychological properties they might share.

Either way, debates about the metaphysical status of natural quantities aren't directly germane to the applications I want to make of them. All that's required for my purposes is that test statistics in significance tests (such as the difference in average time to recovery between a control group and a treatment group) measure natural quantities, and that our best hypotheses typically induce smooth probability distributions on potential values for natural quantities. How does this help significance testing with its putative problem of weakened evidence?

If my claims about smoothness and natural quantities are right, then typical cases won't have alternatives to the null that induce probability distributions that take nose dives over small regions in the tails of the null hypothesis—typical cases won't look like figure 4. Rather, typical cases will be such that if the alternative hypothesis makes it much more likely that the value for the test statistic should be in the rejection region for the null, each of the sub-regions within the rejection region for the null will also be such that it's a good deal more likely that the value for the test statistic should be in that sub-region on the alternative hypothesis than on the null. In such cases, whether we reason with our actual evidence or weakened evidence of the sort used in significance tests won't substantially affect our conclusions.[29]

[29] What about atypical cases where some of our best hypotheses don't induce smooth probability distributions on the test statistic? Whether this is because the test statistic doesn't measure a natural quantity, or because some of our best hypotheses induce non-smooth distributions on a natural quantity, I'm inclined to think that such cases will be cooked-up philosophers examples. While possible, they'll be recognizably strange enough that realistically, wouldn't be inclined to use straightforward applications of significance tests to analyze their data.

Actually, that last step was a bit quick. It's not obvious that ruling out cases where some of the hypotheses under consideration display steep peaks and valleys is sufficient to rule out the possibility of hypotheses receiving spurious support from weakened evidence. While it is sufficient to rule out cases like the one represented in figure 4, mightn't there be other cases where weakening the evidence is a problem, but where non-smooth probability distributions aren't involved? At first blush, the answer seems to be yes. Consider the following figure:

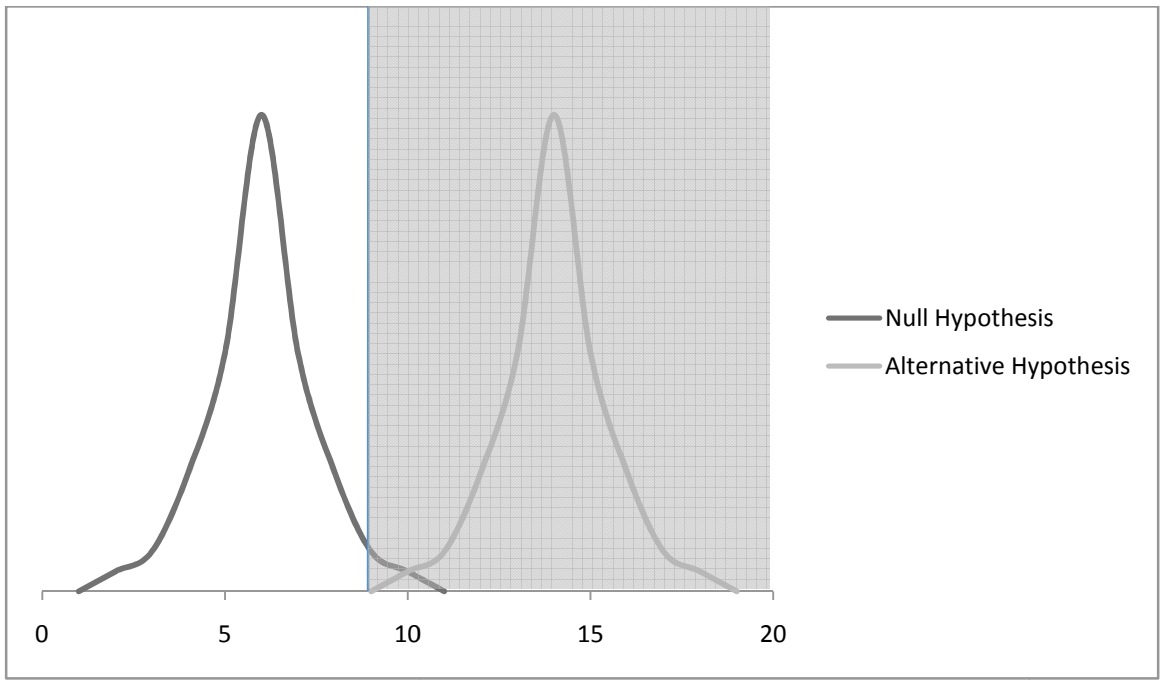

Figure 6

The shaded region represents the rejection region for the null hypothesis, and the lefthand border of the shaded region represents the measured value for the test statistic. On the assumption of the null hypothesis, it's highly unlikely that the test statistic should take a value in the shaded region, while on the assumption of the alternative hypothesis, this is highly likely. That is, the weakened evidence that the measured value for the test statistic was in the shaded

