Expressing Prior Ignorance of a Probability Parameter

George P. Smith

Division of Biological Sciences

Tucker Hall

University of Missouri

Columbia, MO 65211-7400

(573) 882-3344; smithgp@missouri.edu

AUTHOR'S FOOTNOTE

George P. Smith is Professor, Division of Biological Sciences, University of Missouri, Columbia, MO 65211 (e-mail: smithgp@missouri.edu).

ABSTRACT

If p is an unknown probability parameter, prior ignorance of its value is appropriately expressed by the prior probability density distribution $f(p) \propto 1/p(1-p)$. That is the only distribution that remains invariant under transformations that convert the original inference problem into another that should look identical to a truly ignorant observer. This invariance principle for specifying an ignorant prior distribution, pioneered in the writings of the late E.T. Jaynes, is contrasted with the principle of "data translation" as set forth by Box and Tiao—a principle that typifies alternative approaches to "noninformative" prior distributions.

KEYWORDS

Prior probability distribution; Bayesian statistics; Noninformative probability distribution; Likelihood principle

1. INTRODUCTION

1.1 Bayes's Theorem: The Rule for Rationally Updating Opinions in the Light of Evidence

Bayesians regard the laws of probability as the fundamental rules of empirical reasoning, deriving them from the principle of "coherence": the requirement that rational opinions should be consistent with one another and with other commitments that we find deeply compelling *a priori*. Bayes's theorem is central to the logic of science in this view. It articulates the rule for rationally updating our opinions of contending hypotheses in light of new evidence:

$$\Pr(H \mid \hat{E}) \propto \Pr(\hat{E} \mid H)\Pr(H) \qquad \text{Eq. 1}$$

where H is any of a series of contending scientific theories and $\hat{E}$ is relevant evidence newly at hand. $\Pr(H)$ is the prior probability of hypothesis H, expressing our rational degree of belief in its truth before acquiring (or considering) evidence $\hat{E}$. $\Pr(H \mid \hat{E})$ is the posterior probability, expressing an updated degree of belief in the truth of H in light of the new evidence $\hat{E}$. Bayes's theorem states that the posterior probability of each hypothesis is proportional both to its prior probability and to $\Pr(\hat{E} \mid H)$—the probability of obtaining evidence $\hat{E}$ given that hypothesis H is true. Considered as a function of the various contending hypotheses H for fixed evidence $\hat{E}$—the evidence we actually observe—$\Pr(\hat{E} \mid H)$ is the likelihood of hypothesis H. Bayes's theorem can therefore be epitomized

$$\text{updated opinion} \propto \text{likelihood} \times \text{starting opinion}$$

If the contending hypotheses are mutually exclusive and exhaust all the possibilities, their posterior probabilities can be normalized by imposing the additional constraint that they sum to unity.

The likelihood function $\Pr(\hat{E} \mid H)$ is an *E* transect of the sampling distribution $\Pr(E \mid H)$, whose domain subtends not only the manifold of contending hypotheses *H*, but also the manifold of possible observations *E* (including the evidence actually observed, $\hat{E}$): $\Pr(\hat{E} \mid H) = \Pr(E \mid H)\big|_{E=\hat{E}}$. If the sampling distribution $\Pr(E \mid H)$ describes the anticipated outcomes of some sort of experiment about to be undertaken, it can be said to embody the intended design of the experiment. The way the sampling distribution depends on *E* is highly relevant to the design of an anticipated experiment, but completely irrelevant once the observation has actually been made, revealing the particular data-set $\hat{E}$. As implied in Eq. 1, it is the relative values of the likelihood function for the different hypotheses *H* on fixed data $\hat{E}$ that matter; factors that don't depend on *H* are irrelevant, and can be suppressed without loss of relevant information. This aspect of Bayesian coherence is called the likelihood principle.

1.2 Prior Probability Distributions: The Constraints of Ignorance

While Bayes's theorem provides a strict rule for rationally updating our opinions as new evidence emerges, it doesn't provide the starting point for this chain of reassessments. What prior probability distribution should be chosen for hypotheses about which we know little or nothing in advance? Of course, there will be fortunate cases when the choice of prior probability distribution is of no practical importance because the likelihood function is a much steeper function of *H* than is any colorable prior probability distribution. But there remain applications in which the evidence is more vague, so that the posterior probability distribution is sensitive to the prior distribution.

