Received: 10 March 2022 | Revised: 21 June 2022 | Accepted: 6 July 2022

DOI: 10.1111/phc3.12867

ARTICLE

WILEY

The Bayesian and Classical Approaches to statistical inference

Matthew Kotzen

Department of Philosophy, UNC Chapel Hill, Chapel Hill, North Carolina, USA

Correspondence
Matthew Kotzen, Department of Philosophy, UNC Chapel Hill, Chapel Hill, NC, USA.
Email: kotzen@email.unc.edu

Abstract

The Bayesian Approach and the Classical Approach are two very different families of approaches to statistical inference. There are many different versions of each view, often with very substantial differences among them. But I will here endeavor to explain the philosophical core of each family of approaches, as well as to identify four main philosophical differences between them.

1 | THE BAYESIAN APPROACH

On the Bayesian Approach to inference, the rational agent begins with a prior credence distribution over a set of "basic" hypotheses, which obeys the axioms of the probability calculus.[1] When the agent collects new evidence that involves learning (exactly) the proposition E, they update their prior credence distribution according to the Bayesian Rule of Conditionalization. According to the Rule of Conditionalization, when the agent learns E, their *new* credence in an arbitrary hypothesis H should be identical to the ratio of their *old* credence in the conjunction $H \wedge E$ to their *old* credence in E alone. This ratio of $p(H \wedge E)$ to $p(E)$ is commonly abbreviated "$p(H|E)$."

One way of thinking about the Rule of Conditionalization is that, when the agent learns (exactly) E, they thereby "narrow down" which world is the actual world to one of the E-worlds, and thus "rule out" all of the $\neg E$-worlds. But since learning that E is true does not distinguish among any of the E-worlds, the agent has no reason to alter the *ratio* of credences that they assign to any two sets of E-worlds. In particular, the agent has no reason to change their credence ratio between the $H \wedge E$-worlds and the set of all E-worlds (or, equivalently, between the $H \wedge E$-worlds and the $\neg H \wedge E$-worlds). So, for example, if the credence that the agent initially assigns to the $H \wedge E$-worlds constitutes 70% of the total credence that they initially assign to all of the E-worlds, then that ratio should remain unchanged when the agent learns that E is true. But since the agent's credence in E *after* learning E is 1, the result is that the agent's new credence in $H \wedge E$, and hence in H, should just be 0.70. In other words: the agent's *new* $p(H)$, after learning E, should be their *old* $\frac{p(H \wedge E)}{p(E)}$ (i.e, their *old* $p(H|E)$), just as required by the Rule of Conditionalization.[2]

On an only slightly different way of thinking about the Rule of Conditionalization, upon learning E, the agent's prior credence distribution is "renormalized" in proportion both to the prior credence $p(H_i)$ of each hypothesis and to the likelihood $p(E|H_i)$ that each hypothesis assigns to E, as follows: For any hypothesis that assigns a likelihood of 0

to E, that hypothesis is assigned a credence of 0 in the agent's posterior credence distribution, regardless of its prior credence; since the hypothesis assigns a likelihood of 0 to an event that actually occurred, the agent permanently rules that hypothesis out from further consideration. If two hypotheses A and B both assign the same likelihood to E (i.e., if $p(E|A) = p(E|B)$), then the ratio between the agent's credences in A and B in their prior credence distribution (i.e, their $\frac{p(A)}{p(B)}$) will be unchanged in the posterior credence distribution. Similarly, if A assigns a likelihood to E that is twice is high as the likelihood that B assigns to E (i.e., if $p(E|A) = 2 * p(E|B)$), then the ratio between the agent's credences in A and B in the posterior credence distribution will be twice as high as it was in their prior credence distribution. And so on. Together with the constraint that the agent's credences must always obey the axioms of the probability calculus, the agent's posterior credence distribution is uniquely determined.[3]

For example, suppose that the agent's credences are defined over (an algebra of) three mutually exclusive and jointly exhaustive hypotheses: H_1, H_2, and H_3. Further suppose that the ratio of the agent's credences in those hypotheses in their prior credence distribution is 1:2:3, so that their $p(H_1) = \frac{1}{6}$, $p(H_2) = \frac{1}{3}$, and $p(H_3) = \frac{1}{2}$. Finally, suppose that the likelihoods that each hypothesis assigns to E are as follows: $p(E|H_1) = 0$, $p(E|H_2) = \frac{1}{2}$, and $p(E|H_3) = \frac{1}{4}$. When the agent learns exactly E, H_1 is thereby ruled out, since it assigned a likelihood of 0 to E; the agent's posterior credence in H_1 is therefore 0. The ratio between the agent's credences in H_2 and H_3 in their prior credence distribution was 2:3. Since the likelihood that H_2 assigns to E is twice as high as the likelihood that H_3 assigns to E ($\frac{1}{2}$ vs. $\frac{1}{4}$), that ratio must increase by a factor of 2, from 2:3 to 4:3, in the agent's posterior credence distribution. Since the agent's credences in H_1, H_2, and H_3 must continue to sum to 1 in the agent's posterior credence distribution, the agent's posterior credence in H_2 is $\frac{4}{7}$, and their posterior credence in H_3 is $\frac{3}{7}$.

