Ockham's Razor and Bayesian Analysis

The intuitive idea that simple explanations are usually better than complicated ones now gets quantitative support from statistical methods

William H. Jefferys and James O. Berger

The principle known as Ockham's razor has high standing in the world of science, buttressed by its strong appeal to common sense. William of Ockham, the 14th-century English philosopher, stated the principle thus: *Pluralitas non est ponenda sine necessitate*, which can be translated as: "Plurality must not be posited without necessity." It is not entirely certain what Ockham meant by this rather opaque saying, but later versions of the principle, which have been traced to various authors other than Ockham, have a clear enough interpretation. The idea has been expressed as "Entities should not be multiplied without necessity" and "It is vain to do with more what can be done with less"; a modern rendering might be "An explanation of the facts should be no more complicated than necessary," or "Among competing hypotheses, favor the simplest one." Over the years Ockham's razor has proved to be an effective device for trimming away unprofitable lines of inquiry, and scientists use it every day, even when they do not cite it explicitly. See Thorburn (1918) for a history of the principle.

Ockham's razor is usually thought of as a heuristic principle—a rule of thumb that experience has shown to be a useful tool, but one without a firm theoretical or logical foundation. Under some circumstances, however, Ockham's razor can be regarded as a consequence of deeper principles. Specifically, it has close connections to the Bayesian method of statistical analysis, which interprets a probability as the degree of confidence or plausibility one is willing to invest in a proposition.

Ockham's razor enjoins us to favor the simplest hypothesis that is consistent with the data, but determining which hypothesis is simplest is often no simple matter. Bayesian analysis can offer concrete help in judging the degree to which a simpler model is to be preferred. Ironically, whereas Bayesian methods have been criticized for introducing subjectivity into statistical analysis, the Bayesian approach can turn Ockham's razor into a *less* subjective and even "automatic" rule of inference.

Galileo's Problem

The connection between Bayesian statistics and Ockham's razor is implicit in the work of Harold Jeffreys of the University of Cambridge, whose book *Theory of Probability*, published in 1939, was an important landmark in the modern revival of Bayesian methods. The connection has since been made explicit by a number of others: see Good (1968, 1977), Jaynes (1979), Smith and Spiegelhalter (1980), Gull (1988), Loredo (1989) and MacKay (1991).

An example that Jeffreys discussed in 1939 provides an illuminating introduction to the problems that can arise when Ockham's razor is put to the test as an implement of scientific methodology. Suppose you are collecting some data on the motion of falling bodies, as Galileo supposedly did in his legendary experiments at the Tower of Pisa. You drop a weight and record its position, s, at several moments, t, during the fall. The challenge then is to devise a mathematical law describing the motion.

The law proposed by Galileo, and familiar to students of physics, can be expressed as a quadratic equation:

$$s = a + ut + \tfrac{1}{2}gt^2$$

Here a, u and g are adjustable parameters, or in other words constants that can be assigned arbitrary values in order to fit the empirical data. (In this case a is interpreted as the initial position of the falling object, u is the initial velocity, and g is the acceleration due to gravity.) There are straightforward methods for finding values of a, u and g that minimize some measure of the error between the predicted and the observed positions of the body. If Galileo's task is merely to identify those optimum parameter values, then the problem is a standard exercise in estimation theory.

But Galileo did not have to confine his attention to quadratic laws. He could instead have proposed a cubic equation, such as

$$s = a + ut + \tfrac{1}{2}gt^2 + bt^3$$

where the coefficient b is a fourth adjustable parameter. And of course there is no reason to stop with cubic polynomials. By adding further terms the equation could be extended to fourth, fifth or sixth powers of t. Indeed, an infinite sequence of equations could be formed in this way. Why is it, then, that the quadratic law is the choice of physicists everywhere?

The answer is not that a quadratic law offers closer agreement with the empirical data. On the contrary, for any given data set, going to a higher-degree polynomial can always reduce the total error (unless the fit is already perfect). If there are n measured data points, then an equation of degree $n - 1$ specifies a curve that can be made to pass through all of the data points exactly, so that the measured error is zero. Thus there must be something other than accuracy in fitting data that leads people to prefer the quadratic law over any of the higher-degree equations.

One possible explanation is that any coefficients beyond a, u and g are generally very small, so that higher powers of t contribute little to the structure of the physical law. Another interesting point is that even when a high-degree equa-

William H. Jefferys is the Harlan J. Smith Centennial Professor of Astronomy at the University of Texas at Austin. He received his Ph.D. in astronomy from Yale University in 1965. James O. Berger is the Richard M. Brumfield Distinguished Professor of Statistics at Purdue University. He received his Ph.D. in mathematics from Cornell University in 1974. Address for Jefferys: Department of Astronomy, RLM 15.308, University of Texas, Austin TX 78712. Internet: bill@bessel.as.utexas.edu

