Who's Afraid of the Base-Rate Fallacy?

Corey Dethier

[pre-print; forthcoming at *Philosophy of Science*]

Abstract

This paper evaluates the back-and-forth between Mayo, Howson, and Achinstein over whether classical statistics commits the base-rate fallacy. I show that Mayo is correct to claim that Howson's arguments rely on a misunderstanding of classical theory. I then argue that Achinstein's refined version of the argument turns on largely undefended epistemic assumptions about "what we care about" when evaluating hypotheses. I end by suggesting that Mayo's positive arguments are no more decisive than her opponents': even if correct, they are unlikely to compel anyone not already sympathetic to the classical picture.

0 Introduction

The *base-rate fallacy* involves conflating the likelihood of the evidence on a hypothesis—$P(E|H)$—with the posterior probability of said hypothesis on said evidence, or $P(H|E)$. It is sometimes suggested that classical methods that rely on error probabilities—particularly significance testing and hypothesis testing—run afoul of the base-rate fallacy.[1] Howson (1997, S188-89) offers a striking claim to this effect, arguing that that the explication of classical reasoning offered by Mayo (1996) is "demonstrably unsound" (see also Howson 2000, 51–54).

[1] Philosophers who have endorsed some form of this argument include Achinstein (2001, 2010), Howson (1997, 2000), Korb (1991), Rosenkrantz (1977), Spielman (1973, 1974), Sprenger (2017), and Titelbaum (2022). So far as I can tell, the argument is absent from the statistical literature.

Over the years, Mayo (1997a,b, 2005, 2010, 2018) has offered a number of responses to Howson's argument (see also Spanos 2010). Bayesians seem to find these responses uncompelling, however. Achinstein (2001, 132–40, 2010, 182, 186–88), Sprenger (2017, 390), and Titelbaum (2022, 464) all employ a version of the argument to illustrate the (alleged) deficiencies of either Mayo's account or "frequentist" statistics more broadly. And while Achinstein and Titelbaum offer more subtle criticisms than Howson, Sprenger (2017, 390) concludes his discussion of the example with the claim "significance testing is logically invalid."[2]

Against the backdrop of the continuing popularity of the objection, I reevaluate three arguments that Mayo gives in response: first against the narrow claim of invalidity or unsoundness, second against the more subtle contention that error probabilities are not "what we care about" when testing a hypothesis, and third that it's the Bayesians who run into trouble in the relevant cases. The first response is simply correct: the claim of invalidity or unsoundness rests on a misunderstanding of both Mayo's work and classical statistical theory more broadly. The second response is not definitive, but gains substantial support from epistemology: at best, the critics haven't substantiated their claim that error probabilities are less epistemically relevant than posterior probabilities, let alone that they are irrelevant. The third response is less decisive: plausibly, Mayo has identified a set of cases in which classical methods should be preferred to those advanced by her critics. But these cases are unlikely to compel anyone not already sympathetic to the classical picture.

One note before I begin. In what follows, I eschew diving into the technicalities of classical theory whenever possible. This choice contrasts with Mayo and (even moreso) Spanos, who both get deep into the technical weeds. In staying away from the technicalities, I'm sacrificing some details relevant to the argument of §2, though I flag these as they arise. The benefit is a more accessible presentation and a demonstration of something that I think has been lost in previous discussions: *this particular debate* doesn't turn on the technicalities or subtleties of classical statistics. You don't need an education in statistics to know that classical theory doesn't commit the base-rate fallacy.

1 The narrow argument

Consider the following scenario:

[2]Sprenger bases this claim on more than one argument; as he sets it up, however, the base-rate fallacy plays the central crowning role and is the sole response to contrastive versions of significance testing.

DISEASE TESTING

We have invented a new binary test for the presence of a particular disease. The probability that a person with the disease receives a negative result on the test and the probability that a person without the disease receives a positive result are both .05. The incidence of the disease in the population is one in a thousand, or .001. When we select a person from the population at random and test them, they test positive.

What is the probability that we have randomly selected a person with the disease? The correct answer to this question can be calculated easily. Let H be the hypothesis that we've randomly selected a person with the disease and $+$ be the positive result. Then by a straightforward application of Bayes' rule:

$$\begin{aligned} P(H|+) &= \frac{P(H)P(+|H)}{P(H)P(+|H) + P(\neg H)P(+|\neg H)} \\ &= \frac{.001 * .95}{.001 * .95 + .999 * .05} \\ &\approx .019 \end{aligned}$$

People often give a different answer when asked similar questions (see, e.g, Tversky and Kahneman 1982).[3] In particular, they often respond to these kinds of questions by saying that the probability is .95—that is, they report the likelihood of $+$ given H or (in other words) the probability of a positive result given the assumption that the randomly selected person has the disease. This answer illustrates the *base-rate fallacy*: it confuses the likelihood $P(+|H)$ with the posterior probability $P(H|+)$.

Howson argues that classical statistics is committed to precisely this fallacy (see Howson 1997, S188-89, 2000, 52–54). To make this case, he presents examples like DISEASE TESTING as cases of hypothesis/significance testing.[4] The classical recipe for hypothesis testing involves administering a test that is designed so that some outcomes are very improbable if the hypothesis is true. Should we then observe those outcomes, the classical statistician takes this to be powerful evidence against the hypothesis; a good reason to "reject" it. How exactly we should interpret "rejection" here depends on whether

[3] Many of these survey questions are not tightly analogous, however. As Levi (1981, 1983) stresses, for instance, survey questions usually involve subjects who aren't randomly sampled from the relevant populations. I'll come back to this point in §5.

[4] In what follows, I'll play a little fast and loose with the differences between Fisherian significance testing, NP hypothesis testing, and Mayo's error-theoretic reconstruction of classical statistics. Versions of the argument presented below have been directed against all three, and if the argument is successful, it applies to all three.

we're working within the broadly falsificationist approach of Fisher, the more behavioristic approach of Neyman and Pearson, or the evidential approach outlined by Mayo. It doesn't matter for the present discussion which of these interpretations we prefer, however, so we can be sloppy in this respect and say that "rejecting" a hypothesis just involves believing that it's false—this (ahistorical) interpretation of "rejection" is the most amenable to Howson's criticisms.

