

James L. Johnson
Probability and Statistics,
The Wiley Encyclopedia of Computer Science and Engineering,
Vol. 4, pp. 2208-2221, January, 2009.

Probability and Statistics

As an introductory definition, we consider probability to be a collection of concepts, methods, and interpretations that facilitate the investigation of nondeterministic phenomena. While more mathematically precise definitions will appear in due course, this broadly inclusive description certainly captures the nonspecialist's meaning of probability, and it also aligns rather closely with the intuitive appreciation shared by engineering and computer science professionals. That is, probability is concerned with chance. In a general context, probability represents a systematic attempt to cope with fluctuations of the stock market, winning and losing streaks at the gaming table, and trends announced by economists, sociologists, and weather forecasters. In more scientific contexts, these examples extend to traffic flow patterns on computer networks, request queue lengths at a database server, and charge distributions near a semiconductor junction.

Modern life, despite the predictability afforded by scientific advances and technological accomplishments, offers frequent exposure to chance happenings. It is a safe conclusion that this situation has been part of human existence from prehistoric times. Indeed, despite the associated anxiety, it appears that humans actively seek chance encounters and that they have done so from earliest times. For example, archaeological evidence confirms that gambling, typically in the form of dice games, has been with us for thousands of years. Given this long history, we might expect some accumulation of knowledge, at least at the level of folklore, that purports to improve the outcomes of chance encounters. In Europe, however, no evidence of a science of gambling appears before the fifteenth century, and it is only in the middle of the seventeenth century that such knowledge starts to encompass chance phenomena outside the gaming sphere.

A science of chance in this context means a set of concepts and methods that quantifies the unpredictability of random events. For example, frequency of occurrence provides a useful guideline when betting on a roll of the dice. Suppose your science tells you, from observation or calculation, that a seven occurs about once in every 6 throws and a two occurs about once in every 36 throws. You would then demand a greater payoff for correctly predicting the next throw to be a two than for predicting a seven. Although the dice rolls may be random, there is a pattern that applies to repeated rolls. That pattern governs the relative abundance of sevens and twos among the outcomes.

Probability as a science of chance originated in efforts to quantify such patterns and to construct interpretations useful to the gambler. Many historians associate the birth of probability with a particular event in mid-seventeenth

century France. In 1654, the Chevalier de Mère, a gambler, asked Blaise Pascal and Pierre Fermat, leading mathematicians of the time, to calculate a fair disposition of stakes when a game suffers a premature termination. They reasoned that it was unfair to award all stakes to the party who was leading at the point of interruption. Rather, they computed the outcomes for all possible continuations of the game. Each party is allocated that fraction of the outcomes that result in a win for him. This fraction constitutes the probability that the party would win if the game were completed. Hence, the party's "expectation" is that same fraction of the stakes and such should be his fair allocation in the event of an interrupted game. We note that this resolution accords with the intuition that the party who is ahead at the time of interruption should receive a larger share of the stakes. Indeed, he is "ahead" precisely because the fraction of game continuations in his favor is noticeably larger than that of his opponent.

For his role in gambling problems, such as the one described above, and for his concept of dominating strategies in decision theory, Pascal is sometimes called the father of probability theory. The title is contested, however, as there are earlier contributors. In the 1560s, Girolamo Cardano wrote a book on games of chance, and Fra Luca Pacioli described the division of stakes problem in 1494. In any case, it is not realistic to attribute such a broad subject to a single parental figure. The mid-seventeenth century date of birth, however, is appropriate. Earlier investigations are frequently erroneous and tend to be anecdotal collections, while Pascal's work is correct by modern standards and marks the beginning of a systematic study that quickly attracted other talented contributors.

In 1657, Christian Huygens introduced the concept of expectation in the first printed probability textbook. About the same time John Wallis published an algebra text that featured a discussion of combinatorics. Appearing posthumously in 1713, Jacques Bernoulli's *Ars Conjectandi* (The Art of Conjecture) presented the first limit theorem, which is today known as the Weak Law of Large Numbers. Soon thereafter, Abraham de Moivre demonstrated the normal distribution as the limiting form of binomial distributions, a special case of the Central Limit Theorem. In marked contrast with its earlier focus on gambling, the new science now found application in a variety of fields. In the 1660s, John Hudde and John de Witt provided an actuarial foundation for the annuities used by Dutch towns to finance public works. In 1662, John Graunt published the first set of statistical inferences from mortality records, a subject of morbid interest in those times when the plague was devastating London. Around 1665, Gottfried Leibniz brought probability to the law as he attempted to assess the credibility of judicial evidence.

Excepting Leibniz' concern with the credibility of evidence, all concepts and applications discussed to this point have involved patterns that appear over repeated trials of some nondeterministic process. This is the frequency-of-occurrence interpretation of probability, and it presents some difficulty when contemplating a single trial. Suppose a medical test reveals that there is a 44% chance of your having a particular disease. The frequency-of-occurrence interpretation is that 44% of a large number of persons with the same test results

have actually had the disease. What does this information tell you? At best, it seems to be a rough guideline. If the number were low, say 0.01%, then you might conclude that you have no medical problem. If it were high, say 99.9%, then you would likely conclude that you should undergo treatment. There is a rather large gap between the two extremes for which it is difficult to reach any conclusion. Even more tenuous are statements such as: There is a 30% chance of war if country X acquires nuclear weapons. In this case, we cannot even envision the “large number of trials” that might underlie a frequency-of-occurrence interpretation. These philosophical issues have long contributed to the uneasiness associated with probabilistic solutions. Fortunately for engineering and computer science, these questions do not typically arise. That is, probability applications in these disciplines normally admit a frequency-of-occurrence interpretation. If, for example, we use probability concepts to model query arrivals at a database input queue, we envision an operating environment in which a large number of such queries are generated with random time spacings between them. We want to design the system to accommodate most of the queries in an economic manner, and we wish to limit the consequences of the relatively rare decision to reject a request.

In any case, as probability developed a sophisticated mathematical foundation, the mathematics community took the opportunity to sidestep the entire issue of interpretations. A movement, led primarily by Andrei Kolmogorov, developed *axiomatic* probability theory. This theory defines the elements of interest, such as probability spaces, random variables, distributions and their transforms, as abstract mathematical objects. In this setting, the theorems have specific meanings, and their proofs are the timeless proofs of general mathematics. The theorems easily admit frequency-of-occurrence interpretations, and other interpretations are simply left to the observer without further comment. Modern probability theory, meaning probability as understood by mathematicians since the first third of the twentieth century, is grounded in this approach, and it is along this path that we now begin a technical overview of the subject.

