THE BAYESIAN VIEW

N. SRINIVASAN

¨ In this respect it is analogous to the unwritten medical code that requires a doctor to make his diagnosis and treatment of a patient dependent wholly on (i) the case history of and the outcome of some diagnostic tests carried out on that particular patient, and (ii) all the background information that the doctor (and his consultants) may have on the particular problem on hand. It is this same unwritten code that disallows a doctor to include a symmetric die or a Table of random digits as part of his diagnostic gadgets. It also forbids him to allow his judgement about a particular patient to be colored by any speculations on the types and number of patients that he may have later in the week.¨

-Basu [66]

1. Background

We fondly and piously hope that your doctor does not start his diagnosis by tossing a symmetric die or by tossing a coin! This document provides a brief overview of Probability Theory and Statistics and explicates the Bayesian view. This subject is vast and a multitude of authors have written ample number of books and articles on this subject. You may want to consult [7,8,9,10,11,12,13,36,37,40,41,42,43,44,51]. This author is neither a Bayesian nor a Frequentist, but a pragmatist.

The concept of probability originated in gambling or games of chance. Among others who have contributed vastly, the names of the following are intimately interwoven with the lore: Pascal, Fermat, Bernoulli, Laplace, Poisson, Gauss, Markov, Von Mises, Keynes, Kolomogorov, Borel, Bayes, de Finetti, Savage, Levy, Frecht, Ramsey, Jeffreys, Renyi, Fisher, Neyman, Pearson, Ito, Mahalanobis, C.R. Rao, Albert N. Shiryaev, Bradley Efron, Srinivasa Varadhan, Bose, Basu, Lindley, Nozer Singpurwalla. Probability and Statistical theory have found applications in diverse areas such as the physical and biological sciences, engineering, social sciences, psychology, artificial intelligence, forensic science, legal studies, disputed authorship, and even court room testimony.

What is probability? The meaning of probability has been elusive and even today, there is considerable controversy, even though the calculus of probability is well developed. The following quotes illustrate to some extent the differences of opinion on this subject.

This work was partly funded by the U.S. Government. The author wishes to thank Prof. Nozer Singpurwalla, Prof. John Kulesza, Prof. Hal Stern, Prof. Allan Gut, Prof. Ed Wegman, Mr. David Overton (DoJ), and Mr. David Nemecek, Esq. (DoJ).

> ¨.....It is my tentative view that the concept of personal probability ..... except for slight modifications, the only probability concept essential to science and other activities that call upon probability.¨
>
> ¨Fisher's school, with its emphasis on fiduciary probability – a bold attempt to make the Bayesian omelet without breaking the Bayesian eggs– may be regarded as an exception to the rule that frequentists leave great latitude for subjective choice in statistical analysis.¨

Underlying the concept of probability is the notion of randomness or uncertainty. Randomness has been difficult to define [33]. To date there has been no satisfactory definition of randomness. Randomness has been variously defined as nonrepeatability, unpredictability, patternlessness, and subjectively by de Finetti [8] as not known to the observer due to ignorance, even though the event itself may be well determined, such as a historical event that happened some time ago about which one may not have enough knowledge, the numbers generated by random number generators whose algorithms are at least as long as the number sequence. However, as can easily be seen, if a sequence can be computed, it is by definition nonrandom. Operationally, the last definition is convenient and the whole field of simulation is based on it. The various interpretations and the circular nature of some of these definitions show that there is no agreement on a single definition of randomness, and even more, each may have its own definition of applicability. If the core idea is taken to be ¨unknown¨ or ¨unpredictable¨ in defining randomness, one is confronted with defining ¨unknown¨ or ¨unpredictable¨. We have exchanged one undefinable concept with another. The only way a successful theory can be built at this point is to take ¨randomness¨ as undefined, but ¨self-evident¨ concept, until a good all encompassing definition comes along that is acceptable to the various factions. It is important to note that such a lack of definition has not impeded the growth of the probability theory. This has been accomplished by anchoring the theory of probability theory on a set of axioms by Kolomogorov [51], who is considered the founder of the modern probability theory.

Laplace once observed that probability is nothing but common sense reduced to calculations. There are also those who believe that probability is just another branch of mathematics, measure theory in particular. However, a closer examination may suggest that probability is neither mathematics nor science, but uses mathematics and finds applications in science. The concepts of probability transcend that of mathematics and science and the question of ¨What is probability?¨ is a continuing debate in the philosophy of science and epistemology.

Looking at the landscape of applications, it seems probability and statistics are at best certain tools that are useful in getting a grip on the underlying reality which does not reveal itself in its entirety because of its complexity, or as a means of quantifying uncertainty. The field of quantum mechanics is but one example. What baffles many, and this includes those who are new to this field and those who have been working in this field but have not lost their sense of wonder, is the ability of the theory to supply almost precise answers under suitable conditions, that are both empirically and experimentally verifiable, which cannot be even contemplated in a

deterministic context. Another example in this context is the ability to quantify the error rate in a communication channel subjected to random disturbances or noise. There are precise mathematical expressions that give a quantitative measure of the error rate and there is nothing random about these equations. How is it possible to derive such a deterministic result in a nondeterministic environment? An answer can only be supplied through an analogy. When human beings try to communicate an idea or thought, certain words, expressions and sentences are used. To a listener, and in many instances, even to the speaker, the words yet to be uttered are unpredictable or unknown, even though the underlying idea behind this process may indeed be very concrete. Yet, at the end of this exchange, a definite or concrete idea has been communicated through random means. Indeed, it has been argued that if there were no randomness or uncertainties, there will not be any need to communicate at all– either oral or written.

2. Concepts

¨In England, there were three mathematicians:
Hardy, Littlewood, Hardy and Littlewood.¨
–Unknown.

Tossing a fair coin, throwing a fair die, drawing a ball at random from an urn are all protypical random experiments–it is impossible to predict what the outcome will be. However, it should be noted that Persi Diaconis has proved that, through controlled experiments at the Stanford University, a machine can produce a series of heads only or tails only by fully accounting for the dynamics of the coin throwing. His wife also claims that Professor Diaconis can toss coins and make them come up heads or tails– because he is a magician. Since we do not have a well-designed machine that can control the dynamics of the coin throwing, nor are we magicians, we are going to assume that the coin toss is indeed a random experiment whose outcome is unpredictable. It is intuitively obvious that if the coin is fair, the long run proportion of heads or tails will be roughly $\frac{1}{2}$. The long run proportion of any one number showing up in the case of a fair die will be approximately $\frac{1}{6}$. The long run proportion of a particular colored ball being drawn from an urn will be roughly equal to the compositional proportion in the urn, assuming that we replace the ball back in the urn after each draw. The experiments or trials are required to be conducted under ¨similar¨ or ¨identical¨ conditions. It is at once obvious that no experiment can really be repeated under ¨identical¨ conditions, and so one view of probability is willing to settle for ¨similar¨.

Consideration of the above examples suggests two possible ways of defining probability:

(1) The principle of equally likely or eqiprobable events,
(2) Long run relative frequency.

The basis for defining probability based on equally likely events can be explained as follows. If a coin is tossed, there are two possible outcomes, head or tail. If the coin is assumed to be fair (note that this assumption has tacitly introduced subjectivity!), by the symmetry of the coin, it is clear that each outcome is equally likely to occur, thus attributing a probability of $\frac{1}{2}$ to each outcome. If the die is symmetrical, well-balanced, and each face of the die is marked with a different number from 1 to 6– all of which are attributes of fairness– one may conclude that each outcome is equally likely and attribute a probability of $\frac{1}{6}$ to each. A similar reasoning can be applied to the urn example in assigning probabilities.

Now, let us suppose that the coin and the die are not ¨fair¨. For example, as we toss the coin, heads show up more often than tails, or in the case of the die, the number 2 shows up more often than the other numbers. Obviously, assigning probabilities based on eqiprobable or symmetry arguments fail. Assigning probability as the limit of a relative frequency affords a way around this difficulty. According to this definition, the probability of an event is computed as the limit of a ratio of the number of outcomes favorable to the event to the total number of trials, as the number of trials gets arbitrarily large. In fact, the Bernoulli's Law of Large Numbers can be used to justify this definition. This definition has the attributes of

a mathematical definition. However, it cannot be applied in practice, because one can never really complete an indefinitely large number of experiments(in principle, the definition requires an infinite repetition of the experiment before a probability can be determined.) Even though this definition may have an intuitive appeal or the patina of ¨mathematical rigor¨, it is not operationally suitable for the determination of probabilities.

