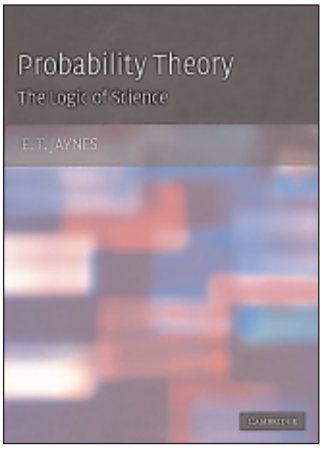
Probability Theory: The Logic of Science

Reviewed by William G. Faris

Probability Theory: The Logic of Science
E. T. Jaynes
Edited by G. Larry Bretthorst
Cambridge University Press
Cambridge, 2004, $70.00
xxix + 727 pages, ISBN 0-521-59271-2

Bayes's Theorem. Bayes's theorem is elementary probability; its application to the world involves a conceptual struggle of the first magnitude. Several hypotheses compete to explain the available experimental evidence. The data are accumulated; the evidence is at hand. Which hypothesis is indicated? The decision can be quite serious.



Here is an artificial but illuminating example. Two kinds of structural material are made. The good one is reasonably strong; the bad one fails considerably more often. A sample of material is delivered, but its composition is unknown. Thus there are two hypotheses: G that it is good, or B that it is bad. It is vital to know which hypothesis is true.

William G. Faris is professor of mathematics at the University of Arizona, Tucson. His email address is faris@math.arizona.edu.

The probability that the good material survives a crush test is known to be $P(S \mid G) = 5/6$. The probability that the bad material survives the same crush test is only $P(S \mid B) = 1/3$. The experimental fact is at hand: the sample passed the crush test. What can we say about the material?

To get some mathematics out of the way, here is Bayes's theorem in its most elementary form. (A more general version comes later in this review.) There are hypotheses θ with "prior" probabilities $P^H(\theta)$. For each hypothesis θ there is a conditional probability $P(x \mid \theta)$ for each data point x. The unconditional probabilities for the data points are then

$$P(x) = \sum_\theta P(x \mid \theta) P^H(\theta). \tag{1}$$

Bayes's theorem states that the "posterior" conditional probability of hypothesis θ given the observed data point x is

$$P^H(\theta \mid x) = \frac{P(x \mid \theta) P^H(\theta)}{P(x)}. \tag{2}$$

It is little more than the definition of conditional probability.

The theorem can also be written in terms of the normalized likelihood function $\hat{f}(x, \theta)$ defined by

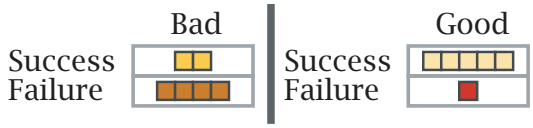
$$P(x \mid \theta) = \hat{f}(x, \theta) P(x). \tag{3}$$

The Bayes result is then that

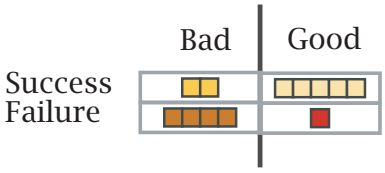
$$P^H(\theta \mid x) = \hat{f}(x, \theta) P^H(\theta). \tag{4}$$

Bayes's Theorem at a Glance

The Sample Distribution. A sample of material is Bad or Good; which one is not known. The result of a crush test is Success or Failure. The starting point is probabilities of Success or Failure given Bad or Good. In the example $P[S \mid B] = 2/6 = 1/3$ and $P[S \mid G] = 5/6$. The relative probabilities are indicated in the two graphs below:



The Joint Distribution. Suppose that the prior probabilities of the two hypotheses are equal, that is, $P^H[B] = 1/2$ and $P^H[G] = 1/2$. Then the combined graph gives equal weight to the two columns. This is shown below:



The Posterior Distribution. Bayes's theorem gives the probabilities for the hypotheses Bad and Good given Success or Failure. For example, $P^H[G \mid S] = 5/7$. The relative probabilities are indicated in the two graphs below:

Once the experiment has indicated Success or Failure, only one of these graphs is relevant.

—Bill Casselman and Bill Faris

The theorem says that the posterior probability of a hypothesis in light of the new evidence is obtained from the prior probability of the hypothesis by multiplication by the normalized likelihood function. This function incorporates the effect of the new evidence.

To apply this to the example, suppose that the two hypotheses good and bad have prior probabilities $P^H(G)$ and $P^H(B)$, which sum to one. The outcome was a success S. Does this indicate that the material is good? The theorem gives

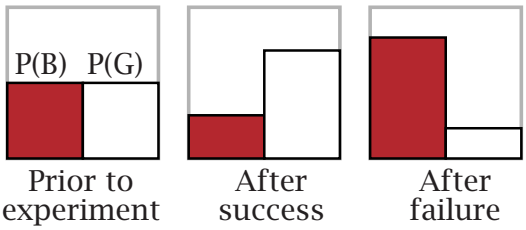


Figure 1. Crush test. The Bayes prior distribution with two possible Bayes posterior distributions. After the experiment only one of the two remains relevant.

$$(5) \qquad P^H(G \mid S) = \frac{\frac{5}{6}}{\frac{5}{6}P^H(G) + \frac{1}{3}P^H(B)} P^H(G).$$

This is tempting. Can we argue, for instance by symmetry, that $P^H(G) = P^H(B) = \frac{1}{2}$? If so, then $P^H(G \mid S) = \frac{10}{7}\frac{1}{2} = \frac{5}{7}$. The new evidence that the sample passed the test made it much more likely that the material is good. This seems satisfying and exact. There is a nice picture, given in Figure 1.