2 | THE CLASSICAL APPROACH

Here, I will focus on Fisherian significance tests, though in many cases analogous points apply to, e.g., Neyman-Pearson tests, as well as to other "classical" statistical methods.[4]

The general strategy of a Fisherian significance test is as follows:

1. Choose a hypothesis H_0, referred to as the Null Hypothesis, which you are going to investigate whether your experiment gives you good grounds to *reject*.
2. Figure out the possible outcomes of the experiment, and assign a likelihood to each outcome on the assumption that H_0 is true.
3. Once you obtain the actual outcome, calculate the likelihood (on the assumption of H_0) that that outcome or an outcome at least as unlikely would occur, by summing the likelihoods (on H_0) of each outcome that is at least as unlikely as the actual outcome (including the actual outcome itself).
4. Use this sum (called a p-value) as a guide to the rejection of H_0. In other words, if the p-value $< \alpha$, then your results are statistically significant at level α and you may reject H_0 at that significance level. The lower the value of α is, the "stronger" the rejection of H_0 is.

The intuitive thought here is that when a particular Null Hypothesis entails that it was very unlikely that an outcome at least as "extreme" as the actual outcome would occur, the actual outcome constitutes reason to reject that Null Hypothesis. And, the less likely that an outcome at least as extreme as the actual outcome is, according to the Null Hypothesis (i.e., the lower the p-value of the experiment), the stronger grounds we have to reject that Null Hypothesis. By contrast, if a Null Hypothesis entails that it was fairly likely that an outcome as least as extreme as the actual outcome would occur (i.e., if the p-value of the experiment is relatively high), then we have comparatively weak reason to reject the Null Hypothesis.

"Rejection" of a Null Hypothesis is a central but often misunderstood attitude in the Classical paradigm.[5] Rejection of a Null Hypothesis at the 0.05 level, for example, definitely does *not* correspond to a credence of 0.95 that the Null Hypothesis is false.[6] The standard account here is that rejection of the Null Hypothesis at the 0.05 level corresponds only to being committed to the view that, if the Null Hypothesis is true, the probability is less that 5% that an outcome at least as extreme as the actual outcome would occur. The standard story typically continues with the thought that, if we were to act as though the Null Hypothesis is false in exactly those situations in which $p < 0.05$, we would commit a Type I Error (i.e., the error of rejecting the Null Hypothesis when it is actually true) in less than 5% of the situations in which the Null Hypothesis is true.[7]

Suppose, for instance, that we are trying to assess whether a particular coin is fair—i.e., whether its probability of landing heads is $\frac{1}{2}$. So, we define our Null Hypothesis, H_0, to be the hypothesis that the coin is fair, and we endeavor to determine whether our experiment gives us sufficient grounds to reject that hypothesis at a particular level of α. Suppose that we design our experiment to involve flipping the coin 20 times under particular conditions and recording the total number of heads and tails outcomes. Next, we must think about what the possible outcomes of the experiment are. Since we are recording the total number of heads and tails flips, there are 21 possible outcomes: {0 heads and 20 tails, 1 heads and 19 tails, 2 heads and 18 tails, …, 20 heads and 0 tails}. It is straightforward to calculate the likelihood of each of these outcomes, on the assumption that H_0 is true. For the 0-heads outcome, the likelihood is simply $\left(\frac{1}{2}\right)^{20}$, since the only way for that outcome to occur is for the coin to land tails 20 times in a row; on the assumption of H_0, the probability of each such tails flip is (independently) $\frac{1}{2}$. For the 1-heads outcome, the total likelihood is the likelihood (on the assumption of H_0) of any one particular 1-heads outcome—say, TTTTTTTTTTHTTTTTTTTT—multiplied by the number of (equiprobable) 1-heads outcomes—here, 20, since the one heads flip could occur on any of the 20 flips. More generally, on the assumption of H_0, the likelihood that we will observe some r-heads outcome or other is: $\left(\frac{1}{2}\right)^{20} \frac{20!}{r!(20-r)!}$. The approximate values of this likelihood for all 21 values of r are given in the Table 1 below.