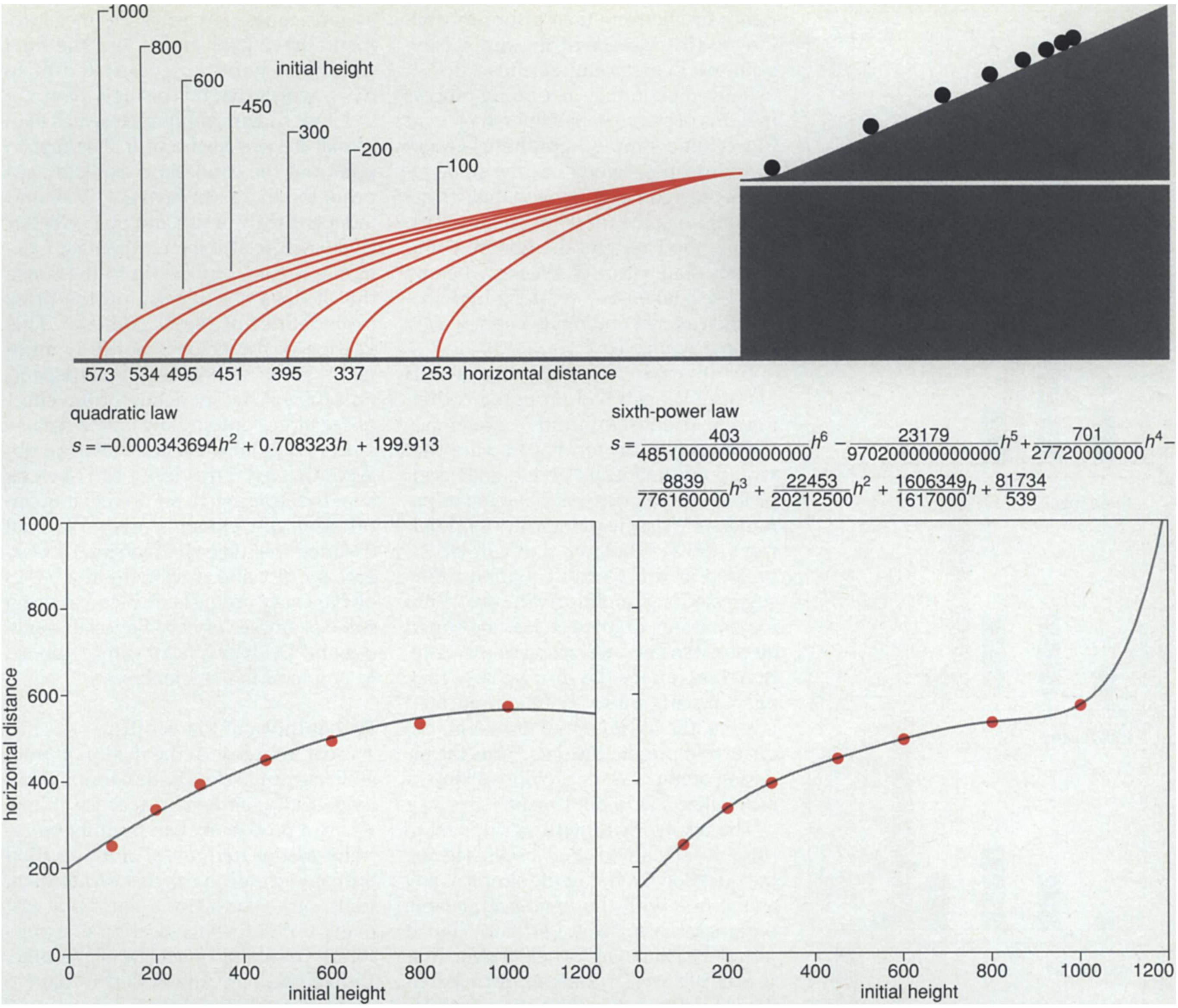

Figure 1. Experiment conducted by Galileo in 1608 offers an illustration of how Ockham's razor and Bayesian analysis can aid scientific inference. Galileo's demonstration that a ballistic trajectory is a parabola relied on experiments with a ball rolling down an inclined plane and then continuing in free-fall. He released the ball at various heights on the plane and measured the horizontal distance it flew. Data recorded during some of these experiments were rediscovered in the 1970s by Stillman Drake of the University of Toronto (Drake and MacLachlan 1975). The seven numbers labeled "horizontal distance" on the diagram above appear on a similar sketch in one of Galileo's notebooks; the corresponding initial heights were inferred from a reconstruction of the experiment. The challenge for the modern analyst, as for Galileo, is to deduce a mathematical law giving the horizontal distance *s* as a function of the initial height *h*. Two candidate laws are shown here. A quadratic law offers a good approximation to the data, but a sixth-degree polynomial is even more accurate: It fits the seven data points exactly. Nevertheless, the higher-degree equation is not the preferred physical law. One weakness of the sixth-degree equation is that it makes reliable predictions only in the immediate vicinity of the data points. Whereas the quadratic law extrapolates reasonably well, the predictions of the sixth-degree law for large values of *h* are implausible. A more fundamental objection to the sixth-degree law is that it is unnecessarily complicated.

tion fits a given set of data exactly, the equation may do very poorly as a predictor of new data. For example, given seven experimental measurements, a sixth-degree polynomial can fit the data exactly, whereas a quadratic equation will generally have some residual error; but if additional measurements are made, perhaps at larger values of t, the higher-degree law is likely to yield much larger errors than the quadratic one. Looked at another way, a single quadratic law can explain a variety of data sets reasonably well, whereas many data sets would require quite different sixth-degree polynomials.

These observations might well serve as an after-the-fact justification for rejecting a law of accelerated motion based on a sixth-degree equation, but they fail to account for a more fundamental fact: Neither Galileo nor a modern student of physics would even consider a sixth-degree equation in the first place. They would favor the quadratic law because is it simpler, whereas all higher-degree polynomials are *unnecessarily* complicated.

Probabilities Prior and Posterior

Jeffreys suggested that the reason for favoring the simpler law is that it has a higher *prior probability*; in other words, it is considered the likelier explanation at the outset of the experiment, before any measurements have been made. This is certainly a reasonable idea. Scientists know from experience that Ockham's razor works, and they reflect this expe-

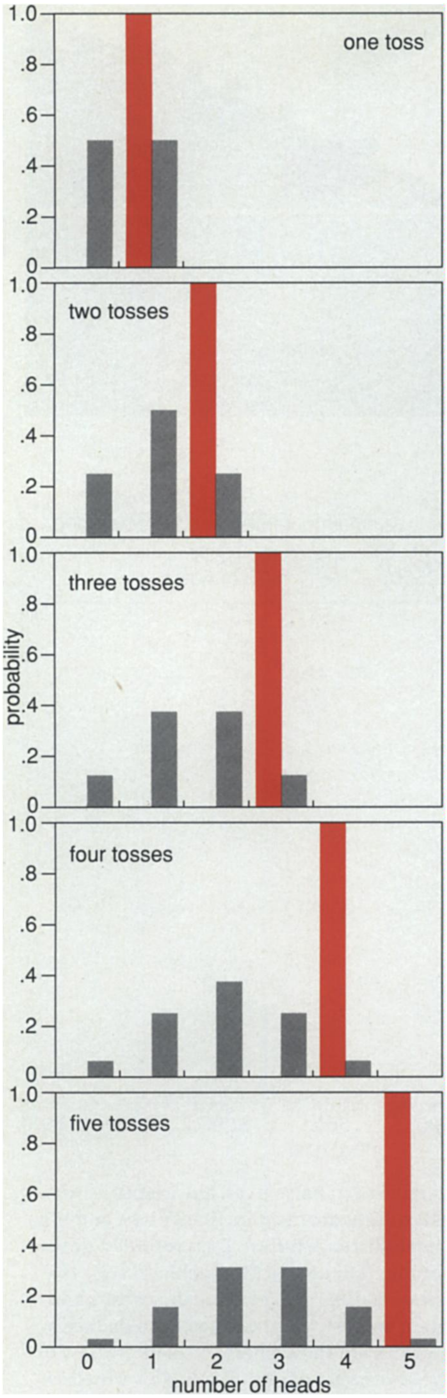

Figure 2. Series of coin tosses can be explained by either of two hypotheses: that the coin is fair or that it has two heads. Depending on who is doing the tossing, the latter hypothesis may initially be accorded a low probability, but if heads appears invariably in a long series of tosses, the hypothesis of a rigged coin becomes more attractive. These graphs show the predictions of the fair-coin hypothesis *(gray)* and the two-headed-coin hypothesis *(red)* for various numbers of tosses. The fair-coin hypothesis is consistent with every conceivable observation; the two-heads hypothesis, in contrast, would be falsified by a single appearance of tails. Because the hypothesis of fraud makes such sharp predictions, it is given greater credence when those predictions come to pass.

rience by choosing their prior probabilities so that they favor the simpler hypothesis. Even though scientists do not usually explain their reasoning process in terms of prior probabilities, they tend to examine simple hypotheses before complex ones, which has the same effect as assigning prior probabilities according to some measure of simplicity. The method reflects the tentative and step-by-step nature of science, whereby an idea is taken as a working hypothesis, then altered and refined as new data become available.