If the definition cannot be based on an infinite number of trials, why not base it on a finite number of trials? For example, let us throw a die 100 times and count how many times the number 6 shows up, and compute the probability of a 6 as the ratio of the number of times 6 occurred to the number of trials, in this case 100. This definition works fine until we encounter an ¨unfair¨ die in the extreme. Let us say that in 100 throws of a die, the number 1 has not shown up even once. Does the probability of 1 occurring equal to zero? What will be our conclusion if we conducted our experiment 1000 times instead of 100 times? May be the number 1 will show up at least once in 1000 trials. So, now the question is, how large the number of trials should be to determine the probability of all outcomes? This question can reasonably be answered in one of the following ways. Either determine the number of trials subjectively based upon prior knowledge such as a close examination of the die, or set a cut off point for the number of trials that are ¨sufficiently¨ large such that if an outcome has not occurred in these number of trials, take its probability to be zero or some small number as appropriate, or continue the trials at least until all ¨possible¨ outcomes have occurred. This approach introduces an element of subjectivity in some form, or we are back to the definition of probability based on an infinite number of trials without explicitly saying so.

We can also contemplate another possible situation of interest. Let us take the case of finite trials of a ¨fair¨ coin. If experimenter 1 tosses the coin 100 times and observes that heads show up m times, and according to our agreed definition, assigns the probability of heads to be $\frac{m}{100}$. If the same coin is assigned to experimenter 2, who tosses the coin 150 times and observes heads n times, experimenter 2 will assign a probability of $\frac{n}{150}$ for the same event. Common sense tells us that the probabilities computed by experimenter 1 and experimenter 2 will be different. In fact they can be quite different. Indeed, they may rarely be equal. Some may even be willing to bet that the whole thing was rigged if indeed the two probabilities turned out to be equal! Under these circumstances, is probability a property of the object, in this case that of the coin (or that of an electron for that matter, to pentrate the potential barrier in the case of a tunnel diode) or a property, if it can be called a property at all, attributed by the observer to the object? Is this ¨property¨ vested in the object or the subject?

Underlying all the approaches to the definition of probability has been the notion that probability is something that is objective, endowed in each object under consideration, something akin to mass, length or temperature, and the only problem lies in its determination according to some normative rules. Probabilities determined according to this line of reasoning are called ¨objective¨ probabilities or ¨scientific¨ probabilities, or frequentist view of probability. All these terminologies

may simultaneously be applied, or selectively according to the application at hand.

Some of the above definitions tacitly introduced subjectivity as was pointed out in the discussion. Besides, there is another notion of probability that does not lend itself to be defined through repeated trials, either finite or infinite, let alone under ¨identical¨ or ¨similar¨ conditions. Even in the case of a fair coin, it is obvious that the coin toss cannot be repeated under ¨identical¨ or ¨similar¨ conditions, even if a machine is built to do the tossing– notwithstanding the machine that has been designed at Stanford– as the machine is bound to suffer some degradation after each toss, and hence the situation cannot be considered identical from toss to toss. The momentum imparted to the coin by the machine is bound to vary even microscopically so that the conditions are likely to be different from trial to trial. Instances where experiments cannot be repeated twice are quite common. When we speak of the probability of rain in the next one hour, or the probability of snow tomorrow, or the probability of a specific car breaking down within a month, it is clear these experiments cannot be repeated. There is only one ¨next hour¨, only one tomorrow, there is one specific car whose reliability is required in a specified month. Obviously, all the previous definitions will fail to have meaning for these one of a kind events, and yet when the concept of probability is used under these contexts, meaningful information is definitely exchanged. Probability for these events can only be assigned subjectively. As we have seen already, some of the definitions that call for repeated trials also introduced subjectivity tacitly or surreptitiously. These arguments lead to another school of thought that declared there is no such thing as objective probability and all probabilities are subjective inherent in the observer. The subjectivist school has built the theory on its own set of axioms [7,8,9]. In this approach, probability of an event is determined by comparing it to the betting behavior of rational individuals or how a rational individual will be prone to bet. In this sense, meaning can be attributed to events such as the probability of rain in the next one hour. To illustrate, let us take an event E. Consider a lottery (or a betting situation) in which the individual is paid $1 if the event occurs and $ 0 if the event does not occur. Let P(E) denote the price the individual is willing to pay for the ticket to this lottery– that is to participate in this lottery. According to de Finetti, if the individual is coherent (consistent or rational), the price P(E) he or she will pay for this ticket or for the privilege to play is a finitely additive probability measure. The word coherence in this context means that a ¨dutch book¨ cannot be made against the individual. A ¨dutch book¨ is an unfair lottery in which the player is made a sure loser if he continues to play the game.

Regardless of whether the notion of probability has been developed based on the objective or subjective view, the probability measure so developed essentially obeys the same set of axioms to be discussed below. The one area where there seems to be a difference is finite additivity versus countable additivity. There is a pragmatic school that admits all views of probability, conceding that every view has its own domain of applicability. This school of thought is predominantly populated by engineers and scientists because missions have to be undertaken, bridges have to be built, communications networks either terrestrial or in space have to be deployed.Engineers and scientists tend to view probability theory as a tool to get a grip on reality or a means to quantify uncertainty, whether this view is explicitly

acknowledged or not.

We are still left with the question: What is probability and what is its meaning? It is still a continuing debate in the philosophy of science. However, no matter how the probabilities are derived– objectively or subjectively– they all obey a set of axioms. Once probability numbers are assigned to events (these events have to obey certain conditions), if these numbers are manipulated according to these axioms, no contardictions result. These axioms and additional results are discussed in the subsequent sections.

Another concept known as Interval Valued Probability, or Upper and Lower Probabilities have been discussed in the literature [58, 59, 60] to explain what is known as flicker noise, or $\frac{1}{f}$ noise. This concept has its own set of axioms different from the ones discussed here and will not be explored here further. Upper and Lower probabilities are not probabilities in the sense of Kolomogorov even though the word probability is appended to them. Particular cases of this concept have been investigated under the name of Belief Functions [61]. They do not obey the axioms of probability theory.

3. Axioms and A Few Related Topics

> Luther was asked: ¨In a situation where it appears that whatever we chose will to some extent be sinful, what are we to do?¨ Luther's response was: ¨Sin bravely.¨
>
> - As quoted by Dr. James Mayfield.

What has come to be called modern mathematics is founded on set theory. Because probability theory uses mathematics, set-theoretic concepts are basic to its exposition. For our purposes, not much is needed, and what is needed is introduced rather implicitly.

All possible outcomes of an experiment or trial constitute a sample space S. S is exhaustive in the sense that it consists of all possible outcomes. It immediately follows that upon the performance of the experiment, one of the outcomes of the sample space S must always occur. A subset of the sample space is called an event. In the familiar coin tossing experiment of tossing a single coin, the sample space S = {H,T} and each outcome {H} or {T} is an event. The sample space for a single throw of a die is

{1, 2, 3 , 4, 5, 6}.

Some of the events in this case are:

{1}, {2}, {3}, {4}, {5}, {6}, {1, 2, 3}, {1, 3, 5}, {2, 4, 6}, {1, 2, 3, 4, 5, 6}, etc.

A pair of events are called mutually exclusive if they cannot occur simultaneously. The outcomes {H} and {T} are mutually exclusive. So are {1}, {2}, {3}, {4}, {5}, {6}in the case of a single throw of a die.

With each event E that is a subset of the sample space S, we associate a real number P, called the probability, whose value lies in the interval
$0 \leq P \leq 1$. The dependence of P on the event E is usually reflected in the notation P(E), instead of just P. In the case of a coin toss, or a throw of the die, the following are easily verified:
P{H, T} = P{S} = 1

P{1, 2, 3. 4, 5, 6} = P{S} = 1

i.e. One of the outcomes of the sample space is bound to occur.

$$P\{H\} + P\{T\} = 1$$

$$P\{1\} + P\{2\} + P\{3\} + P\{4\} + P\{5\} + P\{6\} = 1$$

i.e. The sum of the probabilities of mutually exclusive and exhaustive events is equal to 1.

$$P\{2, 4, 6\} = P\{2\} + P\{4\} + P\{6\}$$

i.e. The probability of the union of mutually exclusive events is the sum of their individual probabilities.

If the die is assumed fair, a probability of $\frac{1}{6}$ will be assigned to each elementary outcome. Therefore,
$P\{2,4,6\} = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$

Or, the probability of an even number occurring in any given throw is $\frac{1}{2}$.