In his *Théorie Analytique des Probabilités* of 1812, which set the stage for modern statistical inference, Laplace invoked what is now known as the principle of indifference: if nothing is known in advance, all alternatives are “equally possible” and should be assigned equal prior probability. When the alternatives in question are discrete, this principle is compelling: the names of the alternatives are in that case arbitrary labels that could be reshuffled without altering the problem from the point of view of the ignorant observer—a constraint that can only be met if the prior probabilities are equal. But what if the alternatives in question are the different possible values of a continuous parameter? Should we assign uniform prior probability densities to the different values of the parameter? to the different values of its logarithm? its reciprocal? The principle of indifference seems on the face of it inapplicable to this case.

Myriad plausible but mutually incompatible desiderata for “noninformative” prior distributions for continuous parameters have been put forth over the years. The choice among them would appear to be merely a matter of personal preference—an exceedingly unwelcome clash with the principle of coherence. A foundational article by Jaynes (1968; 2003, pp. 372–396) addressed this problem in a new and profoundly illuminating way that has yet to be fully absorbed by the Bayesian community. He argued that ignorance imposes stringent constraints on prior probability distributions by requiring that they be invariant under transformations that convert the original prior inference problem into another prior inference problem that should look identical to a truly ignorant observer. From this requirement he derives ignorant prior probability density distributions for the three most common types of continuous parameter:

$$f(\mu) \propto 1, \quad -\infty < \mu < +\infty \text{ for a location parameter } \mu \qquad \text{Eq. 2}$$

(for example, the mean of a normal distribution);

$$f(\sigma) \propto \frac{1}{\sigma}, \quad 0 \leq \sigma < +\infty \text{ for a scale parameter } \sigma \qquad \text{Eq. 3}$$

(for example, the standard deviation of a normal distribution); and

$$f(p) \propto \frac{1}{p(1-p)}, \quad 0 \le p \le 1 \text{ for a probability parameter } p. \qquad \text{Eq. 4}$$

It is the latter that will be the focus of this article.

2. IGNORANT PRIOR DISTRIBUTION FOR A PROBABILITY PARAMETER

2.1 Jaynes's Argument

The random variable p is an unknown probability parameter; we can consider it to be the probability of "success" in some sort of repeatable trial. What prior probability density function $f(p)$ properly represents prior ignorance of this parameter? Jaynes's (1968; 2003, pp. 382–386) argument for Eq. 4 relies on a peculiar rhetorical device he calls "split personalities taken to the extreme." The various probabilities p are regarded as prior opinions held by the members of a huge population of Bayesian analysts. These Bayesians have conflicting prior information and that accounts for their conflicting prior probabilities. The function $f(p)$ describes the distribution of Bayesians in the population holding the various prior probabilities. Jaynes seeks a distribution function $f(p)$ that would appropriately represent a state of "total confusion" in the population. He defines confusion thus: if you give all Bayesians in the population one additional piece of relevant evidence *E*, the distribution of their posterior probabilities is the same as the distribution of their prior probabilities (though of course each of their individual posterior probabilities differs from their prior probabilities according to Bayes's Theorem). From this requirement Jaynes easily deduces that the distribution must have the form of Eq. 4.

I'm unsatisfied with this argument. Although it is certainly an interesting property of the distribution, it is not intuitively clear that "total confusion" as Jaynes defines it can be directly equated with prior ignorance. I present here two alternative arguments that seem to me much clearer, more direct applications of Jaynes's general principles.

It will be mathematically convenient in what follows to consider the random variable as the success:failure ratio

$$X \equiv \frac{p}{1-p} = \frac{\Pr(S)}{1-\Pr(S)} = \frac{\Pr(S)}{\Pr\left(\overline{S}\right)} \qquad \text{Eq. 5}$$

where S and $\overline{S}$ represent success and failure, respectively. Jaynes's ignorant prior distribution for *p*, Eq. 4, corresponds to the probability density function

$$f(X) \propto \frac{1}{X} \qquad \text{Eq. 6}$$

for X. I will put forth two arguments for this distribution as the appropriate expression of prior ignorance.