In an earlier work Jeffreys and Dorothy Wrinch had proposed codifying the scientist's intuitive preference for simplicity in terms of a rule that would automatically give higher prior probability to laws that have fewer parameters (Wrinch and Jeffreys 1921; Jeffreys 1939). For laws that can be expressed as differential equations, they suggested a straightforward algorithm for counting parameters. Having sorted all possible laws according to this criterion, one can try the simpler laws first, only moving on to more complicated laws as the simple ones prove inadequate to represent the data. Thus the ordering of hypotheses provides a kind of rationalized Ockham's razor.

The trouble with Jeffreys's appeal to prior probabilities is that it seems to beg the question. Defining the simplest law as the one with the fewest adjustable parameters is a useful strategy, but it cannot be extended to yield a clear, universal rule for assigning prior probabilities, as Jeffreys himself points out (Jeffreys 1939, page 49). He writes: "I do not know whether the simplicity postulate will ever be stated in a sufficiently precise form to give exact prior probabilities to all laws; I do know that it has not been so stated yet. The complete form of it would represent the initial knowledge of a perfect reasoner arriving in the world with no observational knowledge whatever." Needless to say, no real scientist qualifies as such an unbiased perfect reasoner.

But Jeffreys also suggested a measure of simplicity that does not depend on prior probabilities; instead it is grounded in tests of statistical significance. Basically, if a law has many adjustable parameters, then it will be significantly preferred to the simpler law only if its predictions are considerably more accurate. Indeed, if the predictions of the two models are roughly equivalent, the simpler law can have greater *posterior* probability (the probability an observer assigns to the law after the measurements have been made and the data collected). Jeffreys never stated in so many words that this result is a form of Ockham's razor, although it seems likely that he was aware of it. The first to point out the connection explicitly appears to have been Jaynes (1979), and independently Smith and Spiegelhalter (1980), who called it an "automatic Ockham's razor," automatic in the sense that it does not depend on the prior probabilities of the hypotheses. This version of the razor is not fully automatic, however, because it does depend on probabilistic modeling of the effect of the more complex law on the data.

In Berger and Jefferys (1992) we observe that even this input can often be avoided, leading to an objective quantification of Ockham's razor. We shall describe this objective version of Ockham's razor after reviewing the basics of Bayesian analysis and considering some examples of how Bayesian methods and Ockham's razor can be applied to problems in the sciences.

Probability and Plausibility

The earliest ideas in the theory of probability arose to deal with various problems in the mathematics of gambling, where a probability can usefully be defined as the frequency of a specified outcome in a long series of identical trials. For example, if a fair die is cast many times, the face bearing four pips comes up about one-sixth of the time, and so this outcome is said to have a probability of one-sixth. This *frequentist* formulation of probability theory works well in many contexts, but there are also questions it cannot readily answer. For example, a geologist might ask: What is the probability of an earthquake, given certain precursory seismic signals? Obviously, it is not possible to calculate this probability by performing many trials under identical conditions.

In Bayesian analysis, *probability* is used in another sense: A probability is a measure of the plausibility of a hypothesis or proposition. This alternative definition is particularly useful in the sciences. When a paleontologist states that the dinosaurs *probably* died out as a result of climatic change, or when an astronomer says that Mars is *probably* lifeless, the probabilities cannot readily be understood as frequencies, but they have a natural interpretation as indicating the speaker's degree of belief or confidence in the statement, given the available evidence.

student \ question number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
key	a	d	b	d	b	d	c	d	d	e	d	d	c	d	d	a	d	d	c	c	b	d	d	e	c	b	b	a	a	a
1	b	d	b	b	d	d	c	b	b	e	d	a	c	d	d	a	d	e	a	c	b	b	a	c	c	c	b	c	d	c
2	a	d	b	d	b	d	c	a	d	e	b	a	c	e	d	a	d	a	a	c	e	b	a	a	c	b	b	c	a	a
3	a	d	b	b	b	d	c	d	d	e	d	a	c	d	e	a	d	d	a	c	a	b	b	a	c	c	b	d	a	c
4	a	d	b	b	b	d	c	a	d	e	d	a	c	a	d	a	d	c	a	c	b	b	b	a	c	b	b	c	a	d
5	a	d	b	d	b	d	c	d	d	e	d	d	c	d	d	a	d	d	c	c	b	d	d	e	c	b	b	a	a	a
6	a	c	b	b	b	d	a	a	a	b	a	a	c	d	d	a	d	d	a	c	a	a	b	a	c	b	b	c	a	c
7	a	d	b	b	b	a	c	a	d	e	d	a	c	d	d	a	d	d	a	c	b	b	c	b	c	b	b	c	a	c
8	b	d	b	b	b	d	c	a	d	d	b	a	a	d	b	a	d	a	a	c	a	b	b	a	c	b	b	c	c	b
9	a	d	b	c	a	d	c	a	d	e	d	a	c	d	d	a	d	d	a	c	d	b	b	a	c	b	b	c	a	c
10	a	d	b	d	b	d	c	d	d	e	d	d	c	d	d	a	d	d	c	c	b	d	d	e	c	b	b	a	a	a
11	a	a	b	b	b	d	c	c	d	e	a	a	c	d	d	a	d	a	e	c	b	b	b	a	c	b	b	a	a	c
12	a	d	a	d	b	d	c	a	d	e	d	a	c	a	d	a	b	a	a	c	b	a	b	c	c	b	b	c	d	a
13	a	b	c	d	d	d	b	a	d	d	d	a	c	d	d	a	d	a	a	c	a	b	b	a	c	a	b	c	a	c
14	a	d	b	c	a	d	c	a	d	e	d	a	c	d	d	a	d	d	a	c	d	b	b	a	c	b	b	c	a	c

Figure 3. Detecting plagiarism on a multiple-choice examination is a more serious challenge to Bayesian analysis. Presented here are the answers of 14 hypothetical students to a 30-question test. Each answer is color-coded for clarity; correct answers are shown in white and incorrect answers in black. Similarities in the answers of two students could be explained by either of two hypotheses: coincidence or cheating. As in the case of the coin-tossing experiment, the hypothesis of cheating makes sharper predictions; coincidence can explain anything at all. The analysis is complicated, however, because the answers to each question have different probabilities. Students 5 and 10, for example, should not be accused of collusion even though their answers are identical: They both have a perfect score. But two other students in this sample might be viewed with suspicion. David Harpp and James Hogan of McGill University have written a computer program to perform such analysis.

The foundation of Bayesian statistics is a theorem proved by the Rev. Thomas Bayes, an English clergyman and amateur mathematician, in 1761, the year of his death; the proof was published posthumously (Bayes 1763). At its core, Bayes's theorem represents a way—Bayesians would argue the most consistent way—of incorporating new data into your understanding of the world.

Suppose you have a series of hypotheses about some natural phenomenon. The hypotheses are known to be mutually exclusive and exhaustive, so that exactly one hypothesis must be true. Based on all the information available to you, you assign each hypothesis a probability. These are the prior probabilities mentioned above in connection with Galileo's experiment. Now suppose some new item of data comes to your attention, such as the result of an experiment. The question is: How should you revise the probabilities you ascribe to the various hypotheses in light of the new data? Bayes's theorem offers a mathematical procedure for answering this question.