Suppose that we observe an actual outcome of TTTHTHTTTHTTTTTHHTTH, which contains 6 heads and 14 tails. The likelihood, on the assumption of H_0, of a 6-heads outcome is approximately 0.037. Accordingly, we then look to find all of the outcomes that are at least as "extreme" as the actual outcome—i.e., those outcomes which have likelihoods, on the assumption of H_0, of 0.037 or lower. There are 14 such outcomes: $r = 0$, $r = 1$, $r = 2$, $r = 3$, $r = 4$, $r = 5$, $r = 6$, $r = 14$, $r = 15$, $r = 16$, $r = 17$, $r = 18$, $r = 19$, and $r = 20$. (Note again that, since we are looking for outcomes *at least as extreme* as the actual outcome, we count the actual outcome as well.) Summing the likelihoods for these outcomes, we obtain a p-value of (approximately) 0.000 + 0.000 + 0.000 + 0.001 + 0.005 + 0.015 + 0.037 + 0.037

TABLE 1 The approximate likelihoods of r heads in 20 flips, on the assumption that the coin is fair

r	$p(r$ **heads**$\|H_0)$	r	$p(r$ **heads**$\|H_0)$
0	0.000	11	0.160
1	0.000	12	0.120
2	0.000	13	0.074
3	0.001	14	0.037
4	0.005	15	0.015
5	0.015	16	0.005
6	0.037	17	0.001
7	0.074	18	0.000
8	0.120	19	0.000
9	0.160	20	0.000
10	0.176		

+ 0.015 + 0.005 + 0.001 + 0.000 + 0.000 + 0.000 = 0.116. Since the p-value of this experiment is (approximately) 0.116, the Null Hypothesis that the coin is fair can be rejected at all and only the significance levels greater than or equal to 0.116; thus, H_0 could *not* be rejected at the 0.01, 0.05, or 0.10 levels. If the actual outcome had instead contained 4 heads ($r = 4$) (for instance, if it had been TTTHTTTTTHTTTTTTHTTH), then the p-value of the experiment would have been (approximately) 0.000 + 0.000 + 0.000 + 0.001 + 0.005 + 0.005 + 0.001 + 0.000 + 0.000 + 0.000 = 0.012. In that case, H_0 could have been rejected at the 0.05 (or any $\alpha > 0.012$) level, but not at the 0.01 level (since $0.012 > 0.01$).

3 | ACTUAL AND NON-ACTUAL LIKELIHOODS

The first of the main philosophical differences between the Bayesian and the Classical approaches to statistical inference is that Bayesians think that only the likelihoods of the *actual* outcome on various hypotheses matter inferentially, whereas Classicalists think that the likelihoods of various *non-actual* outcomes can matter too, even once the likelihoods of the actual outcome are fixed.

In the Bayesian analysis from §2 above, note that the focus was exclusively on the likelihood of the *actually observed* evidence, E, on the various hypotheses under consideration (H_1, H_2, and H_3). On the Bayesian approach, when the agent learns precisely E, the likelihoods of propositions other than E are irrelevant; all that matters is the likelihood of E itself, on the various hypotheses over which the agent's credence is defined. When a particular hypothesis assigns a likelihood to E of 0, that hypothesis is ruled out and forever relegated to zero-credence status, regardless of what likelihood the hypothesis assigns to *other* propositions. Similarly, insofar as a particular hypothesis assigns a higher likelihood to E than its competitors do, it gets a proportionally larger credence "boost" than do its competitors, regardless of what likelihoods any of the hypotheses under consideration assigns to propositions *other* than E. At no point in this process is any consideration given to the values of the likelihoods that the various hypotheses under consideration assigned to any outcome that *did not actually occur*, or to any proposition that describes the experiment's outcome in weaker terms.[8]

By contrast, on the Classical Approach, the likelihoods of *non-actual* outcomes can matter too, even once the likelihoods of the actual outcome are fixed. For example, consider two cases in which the likelihood of the actual outcome on the supposition of the Null Hypothesis is 0.01. In Case 1, there are no other possible outcomes with such a low likelihood on the supposition of the Null Hypothesis; in this case, the likelihood of each other possible outcome, on the supposition of the Null Hypothesis, is 0.05. In Case 2, suppose that there are 20 other possible outcomes that have the same likelihood, on the supposition of the Null Hypothesis, as the actual outcome does—i.e., 0.01. Thus, although the likelihood of the *actual* outcome, on the supposition of the Null Hypothesis, is identically 0.01 in both cases, the likelihoods of the various *non-actual* outcomes in the two cases differ; in Case 1, the non-actual outcomes all have higher likelihoods than the actual outcome does, whereas in Case 2, there are several non-actual outcomes that have the same likelihood as the actual outcome. As a result of that difference, the p-values that we calculate in Case 1 and Case 2 will be different. In Case 1, since there are no other possible outcomes with as low a likelihood, on the supposition of the Null Hypothesis, as the actual outcome, the p-value will just be the likelihood of the actual outcome itself—i.e., 0.01. In Case 2, since there are 20 other possible outcomes that have at least as low a likelihood, on the supposition of the Null Hypothesis, as the actual outcome, the p-value will be the sum of all of those likelihoods—i.e., 0.21. This difference in p-values is large and consequential; a p-value of 0.01 corresponds to a statistically significant rejection of the Null Hypothesis in many contexts, whereas a p-value of 0.21 is almost never statistically significant.

