13

The Problem of the Priors and Alternatives to Bayesianism

A typical high-school statistics course offers a package of familiar techniques: significance tests, regressions, etc. While the high-school student may think these have been around forever (along with the teacher offering the instruction), most of the mathematics presented is less than a century old. Moreover, a student who pursues more advanced statistics at the university level may encounter a recent vogue for "likelihoodist" and "Bayesian" statistical methods. Why, then, this fixity of "frequentist" methods at the lower levels of instruction, and (until very recently) in the publishing practices of almost all scientific journals?

As always with these questions, part of the answer has to do with historical contingencies. But it's also important that for a long time (and in many quarters still), frequentism was thought to be the only statistical regime by which the quantitative assessment of hypotheses by data could be put on a firm, objective footing. Bayesianism, in particular, was thought to depend on subjective prior commitments (such as an agent's hypothetical priors—Section 4.3) that could not be defended in a scientifically responsible way. This was the Problem of the Priors, and it seemed for a long time to justify the dominance of frequentist statistics.[1]

This chapter begins with the Problem of the Priors, usually cited as the most important objection to Bayesian epistemology and its theory of confirmation. We consider where the problem originates in the Bayesian formalism, how best to understand it, and whether it can be overcome by mathematical results about the "washing out" of priors. Special attention is paid to how exactly we're meant to understand the problem's demand for objectivity.

Then we consider frequentist and likelihoodist statistical paradigms that claim to offer accounts of evidential support free of subjective influences. After offering brief introductions to some mathematical tools these paradigms employ, I describe some of the problems that have been pointed out for each by partisans of rival camps. My central question is whether these approaches can yield the kinds of confirmation judgments Bayesians are after without sneaking in subjective influences themselves.

Fundamentals of Bayesian Epistemology 2: Arguments, Challenges, Alternatives. Michael G. Titelbaum, Oxford University Press. © Michael G. Titelbaum 2022. DOI: 10.1093/oso/9780192863140.003.0013

13.1 The Problem of the Priors

Suppose you're a scientist, and you've just run an experiment to test a hypothesis. Let H represent the hypothesis, and E represent the outcome of the experiment. According to Bayesian orthodoxy, your new degree of belief in H at time t_2 (just after you've observed E) should be generated from your old, t_1 degrees of belief by Conditionalization. So we have

$$\mathrm{cr}_2(H) = \mathrm{cr}_1(H \mid E) \tag{13.1}$$

Bayes's Theorem, which can be derived from the probability axioms and Ratio Formula, then gives us

$$\mathrm{cr}_2(H) = \mathrm{cr}_1(H \mid E) = \frac{\mathrm{cr}_1(E \mid H) \cdot \mathrm{cr}_1(H)}{\mathrm{cr}_1(E)} \tag{13.2}$$

What determines the values on the far right of this equation? Perhaps H deductively entails that when an experiment like yours is run, the result will be E (or $\sim E$, as the case may be). In that case your likelihood $\mathrm{cr}_1(E \mid H)$ should be set at 1 (or 0, respectively). Yet the implications of H need not be that strong; well-defined scientific hypotheses often assign intermediate chances to experimental outcomes. In that case, something like the Principal Principle (Section 5.2.1) will require your likelihood to equal the chance that H assigns to E.

So the likelihood $\mathrm{cr}_1(E \mid H)$ may be set by the content of the hypothesis H. But what drives your values for $\mathrm{cr}_1(H)$ and $\mathrm{cr}_1(E)$? These unconditional initial credences are your priors in H and E. $\mathrm{cr}_1(E)$ can actually be eliminated from the equation, by applying the Law of Total Probability to yield

$$\mathrm{cr}_2(H) = \mathrm{cr}_1(H \mid E) = \frac{\mathrm{cr}_1(E \mid H) \cdot \mathrm{cr}_1(H)}{\mathrm{cr}_1(E \mid H) \cdot \mathrm{cr}_1(H) + \mathrm{cr}_1(E \mid \sim H) \cdot \mathrm{cr}_1(\sim H)} \tag{13.3}$$

We now need only the prior $\mathrm{cr}_1(H)$ and the likelihoods $\mathrm{cr}_1(E \mid H)$ and $\mathrm{cr}_1(E \mid \sim H)$. (Your prior for the catchall hypothesis $\sim H$ is just $1 - \mathrm{cr}_1(H)$.) But eliminating $\mathrm{cr}_1(E)$ from the expression and adding the likelihood of the catchall—$\mathrm{cr}_1(E \mid \sim H)$—may not be much of an improvement. If, for instance, our hypothesis is the General Theory of Relativity (GTR), it's not clear what would follow about experimental results from the *negation* of GTR. So it's difficult to determine how confident we should be in a particular experimental outcome E on the supposition of $\sim H$.

In some special cases, we have an idea what to think if H is false because it's one member of a finite partition of hypotheses $\{H_1, H_2, \ldots, H_n\}$. Bayes's Theorem can then be rewritten once more for a given hypothesis H_i to yield

$$\mathrm{cr}_2(H_i) = \mathrm{cr}_1(H_i \mid E) = \frac{\mathrm{cr}_1(E \mid H_i) \cdot \mathrm{cr}_1(H_i)}{\mathrm{cr}_1(E \mid H_1) \cdot \mathrm{cr}_1(H_1) + \mathrm{cr}_1(E \mid H_2) \cdot \mathrm{cr}_1(H_2) + \ldots + \mathrm{cr}_1(E \mid H_n) \cdot \mathrm{cr}_1(H_n)} \tag{13.4}$$

In this case, your posterior credence $\mathrm{cr}_2(H_i)$ after observing the outcome of an experiment can be determined entirely from the likelihoods of the hypotheses—which are hopefully clear from their contents—and those hypotheses' priors.[2]

Whatever mathematical manipulations we make, an agent's priors will still play a role in determining her credences after conditionalizing. And that role may be significant. In Section 4.1.2, on the Base Rate Fallacy, we considered a highly reliable medical test for a particular disease. A positive result from this test multiplies your odds that the patient has the disease by 9. If you start off 50/50 whether the patient has the disease (odds of 1 : 1), the test will leave you 90% confident of illness. But if your prior is only 1 in 1,000 confident of disease, then your posterior will be under 1%.

This ineliminable influence generates the **Problem of the Priors** for Bayesian epistemology: Where are an agent's priors supposed to come from? Obviously an agent's priors are just the credences she has in various hypotheses before the experiment is run. But the challenge is to justify those credences, and the role they play in Bayesian inference.[3]

Of course, an agent's opinions about various hypotheses before she runs an experiment are influenced by other experiments she has run in the past. Presumably those opinions were formed by earlier conditionalizations on earlier experimental outcomes. But the problem recurs when we ask what provided the priors for those earlier conditionalizations. Ultimately the Bayesian recognizes *two* influences on an agent's credences at any given time: the agent's total evidence accumulated up to that time, and her epistemic standards. The agent's epistemic standards capture how she assigns attitudes on the basis of various bodies of total evidence; by definition, they are independent of the evidence itself.

As we saw in Chapter 4, the Bayesian formalism provides a convenient mathematical representation of each of these influences. An agent's total evidence at a given time is a set of propositions to which she assigns credence 1.

Her epistemic standards are represented by a hypothetical prior distribution, which can be conditionalized on her total evidence at a given time to yield her credences at that time. So the question "Whence the priors?" ultimately becomes a question about the origins of hypothetical priors, representing epistemic standards.

An Objective Bayesian (in the normative sense, as defined in Section 5.1.2) responds to the Problem of the Priors by maintaining that only one set of hypothetical priors is rational. All rational agents apply the same epistemic standards, in light of which their evidence uniquely dictates what to believe. The Objective Bayesian then has to work out the details of her view: she has to specify something like a Principle of Indifference (Section 5.3) or Carnapian logical probability approach (Section 6.2) that yields plausible results. If that can be accomplished, the Problem of the Priors is met.

For the Subjective Bayesian, however, the problem looms larger. On this view, more than one set of epistemic standards is rationally permissible. Two rational agents may observe the same outcome of an experiment, and assign wildly different credences to the same hypothesis as a result. They may even disagree about whether the outcome confirmed or disconfirmed that hypothesis.[4] In many cases, this difference of opinion will be traceable to relevant differences in the agents' total evidence before they arrived at the experiment. But according to the Subjective Bayesian, there will also be cases in which the difference results entirely from differences in the agents' epistemic standards. Each agent will be rational in her conclusions, and there will be no difference in either the evidence from this experiment or evidence from previous experiments that accounts for the difference between those conclusions. This is a problem for our epistemology and philosophy of science—isn't it?

13.1.1 Understanding the problem

Let's see if we can be a bit more clear about why the Problem of the Priors is supposed to be a *problem* for Subjective Bayesians. The flat-footed Subjective Bayesian position is just that an agent should base her updates on whatever prior opinions she has about the hypotheses in question (as long as those satisfy requirements of rationality such as the probability axioms), plus whatever new evidence she's updating upon. Why is this position considered insufficient, and what's the worry about letting prior opinions affect inductive inference?

Characterizing the Problem of the Priors, Howson and Urbach write:

> 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. The subjectivity of the premises might suggest that the conclusion of a Bayesian induction is similarly idiosyncratic, subjective and variable, which would conflict with a striking feature of science, namely, its substantially objective character.
>
> (2006, p. 237)

Here Howson and Urbach mention a couple of strands that are nearly ubiqitous in Problem of the Priors discussions. First, the concern is almost always about the role of priors in scientific reasoning. Second, that concern contrasts the subjectivity of priors with a desired objectivity of science. As Reiss and Sprenger (2017) put it, "Objectivity is often considered as an ideal for scientific inquiry, as a good reason for valuing scientific knowledge, and as the basis of the authority of science in society." Yet we've already seen that "subjective" and "objective" admit of diverse philosophical interpretations (Section 5.1). What does it mean in this context for priors to be subjective, and why would it be bad for science to be subjective in that sense?

