Hume, Popper, Bayes:
An essay on how we come to know stuff

Brian Powell*

(Dated: Spring 2015)

We review the major achievements and challenges in the development of modern scientific inference, from Hume's articulation of the problem of induction, through Popper's attempts at a cure through strict falsification, to a modern synthesis of these genealogies via Bayes' Theorem. Methods of model selection within Bayesian inference are seen to combine inductive, experience-based learning with Popper's demand for testability into a useful measure of confirmation.

1. Introduction

Three million years ago, the grandfathers of our genus used sharpened stones to chop wood and cut bone. Today, surgeons use lasers to cauterize tiny blood vessels in the eye. Within our own lives, we've gone from burning our mouths on hot chocolate as children to cautiously blowing on our tea as adults. These are examples of *learning*, of individuals and species adapting their behaviors and modifying their worlds in response to life experience. That we have learned is undeniable, but 300 years ago David Hume argued convincingly that this process of learning—of generalizing about the world from individual experiences—is not logically founded, that no formal proof of its effectiveness exists. What then of humanity's most prized system of inquiry, the scientific method? Most of us are convinced that astronomy gets closer to the true nature of things than astrology, but Hume's argument not only blurs the distinction, it claims that no such comparison can be made. To the practicing scientist who is primarily concerned with collecting data, publishing papers, going to great locales for conferences, and so on, these are mere academic concerns to trouble only philosophers; after all, scientists are just driving the car, they don't care about what's buzzing under the hood. But Hume's challenge is a deep stab at the heart of the scientific enterprise: it's a paradox that flies in the face of basic experience, one that has resisted centuries of concerted effort at finding a resolution. Even if we don't care about the formalities, in exploring this problem we'll come to better understand the practice of science and how we make inferences about the world.

2. "Logical induction ain't logical" – David Hume

Most of us are familiar with logical arguments like:

1. All priests are men.
2. Father Jim is a priest.

∴ Father Jim is a man.

This is an example of a logical *deduction*: the truth of the conclusion is guaranteed by the truth of the premises so long as we follow the rules of deductive logic. The thing about deduction is that the conclusions never tell us anything new, since they are always fully entailed, if only implicitly, by the premises. In the example above, the fact that Father Jim is a man is already contained within the

*bpowell@ida.org

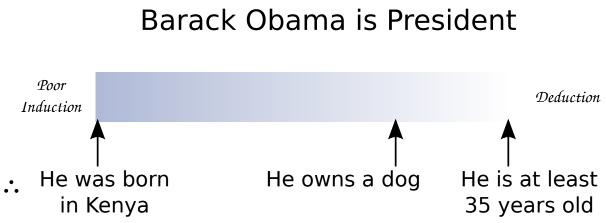

FIG. 1. The sliding scale of induction, from poor inductive reasoning to deduction.

two premises—the conclusion merely makes the statement explicit. Because deductions never tell us anything new, they cannot increase, or amplify, our knowledge—they are said to be *non-ampliative.* How is it that we know the premises are true? Because, among other things, being a man is a necessary requirement for being a priest—this deduction is true *by definition.*

What about this:

The sun has risen on Earth
every day for the past 4.5 billion years.

∴ The sun will rise tomorrow.

Here the premise is true—it is based on historical knowledge—but the conclusion, which refers to the future, does not follow. Regardless of how certain we feel, the conclusion is not guaranteed because it contains information not provided in the premises. This kind of inference, in which the premises do not necessarily entail the conclusion, is called *induction.* You can think of it as the less certain cousin of deduction. The above conclusion is tantamount to saying that the Sun will rise everyday, because there’s nothing in the premises that singles out “today” as particularly special. This is the essence of induction: *from the observation of individual particular cases we project universal generalizations.* In the case of the rising sun, we have many, many individual observations and so we might suspect, though maybe not logically guaranteed, that our future projection is solid.

But what about this:

5 jellybeans, all of which have been black,
have been selected from a jar containing 100 jellybeans.

∴ All jellybeans in the jar are black.

We feel less sure of this conclusion because there seem to be too few observed cases to speculate with much confidence about the larger collection of jellybeans. Inductive reasoning spans the spectrum of inference: from deduction (truth by logical necessity), to flimsy assertions, to the most egregious *non sequiturs* (see Figure 1). Ideally, our inferences should be inductively strong, as close to the deductive end of the scale as possible. Strong inductive arguments are ones in which the conclusions follow from the premises with high probability, like the case of the rising sun. Sometimes the conclusion forms the basis of a hypothesis—a more general statement or principle that includes the premises as particular cases. This generality is achieved through inductive reasoning by extrapolating particular observations across time, space, or jellybeans in jars. The problem is that we have no way of knowing how reliable it is.

Consider the following:

Induction has worked reasonably well in the past

$\therefore$ Induction will work reasonably well in the future.

This is an inductive argument concerning induction itself. Assuming the premise is true, we expect the conclusion, that induction will continue to work reasonably well in the future, to be true. Seem circular? It is, badly so. This is the problem of induction. Hume nailed it in 1748 [1]

> ...there can be no demonstrative arguments to prove, that those instances, of which we have no experience, resemble those, of which we have had experience...even after the observation of the frequent or constant conjunction of objects, we have no reason to draw any inference concerning any object beyond those of which we have had experience.

