Probably not that improbable: on inverse probabilities and the problem of the priors

October 21, 2017

Abstract

In this paper, I offer a new response to the problem of the priors for Bayesianism. Or, at least, I offer a way of proceeding in the face of that problem – a way to constrain the range of prior probability distributions which count as reasonable in light of our data. The idea is this: in certain situations, our data's resulting from certain prior probability distributions would entail that our epistemic lives are more exciting than they probably are. And we should, on those grounds I argue, therefore be confident that our data did not result from those distributions.

The paper proceeds in two parts. The first is devoted to laying out the problem, and considering and rejecting responses to it offered by Bayesians, likelihoodists, and the classical statisticians. The second is devoted to describing and defending my own solution.

1 Introduction

You go to the doctor. She tests you for a disease. The true positive and negative rates of the test you undergo are both 95 percent; i.e.,[1]

$$P(+|D) = .95$$
$$P(-|\neg D) = .95.$$

You test positive. How confident should you be that you are afflicted?

My students would respond that there is not enough information. To answer the question,

[1]Here 'D' is the proposition you have the disease, '+' the proposition that you test positive, and '-' the proposition that you test negative.

they would need to know the base rate. If the disease is rare, then it's unlikely you are afflicted. But if it's common, then you should worry.

But here is a puzzle. What if you, the doctor, and everyone else, have no idea what is the base rate? What if your ignorance is so acute that you cannot place non-trivial upper and lower bounds on what it might be? It is possible everyone has the disease, or that no one does. In fact, for every rate r, it's possible the base rate is r. What then is the rational response to a positive result? This question might seem obtuse, because we are never so entirely ignorant. But first that's not so clear. And second the question I'm really asking is 'how does empirical inquiry *begin*, given that it must ultimately begin from a state of empirical ignorance?'

The entirety of statistics and scientific inference hangs on this question. The specific example is not important. I have in mind an idealized picture of the scientific enterprise. We confront a universe of random mechanisms. Those mechanisms distribute disease, or inherited traits, or economic gains and losses. And our goal is to understand the probability distributions which govern them. But there is always a first step: a first piece of data, a first positive result. And it is only if we know how to take the first step that we will be able to take the rest of them.

This, then, is a paper in the foundations of statistics. It is a defense of a classical outlook much like the one defended by R.A. Fisher in the middle of the last century. But I am not obsequious; I argue here that Fisher got a lot wrong.

In the philosophical circles I run in, Bayesianism is ascendant. But while the Bayesian apparatus is elegant and powerful, I can't make sense of it. This, for familiar reasons: Bertrand's paradox, and the problem of the priors more generally. The other main alternative – besides Bayesianism and classicism, that is – is likelihoodism. But, though the central claim of likelihoodism is right, likelihoodism is too austere. It seems to me unable to do the work that I had hoped statistics could do.

This paper proceeds in two parts. The first is destructive. I argue that each of the views mentioned fails to solve our motivating puzzle. The second is constructive; it is devoted to solving it.

Part I

Destruction

2 Bayesianism

Everyone wants to be a Bayesian. For Bayesians, uncertainty always manifests itself in a probability distribution. If you are ignorant of the base rate, there is a probability distribution which expresses or captures that ignorance. But the difficult question is: which distribution?

In answering that question, Bayes considered, and Laplace championed, the Principle of Indifference: evidential parity implies equiprobability. If you are ignorant of the base rate, then you have as much evidence that you are afflicted as that you are not. Hence, before taking the test, the probability that you had the disease must have been 1/2.

But there are damning, perennial, recalcitrant objections to that thought. Suppose that the disease you will be tested for is one of a set of one hundred mutually exclusive diseases. And you are ignorant of each of their base rates. Then for any disease, you have as much evidence you are afflicted as that you are not. (You have exactly zero evidence, either way.) Thus, if evidential parity implies equiprobability, the probability you have each particular disease must be 1/2. Hence the probability that you have at least one of the diseases is given by

$$
\begin{aligned}
&P(\text{You have the first disease}) + P(\text{You have the second}) + \ldots + P(\text{You have the hundredth}) \\
&= 1/2 + 1/2 + \ldots + 1/2 \\
&= 50
\end{aligned}
$$

But 50 is not a probability. Probabilities are less than one. The Principle of Indifference entails a contradiction.

Now, Bayesians have thought about this problem. One response is the following. You should only consider the finest available partition of the possibility space. So, if there are 100 exclusive diseases (and you know you have one of them), then according to the

Principle of Indifference, the probability you have any particular one is $\frac{1}{100}$.

But, first, why? Evidential parity is evidential parity. If evidential parity implies equiprobability, it ought to regardless of the chosen partition. [2] Second, suppose we grant that we should only consider the finest available partition of the possibility space. Returning to our disease with mystery base rate, that rate might take any real value between zero and one. Hence the finest partition is infinitely fine; it contains a cell for every possible rate r. Presumably, the Principle of Indifference recommends a uniform density over that partition.[3] But this doesn't help. It makes things worse; the problem transforms into Bertrand's paradox.

According to the uniform density, the probability is 1/2 that the base rate is less than 1/2. But a uniform density over the base rate entails a nonuniform density over the *squared* base rate. It entails that the probability is greater than 1/2 that the squared rate is less than 1/2.[4] But why should this be? What justifies confidence, in ignorance, that the squared rate is less than 1/2? More importantly, why assign a uniform density over the rate, but not over the squared rate? After all, we have exactly as much evidence that the squared rate is less than 1/2 as we do that the rate is. A similar problem arises for the cubed rate, the quadrupled rate, the square root of the rate, etc.[5]

Now, Harold Jeffreys and Edwin Jaynes did offer (what they took to be) a solution to this problem. But that solution requires adopting an unbounded distribution as your prior. The sum (technically, integral) of the probabilities of each rate hypothesis is infinite.[6] But

[2] Many people object at this point. Surely Laplace did not intend that *wherever* you have evidential parity you have equiprobability. Fair enough. You can see this as an invitation to specify exactly what the Principle of Indifference *is* such that it avoids this consequence.

