1

The subjective interpretation of probability

Reverend Thomas Bayes (born circa 1702; died 1761) was the oldest son of Reverend Joshua Bayes, who was one of the first ordained nonconformist ministers in England. Relatively little is known about the personal life of Thomas Bayes. Although he was elected a Fellow of the Royal Society in 1742, his only known mathematical works are two articles published posthumously by his friend Richard Price in 1763. The first dealt with the divergence of the Stirling series, and the second, "An Essay Toward Solving a Problem in the Doctrine of Chances," is the basis of the paradigm of statistics named for him. His ideas appear to have been independently developed by James Bernoulli in 1713, also published posthumously, and later popularized independently by Pierre Laplace in 1774. In their comprehensive treatise, Bernardo and Smith (1994, p. 4) offer the following summarization of *Bayesian statistics*:

> Bayesian Statistics offers a rationalist theory of personalistic beliefs in contexts of uncertainty, with the central aim of characterizing how an individual should act in order to avoid certain kinds of undesirable behavioral inconsistencies. The theory establishes that expected utility maximization provides the basis for rational decision making and that Bayes' Theorem provides the key to the ways in which beliefs should fit together in the light of changing evidence. The goal, in effect, is to establish rules and procedures for individuals concerned with disciplined uncertainty accounting. The theory is not descriptive, in the sense of claiming to model actual behavior. Rather, it is prescriptive, in the sense of saying "if you wish to avoid the possibility of these undesirable consequences you must act in the following way."

Bayesian econometrics consists of the tools of Bayesian statistics applicable to the models and phenomena of interest to economists. There have been numerous axiomatic formulations leading to the central unifying Bayesian prescription of maximizing subjective utility as the guiding principle of Bayesian statistical analysis. Bernardo and Smith (1994, Chapter 2) is a valuable segue into this vast literature. Deep issues are involved regarding meaningful separation of probability and utility assessments, and we do not address these here.

Non-Bayesians, who we hereafter refer to as *frequentists*, argue that situations not admitting repetition under essentially identical conditions are not within the realm of statistical enquiry, and hence "probability" should not be used in such situations. Frequentists define the probability of an event as its long-run relative frequency. This frequentist interpretation cannot be applied to (i) unique, once-and-for-all type of phenomenon, (ii) hypotheses, or (iii) uncertain past events. Furthermore, this definition is nonoperational since only a finite number of trials can ever be conducted. In contrast, the desire to expand the set of relevant events over which the concept of probability can be applied, and the willingness to entertain formal introduction of "nonobjective" information into the analysis, led to the subjective interpretation of probability.

Definition 1.1 (Subjective interpretation of probability) Let κ denote the body of knowledge, experience, or information that an individual has accumulated about the situation of concern, and let A denote an uncertain event (not necessarily repetitive). The *probability of A afforded by κ* is the "degree of belief" in A held by an individual in the face of κ.

Since at least the time of Ramsey (1926), such degrees of belief have been operationalized in terms of agreed upon reference lotteries. Suppose you seek your degree of belief, denoted $p = P(A)$, that an event A occurs. Consider the following two options.

1. Receiving a small reward \$$r$ if A occurs, and receiving \$0 if A does not occur.
2. Engaging in a lottery in which you win \$$r$ with probability p, and receiving \$0 with probability $1 - p$.

If you are indifferent between these two choices, then your degree of belief in A occurring is p. Requiring the reward to be "small" is to avoid the problem of introducing utility into the analysis; that is, implicitly assuming utility is linear in money for small gambles.

Bruno de Finetti considered the interesting situation in which an individual is asked to quote betting odds (ratios of probabilities) on a set of uncertain events and accept any wagers others may decide to make about these events. According to de Finetti's *coherence principle* the individual should never assign "probabilities" so that someone else can select stakes that guarantee a sure loss *(Dutch book)* for the individual whatever the eventual outcome. A sure loss amounts to the "undesirable consequences" contained in the earlier quote of Bernardo and Smith. This simple principle implies the axioms of probability discussed in Abadir, Heijmans, and Magnus (2006, Chapter 1) except that the additivity of probability of intersections for disjoint events is required to hold only for *finite* intersections. Nonetheless, for purposes of convenience, we consider only countably additive probability in this volume.