Hume's challenge is that the success of induction, arguably *the* method of establishing natural law since the Enlightenment, cannot be formally demonstrated. Philosopher Willard V. Quine offers a 20th-century view: "The brute irrationality of our sense of similarity, its irrelevance to anything in logic and mathematics, offers little reason to expect that this sense is somehow in tune with the world...Why induction should be trusted, apart from special cases such as the ostensive learning of words, is the perennial philosophical problem of induction." [5]

The solution is at least intuitively obvious: there must be some fundamental uniformity within nature with which to secure our extrapolations. But there are so many ways that nature might appear uniform: so many true regularities, so many perceived but unrelated correlations, so many random occurrences. Unless the requisite uniformity exists *a priori* (a likely untenable proposition, regardless of what Herr Kant might think), it must be established empirically—a feat that necessarily involves induction, taking us right back to where we started. There have been many contemporary attempts to solve Hume's problem (see, for example, [2] and [3]), but all are inextricably self-referential and question begging; all ultimately fail. This is not surprising, because *there can be no resolution to the problem in general.* How do we know whether an inductive argument is likely correct? We don't—we can't—because the conclusion is not based on experience nor a logical consequence of it. Following philosopher Nelson Goodman we adopt the view that Hume's problem is "not a problem of *demonstration*, but a problem of *defining* the difference between valid and invalid predictions" [4] (emphasis mine), that is, the method of induction must be taken as given: the challenge is the more practical one of knowing when we can and cannot apply it.

To illustrate this point, say a die is rolled one million times with each face coming up very nearly 1/6 of the time. We are pretty confident that the die is fair, and if asked to bet on whether the next two rolls will be sixes, we'd bet odds 36-to-1 against. And if the existence of the gambling industry is any proof, we'd be substantiated in this judgement. The roll of the die is a relatively simple phenomenon with readily recognizable symmetry—we can *apply* induction with high confidence. Meanwhile, Hume's concern seems quite immaterial, amounting to a warning that the universe might one day change, abruptly, in such a way that those odds shift, that the die roll no longer conforms to the same long-running averages. This is obviously impossible to guard against—the problem of induction is not about our troubles in dealing with a potentially arbitrary and capricious universe. The problem is instead one of how to isolate the regularities relevant to the system under investigation in the face of incomplete knowledge. We cannot prescribe rules for how to do this *in general* (echoing Goodman, the problem is not one of *demonstration*), but the validity of individual hypotheses about the world can be decided from *within* the system of inductive inference.

Because the uniformity emerges clearly after many trials, the roll of the die is a simple and notable success story. Not all problems are this easy. Suppose we are looking for a mathematical expression governing a certain physical process. As data we have the numbers $1, 3, 5, 7, 9, \cdots$. The sequence is generated by the expression $x_n = 2n + 1$, starting from $n = 0$, and our prediction for the 6th number is $x_5 = 11$. Meanwhile, a rival team of researchers advances the prediction $x_5 = 3845$, based on the relation $x_n = (1 - 2n - 1)(3 - 2n - 1)(5 - 2n - 1)(7 - 2n - 1)(9 - 2n - 1) + 2n + 1$. Both sequences describe the regularity of the data equally well, but they make wildly different extrapolations beyond them. We can think of the alternative sequences as two competing hypotheses, and the challenge is to select the one that best generalizes the given data.

Nelson Goodman found an insightful way of articulating this problem. All emeralds that have ever been observed have been green. Suppose there is some other disposition of color called "grue", such that something that is grue is green before, say, July 1 2100, and blue afterwards. All emeralds that have ever been observed have also then been grue. Now, which color do we predict for yet to be observed emeralds? If we project green we will be correct, whereas a projection of grue will throw us a curve ball—it will not support a strong inductive inference once the calendar ticks past July 1, 2100. This example is deliberately artificial, but Goodman's point is an important one: there is no clear way to differentiate projectable properties, like green, from non-projectable ones, like grue. Goodman suggests past experience as a guide: "Plainly, 'green', as a veteran of earlier and many more projections than 'grue', has the more impressive biography. The predicate 'green', we may say, is much better *entrenched* than the predicate 'grue"' [4]. So while there is no hard rule for distinguishing projectable from non-projectable characteristics, we expect that our acquaintance with the world should inspire our ability to tease out the essential characteristics of the system under study. Not foolproof though, and not always helpful: for example, experience doesn't suggest which expression for x_n above is projectable. The problem of projectability is pervasive and formidable: Goodman named it the "new riddle of induction".

Our battle against Hume's problem of induction has moved from the philosophical high ground to the front lines of applied science. We have passed on solving the meta-problem of induction as a method, focusing instead on our ability to discern projectable patterns in the data log book in support of strong inductive inference. In science, we are given an incomplete glimpse of a small part of the world, and it's easy to suppose that induction, however imperfect, is the only way to complete the picture. This view is not without its challengers.

3. "We don't need no stinkin' induction" – Sir Karl Popper

The traditional scientific method supposedly employs induction both in the course of forming hypotheses and empirically confirming them. The examples of the previous section—whether the sun will rise tomorrow, or whether the sequence $1, 3, 5, 7, 9, \cdots$ anticipates all odd numbers—were discussed in the context of hypothesis formation. They were simplified illustrations, but there are real-world examples. Take Charles Darwin: it was through his careful study and direct observation of individual animal specimens that he came to discern the hints of a much grander and extensive law of nature. By examining the beak shapes of particular finches in Galapagos, Darwin was able to induce a wide ranging principle that applied not only to other species, but across eons of time. Incidentally, this is an example of strong induction, as evidenced by the wild success of the theory of evolution by natural selection. There are other great hypotheses, however, that do not share this inductive genealogy: Schrödinger's non-relativistic wave equation comes to mind, as does Einstein's relativity theory. In Schrödinger's case, the equation was the result of a creative, non-logical affair incorporating theoretical aesthetic, bold conjecture, and apparently a skiing holiday in the Swiss Alps; in Einstein's case, it was a thought experiment that got terribly out of hand.