The above examples suggest that probability conforms to the following rules or axioms:

If S is the sample space, for any set A that is a subset of S, we denote by P(A) the probability that the event A occurs.

$$0 \leq P(A) \leq 1 \tag{3.1}$$

$$P(S) = 1 \tag{3.2}$$

$$P(A_1 \cup A_2 \cup A_3 \cup \cdots \cup A_n) = P(A_1) + P(A_2) + \cdots + P(A_n) \quad if A_i \bigcap A_j = 0 \; for \; i \neq j \tag{3.3}$$

Probability theory is founded on the above three axioms.

Axiom (3.3) is called the finite additivity of the probability measure P.

Verification that the above axioms hold true in the case of finite trials or a finite number of outcomes such as the ones discussed in the above examples is straight forward.

It is also clear that the axioms hold true (or can be ¨proved¨) under the relative frequency (or frequentist) interpretation of probability, either finite or tending towards infinite. In the case of subjective probability (or probabilities assigned subjectively), it can be proven that the probability numbers so assigned have to conform to the above axioms to be ¨coherent¨. Hence, these axioms are also called coherence axioms. References [7, 8, 57] discuss probability from a subjective point of view. It is important to note that whether objective or subjective, probability numbers essentially conform to the same set of axioms.

When the sample space S is uncountably infinite, the above axioms need to be modified.

A sequence or set is countably infinite if all the elements of the sequence or set can be put into one-to-one correspondence with the set of natural numbers. All

infinities are not created equal. The set of all natural numbers or the set of all integers are examples of countably infinite sets. The set of all numbers comprising the real line are uncountable, so are all numbers in the interval (0,1). The set of all real numbers is said to have the power of the continuum.

Let us consider an experiment whose sample space consists of an infinite number of outcomes or points, as opposed to a finite number. Consider a coin being tossed n times. The sample space consists of 2^n points.

$$S = \{w : w = (a_1, a_2, ...a_n), a_i = H \; or \; T\} \tag{3.4}$$

If we choose to denote H by 1 and T by 0,

$$S = \{w : w = (a_1, a_2, ...a_n), a_i = 1 \; or \; 0\} \tag{3.5}$$

That is, each w consists of a sequence of ones and zeroes.

Suppose if we redefine w as $w = 0.a_1a_2a_3.....a_n$, as the number of trials tends to infinity, w will represent a binary expansion of all real numbers in the interval (0,1). We have already noted that the points of this interval are uncountably infinite. Hence, if we pick a point at random in this interval, we have to assign a probability $p(w) = 0$ for this point. However, if we ask a question such as "What is the probability of an outcome lying in the interval $(0, \frac{1}{2})$, one has to give an intuitively obvious answer that this probability is 1/2.

$if \; p(w) = 0, \sum p(w) = 0$ for all w in the interval $(0, \frac{1}{2})$.

The simple experiment of tossing a coin is no longer simple. This example suggests that the remedy may lie in assigning probabilities not to every elementary outcome of the sample space, but only to ¨events¨ that are the subsets of the sample space and which fulfill certain properties to be discussed below. That is, probability is a set function, and the sets to which probabilities may be assigned have to fulfill certain conditions. We get into this below.

A set of subsets F of S is called an algebra (or a field) if:

$$S \; \epsilon \; F \tag{3.6}$$

$$A, B \; \epsilon \; F \;\; implies \;\; A \cup B \;\; \in \;\; F \tag{3.7}$$

$$A \; \epsilon \; F \;\; implies \;\; A^c \;\; \in \;\; F \tag{3.8}$$

Where A^c is the complement of A, defined as the set of all points in S but not in A.

An algebra of sets is considered a σ algebra (sigma algebra), if in addition to the above, they fulfill an additional property:

If $A_n \, \epsilon \, F \, , (n = 1, 2, 3.....)$ implies $\bigcup_{n=1}^{\infty} A_n \, \in \, F$

A σ algebra is also called a Borel Field.

A set function $\mu = \mu(A)$, $A \, \in \, F$, taking values in $[0, \infty)$, is called a finitely additive measure defined on F if,

$\mu(A \cup B) = \mu(A) + \mu(B)$ for every pair of disjoint sets in $A \, and \, B$.

An additive measure μ defined on an algebra F of subsets of S is countably additive (or σ additive), if for all pairwise disjoint subsets $A_1, A_2, A_3 \cdots$ of F, if

$$\mu(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mu(A_n)$$

If a countably additive measure on the algebra F satisfies P(S) = 1, it is called a probability measure, or probability, defined on the sets belonging to the algebra F.

There is always the contentious debate between those who subscribe to the frequency interpretation of probability and the subjective interpretation of probability on the issue of finite additivity versus the countable additivity of the probability measure [7, 8, 57]. Some subjectivists hold that finite additivity is enough, while yet others are of the view that since countable additivity implies finite additivity, in the interests of generality, the theory should be based upon countable additivity. It should be noted that the Kolomogorov's axioms are based on countable additivity.

To summarize:

Definition:

An ordered triple (S, F, P), that fulfills the following three conditions is called a probability model or probability space:

(3.9) S is a set of points w

(3.10) F is a σ *algebra*

(3.11) P is a probability measure (or just probability for short) on F

S is the sample space or the space of elementary events w, the sets A in F are the events for which probability is defined and $P(A)$ is the probability of the event A.

Probability of any event A obeys the three following axioms:

$$0 \leq \ P(A) \ \leq 1 \tag{3.12}$$

$$P(S) = 1 \tag{3.13}$$

$$P(\bigcup_{n=1}^{\infty}(A_n)) = \sum_{n=1}^{\infty} P(A_n), \ if \ A_i \cap A_j = 0 \text{ for } i \neq j \tag{3.14}$$

The concept of conditioning or conditional probability plays a significant role in the development of the theory and its applications.

Given a sample space S, consider an event B for which the probability is defined and $P(B) > 0$. If a new sample space is defined consisting of only those points belonging to B, we can ask, ¨What is the probability of any event A with respect to the new sample space B?¨ Or the probability of any event A with the sample space restricted to B. This is also phrased as the probability of the event A, given the event B, or the probability of the event A, given that the event B has occurred. This called the conditional probability of the event A with respect to the event B and is denoted by $P(A|B)$.

If the frequency interpretation of probability is used, it can easily be seen that:

$$P(A|B) \ = \ \frac{P(A \cap B)}{P(B)}, \ \ P(B) > 0 \tag{3.15}$$

Let the total number of trials be N, N_{ab} the number of occurrences of the joint event A and B, and N_b the number of occurrences of the event B.

By the definition of $P(A|B)$, $P(A|B) = \frac{N_{ab}}{N_b} = \frac{\frac{N_{ab}}{N}}{\frac{N_b}{N}} = \frac{P(A \cap B)}{P(B)}$. To simplify the notation, henceforth we will use $P(AB)$ instead of $P(A \cap B)$.

It easily follows that :

$$0 \leq \ P(A|B) \ \leq 1 \tag{3.16}$$

$$P(B|B) = 1 \tag{3.17}$$

$$P((A \cup C|\, B) = P(A|B) \ + P(C|B), \ if \ A \cap C = 0 \tag{3.18}$$

Hence, it follows that $P(A|B)$ is a probability that may be considered as an induced probability measure with the sample space restricted to the event B.

With the foregoing discussion in mind, at this point, it can be stated that all probabilities are indeed conditional. For, when we speak of an event A, we are indeed talking about an event A conditioned on the event S, which is our sample space, and whose probability = $P(A|S)$. That is the probability of the event A with respect to the sample space S. Also, note that $P(S|S) = 1$. The implication is that when we write $P(A)$, in reality we mean $P(A|S)$, but with the tacit understanding,

we choose to drop S.

By the definition of $P(A|B)$,

$$P(AB) = P(A|B)P(B) \tag{3.19}$$

In the development of a theory of subjective probability, the conditional probability $P(A|B)$ is itself taken to be one of the axioms. $P(A|B)$ can be defined as the price one is willing to pay to win \$1 if the event A occurs with the added condition that the transaction is called off if the event B does not occur.

An important concept in the theory of probability and Statistics is the notion of independence. The idea of independence can be given meaningthrough the interpretation of the conditional probability. Intuitively, it makes sense to say that two events are are independent if the occurrence of one event does not influence the occurrence of the other event. That is,

$$P(A|B) = P(A), \textit{ or} \tag{3.20}$$

$$P(AB) = P(A)P(B) \tag{3.21}$$

Either one is taken to be the definition of independence.