The real problem in this analysis is of course with the assignment of the prior probabilities. Should they come from experiment, from general previous experience, or from subjective judgement? Can they be the result of pure reason, as suggested by the symmetry argument? Or should they be regarded as a mere technical device for directing a practical decision? This problem has a long history, and its status is still controversial.

Laplace's Law of Succession. In a passage in his *Théorie Analytique des Probabilités* (1812) Laplace calculated the probability that the sun will rise tomorrow, given that it has risen every day for the past 5,000 years. If one considers 5,000 years to be $n = 1,826,213$ days, then his probability is $(n+1)/(n+2) = 1,826,214/1,826,215$, enough to give considerable reassurance.

Here is one way to think of his calculation. Consider a Solar System. Each day it flips a biased coin with probability θ of heads to decide whether to have the sun rise. So given this Solar System, the probability that the sun will rise tomorrow is θ itself, quite independent of whatever has happened before.

Laplace asked: What if we know that we are in a Solar System where the sun has already risen n days in a row? A first thought might be that θ is the value that makes this most likely to happen. This would imply that $\theta = 1$. However, for Laplace this answer seemed to express almost too much confidence in what, after all, is a limited amount of data.

Here is one possible line of reasoning that Laplace could have used. A Solar System Construction Utility has created a huge (perhaps infinite) number of Solar Systems that have various values of θ. Furthermore, these values are evenly spread over the unit interval. Whether the sun rises

tomorrow depends on which Solar System is taken and on how the corresponding coin flip turns out. Without any further information we would say that this probability is $1/2$, because when we look at all Solar Systems and all corresponding coin flips, the fraction with heads is the same as the fraction with tails. But not every Solar System has experienced n days in a row where the sun rose. Knowing that we have such a Solar System gives us useful information. With this extra information the values of θ are not spread out uniformly any more; in fact, they are concentrated rather near (but not equal to) $\theta = 1$. What then is the probability of the sun rising tomorrow for this restricted class of Solar Systems? To find this, look at the fraction of these Solar Systems and corresponding coin flips for which the result is heads. Since in most of these Solar Systems the θ is very close to one, the fraction is itself very close to one. Laplace's calculation showed that it is $(n+1)/(n+2)$.

Laplace did not need to believe in multiple Solar Systems to make his argument. Another possible interpretation of the same mathematics is that there is only one Solar System. An initial state of ignorance is followed by increasing knowledge at each sunrise. In this view the probability calculation is an indication of the current state of knowledge.

Some have considered Laplace's argument absurd, but E. T. Jaynes, the author of the book under review, regards the technique as important and useful and defends Laplace against his critics. He admits that Laplace's choice of example is unfortunate, but he quotes Laplace himself stating that this particular calculation omits relevant information from celestial mechanics. As for the 5,000 years, Jaynes writes in a footnote (p. 564):

> Some passages in the Bible led early theologians to conclude that the age of the world is about 5,000 years. It seems that Laplace at first accepted this figure, as did everyone else. But it was during Laplace's lifetime that dinosaur remains were found almost under his feet (under the streets of Montmartre in Paris), and interpreted correctly by the anatomist Cuvier. Had he written this near the end of his life, we think that Laplace would have used a figure vastly greater than 5,000 years.

A more mathematical derivation of the Laplace "law of succession" is as follows. Suppose the probability that the sun will rise each day is θ, and given this θ the events for different days are independent. Consider a sequence of n consecutive days. A day counts as a success if the sun rises that day. The probability of success on a certain specified subset of r days and failure on the remaining

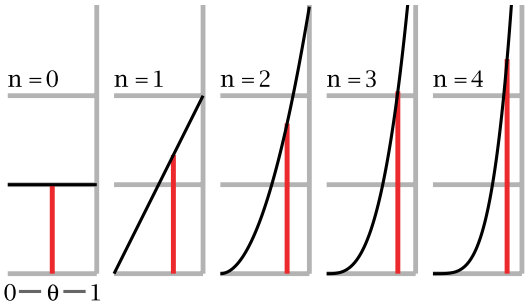


Figure 2. Sunrises. According to Laplace, each day the sun rises the distribution for the probability of another sunrise moves to the right and becomes more peaked near one.