The notation $P(X \mid Y)$ represents a conditional probability: the probability that hypothesis X is true, given the available information Y. With probabilities expressed in this way, Bayes's theorem can be stated as follows:

$$P(H_i \mid D\&I) = \frac{P(D \mid H_i\&I)\, P(H_i \mid I)}{P(D \mid I)}$$

This equation can be used to calculate $P(H_i \mid D\&I)$, or the probability that H_i is true, given both the prior information I and the new data D. Three factors enter into the calculation. $P(H_i \mid I)$ is the prior probability ascribed to hypothesis H_i, or in other words the probability of H_i given the initial information I. $P(D \mid H_i\&I)$ is the probability of observing the new data D, given the initial information I and assuming that H_i is true. Finally, $P(D \mid I)$ is the total probability of observing D given I, no matter which of the hypotheses turns out to be true. Thus the final probability of H_i given both D and I increases if the prior probability of H_i increases or if D is more strongly predicted by H_i and I. Conversely, the final probability of H_i is reduced if D is predicted more generally, by all possible hypotheses.

The use of Bayes's theorem in statistical and scientific reasoning has had a long and controversial history; see Edwards, Lindman and Savage (1963) or Berger (1985) for discussions of the controversies. There are two main points of contention between Bayesians and traditional (frequentist) statisticians. The first is philosophical: Some argue that since only one of the hypotheses H_i can be true, it makes no sense to talk about the "probability" that H_j is true. This has a certain logic if one interprets probabilities as frequencies, but the objection is beside the point if "probability" refers to the degree of plausibility of a hypothesis. This is the way most working scientists use the term.

The second point is that there are no universally accepted ways of assigning the prior probabilities $P(H_i \mid I)$ that Bayes's theorem requires. Hence different scientists, faced with the same data, may come to different conclusions. Bayesians have several responses to this complaint. One school believes that there is nothing inherently wrong with subjectivism, and, indeed, that the frequentist approach is really no more objective, although it has successfully disguised this fact (Berger and Berry 1988). Subjectivist Bayesians point out that it is common for scientists to disagree about the plausibility of hypotheses, and contend that this is a natural, and indeed inescapable, state of affairs.

Another school (Laplace 1812, Jeffreys 1939) has developed methods of choosing and utilizing "objective" prior probability distributions for a wide class of problems. With problems for which such methods are available, Bayesian analysis can claim to be as objective as any other statistical method. Still, there remain problems for which these objective methods do not work. Some of the examples discussed below fall into this troublesome class.

To Catch a Cheat

The key idea linking Bayesian analysis to Ockham's razor is the notion of simplicity in a hypothesis. In quantifying this notion, it is useful to observe that a simpler hypothesis divides the set of observable outcomes into a small set that has a high probability of being observed and a large set that has a small probability of being observed; the more complex hypothesis tends to spread the probability more evenly among all the outcomes. Thus the simpler hypothesis makes sharper predictions about what data will be observed, and it is more readily falsified by arbitrary data. In the case of Galileo's problem, the more complex hypotheses have more parameters, which can be adjusted to accommodate a larger range of data. In other cases, the number of adjustable parameters is not at issue, but, nonetheless, one hypothesis restricts the possible outcomes of an experiment more than another does.

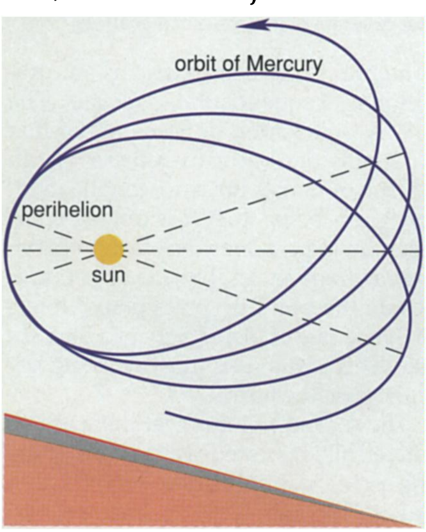

Figure 4. Anomaly in the orbit of Mercury was the subject of a celebrated controversy in the 1920s, which might have been settled by Bayesian reasoning. As observed from the earth, Mercury's perihelion, or point of closest approach to the sun, appears to advance slightly on each of the planet's orbits. The total advance is 5,599 arc-seconds per century, or about a degree and a half. Of this amount some 5,025 arc-seconds results from the precession of the equinoxes on the earth *(orange)*, and another 531 arc-seconds can be attributed to the gravitational influence of the other planets on Mercury's motion *(gray)*. That leaves something more than 40 arc-seconds per century in need of explanation *(red)*. Angles in this diagram are greatly exaggerated.

Suppose a friend who has a reputation as a prankster offers to flip a coin to decide who will perform a little chore: heads he wins, tails he loses. Knowing your friend's reputation, you might well be concerned that he would use trickery (perhaps a two-headed coin) to win the toss. The hypothesis H_{HH} that the coin has two heads is, under this understanding, a simpler one than the hypothesis H_{HT} that the coin is fair. In a series of many coin tosses, H_{HH} will be falsified if tails comes up even once, whereas any sequence of heads and tails could arise under H_{HT}.

Before the coin is flipped, you might believe that the hypotheses H_{HH} and H_{HT} are equally likely. Then the coin is tossed, and it indeed comes up heads. Your degree of belief in the two hypotheses will change as a result of this information, and (by Bayes's theorem) the posterior probability that you assign to H_{HH} should now be twice what you assign to H_{HT}. Still, the evidence that your friend is trying to fool you is not very strong at this point, perhaps not strong enough to challenge him for a close look at the coin. On the other hand, if the coin comes up heads on five occasions in a row, you will be rather inclined to think that your friend is playing a joke on you. Even though both hypotheses remain consistent with the data, the simpler one is now considerably more credible.

In the days before electronic computers, when publishing mathematical tables was still a viable business, the compiler of a table had to contend with possible copyright infringement. If someone published a table identical to your own work, how could you demonstrate to the satisfaction of a court that the new table was copied from yours rather than calculated *de novo*? To guard against plagiarism, compilers frequently took advantage of the fact that numbers ending in the digit 5 can be rounded either up or down without significantly altering the result of a calculation. By rounding such numbers randomly, the compiler could embed a secret code in the table that identified the table as his work, while not significantly affecting the accuracy of the results obtained when using the table.

For example, suppose you published a table of sines with 1,000 entries. You calculated each value to five decimal places, then rounded to four places. About 100 of the entries would have ended in the digit 5 and would have been rounded either up or down at random. Another compiler of a table would be very unlikely to happen on the same pattern of rounding, since there are 2^{100}, or approximately 10^{30}, ways to round the 5s in the table.