If we treat Disease Testing as a case of hypothesis testing, a positive result is very improbable if the hypothesis that the randomly selected person doesn't have the disease is true (that is, $P(+|\neg H)$ is very low). Since we observed this outcome, we have good reason to reject $\neg H$—that is, to believe that H is true. This is the result that we derive even though, as Howson stresses, it's very likely that the H is in fact false. In effect, then, we've committed the base-rate fallacy: we've believed something that is very unlikely to be true because the classical recipe for hypothesis testing tells us to reject the hypothesis when the likelihood of the evidence is sufficiently low. So, since we've just followed the rules of classical statistics here, classical statistics is in some important sense intrinsically committed to the base-rate fallacy. Howson concludes that the rules of classical statistics are "demonstrably unsound" (Howson 1997, S188), and remarks that "It can only be in a Pickwickian sense of 'good', therefore, that a 2 per cent chance constitutes 'good grounds for H'" (Howson 2000, 54).

Call this formalization of Howson's reasoning the "narrow argument":

(N1) Classical statistics recommends using hypothesis testing to determine whether the person randomly selected in Disease Testing has the disease.
(N2) If (N1), then classical statistics recommends believing that the randomly selected person has the disease.
(N3) The probability that the randomly selected person has the disease is approximately .02.
(N4) Any theory that recommends believing P when P has a sufficiently low probability is unsound.
$\therefore$ (NC) Classical statistics is unsound.

Howson's claim that classical statistics commits the base-rate fallacy is essentially a *diagnosis* of why it's unsound: classical statistics takes the low likelihood of the observed results on the hypothesis that the randomly selected person lacks the disease as grounds for belief that they have the disease, whereas the (only?) proper grounds for belief is a high posterior probability. As Howson puts it:

> The counterexample shows clearly that despite the test's being as severe as you like, it is a mistake to suppose that the very (small) chance of a test's passing a hypothesis h when h is false is by itself any indicator of the correctness of h if h passes the test. ... Indeed, if you infer from the test's positive diagnosis to the presence of the disease you will be wrong nearly all the time. (Howson 1997, S189)

Presumably, a probability of approximately .02 is "sufficiently low"; that granted, the argument is valid. The preceding paragraphs demonstrate that (N2) and (N3) are both true. As we'll see in §4, the idea behind (N4) is contentious, but we'll assume that it's true for the sake of argument for now. And the diagnosis is tendentious and dismissive—Howson (2000, 54) explicitly urges us to think that the "unsoundness" of Mayo's position is owed to an obvious fallacy rather than a philosophically-motivated disagreement about the proper approach to statistics—but (if the argument is correct) it does seem to get at the heart of the problem.

That leaves (N1).

2 The narrow argument rebuffed

The first of Mayo's three responses that we'll consider is that (N1) is false. In Mayo (1997a), for instance, she argues that classical hypothesis testing is only appropriate when there aren't (classically acceptable) priors. Where such priors exist, the classical statistician should use them:

> [R]ecall that the error statistical account is based upon frequentist methods such as NP tests, and these methods developed precisely for situations in which no frequentist prior is available or even meaningful, as with the majority of scientific hypotheses of interest.
> ...
> But if H is a random variable, and a frequentist prior is available, the error statistician can use it too. (Mayo 1997a, S205-6)

In another article appearing in the same year, she points out that if we take the perspective of classical statistics, the hypothesis in question should not be considered a proper statistical hypothesis at all:

> However, in all such examples, the hypotheses are forced to be statements about the particular *sample* and are not *statistical hypotheses*. (Mayo 1997b, 326)

According to Mayo, in other words, classical statistics views hypothesis testing and calculating the probability that a random variable takes on some value as distinct problems. (N1) is false, therefore, because Disease Testing involves calculating the probability that a random variable takes on some value and is thus not a problem where classical statistics recommends hypothesis testing.[5]

Is Mayo's characterization accurate? To answer this question, we need look no further than Fisher's *Statistical Methods and Scientific Inference*, which is unequivocal on the matter:

> the different situations in which uncertain inferences may be attempted admit of logical distinctions which should guide our procedure. That it may be the data are such as to allow us to apply Bayes' theorem leading to statements of probability; or secondly, that we may be able validly to apply a test of significance to discredit our hypothesis the expectations from which are widely at variance with ascertained fact. (Fisher 1973, 37)

One can find similar comments elsewhere in Fisher's work (e.g. Fisher 1958, 9–10), as well in that of the other "fathers" of classical statistics, Neyman (1971, 3) and Pearson (1962, 395–96). Indeed, these quotes are unsurprising. Throughout the first half of the 20th century, Fisher's "logical distinction" was typically understood to be the distinction between "direct" inferences from populations to (appropriately selected) samples—which admit of straightforward mathematical reasoning—and "inverse" inferences that go the other direction.[6] Only the latter were understood to be the proper domain of hypothesis testing. This framework is largely presupposed in the canonical work of both Fisher (1922, 313–14, 1958, 7) and Neyman and Pearson (1928, 175, 1933, 291).

So neither the canonical statements of the fathers of classical statistics nor Mayo's philosophically sophisticated reconstruction recommend applying hypothesis testing in Disease Testing. Modern textbooks are no more supportive of (N1). Lehmann and Romano (2022), widely regarded as *the* textbook on hypothesis testing, open with a specification of the problem of statistical inference that excludes Disease Testing:

[5]Spanos (2010) rightly points out that there are quite a few other problems with the use of cases like Disease Testing in arguments against hypothesis testing, *particularly* in the context of Mayo's elaboration of it. This one will do for illustration, however.

[6]While this way of framing the debate has largely dropped out of the literature—not without reason—we can find it centered as late as 1979, when Seidenfeld (1979, 2) opened his book on statistical inference with the distinction between direct and inverse inference, describing the former as "uncontroversial" and finding divergence only in the latter domain.