Surely these ideas came from *somewhere*; after all, not since Locke introduced the world to em-

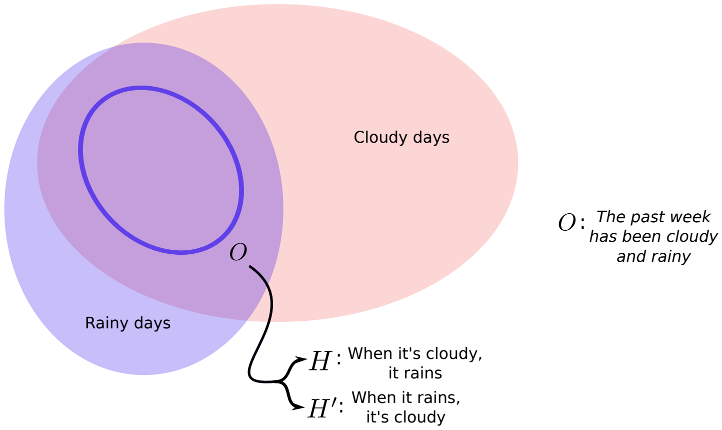
Cloudy days
O: The past week has been cloudy and rainy
O
Rainy days
H: When it's cloudy, it rains
H': When it rains, it's cloudy

FIG. 2. The danger of working from observation, $\boldsymbol{O}$, to hypothesis, $\boldsymbol{H}$. Given the observation (circled in blue) that the past week as been both cloudy and rainy, the causal relationship between rain and clouds is unclear—it is not known whether clouds are a necessary or sufficient condition for rain. The observation is therefore compatible with at least two hypotheses: $\boldsymbol{H}$, which asserts that clouds are a sufficient condition for rain, and $\boldsymbol{H'}$, which asserts that clouds are a necessary condition.

piricism have people actually believed that facts about nature can be known *a priori*[1]. Indeed, they come from observations, but not systematically and not without man’s psychological propensity for pattern seeking or his in-born sense of causal relation[2]. Philosopher of science, Karl Popper, rejected hypothesis-by-induction outright, arguing instead that science progresses not through the steady accumulation of evidence to universals, but instead as a series of conjectures and refutations, of “inventions – conjectures boldly put forward for trial, to be eliminated if they clashed with observations...” that arise “psychologically or genetically *a priori*, prior to all observational experience.”[6] It’s easy to misinterpret Popper’s statement: he does not literally mean *all* observation experience, but that precursor experience (“genetically *a priori*”) from which inductive generalizations are reached. We’ll grant that induction has nothing necessarily to do with the *formation* of hypotheses and examine instead the main theme of this essay: its role in *justifying* hypotheses.

If we consult again the examples from the last section, they are just as much acts of hypothesis justification as they are examples of hypothesis formation. For example, the observation of a few black jellybeans motivated the hypothesis (if unwisely) that all jellybeans in the jar were black. Equivalently, the observation of a few black jellybeans also constitutes a certain amount of evidence in favor of the proposed hypothesis that all jellybeans in the jar are black. And of course, the comparison of the data with the hypothesis fundamentally involves induction. Schematically, the process of hypothesis testing goes as follows: from hypothesis, H, an observable prediction, O, is deduced: $H \to O$. This prediction will be about about a track in a cloud chamber, or a fossil in the Cambrian. We then set up an experiment or organize an expedition to test the prediction. To confirm the hypothesis, we must use inductive inference: we generalize the observation and compare it against the hypothesis: $O \to H$. Now, if O agrees with the prediction, have we confirmed H? Sadly, no, because $O \to H$ is not a deductively valid move. It is an example of the logical fallacy of *affirming the consequent*: the assertion $O \to H$ does not follow from $H \to O$ because it is in general possible to find some other hypothesis, say, H', that is also compatible with the observation, $H' \to O$. This is the problem of projectability that we

[1] Think babies born with electrodynamics already in their heads.
[2] Think Kant’s synthetic *a priori* causal sense, whatever that’s supposed to be.

just discussed: for example, the observation of the numbers $1, 3, 5, 7, 9$ agrees with both the hypothesis $x_n = 2n + 1$ and $x_n = (1 - 2n + 1)(3 - 2n + 1)(5 - 2n + 1)(7 - 2n + 1)(9 - 2n + 1) + 2n + 1$.

Sometimes we *can* legally get away with affirming the consequent if we know that a particular hypothesis describes all the *necessary* and *sufficient* conditions for the observation. The only way to know is to do comprehensive experiments that account for all the relevant conditions, a feat that is seldom possible in practice: as Figure 2 shows, the data might fail to reveal the causal relationships between the experimental conditions and the observations, leaving multiple hypotheses in the running.[3] In general, though, certain and exact confirmation of unique hypotheses is not possible.

In the 1930’s, Karl Popper was moved to abandon confirmation as an instrument of scientific discovery because he did not accept the inductive scaffolding on which it is built. He recognized an important asymmetry between the acts of confirmation and refutation: while it is impossible to verify a universal statement by observing singular instances, universal statements can be *contradicted* by individual observations. Take the prototypical sample hypothesis that all swans are white—though the number of swans is arguably finite, we cannot search the whole world over in an attempt to verify this hypothesis (and if we could, it would not result in a predictive theory). We can, however, falsify it, swiftly and assuredly by observing just a single black swan. Unlike confirmation, the falsification of hypotheses is a deductively-sound procedure employing the classical rule of *modus tollens*: if $H \to O$, then $\overline{O} \to \overline{H}$ (where $\overline{O}$ is read “not O”). While in practice statistical limitations might require several disconfirming trials before a hypothesis or theory is discarded, sometimes the evidence is resounding: for example, the discovery of the cosmic microwave background in 1965 decapitated Hoyle’s steady-state universe model in one fell swoop. By proposing falsification as *the* means of testing hypotheses, Popper sought to establish of a logically-valid, *induction-free* approach to the justification of all scientific proposals based on the simple maxim: “it must be possible for an empirical scientific system to be refuted by experience” [7].