A third strand in the Howson and Urbach quote is that (Subjective) Bayesianism allows "elusive, idiosyncratic influences" on scientific practice. Elliott Sober pursues this line in his critique of Bayesianism:

> It is important to recognize how important it is for prior probabilities to be grounded in evidence. We often calculate probabilities to resolve our own uncertainty or to persuade others with whom we disagree. It is not good assigning prior probabilities simply by asking that they reflect how certain we feel that this or that proposition is true. Rather, we need to be able to cite reasons for our degrees of belief. Frequency data are not the only source of such reasons, but they are one very important source. The other source is an empirically well-grounded theory. When a geneticist says that Pr(offspring has genotype Aa | mom and dad both have the genotype Aa) = 1/2, this is not just an autobiographical comment. Rather, it is a consequence of Mendelism, and the probability assignment has whatever authority the Mendelian theory has. That authority comes from empirical data. (2008, p. 26)

For Sober it's crucial that scientific conclusions be based on empirical data. But Bayesians allow priors to influence scientific conclusions, and the choice among priors is not entirely determined by empirical evidence. (Bayesian priors combine both evidence and epistemic standards, and the latter are independent of evidence by definition.)

It's important not to take this critique of Subjective Bayesianism too far. Critics of Bayesianism often depict subjective priors as *mere* opinions (Sober's "how certain we *feel* that this or that proposition is true"). But priors—even hypothetical priors—need not be groundless, expressions of nothing beyond the agent's whim. They also need not be a vehicle for political or moral values to interfere with science, threatening the value-neutrality some scientists have sought to secure for their discipline. An agent's prior in a hypothesis may reflect commonly accepted practices of theorizing in her culture, or her discipline; it may reflect a considered trade-off among such purely epistemic values as simplicity and strength. It's just that the priors aren't entirely driven by her *evidence*; instead, they shape how she responds to that evidence.

Why reject extra-evidential influences on scientific reasoning? One concern is that such influences will not be shared among practitioners, and so will lead to undesirable variation in opinion. Characterizing his opponents' views,[5] Savage writes:

> It is often argued by holders of necessary and objectivistic views alike that that ill-defined activity known as science or scientific method consists largely, if not exclusively, in finding out what is probably true, by criteria *on which all reasonable men agree*. The theory of probability relevant to science, they therefore argue, ought to be a codification of *universally acceptable criteria*. Holders of necessary views say that, just as there is no room for dispute as to whether one propopsition is logically implied by others, there can be no dispute as to the extent to which one proposition is partially implied by others that are thought of as evidence bearing on it, for the exponents of necessary views regard probability as a generalization of implication.
>
> (1954, p. 67, emphases added)

Similarly, E.T. Jaynes declares:

> The most elementary requirement of consistency demands that two persons with the same relevant prior information should assign the same prior probability.… The theory of personalistic probability has come under severe criticism from orthodox statisticians who have seen in it an attempt to destroy

> the "objectivity" of statistical inference by injecting the user's personal opinions into it. (1968, Sect. I)

There is certainly a strong current of opinion in modern science that evidence should be (at least in principle) shareable. But epistemic standards are presumably shareable as well—they needn't be inscribed in some mysterious, Wittgenstinian private language. Beyond being share*able*, why insist that they be *shared*?

Here we enter deep waters in the philosophy of science. I will simply comment that the absence of dispute is not obviously a hallmark of reasonable or even desirable scientific inquiry. Good scientists disagree about how to draw conclusions from their experiments just as much as they disagree about the conclusions themselves. So it's unclear whether a proper account of science must provide a path for the ultimate reconciliation of all rational differences. There may be pressure to explain what consensus *does* exist, but the Subjective Bayesian can chalk this up to intersubjective agreement in the priors of human agents who share a biological and cultural background. Such commonalities of perspective can be real without being rationally required.

Going in a different direction, one might worry that if an agent's opinions are not properly grounded in something outside herself, she will have insufficient reason to remain committed to those opinions. If nothing objective like empirical evidence ties the agent to a particular hypothetical prior, then when the evidence leads her to an opinion she doesn't like, why not switch to another prior on which that evidence generates a rosier conclusion? (See White 2005.)

Again, we should emphasize that a Subjective Bayesian's adherence to a particular prior need not be based on *nothing at all*. But we should also point out that attitudinal flip-flopping is banned by Conditionalization. When an agent updates by Conditionalization, her final credences are entirely dictated by the combination of her priors and her new evidence. And as we saw in Chapter 4, an agent who always updates by Conditionalization will be representable as having stuck with a single set of hypothetical priors across those updates. If Conditionalization is a rational requirement, then it's rationally impermissible for an agent to jump from one set of epistemic standards to another. And notice that the arguments for Conditionalization in Chapters 9 and 10 (based on Dutch Books and accuracy considerations, respectively) were perfectly consistent with Subjective Bayesianism. Even if multiple priors are rationally acceptable, it will still be the case that an agent with a particular prior who violates Conditionalization will face an expected accuracy cost or

a diachronic Dutch Book. So Bayesians can argue against switching among hypothetical priors without arguing that some such priors are objectively better than others.[6]

Finally, there's the concern that if objective evidence doesn't fully constrain an agent's opinions, she'll be free to believe any old crazy thing she wants. Just as moralists fear the rational defector, Bayesians have long been haunted by the rational counterinductivist. Suppes writes:

> Given certain prior information is one a priori distribution as reasonable as any other? As far as I can see, there is nothing in my or Savage's axioms which prevents an affirmative answer to this question. Yet if a man bought grapes at [a certain] store on fifteen previous occasions and had always got green or ripe, but never rotten grapes, and if he had no other information prior to sampling the grapes I for one would regard as unreasonable an a priori distribution which assigned a probability 2/3 to the rotten state.
>
> (1955, p. 72)

In the same vein, D.H. Mellor suggests that we:

> Take Othello, whom Iago makes so jealously suspicious of Desdemona that conditionalizing on whatever she then says only strengthens his suspicion: reactions whose Bayesian rationality makes him no less mad. If this does not make Bayesianism false, it does at least make it seriously incomplete as an epistemology. (2013, p. 549)

Yet Subjective Bayesianism need not be anything-goes. Just because the Subjective Bayesian thinks rational constraints are insufficient to single out a unique hypothetical prior, this needn't prevent her from developing strong requirements on rational credence. Suppes and Mellor are right that a Bayesianism constrained only by the probability axioms would deliver severely restricted epistemological verdicts. But we saw in Chapter 5 that all sorts of further constraints are available and defensible for Bayesians. Agents may be required to respect frequencies, chances, their own future opinions, or even the opinions of others. So the Subjective Bayesian is as free to deride counterinductive inferences as anyone else.

We have now worked our way through a number of possible concerns about Subjective Bayesianism—and responses to those concerns available to Bayesians—without settling on a unitary Problem of the Priors. I think that's

representative of the literature; rather than a single, well-defined problem, Subjective Bayesians face a cluster of interrelated issues felt to be problematic.[7] Yet despite this vagueness, "the Problem of the Priors" is consistently cited as a crucial challenge to Bayesian epistemology, and a key motivation for endorsing other statistical approaches. Later in this chapter, we'll consider approaches to evidential support that seek to secure objectivity by eliminating the influence of priors. First, however, we'll investigate the claim that Subjective Bayesians may recover the desired objectivity through the application of formal convergence results.[8]

13.1.2 Washing out of priors

We'll begin our discussion of Bayesian convergence results with a simple example. Suppose two investigators, Fiona and Grace, have been given an urn containing ten balls. Each ball is either black or white, and the investigators know there is at least one of each color in the urn. Fiona and Grace then have nine hypotheses to consider: the urn contains one black ball; the urn contains two black balls; etc. on up to nine black balls. Fiona and Grace do not have any further evidence about the urn's composition, but each of them has a prior biasing her toward an opinion about the contents of the urn. The left-hand graph in Figure 13.1 depicts Fiona's and Grace's prior distributions over the urn hypotheses; Fiona's with solid lines and Grace's with dashed. Fiona expects more of the balls to be white than black, while Grace expects the opposite.

Now some evidence is gathered: one at a time, a ball is drawn at random, shown to the investigators, then returned to the urn. Let's say that out of the first four balls drawn, two are black and two are white. Fiona and Grace update by conditionalization, generating the new distributions shown on the right-hand side of Figure 13.1.

Notice that the posterior distributions have begun to converge; Fiona's and Grace's opinions are closer together than they were before the evidence came in. Why is that? If an agent updates by conditionalization on some evidence E between two times t_i and t_j, her credences in hypotheses H_1 and H_2 will evolve as follows:

$$\frac{\mathrm{cr}_j(H_1)}{\mathrm{cr}_j(H_2)} = \frac{\mathrm{cr}_i(H_1)}{\mathrm{cr}_i(H_2)} \cdot \frac{\mathrm{cr}_i(E \mid H_1)}{\mathrm{cr}_i(E \mid H_2)} \tag{13.5}$$

The final fraction in this equation—the ratio of the likelihood of the evidence on H_1 to the likelihood of the evidence on H_2—is known as the **likelihood**

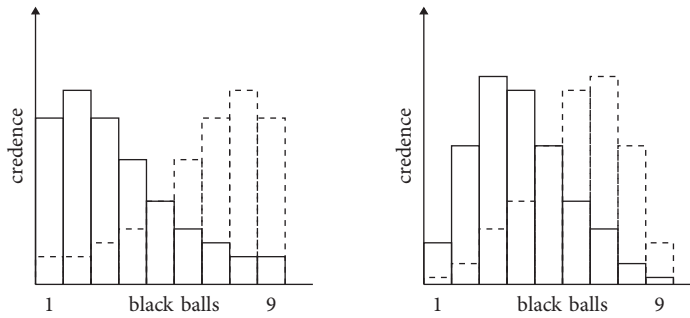

Figure 13.1 Convergence in response to evidence

ratio for H_1 and H_2.[9] When H_1 assigns E a higher likelihood than H_2 does, this ratio will be greater than 1, and the ratio of the agent's posteriors in H_1 and H_2 will be greater than the ratio of her priors in those hypotheses. So relatively speaking, the agent will become more confident in H_1 and less confident in H_2.

More generally, when a probabilistic agent updates by conditionalization, her credence shifts toward hypotheses that assign higher likelihoods to the evidence that was observed. In our urn example, the more equally a hypothesis takes the urn to be split between black and white balls, the higher a chance it will assign to randomly drawing two black balls and two white. So if Fiona and Grace satisfy the Principal Principle (or some other principle of direct inference), they will assign higher t_i likelihoods to a 2-2 observation on hypotheses that posit a fairly equal split. Once that observation is made, each investigator will shift some of her credence towards such hypotheses, which in turn will push their distributions closer together.