[3] According to the uniform density it is equally probable that r is in the interval (a, b) and in (c, d) just when (a, b) and (c, d) are of the same length.

[4] $P(r^2 \leq \frac{1}{2}) = P(r \leq \frac{1}{\sqrt{2}}) > P(r \leq \frac{1}{2})$ because $\frac{1}{\sqrt{2}} > \frac{1}{2}$.

[5] Bas van Fraassen made this problem famous in the philosophical community with his cube factory example (van Fraassen, 1989, 302-307). The problem was originally noticed by Bertrand (1889).

[6] Jaynes (1968, 20) recommends and Jeffreys (1961, 123-125) considers

$$f(p) \propto \frac{1}{p(1-p)}$$

as the prior representative of "total confusion or complete ignorance" (Jaynes, 1968, 20) when an unknown parameter p is restricted to lie between 0 and 1. But the integral of that function does not converge. (Because the integral of $\frac{1}{p}$ diverges over (0, 1), and because $0 \leq \frac{1}{p} \leq \frac{1}{p(1-p)}$ over (0, 1), the integral of $\frac{1}{p(1-p)}$ must also diverge over that interval.)

Jeffreys attributes the view that $f(p)$ is appropriate to J.B.S. Haldane (123), but notes that, for example,

an unbounded distribution is not a probability distribution. Probabilities are bounded; they are always less than one.

Other Bayesians have taken another tack. On their view, Bertrand's paradox reveals only that there is no objective starting point in inductive inference. Instead, they contend, "the prior distribution from which a Bayesian analysis proceeds reflects a person's beliefs before the experimental results are known. Those beliefs are subjective, in the sense that they are shaped in part by elusive, idiosyncratic influences, so they are likely to vary from person to person." And "trying to force this ... entirely legitimate diversity of opinions into a singlue unifom one is misguided Procrusteanism."(Howson and Urbach, 2006, 237)

But I do not think this helps with Bertrand's paradox. The problem is not an interpersonal problem. The problem is not that some people want to place a uniform density on the rate, while others want to place a uniform density on the squared rate. The problem is intra-personal. When I ask myself, 'should I adopt a uniform density on the rate, or on the squared rate?' the "elusive, idiosyncratic influences" on my beliefs which are supposed to choose between them do not yield a verdict. Both seem to equally well reflect my evidential station. But I cannot adopt them both; they are inconsistent.

In sum, the question of how to represent ignorance in the Bayesian framework is a vexed one. And I know of no satisfying answer to it. These are not new points; this is just the problem of the priors, the problem that made the Reverend Bayes himself reluctant to be a Bayesian.[7]

---

were the first two people we investigated to have a disease with unknown base rate, our posterior credence that the population wide rate is 1 would be 1. In other words, we would be certain that everyone has the disease. Jeffreys writes, "The rule $\frac{1}{p(1-p)}$ ... would lead to the conclusion that if a sample is of one type with respect to some property, there is probability 1 that the whole population is of that type" (124). After considering some alternatives, and also finding them wanting, Jeffreys ultimately concludes "we may as well use the uniform distribution ... in the present state of knowledge, that is enough to be going on with" (125). This is just to say that Jeffreys ultimately does not offer a solution to the version of Bertrand's paradox I've presented here.

[7]Fisher, at least, thought so

> Bayes' introduction of an expression representing probability *a priori* contained an arbitrary element, and it was doubtless some consciousness of this that led to his hestitation in putting his work forward. (Fisher, 1956, 17)

3 Likelihoodism

I turn then to likelihoodism. Likelihoodists[8] abandon the central Bayesian dogma. Some evidential states – for example empirical ignorance – cannot be captured by a probability distribution. Sometimes no prior is available.

But to give up priors comes at a cost. Bayesianism allows us to model the acquisition of evidence via Bayes' Theorem. We can condition on our evidence and arrive at posterior probabilities. Thus, we can say how probable it is you have a disease, if you test positive for it in ignorance of the base rate. But the machinery of Bayes' Theorem runs on a prior; it cannot get started if you do not supply one.

And Likelihoodists accept this. They conclude that because we sometimes lack priors, we cannot in those cases locate posterior probabilities. Here is Edwards, accepting that consequence with relish

> It is indeed true that [likelihoodism] ... does not make any assertion about the probability of a hypothesis being correct [in light of the data]. And for good reason: the [view] has been developed by people who explicitly deny that any such statement is generally meaningful in the context of a statistical hypothesis. (Edwards, 1972, 33)

So consider again our disease with mystery base rate. The likelihoodist accepts the upshot of Bertrand's paradox. To settle on a uniform density on the base rate, instead of on the squared base rate, would be arbitrary. But without a prior, we cannot arrive at posterior probabilities. And hence we cannot answer the question of how probable it is you have the disease, upon testing positive. The best we can do is consider the relative likelihoods of this or that hypothesis – i.e., the relative probabilities of our data given this or that hypothesis. So the hypothesis that you are afflicted is best, if you test positive, because given you are afflicted, the probability of a positive test is higher than given you are not. But to say that hypothesis is best is not to say you should be confident it's true, nor that it's probably true.

My main complaint about likelihoodism is not that it's wrong but that it's too austere. Of course we can rank hypotheses by the probabilities each assigns to our observed data. But

[8] like Hacking (1965), Edwards (1972), and Sober (2008) (in certain moods).

I don't think that's the question we were interested in.

Imagine, for example, you test yourself for the disease with mystery base rate one million times. Assume, again, true positive and negative rates of 95 percent. Now suppose that nine hundred and fifty thousand of those one million tests are positive. Rationality requires, it seems to me, that you be confident that you have the disease. It requires that you think it highly probable that you have the disease. But likelihoodism does not deliver this result. The likelihoodist will agree that the hypothesis that you are afflicted is best. (After all, that hypothesis assigns the highest probability to your data.) But without a prior, likelihoodism "does not make any assertion about the probability" that you are afflicted. Thus even after nine hundred and fifty thousand positive tests, it is silent on the question of whether you should be confident you have the disease.