region strongly supports the alternative hypothesis. However, while the observed value is in the shaded region and is unlikely on the null hypothesis, is even less likely on the alternative hypothesis. The stronger evidence—that the test statistic took the value it actually did—supports the null hypothesis. In this respect, the case represented in figure 6 is like the case represented in figure 4—weakening the evidence leads to spurious rejection of the null. However, neither the null hypothesis nor the alternative hypothesis induces a non-smooth distribution on the test statistic.

This case is inspired by one discussed by Howson and Urbach (1993, p. 206) in a section of their book entitled "A Well-Supported Hypothesis Rejected in a Significance Test." Their discussion is of Neyman-Pearson significance testing, but the example is relevant to my discussion with only slight modifications. A researcher knows that a coin has either bias 0.4 in favor of heads, or 0.6 in favor of heads. He decides to flip the coin 1,000,000 times and observe the proportion of heads to find out which. He designates the hypothesis that the coin is biased 0.4 in favor of heads the null hypothesis—so he determines to reject the null if the proportion of heads observed is such that it is less than 5% likely (on the assumption that the bias of the coin is 0.4) that there should be at least that many heads in the sequence. He performs his experiment and observes that the proportion of heads in the sequence is 0.45. Intuitively, he has obtained very strong evidence that the bias of the coin is 0.4. After all, while it was extremely unlikely on the null hypothesis that there should be so many heads, it was far, *far* less likely on the alternative hypothesis that there should be so few heads. However, the researcher's significance test would have him reject the null hypothesis because it was far less than 5% likely on the null hypothesis that the obtained frequency of heads should be at least 0.45—under my interpretation of rejection, significance testing would have him become less confident that the coin's bias is

0.4, surely an absurd result. Furthermore, if I'd designated the hypothesis that the coin's bias was 0.6 the null, I'd have obtained the opposite result. Not only absurdity, but arbitrariness too.

Were we to draw out the probability distributions for the proportion of heads on the null hypothesis and on the alternative hypothesis, they'd look something like figure 6. What's distinctive about cases like figure 6 is that there's a sizable region between the peaks of the two probability distributions such that both hypotheses imply that it's unlikely that the test statistic should take a value in this region. This means that the rejection region for the null overlaps with the left tail of the probability distribution on the alternative hypothesis. This wasn't the case in figures 2 and 3, the ones we took to represent typical cases. In those cases, the probability density on the alternative hypothesis is already relatively high by the time we reach the rejection region for the null. There is no region between the peaks of the hypotheses such that both hypotheses imply that it is very unlikely that the test statistic should take a value in that region. I've already argued that typical cases won't look like figure 4. But why shouldn't they look like figure 6? At first blush, these seem like very different questions, and it seems implausible that a solution to one should imply a solution to another. Ultimately, however, I think that smoothness considerations along much the same lines as those discussed above can also explain why cases represented by figure 6 are atypical. Before arguing for this, I want to try to motivate this strategy by giving some examples of just what real world cases (as opposed to idealized ones like Howson and Urbach's coin case) would have to be like for them to be accurately represented by figure 6. I suspect that not only will it be plausible that such cases are atypical, but it will be plausible that the reasons that they're atypical are by and large the same as the reasons that cases accurately represented by figure 4 are atypical.

Let's consider the pollution, education, and medical cases. What would it take for cases like these to be accurately represented by figure 6? In the pollution case, it would have to be that there was some *n* with the following properties. On the hypothesis that children in Horseshoe Creek don't have their growth stunted by pollution in the creek, it is very unlikely that they should be on average *n* inches shorter than the national mean. However, it's also very unlikely on the hypothesis that their growth *is* stunted by the creek that they should be *n* inches shorter than the national mean. But there is an $m > n$ such that, on the hypothesis that their growth is stunted by the creek, it's quite likely that they should be *m* inches shorter than the national mean. It's hard to imagine what realistic body of evidence could make it reasonable to think that if the creek was stunting their growth, it might make them 4 inches shorter than the national mean, but it couldn't make them only two inches shorter.