Probability Spaces

Formally, a probability space is a triple $(\Omega, \mathcal{F}, \mathcal{P})$. The first component, Ω , is simply a set of outcomes. Examples are $\Omega = \{\text{heads, tails}\}$ for a coin-flipping scenario, $\Omega = \{1, 2, 3, 4, 5, 6\}$ for a single-die experiment, or $\Omega = \{x : 0 \leq x < 1\}$ for a spinning pointer that comes to rest at some fraction of a full circle. The members of Ω are the occurrences over which we wish to observe quantifiable patterns. That is, we envision that the various members will appear nondeterministically over the course of many trials, but that the relative frequencies of these appearances will tend to have established values. These are the values assigned by the function P to outcomes or to collections of outcomes. These values are known as *probabilities*. In the single-die experiment, for example, we might speak of $P(3) = 1/6$ as the probability of obtaining a three on a single roll. A composite outcome, such as three or four on a single roll, is the probability $P\{3, 4\} = 1/3$.

How are these values determined? In axiomatic probability theory, probabilities are externally specified. For example, in a coin-tossing context, we might know from external considerations that heads appears with relative frequency 0.55. We then simply declare $P(\text{heads}) = 0.55$ as the probability of heads. The probability of tails is then $1.0 - 0.55 = 0.45$. In axiomatic probability theory, these values are assigned by the function P , the third component of the probability space triple. That is, the theory develops under the assumption that the assignments are arbitrary (within certain constraints to be discussed shortly), and therefore any derived results are immediately applicable to any situation where the user can place meaningful frequency assignments on the outcomes.

There is, however, one intervening technical difficulty. The function P does not assign probabilities directly to the outcomes. Rather, it assigns probabilities to *events*, which are subsets of Ω . These events constitute \mathcal{F} , the second element of the probability triple. In the frequency-of-occurrence interpretation, the probability of an event is the fraction of trials that produce an outcome in the event. A subset in \mathcal{F} may contain a single outcome from Ω , in which case this outcome receives a specific probability assignment. However, there may be outcomes that do not occur as singleton subsets in \mathcal{F} . Such an outcome appears only within subsets (events) that receive overall probability assignments. Also, all outcomes of an event $E \in \mathcal{F}$ may constitute singleton events, each with probability assignment zero, while E itself receives a nonzero assignment. Countability considerations, to be discussed shortly, force these subtleties in the general case in order to avoid internal contradictions. In simple cases where Ω is a finite set, such as the outcomes of a dice roll, we can take \mathcal{F} to be all subsets of Ω , including the singletons. In this case, the probability assignment to an event must be the sum of the assignments to its constituent outcomes.

Here are the official rules for constructing a probability space (Ω, \mathcal{F}, P) . First, Ω may be any set whatsoever. For engineering and computer science applications, the most convenient choice is often the real numbers, but any set will do. \mathcal{F} is a collection of subsets chosen from Ω . \mathcal{F} must constitute a *σ -field*, which is a collection containing ϕ , the empty subset, and closed under the operations of complement and countable union. That is,

- $\phi \in \mathcal{F}$
- $A \in \mathcal{F}$ forces $\bar{A} = \Omega - A \in \mathcal{F}$
- $A_n \in \mathcal{F}$ for $n = 1, 2, \dots$ forces $\cup_{n=1}^{\infty} A_n \in \mathcal{F}$.

Finally, the function P maps subsets in \mathcal{F} to real numbers in the range zero to one in a countably additive manner. That is,

- $P : \mathcal{F} \rightarrow [0, 1]$
- $P(\phi) = 0$
- $P(\Omega) = 1$
- $P(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n)$ for pairwise disjoint subsets A_1, A_2, \dots

The events in \mathcal{F} are called *measurable* sets because the function P specifies their sizes as probability allotments. You might well wonder about the necessity of the σ -field \mathcal{F} . As noted above, when Ω is a finite outcome set, \mathcal{F} is normally taken to be the collection of all subsets of Ω . This maximal collection clearly satisfies the requirements. It contains all the singleton outcomes as well as all possible groupings. Moreover, the rule that P must be countably additive forces the probability of any subset to be the sum of the probabilities of its members. This happy circumstance occurs in experiments with dice, coins, and cards, and subsequent sections will investigate some typical probability assignments in those cases.

A countable set is one that can be placed in one-to-one correspondence with the positive integers. The expedient introduced for finite Ω extends to include outcome spaces that are countable. That is, we can still choose \mathcal{F} to be all subsets of the countably infinite Ω . The probability of a subset remains the sum of the probabilities of its members, although this sum now contains countably many terms. However, the appropriate Ω for many scientific purposes, is the set of real numbers. In the nineteenth century, Georg Cantor proved that the real numbers are uncountable, as is any nonempty interval of real numbers.

When Ω is an uncountable set, set-theoretic conundrums beyond the scope of this article force \mathcal{F} to be smaller than the collection of all subsets. In particular, the countable additivity requirement on the function P cannot always be achieved if \mathcal{F} is the collection of all subsets of Ω . On most occasions associated with engineering or computer science applications, the uncountable Ω is the real numbers or some restricted span of real numbers. In this case, we take \mathcal{F} to be the *Borel* sets, which is the smallest σ -field containing all open intervals of the form (a, b) .

Consider again the example where our outcomes are interarrival times between requests in a database input queue. These outcomes can be any positive real numbers, but our instrumentation cannot measure them in infinite detail. Consequently, our interest normally lies with events of the form (a, b) . The probability assigned to this interval should reflect the relative frequency with which the interarrival time falls between a and b . Hence, we do not need all subsets of the positive real numbers among our measurable sets. We need the intervals and also interval combinations achieved through intersections and unions. Because \mathcal{F} must be a σ -field, we must therefore take some σ -field containing the intervals and their combinations. By definition the collection of Borel sets satisfies this condition, and this choice also has happy consequences in the later development of random variables.

While not specified explicitly in the defining constraints, closure under countable intersections is also a property of a σ -field. Moreover, we may interpret countable as either finite or countably infinite. Thus, every σ -field is closed under finite unions and finite intersections. From set-theoretic arguments, we obtain the following properties of the probability assignment function P .