To give an example of the role played by independence, the example of two computers operating in parallel in a random load sharing environment can be cited. Let us say that two computer systems, each with its own copy of the operating system, are assigned to process the incoming transactions. The transactions are assigned to the system that is available to accept the work load–that is, to the system that is less loaded. This scheduling policy, coupled with the nature, composition, and arrival rate of the transactions, makes the processing requirements on each computer random. Under these circumstances, if someone informs us that one of the systems has crashed, what can we say about the status of the second system? Failures are precipitated by the underlying faults. Before the failures can occur, the processing requirements imposed by the workload must force the software to traverse a path through the software that contains one or more faults. The internal path traversed must depend on the processor states at the time of arrival of the transaction for processing, and the nature and composition of the transaction. Because of the randomness induced by the nature of the scheduling, the internal path executed by the software will be different in each processor. The chances of encountering the same path in both the computers at the same time are indeed quite slim. Even if we strain ourselves to impose an additional constraint that in a given instance the same transaction (or a similar transaction) will arrive at each computer when its internal states are identical, the software may be forced to traverse the same path, but at different times, because of the differences in the utilization levels. One can conclude that, even when there is a failure, it will occur at different times. It should be pointed out that the number of internal states of a system are too numerous after being put into operation, the chances of finding two systems with the same internal states are slim indeed. One cannot but conclude that even if we know that one computer has failed, this knowledge will not supply any additional information

that will enable us to to predict (or bet) on the status of the other machine. We are not precluding the possibility of finding the other machine in a failed state–it is entirely possible that the second machine has failed too. We only claim that the knowledge about the state of one does not reveal much about the state of the other. That is, the two systems can be treated as stochastically independent.

[A note: In the foregoing analysis, we have not mentioned anything about the hardware reliability, but focussed on software aspects only. The reason is that the hardware technologies have advanced significantly. By operating under designed operating conditions, and also through the introduction of redundancies, it is possible to achieve a very high level of hardware availability. One wishes that software reliability also will achieve this level of availability/reliability some day. That some day is not today. It is also important to note that ¨Reliability¨ and ¨Availability¨ are also quantified as probabilities. Reliability is a function of operational duration, whereas Availability is the probability that the system is operating successfully at a future time t.]

A lengthy justification for the use of independence in the above instance has been provided, because quite often the concept of independence is freely and sometimes instinctively used even when its use cannot be justified. For example, if the integrated circuits within the computer employ liquid cooling, the reliability of the integrated circuits cannot be regarded as independent. Their reliability is dependent upon whether the coolant pump and the associated ¨plumbing¨ is functional or not. However, given the event that the coolant pump is functioning and there is enough coolant in the system, the reliability of the ICs may be regarded as independent. The dependency can be reduced to an insignificant level by providing adequate back-up or redundancy to the cooling system.

4. Convergence Concepts

¨ Nowhere have I seen such a band of determined determinists.¨
-Dr. G. J. MacDonald

The convergence of a sequence of random variables plays an important role in probability and statistics. The well known Central Limit Theorem (CLT) and the Law of Large Numbers are prime examples. Certain definitions are given below to the extent needed for the material discussed in this text.

Definition: Convergence in probability.

The sequence of random variables $X_1, X_2, X_3, \cdots$ converges in probability to the random variable X, if for every $\epsilon > 0$,

$$P(|X_n - X| > \epsilon) \to 0, \; as \; n \to \infty \tag{4.1}$$

This is also known as convergence in measure.

Definition: Convergence with probability 1.

The sequence of random variables $X_1, X_2, X_3, \cdots$ converges with probability one to the random variable X, if,

$$P(X_n \to X) = 1 \quad as \; n \to \infty \tag{4.2}$$

This is also known as almost sure (a.s) convergence, or almost everywhere (a.e) convergence.

Convergence with probability one implies convergence in probability.

5. Law of Large Numbers

¨In questions of science the authority of a thousand is not worth the humble reasoning of an individual.¨

-Galileo (quoted in The World of Mathematics, by J.R. Newman.)

1. Bernoulli's Law of Large Numbers:

If S_n represents the number of occurrences of an event in n trials,

$$P(|\frac{S_n}{n} - p|) \ > \ \epsilon) \ \rightarrow \ 0, \ as \ n \ \rightarrow \infty \tag{5.1}$$

The above is also known as the Weak Law of Large Numbers.

2. Kolomogorov' Law of Large Numbers:

Let $X_1, X_2, X_3, \cdots$ be a sequence of independent, identically distributed random variables with $E[X_1] \ < \ \infty$. If $S_n = X_1 + X_2 + X_3 + \cdots + X_n$, then

$$P\left(\frac{S_n}{n}\right) \ \rightarrow \ m \ = \ 1, \ \text{as} \ n \ \rightarrow \infty \tag{5.2}$$

Where $m \ = \ E[X_1]$. The convergence is with probability 1.

For independent, identically distributed random variables, the condition $E[X_1] \ < \ \infty$ is also necessary and sufficient for convergence with probability one of the ratio $\frac{S_n}{n}$ to a finite limit.

Note: $E[X]$ is the expected value of the random variable X. It is also called the average value or the arithmetic mean of the random variable X.

Kolomogorov's form of the Law of Large Numbers implies the Bernoulli's law of Large Numbers.

6. Bayes' Theorem

¨I was born not knowing and have only had a little time to change that here and there.¨

–Richard Feynman.

If events $A_1, A_2, A_3, \cdots A_n$ are a partion of the sample space S, i.e. the events are mutually exclusive and exhaustive, and B is any event that is a subset of S, then:

$$P(A_i|B) = \frac{P(B|Ai)P(A_i}{\sum\limits_{all\ i} P(B|A_i)P(A_i)} \tag{6.1}$$

The denominator is the probability of the event B, i.e.

$$P(B) = \sum_{all\ i} P(B|A_i)P(A_i) \tag{6.2}$$

Bayes' Theorem plays an important result that plays the central role in Bayesian Inference, as the name implies.

Bayes' theorem can be given an interpretation as follows. If B represents the results of an experiment (or evidence), $P(A_i)$ is the probability of the event A_i before the experiment (i.e. prior to the experiment or a priori), $P(A_i|B)$ is the probability attributed to the same event A_i in the light of the evidence presented by the evidence B (or after the data B resulting from the experiment becomes known). $P(A_i|B)$ is called the posterior probability of the event A_i.

If the summation in the above equation is replaced by an integral-integration is summation- and the probabilities are replaced by their respective densities, we get the corresponding statement of Bayes' theorem in its continuous form:

$$\pi(y|x) = \frac{f(x|y)\pi(y)}{\int\limits_{all\ y} (f(x|y)\ \pi(y)dy} \tag{6.3}$$

where,
$\pi(y)$ is the prior probability density,
$\pi(y|x)$ is the posterior probability density,
$f(x|y)$ when regarded as a function of y for a given x, is called the likelihood function of y.

The denominator is the marginal density, usually denoted by $m(x)$, and acts like a normalizing constant to make the left hand side a proper probability density.

Note that to be a probability density function, it must satisfy:

$$\int_{-\infty}^{\infty} f(x)\ dx = 1 \tag{6.4}$$

The integrand need only be piecewise continuous, the discontinuities at countably infinite points.

6.1. **Proof of Bayes' Theorem.**
Let $S = \{A_1\ , A_2\ , A_3\ , \cdots A_n\}$, where $A_1\ , A_2\ , A_3\ , \cdots A_n$ is a partition of the sample space S. By a partition is meant that the sets are mutually exclusive and exhaustive. Every point in the partition is in S. Let B be an arbitray event such that $B \subset S$.

It follows,

$$B = BS = B(A_1\ + A_2\ + A_3\ + \cdots + A_n) = BA_1 + BA_2 + BA_3 + \cdots + BA_n$$

Because the events $\{A_1\ , A_2\ , A_3\ , \cdots A_n\}$ are mutually exclusive,

$$P(B)\ = P(BA_1)\ + P(BA_2)\ + P(BA_3 + \cdots + P(BA_n)$$

Because, $P(BA_i)$ and $P(BA_j)$ are mutually exclusive for $i \neq j$, and indeed they are mutually exclusive by the definition of being a partition.

We have already shown,

$$P(BA_i)\ = P(B|A_i)P(A_i)$$

Therefore,

$$P(B)\ = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3) \cdots + P(B|A_n)P(A_n))$$

The above equation is known as the Total Probability Theorem.