One consequence for the Classical Approach of the relevance of non-actual likelihoods is that the statistical significance of a particular actual outcome can depend on the "stopping rule" that was used when the outcome was observed.[9] Suppose, for instance, that Anne and Bob decide that they are going to flip a coin several times in order to determine whether they can reject the Null Hypothesis that the coin is fair. However, they have divergent

plans for how to collect the relevant evidence; Anne's plan is to flip the coin 20 times, whereas Bob's plan is to flip the coin until it has landed heads 6 times. Notwithstanding their different plans, they begin flipping the same coin together, each planning to stop when their individual stopping condition is satisfied and to perform their own Classical statistical analysis. But, imagine that, as things happen to turn out, the coin lands heads for the 6th time on the 20th flip, at which point *both* Anne and Bob stop flipping, since each of their separate stopping conditions has been simultaneously met. For the sake of concreteness, suppose that the outcome is the same as the one described in §3: TTTHTHTTTHTTTTTHHTTH.

As a result of their different stopping rules, Anne and Bob will calculate different p-values. Since Anne was going to flip the coin 20 times regardless of the results of each flip, the possible outcomes *for her* include all and only the 20-flip outcomes; by contrast, since Bob was going to flip the coin until it landed heads for the 6th time, the possible outcomes *for him* include all and only the 6-heads outcomes. So, for example, a 20-heads-0-tails outcome was possible for Anne, whereas it was not for Bob; similarly, a 6-heads-0-tails outcome was possible for Bob, but not for Anne. (Of course, the actual outcome of TTTHTHTTTHTTTTTHHTTH is in the set of outcomes that are possible for both of them.) Thus, when Anne and Bob calculate the total likelihood, on the supposition of the Null Hypothesis, that an outcome at least as unlikely as the actual outcome would occur, they will be looking to distinct (though partially overlapping) sets of possible outcomes. And so it is possible for them to calculate different p-values. In this case, whereas Anne will calculate a p-value of approximately 0.116 (as before), Bob will calculate a p-value of approximately 0.032.[10] Thus, Bob can reject the Null Hypothesis that the coin is fair at, for instance, the $\alpha = 0.05$ level (since $0.032 < 0.05$), whereas Anne cannot (since $0.116 > 0.05$).

By contrast, on the Bayesian approach, since the *actual* observed outcome was the same for both Anne and Bob, their inferential processes will be identical as long as their prior probabilities in the relevant hypotheses were the same.[11] Once the priors are fixed, all that matters is the likelihood of the *actual* outcome on the various hypotheses under consideration; likelihoods of *non-actual* outcomes on the various hypotheses under consideration are irrelevant. Since the likelihoods of the actual outcome on the various hypotheses under consideration is not (typically) impacted by the choice of stopping rule, differences in stopping rules are not the source of an inferential difference on the Bayesian approach.[12]

Different intuitions can be marshaled to support each side of this particular dispute about stopping rules. On the Bayesian side, there is a fairly clear intuitive sense in which Anne and Bob have made the "same observation" or collected the "same data" in the case above, and moreover there is a natural intuition that two individuals who are in the same initial epistemic state and who make the same observation should reach identical conclusions. In addition, there is some intuitive force to the thought that the "data" on which we base scientific conclusions is public and shareable, and that it does not depend on the private psychological plans of the individuals who collect that data. For instance, it is plausible to think that anyone who has access to an accurate record of the coin flipping (e.g., a video or a laboratory notebook) would be well-situated to evaluate the epistemic force of the results, without having to engage with questions about the identity of the coin-flipper or the counterfactual circumstances under which that person would have stopped the experiment (say, through conducting interviews with that person's friends and colleagues).

On the Classical side, however, there is the competing thought that Anne and Bob did *not* actually observe the same thing: Anne observed a six-heads outcome *relative to a stopping rule that terminated on the 20th flip*, whereas Bob observed a 6-heads outcome *relative to a stopping rule that terminated on the 6th heads result*. Relatedly, there is an intuition that experimental design *matters*, even once we fix the results of the experiment. One hesitation about so-called "optional stopping rules," for example, is that they permit a scientific investigator to continue collecting data until they observe their favored result, at which point they can promptly end the experiment, leading to results that are biased by the investigator's goals. Rather, the Classical thought goes, scientific objectivity requires that the experimental design be settled *in advance*, and that the investigators not be given discretion to end the experiment when it suits their intellectual or professional goals. But in order to rule that kind of perverse stopping rule out as epistemically illegitimate, it seems that we must adopt an approach (like the Classical one) on which the details of the relevant stopping rule can have an impact on the inferential significance of the observed data.

4 | LIKELIHOOD VALUES VS. LIKELIHOOD RATIOS

Another core difference between the Bayesian and Classical approaches is that Classicalists care about likelihood *values*, whereas Bayesians care about likelihood *ratios*.