- $P(\bar{A}) = 1 - P(A)$.
- $P(\cup_{n=1}^{\infty} A_n) \leq \sum_{n=1}^{\infty} P(A_n)$, without regard to disjointness among the A_n .

- If $A \subseteq B$, then $P(A) \leq P(B)$.
- If $A_1 \subseteq A_2 \subseteq \dots$, then $P(\cup_{n=1}^{\infty} A_n) = \lim_{n \rightarrow \infty} P(A_n)$.
- If $A_1 \supseteq A_2 \supseteq \dots$, then $P(\cap_{n=1}^{\infty} A_n) = \lim_{n \rightarrow \infty} P(A_n)$.

The final two entries above are collectively known as the Continuity of Measure Theorem. A more precise continuity of measure result is possible with more refined limits. For a sequence a_1, a_2, \dots of real numbers and a sequence A_1, A_2, \dots of elements from \mathcal{F} , we define the limit supremum and the limit infimum as follows.

$$\begin{aligned} \limsup_{n \rightarrow \infty} a_n &= \lim_{k \rightarrow \infty} \sup_{n \geq k} a_n \\ \limsup_{n \rightarrow \infty} A_n &= \cap_{k=1}^{\infty} \cup_{n=k}^{\infty} A_n \\ \liminf_{n \rightarrow \infty} a_n &= \lim_{k \rightarrow \infty} \inf_{n \geq k} a_n \\ \liminf_{n \rightarrow \infty} A_n &= \cup_{k=1}^{\infty} \cap_{n=k}^{\infty} A_n \end{aligned}$$

We adopt the term extended limits for the limit supremum and the limit infimum. While a sequence of bounded real numbers, such as probabilities, may oscillate indefinitely and therefore fail to achieve a well-defined limiting value, every such sequence nevertheless achieves extended limits. We find that $\liminf a_n \leq \limsup a_n$ in all cases, and if the sequence converges, then $\liminf a_n = \lim a_n = \limsup a_n$. For a sequence of subsets in \mathcal{F} , we adopt equality of the extended limits as the definition of convergence. That is, $\lim A_n = A$ if $\liminf A_n = \limsup A_n = A$.

When applied to subsets, we find $\liminf A_n \subseteq \limsup A_n$. Also, for increasing sequences, $A_1 \subseteq A_2 \subseteq \dots$, or decreasing sequences, $A_1 \supseteq A_2 \supseteq \dots$, it is evident that the sequences converge to the countable union and countable intersection respectively. Consequently, the Continuity of Measure Theorem above is actually a statement about convergent sequences of sets: $P(\lim A_n) = \lim P(A_n)$. However, even if the sequence does not converge, we can derive a relationship among the probabilities: $P(\liminf A_n) \leq \liminf P(A_n) \leq \limsup P(A_n) \leq P(\limsup A_n)$.

We conclude this section with two more advanced results that are useful in analyzing sequences of events. Suppose A_n , for $n = 1, 2, \dots$, is a sequence of events in \mathcal{F} for which $\sum_{n=1}^{\infty} P(A_n) < \infty$. The Borel Lemma states that, under these circumstances, $P(\liminf A_n) = P(\limsup A_n) = 0$.

The second result is the Borel-Cantelli Lemma, for which we need the definition of independent subsets. Two subsets A, B in \mathcal{F} are *independent* if $P(A \cap B) = P(A)P(B)$. Now suppose that A_i and A_j are independent for any two distinct subsets in the sequence. Then the lemma asserts that if $\sum_{n=1}^{\infty} P(A_n) = \infty$, then $P(\limsup A_n) = 1$. A more general result, the Kochen-Stone Lemma, provides a lower bound on $P(\limsup A_n)$ under slightly more relaxed conditions.

Combinatorics

Suppose a probability space (Ω, \mathcal{F}, P) features a finite outcome set Ω . If we use the notation $|A|$ to denote the number of elements in a set A , then this condition is $|\Omega| = n$ for some positive integer n . In this case, we take \mathcal{F} to be the collection of all subsets of Ω . We find that $|\mathcal{F}| = 2^n$, and a feasible probability assignment allots probability $1/n$ to each event containing a singleton outcome. The countable additivity constraint then forces each composite event to receive probability equal to the sum of the probabilities of its constituents, which amounts to the size of the event divided by n . In this context combinatorics refers to methods for calculating the size of Ω and for determining the number of constituents in a composite event.

In the simplest cases, this computation is mere enumeration. If Ω contains the six possible outcomes from the roll of a single die, then $\Omega = \{1, 2, 3, 4, 5, 6\}$. We observe that $n = 6$. If an event E is described as “the outcome is odd or it is greater than 4,” then we note that outcomes 1, 3, 5, 6 conform to the description, and calculate $P(E) = 4/6$. In more complicated circumstances, neither n nor the size of E is so easily available. For example, suppose we receive five cards from a shuffled deck of fifty-two standard playing cards. What is the probability that we receive five cards of the same suit with consecutive numerical values (a straight flush)? How many possible hands exist? How many of those constitute a straight flush?

A systematic approach to such problems considers sequences or subsets obtained by choosing from a common pool of candidates. A further distinction appears when we consider two choice protocols: choosing with replacement and choosing without replacement. Two sequences differ if they differ in any position. For example, the sequence 1, 2, 3 is different from the sequence 2, 1, 3. However, two sets differ if one contains an element that is not in the other. Consequently, the sets 1, 2, 3 and 2, 1, 3 are the same set.

Suppose we are choosing k items from a pool of n without replacement. That is, each choice reduces that size of the pool available for subsequent choices. This constraint forces $k \leq n$. Let $P_{k,n}$ be the number of distinct sequences that might be chosen, and let $C_{k,n}$ denote the number of possible sets. We have

$$\begin{aligned} P_{k,n} &= (n)_k^{\downarrow} \equiv n(n-1)(n-2)\cdots(n-k+1) = \frac{n!}{k!} \\ C_{k,n} &= \binom{n}{k} \equiv \frac{n!}{k!(n-k)!}. \end{aligned}$$

Consider again the scenario mentioned above in which we receive five cards from a shuffled deck. We receive one of $(52)_5^{\downarrow} = 311875200$ sequences. To determine if we have received a straight flush, however, we are allowed to reorder the cards in our hand. Consequently, the size of the outcome space is the number of possible sets, rather than the number of sequences. As there are $\binom{52}{5} = 2598960$ such sets, we conclude that the size of the outcome space is $n = 2598960$. Now, how many of these possible hands constitute a straight flush?