What happens if the observations continue? Let's suppose that, unbeknownst to our investigators, the urn is actually split 5-5 between black and white balls. The law of large numbers (Section 7.1) assures us that with probability 1, the ratio of black to white balls observed as the number of draws approaches the limit will approach 1 : 1. As this occurs, Fiona's and Grace's distributions will assign more and more credence to equal-split hypotheses and less and less credence to unequal splits. As a result, their distributions will move arbitrarily close to one another. In his (1954), Savage proved that with probability 1 (in a sense we'll clarify shortly), as the number of draws approaches infinity, the difference between the investigators' credences in any given urn hypothesis will approach 0.

It's important to understand why this convergence occurs. The underlying cause is that with probability 1, as the number of draws approaches infinity,

each investigator's credence in the *true* urn hypothesis will approach 1. Since both investigators' opinions approach the truth (and the truth is the same for both of them), their opinions approach each other's as well.

This is the simplest of many formal convergence results that have been proven since Savage's work. The general thrust of all of them is that conditionalizing on more and more evidence will eventually **wash out** the influence of agents' priors. As the evidence piles up, its effect overwhelms any differences in the priors, and posterior opinions grow arbitrarily close together. If priors are a pernicious, subjective influence, the objectivity of evidence will cleanse scientific opinion of that influence if we just wait long enough.

Later results generalized Savage's work in a variety of ways. While we gave our investigators regular distributions over the available hypotheses, Regularity is not strictly required. Convergence results obtain as long as the investigators initially rule out the same possibilities by assigning them credence 0 (and of course don't assign credence 0 to the hypothesis that is true)[10]. Generalized results also tend to work with distributions over a continuous parameter, as opposed to the finite partition of hypotheses in our simple example. And the evidence-gathering process need not be so highly regimented as drawing balls with replacement from an urn. Gaifman and Snir (1982) proved a highly general convergence result which assumes only that the evidence and hypotheses are stated in a standard propositional language.

While Bayesian convergence results are technically impressive, their philosophical significance has come in for a lot of abuse over the years. I will present some of the criticisms in just a moment. But first I want to note that these results do establish something important. Observing the wide variety of priors permitted by the Subjective Bayesian approach, one might have worried that some priors could trap agents in a rut—by contingency and sheer bad luck, some agents might receive priors that forever doom them to high credences in wildly false hypotheses. (And then the skeptic asks how *you* know *you* aren't one of these doxastically tragic figures....) Yet the convergence theorems guarantee what Earman calls a "long-run match between opinion and reality, at least ... for observational hypotheses" (1992, p. 148). And this guarantee makes minimal demands (the probability axioms, perhaps the Principal Principle) on the priors with which investigators begin.

Of course, we need to be careful what sort of guarantee we've been given. Both Savage's result and the law of large numbers provide for convergence "with probability 1". What sort of probability do we mean here; is it something like objective chance, or is it subjective credence? For the simple sampling situations Savage considered—like our urn example—either interpretation

of "probability" will suffice. Applying the law of large numbers to chance values shows that in the limit there's a chance of 1 that the ratio of black to white balls in the sample will approach the true ratio in the urn (and therefore that conditionalizing agents will converge on the true hypothesis). A similar conclusion can also be drawn from within the scientists' point of view. Assuming she meets the minimal requirements for the theorem, each scientist will assign credence 1 that as the number of samples approaches the limit, her opinions will ever more closely match both the truth and the opinions of her peer.[11] Savage writes, "To summarize informally, it has now been shown that, with the observation of an abundance of relevant data, the person is almost certain to become highly convinced of the truth, and it has also been shown that he himself knows this to be the case" (1954, p. 50).

Since we're dealing with an infinitistic case here (number of observations approaching the limit), "probability 1" does not mean "necessarily must occur". As we saw in Section 5.4, when it comes to the infinite, probability-0 outcomes may be possible, and probability-1 outcomes may fail to occur. But many of us will be content to have probability 1 of reaching the truth even without an *absolute* guarantee. What's more problematic is that once we get past Savage and into the more advanced convergence results, the "probability 1" guarantee admits of only the subjective interpretation. It can no longer be proven to be a fact out in the world that the agent will converge on the truth with objective chance 1; we can only prove that the agent should have confidence 1 that she will approach the truth. It's unclear how reassuring we should find such predictive self-satisfaction, especially in the face of skeptical challenges.

One might also worry that many of the advanced results rely on Countable Additivity, to which we saw objections in Section 5.4. But the biggest problems for Bayesian convergence results center around their being results in the limit. The theorems demonstrate that in the long run, accumulating evidence will eventually overcome our priors and send us to the truth. But as Keynes (1923, p. 80) famously quipped, "In the long run we are all dead." (Woody Allen concurs that "eternity is very long, especially towards the end.")[12] Some of the simpler convergence results offer hopeful tidings for the shorter term: In straightforward sampling situations, we can assign a precise probability that opinions will converge to a certain degree within a certain number of samples. But for wider-ranging results, no such bound may be available. In that case we know only that objective influences will overtake our subjective priors if we hold out long enough.

While we wait, science may appear a subjective, conflicted mess. Given any body of total evidence, any hypothesis neither entailed nor refuted by

that evidence, and any nonextreme posterior value, one can write down a probabilistic prior assigning that hypothesis that posterior in light of that total evidence. So no matter how much evidence we collect, it will be possible to find (or at least imagine) a probabilistic agent who disagrees with us by any arbitrary amount in light of that evidence. As Savage puts it, "It is typically true of any observational program, however extensive but prescribed in advance, that there exist pairs of opinions, neither of which can be called extreme in any precisely defined sense, but which cannot be expected, either by their holders or any other person, to be brought into close agreement after the observational program" (1954, p. 68).

In practice such disagreement may be cut down by Subjective Bayesians' imposing stronger constraints on hypothetical priors than just the probability axioms. But the basic issue is that no matter how much evidence we gain, until we pass to the limit our hypothetical priors' influence will never *entirely* disappear. Subjective Bayesians must concede that as long as our evidence is finite, the total objectivity and perfect scientific consensus sought by those who press the Problem of the Priors will never be achieved. de Finetti concludes:

> When the subjectivistic point of view is adopted, the problem of induction receives an answer which is naturally subjective but in itself perfectly logical, while on the other hand, when one pretends to *eliminate* the subjective factors one succeeds only in *hiding* them (that is, at least, in my opinion), more or less skillfully, but never avoiding a gap in logic. It is true that in many cases . . . these subjective factors never have too pronounced an influence, provided that the experience be rich enough; this circumstance is very important, for it explains how in certain conditions more or less close agreement between the predictions of different individuals is produced, but it also shows that discordant opinions are always legitimate. (1937/1964, p. 147)

13.2 Frequentism

We will now discuss two statistical approaches to data analysis distinct from Bayesianism: frequentism and likelihoodism. Frequentists and likelihoodists often contrast their approaches with Bayesianism by noting that their methods do not invoke priors. Instead, they assess the significance of evidence by making various calculations involving likelihoods. As we saw earlier, likelihoods can often be established by objective, non-controversial methods: likelihoods for diagnostic tests can be obtained by trying them on individuals

whose condition is known; likelihoods associated with well-defined scientific hypotheses can be read off of their contents; etc. So these approaches claim to secure a kind of objectivity for scientific inference unavailable to Bayesians.

Assessing such claims is made difficult by the fact that each camp offers a variety of statistical tools, designed to do different things. Here it's helpful to follow Richard Royall (1997, p. 4) in distinguishing three questions one might ask about a particular observation:

1. What do I believe, now that I have this observation?
2. What should I do, now that I have this observation?
3. What does this observation tell me about *A* versus *B*? (How should I interpret this observation as evidence regarding *A* versus *B*?)

Distinguishing these questions helps us clarify the aims of various approaches. For instance, the frequentists Neyman and Pearson explicitly motivate their techniques with prudential considerations of what to *do* after receiving some data. On the other hand Royall (a likelihoodist) sets out to address the contrastive question of what the evidence says about *A* versus *B*.

We've seen over the course of this book that Bayesianism offers tools for answering all three questions. What an agent should believe after making an observation is her posterior credence upon conditionalization. What to do is addressed by decision theory, as described in Chapter 7. And whether the observation tells more strongly in favor of hypothesis *A* or *B* is revealed by the degree to which it confirms each (Chapter 6).

But addressing all three of Royall's questions doesn't automatically make Bayesianism the victor among statistical schools. Each approach has some tasks it accomplishes better than the others. For example, while frequentism and Bayesianism each offer tools for interpreting data after it's been collected, frequentism historically has had much more to say about the proper design of experiments to generate that data.

To describe every purpose one might try to achieve with statistical analysis, then assess the strengths and weaknesses of all of an approach's offerings with respect to each of those purposes, would be far too vast a project for this book.[13] Instead, for each broad approach I will focus on one or two of the tools it offers. For each tool, I will ask what it reveals about relations of evidential support. I do this for two reasons: First, this book is primarily about epistemology, and questions of evidential support (or justification) are central to epistemology. Second, frequentists and likelihoodists attack Bayesian confirmation theory on the grounds that evidential support is an important

concept in science and so must be suitably objective. In light of this attack, it's fair to ask whether the methods promoted by frequentists and likelihoodists offer promising accounts of evidential support to take Bayesianism's place.

13.2.1 Significance testing

Frequentist methods are the methods taught in traditional statistics classes—significance tests, regression analyses, confidence intervals, and the like. If you've taken such a class, you'll know that frequentism is not a single, unified mathematical theory; instead, it's a grab-bag of tools for the analysis of data. Some of these tools address different questions from each other (say, finding a correlation coefficient versus significance testing), while other tools offer multiple ways of accomplishing roughly the same task (for instance, the variety of significance tests).

What do these tools have in common, such that they embody a common statistical approach? The term "frequentism" provides a clue. While many frequentists have historically adopted a frequency interpretation of "probability" (see Section 5.1.1), that isn't really what the "frequentist" moniker means. Also—contrary to what one sometimes hears—frequentism isn't really about basing one's opinions exclusively on observed frequency data. Instead, the hallmark of **frequentism** is to assess a given inference tool by asking, were it repeatedly applied, how frequently it would be expected to yield verdicts with a particular desirable (or undesirable) feature.