De Finetti's Dutch book arguments also lead to the standard rule for conditional probability. Consider two events A and B. By using the *factorization rule for conditional probability* [Abadir et al. (2006, p. 5)],

$$P(A \text{ and } B) = P(A)P(B|A) = P(B)P(A|B),$$

the simplest form of Bayes' theorem follows immediately:

$$P(B|A) = \frac{P(B)P(A|B)}{P(A)}.$$

In words, we are interested in the event B to which we assign the *prior* probability $P(B)$ for its occurrence. We observe the occurrence of the event A. The probability of B occurring given that A has occurred is the *posterior* probability $P(B|A)$. More generally, we have the following result.

Theorem 1.1 (Bayes' theorem for events) Consider a probability space $[S, \tilde{A}, P(\cdot)]$ and a collection $B_n \in \tilde{A}$ $(n = 1, 2, \ldots N)$ of mutually disjoint events such that $P(B_n) > 0$ $(n = 1, 2, \ldots, N)$ and $B_1 \cup B_2 \cup \cdots \cup B_N = S$. Then

$$P(B_n|A) = \frac{P(A|B_n)P(B_n)}{\sum_{j=1}^{N} P(A|B_j)P(B_j)} \quad (n = 1, 2, \ldots, N) \tag{1.1}$$

for every $A \in \tilde{A}$ such that $P(A) > 0$.

Proof: The proof follows directly upon noting that the denominator in (1.1) is $P(A)$.

An important philosophical topic is whether the conditionalization in Bayes theorem warrants an unquestioned position as the model of learning in the face of knowledge of the event A. Conditional probability $P(B|A)$ refers to *ex ante* beliefs on events not yet decided. *Ex post* experience of an event can sometimes have a striking influence on the probability assessor (e.g., experiencing unemployment, stock market crashes, etc.), and the experience can bring with it more information than originally anticipated in the event. Nonetheless, we adopt such conditionalization as a basic principle.

The subjective interpretation reflects an individual's personal assessment of the situation. According to the subjective interpretation, probability is a property of an individual's perception of reality, whereas according to classical and frequency interpretations, probability is a property of reality itself. For the subjectivist there are no "true unknown probabilities" in the world out there to be discovered. Instead, "probability" is in the eye of the beholder.

Bruno de Finetti assigned a fundamental role in Bayesian analysis to the concept of *exchangeability*, defined as follows.

Definition 1.2 A finite sequence Y_t $(t = 1, 2, \ldots, T)$ of events (or random variables) is *exchangeable* iff the joint probability of the sequence, or any subsequence, is invariant under permutations of the subscripts, that is,

$$P(y_1, y_2, \ldots, y_T) = P(y_{\pi(1)}, y_{\pi(2)}, \ldots, y_{\pi(T)}), \tag{1.2}$$

where $\pi(t)(t = 1, 2, \ldots, T)$ is a permutation of the elements in $\{1, 2, \ldots, T\}$. An infinite sequence is exchangeable iff any finite subsequence is exchangeable.

Exchangeability provides an operational meaning to the weakest possible notion of a sequence of "similar" random quantities. It is "operational" because it only requires probability assignments of *observable* quantities, although admittedly this becomes problematic in the case of infinite exchangeability. For example, a sequence of Bernoulli trials is *exchangeable* iff the probability assigned to particular sequences does not depend on the order of "successes" (S) and "failures" (F). If the trials are exchangeable, then the sequences FSS, SFS, and SSF are assigned the same probability.

Exchangeability involves recognizing symmetry in beliefs concerning only observables, and presumably this is something about which a researcher may have intuition. Ironically, subjectivists emphasize observables (data) and objectivists focus on unobservables (parameters). Fortunately, Bruno de Finetti provided a subjectivist solution to this perplexing state of affairs. De Finetti's representation theorem and its generalizations are interesting because they provide conditions under which exchangeability gives rise to an isomorphic world in which we have iid observations conditional on a mathematical construct, namely, a parameter. These theorems provide an interpretation of parameters that differs substantively from the interpretation of an objectivist.