For this calculation, it is convenient to introduce another useful counting tool. If we can undertake a choice as succession of subchoices, then the number of candidate choices is the product of the number available at each stage. A straight flush, for example, results from choosing one of four suits and then one of nine low cards to anchor the ascending numerical values. That is, the first subchoice has candidates: clubs, diamonds, hearts, spades. The second subchoice has candidates: 2, 3, ..., 10. The number of candidate hands for a straight flush, and the corresponding probability of a straight flush, are then

$$\begin{aligned} n_{\text{straight flush}} &= \binom{4}{1} \binom{9}{1} = 36 \\ P(\text{straight flush}) &= \frac{36}{2598960} = 0.0000138517. \end{aligned}$$

The multiplicative approach that determines the number of straight flush hands amounts to laying out the hands in four columns, one for each suit, and nine rows, one for each low card anchor value. That is, each candidate from the first subchoice admits the same number of subsequent choices, nine in this example. If the number of subsequent subchoices is not uniform, we resort to summing the values. For example, how many hands exhibit either one or two aces? One-ace hands involve a choice of suit for the ace, followed by a choice of any four cards from the forty-eight non-aces. Two-ace hands require a choice of two suits for the aces, followed by a selection of any three cards from the forty-eight non-aces. The computation is

$$\begin{aligned} n_{\text{one or two aces}} &= \binom{4}{1} \binom{48}{4} + \binom{4}{2} \binom{48}{3} = 882096 \\ P(\text{one or two aces}) &= \frac{882096}{2598960} = 0.3394. \end{aligned}$$

When the selections are performed with replacement, the resulting sequences or sets may contain duplicate entries. In this case, a set is more accurately described as a multiset, which is a set that admits duplicate members. Moreover, the size of the selection, k , may be larger than the size of the candidate pool, n . If we let $\overline{P}_{k,n}$ and $\overline{C}_{k,n}$ denote the number of sequences and multisets, respectively, we obtain

$$\begin{aligned} \overline{P}_{k,n} &= n^k \\ \overline{C}_{k,n} &= \binom{n+k-1}{k}. \end{aligned}$$

These formulas are useful in occupancy problems. For example, suppose we have n bins into which we must distribute k balls. As we place each ball, we choose one of the n bins for it. The chosen bin remains available for subsequent balls; so we are choosing with replacement. A generic outcome is (n_1, n_2, \dots, n_k) , where n_i denotes the bin selected for ball i . There are n^k such outcomes corresponding to the number of such sequences.

If we collect birthdays from a group of ten persons, we obtain a sequence n_1, n_2, \dots, n_{10} , in which each entry is an integer in the range 1 to 365. As each such sequence represents one choice from a field of $\overline{P}_{365,10} = 365^{10}$, we can calculate the probability that there will be at least one repetition among the birthdays by computing the number of such sequences with at least one repetition and dividing by the size of the field. We can construct a sequence with *no* repetitions by selecting, without replacement, ten integers from the range 1 to 365. There are $P_{365,10}$ such sequences, and the remaining sequences must all correspond to least one repeated birthday among the ten people. The probability of a repeated birthday is then

$$P(\text{repeated birthday}) = \frac{365^{10} - P_{365,10}}{365^{10}} = 1 - \frac{(365)(364) \cdots (356)}{365^{10}} = 0.117.$$

As we consider larger groups, the probability of a repeated birthday rises, although many people are surprised by how quickly it becomes larger than 0.5. For example, if we redo the above calculation with 23 persons, we obtain 0.5073 for the probability of a repeated birthday.

Multisets differ from sequences in that a multiset is not concerned with the order of its elements. In the bin-choosing experiment above, a generic multiset outcome is k_1, k_2, \dots, k_n , where k_i counts the number of times bin i was chosen to receive a ball. That is, k_i is the number of occurrences of i in the generic sequence outcome n_1, n_2, \dots, n_k , with a zero entered when a bin does not appear at all. In the birthday example, there are $\overline{C}_{365,10}$ such day-count vectors, but we would not consider them equally likely outcomes. Doing so would imply that the probability of all ten birthdays coinciding is the same as the probability of them dispersing across several days, a conclusion that does not accord with experience.

As an example where the collection of multisets correctly specifies the equally likely outcomes, consider the ways of writing the positive integer k as a sum of n nonnegative integers. A particular sum $k_1 + k_2 + \dots + k_n = k$ is called a *partition* of k into nonnegative components. We can construct such a sum by tossing k ones at n bins. The first bin accumulates summand k_1 , which is equal to the number of times that bin is hit by an incoming one. The second bin plays a similar role for the summand k_2 and so forth. There are $\overline{C}_{3,4} = 15$ ways to partition 4 into 3 components:

$$\begin{array}{cccccc} 0 + 0 + 4 & 0 + 1 + 3 & 0 + 2 + 2 & 0 + 3 + 1 & 0 + 4 + 0 & \\ 1 + 0 + 3 & 1 + 1 + 2 & 1 + 2 + 1 & 1 + 3 + 0 & & \\ 2 + 0 + 2 & 2 + 1 + 1 & 2 + 2 + 0 & & & \\ 3 + 0 + 1 & 3 + 1 + 0 & & & & \\ 4 + 0 + 0. & & & & & \end{array}$$

We can turn the set of partitions into a probability space by assigning probability $1/15$ to each partition. We would then speak of a random partition as one of these 15 equally likely decompositions.

When the bin-choosing experiment is performed with distinguishable balls, then it is possible to observe the outcome n_1, n_2, \dots, n_k , where n_i is the bin chosen for ball i . There are $\overline{P}_{n,k}$ such observable vectors. If the balls are

not distinguishable, the outcome will not contain enough information for us to know the numbers n_i . After the experiment, we cannot locate ball i , and hence we cannot specify its bin. In this case, we know only the multiset outcome k_1, k_2, \dots, k_n , where k_i is the number of balls in bin i . There are only $\overline{C}_{n,k}$ observable vectors of this latter type. In certain physics contexts, probability models with $\overline{P}_{n,k}$ equally likely outcomes accurately describe experiments with distinguishable particles across a range of energy levels (bins). These systems are said to obey *Maxwell-Boltzmann* statistics. On the other hand, if the experiment involves indistinguishable particles, the more realistic model use $\overline{C}_{n,k}$ outcomes, and the system is said to obey *Bose-Einstein* statistics.