$n - r$ days is $\theta^r(1-\theta)^{n-r}$. Suppose (with Bayes and Laplace) that this θ is random and uniformly distributed on the unit interval. That is, the prior measure expressing the state of ignorance is $d\theta$. The joint probability for a particular pattern x and a particular probability near θ is the product $\theta^r(1-\theta)^{n-r}\,d\theta$, where r is the the total number of successes. Given the extra information of a particular pattern x with exactly r successes, the conditional probabilities for θ are calculated with

$$dP^H(\theta \mid x) = \frac{\theta^r(1-\theta)^{n-r}}{\int_0^1 \theta^r(1-\theta)^{n-r}\,d\theta}\,d\theta = \frac{1}{B(r+1, n-r+1)}\theta^r(1-\theta)^{n-r}\,d\theta. \tag{6}$$

The constant $B(r+1, n-r+1)$ is the conventional notation for the value of the integral in the denominator; it ensures that $dP^H(\theta \mid x)$ defines a probability measure. So given the historical information, the probability of yet one more sunrise is

$$\int_0^1 \theta\, dP^H(\theta \mid x) = \frac{1}{B(r+1, n-r+1)} \times \int_0^1 \theta^{r+1}(1-\theta)^{n-r}\,d\theta = \frac{r+1}{n+2}. \tag{7}$$

In the example of Laplace the sun rose each of the n days, so $r = n$. The picture for this case is given in Figure 2.

A main objection to Laplace's calculation was to his use of prior probability to express knowledge (or ignorance). If the probability of the sun rising tomorrow belonged to a sequence of numbers in the unit interval generated by a well-constructed random number generator (the Solar System Construction Utility), then there would be no problem. But here it seems to be a case of probability without a frequency interpretation.

Frequency Versus State of Knowledge. The mathematical theory of probability is consistent independent of any interpretation. However, many consumers of this theory regard *probability* as a mathematical construct that predicts *frequency*.

Frequency in this context means relative frequency or sample proportion, the number of successes divided by the total number of trials. The law of large numbers (which models the situation where the number of trials is large) shows that this interpretation has internal consistency.

Jaynes considers the frequency interpretation of probability as far too limiting. Instead, probability should be interpreted as an indication of a state of knowledge or strength of evidence or amount of information within the context of inductive reasoning. The foundation of his discussion is an axiomatic derivation (following R. T. Cox) of rules of inference for degrees of plausibility. At one point he suggests that instead of using a probability p in the interval $[0, 1]$, one could use an odds ratio $p/(1-p)$ in the interval $[0, +\infty]$, or even a measure of evidence $\log(p/(1-p))$ in the interval $[-\infty, +\infty]$. In the Laplace example, where the odds ratio is $n+1$ with n well above a million, the evidence $\log(n+1)$ (using base 10) would be above 6 bels (60 decibels). Such a number represents extremely strong evidence.

Furthermore, he argues that probability as empirical frequency hardly ever occurs in a pure form. Thus he writes (p. 337):

> In most recent treatments of probability theory, the writer is concerned with situations where a strong connection between probability and frequency is taken for granted—indeed, this is usually considered essential to the very notion of probability. Nevertheless, the existence of such a strong connection is clearly only an ideal limiting case, unlikely to be realized in any real application. For this reason, the laws of large numbers and limit theorems of probability theory can be grossly misleading to a scientist or engineer who naively supposes them to be experimental facts, and tries to interpret them literally in his problems.

The goal of his presentation is to blur the distinction between probability theory (where one goes from the mathematical model to data) and statistical inference (where one goes from data to a mathematical model). Each probability value is to be a conditional probability of an event given other events. The emphasis is on discrete probabilities. The purpose of such a probability is to express a "judgement", not a physical fact (p. 86). All existing evidence must be taken into account, and when there is little existing evidence one must use prior probabilities for the hypotheses that express this ignorance. The fundamental tool in statistical inference is Bayes's theorem, which gives a method of passing from the conditional probability of data given a hypothesis to the conditional probability of a hypothesis given the data. In his view, there are heros (R. T. Cox, H. Jeffreys) and villains (R. A. Fisher and subsequent frequentists). The tone is polemical.

Jaynes is not the first writer to make such an argument; it is typical of many Bayesian statisticians. However, his background was not in statistics, but in physics. As a physicist he was prominent in a number of areas; in particular he is noted for promoting the method of maximum entropy, a technique for assigning prior probability distributions in order to reason about physical systems about which one has incomplete information. This led him to a more general interest in Bayesian inference and to the project of writing a book on the subject that would serve both as a reference and as a text book. At the time of his death in 1998 the book was partially finished, with some chapters still missing. The manuscript was edited by G. Larry Bretthorst, and now it appears as a published book of well over seven hundred pages. It should be mentioned that Bayesian statistics is not a new subject: it is treated in many other books (for instance [4] or the more advanced [1], [3]).

Maximum Likelihood Parameter Estimation. Laplace's calculation may seem frivolous, but consider the following example. A disease has jumped from an animal host to humans. An international health conference is convened to recommend action. The participants need to know the survival rate (survival probability) to high accuracy in order to make further decisions. This rate is a number between 0 and 1; these numbers constitute a set of infinitely many hypotheses. The existing sample of infected humans is rather small, of size only $n = 16$. Of these a certain number r survived. This experimental number should be available in a few hours. What is the appropriate estimate, based on this evidence, of the survival rate from this disease in humans? There is no time to collect new evidence; only one day is left to evaluate the existing data.