> The raw material of a statistical investigation is a set of observations; these are the values taken on by random variables X whose distribution P_θ is at least partly unknown. Of the parameter θ, which labels the distribution, it is assumed known only that it lies in a certain set Ω, the *parameter space*. *Statistical inference* is concerned with methods of using this observational material to obtain information concerning the distribution of X or the parameter θ with which it is labeled. (Lehmann and Romano 2022, 3)

Other textbooks are similar. Wasserman (2005, ix), for example, contrasts "probability," which involves calculating the probability of outcomes of a given data-generating process, with "statistical inference" where the problem is the "inverse" (again) of that for which the calculations of probability are appropriate.[7] Neither formulation licenses treating the question of whether a random variable takes on a particular value as a hypothesis in the sense of statistical inference generally speaking, let alone hypothesis testing.

Howson is aware that what we've termed (N1) is open to objection. Here is his discussion, quoted in full:

> It might be objected that the hypothesis in the example is a random variable, whereas a hypothesis of the sort philosophers of science usually discuss is not (Mayo several times claims that hypotheses are not random variables). The objection is both wrong and beside the point. It is wrong because there are models of Kolmogorov's axioms in which hypotheses are random variables (measurable functions): any hypothesis is a two-valued random variable in the appropriate space. The objection is beside the point since the error-probability conditions for a severe test of that particular hypothesis H are clearly satisfied; equally clearly, passing the test provides no indication of H's correctness. Indeed, the counterexample is so telling precisely because H is a random variable,

[7] Here are a few more examples. Agresi, Franklin, and Klingenberg (2017, 10, 387) tell us that statistical inference is concerned with inferences from samples to populations and that the hypotheses tested in significance tests must concern the population. After introducing an example in which a sample is drawn from a box, Freedman, Pisani, and Purves (2007, 478) state that "a test of significance only makes sense in a debate about the box." Casella and Berger (1990, 345) define a statistical hypothesis as "a statement about a population parameter," emphasizing that the "important point is that a hypothesis makes a statement about the population." Even explicitly Bayesian treatments follow this pattern: Robert (2001, 224) introduces hypothesis testing as matter of determining whether the true value of a population parameter falls within the specified region.

> possessing an empirically-based prior distribution. (Howson 1997, S190)

There are two rejoinders here. The second—that the objection is "beside the point"—finds no real purchase against Mayo's response: the problem is that the account that he is criticizing says that it would be inappropriate to apply hypothesis testing to the alleged example; whether the conditions are met for the inappropriate test to deliver a particular verdict is immaterial. In determining whether a rule or procedure is "sound" or "valid," what that rule would say outside its explicitly specified domain of proper application is irrelevant.[8] The narrow argument is akin to claiming that first-order logic is unsound because of what it says (or doesn't say) when applied to strings of symbols that aren't well-formed formulas.

The first rejoinder—which is essentially that the objection is confused—is more interesting. In the present context, Howson can be read as pointing out that the distinction between inferences from populations to samples and inferences from samples to populations is not recognized by probability theory. That's correct, of course, and at face value it puts pressure on the distinction that the classical statistician wants to draw. Certainly, it undermines Fisher's characterization of the distinction as a "logical" one. It does not undermine the distinction itself, however, at least not absent further argument. The classical statistician has at least two routes for response.

The first of these is to appeal to different interpretations of probability. According to a "frequentist" interpretation, as Mayo (1997a, S205) highlights in her response, we cannot coherently assign probabilities to certain kinds of hypotheses because those hypotheses are either true or false; there's no sampling procedure or long-run frequency to make the probability claim "meaningful" (see also Spanos 2010, 577). As the hypothesis in Disease Testing is the result of a random sampling process, it can be properly characterized as a random variable according to frequentist strictures. So we can and should calculate its probability distribution in a straightforward mathematical way rather than employing hypothesis testing. In other words, a frequentist interpretation of probability blocks Howson's argument. Whatever the defects of frequentism, it does license a distinction between Disease Testing and those cases where hypothesis testing is applied.

While talk of different interpretations of probability is familiar to philosophers, it is often misleading when mapped onto the concerns of statisticians. The alternative is to ground the distinction in a view about what's required

[8] Though see the next section, where I offer a more "charitable" reading of Howson that has the interpretive disadvantage of not validating his conclusion.

for the justified use of probability distributions *in science*. Roughly: there are some cases where we have sufficient information to specify a intersubjectively acceptable prior probability distribution. Disease Testing is a paradigm example: we know the prevalence of the disease in the population and can treat it as common knowledge. At the same time, there are other cases—thought by the classical statistician to be more common in interesting scientific contexts—where the prior probability distribution is so underdetermined that it's better to proceed without one. Pearson seems to endorse a version of this view in recollecting about his early work with Neyman:

> We were certainly aware that inferences must make use of prior information and that decisions must take account of utilities, but ... we came to the conclusion, rightly or wrongly, that it was so rarely possible to give sure numerical values to these entities, that our line of approach must proceed otherwise. Thus we came down on the side of using only those probability measures that could be related to relative frequency. (Pearson 1962, 395–96)

Pearson's suggestion, essentially, is not that probabilities are relative frequencies, but that relative frequencies are a way of grounding or justifying the choice of probability distribution.[9] Like the first response, this one blocks Howson's rejoinder: the classical statistician may be wrong about what's required for justification, but they aren't committed to (N1).

The narrow argument rests on the assumption that classical statistics recommends applying the methods of hypothesis testing in cases like Disease Testing, an assumption I've labeled (N1). Mayo's response is that (N1) is false. On this point, she's clearly correct: given that both the founders of classical statistics and Mayo herself explicitly recommend against doing so, both the founders and modern textbooks adopt frameworks that preclude doing so, and all three do so on grounds that are broadly consistent with their other views on the interpretation and justification of probability functions, there is simply no question of maintaining this assumption. The idea that classical statistics is committed to the base-rate fallacy is simply a misunderstanding of classical theory.