The discussion above presents only an introduction to the vast literature of counting methods and their interrelationships that is known as combinatorics. For our purposes, we take these methods as one approach to establishing a probability space over a finite collection of outcomes.

Random Variables and their Distributions

A random variable is a function that maps a probability space into the real numbers (\mathcal{R}). There is a rather subtle constraint. Suppose (Ω, \mathcal{F}, P) is a probability space. Then $X : \Omega \rightarrow \mathcal{R}$ is a random variable if $X^{-1}(B) \in \mathcal{F}$ for all Borel sets $B \subset \mathcal{R}$. This constraint ensures that all events of the form $\{\omega \in \Omega | a < X(\omega) < b\}$ do indeed receive a probability assignment. Such events are typically abbreviated $(a < X < b)$ and are interpreted to mean that the random variable (for the implicit outcome ω) lies in the interval (a, b) . The laws of σ -fields then guarantee that related events, those obtained by unions, intersections, and complements from the open intervals, also receive probability assignments. In other words, X constitutes a *measurable* function from (Ω, \mathcal{F}) to \mathcal{R} .

If the probability space is the real numbers, then the identity function is a random variable. However, for any probability space, we can use a random variable to transfer probability to the Borel sets \mathcal{B} via the prescription $P'(B) = P(\{\omega \in \Omega | X(\omega) \in B\})$, and thereby obtain a new probability space $(\mathcal{R}, \mathcal{B}, P')$. Frequently, all subsequent analysis takes place in the real number setting, and the underlying outcome space plays no further role.

For any $x \in \mathcal{R}$, the function $F_X(x) = P'(X \leq x)$ is called the *cumulative distribution* of random variable X . It is frequently written $F(x)$ when the underlying random variable is clear from context. Distribution functions have the following properties.

- $F(x)$ is monotone nondecreasing.
- $\lim_{x \rightarrow -\infty} F(x) = 0$; $\lim_{x \rightarrow \infty} F(x) = 1$.
- At each point x , F is continuous from the right and possesses a limit from the left. That is, $\lim_{y \rightarrow x^-} F(y) \leq F(x) = \lim_{x \rightarrow x^+} F(y)$.
- The set of discontinuities of F is countable.

Indeed, any function F with these properties is the distribution of some random variable over some probability space. If there exists a nonnegative function f such that $F(x) = \int_{-\infty}^x f(t)dt$, then f is called the *density* of the underlying random variable X . Of course, there are actually many densities, if there is one, because f can be changed arbitrarily at isolated points without disturbing the integral.

Random variables and their distributions provide the opening wedge into a broad spectrum of analytical results because at this point the concepts have been quantified. In working with distributions, we are working with real-valued functions. The first step is to enumerate some distributions that prove useful in computer science and engineering applications. In each case, the underlying probability space is scarcely mentioned. After transferring probability to the Borel sets within the real numbers, all analysis takes place in a real-number setting. When the random variable takes on only a countable number of discrete values, it is traditionally described by giving the probability assigned to each of these values. When the random variable assumes a continuum of values, it is described by its density or distribution.

The *Bernoulli* random variable models experiments with two outcomes. It is an abstraction of the coin-tossing experiment, and it carries a parameter that denotes the probability of “heads.” Formally, a Bernoulli random variable X takes on only two values: 0 or 1. We say that X is a Bernoulli random variable with parameter p if $P(X = 1) = p$ and (necessarily) $P(X = 0) = 1 - p$.

The *expected value* of a discrete random variable X , denoted $E[X]$, is

$$E[X] = \sum_{n=1}^{\infty} t_n \cdot P(X = t_n),$$

where t_1, t_2, \dots enumerates the discrete values that X may assume. The expected value represents a weighted average across the possible outcomes. The *variance* of a discrete random variable is

$$\text{Var}[X] = \sum_{n=1}^{\infty} (t_n - E[X])^2 \cdot P(X = t_n).$$

The variance provides a summary measure of the dispersion of the X values about the expected value with low variance corresponding to a tighter clustering. For a Bernoulli random variable X with parameter p , we have $E[X] = p$ and $\text{Var}[X] = p(1 - p)$.

An *indicator* random variable is a Bernoulli random variable that takes on the value 1 precisely when some other random variable falls in a prescribed set. For example, suppose we have a random variable X which measures the service time (seconds) of a customer with a bank teller. We might be particularly interested in lengthy service times, say those that exceed 120 seconds. The indicator

$$I_{(120, \infty)} = \begin{cases} 1, & X > 120 \\ 0, & X \leq 120 \end{cases}$$

is a Bernoulli random variable with parameter $p = P(X > 120)$.

Random variables X_1, X_2, \dots, X_n are *independent* if, for any Borel sets B_1, B_2, \dots, B_n , the probability that all n random variables lie in their respective sets is the product of the individual occupancy probabilities. That is,

$$P\left(\bigcap_{i=1}^n (X_i \in B_i)\right) = \prod_{i=1}^n P(X_i \in B_i).$$

This definition is a restatement of the concept of independent events introduced earlier; the events are now expressed in terms of the random variables. Because the Borel sets constitute a σ -field, it suffices to check the above condition on Borel sets of the form $(X \leq t)$. That is, X_1, X_2, \dots, X_n are independent if, for any n -tuple of real numbers t_1, t_2, \dots, t_n , we have

$$P\left(\bigcap_{i=1}^n (X_i \leq t_i)\right) = \prod_{i=1}^n P(X_i \leq t_i) = \prod_{i=1}^n F_{X_i}(t_i).$$

The sum of n independent Bernoulli random variables, each with parameter p , exhibits a *binomial* distribution. That is, if X_1, X_2, \dots, X_n are Bernoulli with parameter p and $Y = \sum_{i=1}^n X_i$, then

$$P(Y = i) = \binom{n}{i} p^i (1-p)^{n-i},$$

for $i = 0, 1, 2, \dots, n$. This random variable models, for example, the number of heads in n tosses of a coin for which the probability of a head on any given toss is p . For linear combinations of independent random variables, expected values and variances are simple functions of the component values.

$$\begin{aligned} E[a_1 X_1 + a_2 X_2 + \dots] &= a_1 E[X_1] + a_2 E[X_2] + \dots \\ \text{Var}[a_1 X_1 + a_2 X_2 + \dots] &= a_1^2 \text{Var}[X_1] + a_2^2 \text{Var}[X_2] + \dots \end{aligned}$$

For the binomial random variable Y above, therefore, we have $E[Y] = np$ and $\text{Var}[Y] = np(1-p)$.