Of course, Likelihoodism could be amended. We could insist on certain thresholds – after 100 positive tests, be confident you are afflicted. But those thresholds and their justification need explication. Until that's supplied, the view looks too thin.[9]

4 Classicism

I turn now to the classicists, typically represented by the trio of Ronald Fisher, Jerzy Neyman and Egon Pearson.

In philosophical circles, classical statisticians get a bad rap. Howson and Urbach, for example, claim that classical significance tests yield conclusions which "often flatly contradict those which an impartial scientist or ordinary observer would draw."(Howson and Urbach, 2006, 154 quoted in Greco, 2011).

The classicists did make mistakes,[10] but they are often read uncharitably. They were aware of the problem of the priors for Bayesianism.[11] And they were aware that we could interpret

[9]This is, in essence, the critique of likelihoodism, pressed convincingly and thoroughly, by Greg Gandenberger (2016).

[10]Interested readers are referred to Sober, Chapter 1, and Howson and Urbach, Chapter 5.

[11]About Bayesianism, Fisher wrote

> Certainly cases can be found, or constructed, in which valid probabilities *a priori* exist, and can be deduced from the data. More frequently, however, and especially when the probabilities of contrasted scientific theories are in question, a candid examination of the data at the disposal of the scientist shows that nothing of the kind can be claimed. (Fisher, 1956, 17)

evidence via likelihoods.[12] I think we should read them as trying to push likelihoodism further than its austere beginnings. They wanted to see how far they could get without having to invoke prior probabilities. Now perhaps they didn't get very far, but at least some of the criticisms leveled at classical methods fall flat when those methods are viewed in that light.

So what is classicism? I want to abstract away from the details as much as possible. But consider again our disease with mystery base rate, and our test, whose true positive and negative rates are both 95%:

$$P(+|D) = .95$$
$$P(-|\neg D) = .95$$

Here is a general fact that the classicists noticed. Let the unknown base rate be r, and ask: supposing you are about to take the test, how probable is it that you will get an *accurate* result? It's surprising, but that question has a fixed and knowable answer, even though the base rate is unknown. For the test is accurate just in case you test positive, and have the disease, or test negative and lack it. And the chance of *that* is given by

$$\begin{aligned} P((+ \wedge D) \vee (- \wedge \neg D)) &= P((+ \wedge D)) + P(- \wedge \neg D) \\ &= P(+|D)P(D) + P(-|\neg D)P(\neg D) \\ &= .95 \times r + .95 \times (1 - r) \\ &= .95 \end{aligned}$$

Thus, you do not know how probable it is that a positive test will be accurate. And you do not know how probable it is that a negative test will be accurate. But you do know, *in general*, how probable it is that you will receive an accurate result. Roughly put, if one

[12] In certain places, Fisher seemed to explicitly endorse Likelihoodism:

> Although some uncertain inferences can be rigorously expressed in terms of mathematical probability, it does not follow that mathematical probability is an adequate concept for the rigorous expression of uncertain inferences of every kind. ... More generally, a mathematical quantity of a different kind, which I have termed mathematical likelihood, appears to take [the place of probability] as a measure of rational belief. (Fisher 1935, 474, quoted in Lehmann 1993)

billion people took the test, and everyone trusted her result, 95 percent of them would be right. And this is true, again, *no matter what the base rate is*. The point is sometimes put this way: we can know the *pre-data* (or *pre-result*) probability that your test will be accurate, even though we do not know the *post-data* (or *post-result*) probability that it was.

Now, at this point, a rift opens up between our principals. Fisher stands on one side of it, and Neyman and Pearson stand on the other. Allow me to describe their views and my misgivings about each in turn.

4.1 Fisher

I begin with Fisher. As I read Fisher, he wanted to simply pivot on the accuracy of the test and adopt it as a posterior credence. If you test positive, Fisher thought you should be 95 percent confident that you have the disease. And he seemed to argue for this as follows: we have no idea what the base rate is. But we do know that the test will be accurate 95 percent of the time. And in ignorance, it seems rational to *fall back* on the fact that 95 percent of tests are accurate – to use that to determine our posterior credences.

Consider, Fisher might offer, an analogy. Suppose you select Bob at random from a group of people, 95 percent of whom have a heart condition. You are thus 95 percent confident that Bob has a heart condition. But you then learn that Bob is from San Diego. In light of that information, how confident should you be that Bob has a heart condition? Should you, in other words, revise your credences upon learning that Bob is from San Diego?

It's less obvious, but a base rate problem arises here. To know the probability that Bob has a heart condition, it seems you need to know the rate of that heart condition *among San Diegans*. Or at least, you need to know the rate among the San Diegans in the group from which Bob was selected. If they all have it, the probability Bob does is one. If half have it, the probability Bob does is one half. That said, in ignorance of the proportion of San Diegans who are afflicted, it seems reasonable to fall back. It seems reasonable to let the fact that Bob was selected from a group of people – 95 percent of whom are afflicted – guide and determine your credences.

Fisher wanted to say the same thing about accuracy and our test results. We *know* that 95 percent of results are accurate. It's true that we do not know what proportion of the

positive tests are accurate. But, in ignorance, it seems perfectly reasonable to fall back, and let the 95 percent accuracy of the test guide and determine our credences. Hence, if you test positive in ignorance of the base rate, you ought be 95% confident you have the disease.[13]

Fisher's view is under-appreciated. He is offering us a middle ground in the debate between likelihoodists and Bayesians. We can agree with the likelihoodist that the problem of the priors is fatal to Bayesianism. But at the same time, we can deny the likelihoodists' pessimistic conclusion – that posterior probabilities are out of reach. In fact, we *can* locate and endorse posterior probabilities. We need only find tests which are accurate with a known probability. And our posterior credences can then be guided by the accuracy of

[13]This is, I think, the easiest way to understand Fisher's *fiducial* argument. He noticed that, in certain circumstances, we could make general probability statements which are true regardless of prior probability distributions. He considers, for example, a random variable X which follows an exponential distribution with rate parameter θ, and notes that the quantity $2\theta X$ will follow a χ^2 distribution regardless of the value of θ. From that, it follows that, if $\chi^2(P)$ is the P-th percentile of a χ^2 distribution, θ will exceed $\frac{\chi^2}{X}$ with probability P.