[9] On this point, it's worth comparing the longer discussion in Neyman (1952, 22–27), where the question is not "what is the proper account of probabilities?"—indeed, Neyman (1952, 1) earlier suggests that this question is ill-posed—but "under what conditions can we (justifiably) apply the mathematical theory of probability to the real world?"

3 The refined argument

With the exception of Sprenger (2017), more recent discussions of the relationship between classical statistics and the base-rate fallacy have shied away from insisting that the theory is actually committed to a fallacious inference. Titelbaum, for instance, directs his discussion of what is essentially Howson's example against the idea that it's appropriate to (automatically or mechanically) reject the null hypothesis when the alternative passes a test (Titelbaum 2022, 464)—a conclusion that everyone involved accepts.

For our purposes, the more interesting argument is raised by Achinstein:

> Suppose that Mayo refuses to assign a posterior probability to h. Suppose she claims that we do not want, need, or have any such probability, whether or not this is epistemic probability. Then I, and I believe Mill, would have a problem understanding what passing a severe test has to do with something we regard as crucial in induction – namely providing a good reason to believe h. ... passing the test is not a good reason, or a good enough reason, to believe the hypothesis. (Achinstein 2010, 183)

The idea behind Achinstein's arguments is apparently simple: classical statistics may not be committed to the base rate fallacy, but it doesn't give us a high posterior probability. And a high posterior probability is what we really care about.

Note that Howson can also be read in this way, though doing so requires treating his claim that classical theory is "demonstrably unsound" as mere rhetoric. For instance, he complains that "Error statisticians ... have for decades given us something quite different from what we want, which is a way of computing the degree of confidence we should invest in hypotheses given empirical evidence" (Howson 1997, S189). And his response to the objection that his case involves a random variable could, with great charity, be understood as granting the narrow point but holding that the fact that the kind of properties that the classical statistician cares about are so misleading in that scenario indicates that something is deeply wrong with the approach.

Call this more subtle objection the "refined argument":

(R1) (Even if hypothesis testing cannot strictly be applied to Disease Testing, the example still illustrates that:) Error probabilities and posterior probabilities come apart.

(R2) The methods of classical statistics deliver error probabilities rather than posterior probabilities.

(R3) Posterior probabilities are "what we care about."
∴ (RC) The methods of classical statistics don't deliver "what we care about."

(R1) and (R2) are simply facts about the relevant quantities and methods. The only point of attack, therefore, would seem to be (R3). I suspect that most philosophers sufficiently enmeshed in Bayesian confirmation theory will find (R3) plausible, perhaps even obviously true—once it's demonstrated that error probabilities and posterior probabilities come apart, the literature on the subject has tended to assume that no further argument is needed to show the classical statistics doesn't give us what we care about. And while appeals to the base-rate fallacy are both recent and limited to philosophy, the idea that classical statistics doesn't give us what we care about is an older and broader one: Kyburg (1961, 27) calls essentially the same complaint "the most common objection to frequency theory."

Still, the fact that (R3) is intuitive is no guarantee that it is true. As we'll see, Mayo argues that it isn't.

4 The refined argument rebuffed

The second of Mayo's three responses is what initially appears to be an entirely unpromising decision to bite the bullet. She repeatedly asserts that what we really want (or at least what scientists really want) is not information about how likely it is that the hypothesis is true, but information about how well-"probed" it is; we want to know if it has been subjected to tests that are discriminating and that can be expected to reliably distinguish true hypotheses from false ones (Mayo 1997b, 328, 2005, 109, 121, 2010, 197, 2018, 226–27). Why is that? Mayo clearly sees her response as grounded in the entirety of the picture that she presents: the point of the various arguments that she offers throughout her work is to convince us that probity is something that we should care about. In some sense, then, Mayo and the Bayesian are simply at loggerheads. Evaluating the refined argument is then impossible without evaluating the position as a whole.

I do not think things should be left here, however. Though Mayo does not make this connection in her published work, the idea that we care about probity has substantial support within contemporary epistemology. In other words, widely-held views in epistemology provide us with strong grounds for rejecting (R3) and thus the refined argument. Here, briefly, is the case. First, it's commonly thought that "what we care about" is knowledge, which comes apart from posterior probability. Second, it's also commonly thought that knowledge

requires truth-tracking and, as we'll see, error probabilities are best understood as measures of the degree to which our method tracks the truth. Together, these two common views imply the error probabilities are something we care about in at least the same way that posterior probabilities are: they measure something that is necessary but not sufficient for our central epistemic aim.[10]

Lottery paradoxes of the sort originally put forward by Kyburg (1961) can be used to illustrate both of my main points in this section. So consider a lottery in which one winner will be chosen at random from a set of 10,000 tickets. The probability of any individual ticket winning is then 1 in 10,000. Kyburg points out that it is easy to generate a contradiction from these facts using only extremely reasonable inferences. The probability that any individual ticket wins is so low that you can reasonably conclude of any individual ticket that it will not win. If you can conclude of any individual ticket that it will not win, however, then it seems reasonable to conclude that no ticket will win. But this contradicts our assumption that one of the tickets would win.

One common response to the lottery paradox is to hold that no matter how high the probability that an individual ticket loses, there's something defective about believing that the ticket will lose. The problem is that even if the belief that the ticket loses is true, it isn't *tracking* the truth in a way that's important to knowledge. If the ticket was the winner, you would still think it wasn't (Dretske 1971). If you were to believe that it was, it still wouldn't be (Pritchard 2005). And there's some important sense in which no explanation would be needed for you to have an incorrect belief in this case—it would be very *normal* for you to be wrong (Smith 2016). Regardless of how we choose to cash out truth-tracking—the three options above are sensitivity, safety, and normalcy, respectively—it's widely felt that beliefs that track the truth are more valuable than those that don't, even in situations where both beliefs are true or even both true and justified. Which is essentially to say that "what we care about" is not just truth / accuracy but (modally stable) knowledge. From this perspective, the lottery paradox is a particularly nice illustration of the fact that high posterior probabilities can come apart from "what we care about."