To illustrate, I'll focus on a particular frequentist statistical tool: the *p*-value significance test, most commonly associated with the work of R.A. Fisher. This type of test is applied when we have collected a body of data with a particular attribute. The significance test helps us determine whether the data's displaying that attribute was merely the result of chance, or whether it instead indicates something important about the underlying process that generated the data.

Let's take a specific example: We perform IQ tests on each of the sixteen members of Ms. B's second-grade class. We find that the average (mean) IQ score in the class is 110. This is interesting, because IQ tests have a mean in the general population of 100. We start to wonder: Is there some underlying explanation why Ms. B's class has a high average IQ score? Perhaps Ms. B is doing something that improves students' scores, or perhaps students were assigned to Ms. B's class because they possess certain traits that are linked to high IQs. On the other hand, the high average may have no explanation beyond the vagaries of chance. Even if a teacher's students are selected randomly from

the general population, and nothing that teacher does affects their IQs, the luck of the draw will sometimes give that teacher a class with a higher mean IQ than average.

This "luck of the draw" hypothesis about the high IQ scores will be our **null hypothesis**. As a matter of definition, the null hypothesis is whatever hypothesis one sets out to assess with a significance test. In practice, though, the null is usually a hypothesis indicating "that the probabilistic nature of the process itself is sufficient to account for the results of any sample of the process. In other words, nothing other than the variation that occurs by chance need be invoked to explain the results of any given trial, or the variation in results from one trial to another" (Dickson and Baird 2011, p. 212).

How might we test the null hypothesis in this case? Here it helps that IQ scores are calculated so as to have a very specific statistical profile. In the general population, IQ scores have a normal distribution, with a mean of 100 and a standard deviation of 15. I won't explain what all that means—you can find an explanation in any traditional statistics text—but suffice it to say that this statistical profile allows us to calculate very precisely the probability of various outcomes. In this case, we can calculate that under the null hypothesis, the probability is less than 0.4% that the sixteen members of Ms. B's class would have an average IQ score at least as high as 110.[14]

This statistic is called the ***p*-value** of the null hypothesis on our data. The *p*-value records how probable it would be, were the null hypothesis true, that a sample would yield at least as extreme an outcome as the one that's observed. In our case, we add up the probability on the null hypothesis that the average IQ in Ms. B's class would be 110, the probability that it would be 111, that it would be 112, and so on for every IQ score at least as great as 110. The sum of these probabilities is the *p*-value of the null. The *p*-value in our example is quite low; if we drew a class of sixteen students at random from the general population, we should expect to get an average IQ score at least as high as that of Ms. B's class less than 0.4% of the time.[15]

Once we've calculated our *p*-value, we apply the significance test. Before examining the data, a statistician will usually have picked a numerical threshold α—typically 5% or 1%—to serve as the **significance level** for the test. If the *p*-value for the null hypothesis determined from the data is less than α, the result is deemed significant and we reject the null. In the case of Ms. B's class, a statistician working with a significance level of 1% will notice that the *p*-value associated with our sample is less than that. So the statistician will recommend "rejecting the null hypothesis at the 1% significance level."

Notice that the calculations leading to this recommendation worked exclusively with the data and with likelihoods relative to the null hypothesis. In our case, the null hypothesis says that the IQ scores in Ms. B's class are the chance result of a normal distribution with mean 100 and standard deviation 15. That hypothesis yields a particular likelihood that a class of sixteen will have an average score of 110, a likelihood that such a class will average 111, a likelihood of 112, etc. The *p*-value is the sum of these likelihoods. And all of them are determined directly from the content of the hypothesis being tested; no priors or catchalls are required.

Moreover, these calculations tell us something important about the frequency profile of the test. Suppose we make it our general policy to reject the null hypothesis if and only if our data exhibits a *p*-value for that hypothesis less than 1%. Focusing on cases in which the null hypothesis is true, in what percentage of those cases should we expect this policy to make the mistake of rejecting the null? Answer: less than 1% of the time. Call any body of data that leads us to reject the null hypothesis under this policy "rejection data". If the null hypothesis is true, the probability is less than 1% that a given sample will yield rejection data. So given the law of large numbers, we should expect in the long run to reject the null hypothesis in less than 1% of cases where the null hypothesis is true.

A frequentist statistical test is one for which we can use uncontroversial probabilities to generate an expectation for how frequently the test will display a particular desirable or undesirable feature. The previous paragraph performs that analysis for a Fisherian significance test. The probabilities in question are likelihoods derived from a well-defined null hypothesis. The (undesirable) feature in question is rejecting the null hypothesis when it's true.

13.2.2 Troubles with significance testing

Perhaps the biggest real-world problem with significance tests is how often their results are misinterpreted. Misinterpretation of *p*-values is so common that in February of 2016 the American Statistical Association (ASA) felt compelled to publish a "Statement on Statistical Significance and P-values" (Wasserstein and Lazar 2016), the first time in its 177-year history the association had taken an official position on a specific matter of statistical practice. The statement listed six "widely agreed upon principles underlying the proper use and interpretation of the p-value":

1. P-values can indicate how incompatible the data are with a specified statistical model.
2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

The misconceptions this list warns against are common not only among scientific practitioners in the field[16] but even in introductory statistics texts.[17] In elaborating the second principle, the ASA warned in particular:

> Researchers often wish to turn a p-value into a statement about the truth of a null hypothesis, or about the probability that random chance produced the observed data. The p-value is neither. It is a statement about data in relation to a specified hypothetical explanation, and is not a statement about the explanation itself. (Wasserstein and Lazar 2016, p. 131)

The point here is that the *p*-value associated with the null hypothesis is not a statement about that hypothesis by itself—it doesn't, for example, establish the unconditional probability of the null hypothesis. Instead, the *p*-value establishes a specific relation between the data and the null hypothesis, which we can roughly characterize with the following conditional probability:

$$p\text{-value} = \Pr(\text{outcome at least as extreme as what was observed} \mid \text{null hypothesis}) \tag{13.6}$$

Perhaps it's unfair to criticize frequentism based on the behavior of those who misuse it. So what's the proper way to understand a significance test's conclusions? A *p*-value doesn't say anything directly about the truth of the null hypothesis, but can we *use* it to adopt some attitude toward the null? When data yields a very low *p*-value for the null, a careful statistician will recommend "rejecting the null hypothesis at the 1% significance level". What does "reject"

mean in this context? Is this a recommendation that we *disbelieve* the null hypothesis, or *believe* the null to be false? Probably not, for two reasons: First, it's unclear what it would mean to believe or disbelieve a proposition "at the 1% significance level".[18] Second, in many statistical paradigms "reject" contrasts with "accept" rather than "believe". Philosophers of science and statisticians have taken pains to clarify that "acceptance" is a technical term in this context, distinct from "belief". We should distinguish rejection from disbelief as well.

So if a low p-value doesn't justify *believing* the null hypothesis is false, what does it tell us? Fisher (1956, p. 39) suggested we conclude from a low p-value that *either the null hypothesis is false, or something very unlikely occurred.* For instance, in the case of Ms. B's third-grade class, if her students' IQs were random samples from the general population, the probability is less than 0.4% that an average IQ score at least as great as 110 would be observed. So if the null hypothesis is true, observing such a score is extremely unlikely. When the observation actually comes to pass, we conclude that either the null hypothesis is false, or something very unlikely occurred.[19]

Given this disjunction, we might apply Cournot's Principle (attributed to the nineteenth-century philosopher and mathematician Antoine Augustin Cournot), which recommends treating sufficiently unlikely outcomes as practical impossibilities. That would license a disjunctive syllogism allowing us for practical purposes to treat the null as false.[20] But this inference looks dubious on its face, as illustrated by an example from Dickson and Baird (2011, pp. 219–20). Consider the hypothesis that John plays soccer. Conditional on this hypothesis, it's unlikely that he plays goalie. (For the sake of the example we can make this as unlikely as we want.) But now suppose we observe John playing goalie. By the logic of significance testing, we know that either John doesn't play soccer or something unlikely has happened. Yet it would be a mistake to conclude that John doesn't play soccer.

A frequentist might reply that while the relevant inference pattern will sometimes yield this kind of mistaken rejection, for true hypotheses that will happen very infrequently. (After all, we started off by stipulating that goalies are rare soccer players.) But this reply misses the underlying issue: unlikely events occur, and occur all the time. We can even find ourselves in a situation in which we're *certain* some unlikely event has occurred. Consider a case in which we start with a partition of hypotheses, then observe some data that receives a low p-value on *every* hypothesis in the partition. Whichever hypothesis is true, something unlikely has occurred. In such a case it would be oddly biased to reject the null because of *its* low p-value.

Or consider a case in which we're comparing the null hypothesis with some alternative, our experimental result receives a low *p*-value on the null, but the alternative hypothesis had a low probability going into the test. Should we reject the null in such a case?[21] Admittedly, the framing of this example brings prior probabilities into the discussion, which the frequentist is keen to avoid. But ignoring such considerations encourages us to commit the Base Rate Fallacy (Section 4.1.2). Suppose a randomly selected member of the population receives a positive test result for a particular disease. This result may receive a very low *p*-value relative to the null hypothesis that the individual lacks the disease. But if the frequency of the disease in the general population is even lower, it would be a mistake to reject the null.

Perhaps when the frequency of a disease in the general population is known, the frequentist can factor that into her testing methodology. But reflecting on the role of priors can lead us to question significance testing in more subtle ways. Many professional scientific journals select a particular threshold (often 5%), then consider an experimental result reliable enough to publish just in case it's significant to that level. Of course reliability of a result is not the only criterion for publication—the hypothesis offered as an alternative to the null must be sufficiently novel, germane to scientific topics of interest, etc. But presumably scientists can determine whether a hypothesis has those sorts of qualities without empirical testing. So let's imagine that a group of scientists, without a strong sense of where the truth lies in their discipline, generates 1,000 hypotheses meeting the criteria of novelty, level of interest, etc. for publication. And let's imagine that, to be generous, roughly 5% of the hypotheses generated are true. There will be approximately 950 cases in which the suggested hypothesis is false and the null hypothesis is true. When the scientists run empirical experiments in these cases, a significance test at the 5% level will reject the null around forty-seven or forty-eight times. On the other hand, in the cases involving the fifty true hypotheses, not every empirical test will yield results significant enough to reject the null. Taking all these cases togther, the scientists will clear the bar for publication on something like ninety to 100 of their hypotheses, and almost half of those hypotheses will be false.