As in the case of iid sequences, the individual elements in an exchangeable sequence are identically distributed, but they are *not* necessarily independent, and this has important predictive implications for learning from experience. The importance of the concept of exchangeability is illustrated in the following theorem.

Theorem 1.2 (de Finetti's representation theorem) Let Y_t $(t = 1, 2, \ldots)$ be an infinite sequence of Bernoulli random variables indicating the occurrence (1) or nonoccurrence (0) of some event of interest. For any finite sequence Y_t $(t = 1, 2, \ldots, T)$, define the average number of occurrences

$$\overline{Y}_T = \frac{1}{T}\sum_{t=1}^{T} Y_t. \tag{1.3}$$

Let $h(y_1, y_2, \ldots, y_T) = \Pr(Y_1 = y_1, Y_2 = y_2, \ldots, Y_T = y_T)$ denote a probability mass function (p.m.f.) reflecting *exchangeable* beliefs for an arbitrarily long finite sequence Y_t $(t = 1, 2, \ldots, T)$, and let $H(y) = \Pr(Y \leq y)$ denote its associated cumulative distribution function (c.d.f.). Then $h(\cdot)$ has the representation

$$h(y_1, y_2, \ldots, y_T) = \int_0^1 L(\theta) dF(\theta), \tag{1.4}$$

where

$$L(\theta) = \prod_{t=1}^{T} \theta^{y_t}(1-\theta)^{(1-y_t)}, \tag{1.5}$$

$$F(\theta) = \lim_{T\to\infty} P_H(\overline{Y}_T \leq \theta), \tag{1.6}$$

and $P_H(\cdot)$ denotes probability with respect to the c.d.f. $H(\cdot)$ corresponding to p.m.f. (1.4).

Proof: See de Finetti (1937) or the simpler exposition of Heath and Sudderth (1976).

Theorem 1.1 implies that it is *as if*, given θ, Y_t $(t = 1, 2, \ldots, T)$ are iid Bernoulli trials where the probability of a success is θ, and the "parameter" θ is assigned a probability distribution with c.d.f. $F(\cdot)$ that can be interpreted as belief about the long-run relative frequency of $\overline{Y}_T \leq \theta$ as $T \rightarrow \infty$. From de Finetti's standpoint, both the quantity θ and the notion of independence are "mathematical fictions" implicit in the researcher's subjective assessment of arbitrarily long *observable* sequences of successes and failures. The parameter θ is of interest primarily because it constitutes a limiting form of predictive inference about the observable $\overline{Y}_T$ via (1.6). The mathematical construct θ may nonetheless be useful. However, Theorem 1.2 implies that the subjective probability distribution need not apply to the "fictitious θ" but only to the *observable* exchangeable sequence of successes and failures. When the c.d.f. is absolutely continuous, so that $f(\theta) = \partial F(\theta)/\partial\theta$ exists, then (1.4) becomes

$$h(y_1, y_2, \ldots, y_T) = \int_0^1 \prod_{t=1}^{T} \theta^{(y_t)}(1-\theta)^{(1-y_t)} f(\theta)d\theta. \tag{1.7}$$

It is clear from (1.4) and (1.7) that exchangeable beliefs assign probabilities acting as if the Y_t's are iid Bernoulli random variables given θ, and then average over values of θ using the weight $f(\theta)$ to obtain a marginal density for the Y_t's. Let $S_T = T\overline{Y}_T$ be the number of successes in T trials. Since there are $\binom{T}{r}$ ways in which to obtain $S_T = r$ successes in T trials, it follows immediately from (1.4) and (1.5) that

$$\Pr(S_T = r) = \binom{T}{r} \int_0^1 \theta^r (1-\theta)^{T-r} dF(\theta) \quad (r = 0, 1, \ldots, T), \tag{1.8}$$

where

$$F(\theta) = \lim_{T\rightarrow\infty} \Pr(T^{-1}S_T \leq \theta). \tag{1.9}$$

Thus, given θ, it follows from (1.8) that exchangeable beliefs assign probabilities acting as if S_T has a binomial distribution given θ, and then average over values of θ using the weight $f(\theta) = \partial F(\theta)/\partial\theta$. Bayes and Laplace suggest choosing the "mixing" distribution $F(\theta)$ for θ to be uniform over $[0, 1]$, in which case (1.8) reduces to