Similar considerations apply in the education and medical cases. Realistic bodies of evidence won't make it reasonable to think that a new instructional method might improve scores on average by 20 points, but would almost certainly not improve scores on average by only 10 points. Lastly, typical cases will be such that if it's reasonable to think that a treatment for a disease might shorten recovery time by *n* days, there won't be an $m < n$ such that the treatment might not instead shorten recovery time by only $n - m$ days.

Why think that a smoothness-based strategy along the lines laid out above could help explain why these cases are atypical? The short answer is as follows—just as it's plausible that quantities like heights or lengths of recovery times are natural quantities (and so are typically smoothly distributed), it's also plausible that quantities like the degree to which pollution stunts children's growth, or the degree to which a treatment shortens recovery times, are also natural

quantities and should also be smoothly distributed. This claim is quite plausible when we think about variations on Howson and Urbach's coin case. They stipulate that the researcher only allows for the possibilities that the bias is 0.4 or 0.6. Given the right background knowledge, this could be reasonable. But imagine a case where one comes across an ancient Greek coin on a beach—I stipulate that it's an ancient Greek coin because I know nothing about whether or not ancient Greek methods were precise enough to produce symmetrical coins, and I suspect my reader doesn't either. If it's plausible that natural quantities are typically smoothly distributed, I submit that it's also plausible in this case that one's probability distribution over possible values for the bias of the coin should also be smooth. Why is this? Well, presumably the sorts of factors that contribute to the coin's having the bias that it does—e.g., the distribution of mass—are themselves expressible as natural quantities, and are such that our best hypotheses will induce smooth probability distributions on them. If the reader doesn't find this plausible, imagine that the coin is actually just a roughly disc-shaped object produced by some natural process—perhaps a very smooth, very thin rock—imagine that it looks roughly symmetrical, but one doesn't have any reason to think that it was produced by somebody trying to guarantee this. If the quantities that determine the coin/rock's bias are themselves smoothly distributed, so will the bias for the coin/rock itself (barring any unusual dependencies).[30]

I take it that similar considerations make it plausible that quantities like the degree to which an educational method improves test scores, or the degree to which a treatment shortens recovery times, should also be smoothly distributed. The factors that affect just how the

[30] Imagine a case where some quantity x is determined by two other quantities y and z, such that $x = |y - z|$. One might have a smooth distribution for y that was symmetrical around some value n, and likewise for z, but have an unsmooth distribution for x that assigned probability 1 to x's taking the value n, and assigned probability 0 to x's taking any other value. This would be possible if one were sure that for every m, whenever y took the value $n + m$, z took the value $n - m$. However, I take it that such dependencies would be implausible in the coin case, and in most natural cases.

educational method affects test scores—how much more information the students retain when taught using the method, how much more focused they are, etc.—are themselves plausibly expressible as natural quantities, and should themselves be smoothly distributed. Mutatis mutandis for the other cases.

Why does this mean that typical cases won't be like that represented in figure 6? Consider the following figure, which represents an unrealistic version of the medical treatment case in which it's quite likely that the treatment should shorten recovery times by 2 days on average, but not by 1.