A *Poisson* random variable X with parameter λ has $P(X = k) = e^{-\lambda} \lambda^k / k!$. This random variable assumes all nonnegative integer values, and it is useful for modeling the number of events occurring in a specified interval when it is plausible to assume that the count is proportional to the interval width in the limit of very small widths. Specifically, the following context gives rise to a Poisson random variable X with parameter λ . Suppose, as time progresses, some random process is generating events. Let $X_{t,\Delta}$ count the number of events that occur during the time interval $[t, t + \Delta]$. Now, suppose further that the generating process obeys three assumptions. The first is a homogeneity constraint:

- $P(X_{t_1,\Delta} = k) = P(X_{t_2,\Delta} = k)$ for all integers $k \geq 0$.

That is, the probabilities associated with an interval of width Δ do not depend on the location of the interval. This constraint allows a notational simplification,

and we can now speak of X_Δ because the various random variables associated with different anchor positions t all have the same distribution. The remaining assumptions are

- $P(X_\Delta = 1) = \lambda\Delta + o_1(\Delta)$
- $P(X_\Delta > 1) = o_2(\Delta)$,

where the $o_i(\Delta)$ denote anonymous functions with the property that $o_i(\Delta)/\Delta \rightarrow 0$ as $\Delta \rightarrow 0$. Then the assignment $P(X = k) = \lim_{\Delta \rightarrow 0} P(X_\Delta = k)$ produces a Poisson random variable.

This model accurately describes such diverse phenomena as particles emitted in radioactive decay, customer arrivals at an input queue, flaws in magnetic recording media, airline accidents, and spectator coughing during a concert. The expected value and variance are both λ . If we consider a sequence of binomial random variables, $B_{n,p}$, where the parameters n and p are constrained such that $n \rightarrow \infty$ and $p \rightarrow 0$ in a manner that allows $np \rightarrow \lambda > 0$, then the distributions approach that of a Poisson random variable Y with parameter λ . That is, $P(B_{n,p} = k) \rightarrow P(Y = k) = e^{-\lambda}\lambda^k/k!$.

A *geometric* random variable X with parameter p exhibits $P(X = k) = p(1-p)^k$ for $k = 0, 1, 2, \dots$. It models, for example, the number of tails before the first head in repeated tosses of a coin for which the probability of heads is p . We have $E[X] = (1-p)/p$ and $\text{Var}[X] = (1-p)/p^2$. Suppose, for example, that we have a hash table in which j of the N addresses are unoccupied. If we generate random address probes in search of an unoccupied slot, the probability of success is j/N for each probe. The number of failures prior to the first success then follows a geometric distribution with parameter j/N .

The sum of n independent geometric random variables displays a *negative binomial* distribution. That is, if X_1, X_2, \dots, X_n are all geometric with parameter p , then $Y = X_1 + X_2 + \dots + X_n$ is negative binomial with parameters (n, p) . We have

$$\begin{aligned} P(Y = k) &= C_{n+k-1, k} p^n (1-p)^k \\ E[Y] &= \frac{n(1-p)}{p} \\ \text{Var}[Y] &= \frac{n(1-p)}{p^2}, \end{aligned}$$

where $C_{n+k-1, k}$ is the number of distinct multisets available when choosing k from a field of n with replacement. This random variable models, for example, the number of tails before the n^{th} head in repeated coin tosses, the number of successful flights prior to the n^{th} accident in an airline history, or the number of defective parts chosen (with replacement) from a bin prior to the n^{th} functional one. For the hash table example above, if we are trying to fill n unoccupied slots, the number of unsuccessful probes in the process will follow a negative binomial distribution with parameters $n, j/N$. In this example, we assume that

n is significantly smaller than N , so that insertions do not materially change the probability j/N of success for each address probe.

Moving on to random variables that assume a continuum of values, we describe each by giving its density function. The summation formulas for the expected value and variance become integrals involving this density. That is, if random variable X has density f , then

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} tf(t)dt \\ \text{Var}[X] &= \int_{-\infty}^{\infty} [t - E[X]]^2 f(t)dt. \end{aligned}$$

In truth, precise work in mathematical probability uses a generalization of the familiar Riemann integral known as a *measure-theoretic* integral. The separate formulas, summation for discrete random variables and Riemann integration against a density for continuous random variables, are then subsumed under a common notation. This more general integral also enables computations in cases where a density does not exist. When the measure in question corresponds to the traditional notion of length on the real line, the measure-theoretic integral is known as the *Lebesgue* integral. In other cases, it corresponds to a notion of length accorded by the probability distribution: $P(a < X < t)$ for real a and b . In most instances of interest in engineering and computer science, the form involving ordinary integration against a density suffices.

The *uniform* random variable U on $[0, 1]$ is described by the constant density $f(t) = 1$ for $0 \leq t \leq 1$. The probability that U falls in a subinterval (a, b) within $[0, 1]$ is simply $b - a$, the length of that subinterval. We have

$$\begin{aligned} E[U] &= \int_0^1 tdt = \frac{1}{2} \\ \text{Var}[U] &= \int_0^1 \left(t - \frac{1}{2}\right)^2 dt = \frac{1}{12}. \end{aligned}$$

The uniform distribution is the continuous analog of the equally likely outcomes discussed in the combinatorics section above. It assumes that any outcome in the interval $[0, 1]$ is equally likely to the extent possible under the constraint that probability must now be assigned to Borel sets. In this case, all individual outcomes receive zero probability, but intervals receive probability in proportion to their lengths. This random variable models situations such as the resting position of a spinning pointer, where no particular location has any apparent advantage.