3.1 Attempts at falsification: the p-value catastrophe

Falsification is the presumed mode of inference employed by much of contemporary science based on frequentist, or classical, statistics. The weapon of choice is the most well-worn instrument of statistical terror ever to be foisted upon the scientific process—the *p-value*. Suppose we wish to test some hypothesis, H_0. The subscript ‘0’ indicates that this hypothesis generally refers to some fiducial model, perhaps one for which the effect we’re seeking is absent. In this case it is appropriately called the *null hypothesis*, and the purpose of the p-value is to determine the agreement between the observed data, O, and the null hypothesis, H_0, when there are other possible explanatory hypotheses. Specifically, it provides the probability that we would observe the data O *under the assumption that* H_0 *is correct*, $p(O|H_0)$.[4] A small p-value, say less than 0.05, means that there is a smaller than 5% chance that we would have measured the data O given that H_0 is true. In other words, H_0 is correct, and the tension with the data arises instead due to chance (whether it’s from noise in the measuring device or some fundamental uncertainty of the physical process being measured.) Often, however, in the statistical parlance one reads that the p-value as the probability of “falsely rejecting the null hypothesis”. This agrees with our understanding of the p-value as the chance of a statistical fluke, but we are actually in no way licensed to reject, or *falsify*, the null hypothesis if the p-value happens to be small.

[3] Charlatans and conspiracy theorists alike adore this kind of situation, because they believe it gives faith healing and UFO abductions equal room on the stage of potential explanations for a given piece of evidence. It doesn’t, and we’ll see why shortly.

[4] The notation $p(x|y)$ should be read “the probability of x given that y is true.”

Let’s see what goes wrong when we use p-values to make inferences. Upon making observation O, we compute $p(O|H_0)$, the probability that we measure O given the truth of H_0. Suppose the p-value is found to be < 0.05. Next, we equate $p(O|H_0)$ with $p(H_0|O)$, the probability of the truth of H_0 given the observation, O. Then, we conclude $p(\overline{H}_0|O) = 1 - p(H_0|O)$, which asserts that some alternative hypothesis (or group of hypotheses), $\overline{H}_0$, which is typically a model (or group of models) that supports the observed effect, is correct at 95% confidence. Can you spot the error? In the very first step we took $p(O|H_0) = p(H_0|O)$. This is wrong—it’s called the *fallacy of the transposed conditional.* It’s not as bad as affirming the consequent (for which we’d always have $p(O|H_0) = p(H_0|O) = 1$), but it can really get us into trouble. An especially striking example of this confusion is discouragingly called the *prosecutor’s fallacy*, because it has been used by the prosecution in criminal court to argue for the guilt of the defendant: the probability of getting certain evidence, say DNA evidence (E), given the innocence (I) of the defendant, $P(E|I)$ (which is typically low) is equated with the probability of innocence given the evidence, $p(I|E)$ (which is then likewise low). What is neglected here is the chance that the defendant is innocent *in the first place, before the evidence is brought.* Catastrophic when misused in the court of law, it is also tragic that it provides the statistical foundation of many important scientific results, from drug efficacy to climate change.

In practice, we see that the p-value is not *really* about falsification: it’s used ostensibly to *confirm* a hypothesis alternative to the null hypothesis, which it must falsify in the course of doing business. And we must commit a logical fallacy in order to accomplish this inference. Popper would not be pleased, nor would anyone seeking a coherent scientific inference. P-value catastrophe aside, Popper’s model of falsification still has its dissenters. After all, what good is the scientific method if it can only tell us when we get stuff wrong?

3.2 Corroboration without induction?

Popular objections to Popper’s program tend to center around the perceived negativity associated with the act of falsification; call it a psychological aversion to failure. More practically, the concern is that science should not be about what doesn’t work—how can we hope to increase our knowledge of the world if we cannot confirm, or verify, scientific hypotheses? Shouldn’t science be a constructive, rather than destructive, pursuit? According to Popper, the only tenable position is that all of our theories are basically wrong, awaiting falsification. Importantly, though, Popper views the cycle of conjecture and refutation as generative, as one that puts scientific hypotheses through a kind of optimization process in order to develop theories that, though ultimately incorrect, are the very best possible prototype of the truth: “Theories are nets cast to catch what we call ‘the world’: to rationalize, to explain, to master it. We endeavor to make the mesh ever finer and finer.” [7] This optimization process must be a relentless, vigorous assault on all hypotheses and scientific proposals bent on striking down those that miss the mark, weeding out the “unfit” hypotheses by “exposing them all to the fiercest struggle for survival” [7].

While Popper advocates for falsification, those hypotheses that escape it are not all considered equal. Those that pass the most strenuous attempts at falsification are in some sense preferred, and were said by Popper to be *corroborated.* Corroboration is not an absolute condition: a theory is either more corroborated than a competitor, or new data can further corroborate an existing theory. Corroboration is essentially a measure of the testability, or falsifiability, of a hypothesis relative to alternatives. And testability, Popper argues, is logically related to the *empirical content* of the theory: “the more a theory forbids, the more it says about the world of experience.” What this means is that theories with more universal or precise statements have greater testability because there are more opportunities for a misstep, more places for them to go wrong.