$$\Pr(S_T = r) = (T+1)^{-1}, \quad r = 0, 1, \ldots, T. \tag{1.10}$$

In words, (1.10) describes beliefs that in T trials, any number r of successes are equally likely. In the degenerate case in which the distribution of θ assigns probability one to some value θ_0, then de Finetti's theorem implies that S_T follows the standard binomial distribution

$$\Pr(S_T = r) = \binom{T}{r} \theta_0^r (1-\theta_0)^{T-r}, \tag{1.11}$$

and (1.9) implies

$$\lim_{T\rightarrow\infty} \overline{Y}_T = \theta_0 \tag{1.12}$$

with "probability one." This last result, as a special case of de Finetti's Theorem, is equivalent to the *strong law of large numbers*.

De Finetti's representation theorem has been generalized by seeking more stringent forms of "symmetry" than simple exchangeability, in the process rationalizing sampling models other than the binomial [see Bernardo and Smith (1994, Chapter 4)]. Although these theorems do not hold exactly for infinite sequences, they hold approximately for sufficiently large finite sequences.

The pragmatic value of de Finetti's theorem depends on whether it is easier to assess the left-hand side of (1.8), which involves only observable quantities, or instead, the integrand on the right-hand side of (1.8), which involves two distributions and the mathematical fiction θ. Most statisticians think in terms of the right-hand side. Frequentists implicitly do so with a degenerate distribution for θ that in effect treats θ as a constant, and Bayesians do so with a nondegenerate "prior" distribution for θ. What is important to note here, however, is the isomorphism de Finetti's theorem suggests between two worlds, one involving only observables and the other involving the parameter θ. De Finetti put parameters in their proper perspective: (i) They are mathematical constructs that provide a convenient index for a probability distribution, and (ii) they induce conditional independence for a sequence of observables.

Exercise 1.1 (Let's make a deal) Consider the television game show "Let's Make a Deal" in which host Monty Hall asks contestants to choose the prize behind one of three curtains. Behind one curtain lies the grand prize; the other two curtains conceal only relatively small gifts. Assume Monty knows what is behind every curtain. Once the contestant has made a choice, Monty Hall reveals what is behind one of the two curtains that were not chosen. Having been shown one of the lesser prizes, the contestant is offered a chance to switch curtains. Should the contestant switch?

Solution

Let C denote which curtain hides the grand prize. Let $\hat{C}$ denote the curtain the contestant chooses first, and let M denote the curtain Monty shows the contestant. Assume $\Pr(C = i) = 1/3,\ \ i = 1, 2, 3$, $\Pr(\hat{C} = k|C) = 1/3,\ \ k = 1, 2, 3$, and that C and $\hat{C}$ are independent. Without loss of generality, suppose $C = 1$ and $M = 2$. Then use Bayes' theorem for events to compute the numerator and denominator of the following ratio:

$$\frac{\Pr(C=3|M=2,\hat{C}=1)}{\Pr(C=1|M=2,\hat{C}=1)} = \frac{\frac{\Pr(M=2,\hat{C}=1|C=3)\Pr(C=3)}{\Pr(M=2,\hat{C}=1)}}{\frac{\Pr(M=2,\hat{C}=1|C=1)\Pr(C=1)}{\Pr(M=2,\hat{C}=1)}} \tag{1.13}$$

$$= \frac{\Pr(M=2,\hat{C}=1|C=3)}{\Pr(M=2,\hat{C}=1|C=1)}$$

$$= \frac{\Pr(M=2|\hat{C}=1,C=3)\Pr(\hat{C}=1|C=3)}{\Pr(M=2|\hat{C}=1,C=1)\Pr(\hat{C}=1|C=1)}$$

$$= \frac{\Pr(M=2|\hat{C}=1,C=3)}{\Pr(M=2|\hat{C}=1,C=1)}.$$

The numerator of the last line of (1.13) is one because Monty has no choice but to choose $M = 2$ when $\hat{C} = 1$ and $C = 3$. The denominator of (1.13), however, is ambiguous because when $\hat{C} = 1$ and $C = 1$, Monty can choose either $M = 2$ or $M = 3$. The problem formulation does not contain information on Monty's choice procedure in this case. But since this probability must be less than or equal to one, ratio (1.13) can never be less than one. Unless Pr$(M = 2|\hat{C} = 1, C = 1) = 1$, the contestant is better off switching curtains. If Pr$(M = 2|\hat{C} = 1, C = 1) =$ Pr$(M = 3|\hat{C} = 1, C = 1) = 1/2$, then the contestant doubles the probability of winning the grand prize by switching.