If you learn that a newly published table has the same rounding pattern as your own, Bayesian analysis can quantify your suspicions of plagiarism. Let H_P be the hypothesis that the second table was plagiarized from the yours, and H_I be the hypothesis that the second table was generated independently and just happens to have the same pattern of roundings. On the data D that the rounding patterns are identical, we can calculate that $P(D \mid H_P) = 1$ and $P(D \mid H_I) = 2^{-100}$. Assuming equal prior probabilities for the two hypotheses, Bayes's theorem shows that the posterior probability of plagiarism differs only negligibly from 1.

The reason for this clear outcome is that H_P makes a precise prediction about what will be seen, and is inconsistent with almost all possible data, whereas H_I is consistent with any observation. H_I "hedges its bets" by trying to accommodate all possible data; in contrast, H_P risks everything on a single possibility. As a result, when that single possibility turns out to be true, H_P is rewarded for the greater risk it takes by being given a very high posterior probability compared to H_I, even though H_I is also consistent with the data.

It is now routine for authors of directories, maps, mailing lists and similar compilations to deliberately introduce innocuous errors into the material. When plagiarism or other unauthorized use of the material takes place, the presence of these errors in the copied material serves as very strong evidence of copyright violation.

David Harpp and James Hogan of McGill University have used a similar idea to detect cheating on multiple-choice tests. They wrote a computer program to compare the answers given by each pair of students in the class and look for a near-match between correct and incorrect answers. Of course, as teachers they hope and expect students to know the subject material, so that conclusions about cheating cannot be drawn from a student's correct answers. But if two students make the same errors, the evidence of cheating can be compelling. The analysis of the data in this problem is more complicated than it is in the case of a plagiarized mathematical table because different questions are answered incorrectly with differing frequencies and because the various incorrect responses for each question can be expected to draw different numbers of responses. But there are practical solutions for these complications.

Another application of this principle comes from evolutionary biology. When the DNA of two organisms is compared, similarities in sequence can be taken as evidence of descent from a common ancestor. For DNA within a functioning gene, however, the strength of such evidence is compromised, because the nucleotide sequence could not diverge too far without impairing the function of the gene product. This constraint is removed in the case of a pseudogene, which is a region of DNA that has most of the characteristics of a gene but because of some defect does not give rise to a functioning protein or other product (Max 1986, Watson et al. 1988). A pseudogene can be passed on to an organism's progeny, even though it has lost its function. If two species have identical or nearly identical pseudogenes (as human beings and chimpanzees do, for example), this constitutes very powerful evidence in favor of the hypothesis that the species have a common ancestor. Just as with cheating on multiple-choice tests, or plagiarism of compiled materials, it is the verbatim or near-verbatim repetition of a "mistake" that gives the hypothesis of copying—in evolutionary terms, descent from a common ancestor—a high posterior probability.

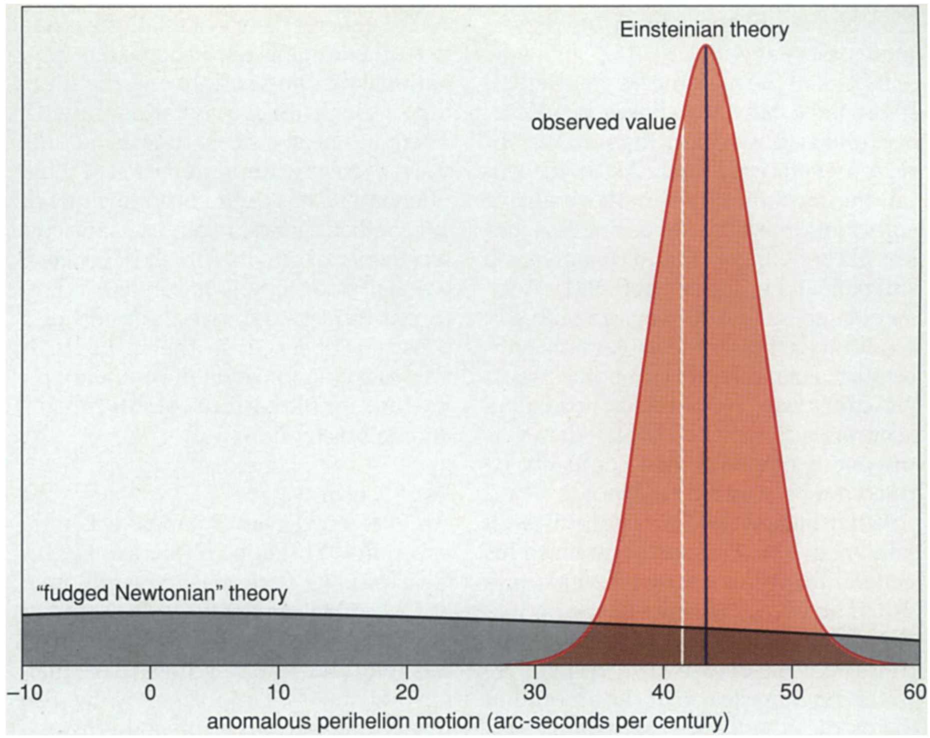

Figure 5. Two competing explanations of Mercury's anomalous motion can be evaluated through an application of Ockham's razor. Einstein's general theory of relativity makes a sharp prediction that the perihelion advance is equal to 42.9 arc-seconds per century *(purple line)*. A "fudged Newtonian" theory, on the other hand, can be adjusted to accommodate almost any observation. In this analysis the predictions of the fudged Newtonian theory are modeled by a normal distribution with a mean of zero and a standard deviation of 50.04. The actual value of the anomalous advance—as measured in the 1920s—is 41.6 arc-seconds per century *(white line)*. This observation is consistent with either hypothesis, but the much narrower probability distribution for Einstein's theory favors it by a ratio of 28.6 to 1.

A Planetary Puzzle

Our last example illustrating the predictive power that simplicity confers on a hypothesis merits somewhat more detailed analysis. It concerns a celebrated controversy in astronomy and celestial mechanics.

Beginning with the work of the French astronomer Urbain Leverrier in the 1840s, astronomers were aware of a serious problem in explaining the motion of the planet Mercury. Newtonian theory, which had been extraordinarily successful in accounting for most of the motions in the solar system, had run up against a small discrepancy in the motion of Mercury that it could not explain easily. After all the perturbing effects of the other planets had been taken into account, there remained an unexplained residual motion of Mercury's perihelion (the point in its orbit where the planet is closest to the sun) in the amount of approximately 43 seconds of arc per century.

It seemed something had been overlooked. One appealing possibility was the proposal that another planet might exist, closer to the sun than Mercury. Leverrier himself, along with the English astronomer John Couch Adams, had recently met with brilliant success by predicting that a previously unknown planet was responsible for discrepancies in the motion of Uranus. When Johann Gottlieb Galle, a young astronomer at the Berlin Observatory, looked where Leverrier suggested, the planet Neptune was discovered. It seemed possible that a similar phenomenon might explain the anomaly in Mercury's motion.

A number of astronomers duly set out to find the new planet, dubbed Vulcan in anticipation of its discovery, and some sightings were announced. The sightings could not be confirmed, however, and over time interest in the Vulcan hypothesis waned.