Since,

$$P(BA_i)\ =\ P(A_i)|B)P(B),$$

$$P(A_i|B)\ =\ \frac{P(BA_i)}{P(B)}\ =\ \frac{P(B|A_i)\,P(A_i)}{P(B)}$$

Substituting the relavant quantities,we have the Bayes' Theorem

(6.5)
$$P(A_i|B)\ =\ \frac{P(B|A_i)P(A_i)}{P(B|A_1)P(A_1)\ +\ P(B|A_2)P(A_2)\ +\ P(B|A_3)P(A_3)\ \cdots + P(B|A_n)P(A_n)}$$

■

We will now proceed to prove Bayes' Theorem for continuous probability distributions, where the probability densities exist. It is not obvious directly from the

Bayes' Theorem for probabilities.

$P\{X \leq x\} = P\{X \leq x|A_1\}P(A_1) + \cdots + P\{X \leq x|A_n\}P(A_n)$

Note, $F(x) = P\{X \leq x\}$

Hence,

$F(x) = F(x|A_1)P(A_1) + F(x|A_2)P(A_2) + \cdots + F(x|A_n)P(A_n)$

Assuming $F(x)$ is differentiable,

$f(x) = f(x|A_1)P(A_1) + f(x|A_2)P(A_2) + \cdots + f(x|A_n)P(A_n)$

Note: $A_1\,, A_2\,, A_3\,\cdots\,, A_n$ form a partition of S

We have already established,

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} \tag{6.6}$$

Let,

$B = \{x_1 < X \leq x_2\}$ and $A_i = \{y_i < Y \leq y_i + \delta_i\}$

The left hand side of the above equation:

$P(A_i|B) = P(y_i < Y \leq y_i + \delta_i \,|\, x_1 < X \leq x_2)\, P(x_1 < X \leq x_2)$

The numerator of the right hand side of the above equation:

$P(B|A_i)P(A_i) = P(x_1 < X \leq x_2 \,|\, y_i < Y \leq y_i + \delta_i)P(y_i < Y \leq y_i + \delta_i)$

The denominator of the right handside of the above equation:

$$P(B) = \sum_1^n P(B|A_i)P(A_i) \tag{6.7}$$

$$= \sum_1^n P(x_1 < X \leq x_2 \,|\, y_i < Y \leq y_i + \delta_i)P(y_i < Y \leq y_i + \delta_i) \tag{6.8}$$

$\delta_i, \to 0\,, n \to \infty, P(B) = \int_{-\infty}^{\infty} P(x_1 < X \leq x_2 \,|\, Y) f_Y(Y)\, dY$

Combining all the equations together,

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} \tag{6.9}$$

(6.10)

$$P(y_i < Y \leq y_i + \delta_i \,|\, x_1 < X \leq x_2) = \frac{P(x_1 < X \leq x_2 \,|\, y_i < Y \leq y_i + \delta_i)\, P(y_i < Y \leq y_i + \delta_i)}{\int_{-\infty}^{\infty} P(x_1 < X \leq x_2 \,|\, Y) f_Y(Y)\, dY}$$

Let $\delta_i \to 0$ on both sides,

$$f_Y(Y \,|\, x_1 < X \leq x_2)\, dY = \frac{P(x_1 \leq X \leq x_2 |\, Y) f_Y(Y)\, dY}{\int_{-\infty}^{\infty} P(x_1 < X \leq x_2 \,|\, Y) f_Y(Y)\, dY} \tag{6.11}$$

Cancelling dY on both sides,

$$f_Y(Y \,|\, x_1 < X \leq x_2) = \frac{P(x_1 < X \leq x_2 |\, Y) f_Y(Y)}{\int_{-\infty}^{\infty} P(x_1 < X \leq x_2 \,|\, Y) f_Y(Y)\, dY} \tag{6.12}$$

Let $x_2 = x_1 + h$,

$$f_Y(Y \,|\, x_1 < X \leq x_1 + h) = \frac{P(x_1 < X \leq x_1 + h \,|\, Y) f_Y(Y)}{\int_{-\infty}^{\infty} P(x_1 < X \leq x_1 + h |\, Y) f_Y(Y)\, dY} \tag{6.13}$$

$$f_Y(Y \,|\, x_1 < X \leq x_1 + h) = \frac{[F(x_1 + h \,|\, Y) - F(x_1 \,|\, Y)] f_Y(Y)}{\int_{-\infty}^{\infty} [F(x_1 + h \,|\, Y) - F(x_1 \,|\, Y)] f_Y(Y)\, dY} \tag{6.14}$$

As x_1 is arbitrary, we can replace it with X and let $h \to 0$,

$$f_{Y|X}(Y \,|\, X) = \frac{f_{X|Y}(X|Y)\, dX\, f_Y(Y)}{\int_{-\infty}^{\infty} f_{X|Y}(X|Y)\, dX\, f_Y(Y)\, f_Y(Y)\, dY} \tag{6.15}$$

As the integration in the denominator is with respect to Y, the dX in the numerator and the denominator cancels out, leaving,

$$f_{Y|X}(Y \,|\, X) = \frac{f_{X|Y}(X|Y)\, f_Y(Y)}{\int_{-\infty}^{\infty} f_{X|Y}(X|Y)\, f_Y(Y)\, dY} \tag{6.16}$$

The above equation is the Bayes' Theorem for continuous functions.

■

An alternative proof:

$$P(Y|X) \;=\; \frac{P(X,Y)}{P(X)}, \;\; P(X) \;>\; 0 \text{ , where } X \text{ and } Y \text{ are } \textit{events.} \tag{6.17}$$

The above equation follows directly from the definition of conditional probability for *events.*

Define the events thus:

Assume X and Y are continuous and differentiable.

$X \;=\; \{x_1 \;<\; X \;<\; x_1 \;+\; \delta_x\}\,, \;\; Y \;=\; \{y_1 \;<\; Y \;<\; y1 \;+\; \delta_y\}$

$P(Y|X)\,P(X) \;=\; P(X,Y)$

$P(y_1 \;<\; Y \;<\; y1 \;+\; \delta_y|x_1 \;<\; X \;<\; x_1 \;+\; \delta_x)\,P(x_1 \;<\; X \;<\; x_1 \;+\; \delta_x) \;=$

$P(x_1 \;<\; X \;<\; x_1 \;+\; \delta_x,\, y_1 \;<\; Y \;<\; y1 \;+\; \delta_y)$

$dF_{Y|X}(Y|X)\;dF_X(X) \;=\; dF_{X,Y}(X,Y)$

$f_{Y|X}(Y|X)\,dY\;f_X(X)\,dX \;=\; f_{X,Y}(X,Y)\,dX\,dY$

Thus,

$$f_{Y|X}(Y|X) \;=\; \frac{f_{X,Y}(X,Y)}{f_X(X)} \tag{6.18}$$

Having proved the existence of conditional densities subject to the assumptions of continuity and differentiability, the follwing is immediate:

$$f_{Y|X}(Y|X) \;=\; \frac{f_{X|Y}(X|Y)f_Y(Y)}{\int\limits_{all\,Y} f_{X|Y}(X|Y)\,f_Y(Y)\,dY} \tag{6.19}$$

■

The following is an alternate proof suggested by Prof. Allan Gut of The University of Uppsala, Sweden:

$P\{X \;\leq\; x\} \;=\; P\{X \;\leq\; x|A_1\}P(A_1) \;+\cdots+\; P\{X \;\leq\; x|A_n\}P(A_n)$

Note, $F(x) = P\{X \leq x\}$

Hence,

$$F(x) = F(x|A_1)P(A_1) + F(x|A_2)P(A_2) + \cdots + F(x|A_n)P(A_n)$$

Assuming $F(x)$ is differentiable,

$$f(x) = f(x|A_1)P(A_1) + f(x|A_2)P(A_2) + \cdots + f(x|A_n)P(A_n)$$

Note: $A_1\,, A_2\,, A_3 \cdots, A_n$ form a partition of S

We have already established,

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} \tag{6.20}$$

Let,

$$B = \{x_1 < X \leq x_2\} \quad \text{and} \quad A_i = \{y_i < Y \leq y_i + \delta_i\}$$

The left hand side of the above equation:

$$P(A_i|B) = P(y_i < Y \leq y_i + \delta_i \,|\, x_1 < X \leq x_2)\, P(x_1 < X \leq x_2)$$

The numerator of the right hand side of the above equation:

$$P(B|A_i)P(A_i) = P(x_1 < X \leq x_2 \,|\, y_i < Y \leq y_i + \delta_i)P(y_i < Y \leq y_i + \delta_i)$$