Exercise 1.2 (Making Dutch book) Consider a horse race involving N horses. Suppose a bettor's beliefs are such that he believes the probability of horse n winning is p_n, where $p_1 + p_2 + \cdots + p_N < 1$. Show how to make Dutch book with such an individual.

Solution

Consider a bet with this person of p_n dollars that pays one dollar if horse n wins, and place such a bet on each of the N horses. Then you are guaranteed winning one dollar (since one of the horses has to win) and earning a profit of $1 - (p_1 + p_2 + \cdots + p_N) > 0$.

Exercise 1.3 (Independence and exchangeability) Suppose $Y = [Y_1 \; Y_2 \; \cdots \; Y_T]' \sim N(0_T, \Sigma)$, where $\Sigma = (1 - \alpha)I_T + \alpha\iota_T\iota_T'$ is positive definite for some scalar α and ι is a $T \times 1$ vector with each element equal to unity. Let $\pi(t)$ $(t = 1, 2, \ldots, T)$ be a permutation of $\{1, 2, \ldots, T\}$ and suppose $[Y_{\pi(1)}, Y_{\pi(2)}, \ldots, Y_{\pi(T)}] = AY$, where A is a $T \times T$ selection matrix such that, for $t = 1, 2, \ldots, T$, row t in A consists of all zeros except column $\pi(t)$, which is unity. Show that these beliefs are exchangeable.

Solution

Note that $AA' = I_T$ and $A\iota_T = \iota_T$. Then, $AY \sim N(0_T, \Omega)$, where

$$\begin{aligned}\Omega &= A\Sigma A' \\ &= A[(1-\alpha)I_t + \alpha\iota_T\iota_T']A' \\ &= (1-\alpha)AA' + \alpha A\iota_T\iota_T'A' \\ &= (1-\alpha)I_T + \alpha\iota_T\iota_T' \\ &= \Sigma.\end{aligned}$$

Hence, beliefs regarding $Y_t(t = 1, 2, \ldots, T)$ are exchangeable. Despite this exchangeability, it is interesting to note that if $\alpha \neq 0$, Y_t $(t = 1, 2, \ldots, T)$ are *not independent*.

Exercise 1.4 (Predicted probability of success of a Bernoulli random variable) Suppose a researcher makes a coherent probability assignment to an infinite sequence $Y_t(t = 1, 2, 3, \ldots)$ of exchangeable Bernoulli random variables. Given an observed sequence of T trials with r successes, find the probability that the next outcome, Y_{T+1}, is y_{T+1}.

Solution

Applying the definition of conditional probability and then Theorem 1.2 to both the numerator and denominator yields

$$
\begin{aligned}
\Pr(Y_{T+1}=y_{T+1}|T\overline{Y}_T=r) &= \frac{\Pr(T\overline{Y}_T=r, Y_{T+1}=y_{T+1})}{\Pr(T\overline{Y}_T=r)} \qquad (1.14)\\
&= \frac{\int_0^1 \theta^{(r+y_{T+1})}(1-\theta)^{(T+1-r-y_{T+1})}p(\theta)d\theta}{\int_0^1 \theta^r(1-\theta)^{(T-r)}p(\theta)d\theta}\\
&= \frac{\int_0^1 \theta^{(y_{T+1})}(1-\theta)^{(1-Y_{T+1})}p(\theta)L(\theta)d\theta}{\int_0^1 L(\theta)p(\theta)d\theta}\\
&= \int_0^1 \theta^{(y_{T+1})}(1-\theta)^{(1-y_{T+1})}p(\theta|y)d\theta,
\end{aligned}
$$