Other mechanisms were also proposed. It was suggested that rings of material around the sun could produce the observed effect; or the sun might be slightly oblate, due to its rotation on its axis; or, finally, the Newtonian law of gravitation might not be exactly right. For example, the American astronomer Simon Newcomb (1895) proposed that the exponent in Newton's law of gravitation might not be exactly 2, but instead might be $2 + \varepsilon$.

All of these hypotheses had one characteristic in common: They had parameters that could be adjusted to agree with whatever data on the motion of Mercury existed. In modern parlance, we would call the presence of such parameters a "fudge factor." The Vulcan hypothesis had the mass and orbit of the putative planet; the ring hypothesis had the mass and location of the ring of material; the solar-oblateness hypothesis had the unknown amount of the oblateness; and all the hypotheses that modified Newton's law of gravitation had an adjustable parameter (such as

Newcomb's ε) that could be chosen more or less at will.

Not all of the hypotheses were equally probable, however (Roseveare 1982). As noted above, sightings of Vulcan were never confirmed. As time went on, the hypothesis of matter rings of sufficient density also became less and less likely (Jeffreys 1921), although some still believed in them (Poor 1921). A solar oblateness of sufficient size probably would have been detectable with 19th-century techniques. The hypothesis that Newton's law of gravitation needed an arbitrary adjustment to fit the data was the one explanation that could not be ruled out by existing evidence.

What happened historically is well known. In 1915 Einstein announced his general theory of relativity, which predicted an excess advance in the perihelion motion of the planets. After some confusion (Roseveare 1982, pages 154–159) it became clear that the amount of the predicted advance for Mercury was very close to the unexplained discrepancy in Mercury's motion. The amazing thing was that the predicted value, which is 42.98 seconds of arc per century using modern values (Nobili and Wills 1986), was not a fudge factor that could be adjusted to suit the data but instead was an inevitable consequence of Einstein's theory.

The general theory of relativity made two other testable predictions (the gravitational bending of light and the slowing of clocks in a gravitational field). There has been a lively debate over the years as to how important each of these phenomena has been in convincing scientists that general relativity is the correct theory of gravity (Brush 1989). Here we shall side-step this argument and try to put ourselves inside the mind of a Bayesian observer in the early 1920s who is trying to weigh the evidence for various explanations of Mercury's anomalous motion.

Poor *v.* Jeffreys

An interesting pair of papers was published in 1921 (Poor 1921, Jeffreys 1921). Charles Lane Poor was an astronomer at Columbia University who was not convinced by the evidence for general relativity and who still clung to the matter-ring theory. Unfortunately, he also made some serious errors in his assessment of how a matter ring would affect the other inner planets. Jeffreys, in response, argued persuasively that the ring theory was not viable because sufficient matter did not exist. Jeffreys's paper was published before he made his major contributions to probability theory, and he does not, ironically, make the Bayesian argument that we have outlined above. And so we will make for Jeffreys the argument that he might have made had he returned to this question some years later.

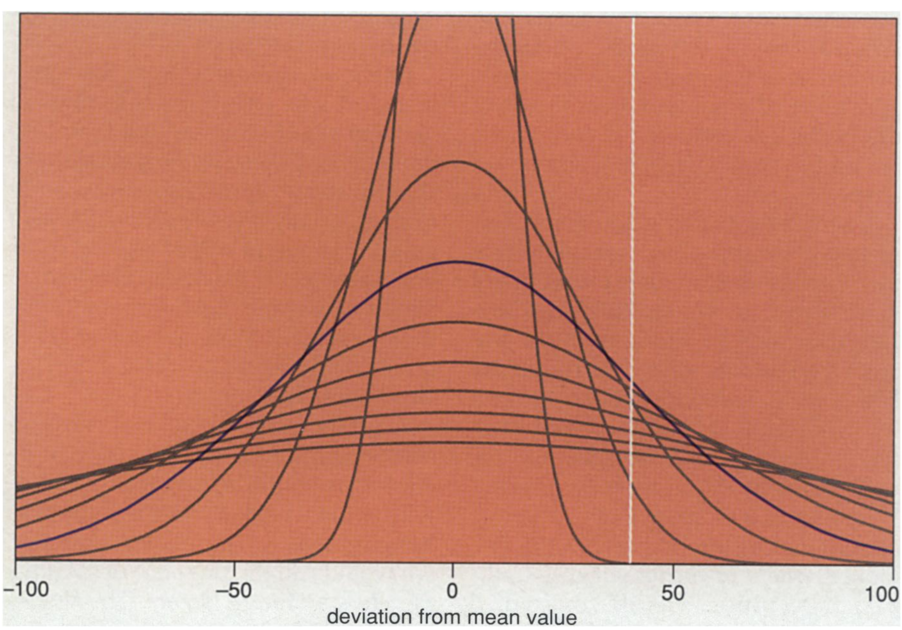

Figure 6. Assumption of a normal distribution with a specific standard deviation is a troubling step in comparing the Einsteinian and the fudged Newtonian theories. But if the distribution is indeed normal, there must be some value of the standard deviation that maximizes the probability of the observed data point at 41.6 *(white line)*. With a very narrow distribution, the data point falls far out on the tail of the curve. With a wide distribution, the probability is spread over such a large range of possible observations that no one value has a very high likelihood. For the case at hand, the optimum distribution has a standard deviation of roughly 40 arc-seconds per century *(purple curve)*.

Poor gives a value of $a = 41.6 \pm 1.4$ arc-seconds per century for the observed anomalous motion of Mercury. The task for Bayesian analysis is to assign a probability, based on this observation, to each of the two candidate explanations of the planetary motion: Einstein's general theory of relativity and a "fudged Newtonian" theory, in which some parameter is adjusted to account for the discrepancy in the observations.

The place to begin is with the measurement's reported uncertainty of ±1.4 arc-seconds per century. Although Poor's paper does not discuss the nature of this uncertainty, it is surely what statisticians designate a probable error, which is equal to 0.6745 times the standard deviation; thus the standard deviation itself is 2.0 arc-seconds per century. It is reasonable to assume that this error has a normal distribution; in other words it is described by a symmetrical, bell-shaped curve, with the total area under the curve equal to 1, and with about two-thirds of the area lying within one standard deviation of the center.

Poor reports the prediction of Einstein's theory as $\alpha_E = 42.9$ arc-seconds per century, which is quite close to the modern value. On the assumption that Einstein's prediction is in fact correct, what is the value of $P(a \mid E)$, the probability of observing a value of $a = 41.6$ arc-seconds? The answer can be determined by evaluating the appropriate normal curve (namely the curve centered at Einstein's prediction of 42.9 and having a standard error of 2.0) at the observed data value $a = 41.6$ *(Figure 5)*. The resulting value, called the probability density of $a = 41.6$, is about 0.16, which is reasonably high in this context. If the observed value of a were 42.9, exactly equal to the predicted value, the probability density would rise only to 0.20.