He then writes,

> The probability statement [– that θ will exceed $\frac{\chi^2}{X}$ with probability P –] had as a reference set all the values of X which might have occurred in unselected samples for a particular value of θ. It has, however, been proved for all values of θ, and so is applicable to the enlarged reference set of all pairs of values (X, θ) obtained from all values of θ. **The particular pairs of values θ and X appropriate to a particular experimenter certainly belongs to this enlarged set,** and within this set the proportion of cases satisfying the inequality
>
> $$\theta > \frac{\chi^2(P)}{X}$$
>
> ... is certainly equal to the chosen probability P. (Emphasis added.) (Fisher, 1956, 54)

The comment I emphasized above is, I think, crucial. When Fisher refers to the pairs of values "appropriate to a particular experimenter", he just means that the general probability statement is true regardless of the experimenter's prior. In other words, no matter what your prior over θ is, before the data X comes in, you should think that θ will exceed $\frac{\chi^2(P)}{X}$ with probability P.

Fisher then goes on to write,

> It might have been true ... that in some recognizable subset of pairs (X, θ) ... the proportion of cases in which θ exceeds $\frac{\chi^2(P)}{X}$ should have had some value other than P. It is the stipulated absence of knowledge *a priori* of the distribution of θ ... that makes the recognition of any such subset impossible, and so guarantees that in [the experimenter's] particular case ... the general probability is applicable. (Fisher, 1956, 55)

Here, Fisher is just saying that if you *knew* the appropriate prior probability distribution to adopt over θ, you might, upon observing your data, have reason to abandon the general probability statements, which you recognized as true before the data came in. But, he contends, our ignorance somehow "guarantees" that the general probability statement must remain applicable, after you see the data.

those tests.

Now, some authors have claimed that this doesn't really count as progress. Fisher is implicitly invoking a prior, they contend, even if he is unwilling to admit it. (Edwards (1972, 208-209) and Bulmer (1967, 179) both raise versions of this criticism.) After all, if the posterior probability that you have the disease is .95, then the prior probability must have been 1/2.[14]

But while that's right, it's hard for me to see how this amounts to much of a criticism of Fisher. The whole *problem* with the Bayesian approach is that it is impossible to locate a prior probability. At worst, then, Fisher has supplied us with a method by which we can locate a rational prior. We can say, 'the *reason* this prior is reasonable is that, in using it, our credences will be guided by the objective accuracy of our tests. *That* is why you should adopt a prior of 1/2 in this case; it has nothing to do with evidential parity.' And that, I think, would count as progress.

There is, however, a more difficult problem for Fisher's view. And here's a trivial way to see it. Suppose you have a fair coin. On one side of it, you write 'you have the disease'. On the other, you write, 'you don't.' Then, by the considerations above, it seems you ought to reason as follows. You know that that coin will give you an accurate report 50 percent

[14]

$$\begin{aligned} P(D|+) &= .95 \\ &= \frac{P(+|D) \cdot P(D)}{P(+)} \quad \text{(Bayes' Theorem)} \\ &= \frac{.95P(D)}{P(+|D) \cdot P(D) + P(+|\neg D) \cdot P(\neg D)} \\ &= \frac{.95P(D)}{.95 \cdot P(D) + .05 \cdot (1 - P(D))} \\ &= \frac{.95P(D)}{.9P(D) + .05} \end{aligned}$$

Hence

$$\begin{aligned} .95 &= \frac{.95P(D)}{.9P(D) + .05} \\ .9P(D) + .05 &= P(D) \\ .05 &= .1P(D) \\ P(D) &= .5 \end{aligned}$$

of the time, whether you have the disease or not. So if you flip it, and it says you have the disease, on Fisher's view you ought to adopt a posterior credence of 1/2 that you do. (Same goes, I suppose, if it says you don't have the disease.)

At first blush, maybe that doesn't seem so bad. It looks like an application of the Principle of Indifference. But now imagine you also have a fair three-sided die. On one side, you write 'you have the disease', on another, you write 'you don't and it will rain tomorrow', and on the third, you write 'you don't and it won't rain tomorrow.' (Assume you are ignorant both about the climate, and about whether you have the disease.) You roll the die, and it reads 'you have the disease.' Well, again, regardless of what the truth is, you know the die will report that truth with probability 1/3. And so, by the considerations above, it seems you ought to adopt a posterior credence of 1/3 that you are afflicted.

The problem is that neither the coin nor the die gives you any evidence. So imagine one person flips the coin and reads 'you are afflicted', and another rolls the die and reads the same. Oddly, on Fisher's view, the coin flipper should be more confident that she is afflicted than the die roller.[15] [16]

If we *do* allow our credences to be determined by these mechanisms, problems will percolate into our epistemic futures. Suppose I flip the coin, and arrive at a 1/2 credence I have

[15] And we can make this worse. Imagine a fair million-sided die. And suppose you have a random number generator which generates whole numbers between 1 and 999,999. On one side of the die, you write, 'you have the disease,' on another you write, 'you don't, and this random number generator will generate 1', on another you write, 'you don't, and this random number generator will generate 2', etc. Well then *that* die will again be accurate once every million rolls. And so, if the die says 'you have the disease', on Fisher's picture your posterior credence that you do ought to be one in one million. But again, the die supplies no evidence one way or another. So if you weren't already, you should not become 99.9999% confident that you do not have the disease, if and just because that million sided die reported you do.

[16] Sophisticates might try to save Fisher by saying something like the following: if you read Fisher, you will find that his whole view rested on the notion of a *pivotal quantity.* A *pivotal quantity*, recall, is a function of the possible data and hypotheses which is independent of which hypothesis is true. If, for example, X is a normally distributed random variable with unit variance and with hypothesized but unknown mean μ, then $(X - \mu)$ is a pivotal quantity – the probability that it takes any particular value is independent of the value of μ.