As explained in §3, on the Classical Approach, a p-value is calculated by summing the likelihood *values* for each possible outcome that is at least as unlikely as the actual outcome, on the supposition of the Null Hypothesis. As a consequence, any outcome with an even *slightly* higher likelihood, on the supposition of the Null Hypothesis, than the actual outcome is completely disregarded in calculating a p-value; it is irrelevant whether such an outcome's likelihood exceeds the likelihood of the actual outcome by a large factor (or difference) or a small one. Similarly, any outcome with a likelihood that is no higher, on the supposition of the Null Hypothesis, than the actual outcome, will simply contribute the value of its likelihood to the p-value.

By contrast, on the Bayesian approach, what matters is not the *values* of the various likelihoods per se, but rather their *ratios*. If an outcome that is extremely unlikely on the supposition of some particular hypothesis occurs, that alone is no reason for a Bayesian to lower their credence in that hypothesis at all; all that matters is how the likelihood of that outcome on the supposition of the hypothesis *compares* to the likelihood of that outcome on the supposition of competitor hypotheses. So, for example, if an outcome occurs that is just as unlikely on the supposition of *any* of the hypotheses under consideration, then a Bayesian will reason that the outcome—however unlikely—does not distinguish between the hypotheses, and hence that it does not provide a reason to change their credences in the hypotheses.

One consequence of this difference is that, on the Classical Approach, it is possible to reject two (or more) different hypotheses, even if it is known that one of those hypotheses is true. After all, an observed outcome might be extremely improbable on *any* of the hypotheses under consideration, in which case each hypothesis can be rejected at a statistically significant level. By contrast, since the Bayesian Approach focuses on likelihood ratios rather than values, an observed outcome that has a low likelihood on each of the relevant hypotheses doesn't necessarily constitute a reason to rule out (or to assign low credence to) any of the relevant hypotheses; all that matters to a Bayesian is how much *likelier* the outcome was on the supposition of one hypothesis than it was on the supposition of its competitors.[13]

5 | DESCRIBING THE DATA

A third point of difference between the Bayesian and Classical approaches to inference is the issue of how to describe the data on the basis of which a statistical inference is being made. Bayesians standardly embrace the Requirement of Total Evidence, according to which "to the extent that what it is reasonable to believe depends on one's evidence, what is relevant is the bearing of one's total evidence."[14] The rough idea here is that, when we ignore some of the evidence that is in our possession, this can lead us epistemically astray. For example, someone would violate the Requirement of Total Evidence if they were to evaluate the dangerousness of a particular situation by attending only to their evidence that there is a bear nearby, ignoring the evidence they have that it is a *friendly* (and hence non-dangerous) bear. To be sure, the Requirement of Total Evidence sometimes directs us to take *irrelevant* evidence—say, the fact that a Bob Dylan song was playing on the radio when the instrument detected a particular photon—into account, but this is not epistemically problematic; as long as none of the hypotheses under consideration assigns a different likelihood to Bob Dylan playing on the radio than any other hypothesis does, taking the extra evidence into consideration will yield the same results as ignoring it, just as we would expect for irrelevant evidence.

By contrast, there is an important sense in which Classical approaches to inference often *depend on* violations of the Requirement of Total Evidence. Consider, for instance, the toy example from §3. The calculation of the p-value in that case depended on individuating the outcomes in terms of the *number of heads results* (or, equivalently, in terms of the *number of tails results*), which could range from 0 to 20. On that individuation of outcomes, there are

some outcomes, such as $r = 10$, that have higher likelihoods on the supposition of the Null Hypothesis, and there are some outcomes, such as $r = 0$ and $r = 20$, that have lower likelihoods; as a result, it is possible to observe an outcome (such as $r = 20$) that is associated with a p-value in the "statistically significant" range. However, if we had individuated the possible outcomes more finely—say, in terms of the precise sequence of heads and tails flips, rather than in terms of the total number of heads (or tails) flips—then a statistically significant p-value is no longer possible. After all, on the supposition of the Null Hypothesis, *every single* sequence of heads and tails results has exactly the same likelihood: $\frac{1}{2}^{20}$. Thus, on this individuation of outcomes, *any* possible outcome (including, for example, the all-heads outcome HHHHHHHHHHHHHHHHHHHH) is such that the likelihood, on the supposition of the Null Hypothesis, that an outcome as least as unlikely would occur is 1. And so, relative to this individuation of outcomes, a p-value of 1 will be calculated, regardless of which sequence of heads and tails is actually observed; obviously, this corresponds to the absence of statistical significance at any level.