The most obvious answer is the sample proportion r/n. It turns out, however, that this is not the only possible answer. This raises the question of how one justifies such a calculation.

The framework for a discussion of such issues is a situation where there are a number of possible *parameter* values or hypotheses. For each parameter value θ there is a probability measure $P(\cdot \mid \theta)$, representing the probabilities for various sets of *data* values. This is ordinarily called the *sample distribution*. The question is how to use the data to estimate the parameter. In the example each parameter value is a number θ between zero and one. This is the unknown survival rate. The data point x is the pattern of survival and loss; it has 2^n possible values, each 1 or 0. In what follows the only function of the data that matters is $r = \sum x$,

the number of survivors out of the n patients in the sample. Figure 3 shows the probabilities of the possible r/n values as determined by the sample distribution.

Here is one inference procedure that is simple, universal, and works reasonably well in many practical situations. For fixed θ consider the density of $P(\cdot \mid \theta)$ with respect to some reference measure Q on the data space, so

$$dP(x \mid \theta) = f(x, \theta)\, dQ(x). \tag{8}$$

For each parameter value θ this density $f(x, \theta)$ is a function of the data value x. Alternatively, for each data value x one may consider $f(x, \theta)$ as a function of the parameter value θ; this is called the *likelihood function*. It is uniquely determined up to a factor that depends only on x. The procedure is to take the actual data, compute the parameter value that maximizes the likelihood function, and use this for the estimate of the true parameter value.

In the example the probability of survival for each patient is θ, and the probability of death is $1 - \theta$. The likelihood function defined with respect to counting measure on the 2^n possible outcomes is

$$f(x, \theta) = \theta^r (1 - \theta)^{n-r}, \tag{9}$$

where $r = \sum x$ is the number of survivors. It is easy to see that the maximum likelihood value of θ is r/n, the sample proportion.

Bayes Theory. The new ingredient in Bayes theory is the *Bayes prior distribution*, a probability measure P^H on the parameter space. It is assumed given before the current experiment is performed. (This measure could also be called the *unconditional parameter distribution*.) With this marvelous object one can construct other useful quantities. There is a new probability measure P for the data, called the *Bayes unconditional data distribution*. This is obtained by averaging the sample distribution with respect to the Bayes prior distribution. In this context, the sample distribution $P(\cdot \mid \theta)$ could be called the *conditional data distribution*. There is a *normalized likelihood function* $\hat{f}$. This is a function of the data variable x and the parameter variable θ. It is defined by

$$dP(x \mid \theta) = \hat{f}(x, \theta)\, dP(x). \tag{10}$$

In the language of measure theory $\hat{f}(\cdot, \theta)$ is the density or Radon-Nikodym derivative of the conditional data distribution $P(\cdot \mid \theta)$ with respect to the unconditional data distribution P.

The most important object in Bayesian analysis is the *Bayes posterior distribution*. This could also be called the *conditional parameter distribution*. It is the distribution of the parameters given the data, and it represents the updated account of what is known after the experimental results are in. Mathematically, it is a family of probability measures $P^H(\cdot \mid x)$ on the parameter space that give conditional probabilities for fixed data values x. *Bayes's theorem* says that the Bayes posterior distribution is given by multiplying the Bayes prior distribution by the normalized likelihood function. That is, the posterior probabilities given the data x are given by integrating

$$dP^H(\theta \mid x) = \hat{f}(x, \theta)\, dP^H(\theta). \tag{11}$$

Thus $\hat{f}(x, \cdot)$ is the density or Radon-Nikodym derivative of the conditional parameter distribution $P^H(\cdot \mid x)$ with respect to the unconditional parameter distribution P^H. This remarkable result is named after the Reverend Thomas Bayes, who presented a version of it in 1763.

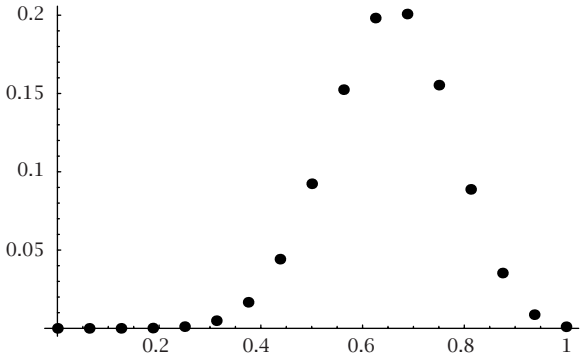


Figure 3. Distribution of the sample proportion. In this example the parameter value is $\theta = 0.65$, and the sample size is $n = 16$. The plot shows the probabilities associated with each value of r/n, for $r = 0, 1, 2, \ldots, n$. These are computed from the binomial distribution. For each value of θ between 0 and 1 there is a corresponding such plot. The method of maximum likelihood says to choose the value of θ that makes the experimental result most probable.