The denominator of the right handside of the above equation:

$$P(B) = \sum_1^n P(B|A_i)P(A_i) \tag{6.21}$$

$$= \sum_1^n P(x_1 < X \leq x_2 \,|\, y_i < Y \leq y_i + \delta_i)P(y_i < Y \leq y_i + \delta_i) \tag{6.22}$$

$$\delta_i, \to 0\,, n \to \infty, P(B) = \int_{-\infty}^{\infty} P(x_1 < X \leq x_2 \,|\, Y) f_Y(Y)\, dY$$

Combining all the equations together,

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} \tag{6.23}$$

(6.24)

$$P(y_i < Y \leq y_i + \delta_i \,|\, x_1 < X \leq x_2) = \frac{P(x_1 < X \leq x_2 \,|\, y_i < Y \leq y_i + \delta_i)\, P(y_i < Y \leq y_i + \delta_i)}{\int_{-\infty}^{\infty} P(x_1 < X \leq x_2 \,|\, Y) f_Y(Y)\, dY}$$

Divide both sides by δ_i and let $\delta_i \rightarrow 0$.

$$f_Y(Y \,|\, x_1 < X \leq x_2) = \frac{P(x_1 \leq X \leq x_2 | Y) f_Y(Y)}{\int_{-\infty}^{\infty} P(x_1 < X \leq x_2 \,|\, Y) f_Y(Y)\, dY} \tag{6.25}$$

Let $x_2 = x_1 + h,\ h > 0$,

$$f_Y(Y \,|\, x_1 < X \leq x_1 + h) = \frac{P(x_1 < X \leq x_1 + h \,|\, Y) f_Y(Y)}{\int_{-\infty}^{\infty} P(x_1 < X \leq x_1 + h | Y) f_Y(Y)\, dY} \tag{6.26}$$

$$f_Y(Y \,|\, x_1 < X \leq x_1 + h) = \frac{[F(x_1 + h \,|\, Y) - F(x_1 \,|\, Y)] f_Y(Y)}{\int_{-\infty}^{\infty} [F(x_1 + h \,|\, Y) - F(x_1 \,|\, Y)] f_Y(Y)\, dY} \tag{6.27}$$

As x_1 is arbitrary, we can replace it with X and let $h \rightarrow 0$,

$$f_{Y|X}(Y \,|\, X) = \frac{f_{X|Y}(X|Y)\, dX\, f_Y(Y)}{\int_{-\infty}^{\infty} f_{X|Y}(X|Y)\, dX\, f_Y(Y)\, f_Y(Y)\, dY} \tag{6.28}$$

As the integration in the denominator is with respect to Y, the dX in the numerator and the denominator cancels out, leaving,

$$f_{Y|X}(Y \,|\, X) = \frac{f_{X|Y}(X|Y)\, f_Y(Y)}{\int_{-\infty}^{\infty} f_{X|Y}(X|Y)\, f_Y(Y)\, dY} \tag{6.29}$$

Once again,the above equation is the Bayes' Theorem for continuous functions.

■

7. Some Probability Models

¨The divergent series are the invention of the devil, and it is a shame to base on them any demonstration whatsoever. By using them, one may draw any conclusion he pleases and that is why these series have produced so many fallacies and so many paradoxes.¨

-Neils Henrik Abel, quoted in ¨To Infinity and Beyond¨ by Eli Maor.

If a coin is flipped n times, the probability of exactly r heads is:

$$P(\ r\ heads)\ =\ \binom{n}{r}\ p^r\ \ (1\ -\ p)^{n-r}\ \ ,0 \leq\ r\ \leq n\ ,0\ \leq p\ \leq 1 \tag{7.1}$$

where p is the probability of heads. Note: This formula takes into account that the coin may be biased. An expression of this form is a probability model of the experiment under investigation. The behavior of this model is determined by the quantity p, which is called the parameter of the model. This one is called a discrete model, because the variable of interest r is a discrete quantity. The parameter p is continuous. The above expression represents the binomial distribution.

An example of a continuous model is the exponential distribution distribution whose probability distribution is given by:

$$f(x|\lambda)\ =\ \lambda\, e^{-\lambda x} \tag{7.2}$$

Another example of a continuous model is the beta distribution, given by:

$$f(x|\alpha,\beta)\ =\ \frac{\Gamma(\alpha\ +\ \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1},\ \ 0\ \leq x \leq\ 1, \alpha\ >\ 0, \beta\ >\ 0 \tag{7.3}$$

Where $\Gamma(.)$ is the Gamma distribution, α , β are the parameters of the distribution.

8. Statistics

¨No human investigation can claim to be scientific if it doesn't pass the test of mathematical proof.¨

-Leonardo Da Vinci
(quoted in Concepts of Mathematical Modeling by Walter J. Meyer)

The words ¨probability¨ and ¨statistics¨ are used together so often that one can infer they are connected in some manner.

Suppose we consider an experiment in which we randomly draw a ball from an urn containing a total of 100 balls of which 30 are blue and 70 are red. We can ask the question, ¨ What is the probability that a ball drawn at random from this urn, after a thorough mixing is red?¨ Given the aforementioned composition, one can assign a probability of 0.7 for this event. However, let us suppose we did not know the composition of this urn, and we are allowed to repeat this experiment of the random drawing as often as we want with the condition that after each draw and noting its color, we replace the ball back in the urn and thoroughly mix the contents of the urn before the next ball is drawn. Based on these observations, we are required to estimate or infer the composition of the urn. Statistics is concerned with this reverse problem of going from the particular to the general or from the sample to the ¨population¨. Statistics is inductive by nature, and uses probability in its methods. If mathematics is viewed as deductive, the phrase ¨mathematical statistics¨ is an oxymoron. Parts of mathematics and most of our daily experience is based on induction. The phrase ¨mathematical statistics¨ can be understood to mean that mathematics is used to do statistics. Induction is a topic in its own right– and a controversial one too– in the philosophy of science and mathematics [62]. Whether we are willing to acknowledge it or not, we are all active practitioners of induction in our daily lives, and indeed our cognitive processes are inductive.

There are in general three manifestations of statistics: Descriptive, Inferential, and Decision Theoretic. Descriptive statistics refers to the summarization of data in the form of tables, charts, averages, etc. Inferential statistics is concerned with drawing conclusions about a larger population based on a limited set of data or a random sample taken from the population. Statistical decision theory is concerned with the methods of decision making in the light of available data, with due consideration placed on the consequences of the various decisions. Most of the statistical theory is concerned with the latter two.

Let us consider a statistical probability model, for example, the normal probability density function given by:

$$f(x|\mu,\sigma) \;=\; \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\,\sigma^2}} \tag{8.1}$$

Where,$-\infty \;<\; x \;<\infty\,, -\infty \;<\mu \;<\; \infty\;,\; 0 \;<\; \sigma \;<\; \infty$, and μ is the mean, and σ is the standard deviation.

The random variable x behaves according to this distribution for the phenomena under investigation. What is observable is usually the quantity x and not the parameters μ and σ. μ and σ are apriori unknown and have to be determined (or estimated) based on certain observations of x. The process of determining the parameters based on the observations is known as inference.

The so called classical statistics regards the parameters unknown, but fixed constants. The Bayesian Statistics takes the position that if some quantity is unknown, it is uncertain, and hence the parameters should also be regarded as random variables. Yes, why not? In addition to this difference, there are also other philosophical and foundational differences between the two schools of thought [7, 8, 10]. In many instances, the results obtained under the Bayesian procedures are close to and may even agree with the results of classical statistical theory. The term ¨Bayesian Inference¨as the name implies, is based on the Bayes' theorem. It is repeated here for easy reference:

$$\pi(y|x) = \frac{f(x|y)\pi(y)}{\int\limits_{all\ y} (f(x|y)\ \pi(y)dy} \tag{8.2}$$

In the above equation, if the random variable y is replaced by the quantity θ, an unknown and hence an uncertain (i.e. a random variable) parameter, the Bayes' formula becomes:

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int\limits_{all\ \theta} (f(x|\theta)\ \pi(\theta)d\theta} \tag{8.3}$$

where $\pi(\theta)$ is the prior probability density of θ. $\pi(\theta|x)$ is the posterior probability density of θ, given the data or the evidence x. $f(x|\theta)$ is the likelihood function, i.e. the probability model regarded as a function of θ, for a given x.

The denominator, after integrating, is independent of θ, and is a function of x only. It is called the marginal density of x, denoted by $m(x)$, and for a given x, it is a constant. $m(x)$ is the normalizing constant.