One Classical strategy for solving this problem is to insist that the test statistics that should be used in significance tests are the "minimal-sufficient" statistics.[15] A statistic t is *sufficient* relative to a particular Null Hypothesis iff, on the supposition that the Null Hypothesis is true, all of the more specific outcomes compatible with t are equally likely. So, for example, the "number of heads" statistic is sufficient in our coin case, relative to the Null Hypothesis that the coin is fair; on the supposition that H_0 is true, each 1-head outcome is equally likely (i.e., $20 \times \left(\frac{1}{2}\right)^{20}$), and similarly for each other n-heads outcome. However, the precise sequence of heads and tails is a sufficient statistic too; each value of that statistic is compatible with only one possible outcome, so sufficiency is trivially secured. A *minimal-sufficient* statistic, then, is a sufficient statistic such that any loss of information would destroy its sufficiency. The precise sequence of heads and tails is not minimal-sufficient, since the "number of heads" statistic contains less information and is still sufficient. The motivating idea behind this response is that, since a minimal-sufficient statistic partitions the outcome space into equivalence classes of outcomes that have equal likelihoods on the supposition of the Null Hypothesis, it contains all of the information that is really needed when we are trying to evaluate that Null Hypothesis.

One potential worry about this appeal to minimal-sufficiency is that, as noted above, the Requirement of Total Evidence mandates using the *most* informative statement of our evidence available, and minimal-sufficient statistics are (at least often) *less* than maximally informative, so there is a prima facie tension with the (independently plausible) Requirement of Total Evidence. Another worry is that, in the particular coin case we've been considering where the Null Hypothesis is that the coin is fair, the "number of heads" statistic is not minimal-sufficient either, since the "empty" statistic that contains no information at all (think of it as the "information" that there were 20 flips) is sufficient too, as it partitions the outcome space into a single equivalence class containing 2^{20} precise sequences, each of which is equally likely on the supposition of the Null Hypothesis.[16] And, of course, when we use that "empty" statistic, we learn nothing at all; the likelihood, on the supposition of the Null Hypothesis, that the empty statistic would take the value it does is 1, and so statical significance at any level is impossible relative to this statistic. Finally, even in cases where the Null Hypothesis involves probabilities other than $\frac{1}{2}$, the strategy of appealing to minimal-sufficient statistics can have the effect of collapsing intuitively different outcomes into the same equivalence class, if the Null Hypothesis "just so happens" to assign them equal likelihoods.[17]

6 | THE ELIMINABILITY OF SUBJECTIVITY

A fourth difference between the Bayesian Approach and the Classical Approach is implicit in the above, but it is worth making explicit: the "location" of subjectivity.

On the Bayesian approach, one large and important question is which constraints, if any, the agent's prior credence distribution is subject to. This issue divides so-called "subjective Bayesians" and "objective Bayesians"; roughly, the former camp thinks that any coherent[18] prior credence distribution that an agent might have is perfectly

rational, whereas the latter camp thinks that there are additional objective constraints on rational prior credence distributions that go beyond mere coherence.[19] But it has been notoriously difficult to formulate a version of objective Bayesianism that is precise about what the objective constraints on rational prior credence functions are, and in a lot of cases it is very difficult to imagine how, even if there are such constraints, they could *uniquely* settle the rationally permissible prior credence distribution to have; for example, it is very hard to imagine what sorts of constraints could make it the case that the uniquely rational prior credence to have in (say) the General Theory of Relativity is (say) 0.28934218, rather than some other non-extreme value.

In light of this, most Bayesians look to be committed to the claim that it is possible for Scientist #1 and Scientist #2 to come to the table with different prior opinions about some subject matter, to collect the same evidence, and to be fully justified in reacting differently to that evidence.[20] This sort of "subjectivity" in the Bayesian program has troubled some philosophers and statisticians, and has inclined them toward statistical methods that are not similarly dependent on subjective factors. However, while it is true that, since the Classical Approach does not rely on a prior credence distribution, it does not embed *exactly the same* subjective element as the Bayesian Approach does, it is not clear that the Classical Approach is any less subjective overall. For example, as discussed above, an evidence-gatherer's choice of stopping rule and an evidence-analyzer's choice of test statistic can have a substantial impact on the results of a Classical statistical analysis, and it is not ultimately clear that these subjective elements can (or should) be eliminated from the Classical Approach. Thus, while the Bayesian and Classical approaches may disagree about the *location* and *role* of subjective elements in statistical analysis, it is not clear that they ultimately disagree about *whether* subjective elements can or should be eliminated from that analysis entirely.

ACKNOWLEDGMENT

Thanks to Dan Greco, Branden Fitelson, Ram Neta, John Roberts, Roger White, an anonymous referee from this journal, and an audience at the Eastern APA for helpful comments and feedback.

ORCID

Matthew Kotzen https://orcid.org/0000-0003-0305-4736

ENDNOTES

[1] Technically, the agent's prior credence distribution is defined over an *algebra* (more specifically, a σ-algebra) of hypotheses, which is formed by taking the closure of this set of hypotheses under the operations of (countable) union, (countable) intersection, and complement. In effect, the requirement that the agent's credences be defined over an algebra of basic outcomes guarantees that, whenever an agent has credences in those basic outcomes, the agent will also have credences in arbitrary conjunctions, disjunctions, and negations of those outcomes.