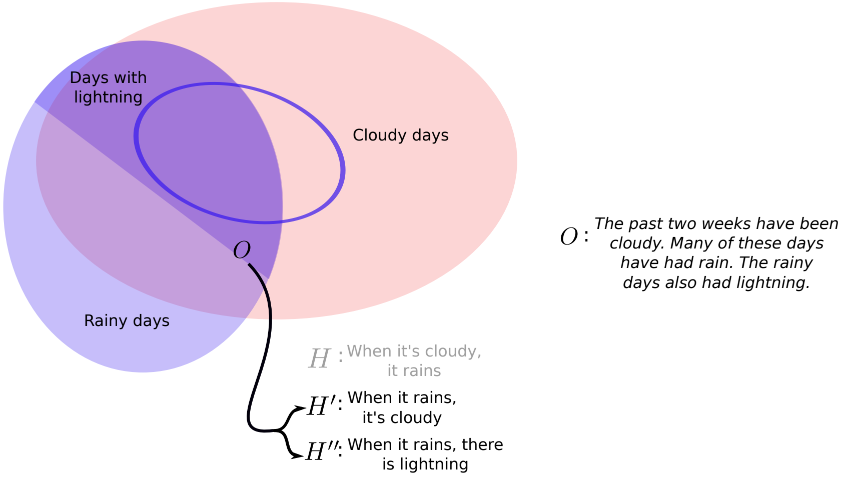
Days with lightning
Cloudy days
Rainy days
O
O : The past two weeks have been cloudy. Many of these days have had rain. The rainy days also had lightning.
H : When it's cloudy, it rains
H': When it rains, it's cloudy
H'': When it rains, there is lightning

FIG. 3. An illustration of corroboration. Both H' and H'' are consistent with the observation, but H'' is more prone to falsification than H' because its "predictive range" is smaller. Hypothesis H'' is more corroborated according to Popper.

But when all is said and done and we end up with our most stream-lined, best-fitting model, there is a strong sense that it has been given positive support by the data. Is this just an illusion? Has Popper really, truly done away with confirmation? I don't think so. Let's revisit the "experiment" conducted in Figure 2, but suppose we've collected more data (Figure 3). Furthermore, our experiment has been improved so that we can also measure the existence of lightning. Our dataset is such that we've observed some cloudy days without rain, falsifying hypothesis H, and it just so happens that every rainy day also had lightning. This latter observation supports a new hypothesis H'', that whenever it rains, there is lightning. On the question of causality, it is an alternative to H': that whenever it rains, there are clouds. Now, while both H' and H'' are consistent with the data, we see H'' is *more easy to falsify* because its "predictive range" is smaller, in fact contained within, that of H'. In other words, *fewer* observations are needed to falsify H'' than H'. H'' is therefore more corroborated than H' and by this we must mean that it gets more things right than H'. Popper called this tendency towards truth *verisimilitude* [6] and, while it is meant to apply to ultimately false hypotheses, it is an accounting of *correct* statements made by the hypothesis, *i.e.* it is an accounting of the *degree of support* that the data lends to the hypothesis. This suggests a tempered view of confirmation, one that surrenders certainty for degrees of confidence instead, in line with an inductive mode of justification. While Popper might wish to view the hypothesis that passes experimental muster as ultimately false and merely awaiting rejection, until that time we consider it the most successful generalization of observed phenomena. Popper discounts the validity of the latter view, but what else is corroboration if not tentative *acceptance*? Both refer to the *same logical relation between evidence and hypothesis*. My impression is that, despite Popper's attempts to exorcise him, Hume's specter still lurks in the idea of corroboration. Let's see then about developing a theory of confirmation.

4. Bayes' Theorem: balancing predictive success with falsifiability

Despite its murky logical pedigree, confirmation is a key part of learning. After all, some of the greatest achievements of science are unabashed confirmations, from the discovery of acquired immunity to the gauge theory of particle physics. But because we cannot isolate a unique hypothesis from the collection of a limited amount of evidence, we must entertain the possibility that several hypotheses agree with a certain experimental result. The task of confirmation is then to find the *most probable* hypothesis.

Philosopher Rudolph Carnap was the first person to take a serious crack at developing a theory of induction based on probability. The first thing he realized was that the probability of classical statistics, defined as the relative frequency of a given outcome in a long run of trials, would need to be jettisoned. Though imminently familiar, this conception of probability is not equipped to compute the chance that this or that hypothesis is true. Classical probability deals in sampling, in repeated measurements from a population of like kinds: it can determine the average number of heads in a series of coin flips, or whether smoking causes lung cancer more readily in men than women. Hypotheses are not coins or people, but singular statements like "there's a 50% chance of rain tomorrow", or wagers paying 2-to-1 odds against the home team in this Sunday's football game. These probabilities are not distilled from long-running averages; rather, they are measures of rational *degrees of belief.* Though perhaps prone to subjectivity, this new type of probability nonetheless conforms to the mathematical notion of a measure of chance. Carnap was not the first to consider probabilities as shades of belief or possibility, but he argued decisively that the degree of confirmation of a hypothesis given evidence *must* be this kind of probability. While classical probability surely has its place *within* scientific statements (in describing objective properties of physical or biological systems), a different kind of probability is needed in order to make "judgments *about* such statements; in particular, judgments about the strength of support given by one statement, the evidence, to another, the hypothesis, and hence about the acceptability of the latter on the basis of the former." [8] Carnap called this new type of probability *inductive*, or *logical* probability (or probability$_1$ in his original notation; the classical probability getting bumped to number 2, probability$_2$.) Carnap's efforts to establish a rigorous inductive logic and a quantitative measure of confirmation were significant but not without difficulty—he wound up discovering a whole bunch of *confirmation functions*, each seeking to compare the probability of H given O, $p(H|O)$, with the prior chances of H before any data is consulted, $p(H)$. The thinking is simple: if a collection of data provides positive evidence for a hypothesis then $p(H|O)$ should be larger than $p(H)$: confirmation of H would therefore be signalled by the condition $p(H|O)/p(H) > 1$. Conceptually solid, but Carnap struggled to apply it, in part due to his insistence that the probability $p(H)$ be an *objective* measure of prior knowledge [8]. This is too restrictive—a full-fledged theory of induction and confirmation will need to be subjective, more free-wheeling and less constrained.