The words ¨prior¨ and ¨posterior¨ are interpretations. Bayes' theorem holds true for all probability densities (or probabilities) that satisfy the required conditions. Bayes' theorem is provable within the axioms of probability theory regardless of whether probability is interpreted according to the frequency theory or subjectively. In the light of the above discussion, the Bayes' theorem is used as a vehicle to modify the ¨prior¨ belief or the ¨prior¨ probability density into ¨posterior belief¨ or the ¨posterior¨ probability density, in the light of the evidence or the data x. i.e. The Bayes' theorem has been used to perform the inference about the unknown parameter θ, given the observation x.

9. Some Observations on the Bayesian View

> ¨The only way to learn mathematics is to do mathematics. That tenet is the foundation of the do-it-yourself, Socratic, or Texas method, the method in which the teacher plays a role of an omniscient but largely uncommunicative referee between the learner and the facts.¨
>
> -Paul Halmos

A Russian said to his Chinese friend, ¨When we dug during an archaeological expedition, we found some copper wires; so we concluded that our ancestors knew about telephony long before the west.¨ To which his Chinese friend responded, ¨We did some digging too; we did not find anything, and so we concluded that our ancestors were using wireless.¨

Lesser the evidence, the more can we conclude!

The Bayesian approach to inference makes it explicit the role played by the data and the prior knowledge or belief as embodied in the prior probability density. Any statement made about the unknown parameter θ based on the posterior density is a true probability statement unlike the confidence intervals of classical statistics. Classical statisticians indeed take great pains to emphasize why their confidence intervals are not probability statements. Notwithstanding these statements, in day-to-day use, confidence intervals are being used as degrees of belief or as probability statements. Only under the Bayesian interpretation, unlike the classical confidence intervals, the Bayesian Confidence Intervals (or the Confidence Sets) become a true probability statement.

Other methods of inference, such as the Maximum Likelihood Estimation (MLE), hypothesis testing, etc., use methods outside of probability theory, whereas the Bayesian approach is centered entirely within the axioms of probability theory. A discussion of various views, pro and con, can be found in [10]. Another criticism centers on the use of a prior on the unknown quantity. If a quantity is unknown in every sense of the word, one will not be tempted to do any investigations on it. The very fact that one is contemplating an inference suggests that something is known to kindle enough interest to launch an investigation. The Bayesian approach provides a vehicle to incorporate such a knowledge explicitly, in a proper manner without sweeping it under the rug or without pretending that nothing is known. The Bayesian position forces the person doing the analysis to make the assumptions explicit. If in reality, one does not know very much to incorporate into the prior, there are a class of distributions known as non-informative priors that may be used. Another advantage of the Bayesian view is that the posterior distribution in turn can be used as a prior distribution as further data are accumulated to refine the estimate of the unknown quantity successively. The debate between the Bayesians and non-Bayesians has been a continuing one, sometimes vitriolic and pungent. Examples and counter examples in support of each position have been given in the literature.

In contrast to the confidence intervals and hypothesis testing of conventional statistics which requires that consideration be given to data or outcomes that did not occur, the Bayesian view requires and indeed forces the attention to be centered on only the data that did occur and the prior knowledge that has been accumulated.

10. Predictive Distributions

¨A central lesson of science is that to understand complex issues (or even simple ones), we must try to free our minds of dogma and to guarantee the freedom to publish, to contradict, and to experiment. Arguments from authority are unacceptable.¨

-Carl Sagan,

A probability distribution function or a probability density function characterizes a random variable. It essentially portrays what values or more likely or less likely than others. In several of these probability distributions, as we have seen already, certain quantities called parameters, whether considered fixed or random, determine a specific distribution among a class of distributions. A great body of statistical theory is concerned with estimating these parameters based on measurements or data. Sometimes, these parameters are the familiar mean and variance, but in general, they need not be. Nevertheless, parameters are not directly observable nor measurable. They have to be calculated or ¨inferred¨ based on certain observable or measured quantities. Parameters such as the mean and the variance can be endowed with meaning as representing certain aggregate or global properties, even though they are neither directly observable or measurable. Sometimes, one is interested, not in the aggregate behavior, but in a specific instance. For example, one can ask, ¨What is the probability that the system will operate without any failure over a specified interval of time?¨ The parameter, ¨Mean Time to Failure (MTTF)¨ cannot be used to answer this question, even though MTTF may be used to characterize the aggregate behavior of the system. To cite another example to clarify the problem and the nature of the solution sought, let us say that data on the failure times of 100 bulbs have been gathered from a specified lot. One can certainly compute the MTTF, but a more pertinent and a natural question will be, ¨What is the probability another light bulb from this lot will survive a specified period of time?¨ In other words, if $x_1, x_2, x_3, \cdots x_n$ are the observations, the interest is centered on $p(x_{n+1}|x_1, x_2, x_3, \cdots x_n)$, the probability distribution of x_{n+1} given the n observations. This is called the predictive distribution [9, 10].

Let us derive an expression for the predictive distribution.

By definition,

$$p(y|x) \;=\; \frac{p(y,x)}{p(x)} \;=\; \frac{\int p(y,x|\theta)p(\theta)\; d\theta}{\int (p(x|\theta)\; p(\theta) d\theta} \tag{10.1}$$

the integration of the numerator and the denominator are over all θ.

If x and y are independent given θ,

$$p(y|x) \;==\; \frac{\int p(y|\theta)p(x|\theta)p(\theta)\; d\theta}{\int (p(x|\theta)\; p(\theta) d\theta} \tag{10.2}$$

For a given x (this is the data that have been collected, hence known), the denominator is a constant.

Therefore,

$$p(y|x) \;=\!=\; K \int_{all\ \theta} p(y|\theta)p(x|\theta)p(\theta)\ d\theta \tag{10.3}$$

If $x = (x_1, x_2, x_3, \cdots x_n) = x^{(n)}$, the vector representing the n observations, and $y \;=\; x_{n+1}$, the unknown quantity of interest,

$$p(x_{n+1}|x^{(n)}) \;=\!=\; K \int_{all\ \theta} p(x_{n+1}|\theta)p(x^{(n)}|\theta)\ p(\theta)\ d\theta \tag{10.4}$$

If the n observations are independent,

$$p(x_{n+1}|x^{(n)}) \;=\; K \int_{all\ \theta} p(x_{n+1}|\theta) \prod_{1}^{n} p(x_i)|\theta)\ p(\theta)\ d\theta \tag{10.5}$$

Where,

$$\prod_{1}^{n} p(x_i)|\theta) \;=\; p(x_1|\theta)\, p(x_2|\theta)\, p(x_3|\theta) \cdots p(x_n|\theta) \tag{10.6}$$

The expression for $p(x_{n+1}|x^{(n)})$ does not involve θ and is called the predictive distribution for X.

> ¨I have learnt silence from the talkative, tolerance from the intolerant, and kindness from the unkind. Yet, strange, I am ungrateful to these teachers.¨
>
> -Kahlil Gibran

References

[1] R. E. Barlow, F. Proschan, *Statistical Theory of Reliability and Life Testing*, To Begin With Publishing, Silver Spring, MD, 1981.

[2] N. R. Mann, R. E. Schaefer, N. D. Singpurwalla, *Methods for the Statistical Analysis of Reliability and Life Data*, John Wiley & Sons, 1974.

[3] K. S. Trivedi, *Probability and Statistics with Reliability, Queuing and Computer Science Applications*, Prentice Hall, 1982.

[4] H. F. Martz, R. A. Waller, *Bayesian Reliability Analysis*, John Wiley & Sons, 1982.

[5] J. Musa, A. Iannino,K. Okumoto, *Software Reliability*, McGraw Hill, 1987.

[6] C. V. Ramamoorthy, F. B. Bastani, ¨Software Reliability–Status and Perspectives¨, SE-84, 1984, pp. 354-370.

[7] L. J. Savage, *The Foundations of Statistics*, Dover, 1972.

[8] B. de Finetti, *Theory of Probability*, Vol I and II, John Wiley, 1970.

[9] J. Aitchison, I.R. Dunsmore, *Statistical Prediction Ananlysis*, Cambridge University Press, 1975.

[10] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag, 2nd Edition, 1985.

[11] S. Karlin, H. M. Taylor, *A First Course in Stochastic Processes*, Volumes I and II, Academic Press, 1975.

[12] S. M. Ross, *Stochastic Processes*, John Wiley, 1983.

[13] W. Feller, *An Introduction to Probability Theory and its Applications*, Volumes I and II, Wiley, 1957 and 1966.

[14] S. Mourad and D. Andrews, ¨On the Reliability of the IBM MVS/XA Operating System¨, IEEE Transactions on Software Engineering, Vol. SE-13, No. 10, pp. 1135-1139, October, 1987.