One version of Bayesian statistical inference is the following. Acquire somehow on the basis of experience or intuition an appropriate Bayes prior distribution P^H. Perform the experiment and acquire the actual data x. The final (or updated) product is the Bayes posterior probability distribution $P^H(\cdot \mid x)$ corresponding to the data. This summarizes the revised state of knowledge and can be used however one wishes.

Bayes Decision Theory. Decision theory is another ingredient in statistical inference. The idea is to be explicit about actions and their consequences. There is a set of *actions*. In *estimation* this may consist of the values of certain parameter components, while in *hypothesis testing* it may be a set with as few as two points. There is also a *loss function* L that takes as inputs an action a and a state of nature θ and calculates a corresponding loss $L(a, \theta)$. This loss function may be measured crudely in dollars, or perhaps better in some sort

of utility units. Finally, a statistical inference procedure is a *decision function* δ from data values to possible actions. Thus decision theory moves from statistical inference as a way of gaining knowledge toward statistical inference as a guide to practical action, taking into account the economic consequences.

It is not difficult to combine Bayes theory and decision theory. Jaynes seems to be willing to do this, though with some reluctance. The procedure is simple. First look at the data x. Then consider the *Bayes posterior risk*

$$r_x(a) = \int L(a,\theta)\, dP^H(\theta \mid x) \tag{12}$$

of action a. The Bayes decision $\delta(x)$ is the a that minimizes this quantity. Its risk is then $r_x(\delta(x))$.

The Bayes posterior risk of action a given data x may be written explicitly as

$$r_x(a) = \int L(a,\theta)\hat{f}(x,\theta)\, dP^H(\theta). \tag{13}$$

For the purposes of finding the Bayes decision $\delta(x)$ by minimization one can replace the normalized likelihood function $\hat{f}(x,\theta)$ by any other likelihood function $f(x,\theta)$, in particular one that is defined independent of the Bayes prior distribution. Then the Bayes prior distribution P^H enters only in combination with the loss function $L(a,\theta)$ as the product $L(a,\theta)\, dP^H(\theta)$. As we shall see below, Jaynes has vigorous comments on this point.

Bayes Versus Frequentist. A frequentist might have no problem using a Bayes decision function δ. However, in this view the Bayes prior distribution P^H is merely a convenient way of selecting the δ. As an index of performance, the natural object is the *risk function* of the decision procedure δ given by

$$R_\delta(\theta) = \int L(\delta(x),\theta)\, dP(x \mid \theta). \tag{14}$$

This makes no reference either to the Bayes prior distribution or to actual experimental results.

A critic might ask: Why show me a risk function defined by a sum or integral over all data values, when you already have the actual data at hand? The frequentist answer is that the risk function shows how the procedure works in the long run, in all possible circumstances. One neither knows the actual circumstance, nor whether today's experiment gave an appropriate decision. A statistician can at best hope to do well most of the time, but on each individual experiment fate works its will.

The decision function δ maps data to actions. Suppose that there is also a reasonable estimation function δ' from data to parameters. (If the action space is the parameter space, then δ may serve as δ'.) Then the frequentist statistician can estimate the risk for the particular experimental data x at hand by reporting the number $R_\delta(\delta'(x))$. This experimental number attempts to describe the risk of the decision; it may well differ from the Bayes posterior risk of the same decision.

One would think that natural selection would have eliminated one of the schools of statisticians by now. Why has this not happened? A mathematical identity may provide a clue. It says that the average of the Bayes posterior risk over all data values is the average of the frequentist risk over all parameter values. Thus

$$\int r_x(\delta(x))\, dP(x) = \int R_\delta(\theta)\, dP^H(\theta). \tag{15}$$

Perhaps if the priors are not too crazy, then the two risk estimates are roughly compatible.

Bayes Parameter Estimation. Return to the example of the patients with the new disease. This example is so simple that it is does not fairly represent Bayesian statistics or indeed statistical inference in general, but it illustrates some of the issues.

The unknown parameter is a number θ between zero and one. This is the survival rate for humans in general. The experimental sample consists of n patients. For each patient survival is indicated by 1 and death by 0, so the data vector x consists of a list of n values, each 1 or 0. The task is to use the data to estimate the parameter θ and to present this estimate to the health conference.

Decision theory needs a loss function. Economics experts assisting in the preparation of the conference have concluded that the problem at hand is an estimation problem with loss function $L(a,\theta) = (a-\theta)^2$. One might suspect that this choice may be more for mathematical convenience than the result of a deep analysis of the consequence of error. However it allows the analysis to proceed.

Suppose that there is prior information that the human survival rate for new diseases that jump from animal hosts is usually above 50 percent. In fact, a survival rate of about 75 percent might be reasonable, but this is not a firm figure. One simple device for arriving at a Bayes prior distribution that is mainly distributed above 50 percent is to consider that it would have about as much influence on the final result as if the sample size were increased by $m = 8$ and the number of successes were increased by $s = 6$. Of course there is no claim that there was ever an actual sample of size $m = 8$ with $s = 6$ successes. This is just a device for defining a Bayes prior distribution that is centered near 75 percent (since $s/m = 6/8 = 0.75$) but has considerable spread (since m is so small).