And the sophisticate will rightly point out that I have not said anything about pivotal quantities here. But, in fact I have been discussing pivotal quantities, just not by that name. Consider again our fair coin; on one side of it, it says 'you have the disease', on the other, it says 'you do not.' And now consider the proposition *the coin reports the truth.* That proposition is a function of the possible data – what the coin reports – and the hypotheses – whether or not you have the disease, into the set $\{\textsc{True}, \textsc{False}\}$. And the probability that that function takes the value $\textsc{True}$ is 1/2, regardless of whether or not you have the disease. Hence that proposition *is* a pivotal quantity, and hence on Fisher's view you ought to pivot on it, and use it to arrive at posterior probabilities.

the disease. You roll the die, and arrive at a 1/3 credence you do. Suppose we then both conditionalize on the same evidence – a collection of positive tests, say. Then we will arrive at different conclusions about the probability we are afflicted.

Thus the problem of the priors re-arises for Fisher. We don't actually need to construct these coins and dice, and see how they behave. We already know that they will provide no further evidence than we already have. And so, if we were to let them guide our credences, we would need to arbitrarily choose a starting point.

4.2 Neyman and Pearson

I turn, lastly, to Neyman and Pearson. Like Fisher, Neyman and Pearson were impressed that we can know the pre-result probability that a test will be accurate, though we do not know the post-result probability that it was. But they were less committal than Fisher about the upshot of that fact.

If Fisher is like a Bayesian, then Neyman and Pearson are like the likelihoodists. They thought that in ignorance of the prior, we should give up on posterior probabilities. Instead, they contended, we should focus on "rules of behavior" that we will follow upon observing our data. "Without hoping to know whether each separate hypothesis is true or false," they wrote, "we may search for rules to govern our behaviour with regard to them, in following which we insure that, in the long run of experience, we shall not too often be wrong."(Neyman and Pearson, 1933, 291)

One rule of behavior, for example, is the following. If you confront a positive test, accept that you have the disease. Otherwise accept that you do not. (Let us, along with Neyman and Pearson, leave the notion of 'acceptance' vague.) If you follow that rule, then when you take our test for the disease with mystery base rate, there is a 95% chance that you will accept the truth on any given occasion.

But for Neyman and Pearson, the probability of a hypothesis is always relativized to a rule of behavior. You cannot say categorically, as Fisher or the Bayesians wanted to, 'the probability I have this disease is 95 percent.' For there are often many rules available to you. And there is no fact of the matter about which rule you are following on a particular occasion. Here, for example, are two rules you might follow upon confronting a positive test:

1. Accept that you have the disease only if you test positive, otherwise, accept that you do not have the disease.

2. Accept that you have the disease regardless of your test result.

Notice that to follow each of these rules is to accept that you have the disease, given you've tested positive. But while following the first will lead you to accept the truth 95 percent of the time, following the second will not. So the question is: is accepting that you have the disease *on this particular occasion* to follow the first or the second rule? Depending on how we answer that question, we will be led to different conclusions about the probability that you have accepted the truth. (This is just the reference class problem.)

As with likelihoodism, I do not think Neyman and Pearson's broad view is wrong. The commitments of the view are *right*. It is true that if we follow these rules, we will in fact accept the truth with some known probability (relative to repeated followings of that rule).

But imagine again you test positive nine hundred and fifty thousand times in one million tests. Then it's true that there is a rule of behavior, which you could follow, and which will very, very often lead you to accept the truth. That rule is this one:

(1) If you test positive 950,000 times in 1 million tests, accept that you have the disease, and if you test negative 950,000 times, accept that you do not.

But so what? Should we, in light of this rule, then be confident that you have the disease or not, and on what grounds? Or is there nothing definitive to be said here? Can, perhaps, the question of how confident we should be *itself* only be answered relative to a rule of behavior?

The intuitive thing to say is this. If you are following a rule which will lead you to accept the truth $p\%$ of the time, then you ought to be $p\%$ confident that you have accepted the truth on this occasion. But that is Fisher's view. Fisher, recall, wanted to pivot on the accuracy of a test to arrive at posterior probabilities. And as we saw, Fisher's view yields inconsistent proscriptions.[17]

[17]The main problem I raised for Fisher is also awkward for Neyman and Pearson. If you flip a fair coin, one rule of behavior available to you is this one: accept that you have the disease if and only if the coin yields heads. And following that rule will lead you to accept the truth fifty percent of the time. But it is unclear how or whether that fact should bear on your confidence that you are afflicted.

Now, as with likelihoodism, Neyman and Pearson's view could be amended. We could insist that, when you follow a rule which will lead you to accept the truth at least p percent of the time, be confident that you've accepted the truth on this particular occasion. But also, as with likelihoodism, that amendment needs explication and then justification.

In sum, in spite of valiant effort, the classicists didn't make much progress on the original problem. Fisher inherits the Bayesian problem of the priors. And Neyman and Pearson inherit the likelihoodist problem of austerity.

Part II

Construction

5 Introduction, redux

I turn finally to the constructive part of this paper, to my answer to the question: how confident should you be that you have a disease, if you test positive for it in ignorance of the base rate?

My aim here is to lay out a view which does not rely on Bayesian subjective priors, but which I think avoids the problems which I raised above for the likelihoodists, Fisher, and Neyman and Pearson. One way to understand the view I offer here is as follows: I'm going to give an argument that certain prior probability distributions are unreasonable, in light of one's data, and that will entail facts about the minimal rational credence you should adopt.

I want to begin by noting that, in the face of one positive test, it seems reasonable to reassure oneself thus: 'It's true I tested positive. But it's also possible this disease is rare, in which case I needn't worry.'

There are two lessons I want to draw from this.