So the lottery paradox illustrates the first of the two main points of this

[10]To be clear about how I see the dialectic, while the arguments below will indicate why the relevant epistemological positions are plausible, I won't offer anything like a full defense of them here—interested parties should see Pritchard, Turri, and Carter (2022) and Ichikawa and Steup (2017) respectively. Instead, my aim is to show that these common views have important consequences for the present debate; that established, I suggest that—in virtue of (R3)'s conflict with common epistemic views—the premise is at best in need of a defense that it has not received in the literature.

section: many contemporary epistemologists hold that what we care about is knowledge, and knowledge comes apart from high posterior probability, which is (at best) a necessary but not a sufficient condition for it. To illustrate the second point—that the error probabilities of classical statistics should be thought of as measuring the degree of truth-tracking—it will be helpful to have a more finely specified case. So consider what we'll call "Lockean lottery" after Foley (1992):

> Lockean Lottery
> John doesn't normally play lotteries, but on a whim he decided to buy a single ticket for a lottery that he knows is fair. Feeling silly, he considers whether or not to simply throw the ticket away, reasoning that it is very likely that he lost. He decides that his decision should be based on the size of the lottery; the larger the pool of tickets, the more likely it is that his is a loser. After some consideration, he decides that it would be unreasonable to believe that his ticket has any worthwhile chance of winning if the number of tickets sold exceeds 10,000. He thus settles on the following decision procedure: he will throw the ticket away if more than 10,000 tickets have been sold; otherwise, he will keep the ticket.

(Those who feel the potential costs of John's decision affect our judgments in this case can imagine that he has voluntary control over his belief and he is simply deciding what to believe; nothing will turn on the actual costs and benefits of throwing away the ticket.)

Notice: John's ticket being the winning ticket doesn't affect how many tickets have been sold. As a consequence, the error probabilities for John's decision procedure are just a function of the cutoff he chooses. More precisely, the probability of judging his ticket to be a loser when it isn't is just the probability that more than 10,000 tickets have been sold; the probability of failing to judge his ticket a loser when it is is the probability that 10,000 or fewer tickets have been sold. Or, more formally, the probability of the first kind of error is $P(n > 10,000)$; the probability of the second kind is $1 - P(n > 10,000)$.

That's *bad*. Normally, we aim for statistical tests in which both probabilities are very near to 0. Infamously, a .05 probability of a false negative is a standard minimum requirement for publication in most disciplines; there's no similar standard for false positives, but .2 is a commonly-cited cutoff. John's method can't achieve anything close to that. In fact, if $P(n > 10,000)$ is anything other than .5—in which case his method is no different than flipping a coin—he's guaranteed that the method will be *anti*-truth-tracking in one respect or the

other. No honest statistician would ever recommend a testing method with this error profile.

As Bayesians have rightly stressed for decades, the error probabilities are not telling us that John's method is *unreliable*; after all, if the number of tickets sold is greater than 10,000, then John's method is extraordinarily reliable. It has a success rate of $(n-1)/n$! Instead, what they're telling us is that John's method is not tracking the truth of the hypothesis. More precisely, the two error probabilities, $P(n > 10,000)$ and $1 - P(n > 10,000)$, are quantitative measures of how sensitive John's method is to the truth and falsity of the hypothesis respectively.[11] If John's method were perfectly sensitive, the probability that John would reject the hypothesis if it were true would be 0, as would the probability that he would accept that the hypothesis if it were false.[12] The poor quality of the actual error probabilities tells us that John's method is not tracking the truth at all, which means that it cannot support knowledge; what the statistical analysis reveals that John's beliefs don't track the truth in the way knowledge is commonly thought to require.

Contrary to what the critics suggest, therefore, classical statistics has a good claim to giving us something that we care about: error probabilities measure one of the necessary components of knowledge. At best, then, (R3) in the refined argument is an unargued-for assumption that is inconsistent with common views in epistemology. As such, I think it's fair to conclude that the refined argument is unsuccessful absent substantial further justification of (R3).

In an important sense, however, this conclusion understates the problems with the refined argument. Recall that the argument relies on a kind of sleight-of-hand: the audience is shown a misleading-at-best example in which the classical statistician allegedly commits to the base-rate fallacy, and it is then argued that while they aren't actually committed to the fallacy, the example still illustrates that the methods of classical statistics are deficient in some deeper sense. The entire intuitive pull of the argument, however, turns on the base-rate case: what makes it compelling is that Disease Testing is the rare situation where we *know* that posterior probabilities track the truth and thus that results that disagree are unreliable—Howson (1997, S190) himself says as much! In other words, the features that make the refined argument compelling are *exactly* the features that make it misleading. To me, it thus seems fair to

[11] For more on the relationship between error probabilities—especially as analyzed by Mayo—and modal conditions, see Fletcher and Mayo-Wilson (2024), Gardiner and Zaharatos (2022), and Mayo-Wilson (2018).

[12] Of course, there's a sense in which John could also be perfectly "sensitive" to the truth by being wrong in all cases, but that's not really the sense that concerns us here.

conclude that even if the assumptions identified above can be defended, the argument itself only has the advantages of theft over honest toil to recommend it.

5 Should the Bayesian fear base rates?

The previous two sections have demonstrated that neither the narrow argument offered by Howson and Sprenger nor the refined argument offered by Achinstein is successful. The former is just straightforwardly unsound; the best thing that can be said about the latter is that it relies on assumptions that are both controversial and for which no argument is provided. Nevertheless, there is one more argument of Mayo's that we have yet to examine—an argument to the effect that it is the Bayesian who should be afraid of base rates.

To makes this case, Mayo appeals to scenarios that look superficially like Disease Testing (Mayo 1997a, S205-6, 1997b, 327–29, 2005, 115–18, 2010, 195–99, 2018, 368–69). Here is a version of her most common example:

> Readiness Testing
> The Test of Aptitude, Scholastic (TAS) is a widely used test for evaluating college readiness, and its error rates are extremely well understood. The probability that a person who is ready for college receives a negative result on the TAS and the probability that a person who isn't ready for college receives a positive result are both .05. A confluence of confounding factors—high rates of crime, poverty, drug use, and homelessness; underfunded schools; and a dearth of previous successful college students—indicates that only one out of every 1000 residents of the town *Fewready* is ready for college. Isaac, a rising senior living in Fewready, takes the TAS and tests positive.