The most famous continuous random variable is the *Gaussian* or *normal* random variable $Z_{\mu,\sigma}$. It is characterized by two parameters, μ and σ , and has density, expected value, and variance as follows.

$$f_{Z_{\mu,\sigma}}(t) = \frac{1}{\sigma\sqrt{2\pi}}e^{-(t-\mu)^2/2\sigma^2}$$

$$\begin{aligned} E[Z_{\mu,\sigma}] &= \mu \\ \text{Var}[Z_{\mu,\sigma}] &= \sigma^2 \end{aligned}$$

The well-known *Central Limit Theorem* states that the average of a large number of independent observations behaves like a Gaussian random variable. Specifically, if X_1, X_2, \dots are independent random variables with identical finite-variance distributions, say $E[X_i] = a$ and $\text{Var}[X_i] = c^2$, then for any t ,

$$\lim_{n \rightarrow \infty} P\left(\frac{1}{\sqrt{nc^2}} \sum_{i=1}^n (X_i - a) \leq t\right) = P(Z_{0,1} \leq t).$$

For example, if we toss a fair coin 100 times, what is the probability that we will see 40 or fewer heads? To use the Central Limit Theorem, we let $X_i = 1$ if heads occurs on the i^{th} toss and zero otherwise. With this definition, we have $E[X_i] = 0.5$ and $\text{Var}[X_i] = 0.25$, which yields

$$\begin{aligned} P\left(\sum_{i=1}^{100} X_i \leq 40\right) &= P\left(\frac{1}{\sqrt{100(0.25)}} \sum_{i=1}^{100} (X_i - 0.5) \leq \frac{40 - 100(0.5)}{\sqrt{100(0.25)}}\right) \\ &\approx P(Z_{0,1} \leq -2) = 0.0288. \end{aligned}$$

The last equality was obtained from a tabulation of such values for the *standard* normal random variable with expected value 0 and variance 1.

Because it represents a common limiting distribution for an average of independent observations, the Gaussian random variable is heavily used to calculate *confidence intervals* that describe the chance of correctly estimating a parameter from multiple trials. We will return to this matter in a subsequent section.

The *Gamma* function is $\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx$, defined for $t > 0$. The Gamma random variable X with parameters (γ, λ) (both positive) is described by the density

$$f(x) = \begin{cases} \lambda^\gamma x^{\gamma-1} e^{-\lambda x} / \Gamma(\gamma), & \text{for } x \geq 0 \\ 0, & \text{for } x < 0. \end{cases}$$

It has $E[X] = \gamma/\lambda$ and $\text{Var}[X] = \gamma/\lambda^2$. For certain specific values of γ , the random variable is known by other names. If $\gamma = 1$, the density reduces to $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$, and X is then called an *exponential* random variable. The exponential random variable models the interarrival times associated with events such as radioactive decay and customer queues, which were discussed in connection with Poisson random variables above. Specifically, if a Poisson random variable with parameter λT models the number of events in interval T , then an exponential random variable with parameter λ models their interarrival times. Consequently, the exponential random variable features prominently in queueing theory.

Exponential random variables possess a remarkable feature; they are *memoryless*. To understand this concept, we must first define the notion of *conditional probability*. We will use the exponential random variable as an example,

although the discussion applies equally well to random variables in general. Notationally, we have a probability space (Ω, \mathcal{F}, P) and a random variable X , for which

$$P\{\omega \in \Omega : X(\omega) > t\} = \int_t^\infty \lambda e^{-\lambda x} dx = e^{-\lambda t},$$

for $t \geq 0$. Let t_1 be a fixed positive real number, and consider a related probability space $(\hat{\Omega}, \hat{\mathcal{F}}, \hat{P})$, obtained as follows.

$$\begin{aligned}\hat{\Omega} &= \{\omega \in \Omega : X(\omega) > t_1\} \\ \hat{\mathcal{F}} &= \{A \cap \hat{\Omega} : A \in \mathcal{F}\} \\ \hat{P}(B) &= P(B)/P(\hat{\Omega}), \quad \text{for all } B \in \hat{\mathcal{F}}\end{aligned}$$

By restricting its domain, we can consider X to be a random variable on $\hat{\Omega}$. For any $\omega \in \hat{\Omega}$, we have $X(\omega) > t_1$, but we can legitimately ask the probability, using the new measure \hat{P} , that $X(\omega)$ exceeds t_1 by more than t .

$$\begin{aligned}\hat{P}(X > t_1 + t) &= \frac{P(X > t_1 + t_2)}{P(X > t_1)} = \frac{e^{-\lambda(t_1+t)}}{e^{-\lambda t_1}} \\ &= e^{-\lambda t} = P(X > t).\end{aligned}$$

The probability $\hat{P}(B)$ is known as the *conditional probability* of B , given $\hat{\Omega}$. From the calculation above, we see that the conditional probability that X exceeds t_1 by t or more, given that $X > t_1$ is equal to the unconditional probability that $X > t$. This is the memoryless property. If X is an exponential random variable representing the time between query arrivals to a database input queue, then the probability that 6 microseconds or more elapses before the next arrival is the same as the probability that an additional 6 microseconds or more elapses before the next arrival, given that we have already waited in vain for 10 microseconds.

In general, we can renormalize our probability assignments by restricting the outcome space to some particular event, such as the $\hat{\Omega}$ in the example. The more general notation is $P(B|A)$ for the conditional probability of B given A . Also, we normally allow B to be any event in the original \mathcal{F} with the understanding that only that part of B that intersects A carries nonzero probability under the new measure. The definition requires that the conditioning event A have nonzero probability. In that case,

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

specifies the revised probabilities for all events B . Note that

$$\begin{aligned}P(B|A) &= \frac{P(A \cap B)}{P(A)} = \frac{P(A \cap B)}{P(A \cap B) + P(A \cap \bar{B})} \\ &= \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|\bar{B})P(\bar{B})}.\end{aligned}$$

This formula, a simple form of *Bayes' Law*, relates the conditional probability of B given A to that of A given B . It finds frequent use in updating probability assignments to reflect new information. Specifically, suppose we know $P(B)$ and therefore $P(\overline{B}) = 1 - P(B)$. Such probabilities are called *prior* probabilities because they reflect the chances of a B occurrence in the absence of further knowledge about the underlying random process. If the actual outcome remains unknown to us, but we are told that event A has occurred, we may want to update our probability assignment to reflect more accurately the chances that B has also occurred. That is, we are interested in the *posterior* probability $P(B|A)$. Bayes' Law allows us to compute this new value, provided we also have the reverse conditional probabilities.

For example, suppose a medical test for a specific disease is applied to a large population of persons known to have the disease. In 99% of the cases, the disease is detected. This is a conditional probability. If we let S be the event "person is sick" and "+" be the event "medical test was positive," we have $P(+|S) = 0.99$ as an empirical estimate. Applying the test to a population of persons known not to have the disease might reveal $P(+|\overline{S}) = 0.01$ as a false alarm rate. Suppose further that the fraction $P(S) = 0.001$ of the general population is sick with the disease. Now, if you take the test with positive results, what is the chance that you have the disease? That is, what is $P(S|+)$. Applying Bayes' Law, we have

$$P(S|+) = \frac{P(+|S)P(S)}{P(+|S)P(S) + P(+|\overline{S})P(\overline{S})} = \frac{0.99(0.001)}{0.99(0.001) + 0.01(0.999)} = 0.0909.$$

You have only a 9% chance of being sick, despite having scored positive on a test with an apparent 1% error rate. Nevertheless, your chance of being sick has risen from a prior value of 0.001 to a posterior value of 0.0909. This is nearly a hundredfold increase, which is commensurate with the error rate of the test.