where

$$p(\theta|y) = \frac{p(\theta)L(\theta)}{p(y)}. \qquad (1.15)$$

Therefore $\Pr(Y_{T+1}=y_{T+1}|T\overline{Y}_T=r)$ is simply

$$E(\theta|y) \text{ if } y_{T+1}=1,$$

or

$$1-E(\theta|y) \text{ if } y_{T+1}=0.$$

The simplicity of this exercise hides its importance because it demonstrates most of the essential operations that characterize the Bayesian approach to statistics. First, the existence of the density $p(\theta)$ is a result of Theorem 1.2, *not* an assumption. Second, the updating of prior beliefs captured in (1.15) amounts to nothing more than Bayes' theorem. Third, although Y_t $(t=1,2,\ldots,T)$ are independent conditional on θ, unconditional on θ they are dependent. Finally, the parameter θ is merely a mathematical entity indexing the integration in (1.14). Its "real-world existence" is a question only of metaphysical importance.

Exercise 1.5 (Independence and conditional independence) Consider three events A_i $(i=1,2,3)$, where $\Pr(A_i)=p_i,\ i=1,2,3$. Show that the following statements are totally unrelated: (a) A_1 and A_2 are independent and (b) A_1 and A_2 are conditionally independent given A_3.

Solution

There are $2^3=8$ possible three-element strings that can occur when considering A_i $(i=1,2,3)$ and their complements A_i^c $(i=1,2,3)$. This leaves assessment of $7=8-1$ probabilities since the eighth is determined by the adding-up condition. These can be assessed in terms of the following probabilities: $\Pr(A_1\cap A_2)=q_{12}$, $\Pr(A_1\cap A_3)=q_{13}$,

$\Pr(A_2 \cap A_3) = q_{23}$, and $\Pr(A_1 \cap A_2 \cap A_3) = s$. Independence of A_1 and A_2 places a restriction on $\Pr(A_1 \cap A_2)$, namely $q_{12} = p_1 p_2$. Conditional independence places a restriction on the remaining probabilities $q_{13},\ q_{23}, p_3$, and s. To see this note $\Pr(A_1 \cap A_2 | A_3) = s/p_3$ by simply expressing the conditional as the joint divided by the marginal, and conditional independence implies $\Pr(A_1 \cap A_2 | A_3) = \Pr(A_1 | A_3)\Pr(A_2 | A_3) = (q_{13}/p_3)(q_{23}/p_3)$. Putting these equalities together implies $s = q_{13} q_{23} / p_3$. Note that the restrictions implied by independence and conditional independence share no common probabilities.

2

Bayesian inference

In this chapter we extend Chapter 1 to cover the case of random variables. By *Bayesian inference* we mean the updating of prior beliefs into posterior beliefs conditional on observed data. This chapter covers a variety of standard sampling situations in which prior beliefs are sufficiently regular that the updating can proceed in a fairly mechanical fashion. Details of point estimation, interval estimation, hypothesis testing, and prediction are covered in subsequent chapters. We remind the reader that the definitions of many common distributions are provided in the Appendix to this book. Further details on the underlying probability theory are available in Chapters 1 and 2 of Poirier (1995).

One of the appealing things about Bayesian analysis is that it requires only a few general principles that are applied over and over again in different settings. Bayesians begin by writing down a joint distribution of all quantities under consideration (except known constants). Quantities to become known under sampling are denoted by the T-dimensional vector y, and remaining unknown quantities by the K-dimensional vector $\theta \in \Theta \subseteq \mathcal{R}^K$. Unless noted otherwise, we treat θ as a continuous random variable. Working in terms of densities, consider

$$p(y, \theta) = p(\theta)p(y|\theta) = p(y)p(\theta|y), \tag{2.1}$$

where $p(\theta)$ is the *prior density* and $p(\theta|y)$ is the *posterior density*. Viewing $p(y|\theta)$ as a function of θ for known y, any function proportional to it is referred to as a *likelihood function*. We will denote the likelihood function as $L(\theta)$. Unless noted otherwise, we will work with $L(\theta) = p(y|\theta)$ and thus include the integrating constant for $y|\theta$ in our description of the likelihood. We also note that

$$p(y) = \int_\Theta p(\theta)L(\theta)d\theta \tag{2.2}$$

is the *marginal density of the observed data* (also known as the *marginal likelihood*).