Mayo takes Readiness Testing to be amenable to the methods of hypothesis testing in a way that Disease Testing isn't (c.f. Spanos 2010, 577–79). The crucial difference? Here our concern is with a particular individual, not a randomly-selected sample, and the information about the population—the prior, essentially—is no longer acceptable by frequentist strictures.[13]

Before turning to the importance of this point, notice that a strict Bayesian can happily agree with Mayo that the previous research doesn't dictate the

[13]Notice that Mayo is here departing from the framework found in the modern textbooks quoted in §2. Since her view is explicitly revisionary, that's not a problem, but it is important.

choice of prior. Indeed, Levi (1981, 1983) makes essentially the same point in his discussion of empirical research on the base-rate fallacy: for a committed subjectivist, there's no necessary connection between the frequency of college readiness in the population and the prior probability that Isaac is college ready; so long as the relevant probability assignments obey the Kolmogorov axioms, the subjectivist is free to assign any prior probability. About this much, at least, the strict frequentist and the strict subjectivist agree.

So the Bayesian should be willing to grant Mayo that there is no necessary connection here; it's consistent for her to employ the population-level information in the first case but not in this one. Still, most Bayesians are liable to say that the population level information is *relevant*; after all, the whole point of including prior distributions in the analysis—or at least of including priors other than the objectivist's minimally informative ones—is to account for information like the general rate of college readiness among Isaac's peers. Here, therefore, we have an example that parallels Disease Testing but where the two accounts actually disagree. Why, then, does Mayo think this example supports the classical viewpoint when she was so careful to reject the earlier application?

To draw out her reasoning, notice an oddity about the Bayesian position in this scenario: for the Bayesian, there is no reason for Isaac to take the TAS. Or, more accurately, there is only reason to carry out the test given a very specific utility distribution: Isaac's text scores matter only if our decision problem is set up so that the very small change in the probability of Isaac's readiness alters which of our options has the highest expected utility. There's something worrisome about this aspect of the Bayesian position: it seems bad that the prior swamps *any* evidence that we could collect using available methods; it seems worse that we have to jury-rig the decision problem to avoid the situation where the Bayesian recommends rejecting Isaac from college purely on the basis of their priors.

It's easy to give this intuition some teeth. In the U.S., a long history of discriminatory practices that encourage (and enforce) segregation has rendered the location of a person's residence a good proxy for their race (Rothstein 2017; Taylor 2019). Widespread reliance on these proxies continues to contribute to racial disparities in a variety of sectors, including incarceration, housing, health, and education. Against this backdrop, it should be disturbing that even apparently innocuous demographic information about Isaac's peers can entirely swamp the information that we have about Isaac himself—indeed, can swamp it to such a great extent that there's no point in Isaac even trying.

It's worth being precise about the worry here. It is not that priors are

"subjective" in some vague sense and therefore guaranteed to be biased.[14] The concern is a narrower one: Mayo contends that when we take the critics' own examples and strip them of the idealizations that ensure the accuracy and/or relevance of the prior, we're left with a scenario in which the Bayesian's recommendation—or at least the critics'—is extremely troubling. On the one hand, we have a prior that is subject to few constraints, influenced by information we have good reason to think is biased, and whose relevance to the present testing situation we have no way of evaluating.[15] On the other hand, we have data from an extremely probabitive test that we know is relevant to the present testing situation. But if we adopt the method that the critics apparently believe we should, we allow the former to dominate both our epistemic conclusions (the posterior probabilities) and our decision making.

Bayesians have two traditional responses to worries about the subjective elements of their methodology. Neither is particularly compelling in this case. The first response involves pointing to results (e.g., Hawthorne 1993) that show that agents who properly conditionalize will eventually converge on the same posteriors under relatively weak assumptions about their priors. In this scenario, however, it's hard to see how eventual convergence helps Isaac—here, at least, the Bayesian is susceptible to their own complaint that what we care about is whether the hypothesis is correct, not whether our method has nice features in the long run. The second response is to (correctly) point out that classical methods also rely on the judgment of the individual statisticians. But Mayo can—and indeed does (see Mayo 2018, Excursion 4)—grant the general point while arguing that it isn't relevant: at least in this specific scenario, we have good reason to think that the priors are susceptible to bias and substantial empirical evidence that constrains the classical approach. Even if the different elements are all generally susceptible to bias, the prior is clearly more susceptible in this specific case.[16]

[14]Both critics and defenders of Bayesianism are guilty of running together a wide variety of different objections under the heading of "subjectivity." Mayo (1996, 2018) brings clarity to at least one line of criticism; Sprenger (2018) brings some to the defense.

[15]Recent (more-or-less qualified) endorsements of Bayesian methods by theoretically-inclined statisticians—e.g., Cox (2006), Kass (2011), and Gelman and Shalizi (2013)—have tended to emphasize that priors can be a tool like any other in the statistician's toolbox. In this respect, at least, their remarks are less in the mold of Howson and Urbach (2006), and more in-line with how Rosenkrantz characterizes the main commitment of his "objectivist Bayesian" position: "The assumptions which underlie a prior distribution are every bit as corrigible as those which underlie a data distribution or probability model of a natural phenomenon. And, by the same token, they are as empirically confirmable in the one case as in the other" (Rosenkrantz 1977, 189).