Understanding how to use the probability $p(H)$ to determine $p(H|O)$ is key—after all, induction is all about basing future expectations on past experience. A boneheadedly simple theorem by Reverend Thomas Bayes published in the 18th-century provides a way of packaging up this past experience, à la Goodman, and using it to compute the kind of probability that Carnap indicated as a route to confirmation. Starting with the tautology $p(O \cdot H) = p(H \cdot O)$ and using the definition of conditional probability, $p(H \cdot O) = p(H|O)p(O)$, we arrive at Bayes' Theorem,

$$p(H|O, \mathcal{M}) = \frac{p(O|H, \mathcal{M})p(H|\mathcal{M})}{p(O|\mathcal{M})}. \tag{1}$$

where I have made reference to $\mathcal{M}$, the underlying model that provides the functional relationship between the observations and the hypotheses—it's role in the above expression is merely informational[5]. Bayes' Theorem facilitates the remarkable feat of transforming the probability of the evidence, O, given hypothesis, H, into the probability of the hypothesis, given the evidence. This quantity, $p(H|O, \mathcal{M})$, is referred to as the *posterior odds* of H given evidence O. There is no strict verification or falsification—Bayes' Theorem does not accept or reject H—it is a *function of H* that assigns degrees of confirmation

[5] Often when there is no risk of confusion, reference to the explicit model is suppressed; however, we *will* have cause to examine alternative models, and so we introduce this notation now.

as Carnap sought. Bayes' Theorem is not concerned with examining individual hypotheses—it provides instead a probability distribution over the entire hypothesis space. Generally, we seek the hypothesis that maximizes this function. The extent to which each hypothesis is confirmed depends on the strength of the data *relative* to what is known about H beforehand, through the *prior probability*, $p(H|\mathcal{M})$; and on possible alternative explanations, through the probability of the data irrespective of hypothesis, $p(O|\mathcal{M})$. Let's go through these quantities one-by-one.

First, the infamous prior: $p(H|\mathcal{M})$. This severely misunderstood quantity has caused much trouble for the Bayesian enterprise, often cited by dissenters as infusing what should be a concrete, no-frills computation with unnecessary subjectivity and speculation. Indeed, it does these things, but they are necessary. After all, to quote the late David MacKay, "*you can't make inferences without making assumptions!*" [9]. The prior brings all of the world's combined knowledge to bear on the assessment of a hypothesis. It's where we include previous test results and guidance from theory. The prior is general—it must incorporate all prior knowledge, widely construed, across a varied portfolio of understanding. Great theories find support in many places: our knowledge of gravity is based on observations of the changing seasons and tides, the retrograde motion of Mars, and the twists of the torsion balance in the Cavendish lab. The prior is what embodies the repeatability of science, and separates the frivolous pseudoscientific proposals from the real meat. For example, the claim that a vastly diluted glass of beet juice protects against polio has virtually zero prior support from our understanding of molecular biology or physics, whereas our prior expectations that a vaccine will yield success rests on a significant body of biological knowledge. The prior helps enforce burden of proof: claims with low prior odds (e.g. that homeopathy works) need overwhelming positive evidence to confirm with high posterior odds. The prior probability is where Goodman's green and grue emeralds get sorted out, and where Hume ultimately finds his rationality for induction—human habit.

Next, the probability of the evidence, $p(O|\mathcal{M})$ (sometimes called the *Bayesian evidence*). This is the chance that we obtain the given dataset irrespective of the hypothesis,

$$p(O|\mathcal{M}) = \sum_{H'} p(O|H', \mathcal{M})p(H'|\mathcal{M}). \tag{2}$$

It considers the possibility that other hypotheses could have given rise to the observed evidence. Consider what happens if there are many competing hypotheses, equally consistent with the data. Then, $p(O|\mathcal{M})$ is large and the chance that any one hypothesis H is *the* true one, $p(H|O, \mathcal{M})$, is small. Meanwhile, if there are *no* alternative hypotheses, we find $p(H|O, \mathcal{M}) = 1$ with deductive certainty.

Lastly, the probability $p(O|H, \mathcal{M})$: it gives the chance that we'd expect to observe the evidence, O, under the assumption that the hypothesis, H, is true. In Bayesian vernacular, it is called the *likelihood* of H. You might remember this quantity from our earlier discussion of p-values, which are given by $p(O|\overline{H})$, where $\overline{H}$ is the null hypothesis, and proceeds by illegally equating $p(O|\overline{H}) = p(\overline{H}|O)$. We have in Bayes' Theorem, by virtue of the added ingredients $p(H|\mathcal{M})$ and $p(O|\mathcal{M})$, the proper way to perform this inference. Bayes' Theorem for the case of two hypotheses, H and $\overline{H}$, is

$$\begin{aligned} p(H|O) &= 1 - p(\overline{H}|O) \\ &= 1 - p(O|\overline{H})\frac{p(\overline{H})}{p(O)} \\ &= 1 - \text{p–value} \times \frac{p(\overline{H})}{p(O|H)p(H) + p(O|\overline{H})p(\overline{H})}. \end{aligned} \tag{3}$$

When written this way, it is clear that the inference $p(H|O) = 1-$ p-value follows only if $p(\overline{H}) = 1$, the assumption that the null hypothesis is certainly true. In this case, it is impossible for H to be anything other than false, and $p(H|O) = 0$!

Now that we've introduced all the players, let's see what Bayes' Theorem can do. As an example, we'll turn to the workhorse of statistics props—the urn filled with black and white balls. Suppose we are given an urn filled with 10 balls, u of which are black and $10 - u$ white. We draw $N = 10$ balls at random, with replacement. $n_B = 4$ of these balls are black. The task is to determine the number, u, of black balls in the urn. The number, u, is the hypothesis.