The mathematical representation of this distribution is

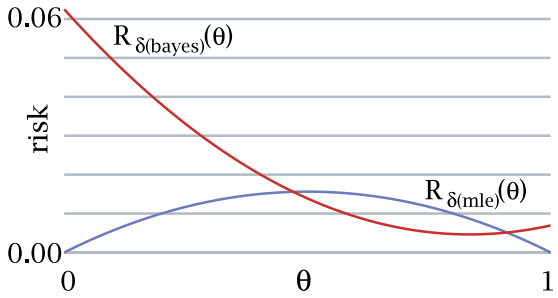


Figure 4. Procedures for estimating survival rates. Risk functions for the maximum likelihood estimator and for a Bayes estimator. Is one less risky than the other?

$$(16)\quad dP^H(\theta) = \frac{1}{B(s, m-s)}\theta^{s-1}(1-\theta)^{m-s-1}\,d\theta.$$

Here is where this comes from. Complete ignorance is expressed by a distribution proportional to $\theta^{-1}(1-\theta)^{-1}$. (Never mind that this starting point is not a probability distribution.) Each success adds one more power of θ, and each failure adds one more power of $(1-\theta)$. The normalization constant makes this a measure with total probability one, at least when s and $m-s$ are both greater than zero. This particular prior distribution is convenient mathematically, and it is indeed spread out over the required range of parameter values.

The effect of the experimental data is captured in the normalized likelihood function, which in this case is

$$(17)\quad \hat{f}(x,\theta) = \frac{B(s, m-s)}{B(r+s, n+m-s)}\theta^r(1-\theta)^{n-r},$$

where $r = \sum x$ is the number of successes. The Bayes posterior distribution is the product of the normalized likelihood function with the Bayes prior distribution, that is,

$$(18)\quad dP^H(\theta \mid x) = \frac{1}{B(r+s, n+m-s)}\theta^{r+s-1}(1-\theta)^{n+m-r-s-1}\,d\theta.$$

This posterior distribution incorporates more information and therefore is more concentrated than the prior distribution. The Bayes estimate of θ (under squared error loss) is the conditional mean of the Bayes posterior distribution, which is given by

$$(19)\quad \delta(x) = \frac{r+s}{n+m} = \frac{n}{n+m}\frac{r}{n} + \frac{m}{n+m}\frac{s}{m},$$

where again r is the number of successes in the data set x. Indeed, it says to act as if the sample size were increased by m and the number of survivals were increased by s. This gives a combination of the maximum likelihood estimate r/n and the nonexperimental ratio s/m.

Even a frequentist may use such an estimator. From this point of view it is reasonable to look at the risk function R_δ, perhaps forgetting that the particular estimator δ has Bayesian origins. The advantage of considering the risk function is that it makes no reference to the Bayes prior distribution or to the experimental data values. With squared error loss this risk function is the sum of the variance with the square of the bias:

$$(20)\quad R_\delta(\theta) = \frac{n^2}{(n+m)^2}\frac{\theta(1-\theta)}{n} + \frac{m^2}{(n+m)^2}\left(\theta - \frac{s}{m}\right)^2.$$

The frequentist may consider various possible choices for the m and s (with fixed sample size n) that define the estimator. The problem is that one is not clearly better than the other. Less risk for one range of θ values is compensated by more risk at other θ values, and θ is unknown. However, if n is large enough, then perhaps at least one of the risk functions is acceptable to the frequentist. For sample size $n = 16$ the maximum likelihood risk function $(s = 0, m = 0)$ is contrasted with the Bayes risk function $(s = 6, m = 8)$ in Figure 4. The Bayes risk function has less risk in the range of θ values that the Bayesian considers most probable, but more risk elsewhere. These are not the only possibilities. A frequentist worried about the worst that could happen might prefer the estimator with $s/m = 1/2$ and $m = \sqrt{n}$, since it has constant risk.

In the example the sample size is $n = 16$, so the decision function that gives the Bayes estimate is $(r+s)/(n+m) = (r+6)/(16+8) = (r+6)/24$. This is a biased estimate, but that may be just what is needed. In fact, the Bayesian might argue that it is better to bias the result from the sample proportion toward the prior value of 0.75. After all, if the new findings deviate greatly from this, an experienced observer might well suspect a misleading run of bad luck in a relatively small sample. In short, it is foolish to ignore relevant evidence.

If the prior hypothesis of something like an average 75 percent survival rate is even roughly correct, then using it to bias the estimate toward this value is perhaps quite helpful. However, if the new disease, unlike its predecessors, is highly fatal (θ near zero), then the use of the prior hypothesis could lead to a nasty bias in the wrong direction. Has anyone thought of this possibility? What is the actual situation? Nobody knows for sure, and, in our scenario, the decisive meeting is just hours away. Suppose the decision is to use the particular Bayes estimate at hand, as the best way of updating what one hopes is the best available information.