[2] This is intended only as an illustration of one intuition behind the Rule of Conditionalization, not as a fully developed *argument* for the Rule of Conditionalization. For arguments for the Rule of Conditionalization, *see*, *e.g.*, Greaves and Wallace (2006), Lewis (1999), Pettigrew (2020), and Savage (1954).

[3] For a much more detailed (but very accessible) overview of Bayesian Confirmation Theory, *see* Strevens (2017).

[4] For a discussion of Neyman-Pearson tests, see Howson and Urbach (1993).

[5] For discussion *see* Cohen (1994).

[6] The level of α in a particular context that is "statistically significant" varies depending on the nature of the data involved, the practical importance of the decision, and a variety of other factors, and there is undoubtedly some degree of arbitrariness in setting a level for some particular decision, or some particular publication, at 0.05 rather than (say) 0.049 or 0.051. Note too that the value of α does not directly speak to the probability of committing a Type II Error (i.e., the error of failing to reject the Null Hypothesis when it is actually false) in situations in which the Null Hypothesis is false; to address this, we must determine the value of a different parameter, β.

[7] One reason that this is not the same as having a credence of 0.95 that the Null Hypothesis is false is that we haven't yet said anything about the "base rate" at which the Null Hypothesis is true in the world. Indeed, whereas the Bayesian Approach explicitly takes this base rate into account in the prior probability distribution, the Classical Approach is designed to *ignore* the base rate and focus instead on p-values. For example, this morning, I saw a student who was on her

way to class catch a wayward frisbee with one hand while she was holding her cell phone in the other hand. Consider the Null Hypothesis that this student is not a professional frisbee player. On the assumption of that Null Hypothesis, the likelihood was fairly low that she would be able to catch an unexpected frisbee one-handed; let's suppose that the likelihood was 5%. Similarly, if 100 students who are not professional frisbee players were to find themselves in the same situation, only around 5% of them would catch the frisbee. But none of that commits me to having a credence of 0.95 that the particular student I saw is a professional frisbee player. After all, the vast majority of students are not professional frisbee players, and it is overwhelmingly likely that the student I saw was simply an ordinary student who made an impressive catch, not a professional frisbee player. So, my credence is not at all high, and nowhere close to 0.95, that that student is a professional frisbee player. And yet, on the assumption that she is not a professional frisbee player, the likelihood was only 0.05 that she would make a catch that was (at least) that impressive. A somewhat less accurate, but more intuitive, way to make this point is to say: to reject a Null Hypothesis is to think that, if that Null Hypothesis is indeed true, then something pretty unusual happened. But there are circumstances in which it's perfectly reasonable to think that something pretty unusual (probably) happened—as when I think that I probably saw a student who is not a professional frisbee player make an improbably nice catch.

[8] This point about the appropriate description of the experimental outcome will be explored in more detail in §6 below. For now, it suffices to note that the only likelihoods that matter for a Bayesian are the likelihoods of *E itself*—nothing stronger or weaker than *E*, nothing with a different content than *E*—on the various hypotheses under consideration.

[9] *See*, *e.g.*, Howson and Urbach (1993, pp. 213–15), Savage (1962, p. 18), Whitehead (1993, pp. 1412–13), and Gillies (1990, p. 94).

[10] The number of 6-heads-n-tails outcomes is $\frac{(5+n)!}{n!5!}$, and the likelihood of a 6-heads-n-tails outcome on the supposition of H_0 is $\frac{(5+n)!}{n!5!}(0.5)^{6+n}$. For the actual value of $n = 14$, this likelihood is approximately 0.011. All and only outcomes with $n \geq 14$ are at least as unlikely on the supposition of H_0 as the $n = 14$ (i.e., the actual) outcome (the likelihood for $n = 0$ is approximately 0.016). The sum of all likelihoods for $n \geq 14$ is approximately 0.032.

[11] For discussion *see* Rouder (2014).

[12] Of course, a difference in stopping rules could matter even for a Bayesian, if the choice of stopping rule impacts what outcome is actually observed. But the important point is that the stopping rule does not have an inferential impact per se; once we fix the priors, the outcome, and the likelihood of that outcome on the various hypotheses, the choice of stopping rule has no further inferential significance.

[13] This challenge for the Classical Approach is known as Lindley's Paradox; it was first raised by Harold Jeffries in 1939 and got its name from Dennis Lindley in 1957. It is discussed in Kyburg (1974) and Howson and Urbach (1993).

[14] Kelly (2008). For a discussion of how to implement the Requirement of Total Evidence, see Kotzen (2012).

[15] For a discussion of sufficiency and minimal-sufficiency, *see* Howson and Urbach (1993, pp. 189–92). For a defense of the minimal-sufficiency move, *see* Seidenfeld (1979).