First, the probability of the data given the hypothesis,

$$p(n_B|u, \mathcal{M}_1) = \binom{N}{n_B} f_u^{n_B}(1 - f_u)^{N-n_B} \tag{4}$$

follows a binomial distribution, where $f_u = u/10$. If we are totally ignorant about the number u, then our prior is noncommittal and uniform: $p(u|\mathcal{M}_1) = 1/10$ (there is either 1 black ball, 2 black balls, up to 10 black balls, each equally probable for a prior probability of 1/10). The model, $\mathcal{M}_1$, refers to our description, u, of the number of black balls in the urn, together with our prior assumptions about u, namely that any number of black balls between 1 and 10 is possible with equal chances. The Bayesian evidence accounts for the chances that we'd select n_B balls under each hypothesis, appropriately weighted by $p(u|\mathcal{M}_1)$ which is constant in this case,

$$p(n_B|\mathcal{M}_1) = \sum_u p(u|\mathcal{M}_1)p(n_B|u, \mathcal{M}_1). \tag{5}$$

Calculating the sum gives 0.00042, which will be important later. Because the Bayesian evidence and the prior probability are constants, the posterior odds are binomial distributed, proportional to Eq. (4). The posterior distribution is shown in Figure 4. As expected, the hypothesis $u = 4$ is most likely, whereas the alternative hypothesis, say, $u = 6$, has odds roughly 2-to-1 against.

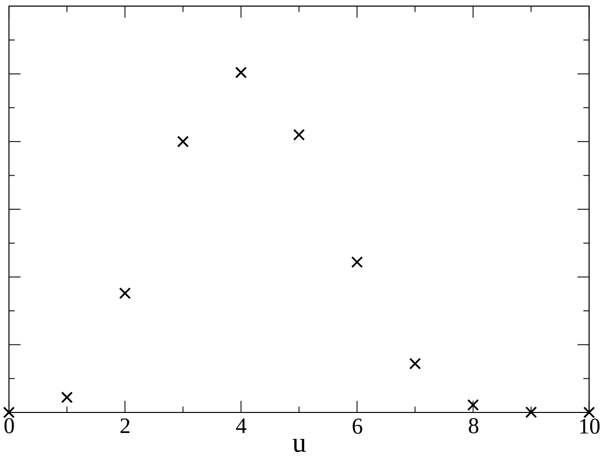

FIG. 4. Posterior odds of u, the number of black balls in the urn, given that $n_B = 4$ black balls were drawn in $N = 10$ trials with replacement.

But suppose we adopt a different prior. Say that we have some knowledge that suggests that there must be an even number of black balls in the urn. Maybe we have done some early experiments that support this assumption, or maybe we have some insight into the process by which the urns are filled[6]. Either way, it's possible to examine the exact same hypothesis, u, under a different set of prior

[6] In analogy with an actual scientific inquiry, knowing about the filling process might be like having some background knowledge from theory, or some other formal constraint on the system.

assumptions—under a new model, $\mathcal{M}_2$. Now $p(u|\mathcal{M}_2) = 1/5$ for u even, and $p(u|\mathcal{M}_2) = 0$ for u odd. The Bayesian evidence is the sum Eq. (5) but now only over even u, giving 0.00054. When we plot $p(u|O, \mathcal{M}_2)$, we find that the posterior odds are larger (for u even) than the odds under the original prior, Figure 5. What might this mean? Is Bayes' Theorem telling us that the hypothesis $u = 4$ under the *second* model is preferred by the data? Not necessarily: it's generally possible to arbitrarily increase the posterior odds of a hypothesis by making that hypothesis more and more sophisticated. This is called over-fitting the data, and so the posterior probability distribution cannot be trusted with this determination.

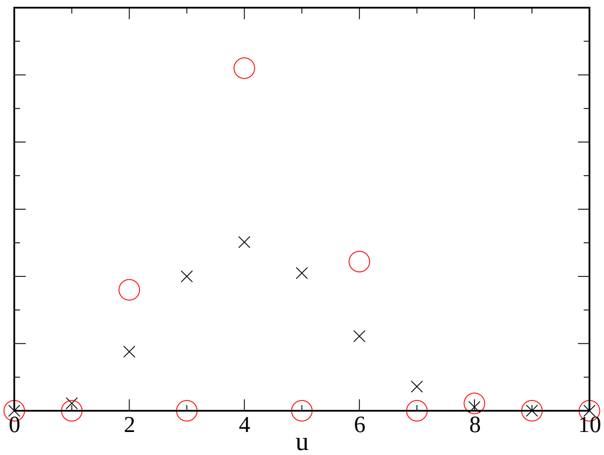

FIG. 5. Posterior odds of u, the number of black balls in the urn, given that $n_B = 4$ black balls were drawn in $N = 10$ trials with replacement. Black crosses assume uniform prior over all u, red triangles assume u even.

To investigate this question more carefully, we can apply Bayes' Theorem to the models themselves,

$$p(\mathcal{M}|O) = \frac{p(O|\mathcal{M})p(\mathcal{M})}{p(O)}. \tag{6}$$

Bayes' Theorem not only calculates posterior odds over the space of competing *hypotheses* (in this case, the value of u), it also yields posterior odds over the space of competing models, $\mathcal{M}_i$. And notice something marvelous: the likelihood of the model, $p(O|\mathcal{M})$, is the denominator of Eq. (1)—it is the Bayesian evidence! Unless we have strong *a priori* feelings about the various choices of model, then $p(\mathcal{M}_1) = p(\mathcal{M}_2)$ and the likelihood, and therefore the Bayesian evidence, is a direct measure of the posterior odds of the model. If we revisit the calculations of $p(u|O)$ above, the evidence under $\mathcal{M}_1$ is 0.00042, while the evidence under $\mathcal{M}_2$ is 0.00054. Indeed, $\mathcal{M}_2$ *is* preferred by the data. What is it about the second prior that results in a better "fit"? It's helpful to visualize the sample space of the urn in order to compare the two models. In Figure 6, each configuration is possible, with equal weight, according to the prior of $\mathcal{M}_1$, while only those in gray boxes—the even numbers of black balls—are possible under $\mathcal{M}_2$. Bayesian inference has a thing for minimal priors, both in the number of free parameters and the intervals over which they can range. The idea is that a model with a more economical parameter space is more exacting in its range of predictions, and hence, more easily falsified. This is exactly the sentiment behind Karl Popper's notion of corroboration that was illustrated in Figure 3. Recall the importance that Popper ascribed to the empirical content of a hypothesis: the more lenient the hypothesis, the less it charges about the system under study. Popper argued that empirical content is inversely related to what he called the *logical probability* of the hypothesis [7], the idea being that we should select the least likely hypothesis given the data. This obviously goes against the principles of Bayesian theory, which seeks the hypothesis with the greatest posterior odds. And so a sizeable rift opened across the field