In this example I simply stipulated that about 5% of the scientists' hypotheses are true. In real life there'd be no way of knowing. But it's worth thinking about what percentage of hypotheses scientists put to the test actually are true, and the downstream effects this has on the percent of results published in significance-test journals that actually support the hypothesis claimed. Many commentators suggested that journals' reliance on significance tests was

responsible for psychology's recent "replication crisis". A group of psychologists (Open Science Collaboration 2015) set out to replicate randomly selected results published in three prominent psychology journals. Of the ninety-seven results with p-values under 5% they attempted to replicate, they obtained only thirty-five outcomes significant at that α level.[22]

To be clear: The mathematics of p-values is entirely uncontroversial. I'm not challenging the relevant math, nor am I challenging the objectivity of the likelihoods underlying frequentist calculations. I'm trying to determine whether these mathematical calculations have any *epistemological* significance. (Recall that frequentists raise a similar question about Bayes's Theorem, which they're happy to concede is a theorem.) Perhaps it's a mistake to use significance tests for hypothesis rejection (in *any* sense). Could we find a more conservative—yet still epistemologically important—lesson to draw from low p-values? Suppose the null hypothesis makes it highly unlikely that something as extreme as a particular observation would occur. When that observation is actually recorded, doesn't this *disconfirm* the null, at least to some extent?

The prospects for this confirmational interpretation of significance testing dim when we think back on the case in which every available hypothesis receives a low p-value from the data. In such a case it may be that *none* of the hypotheses is disconfirmed by the data. But there's an even deeper problem with reading low p-values as disconfirmatory: p-values exhibit a worrying sort of language dependence.

Suppose we have a coin of unknown bias, and our null hypothesis says that the coin is fair. We flip the coin twenty times, and observe the following string of outcomes:

HHHTHHHHHTHHHHTHHHHH

This string of twenty flip outcomes is extremely unlikely on the null. But then again, any particular string of twenty Hs and Ts will be extremely unlikely on the null. To run a meaningful significance test, we need to set aside some of our total evidence and describe the data in a coarse-grained fashion. We might select the number of heads as our test statistic, then calculate a p-value on the null hypothesis for the observation "seventeen heads out of twenty flips." The resulting p-value is under one-half of one percent.

Yet number-of-heads isn't the only statistic we might use to summarize flip data. Instead of asking how many heads were observed in a string of twenty tosses, we might ask for the length of the longest run of heads observed. For example, the string of outcomes above has a longest-heads-run of five.

There's no obvious reason to prefer number-of-heads as a test statistic over longest-heads-run. They each carve up the possibilities with the same fineness of grain (in each case the possible values run from zero to twenty), and each is positively correlated with the coin's bias toward heads. But these statistics yield different *p*-values for the null hypothesis. As I mentioned above, number-of-heads yields a *p*-value for this string of outcomes sufficient to reject the null at the 1% significance level. But the probability on the null of getting a longest-heads-run of at least five is almost exactly 25%. So should we count this observation as significant enough to reject the null hypothesis?

It's easy to cook up examples that work in the other direction, as well: If my flips came out ten heads followed by ten tails, the longest-heads-run *p*-value for this observation would be 0.59%.[23] But using number-of-heads, the null hypothesis would receive a *p*-value over 50%. Set aside questions of rejection for a moment—if *p*-values are our tool for assessing confirmation, should we take this evidence to disconfirm the null hypothesis or not?[24]

It's time to sum up. Despite the purportedly objective character of frequentist methods, we have identified a number of subjective influences on the use of significance tests. First, while a significance test is meant to assess a single null hypothesis in light of experimental data, the epistemological import of such a test can depend heavily on features of the alternative hypotheses with which the scientist contrasts the null.[25] Second, significance tests can operate only when the data is summarized by some test statistic. Yet as we saw above, the very same data can yield wildly divergent *p*-values for the null depending on which statistic the experimenter uses to summarize her data.[26] Finally, while we haven't delved into this issue here, the choice of a significance level α (5%? 1%?) sufficient for rejection seems to have little objective basis.[27]

Fisherian significance testing is hardly the only tool in the frequentist's toolbox. Other frequentist tools, such as Neyman-Pearson statistics, may fare better on the particular examples I've mentioned, but ultimately admit subjective influences for very similar reasons. (The details are spelled out at great length in this chapter's Further Readings.) Moreover, frequentists' abstention from priors leaves them open to the Base Rate problem I discussed above, as well as a bigger-picture problem I'll discuss at the end of Section 13.3.1. One is reminded of the de Finetti remark I quoted earlier:

> When one pretends to *eliminate* the subjective factors one succeeds only in *hiding* them (that is, at least, in my opinion), more or less skillfully, but never avoiding a gap in logic. (1937/1964, p. 147)

13.3 Likelihoodism

Earlier, in Equation (13.5), we saw that when an agent conditionalizes on some evidence E, her relative credences in hypotheses H_1 and H_2 change as follows:

$$\frac{\text{cr}_j(H_1)}{\text{cr}_j(H_2)} = \frac{\text{cr}_i(H_1)}{\text{cr}_i(H_2)} \cdot \frac{\text{cr}_i(E \mid H_1)}{\text{cr}_i(E \mid H_2)} \tag{13.7}$$

The left-most ratio in this equation captues how the agent compares H_1 to H_2 after the update. The middle ratio expresses how the agent made that comparison before the evidence came in. To see how the comparison has changed, we focus on the right-most ratio—the likelihood ratio. The likelihood ratio expresses the evidence's effect on the agent's relative credences in the hypotheses. If the likelihood ratio is greater than 1, the agent's relative confidence in H_1 compared to H_2 increases; less than 1, H_1 fares worse with respect to H_2; and a likelihood ratio of exactly 1 yields no change.

Likelihoodists take this equation *very* seriously: according to them, the likelihood ratio is our best tool for understanding what a piece of evidence says about one hypothesis versus another. They endorse Hacking's (1965)[28]

Law of Likelihood: Evidence E favors hypothesis H_1 over H_2 just in case $\Pr(E \mid H_1) > \Pr(E \mid H_2)$. In that case, the degree to which E favors H_1 over H_2 is measured by the likelihood ratio $\frac{\Pr(E \mid H_1)}{\Pr(E \mid H_2)}$.

The Law of Likelihood has two components: First, it makes the *qualitative* claim that E favors H_1 over H_2 just in case the likelihood ratio is greater than 1; second, it makes a *quantitative* claim that the likelihood ratio measures the degree of favoring.[29]

Notice that while Equation (13.7) worked with credence values (cr), in stating the Law of Likelihood I switched to generic probability notation (Pr). That's because likelihoodists think the likelihood values that drive personal changes in relative confidence should ultimately be grounded objectively. And it's objectively grounded likelihoods that are supposed to figure in the Law of Likelihood. Where do these likelihoods come from? As we've discussed, probabilities of experimental outcomes can be read directly off of well-defined scientific hypotheses; similarly, diagnostic test likelihoods may be based on frequency data from past trials. Likelihoodists stress that in such situations, likelihoods will be independent of priors; the reliability profile of a diagnostic test doesn't depend on how frequently the disease it tests for appears in a

population.[30] So all the values in the Law of Likelihood can be established on a purely objective basis; likelihoodists think this gives their approach to evidential support a strong advantage over the subjectivity of Bayesianism.

Likelihoodists also think they succeed better than frequentists at achieving objectivity. Recall that in order to apply a significance test, the frequentist must summarize her observations with a test statistic. Different test statistics may yield different verdicts (about rejection of the null, etc.), so the choice of test statistic seems to introduce a subjective element into significance testing. The Law of Likelihood, on the other hand, works with the full description of an observation, honoring the Principle of Total Evidence and eschewing summary statistics.

This difference between likelihoodism and frequentism springs from a deeper, underlying difference: When we apply a significance test to a null hypothesis based on some particular observation, we ask how likely the null would make that observation *or any other observation more extreme*. The calculated p-value thus depends on what other observations might have occurred. This is the point at which different methods of partitioning the possible observations make a difference. The Law of Likelihood, however, operates only on the two hypotheses under comparison and the observation that was actually made. So likelihoodism has no troubles with language dependence. Moreover, interpreting the Law of Likelihood's favoring results does not require us to ask what further hypotheses beyond H_1 and H_2 might have been under consideration. Contrast this with significance tests on a hypothesis; we sometimes felt we had to know how other hypotheses might have fared before we could interpret the meaning of a significant p-value.

Richard Royall neatly summed up this case for likelihoodism in his (1997), when he wrote:

> Fortunately, we are not forced to choose either of these two evils, the sample-space dependence of the frequentists or the prior distributions of the Bayesians. Likelihood methods avoid both sources of subjectivity.
>
> (p. 171)

Bayesians and frequentists retort that whatever the Law of Likelihood's objectivist *bona fides*, it fails at a fundamental level: it gets basic cases of evidential support wrong. To borrow one of Royall's own examples (1997, Sect. 1.7.1), suppose you draw an ace of diamonds from a shuffled deck of cards. Consider the hypothesis that this is an ordinary fifty-two-card deck, versus the hypothesis that the deck is all aces of diamonds. The ordinary-deck hypothesis

confers a probability of 1/52 on drawing an ace of diamonds, while the trick-deck hypothesis gives a likelihood of 1. So by the Law of Likelihood, drawing an ace of diamonds favors the trick-deck hypothesis over an ordinary deck. This has struck many commentators as problematic; it seems like no matter what card we draw from an ordinary deck, that card will count against the hypothesis that the deck is ordinary.

To understand the likelihoodist response to this case, we need to clarify a couple of features of likelihoodism. Like Bayesians and frequentists, likelihoodists offer an account of evidential support. But there's a key difference between their approaches. Bayesians and frequentists are happy to analyze two-place relations between a body of evidence and a single hypothesis—whether the evidence suffices to reject the hypothesis, whether the evidence confirms the hypothesis, etc. Likelihoodism restricts itself to *three*-place relations, between a body of evidence and *two* hypotheses.[31] Likelihoodists emphasize that their view is *contrastive*.