2.2 Argument I: Symmetry Between Unconditional and Conditional Success:Failure Ratios

Bayesian I considers the success:failure ratio Eq. 5, and seeks a prior distribution $f(X)$ that expresses his prior ignorance of that ratio. He is then asked to consider

$$Y \equiv \frac{\Pr(S \mid E)}{\Pr\left(\overline{S} \mid E\right)}, \qquad \text{Eq. 7}$$

the success:failure ratio conditional on some event *E*. He is given the value of the likelihood ratio $R \equiv \Pr(E \mid S)/\Pr\left(E \mid \overline{S}\right)$, which by Bayes's theorem implies that $Y = RX$. This change of variable implies a strict relationship between the ignorant distribution $f(X)$ and his probability density distribution for *Y*—let's call that distribution $g(Y)$:

$$g(Y) = f\left(\frac{Y}{R}\right)\frac{1}{R} \qquad \text{Eq. 8}$$

Now let's consider a second Bayesian, II, who is asked at the outset to consider the conditional success:failure ratio *Y* (Eq. 7). The fact that success and failure are both conditional on some unspecified event *E* conveys no useful information about the success:failure ratio. After all, any pair of events whatever could be considered to be conditional on some unspecified background knowledge. So Bayesian II assigns the same ignorant distribution to the conditional success:failure ratio *Y* as Bayesian I did to unconditional success:failure ratio *X*: $f(Y)$. She is then asked to consider the unconditional success:failure ratio *X* (Eq. 5), and (like Bayesian I) is informed of the value of the likelihood ratio $R \equiv \Pr(E \mid S)/\Pr\left(E \mid \overline{S}\right)$, which by Bayes's theorem implies that $X = Y/R$. This change of variable implies a strict relationship between the ignorant distribution $f(Y)$ and her probability density distribution for *X*—let's call that distribution $h(X)$:

$$h(X) = f\left(RY\right)R \qquad \text{Eq. 9}$$

But now Bayesian I and II have precisely the same information, so their distributions must be the same. This means that Bayesian II's <u>deduced</u> distribution for *X*, $h(X)$, must be the same as Bayesian I's <u>original</u> ignorant distribution $f(X)$; and likewise that Bayesian I's <u>deduced</u> distribution for *Y*, $g(Y)$, must be the same as Bayesian II's <u>original</u> ignorant distribution $f(Y)$. In other words, consistency demands that the functions $f(\cdot)$, $g(\cdot)$ and $h(\cdot)$ be all the same function, which we'll call $f(\cdot)$. Bayesian I's expression for $g(Y)$, Eq. 8, therefore imposes the constraint

$$g(Y) = f(Y) = f\left(\frac{Y}{R}\right)\cdot\frac{1}{R} \qquad \text{Eq. 10}$$

and this must hold for any value of R greater than 0. Precisely the same functional constraint is implied by Bayesian II's expression for $h(X)$, Eq. 9, as can be seen by substituting h for g, X for Y and $1/R$ for R in the above equation. The functional equation Eq. 10 has the unique solution given in Eq. 6.

The force of this argument rests on the symmetry between the unconditional and conditional success:failure ratios to the ignorant observer. Either one can serve as the starting-point for inference, and both must therefore be assigned the same "ignorant" prior distribution $f(\cdot)$. Bayes's Theorem then imposes a functional constraint on $f(\cdot)$ that can only be satisfied by Eq. 6.

2.3 Argument II: Success:Failure Ratio Is a Scale Parameter

The prior distribution deduced in the previous section, Eq. 6, is the same as that deduced by Jaynes (1968; 2003, pp. 378–382) when X is a scale parameter, Eq. 3. I will argue now that the success:failure ratio can indeed be considered a fully legitimate scale parameter, so that Jaynes's argument applies. A change in scale in the success:failure ratio results from a change in scale in counting either successes or failures or both. For example, let us suppose that success is redefined, so that one new success is counted for every n old successes (n can be any positive real number). Similarly, suppose that failure is redefined so that one new failure is counted for every m old failures. The new success:failure ratio Y has a simple scalar relationship to the old ratio X: $Y = RX$, where $R = m/n$; the success:failure is now being measured on a new scale whose units are R-fold smaller than the old scale. But if we're totally ignorant in advance of the value of the success:failure ratio, it doesn't matter to us which of the two scales the ratio is measured on: we're equally ignorant in both cases. In order to appropriately express prior ignorance, therefore, the prior probability density distribution for success:failure ratio must be invarant under a change of scale—a stringent constraint that is only satisfied by the distribution in Eq. 6.