[16]A third response—retreating from the potentially biasing demographic information to

What does this argument show? By itself, I think, not a lot. Consider two distinct audiences for this kind of argument. The first views statistics as akin to engineering: the goal is to develop the right test for a messy and complex world where things often don't behave as they're supposed to.[17] If they are not already users of classical statistics, they are at least sympathetic to the idea that there may be cases where Bayesian tools are ineffective or unreliable and, crucially, take that as a reason not to use Bayesian tools in those specific cases. To this first audience, Mayo has offered a compelling scenario in which potentially biased prior information swamps a probative test and thus where it's arguable that the classical approach is simply preferable. But the interesting question for this audience is not whether these cases exist but their commonality: is Readiness Testing more representative of the use of statistics in science than the cases that Bayesians like Spielman (1974, 219) point to when motivating their position? (How are we even to evaluate this question, given that the concern is at least partly about the appropriate questions to ask in different scientific contexts?)

The second distinct audience is more committed to a Bayesian philosophy. This audience, I suspect, is unlikely to be swayed towards a pluralistic—let alone a classical—approach by pointing to specific cases where a reliance on Bayesian tools leads to dubious judgments. After all, one of the common commitments of the Bayesian philosophy is to seeing the problems of statistics as problems of logic; the goal is not a toolbox of different rules, methods, and heuristics specifically suited to their individual situations, but rather something that looks like quantified first-order logic. Howson and Urbach express this view when they remark that the "truth, rationality, objectivity, cogency or whatever of the premises ... are exogenous considerations, just as they are in deductive logic" (Howson and Urbach 2006, 301). They would likely say the same thing about Readiness Testing; on their view, it's simply not the job of a theory of statistics to say anything about where you got the information on which your prior distribution is based. Bad inputs lead to bad outputs—just as they do in logic.

Like the refined argument, then, Mayo's response is indecisive without further supplementation (though one might fairly note that her work more broadly can be viewed as an extended argument for the first of these posi-

a more objective prior—gives up the game. If the appropriate response to cases like this is to ignore the apparent base-rate information and use the prior only as a technical tool, then Mayo has made her point.

[17] On my reading, this metaphilosophy of statistics is not just that of Neyman (1952) and Pearson (1962), but also pluralists such as Cox (2006) and arguably some practicing Bayesians (e.g., Gelman and Shalizi 2013; Kass 2011).

tions). Nor should this be surprising—if the disagreements among statistical approaches could be definitively settled by appealing to simple examples like those found in the philosophical literature, we would have expected it to be settled long ago. Cast in this light, I think Mayo's third response is best read not as an argument against Bayesianism writ broadly—among other things, not all Bayesians would endorse the kind of flat-footed picture of testing that leads to trouble here—but instead as a demonstration that the defender of classical statistics is just as capable of constructing problematic scenarios involving base rates as the Bayesian.

6 Conclusion

It's tempting to view the conclusion here as disappointing: what we've discovered is that cases involving base rates do not, or at least do not obviously, help us resolve debates about statistical methodology in either direction. Against this disappointment, it's helpful to remember where we started, namely with Howson expressing the view that classical inferences are simply unsound in virtue of committing a well-known and obvious fallacy. We've shown that that argument rests on a serious mischaracterization of classical theory, and we've also seen that more subtle appeals to the base-rate fallacy trade on that same mischaracterization for their rhetorical force. Indeed, I think our investigation ends with an unusually clear recommendation: philosophers should stop claiming the classical statistics runs afoul of the base-rate fallacy.

Acknowledgments

My thanks to Mike Titelbaum and two referees for comments on a previous version of this paper.

Funding

Funding for this research was provided by the National Science Foundation under Grant No. 2042366.

References

Achinstein, Peter (2001). *The Book of Evidence*. Oxford: Oxford University Press.

— (2010). Induction and Severe Testing. In: *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science*. Ed. by Deborah Mayo and Aris Spanos. Cambridge: Cambridge University Press: 170–88.

Agresi, Alan, Christine Franklin, and Bernhard Klingenberg (2017). *Statistics: The Art and Science of Learning From Data*. 4th ed. USA: Pearson.

Casella, George and Roger Berger (1990). *Statistical Inference*. Belmont: Wadsworth.

Cox, David R. (2006). *Principles of Statistical Inference*. Cambridge: Cambridge University Press.

Dretske, Fred (1971). Conlcusive Reasons. *Australasian Journal of Philosophy* 49.1: 1–22. DOI: 10.1080/00048407112341001.

Fisher, Ronald A. (1922). On the Mathematical Foundations of Theoretical Statistics. *Philosophical Transactions of the Royal Society of London. Series A* 222.594-604: 309–68. DOI: 10.1098/rsta.1922.0009.

— (1958). *Statistical Methods and for Research Workers*. 13th ed. New York: Macmillan.

— (1973). *Statistical Methods and Scientific Inference*. 3rd ed. New York: Macmillan.

Fletcher, Samuel C. and Conor Mayo-Wilson (2024). Evidence in Classical Statistics. In: *The Routledge Handbook of Evidence*. Ed. by Maria Lasonen-Aarnio and Clayton Littlejohn. New York: Routledge: 515–27.

Foley, Richard (1992). The Epistemology of Belief and the Epistemology of Degrees of Belief. *American Philosophical Quarterly* 29.2: 111–24.

Freedman, David, Rogert Pisani, and Roger Purves (2007). *Statistics*. 4th ed. New York: W.W. Norton & Company.

Gardiner, Georgi and Brian Zaharatos (2022). The Safe, the Sensitive, and the Severely Tested: A Unified Account. *Synthese* 200.369: 1–33. DOI: 10.1007/s11229-022-03731-w.

Gelman, Andrew and Cosma Rohilla Shalizi (2013). Philosophy and the Practice of Bayesian Statistics. *British Journal of Mathematical and Statistical Psychology* 66: 8–38. DOI: 10.1111/j.2044-8317.2011.02037.x.

Hawthorne, James (1993). Bayesian Induction Is Eliminative Induction. *Philosophical Topics* 21.1: 99–138. DOI: 10.5840/philtopics19932117.

Howson, Colin (1997). Error Probabilities in Error. *Philosophy of Science* 64.4: S185–94. DOI: 10.1086/392599.