It may feel like drawing an ace of diamonds shouldn't be any evidence against the ordinary-deck hypothesis. But the likelihoodist never said it was, because the Law of Likelihood doesn't say whether a particular observation is evidence for or against a particular hypothesis. The likelihoodist asks whether the observation offers stronger support for the ordinary-deck hypothesis *than some other particular hypothesis*. If the two hypotheses you're entertaining are ordinary-deck versus trick-deck-of-diamond-aces, drawing an ace of diamonds does look like it speaks in favor of the latter over the former. Notice, by the way, that if we consider other trick-deck hypotheses (trick-deck-of-diamond-kings? trick-deck-of-spade-deuces?), the Law of Likelihood has an ace of diamonds observation favoring the ordinary-deck hypothesis over those. So the observation favors ordinary-deck over some hypotheses, but not over others.

One still might protest that having drawn the ace of diamonds, we would feel no inclination to *believe* the deck was all aces of diamonds or to abandon an ordinary-deck assumption. But this brings us to the second key feature of likelihoodism: likelihoodism isn't about what you should believe. Royall distinguished three questions we can ask about an observation (Section 13.2) specifically to point out that likelihoodism addresses question three (What does this observation tell me about *A* versus *B*?), *not* question one (What do I believe?). What to believe (or what unconditional credence to assign) is the kind of question a Bayesian answers, by calculating a posterior using Bayes's theorem. When first reading the ace of diamonds example, you probably assumed some kind of background setup that gave the ordinary deck a much

higher prior than the trick-deck-of-diamond-aces. And this probably accounts for your high credence in the ordinary deck even after the observation. But that's separate from the question of which hypothesis the observation favors more strongly when considered alone.[32]

13.3.1 Troubles with likelihoodism

There are, however, more worrisome counterexamples to likelihoodism. Fitelson (2007) asks us to consider a card drawn at random from a deck known to be ordinary. Suppose you are told that the card is a spade. Which of two hypotheses is favored by this evidence: that the card is the ace of spades, or that it's black?

Most of us think the answer is obviously the latter. Yet the likelihood of drawing a spade on the ace-of-spades hypothesis is 1, while the likelihood of spade on black is 1/2. So the Law of Likelihood goes in the opposite direction. Here we've focused on a purely contrastive question, about what the evidence says (not about what you should believe), that likelihoodism gets dead wrong. Moreover, we can back up our intuition about this case with a general principle: the Logicality principle from Section 6.4.1. In that section we offered Logicality as an adequacy condition on measures of (two-place) confirmation, but it can be read as a purely comparative principle: Any observation favors a hypothesis it entails over a hypothesis that it doesn't (and treats equally any two hypotheses that are both entailed). The general idea is that deductive entailment is the strongest kind of evidential support we can get. But likelihoodism says that the card's being a spade favors its being the ace of spades (which it doesn't entail) *over* the card's being black (which it does).

Likelihoodists have some available responses to this counterexample. First, some likelihoodists (such as Chandler 2013 and Gandenberger ms) think hypotheses can be *competing* only if they are mutually exclusive, and so apply the Law of Likelihood only in that case. In the present example, the hypothesis that the card is the ace of spades and the hypothesis that it's black are not mutually exclusive, so the Law of Likelihood yields no verdict (much less a counterintuitive one). Second, a likelihoodist might say that examples invoking Logicality are necessarily special cases. Likelihoodism is driven by a quest for objectivity in scientific data analysis. Deductive entailment relations are clearly objective, so likelihoodists may be happy to concede an intuitive condition on favoring that can be applied entirely using entailment facts. Perhaps the likelihoodist will grant that, when the evidence deductively entails

exactly one of the hypotheses, the favoring relations are as Logicality says. But in the vast majority of cases—where Logicality doesn't settle the matter—we should revert to the Law of Likelihood. (We'll presently discuss an interpretation of likelihoodism as this kind of "fallback" position.)

These responses are best addressed by a counterexample that avoids both of them. We want an example in which the hypotheses are mutually exclusive, neither is entailed by the evidence, yet the Law of Likelihood still gets the favoring wrong. We might create such an example by taking Fitelson's spade case and introducing a bit of statistical noise. (Perhaps you *thought* you heard that the card is a spade, but aren't *entirely* certain . . . etc.) Instead, let's work with a new, cleaner example:

> We're playing the card game Hearts (with an ordinary deck). The goal of Hearts is to score as few points as possible. At the beginning of the game, the player to my right passes me one card, face down. Some cards I might receive give me no risk of scoring points. The two of hearts exposes me to some risk, but that risk is fairly low. So if I'm passed the two of hearts, I'm mildly annoyed. However if I receive a different heart, or the queen of spades, those represent real trouble, and I get really pissed off.
>
> Now suppose you catch a glimpse of the card passed to me, and see only that it's a heart. Does this evidence favor the hypothesis that I'm mildly annoyed or that I'm pissed off? (Assume these hypotheses are mutually exclusive.)

In this example the evidence that the card is a heart entails neither hypothesis, and by stipulation the hypotheses are mutually exclusive. So neither likelihoodist response above applies. Yet I submit that the Law of Likelihood gets this case wrong. The probability of heart-passed given mildly annoyed is 1, while the probability of heart-passed given pissed-off is 12/13. So by likelihoodist lights the former is favored. But intuitively, catching a glimpse of a heart favors the hypothesis that I'm pissed off over the hypothesis that I'm mildly annoyed.[33]

One can respond to both this counterexample and Fitelson's spade case using a different likelihoodist tack. Notice that in both examples, Bayesian confirmation calculations track our intuitions. If we apply the log likelihood ratio to measure confirmation (Section 6.4.1), drawing a spade in Fitelson's example favors black over ace-of-spades, while glimpsing a heart favors pissed-off over mildly annoyed.[34] Elliott Sober suggests we accept the Bayesian's confirmation verdicts—and her methods of reaching them—for these particular examples.

Sober is a likelihoodist because he wants scientific analyses to proceed on purely objective grounds. In these card-playing examples the distribution of cards in the deck is available, and the draws are assumed to be random, so it's easy to calculate any prior, likelihood, or catchall values we might desire. When all these values can be established objectively, Sober is happy to let the Bayesian employ all of them in making a judgment of evidential support. But when objective priors are not available, likelihoodism comes into its own. Sober writes:

> These examples and others like them would be good objections to likelihoodism if likelihoodism were not a fallback position that applies only when Bayesianism does not. The likelihoodist is happy to assign probabilities to hypotheses when the assignment of values to priors and likelihoods can be justified by appeal to empirical information. Likelihoodism emerges as a statistical philosophy distinct from Bayesianism only when this is not possible. The present examples therefore provide no objection to likelihoodism; we just need to recognize that the ordinary words "support" and "favoring" sometimes need to be understood within a Bayesian framework in which it is the probabilities of hypotheses that are under discussion; but sometimes this is not so. Eddington was not able to use his eclipse data to say how probable the [General Theory of Relativity] and Newtonian theory each are. Rather, he was able to ascertain how probable the data are, given each of these hypotheses. *That's* where likelihoodism finds its application.
>
> (2008, p. 37, emphasis in original)

There are three problems with Sober's "likelihoodism is a fallback position" response, one technical and two more methodological. First, it's easy to modify our examples so that objective priors are unavailable. Imagine that in either Fitelson's spade example or the Hearts example a few cards from the deck have been lost, but you have no idea which cards they are (except that in the spades example you know the ace of spades is still there, while in the Hearts example the two of hearts and the queen of spades remain). In that case we don't know the distribution of the deck, and so can't calculate any objective priors. According to Sober's fallback position, we should therefore abandon Bayesian confirmation measures and judge favoring relations using the Law of Likelihood. But I submit that our intuitions about favoring in these lost-card cases are the same as in the original examples, and continue to run counter to the Law of Likelihood.

Second, thinking about Sober's fallback methodology yields even bigger problems. Here's an analogy: There's a very popular international magazine I used to read to learn about goings-on in far-flung reaches of the world. Then one day I read an article in that magazine about a local issue I knew quite well. The article got both the facts and their significance fairly starkly wrong. And so I began to wonder: If the magazine is so unreliable about this particular case I understand, can I really trust what it says about matters I know little about?

Sober agrees with the Bayesian's take on favoring in examples where the priors are objective. In some of those examples (Fitelson's spade, the Hearts example), the Law of Likelihood disagrees with Bayesian confirmation theory on where the favoring falls. So Sober must grant that in those examples, the Law of Likelihood gets favoring wrong. Why, then, should we trust the Law of Likelihood to get favoring relations right in the cases where objective priors are unavailable?

Third, and finally, we should have *learned* something from the problems likelihoodism confronts when objective priors are available. If we're willing to grant that Bayesianism gets those cases right, we should see if there's any clear explanation of why likelihoodism sometimes gets them wrong. And here I think the answer is straightforward: likelihoodism relies on tools that are fundamentally backward.

Positive probabilistic relevance is symmetrical between E and H: conditioning on E increases the probability of H just in case conditioning on H increases the probability of E. So if all we want is a judgment on whether E supports H, it doesn't matter whether we work with $\Pr(E \mid H)$ or $\Pr(H \mid E)$. But when it comes to measuring degree of support, or comparing whether evidence supports one hypothesis over another, the order in which those propositions appear can make all the difference. Ask yourself: When it comes to evidential support, should the question of favoring be settled by asking whether one hypothesis entails the evidence (as occurs in both our counterexamples)? Or is it more relevant whether the evidence entails one of the hypotheses?

Royall justifies the Law of Likelihood by writing, "The hypothesis that assigned the greater probability to the observation did the better job of predicting what actually happened, so it is better supported by that observation" (1997, p. 5). This sounds intuitive at a first pass, but we should attend to the transition across that "so". Royall's third question asks what the *evidence* says about the *hypotheses*; likelihoodism examines what the *hypotheses* say about the *evidence*. It would be nice if you could discern the former by examining the latter. But that's exactly the bit that fails in the spade and Hearts examples.[35]

In the end, likelihoodists and frequentists have a common problem. They demand statistical tools built from quantities with a certain sort of objectivity. Sometimes they admit priors and catchalls, but often only likelihoods are sufficiently objective. So they build tools from the pieces they've allowed themselves. (The Law of Likelihood combines likelihoods in a simple ratio; frequentist methods make more baroque likelihood calculations.) Yet these pieces point in the opposite direction from the kinds of conclusions they want to draw. No matter how you combine them, likelihoods don't suffice to measure evidential favoring. You need priors or catchalls to truly capture evidential support.