The experimental result just came in. The data for the sixteen patients are

$$(21)\quad x = (1,0,0,0,1,1,0,0,0,0,1,0,1,0,1,0).$$

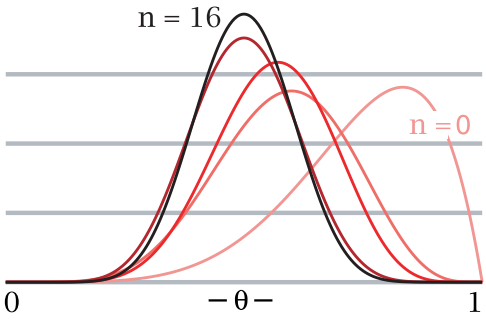


Figure 5. Disease survivors. As information on the seriousness of the disease accumulates, the Bayes distribution for the chance of survival becomes more peaked and moves to the left. Posterior distributions are shown for $n = 4, 8, 12, 16$.

There are only $r = 6$ survivors out of the 16 patients, so the Bayes estimate of the population survival probability is $12/24 = 0.5$. This is considerably above the maximum likelihood estimate $6/16 = 0.375$, but, given the assumptions, it is the best combination of prior knowledge and new evidence.

The prior distribution centered at 0.75 and posterior distributions that reflect accumulating evidence are shown in Figure 5. The final posterior distribution is centered at 0.5, well above the experimental proportion of 0.375, but reflecting its influence.

Ignorance Priors. Since the work of A. Wald relating Bayesian theory to decision theory there is not much dispute about whether Bayesian methods are reasonable; clearly in many cases they are [2]. The misgivings are about the interpretation of the prior probability measure P^H in circumstances when it has no natural frequency interpretation. There are really two problems: How to choose the prior? In what sense is the choice correct?

Jaynes reviews several proposals for how to choose the prior. In the case of complete ignorance one should look for *group invariance* to choose the prior. Thus, for instance, in the case of a location parameter θ_l this would be a multiple of $d\,\theta_l$, while in the case of a scale parameter $\theta_s > 0$ this would be a multiple of $d\theta_s/\theta_s$. The problem that these are not probability measures may be overcome by various devices.

In the case of the unit interval a natural prior for θ with $0 < \theta < 1$ is given by taking the evidence $\theta_l = \log(\theta/(1-\theta))$ as a location parameter, so that a multiple of

$$d\theta_l = \frac{1}{\theta(1-\theta)}\,d\theta \tag{22}$$

is the preferred choice.

This, by the way, is not the choice made by Laplace. He avoided the maximum likelihood estimator r/n, perhaps on the ground that a run of $r = n$ sunrises should not suggest that the sun will certainly rise. Instead his estimator was $(r+1)/(n+2)$. Laplace was using the Bayes prior $d\theta$ for which there is a prior bias toward $1/2$ as strong as that which would come from a sample of size 2. That is, equation (6) corresponds to equations (16) and (18) with $s = 1$ and $m = 2$.

Say that there there is already a non-informative prior P^H, presumably a uniform distribution determined by group invariance. One wants to incorporate additional information to the effect that a certain quantity h has specified expectation equal to E. The statistician needs a new prior P^H_β such that

$$\int h(\theta)\,dP^H_\beta(\theta) = E. \tag{23}$$

Jaynes argues that the least informative such P^H_β is given by the *maximum entropy* principle. The result of using this principle is that

$$dP^H_\beta(\theta) = \frac{1}{Z_\beta}e^{-\beta h(\theta)}\,dP^H(\theta). \tag{24}$$

For each fixed β the constant Z_β is chosen so that P^H_β is a probability measure. The constant β is then to be chosen so that the expectation is equal to E. It is no accident that such formulas are fundamental in equilibrium statistical mechanics. In that context $h(\theta)$ is the energy of configuration θ, while β is inversely proportional to the temperature, and E is the expected energy corresponding to equilibrium at the given temperature.

The obvious objection is that there may be no group invariance available to determine the prior expressing total lack of information. Furthermore, whether or not there is group invariance, there is a general question of how to justify a method of choosing the Bayes prior distribution.

Jaynes discusses writers who are reluctant to introduce the notion of prior probability. He states (p. 419):

> But these same writers do not hesitate to pull a completely arbitrary loss function out of thin air, and proceed with the calculation! Our equations show that if the final decision depends strongly on which prior probability assignment we choose, it is going to depend just as strongly on which particular loss function we use. If one worries about arbitrariness in the prior probabilities, then, in order to be consistent, one ought to worry just as much about arbitrariness in the loss functions. If one claims (as sampling theorists did for decades and as some still do) that uncertainty as to the proper choice of prior probabilities invalidates the Laplace-Bayes theory,

> then, in order to be consistent, one must claim also that uncertainty as to the proper choice of loss functions invalidates Wald's theory.

Jaynes admits that there has been work on the problem of determining loss/utility functions. However he claims (p. 420):

> Such constructions, if one can transfer them into a computer, will be better than nothing; but they are clearly desperation moves in lieu of a really satisfactory formal theory such as we have in the principles of maximum entropy and transformation groups for priors.