Performing the equivalent calculation for the fudged Newtonian theory is not as straightforward. For the very reason that the theory has a fudge factor, it is not easy to say exactly what it predicts. To give the theory an explicit probabilistic form, it is necessary to make some assumptions, although it will become apparent later that the outcome is quite insensitive to these assumptions.

One useful point of departure is the conservative assumption that since the Newtonian theory is well established, large deviations from it are less believable than small ones. (If gravity had an inverse-cube law instead of an inverse-square law, the difference would have been noticed long ago.) Accordingly, it is natural to choose a probability distribution for the unknown anomalous perihelion motion α that makes $\alpha = 0$ the likeliest value, with the probability density diminishing smoothly as α departs from zero. Likewise, it makes sense to give the probability density a symmetrical distribution, at least for those theories in which α could equally well be either positive or negative, so that the perihelion motion of Mercury could be either advanced or retarded. It is important to think *a priori* here; we are discussing predictions of the alternative theory prior to seeing the data.

These considerations would be satisfied by a normal probability distribution with a mean of $\alpha = 0$. But the most difficult question remains: What is the standard deviation of this distribution, which determines the width of the bell-shaped curve? Again the only available guidance is the knowledge that very large values of the anomalous perihelion motion are ruled out by existing observations. For example, if some gravitational effect perturbed the perihelion motion of Mercury by as much as 100 arc-seconds per century, it would also alter the orbits of Venus, the earth and Mars to an extent that could have been detected in the 1920s. For the purposes of rough calculations, a reasonable standard deviation is about 50 arc-seconds per century, which does not contradict any observational data on the inner planets.

We now have in hand the two elements needed to calculate the probability density of the observed data $a = 41.6$, assuming the validity of a fudged Newtonian theory. Assuming that the unknown anomalous perihelion shift α has a normal distribution with mean zero and standard deviation 50, and that the observed a is equal to α plus a random error having standard deviation 2.0, standard methods of probability theory can be used to compute $P(a \mid F)$, the overall probability density of a under the fudged Newtonian theory. In this example, $P(a \mid F)$ itself turns out to have a normal distribution with mean 0 and standard deviation 50.04. This distribution is much flatter than $P(a \mid E)$, so that the probability is distributed over a much wider range. For this reason, the probability density of any one value is greatly reduced. Specifically, the probability density of

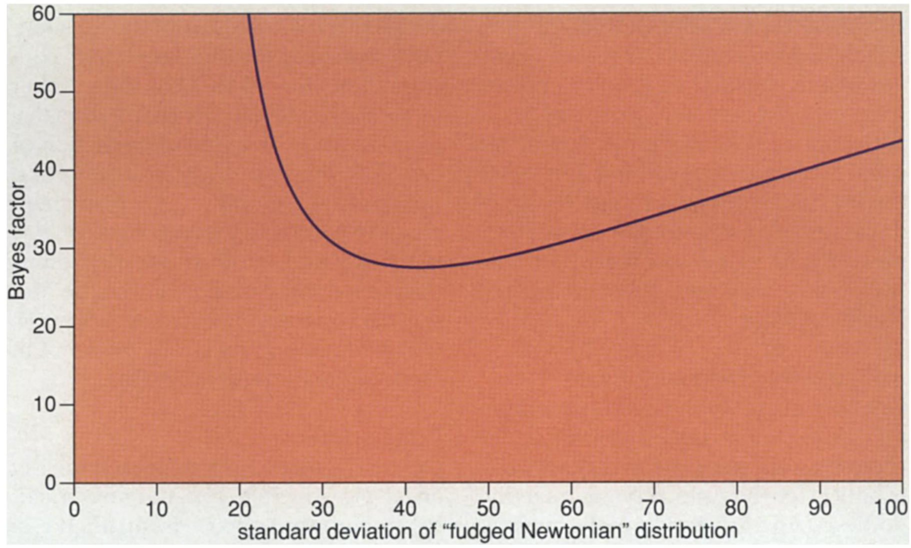

Figure 7. Bayes factor indicates the degree to which Einstein's theory is favored over the fudged Newtonian theory as a function of the standard deviation assumed in the latter theory. For the comparison graphed here the distribution is assumed to be normal. Under this circumstance the minimum Bayes factor is 27.76. An alternative formulation of Ockham's razor can be applied to other distributions as well, provided only that they are symmetric and decreasing with distance from the central value. By this more liberal criterion Einstein's explanation is favored over the fudged Newtonian hypothesis by odds of at least 15 to one.

the actual value $a = 41.6$ is only about 0.0056, compared with the probability density of 0.16 for Einstein's theory.

What is of interest, however, is not the probability density of the data $a = 41.6$ given the various theories, but rather the probabilities of the various theories being true given $a = 41.6$. These latter probabilities could be calculated from Bayes's theorem if one were willing to assign prior (that is, premeasurement) probabilities to the theories. Luckily, the need to choose prior probabilities can be avoided (if desired) by use of the ratio of the probability densities of $a = 41.6$ under the Einsteinian and the fudged Newtonian theories, namely

$$B = \frac{P(a \mid E)}{P(a \mid F)}$$

It can be shown from Bayes's theorem that this ratio, called the Bayes factor, gives the odds favoring E over F arising from the data. When B is greater that 1, the data favor E, and when B is less than 1, they favor F. The overall odds of E over F are found by multiplying B by the prior odds, which is the ratio of the prior probabilities of E and F. The point here is that it may suffice to consider only B; the Bayes factor may well answer the question without the need to formally involve the prior odds.

Plugging in the numbers yields a value of $B = 28.6$, which is moderately strong evidence in favor of the Einsteinian hypothesis. Ironically, the data that Poor himself provides in his paper against general relativity favor the Einsteinian hypothesis over the fudged Newtonian hypothesis.

The calculations leading to this conclusion involve several factors. There is, first, the matter of how well the data fit each hypothesis. Obviously, if the observed data differ sharply from the predictions of a hypothesis, one would expect that hypothesis to be assigned a low probability. In many cases such goodness-of-fit considerations are decisive in choosing among hypotheses. In this instance, however, the predictions of both theories are consistent with the data. Nevertheless, Bayes's theorem offers a clear choice between the theories.

The factor that contributes most to the outcome of the calculations is the width of the probability distribution of a for the fudged Newtonian hypothesis. Because this distribution is relatively wide, the fudged Newtonian hypothesis has to waste a considerable amount of probability on hypothetical values of a that are far from the actual $a = 41.6$.

The fudged Newtonian hypothesis has an additional degree of freedom that allows it to accommodate a much larger range of hypothetical data than does the Einsteinian hypothesis. As a result the fudged Newtonian hypothesis must spread its risk over a larger parameter space in order not to miss the region supported by the data. In this sense, it is a less simple theory than the Einsteinian hypothesis. Einstein's hypothesis makes a sharp prediction about Mercury's perihelion motion, which depends only on the known values of the constant of gravity and the speed of light. Any measurement of the perihelion motion that is not close to the predicted value contradicts Einstein's theory. In contrast, a broad range of data—every value of the anomalous motion not ruled out for other reasons—is consistent with the fudged Newtonian hypothesis.