13.4 Exercises

Unless otherwise noted, you should assume when completing these exercises that the distributions under discussion satisfy the probability axioms and Ratio Formula. You may also assume that whenever a conditional probability expression occurs or a proposition is conditionalized upon, the needed proposition has nonzero unconditional probability so that conditional probabilities are well defined.

Problem 13.1. What do you think it means when someone says that science is objective? Do you think it's important that science be objective in that sense?

Problem 13.2. Some of the exercises below demonstrate how Bayesianism, frequentism, and likelihoodism evaluate a simple coin-toss scenario. To set you up for those exercies, here are a few basic probability calculations you might want to make first:

(a) Suppose I toss a fair coin six times. What is the chance it comes up heads exactly five times? More generally, for each n from zero through six, what is the chance that a fair coin tossed six times will come up heads exactly n times?

(b) Now suppose instead that the coin is biased toward heads, such that its chance of coming up heads is 3/4 (though results of successive tosses are still independent of each other). For this biased coin, what is the chance that it will come up heads exactly n times in six tosses, for each n zero through six?

Problem 13.3. Suppose you entertain exactly two hypotheses about a particular coin. H_1 says that the coin is fair. H_2 says the coin is biased toward heads: it has a 3/4 chance of coming up heads on any given flip, though outcomes of successive tosses are probabilistically independent. Let's say that at t_i, you assign $\text{cr}_i(H_1) = 0.2$ and $\text{cr}_i(H_2) = 0.8$, and you satisfy the Principal Principle.

(a) Between t_i and t_j, I flip the coin six times, and it comes up heads exactly five times. If you update on this evidence by conditionalizing, what will be your $\text{cr}_j(H_1)$?

(b) Imagine that—unbeknownst to you—this coin is actually fair. Out of all the outcomes I might have generated by flipping the coin six times between t_i and t_j, which outcomes would have increased your credence in the true hypothesis?

(c) Given that the coin is in fact fair, when I set out to flip it six times between t_i and t_j, what was the chance that the result of those flips would increase your credence in the true hypothesis?

(d) Explain what this has to do with the idea of priors' "washing out" from Section 13.1.2.

Problem 13.4. Imagine you've just seen me flip a coin six times; it came up heads exactly five times. You suspect that I might be flipping a coin that's biased toward heads.

(a) Let's take it as our null hypothesis that the coin is fair. If the null hypothesis is true, what's the chance that flipping the coin six times would yield *at least* five heads?

(b) Does seeing five heads on six flips suffice to reject the null hypothesis at a 5% significance level? At a 1% significance level? Explain why or why not for each.

(c) Imagine all six of my flips had come up heads. Would that data recommend rejecting the null at a 5% significance level? How about 1%? Explain why or why not for each.

(d) In order to answer parts (b) and (c), did you need to consider the prior probability of the null hypothesis?

Problem 13.5. Returning to page 462, go through each of the ASA's six widely agreed-upon principles one at a time, and explain why it's true.

Problem 13.6. Consider the hypotheses H_1 and H_2 about a particular coin described in Exercise 13.3.

(a) Suppose that between t_i and t_j I flip the coin six times, and it comes up heads exactly five times. According to the Law of Likelihood, does this evidence favor H_1 over H_2?
(b) What possible outcomes of flipping the coin six times between t_i and t_j would have favored H_1 over H_2, according to the Law of Likelihood?
(c) In order to answer parts (a) and (b), did you need to consider $\mathrm{cr}_i(H_1)$ or $\mathrm{cr}_i(H_2)$?

Problem 13.7. Suppose we have a probability distribution Pr, two mutually exclusive hypotheses H_1 and H_2, and a body of evidence E. Here are two conditions that might obtain for these relata:

(i) Conditional on the disjunction of the hypotheses, E is positively relevant to H_1. That is,

$$\Pr(H_1 \mid E \,\&\, (H_1 \vee H_2)) > \Pr(H_1 \mid H_1 \vee H_2)$$

(ii) According to the Law of Likelihood, E favors H_1 over H_2 on Pr.

Prove that these two conditions are equivalent. That is, the relata satisfy condition (i) just in case they satisfy condition (ii).[36]

Problem 13.8. Considering Bayesianism, frequentism, and likelihoodism, which do you think offers the best tools for assessing relationships of evidential support? Explain why you think so.

13.5 Further reading

Introductions and Overviews

John Earman (1992). *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*. Cambridge, MA: The MIT Press

Chapter 6 presents and evaluates a variety of Bayesian convergence results, admittedly at a high level of mathematical sophistication.

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

Chapter 1 provides a highly accessible introduction to Bayesianism, likelihoodism, frequentism, and challenges faced by each—albeit from a likelihoodist's perspective.

Classic Texts

Leonard J. Savage (1954). *The Foundations of Statistics*. New York: Wiley

Contains Savage's basic washing out of priors result, and a keen assessment of its significance.

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

J. Neyman and Egon Pearson (1967). *Joint Statistical Papers*. Cambridge: Cambridge University Press

Canonical sources in the development of frequentism.

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

A.W.F. Edwards (1972). *Likelihood: An Account of the Statistical Concept of Likelihood and its Application to Scientific Inference*. Cambridge: Cambridge University Press

Richard M. Royall (1997). *Statistical Evidence: A Likelihood Paradigm*. New York: Chapman & Hall/CRC

Canonical books making the case for likelihoodism.

Extended Discussion

James Hawthorne (2014). Inductive Logic. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2014. URL: http://plato.stanford.edu/archives/win2014/entries/logic-inductive/

Presents state-of-the-art Bayesian convergence results, many of which have marked advantages over previous efforts in the area. For instance, some of these results eschew Countable Additivity, and some avoid concerns about the irrelevance of the long-run.

Colin Howson and Peter Urbach (2006). *Scientific Reasoning: The Bayesian Approach*. 3rd edition. Chicago: Open Court

Statistically sophisticated defense of the Bayesian perspective, with extensive, detailed critiques of frequentist methods.

Deborah Mayo (2018). *Statistical Inference as Severe Testing: How to Get beyond the Statistics Wars*. Cambridge: Cambridge University Press

An extended defence of frequentism, which (among many other things) offers responses to the objections raised in this chapter.

Prasanta S. Bandyopadhyay and Malcolm R. Forster (2011). *Philosophy of Statistics*. Vol. 7. Handbook of the Philosophy of Science. Amsterdam: Elsevier

A collection of excellent philosophical essays—many accessible to the beginner—exploring and assessing a wide variety of techniques in statistics.

Notes

1. Efron's (1986) article "Why Isn't Everyone a Bayesian?" begins by pointing out that "Everyone used to be a Bayesian. Laplace wholeheartedly endorsed Bayes's formulation of the inference problem, and most 19th-century scientists followed suit." When Efron reaches, at the end of his article, his "summary of the major reasons why [frequentist] ideas have shouldered Bayesian theory aside in statistical practice," he cites as crucially important that "The high ground of scientific objectivity has been seized by the frequentists."
2. Some experiments test the value of a continuous physical parameter. In that case we have an infinite partition of hypotheses—one for each possible value of the parameter. Assuming we can determine the likelihood of E on each possible setting of the parameter, and we have a prior distribution over the possible parameter settings, a calculation like that of Equation (13.4) can be accomplished with integrals.
3. Notice that switching from Conditionalization to Jeffrey Conditionalization (or the other Bayesian updating rules discussed in Chapter 11) would make no difference here. Priors play a role in determining the posterior credences generated by all of those rules.
4. Looking back at the list of historically significant Bayesian confirmation measures in Section 6.4.1, each one invokes either the agent's prior in the hypothesis or some credence involving the catchall.

5. Like Howson and Urbach, Savage was a Subjective Bayesian. It's a consistent theme throughout this part of the book that objections to Bayesianism are often stated more clearly by the targets of those objections than by the objectors themselves.
6. There's an interesting comparison here to the action theory literature on intentions. Most authors agree that there are situations in which either of two incompatible intentions is equally rationally permissible for an agent. (Consider Buridan's ass, for instance.) Once the agent forms one of those permitted intentions, is there rational pressure for the agent to stick with that choice? The lack of an objective reason for the initial selection needn't entail that that selection may be rationally abandoned later on. (See Bratman 1987 and the literature that followed.)
7. A similar point could be made about the general concern for "objectivity" in science. See Reiss and Sprenger (2017).
8. In Chapter 14 we will discuss another approach to recovering objectivity for Bayesianism: modeling agents' confidence using intervals of values instead of single real numbers.
9. The Bayes factor, which we discussed in Section 4.1.2, is a special case of the likelihood ratio in which the two hypotheses are H and $\sim H$. Equation (13.5) generalizes Equation (4.8) from our discussion there. (Recall that the likelihood ratio also furnished our measure of confirmation l in Section 6.4.1.)
10. Actually, some Bayesian convergence results (Lele 2004) work even when the true hypothesis isn't included among those the investigators consider. In that case, opinion converges on the hypothesis in the agents' partition that is *closest* to the truth, in a well-defined sense. (Thanks to Conor Mayo-Wilson for the reference.)
11. One slight difference between the two interpretations of "probability" here: On the chance interpretation, there is one particular composition of the urn (5 black, 5 white), such that the chance is 1 that both scientists' opinions will converge on the hypothesis describing that composition. On the other hand, there is no particular hypothesis such that either scientist assigns credence 1 that the two of them will converge on *that* hypotheses. The scientists are certain that a truth exists, assign credence 1 that they'll eventually converge on whatever it is, but are uncertain where exactly it lies.
12. Quoted at Bandyopadhyay and Forster (2011, p. xi), which in turn attributes the quotation to Rees (2000, p. 71).
13. Though it is arguably accomplished by another book, Howson and Urbach (2006).
14. For those who understand such things, the z-score of our sample is 8/3.
15. If you're new to significance testing, it might seem awfully odd to assess the probability on the null hypothesis that our sample would yield a result *at least as extreme as* what was observed. Why not assess the probability that the sample would yield *exactly* what was observed? Well, consider that in our IQ example, even though IQ scores are reported as whole numbers, Ms. B's class contains 16 students, so the average IQ observed could have been any fraction with a denominator of 16—not just the integers I mentioned in the text. That means there are a *lot* of exact averages that might have arisen from our sample, which in turn means that none of them has a particularly high probability. Even the probability of getting a 16-student sample with an average IQ of exactly 100—the most probable average on the null given the way IQ scores are calculated—is extremely low. Now generalize that thought to more common statistical

tests, which often have hundreds or thousands of individuals in their samples, and you'll see that the probabilities attached to getting a body of data with the *exact* attribute observed grow so low as to be near-useless. (Never mind situations in which your data estimates the value of a continuous parameter, in which case all the probabilities may be 0!)