First, when we are thus reassured, we are focused on a prior probability. To say the disease is rare is just to say that the prior probability you have it is low. But our focus is an objective prior. It must be. You cannot reassure yourself with a subjective prior: 'It's true

I tested positive. But perhaps I was antecedently very confident that I didn't have this disease.' That thought is not coherent. Thus I think the question we want answered, when considering this problem is, 'what is the objective probability I am afflicted, in light of my data?' That is the question that I think we care about. We are not asking after antecedent confidences.

But notice the objective probability I am afflicted, in light of my data, is inescapably sensitive to the objective prior probability that I was afflicted. In our case, the probability I have a disease is inescapably sensitive to the base rate. And hence – if we're going to answer the question we want answered – we must somehow marshal our data in the service of determining objective priors.[18] We must, in other words, somehow marshal our data in the service of determining the base rate. That is the first lesson.

Second, I take it for granted that, at a certain point, it becomes unreasonable to be reassured by the thought that the base rate could be low. Supposing you test positive 950,000 times in 1 million tests, it would be odd to reassure yourself, 'perhaps the base rate is *minuscule* or even *zero*. And thus I needn't be worried.' It is true, of course, that if the base rate is zero, you shouldn't be worried. The posterior chance you are afflicted is also zero. But I take it as a datum that that response is too optimistic. And I think our view must account for this transition. It must explain why it might be reasonable to be reassured after one positive test, but not after 950,000.

6 The Proposal

Let us, then, consider the easier case: suppose you test positive 950,000 times in 1 million tests. This thought, I claim, is too optimistic: 'Perhaps the base rate is *minuscule* or even *zero*. And thus I needn't be worried.' But why is that thought too optimistic?

Here is my answer. Suppose the base rate *is* minuscule. Suppose it is so small that it is unlikely you are afflicted, even after 950,000 positive tests. Then I think there is something objectionably *remarkable* about your data. That hypothesis entails that your epistemic life is more exciting than it probably is.

[18]The discussion to follow puts things in terms of base rates, and objective chances. But I think the argument can apply to any notion of objective probabilities you favor – for example, the evidential probabilities of Williamson (2000, 209-230).

Note that if the base rate is minuscule, then one of two things must be true. Either you are afflicted in defiance of the fact that barely anybody is. Or, you've received a mountain of evidence in support of a hypothesis which is false. More succinctly, if the base rate is minuscule, you've received a mountain of evidence in support of a hypothesis that had a minuscule prior chance of being true.

Now, that on its own need not be remarkable. Suppose, for example, I roll a million-sided die. But you then learn the die came up 17. That had a 1 in 1 million shot of occurring, and so you've received evidence in support of a hypothesis which was antecedently unlikely to be true (namely, that the die *would* come up 17). But there is nothing remarkable about that.[19]

So 950,000 positive tests – in defiance of a minuscule base rate – are not remarkable *simply* because they provide a lot of evidence in support of a hypothesis which was unlikely to be true. Instead, I contend they are remarkable because,*when you confront a test like ours with only two possible outcomes*, it is very unlikely that that test should supply evidence as strong as 950,000 positive tests in support of a hypothesis that had a small prior chance of being true.

Here's another way of expressing the remarkability thought: it is not remarkable when a fair, million-sided die comes up 17. But it *is*, I contend, remarkable if a coin comes up heads that was weighted 999,999 to 1 in favor of tails. This, despite the fact that both events – the die's coming up 17, and the coin's coming up heads – are equally improbable. And here's what I think explains that difference: before you roll a million-sided die, you know you will get evidence in support of some hypothesis which had a one in one million chance of being true. But before you flip a coin biased 999,999 to 1 in favor of tails, it is not at all likely that your evidence will support a hypothesis that had a 1 in 1 million chance of being true. What is overwhelmingly likely to happen is that your evidence will support a hypothesis that was likely to be true to begin with. And it is remarkable when it fails to do so.

[19]Paul Horwich offers the following account of surprising data: data is surprising if a hypothesis which was unlikely to be true is confirmed, and a hypothesis that was likely to be true is disconfirmed. (Horwich, 1982, 100-101)

I take it the example in this paragraph shows that that account fails. Consider the hypothesis 'The die will not come up 17.' Before you roll a million-sided die, that hypothesis is very likely to be true. Now consider the hypothesis 'The die will come up 17.' That hypothesis is very *unlikely* to be true. So when the die does come up 17, a hypothesis which was unlikely to be true is confirmed at the expense of one that was likely to be true. But that's not at all surprising. So Horwich's account is wrong.

So when I say that 950,000 positive tests would be remarkable, were the base rate minuscule, all I mean is this: Prior to taking that many tests, you should think it very unlikely that evidence that strong *would* support a hypothesis which was unlikely to be true. After all, in order for that to occur, you would either (i) have to test positive 950,000 times while the base rate is very small or (ii) test negative 950,000 times while the base rate is very large. But if the base rate is very small, then it is very unlikely that you have the disease. And hence it is very unlikely that you would test positive 950,000 times. Similarly, if the base rate is extremely high, it is very likely that you are afflicted. And hence it is very unlikely that you would test negative 950,000 times. And thus *in general*, in a situation like this one, it is very unlikely that evidence as strong as 950,000 positive tests *would* support a hypothesis that had a minuscule chance of being true in the first place.

(You might object at this point: even if the base rate is *not* small, it is unlikely that I would test positive *exactly* 950,000 times (as opposed to, say, 949,487 times, or 950,132 times, or whatever). That's right, but it is a red herring. When I say 'it is unlikely that evidence as strong as 950,000 positive tests would support a hypothesis that had a small chance of being true,' I mean that even *conditional* on your receiving evidence as strong as 950,000 positive tests, it is very unlikely that your evidence would support a hypothesis that was unlikely to be true in the first place. So it is not merely unlikely in the sense that any possible sequence of test results is unlikely.)

Now, there is one last pillar in my proposal. In ignorance of the base rate, I contend, we should not think ourselves or our data remarkable. We should not take our data to supply evidence that that data supports a hypothesis which had a small chance of being true in the first place.