3. ALTERNATIVE NONINFORMATIVE PRIOR DISTRIBUTIONS

3.1 Data-Translated Prior Distributions

Box and Tiao (1973, pp. 25–60) support an altogether different desideratum for noninformative prior distributions, which will serve to exemplify the many extant alternatives to Jaynes's invariance principles. These authors argue that noninformative prior distributions should to the extent possible be "data-translated." If the continuous parameter in question is *X*, they seek a transformation $Y = \varphi(X)$ such that all *E* transects of the sampling distribution $\Pr(E \mid Y)$ have the same shape but different locations on the *Y* axis when plotted against *Y*. The effect of observing different outcomes $\hat{E}$ is thus to slide ("translate") the likelihood function $\Pr(E \mid Y)\big|_{E=\hat{E}}$ along the *Y* axis without changing its contour. If we choose a prior probability distribution that is uniform in *Y*, the posterior probability function will therefore have the same shape as—and in that sense be "dominated" by—the likelihood function. Box and Tiao contend that a uniform prior probability density in *Y*—or equivalently, the density function $f(X) \propto \left|d\varphi(X)/dX\right|$ in *X*—lets the data speak for themselves, and in that sense qualifies as "noninformative." Exact data-translated prior probability distributions exist only for a limited number of sampling distributions, but nearly exact data-translated prior distributions can be found in a wide variety of circumstances. In particular, Box and Tiao (1973, pp. 34–35) consider a series of $\hat{n}$ identical trials in each of which the unknown probability of success is *p*; the outcome of the trial series is the number of successes observed, *r*. The nearly data-translated prior distribution in that case is

$$f(p) \propto \frac{1}{[p(1-p)]^{1/2}} \ , \qquad \text{Eq. 11}$$

a function that differs from Jaynes's general ignorant prior distribution for a probability parameter (Eq. 4). Accepting data translation as reasonable would therefore undermine Jaynes's invariance arguments as a single unifying principle for expressing prior ignorance.

3.2 Criticisms of Data Translation

But is data translation really reasonable? It bases the prior probability distribution, not on the scientist's state of knowledge about the actual scientific matter at hand (the competing values of the parameter of interest in this case), but rather on his investigative intentions, as embodied in the sampling distribution. Inferences starting with such prior distributions violate the likelihood principle because they must take into account not only the dependence of the sampling distribution $\Pr(E \mid X)$ on the continuous parameter of interest *X*, but also its dependence on the entire manifold of possible outcomes *E*, including all the data we might have observed but didn't.

Consider as an example three investigators who are equally ignorant in advance of the probability *p* of success in a repeatable trial of the sort discussed above. Independently, they design experimental trial series in order to learn more about *p*. Investigator I sets out from the start to complete exactly 22 trials and will score the number of successes *r*; this is the design already described above. Investigator II sets out from the start to continue trials until exactly 7 successes are observed, and will score the number of trials *n* required to achieve this goal. As investigator III is considering his options, he learns he will be handsomely rewarded by a patron if he succeeds in finding evidence highly favorable to the hypothesis that the true probability of success *p* has the predetermined value $1/\pi$. As each new trial is completed, therefore, he plans to analyze the data to see how strongly they support his patron's preconceived notion. If at any point the results to date (the number of trials *n* and successes *r*) seem likely to earn him his reward, he'll terminate the experiment; otherwise, he'll continue until frustration, boredom or fatigue finally compel him to quit. It is stipulated that none of the investigators will interfere with the actual trials or tamper with the results in any way.

For investigator I the nearly data-translated prior distribution has already been given in Eq. 11. For investigator II the nearly data-translated prior distribution is $f(p) \propto p^{-1}(1-p)^{-1/2}$ (Box and Tiao 1973, pp. 44–45). Specifying a sampling distribution for investigator III would be a staggering task, involving as it would a detailed analysis of his degree of avarice for the reward, guess as to what sort of evidence will satisfy his patron, propensity to terminate an exhausting investigation that seems to have no prospect of yielding favorable results, level of commitment to standard canons of proper research procedure, etc. Whatever the sampling distribution's complexity, however, it can be factored into the form

$$\Pr(E \mid p) = \left[\prod_{i=1}^{n_E-1}(1-Q_{Ei})P_{Ei}\right]Q_{En_E}P_{n_E} = \left[\prod_{i=1}^{n_E-1}(1-Q_{Ei})\right]Q_{En_E}\,p^{r_E}(1-p)^{n_E-r_E} ,$$

where *E* is one of the possible outcomes, a particular succession of r_E successes and $r_E - n_E$ failures in a total of n_E trials; P_{Ei} is either *p* or $1-p$ depending on whether the *i*th trial in succession *E* is a success or failure, respectively; and Q_{Ei} is the probability that the investigator will quit after obtaining the results of the first *i* trials of succession *E*. The sampling distribution can therefore be written

$$\Pr(E \mid p) = \alpha(E)\, p^{r_E} (1-p)^{n_E - r_E} , \qquad \text{Eq. 12}$$

where all the psychological complications are confined to the function $\alpha(E)$, which cannot depend on the unknown value of p. The sampling distributions for investigators I and II will also have this form, though in their case the factor $\alpha(E)$ is much simpler: the constant 1 for any of the allowed outcomes ($n_E = 22$ for investigator I; any succession for which $r_E = 7$ and the last trial is a success for investigator II), the constant 0 otherwise. In any case, we'll assume for the sake of argument that a data-translated prior distribution could be specified in principle even for investigator III.