You might also wonder why we calculated the *p*-value by summing the probabilities of only IQ averages greater than or equal to 110, instead of the probabilities of all averages at least as unlikely as 110. (Given the IQ distribution in the general population, an average of 90 is just as unlikely as 110; an average of 89 is even less likely than 110, etc.—so why didn't we add the probabilities of those sub-90 averages into our sum?) Because the alternatives to the null hypothesis we proposed (Ms. B increases students' IQ scores, the class was drawn from a high-IQ population, etc.) all would push the observed IQ in the same direction (that is, upwards), we performed what's known as a one-sided test. If we had calculated the *p*-value using not just the probabilities of averages above 110, but also the probabilities of averages below 90, that would be a two-sided test. Differences between one-sided and two-sided tests—and reasons for using one kind rather than the other—are discussed in standard statistics texts.

16. For instance, a study by Carver (1978) found education researchers espousing the following "fantasies" about the *p*-value: (1) p is the probability that the trial result is due to chance; (2) $1 - p$ indicates the replicability of the trial; and (3) p is the probability that the null hypothesis is true.
17. Picking a popular introductory statistics textbook (Moore, McCabe, and Craig 2009) at random, I found all of the following:

    - "The smaller the *P*-value, the stronger the evidence against [the null hypothesis] provided by the data." (p. 377, in the *definition* of *P*-value)
    - "If the *P*-value is less than or equal to [the significance level], you conclude that the alternative hypothesis is true." (p. 380)
    - "This *P*-value tells us that our outcome is extremely rare. We conclude that the null hypothesis must be false." (p. 381)
    - "The spirit of a test of significance is to give a clear statement of the degree of evidence provided by the sample against the null hypothesis. The *P*-value does this." (p. 395)

    This isn't an instance of cherry-picking—Moore, McCabe, and Craig (2009) was literally the first text I examined in search of examples. I certainly don't mean to beat up on them; similar passages appear in many other statistics texts.
18. "Rejecting the null hypothesis at the 1% significance level" does *not* mean assigning the null a credence of 0.01 or less. To conflate these is a version of the second mistake on the ASA's list.
19. For what it's worth, it's not entirely clear how to derive Fisher's disjunction with a probabilistic disjunct from the kind of conditional probability we equated with *p*-value in Equation 13.6. It looks like we're moving from "the conditional probability of this kind of observation given the null hypothesis is low" to "if the null hypothesis is true, then the probability of what was observed is low" en route to "either the null is false or what was observed is improbable." In light of our discussion of conditionals and conditional probabilities in Section 3.3, the first of these steps is highly suspect.

20. Instead of describing the relevant inference as a disjunctive syllogism, some authors call it "probabilistic *modus tollens*". We begin with the conditional "If the null hypothesis is true, then something very unlikely occurred", take as our minor premise that nothing very unlikely occurred, then conclude that the null is false.
21. A wonderful xkcd comic (that I wish I had the rights to reproduce) involves a neutrino detector designed to determine whether the sun has gone nova. When the button on the device is pressed, it rolls two fair six-sided dice. If the dice come up double-sixes, the device buzzes no matter what. On any other dice result, the device buzzes only if the sun has exploded.

    Someone presses the button, and the device buzzes. A frequentist statistician says, "The probability of this result happening by chance is $1/36 = 0.027$. Since $p < 0.05$, I conclude that the sun has exploded." The Bayesian statistician replies, "Bet you $50 it hasn't" (xkcd.com/1132, by Randall Monroe).
22. Besides suggesting journals decrease their reliance on significance tests, some scholars of scientific methodology have advocated scientists' reporting *all* the hypotheses they test, not just the ones for which they obtain significant results. This is one motivation for item #4 on the ASA's list of principles above.

    For more dramatic reactions, see Woolston (2015), about a psychology journal that stopped publishing papers containing *p*-values, and Amrhein, Greenland, and McShane (2019), a commentary with over 800 signatories calling "for the entire concept of statistical significance to be abandoned."
23. The longest-heads-run *p*-values in this section were calculated using a web application by Max Griffin, which in turn developed out of a discussion dated July 24, 2010 on the online forum askamathematician.com. For further examples of statistical redescriptions that flip a significance test's verdicts about a sequence of coin flips, see Fitelson and Osherson (2015).
24. Frequentists will of course have responses to this type of example. For instance, they might complain that longest-heads-run is not what statisticians call a "sufficient" statistic, while number-of-heads is. In that case, we can shift the example to compare two sufficient statistics. (In any particular experimental setup, many sufficient statistics will be available—for one thing, combining a sufficient statistic with any other statistic will automatically yield another sufficient statistic.) Fisherian statisticians may now complain that only one of the statistics used is a "minimal" sufficient statistic, but it's not clear that there's a good justification for confining our attention to those. (Thanks to Conor Mayo-Wilson for extended discussion on this point.)
25. Also, when more advanced statistical methods are applied, it can sometimes make a difference *which* among the rival hypotheses the scientist chooses to designate as the null.
26. As I hinted earlier, significance testing a statistic that captures only some features of one's data violates the Principle of Total Evidence (Section 4.2.1). Beyond the language-dependence problem discussed in the main text, this generates another problem for significance tests having to do with the optional stopping of experiments. It's sometimes suggested that on a frequentist regime, one can interpret the results of an experiment only if one knows whether the experimenters would have kept collecting data had they

received a different result. This seems to inject the experimenters' subjective intentions into the interpretation of objective results. (See, e.g., Berger and Berry 1988).

27. Deborah Mayo (2018) has defended what she sees as a non-subjective approach to selecting significance levels. She also offers thoughtful, epistemologically sensitive responses to many of the other critiques of frequentism I describe in this section. Unfortunately, engaging with her proposals goes well beyond the scope of this book.
28. Hacking is oft-cited by likelihoodists for having articulated the Law of Likelihood. Yet Reiss and Sprenger (2017) trace the origins of likelihoodism back to Alan Turing and I.J. Good's work on cracking the Enigma code during World War II.
29. The Law of Likelihood should not be confused with another likelihoodist commitment, the **likelihood principle**. The Law of Likelihood considers one experimental observation, and explains how it bears on the comparison between two hypotheses. The likelihood principle considers two different observations, and explains when they should be taken to have the same evidential significance. Birnbaum (1962) gave the latter principle its name, and showed how it could be proven from two other, commonly accepted statistical principles. (See also Gandenberger 2015.)
30. The claimed independence between likelihoods and priors isn't supposed to be a *mathematical* independence; likelihoods and priors have well-defined mathematical relationships captured by equations such as Bayes's Theorem. Instead, the idea is that likelihoods can be *established* without consideration of priors, usually because their values are determined by different physical systems than the values of priors. The likelihood profile of a medical diagnostic test is determined by the biological mechanics of the test and the human body; the prior probability that a given subject has the tested-for condition is determined by the broader health situation in the population. The likelihood that a particular experiment will yield a particular outcome if the General Theory of Relativity is true is determined by the content of that theory; who knows what determined the prior probability that General Relativity would be true in the first place.
31. Perhaps all these relations should include yet another relatum—a background corpus—in which case Bayesians and frequentists would go in for three-place relations while likelihoodists would hold out for four. I will suppress any mention of background corpora in what follows.
32. Royall drives the point home by considering a case in which we start out with two decks, one ordinary and one full of aces of diamonds, then a fair coin flip determines which deck a card is drawn from. Now which hypothesis does the ace of diamonds observation incline you to believe? Is the evidential favoring in this case really any different than it was in the original?
33. A bit of history on this example: I originally proposed it to Greg Gandenberger, who posted it to his blog along with a poll (Gandenberger 2014). Seventy-two percent of respondents to the poll agreed with my intuition about favoring in the example, though admittedly the sample was small and perhaps not broadly representative. Jake Chandler responded to the example with a principle he proposed in Chandler (2013): that a piece of evidence favors one hypothesis over another just in case it favors that hypothesis when we first conditionalize on the disjunction of the hypotheses. If we first conditionalize on mildly-annoyed-or-pissed-off, then heart-passed intuitively favors

mildly annoyed over pissed-off, so Chandler's principle has the Law of Likelihood getting the case right. Steven van Enk, however, replied that Chandler's principle cannot be used to support the Law of Likelihood, because unless one has already accepted the Law, assuming the hypotheses' disjunction looks like it "change[s] our background knowledge in a way that is not neutral between the two hypotheses" (2015, p. 116). van Enk's paper also contains the only published discussion of the Hearts example of which I'm aware.

34. It's easy to get the log likelihood ratio measure of confirmation confused with likelihoodists' use of likelihood ratios, so let me distinguish the two approaches. Bayesians use the log likelihood ratio to calculate a numerical answer to a question about two relata: to what degree does this body of evidence support this hypothesis? They do this by comparing $\Pr(E \mid H)$, the probability of the evidence on the hypothesis, to $\Pr(E \mid \sim H)$, its probability on the negation of the hypothesis (the catchall). If a Bayesian wants to know whether E favors H_1 over H_2, she calculates the log likelihood ratio separately for each hypothesis, then checks which numerical result is greater.

    A likelihoodist, on the other hand, answers the comparative favoring question directly by comparing the likelihood of the evidence on one hypothesis to the likelihood of that evidence on the other (that is, by comparing $\Pr(E \mid H_1)$ to $\Pr(E \mid H_2)$). This can yield different results than the Bayesian approach—as we've already seen, comparisons of log likelihood ratios satisfy Logicality while the Law of Likelihood does not. Also, the likelihoodist approach yields no answer to two-place questions about, say, how strongly E supports H_1 straight out.
35. Frequentism is sometimes described as a probabilified Popperian falsificationism. One might say I'm suggesting that likelihoodism is a probabilified hypothetico-deductivism.
36. This equivalence is reported by Chandler (2013), who says it was pointed out to him by Branden Fitelson.