An Objective Ockham's Razor

One step in our analysis of Poor's argument may seem rather doubtful: the choice of a specific prior probability distribution for the fudged Newtonian hypothesis. And this aspect of the analysis turned out to be particularly important, since it is the great width of that distribution that makes the difference between the two hypotheses. We suggested first that the probability distribution should be symmetric about $\alpha = 0$ and decreasing as the absolute value of α increases; these are reasonable constraints on the shape of the distribution. But the final choice of a normal distribution with a specific standard deviation of 50 arc-seconds per century seems rather arbitrary. The method by which we arrived at the figure of 50 would be difficult to generalize to other problems.

The need to specify a specific standard deviation for $P(\alpha \mid F)$ is easy to overcome. One can simply consider an arbitrary standard deviation—call it τ—and then graph the Bayes factor B as a function of τ. This is done in Figure 7. Of considerable interest is the finding that B has a minimum value; it is always greater than 27.76. Thus there is strong evidence in favor of the Einsteinian theory no matter what value of τ is chosen.

It is less obvious how to overcome the rather arbitrary choice of a normal distribution for $P(\alpha \mid F)$. The solution is given in Berger and Jefferys (1992), where it is shown that the Bayes factor has a lower limit even if the distribution for $P(\alpha \mid F)$ is not a normal one, provided only that the distribution obeys certain rather mild conditions. Specifically, for any $P(\alpha \mid F)$ that is symmetric about $\alpha = 0$ and decreasing in the absolute value of α, B is always less than or equal to the following expression:

$$\sqrt{\frac{2}{\pi}}\left(|D_F| + \sqrt{2\,ln\left(|D_F| + 1.2\right)}\right)\exp\left(\frac{-D_E^2}{2}\right)$$

Here D_E is the number of standard deviations that a deviates from the Einsteinian prediction; for the data under consideration $D_E = -0.65$. D_F is the number of standard deviations that a deviates from the base Newtonian prediction of $\alpha = 0$; in this case $D_F = 20.8$. Adopting this "worst-case" value gives every benefit of the doubt to the hypothesis F; if F is not favored under these conditions, then it is not favored at all. For the present case, the lower bound on B is 15.04, which remains fairly strong evidence in favor of the Einsteinian theory.

Conclusions

Ockham's razor, far from being merely an *ad hoc* principle, can in many practical situations in science be justified as a consequence of Bayesian inference. Bayesian analysis can shed new light on what the notion of the "simplest" hypothesis consistent with the data actually means. We have discussed two ways in which Ockham's razor can be interpreted in Bayesian terms. By choosing the prior probabilities of hypotheses, one can quantify the scientific judgment that simpler hypotheses are more likely to be correct. Bayesian analysis also shows that a hypothesis with fewer adjustable parameters automatically has an enhanced posterior probability, because the predictions it makes are sharp. Both of these ideas are in agreement with the intuitive notion of what makes a scientific theory powerful and believable.

Acknowledgments

We wish to thank our colleagues who have commented on drafts of this paper: Raynor L. Duncombe, David M. Hillis, Myra Samuels and Mark Jefferys. This work was supported by the National Aeronautics and Space Administration, under NASA Contract NAS5-29285, and by the National Science Foundation, Grant DMS-8923071.

References

Bayes, Thomas. 1763. An essay toward solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society* 53:370–418.

Berger, James O. 1985. *Statistical Decision Theory and Bayesian Analysis*. New York: Springer-Verlag.

Berger, James O., and D. Berry. 1988. Statistical analysis and the illusion of objectivity. *American Scientist* 76:159–165.

Berger, James O., and William H. Jefferys. 1992. The application of robust Bayesian analysis to hypothesis testing and Occam's razor. To appear in *Journal of the Italian Statistical Society*.

Brush, Stephen. 1989. Prediction and theory evaluation: The case of light bending. *Science* 246:1124–1129. See also the responses to this article in *Science* 248:422–423.

Drake, Stillman, and James MacLachlan. 1975. Galileo's discovery of the parabolic trajectory. *Scientific American* 232:102–110.

Edwards, W., H. Lindman and L. J. Savage. 1963. Bayesian statistical inference for psychological research. *Psychological Review* 70:193–242.

Good, I. J. 1968. Corroboration, explanation, evolving probability, simplicity, and a sharpened razor. *British Journal of the Philosophy of Science* 19:123–143.

Good, I. J. 1977. Explicativity: a mathematical theory of explanation with statistical applications. *Proceedings of the Royal Society A* 354:303–330.

Gull, S. 1988. Bayesian inductive inference and maximum entropy. In G. J. Erickson and C. R. Smith (eds.) *Maximum Entropy and Bayesian Methods in Science and Engineering* (Vol. 1), 53–74. Dordrecht: Kluwer Academic Publishers.

Harpp, David. 1991. Quoted in "Big Prof is Watching You," *Discover* 12 (April):12–13.

Jaynes, E. T. 1979. Inference, method, and decision: Towards a Bayesian philosophy of science. *Journal of the American Statistical Association* 74:740-41.

Jeffreys, Harold. 1921. Secular perturbations of the inner planets. *Science* 54:248.

Jeffreys, Harold. 1939. *Theory of Probability*. (Third Edition 1983.) Oxford: Clarendon Press.

Laplace, Pierre de Simon. 1812. *Théorie Analytique des Probabilities*. Paris: Courcier.

Loredo, T. J. 1990. From Laplace to Supernova 1987A: Bayesian inference in astrophysics. In P. Fougere (ed.) *Maximum Entropy and Bayesian Methods*, 81–142. Dordrecht: Kluwer Academic Publishers.

MacKay, David J. C. 1991. Bayesian interpolation. Submitted to *Neural Computation*.

Max, Edward E. 1986. Plagiarized errors and molecular genetics: Another argument in the evolution-creation controversy. *Creation/Evolution* XIX:34–46.

Newcomb, Simon. 1895. *The Elements of the Four Inner Planets and the Fundamental Constants of Astronomy*. Washington: Government Printing Office. pages 109–122.

Nobili, A. M., and C. M. Will. 1986. The real value of Mercury's perihelion advance. *Nature* 320:39–41.

Poor, C. L. 1921. The motions of the planets and the relativity theory. *Science* 54:30–34.

Smith, A. F. M., and D. J. Spiegelhalter. 1980. Bayes factors and choice criteria for linear models. *Journal of the Royal Statistical Society B* 42:213–220.

Roseveare, N. T. 1982. *Mercury's Perihelion from Le Verrier to Einstein*. Oxford: Clarendon Press.

Thorburn, W. M. 1918. The myth of Occam's razor. *Mind* 27:345–353.

Watson, James D., Nancy H. Hopkins, Jeffrey W. Roberts, Joan A. Steitz and Alan M. Weiner. 1988. *Molecular Biology of the Gene*, 4th Edition. Menlo Park: Benjamin/Cummings Publishing Company. pages. 649–663.

Wrinch, D., and H. Jeffreys. 1921. On certain fundamental principles of scientific inquiry. *Philosophical Magazine* 42:369–390.