[16] This would not be the case if the Null Hypothesis were that the coin has some particular bias—say, 0.6 in favor of heads—since in that case not all of the more specific outcomes would be equiprobable on the assumption of the Null Hypothesis.

[17] For example, consider the Null Hypothesis that the heads-bias of a particular coin is $\frac{6-\sqrt{6}}{5} \approx 0.710$, and suppose that the coin is flipped four times. On that Null Hypothesis, the likelihood of a four-heads outcome is identical to the likelihood of a two-heads-two-tails outcome: $\left(\frac{6-\sqrt{6}}{5}\right)^4 \approx 0.254$. Thus, on the strategy under consideration, the statistic that can take the values of 0-heads, 1-heads, 2-or-4 heads, or 3-heads is a sufficient statistic, and thus prevents the "number of heads" statistic from being minimal-sufficient.

[18] A "coherent" credence distribution is simply one that obeys the axioms of probability theory.

[19] For some defenses of subjective approaches, *see*, *e.g.*, Jeffrey (1992), de Finnetti (1937), and de Finetti (1974). For some defenses of objective approaches, *see*, *e.g.*, Jaynes (2003) and Rosenkrantz (1981).

[20] Bayesians often appeal to "washing out of the priors" results here to show that any effect of initial differences of opinion will vanish to zero as more evidence is collected. For discussion *see* Hawthorne (1994).

REFERENCES

Cohen, J. (1994). *The world is round: p < 0.05* (Vol. 49, pp. 997–1003). American Psychologist.

de Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives, *Ann. Inst. Henri Poincaré. 7*, pp. 1–68. Translation reprinted in H. E. Kyburg and H. E. Smokler (eds.) (1980), Studies in Subjective Probability, 2nd ed. (Krieger): pp. 53–118.

de Finetti, B. (1974). *Theory of Probability* (Vol. 1). Wiley and Sons.

Gillies, D. (1990). Bayesianism versus Falsificationism. *Ratio (New Series)*, *III*(1), 82–98. https://doi.org/10.1111/j.1467-9329.1990.tb00014.x

Greaves, H., & Wallace, D. (2006). Justifying Conditionalization: Conditionalization Maximizes Expected Epistemic Utility. *Mind*, *115*(459), 607–632. https://doi.org/10.1093/mind/fzl607

Hawthorne, J. (1994). On the Nature of Bayesian Convergence. *PSA*, *1*, 241–249.

Howson, C., & Urbach, P. (1993). Scientific Reasoning: The Bayesian Approach. (2nd ed.).

Jaynes, E. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press.

Jeffrey, R. (1992). *Probability and the Art of Judgment*. Cambridge University Press.

Kelly, T. (2008). Evidence: Fundamental Concepts and the Phenomenal Conception. *Philosophy Compass*, *3*(5), 933–955. https://doi.org/10.1111/j.1747-9991.2008.00160.x

Kotzen, M. (2012). Selection Biases in Likelihood Arguments. *British Journal for the Philosophy of Science*, *63*(4), 825–839. https://doi.org/10.1093/bjps/axr044

Kyburg, H. (1974). *The Logical Foundations of Statistical Inference*. Reidel.

Lewis, D. (1999). Why conditionalize? In *Papers in Metaphysics and Epistemology* (pp. 403–407). Cambridge University Press.

Pettigrew, R. (2020). What is Conditionalization, and Why Should We Do It? *Philosophical Studies*, *177*(11), 3427–3463. https://doi.org/10.1007/s11098-019-01377-y

Rosenkrantz, R. (1981). *Foundations and Applications of Inductive Probability*. Ridgeview Publishing.

Rouder, J. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*(2), 301–308. https://doi.org/10.3758/s13423-014-0595-4

Savage, L. J. (1954). *The Foundations of Statistics*. Wiley.

Savage, L. J. (1962). Subjective Probability and Statistical Practice. In G. A. Barnard & D. R. Cox (Eds.), *The Foundations of Statistical Inference* (pp. 9–35). Wiley and Sons.

Seidenfeld, T. (1979). *Philosophical Problems of Statistical Inference*. Reidel.

Strevens, M. (2017). Notes on Bayesian Confirmation Theory. http://www.strevens.org/bct/BCT.pdf

Whitehead, J. (1993). The Case for Frequentism in Clinical Trials. *Statistics in Medicine*, *12*(15–16), 1405–1413. https://doi.org/10.1002/sim.4780121506

AUTHOR BIOGRAPHY

Matthew Kotzen is an Associate Professor of Philosophy and the Chair of the UNC Department of Philosophy. His research is primarily on issues in epistemology, the philosophy of science, and the law of evidence.

How to cite this article: Kotzen, M. (2022). The Bayesian and Classical Approaches to statistical inference. *Philosophy Compass*, e12867. https://doi.org/10.1111/phc3.12867