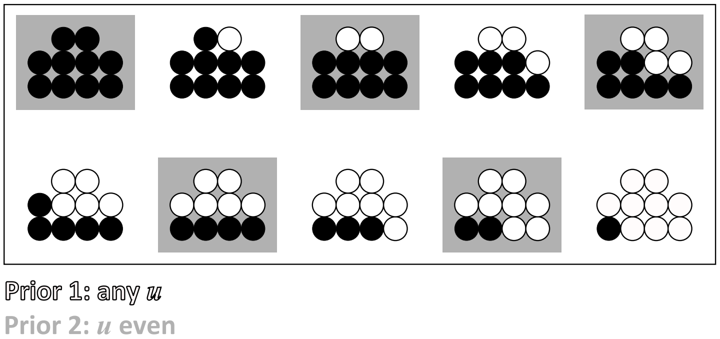

FIG. 6. Graphical depiction of the configuration space of the urn problem: prior 1 includes all configurations, prior 2 includes only those with an even number of black balls.

of scientific inference, with two opposing notions of what makes a hypothesis favorable. Both have an air of correctness about them: we do want our hypotheses to be precise and daring, but we also want them to be well-supported by the data. Can't we have it both ways? I believe we can. Popper isn't talking about the posterior odds, $p(H|O)$, when he argues for *low* probability. Rather, he has in mind what Figure 6 captures: that if we were to throw a dart at the prior sample space, we'd be less likely to hit those urn configurations belonging to model $\mathcal{M}_2$. It is in the *prior space* that we seek Popper's low logical probability, not in the posterior odds of a particular hypothesis once the data arrive[7]. So Popper's corroboration is really a model comparison ($\mathcal{M}_1$ vs. $\mathcal{M}_2$) rather than a measure of support for a particular hypothesis within a given model (the value of u within $\mathcal{M}_1$, for example). Bayes' Theorem is one-stop-shopping for both traditional confirmation and corroboration.

In conclusion, Bayes' Theorem is the thread that connects the key aspects of prior knowledge, predictability, and goodness-of-fit to render a measure of confirmation—a posterior probability and corroboration of hypotheses. It is arguably our best method of scientific inference. Written down by Thomas Bayes in the 18th century, it anticipated much of the work on induction and falsifiability that would challenge the greatest philosophical minds of the past three centuries: Hume, Carnap, Reichenbach, Quine, Hemple, Goodman, Popper, and so many more. Though incomplete, the various elements of Bayes' Theorem have obvious champions:

$$\overbrace{p(H|O)}^{\text{Carnap}} = \frac{\overbrace{p(O|H)}^{\sim\text{p–value}}\ \overbrace{p(H)}^{\text{Hume}}}{\underbrace{p(O)}_{\text{Popper}}}. \tag{7}$$

Because of the problem of induction, we will never be certain of our scientific assertions or conclusions. In fact, science is an enterprise that is *perpetually imperfect* in its compiled knowledge about the world because it must make guesses based on sometimes tenuous patterns and regularities in data sets that are always incomplete. Practical science proceeds by establishing theories as summaries of these regularities, but they are only approximations of the true structure. Eventually, with hope, anomalies or systematics

[7] Popper also does not mean that the prior probability should be low, although he has made various arguments to this effect [6, 7]. The prior probability of u under $\mathcal{M}_2$ is *larger*, $p(u|\mathcal{M}_2) = 1/5$, than under $\mathcal{M}_1$, $p(u|\mathcal{M}_1) = 1/10$. This is because the prior probability is equally distributed across fewer states in the case of $\mathcal{M}_2$. The logical probability is based on the percentage of the configuration space permitted by the given model: $\mathcal{M}_2$, with half as many available configurations as $\mathcal{M}_1$, has half the logical probability.

will emerge in new data. Tantalizing and maybe a little maddening at first, they will reveal hints of new patterns and symmetries signaling a need to adjust the theory—maybe we falsify it, or adjust it. The new theory that emerges is also transitional and ultimately wrong, but it is confirmed in the sense that it assumes the role of the new foundation, the next scaffold in the rise to truth. It is selected just as a favorable trait in an evolving species—both seek to an optimized form. As Popper said so well, "It is not his *possession* of knowledge, of irrefutable truth, that makes the man of science, but his persistent and recklessly critical *quest* for truth."

[1] *An Enquiry Concerning Human Understanding*, David Hume (1748).

[2] *Choice & Chance: An Introduction to Inductive Logic*, Brian Skyrms, Dickenson Pub. Co. (1986).

[3] *The Foundations of Scientific Inference*, Wesley Salmon, Univ. of Pittsburgh Press (1979).

[4] *Fact, Fiction, and Forecast*, Nelson Goodman, Harvard University Press (1954)

[5] Natural Kinds, W. V. Quine, in *Ontological Relativity & Other Essays*, Columbia University Press (1969)

[6] *Conjectures and Refutations: The Growth of Scientific Knowledge*, Karl Popper, Routledge (1963).

[7] *The Logic of Scientific Discovery*, Karl Popper, Routledge (1934).

[8] Statistical and Inductive Probability, Rudolph Carnap, in *Philosophy of Probability: Contemporary Readings*, Routledge (1955)

[9] *Information Theory, Inference, and Learning Algorithms*, David MacKay, Cambridge (2003).