I take that last claim to be fundamental. It's meant to be akin to the Principal Principle or the Principle of Indifference. Bedrock. But allow me at least a bit of rhetoric. When we come into the world, and we confront it in empirical ignorance, nature nevertheless supplies us with an *a priori* guarantee – that our evidence and our epistemic lives will likely be boring and staid. In situations like the one we've been considering, we can be confident that our evidence will support a hypothesis that was fairly likely to be true to begin with. (Or, at least, we can be confident our evidence will not support a hypothesis that was wildly unlikely to be true to begin with.) And it would be odd, in ignorance and after the data comes in, to suddenly think you or your data exceptional.

But if that's right, you should be very confident that your data does not support a hypothesis that was unlikely to be true to begin with. And hence, if you receive very strong evidence in support of your being afflicted, you should be confident that the base rate is *not* minuscule. And hence you should be confident that it is probable you are afflicted, given your test results.

Here's one last, perhaps simpler, way to describe the view. Imagine a debate between Edwards (the likelihoodist) and Fisher. Fisher tests positive 950,000 times for a disease with mystery base rate. Edwards laments, 'There's really nothing to be said about the probability you are afflicted, Fisher, without our knowing the base rate.' Fisher responds, 'But I know this many tests will give me an accurate result with an extremely high probability.' Edwards answers, 'That's true, Fisher, but sometimes very accurate tests indicate hypotheses are true which were very unlikely to be true.'

My contribution just comes in at the end of that conversation. Edwards is right; sometimes extremely accurate tests *do* support hypotheses that had a small prior chance of being true in the first place. But my point is: in the context we've been considering, *that itself* is unlikely to happen. So while we *should worry* about that possibility, I think we should only lend it as much credence as it deserves.

7 The original question answered

Let me, for thoroughness, work through the details of how the view I've offered responds to our original example. I'll then turn to the question of how to generalize the picture.

You test positive in ignorance of the base rate, and your test has fixed true positive and negative rates of 95 percent:

$$P(+|D) = .95$$
$$P(-|\neg D) = .95$$

Now, consider the possibility that the base rate might be less than 1 in 1000. If so, then your evidence supports a hypothesis which had a fairly low chance of being true.

So here is the question I think is relevant: prior to taking the test, what was the probability

that evidence as strong as your evidence should have supported a hypothesis which had less than a 1 in 1,000 chance of being true?

For your evidence to do so you would need to test positive while the base rate $r < .001$ or test negative while the base rate $r > .999$. [20] And the probability of *that* is given by

$$\begin{aligned} c((+ \wedge r < .001) \vee (- \wedge r > .999)) &= c(+ \wedge r < .001) + c(- \wedge r > .999) \\ &= c(+|r < .001)c(r < .001) + c(-|r > .999)c(r > .999) \end{aligned}$$

Considering the left term, $c(+|r < .001)$, note

$$\begin{aligned} c(+|r < .001) &= c((+ \wedge D) \vee (+ \wedge \neg D)|r < .001) \\ &= c(+ \wedge D|r < .001) + c(+ \wedge \neg D|r < .001) \\ &= c(+|D \wedge r < .001)c(D|r < .001) + c(+|\neg D \wedge r < .001)c(\neg D|r < .001) \\ &= c(+|D)c(D|r < .001) + c(+|\neg D)c(\neg D|r < .001) \\ &= .95c(D|r < .001) + .05c(\neg D|r < .001) \\ &= .95c(D|r < .001) + .05(1 - c(D|r < .001)) \\ &= .9c(D|r < .001) + .05 \\ &< .9 \times .001 + .05 \\ &= .0509 \end{aligned}$$

By a symmetric argument

$$c(-|r > .999) < .0509$$

Hence

$$\begin{aligned} c((+ \wedge r < .001) \vee (- \wedge r > .999)) &< .0509c(r < .001) + .0509c(r > .999) \\ &= .0509(c(r < .001) + c(r > .999)) \\ &< .0509 \end{aligned}$$

Thus, you ought to think, prior to taking the test, there is at most a 5.09% chance that

[20] I take it for granted here that, whatever our account of evidential strength, a positive test is as strong evidence that you have the disease as a negative test is that you lack it.

your data will support a hypothesis which has less than a 1 in 1,000 chance of being true. And hence I think it would be mildly remarkable if it did.

And on my view, you ought not take a positive test to be evidence that your data is remarkable. And thus, if you test positive in ignorance of the base rate, you ought be 94.91% confident that the base rate is at least 1 in 1000. And hence, you ought be 94.91% confident that the chance you are afflicted, conditional on your positive test, is at least 1.7%.[21]

Now that is a weak thing to say. But importantly it's not as weak as what the likelihoodists and Neyman and Pearson say. Because as more data comes in, we'll be able to say more. Suppose, for example, you test positive 10 times. Then the chance of receiving evidence as strong as that in support of a hypothesis which had, say, less than a 1 in 1 billion chance of being true is itself approximately 1 in 1 billion. And hence on my view, if you test positive 10 times, you ought to be very confident that the base rate is at least 1 in 1 billion. And hence you ought to be very confident that the posterior chance you are afflicted is high. The view thus captures the natural thought, that after a single positive test, you should be uncertain about the chance you are afflicted, but that after 10 positive tests, or 950,000, you should be worried.

8 Generalizing

Allow me briefly to sketch how I want to generalize the view described above to more complicated cases. What if, for example, we have more than two hypotheses, or more than

[21] This just follows from Bayes' Theorem, and the fact that it is a monotone increasing function of the prior. If $c(D) > .001$, then

$$\begin{aligned} c(D|+) &= \frac{c(+|D)c(D)}{c(+)} \\ &= \frac{.95c(D)}{c(+|D)c(D) + c(+|\neg D)c(\neg D)} \\ &= \frac{.95c(D)}{.9c(D) + .05} \\ &> \frac{.95 \times .001}{.9 \times .001 + .05} \\ &= 1.7\% \end{aligned}$$

two possible pieces of evidence? For now let me simply offer another simple case; I leave the full reckoning to future work.