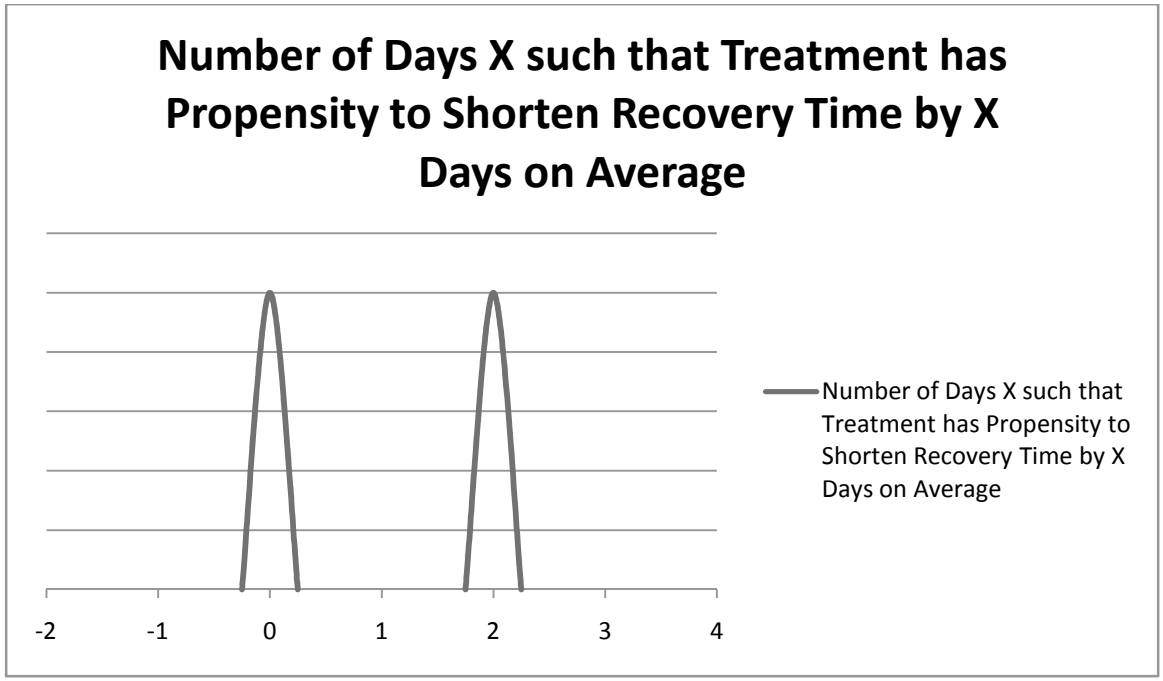

Figure 7

Because the null hypothesis is relatively plausible, there's a peak at 0—there's a decent chance that the treatment doesn't tend to shorten recovery time at all. Because the alternative hypothesis is also relatively plausible, there's a peak at 2—there's a decent chance that the treatment tends

to shorten recovery time by roughly two days. But because the case is unrealistic—like that in figure 6—there's no chance the treatment tends to shorten recovery time by one day. That's not to say there's no chance the difference in recovery times in some particular case will be one day—it's just to say that if it is, it's because such a case was an improbable result of a treatment that, on average, tends to shorten recovery times a by less (0 days) or more (2 days) than one day. In this case, an observed difference in recovery times of half a day could lead to rejecting the null in a significance test, even though it is really evidence for it. However, the probability distribution represented above is clearly unsmooth. A more realistic case would look something like the following:

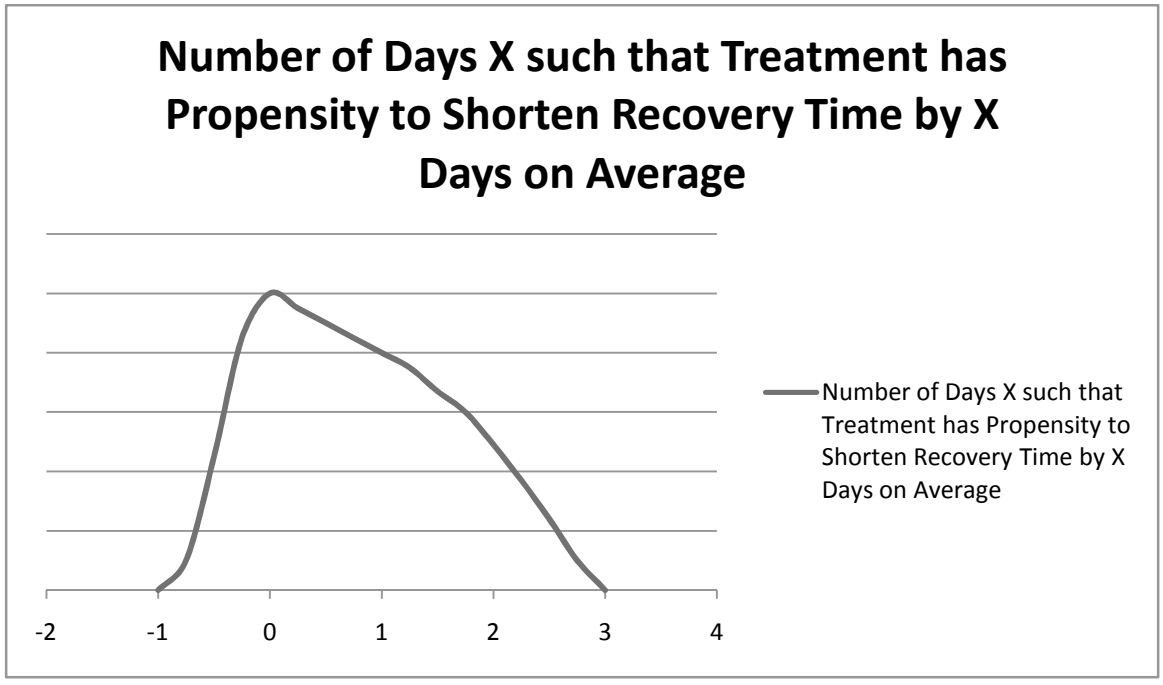

Figure 8

Clearly, in a case like this one, no observation to the effect that the difference in average recovery time between the two groups was *n* days is such that it is very improbable on the null

hypothesis ($X = 0$, in figure 8), but also very improbable on the alternative hypothesis ($X > 0$), and such that there is some $m > n$ such that it was quite probable on the alternative hypothesis that the difference should've been n days. That is, cases like the above one will be cases where weakening the evidence is unproblematic. And the above case is one in which a natural quantity (the number of days such that the treatment has a propensity to shorten recovery times by that many days) is smoothly distributed.