Howson, Colin (2000). *Hume's Problem: Induction and the Justification of Belief*. Oxford: Oxford University Press.
Howson, Colin and Peter Urbach (2006). *Scientific Reasoning: The Bayesian Approach*. 3rd ed. Chicago: Open Court.
Ichikawa, Jonathan Jenkins and Matthias Steup (2017). The Analysis of Knowledge. In: *Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. URL: https://plato.stanford.edu/ENTRIES/knowledge-analysis/ (visited on 08/14/2024).
Kass, Robert E. (2011). Statistical Inference: The Big Picture. *Statistical Science* 26.1: 1–9. DOI: 10.1214/10-STS337.
Korb, Kevin B. (1991). Explaining Science. *British Journal for the Philosophy of Science* 42.2: 239–53. DOI: 10.1093/bjps/42.2.239.
Kyburg, Henry E. (1961). *Probability and the Logic of Rational Belief*. Middletown: Wesleyan University Press.
Lehmann, Erich L. and Joseph P. Romano (2022). *Testing Statistical Hypotheses*. 4th ed. Cham: Springer.
Levi, Isaac (1981). Should Bayesians Sometimes Neglect Base Rates? *Brain & Behavioral Sciences* 4.3: 342–343. DOI: 10.1017/s0140525x00009225.
— (1983). Who Commits the Base Rate Fallacy? *Brain & Behavioral Sciences* 6.3: 502–6. DOI: 10.1017/s0140525x00017209.
Mayo, Deborah G. (1996). *Error and the Growth of Experimental Knowledge*. Chicago, IL: University of Chicago Press.
— (1997a). Error Statistics and Learning From Error: Making a Virtue of Necessity. *Philosophy of Science* 64.4: S195–212. DOI: 10.1086/392600.
— (1997b). Response to Howson and Laudan. *Philosophy of Science* 64.2: 323–33. DOI: 10.1086/392555.
— (2005). Evidence is Passing Severe Tests: Highly Probable versus Highly Probed Hypotheses. In: *Scientific Evidence: Philosophical Theories and Applications*. Ed. by Peter Achinstein. Baltimore: Johns Hopkins University Press: 95–127.
— (2010). Sins of the Epistemic Probabilist: Exchanges with Peter Achinstein. In: *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science*. Ed. by Deborah Mayo and Aris Spanos. Cambridge: Cambridge University Press: 189–201.
— (2018). *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*. Cambridge: Cambridge University Press.
Mayo-Wilson, Conor (2018). Epistemic Closure in Science. *The Philosophical Review* 127.1: 73–114.
Neyman, Jerzy (1952). *Lectures and Conferences on Mathematical Statistics and Probability*. 2nd ed. Washington, D.C.: US Department of Agriculture.

Neyman, Jerzy (1971). Foundations of Behavioristic Statistics. In: *Foundations of Statistical Inference*. Ed. by Vidyadhar P. Godambe and David A. Sprott. Toronto: Holt, Rinehart, and Winston: 1–19.

Neyman, Jerzy and Egon S. Pearson (1928). On the Use and Interpretation of Certain Test Criteria for the Purposes of Statistical Inference. *Biometrika* 20A: 175–240, 263–94. DOI: 10.2307/2331945.

— (1933). On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society Series A* 231: 289–337. DOI: 10.1098/rsta.1933.0009.

Pearson, Egon S. (1962). Some Thoughts on Statistical Inference. *The Annals of Mathematical Statistics* 33.2: 394–403. DOI: 10.1214/aoms/1177704566.

Pritchard, Duncan (2005). *Epistemic Luck*. Oxford: Oxford University Press.

Pritchard, Duncan, John Turri, and Adam Carter (2022). The Value of Knowledge. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. URL: https://plato.stanford.edu/archives/spr2018/entries/knowledge-value/ (visited on 08/14/2024).

Robert, Christian P. (2001). *The Bayesian Choice*. 2nd ed. New York: Springer.

Rosenkrantz, Roger D. (1977). *Inference, Method and Decision: Towards a Bayesian Philosophy of Science*. Dordrecht: D. Reidel.

Rothstein, Richard (2017). *The Color of Law: A Forgotten History of how our Government Segregated America*. New York: Liveright Publishing Corporation.

Seidenfeld, Teddy (1979). *Philosophical Problems of Statistical Inference: Learning from R.A. Fisher*. Dordrecht: D. Reidel.

Smith, Martin (2016). *Between Probability and Certianty: What Justifies Belief*. Oxford: Oxford University Press.

Spanos, Aris (2010). Is Frequentist Testing Vulnerable to the Base-Rate Fallacy? *Philosophy of Science* 77.4: 565–83. DOI: 10.1086/656009.

Spielman, Stephen (1973). A Refutation of the Neyman-Pearson Theory of Testing. *British Journal for the Philosophy of Science* 24: 201–22. DOI: 10.1093/bjps/24.3.201.

— (1974). The Logic of Tests of Significance. *Philosophy of Science* 41: 211–26. DOI: 10.1086/288590.

Sprenger, Jan (2017). Bayesianism vs. Frequentism in Statistical Inference. In: *The Oxford Handbook of Probability and Philosophy*. Ed. by Alan Hájek and Christopher Hitchcock. Oxford: Oxford University Press: 382–405.

— (2018). The Objectivity of Subjective Bayesianism. *European Journal for Philosophy of Science* 8.3: 539–58.

Taylor, Keeanga-Yamahtta (2019). *Race for Profit: How Banks and the Real Estate Industry Undermined Black Homeownership*. Chapel Hill: University of North Carolina Press.

Titelbaum, Michael (2022). *Fundamentals of Bayesian Epistemology 2: Arguments, Challenges, and Alternatives*. Oxford: Oxford University Press.

Tversky, Amos and Daniel Kahneman (1982). Evidential Impact of Base Rates. In: *Judgment under Uncertainty: Heuristics and Biases*. Ed. by Daniel Kahneman, Paul Slovic, and Amos Tversky. Cambridge: Cambridge University Press: 521–52.

Wasserman, Larry (2005). *All of Statistics: A Concise Course in Statistcal Inference*. Cham: Springer.