Let us finally abandon our disease with mystery base rate. Suppose you confront 1,000 identical cups arranged into a circle. One of the cups contains a ball, you do not know which. But suppose that in the center of the circle is something that looks like a compass needle. And that needle behaves as follows: given the ball is in some particular cup, there is a 95 percent chance that the needle will indicate that cup. But the needle sometimes misfires, and when it does so, it does randomly. Now suppose that the needle indicates cup 17.

Here we have 1000 hypotheses and 1000 possible pieces of evidence. So now let us ask, as we did above: prior to spinning the needle, how confident were you that your evidence would, most strongly, support a hypothesis that was unlikely to be true, say, had a 1 in 1000 or less chance of being true?

And notice a problem arises here, which did not arise when considering our disease with mystery base rate. Suppose the ball was placed randomly, and hence the prior chance that the ball would end up in any particular cup was 1 in 1000. In other words, suppose the chance distribution over the space of cups was uniform. Then it was a *certainty* that you would receive evidence as strong as your evidence in support of a hypothesis that had a 1 in 1000 or less chance of being true.

The moral is that, in this case, *you cannot place an upper bound* on the probability that you should have observed evidence as strong as yours in support of a hypothesis that had a 1 in 1000 or less chance of being true. But this is not really a problem for the view I've offered here, for we can still find upper bounds. For example, we can ask, what are the chances that I should have observe evidence in favor of a hypothesis that had less than a 1 in $10,000$ chance of being true? And by a crude calculation,[22] the chance of that is less than 1 in 10. And hence, on my view, you ought to be 90% confident that the posterior chance the ball is in cup 17 is at least 65%. (This last, again, just follows from Bayes' Theorem.)

[22]Let $\uparrow i$ be the propostion 'the needle indicates cup i' and let c_i be the chance the ball is in cup i. Then the chance that our evidence will most strongly support a hypothesis that had less than a 1 in 10,000

9 Conclusion

There is work left to do – the generalization just sketched needs to be formally articulated, and the range of cases to which the view applies needs to be circumscribed. But I'm optimistic that both can be done fairly straightforwardly.

For now, let me just reiterate the main point. In ignorance of the base rate, or, in ignorance of the chance distribution over some space of hypotheses in general, we can nevertheless arrive at confidence that those hypotheses supported by our evidence are true. To do so, all that's required is that we be confident that our epistemic lives are unremarkable ones – that our evidence supports hypotheses which were relatively likely to be true to begin with. And as I've shown here, at least generally speaking, it is very probable that our epistemic lives are unremarkable in that respect.

References

Bertrand, J. (1889), *Calcul des probabilites*. Gauthier-Villars.

---

chance of being true is given by

$$
\begin{aligned}
& c\Big((\uparrow 1 \wedge c_1 < \frac{1}{10,000}) \vee (\uparrow 2 \wedge c_2 < \frac{1}{10,000}) \vee \ldots \vee (\uparrow 1,000 \wedge c_{1,000} < \frac{1}{10,000})\Big) \\
=& c(\uparrow 1 \wedge c_1 < \frac{1}{10,000}) + c(\uparrow 2 \wedge c_2 < \frac{1}{10,000}) + \ldots + c(\uparrow 1,000 \wedge c_{1,000} < \frac{1}{10,000}) \\
=& c(\uparrow 1 | c_1 < \frac{1}{10,000}) c(c_1 < \frac{1}{10,000}) + \ldots + c(\uparrow 1,000 | c_{1,000} < \frac{1}{10,000}) c(c_{1,000} < \frac{1}{10,000}) \\
<& \underbrace{(.95 \times \frac{1}{10,000} + \frac{.05}{999} \times \frac{9,999}{10,000})}_{c(\uparrow i|i)c(i|c_i < \frac{1}{10,000}) + c(\uparrow i|\neg i)c(\neg i|c_i < \frac{1}{10,000})} \quad (c(c_1 < \frac{1}{10,000}) + \ldots + c(c_{1,000} < \frac{1}{10,000})) \\
\approx& \frac{1}{10,000}(c(c_1 < \frac{1}{10,000}) + \ldots + c(c_{1,000} < \frac{1}{10,000})) \\
<& \frac{1}{10,000}(1 + \ldots + 1) \\
<& \frac{1000}{10,000} \\
=& \frac{1}{10}
\end{aligned}
$$

That calculation, as I said, was crude, and we could find a smaller upper bound if we worked harder.

Bulmer, M.G. (1967), *Principles of Statistics*. Dover Publications.

Edwards, A. W. F. (1972), *Likelihood: an account of the statistical concept of likelihood and its application to scientific inference*. Cambridge University Press.

Fisher, R. A. (1935), "Statistical tests." 136.

Fisher, Ronald A. (1956), *Statistical Methods and Scientific Inference*. Oliver and Boyd.

Gandenberger, Greg (2016), "Why I am not a likelihoodist." *Philosopher's Imprint*, 16.

Greco, Daniel (2011), "Significance testing in theory and practice." *British Journal for the Philosophy of Science*, 62, 607–637.

Hacking, Ian (1965), *Logic of Statistical Inference*. Cambridge University Press.

Horwich, Paul (1982), *Probability and Evidence*. Cambridge University Press.

Howson, Colin and Peter Urbach (2006), *Scientific Reasoning, The Bayesian Approach*, third edition. Carus Publishing Company.

Jaynes, Edwin T. (1968), "Prior probabilities." *IEEE Transactions on Systems Science and Cybernetics*, 3, 227–241.

Jeffreys, Harold (1961), *Theory of Probability*. Oxford University Press.

Lehmann, E. L. (1993), "The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two?" *Journal of the American Statistical Association*, 88, 1242–1249.

Neyman, J. and E. S. Pearson (1933), "On the problem of the most efficient tests of statistical hypotheses." *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231, 289–337.

Sober, Elliot (2008), *Evidence and Evolution: The logic behind the science*. Cambridge University Press.

van Fraassen, Bas (1989), *Laws and Symmetry*. Clarendon Press.