Having thus formulated their plans, investigators I–III carry out their respective trial series, and (we'll suppose) obtain exactly the same data: 7 successes and 15 failures in precisely the same order of succession. If they follow the advice of Box and Tiao, they will come to different inferences about p, embodied in different posterior probability density distributions. That's because they have different sampling distributions and therefore start from different prior probability density distributions. But this does not square with plain common sense. The differences among the investigators lie entirely in their mental states before and during data-collection—their intentions, hopes, expectations of reward, susceptibility to fatigue, commitment to investigative propriety, etc. None of them interferes with the actual objective evidence in the case—the succession of successes and failures in the trial series. If they start with the same prior information about p and collect the same objective data, surely their final inferences ought to be the same; their contrasting psychological states during data collection are irrelevant to any scientific judgment. Admirers of data translation try to counter this indictment by claiming that the design of the anticipated experimental trial series is part of the investigator's prior knowledge, and can thus legitimately influence his prior probability distribution for p. But this is not prior knowledge of the parameter p itself; it's prior knowledge of his own psychological condition as he embarks on or carries out the experiment, and thus surely extraneous to rational inference. To incorporate such "knowledge" into the inference problem is a classic case of what Jaynes (1989; 2003, p. 22) calls the "mind projection fallacy": the "temptation to project our private thoughts out onto the real world, by supposing that the creations of one's own imagination are real properties of Nature…" In summary, the principle of data translation commits us to an unpleasant violation of the likelihood principle and consequent conflict with common sense.

This conflict will never arise if, ignoring the counsel of Box and Tiao, investigators I–III adopt a common prior distribution, Eq. 4, that expresses their shared state of prior ignorance without reference to the anticipated trial series; and reason in accord with the likelihood principle. That's because their likelihood functions are proportional to one another regardless of the differences among the sampling distributions (Eq. 12) they're derived from: $\Pr(22,7 \mid p) \propto p^7 (1-p)^{22-7}$ in all three cases. Their results therefore carry precisely the same evidentiary information despite their psychological disparities—in full accord with intuition. Multiplying this likelihood function by Jaynes's ignorant prior distribution Eq. 4, all three investigators arrive at the same posterior probability density distribution: $f(p \mid 22,7) \propto p^6 (1-p)^{14}$. For all three investigators, the "rule of succession"—that is, the probability, given the data, that the outcome of the next trial will be a success—is the same:

$$\text{Expectation of } p \text{ given the data} = \int_0^1 p \cdot f(p \mid 22,7)\,dp = \frac{\int_0^1 p^7 (1-p)^{14}\,dp}{\int_0^1 p^6 (1-p)^{14}\,dp} = \frac{7}{22}.$$

3.3 Jaynes's Invariance Principle: Still the Foundation for Articulating Ignorance

Data translation's clash with common sense should not surprise us. It does not even pretend to meet the core requirement of any Bayesian probability distribution: that it accurately represent the investigator's state of knowledge. The steps leading from prior ignorance to data translation as a desideratum are correspondingly indirect, casual, tenuous; those leading to Jaynes's invariance principle are direct, rigorous, perspicuous. It will take a far more compelling principle than the former to seriously challenge the latter's preeminent claim as a unifying foundation for articulating ignorance. Such a challenge does not seem soon in prospect.

4. REFERENCES

Box, G.E., and Tiao, G.C. (1973), *Bayesian Inference in Statistical Analysis*. Reading: Addison-Wesley.

Jaynes, E.T. (1968), "Prior probabilities," *IEEE Transactions on Systems Science and Cybernetics*, SSC-4, 227–241.

Jaynes, E.T. (1989), "Clearing up mysteries—the original goal," in *Maximum Entropy and Bayesian Methods*, Skilling, J. (ed.), Dordrecht, The Netherlands, Kluwer Academic Publishers.

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