I don't mean to say that there couldn't be any realistic cases with the same structure as Howson and Urbach's coin case. We might imagine a variant on the Horsheshoe Creek case, where it's known that a boat carrying a tremendous quantity of pollutants passed through the creek, and what's unknown is whether or not it had a spill. We might imagine that boats like this one spill a *lot* if they spill at all, and the pollutants are known to have *very* substantial growth-stunting effects. Lastly, we might imagine that it's known that there are no other potential growth stunting factors in the town. In this case, finding that the children were significantly shorter than the national mean might actually be evidence *for* the null hypothesis that nothing is stunting their growth; this would hold if the difference were still too small to be consistent with the hypothesis that the lake was polluted. While the above case is not a science fiction scenario, it did require quite a lot of caveats, and I hope the considerations I suggested above make it plausible that cases like it shouldn't be typical. Moreover, while I don't have the space to defend this claim here, I suspect that in cases that do have the same structure as the Howson and Urbach coin case, savvy researchers could find classical methods other than significance testing to analyze their data; sticking to classical methods would not force them to run a significance test and risk spuriously rejecting the null.

Of course, any smoothness-based argument for the conclusion cases in which weakening the evidence leads to spurious rejections of the null are atypical must be merely qualitative and somewhat rough, since the notion of "smoothness" it appeals to is itself qualitative, and somewhat rough. But I don't believe that this robs my argument of its interest. The cases represented by figures 4 and 6 seem to be paradigm instances where weakening the evidence leads to straightforwardly spurious rejection of the null, and smoothness considerations really do tell against taking these cases to be typical. While I can't provide a general argument that every case in which the null would be spuriously rejected is a case like figure 4 or 6, I suspect that once one recognizes the importance of the smoothness considerations I've discussed, it becomes much less plausible that spurious rejections of null hypotheses based on weakened evidence are an important source of error in statistical practice.

6. Conclusions

I've considered two ways in which significance testing might seem prone to lead to flawed evaluations of the probative force of a body of evidence—ways in which significance testing might have us conclude that some hypothesis is probably not true, when a rational evaluation of the evidence would not. I've argued that neither apparent threat is real—we needn't worry that significance testers irrationally reject hypotheses.

First, I considered the worry that significance testing involves committing the fallacy of probabilistic modus tollens—inferring that a hypothesis is improbable because an event that was improbable according to the hypothesis occurred. I argued that there are conditions under which such events really do count as strong evidence against a hypothesis—when there are plausible

alternative hypothesis that assign the event a higher probability—and that these conditions are also exactly the conditions under which researchers are likely to perform significance tests.

Second, I considered the worry that significance testing involves committing the fallacy of weakening the evidence—concluding that a hypothesis is probably not true because some description of one's evidence that's logically weaker than one's total evidence would count strongly against it. I argued that there are conditions under which concluding that some hypothesis is probably false based on a weakened version of one's evidence is harmless—when each of the potential strengthenings compatible with the weakened version of one's evidence are also much less likely on the assumption of the rejected hypothesis than on some alternative. I argued that because test statistics in significance tests represent natural quantities, and our best hypotheses induce smooth distributions on natural quantities, whether we use the evidence that the value for the test statistic was at least as extreme as x, or the evidence that it was x, our conclusions are unlikely to be substantially affected. This is because paradigm cases in which updating with the evidence that the value for the test statistic was actually x (rather than just at least as extreme as x) changes whether or not we ought become less confident in the null are cases that involve unsmooth probability distributions over a natural quantity. Sometimes that natural quantity is the test statistic itself (as in cases represented by figure 4), and sometimes it is a quantity that represents the propensity of the system being studied to produce particular values for the test statistic (as in cases represented by figure 6). Either way, if typical cases where significance testing is used involve smooth distributions over natural quantities, the most obvious ways for significance testers to commit the fallacy of weakening the evidence are unlikely to arise in practice.