[15] S. Mourad and D. Andrews, ¨The Reliability of the IBM MVS/XA Operating System¨, Fifteenth Annual International Symposium On Fault Tolerant Computing, Ann Arbor, Michigan, June 19-21, 1985.

[16] R. K. Iyer, S. E. Butner, E. J. McClusky, ¨A Statistical Failure/Load Relationship, Results of a Multi-Computer Study¨, IEEE Transactions on Computers, Vol. C-31, pp. 697-706, July, 1982.

[17] R. K. Iyer, D. J. Rosetti, ¨CPU Failures and System Activity: Measurement and Modeling¨, IEEE Transactions on Computers, Vol. C-32, July, 1983.

[18] S. G. Tucker, ¨The IBM 3090 System: An Overview¨, IBM Systems Journal, Vol. 25, No. 1, 1986.

[19] Y. Singh,A. M. King, J.W. Anderson, ¨IBM 3090 Performance: A Balanced System Approach¨, IBM Systems Journal, Vol. 25, No. 1, 1986.

[20] ¨The IBM 3090 Processor Familiy: A Balance of Technology and Design¨, IBM Technology Marketing Center, Essex Junction, Vermont 05452.

[21] ¨IBM 3880 Storage Control Models 3, 21, and 23¨, G520-6449-00, IBM Corporation, Rye Brook, Ny 10573.

[22] ¨IBM 3990 Storage Control Family¨, G520-6441-00, IBM Corporation, Rye Brook, Ny 10573.

[23] ¨IBM 3990 Storage Control Introduction¨, GA32-0098-0, IBM Corporation, September, 1987.

[24] D. R. Cox, ¨A Note on the Analysis of a Type of Reliability Trial¨, J. SIAM. Appl. Math., Vol. 14, No. 5, pp. 1133-1142, September, 1966.

[25] P. A. W. Lewis, ¨A Branching Poisson Process Model for Analysis of Computer Failure Patterns¨, J. Royal Statistical Society, Sr. B, 26, pp. 398-456, 1964.

[26] N. D. Singpurwalla, ¨Foundational Issues in Reliability and Risk Analysis¨, SIAM Review, Vol. 30, No. 2, pp. 264-282, June 1988.

[27] R. L. Winkler, *¨Introduction to Bayesian Inference¨*, Holt, Rinehart, and Winston, 1972.

[28] D. V. Lindley, ¨Scoring Rules and the Inevitability of Probability¨, International Statistical Review, 50, pp. 1-11, 1982.

[29] D. M. Brender, ¨The Prediction and Measurement of System Availability: A Bayesian Treatment¨, IEEE Transactions on Reliability, Vol. R-17, No. 3, pp. 127-138, Sep. 1968.

[30] D. M. Brender, ¨The Bayesian Assessment of System Availability: Advanced Applications and Techniques¨, IEEE Transactions on Reliability, Vol. R-17, No. 3, pp. 127-138, Sep. 1968.

[31] R. L. Winkler, W.L. Hays, *¨Statistics, Probability, Inference and Decision¨,* 2nd Edition, Holt, Rinehart and Winston, 1975.
[32] P. Veraldi, R. K. Iyer, ¨A Study of Software Failure and Recovery in the MVS Operating System¨, IEEE Transactions on Computers, Vol. C-33, No. 6, pp. 564-568, June 1984.
[33] D. P. Gaver, M. Mazumdar, ¨Some Bayes' Estimates on Long-run Availability in a Two-State System¨, IEEE Transactions on Reliability, Vol. R-18,No. 4, pp. 184-189, 1969.
[34] D. P. Gaver, M. Mazumdar, ¨Statistical Estimation in a Problem of System Reliability¨, Naval Research Logistic Quarterly¨, Vol. 4, pp. 473-488, 1967.
[35] W. E. Thompson, M. D. Springer, ¨A Bayes Analysis of Availability for a system consisting of Several Independent Subsystems¨, IEEE Transactions on Reliability, Vol. R-21, No. 4, pp. 212-214, Nov. 1972.
[36] H. Raiffa, R. Schaifer, ¨Applied Statistical Decision Theory¨, Harvard University Press, 1961.
[37] R. V. Hogg, A. T. Craig, *¨Introduction to Mathematical Statistics¨*, 4th Ed., MacMillan, 1978.
[38] H. Weiler, ¨The Use of Incomplete Beta Functions for Prior Distributions in Binomial Sampling¨, Technometrics, Vol. 7, pp. 335-347, 1965.
[39] M. S. Waterman, H.F. Martz, R. A. Waller, ¨Fitting Beta Prior Distributions in Bayesian Analysis¨, Los Alamos Scientific Laboratory, LA-6395-MS, 1976.
[40] *¨Probability Theory¨*, Volumes I and II, Springer-Verlag, 1977 and 1978.
[41] A. N. Shiryayev, Probability, Springer-Verlag, 1984.
[42] P. S. Laplace, *¨A Philosophical Essay on Probability¨*, Dover, NY, 1951.
[43] R. Von Mises, *Mathematical Theory of Probability and Statistics¨*, Academic Press, NY, 1964.
[44] J. M. Keynes, *¨A Treatise on Probability Theory¨*, MacMillan, London, 1952.
[45] J. C. Laprie, ¨Dependability Evaluation of Software Systems in Operation¨, IEEE Transactions on Software Engineering, SE-10, 6, pp. 701-714, 1984.
[46] R. D. Cheung, ¨A User Oriented Software Reliability Model¨, IEEE Trans. on Software Engg., Vol. SE-6,pp. 118-125, March 1980.
[47] B. L. Littlewood, ¨ Software Reliability Model for Modular Program Structure¨, IEEE Trans. on Reliability, Vol. R-28,pp. 241-246, August, 1979.
[48] H. Cramer, *¨Mathematical Theory of Statistics¨*, Princeton University Press, 1946.
[49] L. J. Savage, ¨The Foundations of Statistics Reconsidered¨, Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Vol. I, Berkeley, University of California Press, pp. 575-586, 1961.
[50] A. P. Dawid, ¨Intersubjective Statistical Models¨, in Exchangeability in Probability and Statistics, G. Koch and F. Spizzichino, eds., pp. 217-232, North Holland, 1982.
[51] A. N. Kolomogorov, Foundations of the Theory of Probability, Chelsea Publishing Co., NY, 1950.
[52] ¨On Tables of Random Numbers¨, Sankhya, Series A, 25, pp. 369-376, 1963.
[53] E. J. Wegman, ¨On Randomness, Determinism, and Computability¨, Technical Report, No. 7, George Mason University, Fairfax, Va, 1986.
[54] S. M. Ross, ¨Statistical Estimation of Software Reliability¨, IEEE Trans. on Software Engineering, Vol. SE-11, No. 5, pp. 479-483, May 1985.
[55] D. E. Brown, ¨A Method for Obtaining Software Reliability Measures During Development¨, IEEE Trans. on Reliability, Vol. R-36, No. 5, pp. 573-580, Dec. 1987.
[56] G. J. MacDonald, Private Communications.
[57] P. C. Fishburn, ¨The Axioms of Subjective Probability¨, Statistical Science, Vol. 1, No. 3, August 191986.
[58] Y. L. Grize, T. L. Fine, ¨Continuous Lower Probability-based Models for Stationary Processes With Bounded and Divergent Time Averages¨, The Annals of Probability, Vol. 15, No. 2, pp. 783-803, 1987.
[59] A. Papamarcum, T. L. Fine, ¨A Note on Undominated Lower Probabilities¨, The Annals of Probability, Vol. 14, No. 2, pp. 710-723, 1986.
[60] P. Walley, T. L. Fine, ¨Towards a Frequentist Theory of Upper and Lower Probability¨, The Annals of Statistics, Vol. 10, No. 3, pp. 741-761, 1982.
[61] G. Shafer, *¨A Mathematical Theory of Evidence¨*, Princeton University Press, 1976.
[62] K. Popper, *¨The Logic of Scientific Discovery¨*, Hutchinson, London, 1959.

[63] M. Herlihy, ¨A Quorum Consensus Replication Method for Abstract Data Types¨, ACM Transactions on Computer Systems, Feb. 1986.
[64] D. Eager, K. Sevcik, ¨Achieving Robustness in Distributed Database Systems¨, ACM Trans. on Database Systems, Sep. 1983.
[65] J. Gray, *¨Operating Systems: An Advanced Course¨*, Springer-Verlag, NY, 1979.
[66] D. Basu, ¨Statistical Information and Likelihood¨, Sankhya, A, 37, pp. 1-71, 1975.