He even argues that "a change in prior information which affects the prior probabilities could very well induce a change in the loss function as well" (p.424). In a footnote he gives a literary illustration:

> Quasimodo, condemned by an accident of Nature to be something intermediate between man and gargoyle, wished that he had been made a whole man. But, after learning about the behavior of men, he wished instead that he had been made a whole gargoyle: 'O, why was I not made of stone like these?'

Opinions. Jaynes has little patience with measure theory and scant interest in stochastic processes. For instance, he states that "those who persist in trying to calculate probabilities conditional on propositions of probability zero, have before them an unlimited field of opportunities for scholarly research and publication—without hope of any meaningful and useful results" (p. 485). This is consistent with the fact that he sees little use for the terms "sigma-algebra, Borel field, Radon-Nikodym derivative" (p. 676). As for stochastic processes, he asserts that "the most valuable applications of probability theory are concerned with incomplete information and have nothing to do with those so-called 'random phenomena' which are still undefined in theory and unidentified in Nature" (p. 709). Clearly his book is not the place to learn modern probability theory.

The central thesis of the book occurs in the following passage (p. xxii):

> Our theme is simply: *probability theory as extended logic*. The 'new' perception amounts to the recognition that the mathematical rules of probability theory are not merely rules for calculating frequencies of 'random variables'; they are also the unique consistent rules for conducting inference (i.e. plausible reasoning) of any kind, and we shall apply them in full generality to that end.

For the author, "maximum entropy is the appropriate (safest) tool when we have little knowledge beyond the raw data" (p. xxiv). Another passage summarizes his view of the role of uniform or maximum entropy priors (p. xxiv):

> Bayesian and maximum entropy methods differ in another respect. Both procedures yield the optimal inferences from the information that went into them, but we may choose a model for Bayesian analysis; this amounts to expressing some prior knowledge—or some working hypothesis—about the phenomena being observed. Usually such hypotheses extend beyond what is directly observable in the data, and in that sense we might say that Bayesian methods are—or at least may be—speculative. If the extra hypotheses are true, then we may expect that Bayesian results will improve on maximum entropy; if they are false, the Bayesian inferences will likely be worse.

This prescription of Bayesian analysis as a universal recipe is too enthusiastic. A working hypothesis may be an incorrect hypothesis. Furthermore, the maximum entropy method is to be used in a situation of almost complete ignorance, where only a few moments are known. It depends ultimately on some uniform prior, typically defined by group invariance. From a frequentist point of view, the use of a uniform distribution as a prior seems to represent a definite knowledge claim. If this claim is far off the mark, then what is the virtue of making it? Perhaps a Bayes decision is a procedure to be evaluated in the same way as any other such procedure. If there is enough data the risk can be kept acceptably low.

A positive feature of the book is that the author thinks for himself (within this ideological framework) and writes in a lively way about all sorts of things. It is worth dipping into it if only for vivid expressions of opinion. The annotated References and Bibliography are particularly good for this. For instance, in the annotation to C. J. Preston's excellent book on *Gibbs States on Countable Sets* [5] Jaynes writes (p. 715):

> Here we have the damnable practice of using the word *state* to denote a probability *distribution*. One cannot think of a more destructively false and misleading terminology.

There are many books on Bayesian statistics, but few with this much color. The term "state", by the

way, has its origin in quantum theory, and its transfer to probability is quite natural. On the other hand, for Jaynes quantum theory presents problems (p. 328):

> As Bohr stressed repeatedly in his writings and lectures, present quantum theory can only answer questions of the form: 'If this experiment is performed, what are the possible results and their probabilities?' It cannot, as a matter of principle, answer any question of the form: 'What is really happening when…?' Again, the mathematical formalism of present quantum theory, like Orwellian *newspeak*, does not even provide the vocabulary in which one could ask such a question.

This is an unusual perspective from a physicist, but others have been troubled by this issue [6]. Perhaps Jaynes has a point.

Acknowledgement. I thank my colleagues Rabi Bhattacharya, Hermann Flaschka, Don Myers, and Joe Watkins for comments on drafts of this review. I am particularly grateful to Bill Casselman for his magical multi-colored intervention at the editorial stage.

References

[1] James O. Berger, *Statistical Decision Theory and Bayesian Analysis*, 2nd edition, Springer-Verlag, New York, 1993.
[2] Thomas S. Ferguson, *Mathematical Statistics: A Decision Theoretic Approach*, Academic Press, New York, 1967.
[3] J. A. Hartigan, *Bayes Theory*, Springer-Verlag, New York, 1983.
[4] Peter M. Lee, *Bayesian Statistics: An Introduction*, 3rd edition, Arnold, London, 2004.
[5] Christopher J. Preston, *Gibbs States on Countable Sets*, (Cambridge Tracts in Mathematics 681), Cambridge University Press, London, 1974.
[6] David Wick, *The Infamous Boundary: Seven Decades of Heresy in Quantum Physics*, Copernicus (Springer-Verlag), New York, 1995.