Hopefully, the considerations I've raised make it plausible that Bayesians need not regard significance testing as quite as hopelessly flawed as the Howson and Urbach quotes at the beginning of this paper suggest they should. While Bayesians who accept my arguments may still think that the frequentist approach to statistical hypothesis testing is bad epistemology, they needn't worry that the practical application of this bad epistemology in the form of significance testing is likely to lead us to reject hypotheses when a rational evaluation of the evidence would have us retain them.

This result is congenial to the Bayesian in a number of ways. First, it lets the Bayesian avoid the skeptical threat I mentioned at the beginning of this paper. Furthermore, if other non-frequentist approaches to statistical inference (e.g., Sober's Likelihoodism) are unable to appeal to the arguments I've offered here to explain why significance testing is generally reliable, this may give the Bayesian a leg up over other non-classical schools of thought about statistical inference—it may be that not all non-frequentists can avail themselves of my anti-skeptical escape route.

More generally, it counts in favor of the Bayesian approach to statistical inference if it can not only improve upon orthodox methods where they fail, but also explain why they work when they do. It's a commonplace in the philosophy of science that when a new theory can both accommodate the anomalies that afflict an old one, and explain why the old one works as well as it does in a limited domain (perhaps by showing that the old theory is a limiting case of the new one), that counts strongly in favor of the new theory.[31] That general relativity could both explain the precession of the perihelion of Mercury when Newtonian mechanics could not, *and* explain

[31] See Kuhn (1996)

why Newtonian mechanics worked so well in low mass, low speed environments, counted doubly in its favor. Analogously, it should count doubly in favor of Bayesianism that it can explain why significance testing gets certain cases wrong (e.g., by explaining why probabilistic modus tolens and weakening the evidence are fallacious methods of reasoning), *and* explain why it works as well as it does in the conditions when it's generally used (by appealing to the arguments I've offered in this paper). Ultimately, what at first looked like a potential skeptical problem for Bayesianism turns out to be an opportunity for it to shine.

Works Cited

Achinstein, Peter. (1994) "Explanation v. Prediction: Which Carries More Weight?" in Hull, Forbes, and Burian (eds), pp. 156-54

Bulmer, M.G. (1979) *Principles of Statistics*. New York, Dover

Collins, Robin. (1994) "Against the Epistemic Value of Prediction Over Accommodation," *Noûs* 28, pp. 210-24

Duncan, and Howitt, Laurence. (2004) *The SAGE Dictionary of Statistics: A Practical Resource for Students in the Social Sciences*. London, SAGE

Howson, Colin, and Urbach, Peter. (1993) *Scientific Reasoning: The Bayesian Approach*. Chicago. Open Court

Kuhn, Thomas. (1996) *The Structure of Scientific Revolutions*. Chicago. Chicago University Press

Lipton, Peter. (2004) *Inference to the Best Explanation*. London, Routledge

Lewis, David. (1984) "Putnam's Paradox." *Australasian Journal of Philosophy* 62, pp. 221-236

–(1987)"A Subjectivist's Guide to Objective Chance," in *Philosophical Papers, Volume II*. Oxford. Oxford University Press

Mayo, Deborah. (1996) *Error and the Growth of Experimental Knowledge*. Chicago and London. University of Chicago Press

Mankiewicz, Richard. (2000) *The Story of Mathematics*. Princeton, NJ. Princeton University Press

Mellor, D.H. (2005) *Probability: A Philosophical Introduction*. New York, NY. Routledge

Plato. (1925) *Plato in Twelve Volumes*, Vol. 9 translated by Harold N. Fowler. Cambridge, MA. Harvard University Press.

Quine, W.V. (1969) *Ontological Relativity and Other Essays*. New York: Columbia University Press.

Sider, Ted. (Forthcoming) "Ontological Realism," in David Chalmers, David Manley, and Ryan Wasserman, eds., *Metametaphysics*, Oxford University Press.

Sober, Elliott. (2008) *Evidence and Evolution*. Cambridge, UK. Cambridge University Press

Strevens, Micahel. (1998) "Inferring Probabilities from Symmetries," *Noûs* 32, pp. 231-46

White, Roger. (2000) "Fine Tuning and Multiple Universes," *Noûs* 34, pp. 260-76

–(2003) "The Epistemic Advantage of Prediction Over Accommodation," *Mind* 112, pp. 653-683

–(2007) “Does Origins of Life Research Rest on a Mistake?” *Noûs* 41, pp. 453-477