The full form of Bayes' Law uses an arbitrary partition of the outcome space, rather than a simple two-event decomposition, such as "sick" and "not sick." Suppose the event collection $\{A_i : 1 \leq i \leq n\}$ is a partition of the outcome space Ω . That is, the A_i are disjoint, each has nonzero probability, and their union comprises all of Ω . We are interested in which A_i has occurred, given knowledge of another event B . If we know the reverse conditional probabilities, that is if we know the probability of B given each A_i , then Bayes' Law enables the computation

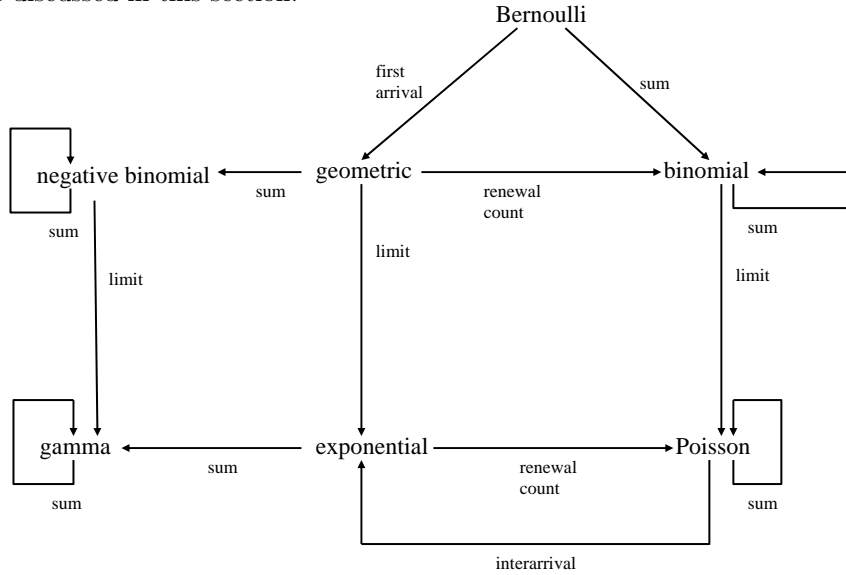
$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^n P(B|A_j)P(A_j)}.$$

Returning to the Gamma random variable with parameters (γ, λ) , we can distinguish additional special cases. If $\gamma = n$, a positive integer, then the corresponding Γ distribution is called an *Erlang* distribution. It models the time necessary to accumulate n events in a process that follows a Poisson distribution for the number of events in a specified interval. An Erlang distribution, for example, describes the time span covering the next n calls arriving at a telephone exchange.

If $\gamma = n/2$ for a positive integer n and $\lambda = 1/2$, then the corresponding Gamma random variable is called a chi-square random variable. It exhibits the distribution of the sum of n independent squares, $Y = \sum_{i=1}^n Z_i^2$, where each Z_i is a Gaussian random variable with $(\mu, \sigma^2) = (0, 1)$. These distributions are useful in computing confidence intervals for statistical estimates.

Gamma distributions are the limiting forms of negative binomial random variables in the same manner that the Poisson distribution is the limit of binomials. That is, suppose we have a sequence C_n of negative binomial random variables. The parameters of C_n are (m, p_n) . As $n \rightarrow \infty$, we assume that $p_n \rightarrow 0$ in a manner that allows $np_n \rightarrow \lambda > 0$. Then the limiting distribution of C_n/n is the Gamma (Erlang) distribution with parameters (m, λ) . In particular, if $m = 1$, the C_n are geometric and the limit is exponential.

The following diagram summarizes the relationships among the random variables discussed in this section.



The *renewal count* arrow from exponential to Poisson refers to the fact that a phenomenon in which the event interarrival time is exponential (λ) will accumulate events in an interval T according to a Poisson distribution with parameter λT . That is, if the sequence X_1, X_2, \dots of random variables measures time between successive events, then the random variable

$$N_T = \max\left\{k \mid \sum_{i=1}^k X_i \leq T\right\}$$

is called a renewal count for the sequence. If the X_i are independent exponentials with parameter λ , then N_T has a Poisson distribution with parameter λT .

A similar relationship holds between a sequence $G_1 + 1, G_2 + 1, \dots$ of geometric random variables with a common parameter p . The difference is that the

observation interval T is now a positive integer. The renewal count N_T then exhibits a binomial distribution with parameters (T, p) .

Convergence Modes

For a sequence of real numbers, there is a single mode of convergence: A tail of the sequence must enter and remain within any given neighborhood of the limit. This property either holds for some (possibly infinite) limiting value, or it does not. Sequences of random variables exhibit more variety in this respect. There are three modes of convergence.

A sequence X_1, X_2, \dots of random variables converges *pointwise* to a random variable Y if $X_n(\omega) \rightarrow Y(\omega)$ as a sequence of real numbers for every point ω in the underlying probability space. We may also have pointwise convergence on sets smaller than the full probability space. If pointwise convergence occurs on a set of probability one, then we say that the sequence converges *almost surely*. In this case, we use the notation $X_n \rightarrow Y$ *a.s.*

The sequence converges *in probability* if, for every positive ϵ , the measure of the misbehaving sets approaches zero. That is, as $n \rightarrow \infty$,

$$P(\{\omega : |X_n(\omega) - Y(\omega)| > \epsilon\}) \rightarrow 0.$$

If $X_n \rightarrow Y$ *a.s.*, then it also converges in probability. However, it is possible for a sequence to converge in probability and at the same time have no pointwise limits.

The final convergence mode concerns distribution functions. The sequence converges *in distribution* if the corresponding cumulative distribution functions of the X_n converge pointwise to the distribution function of Y at all points where the cumulative distribution function of Y is continuous.

The *Weak Law of Large Numbers* states that the average of a large number of independent, identically distributed random variables tends in probability to a constant, the expected value of the common distribution. That is, if X_1, X_2, \dots is an independent sequence with a common distribution such that $E[X_n] = \mu$ and $\text{Var}[X_n] = \sigma^2 < \infty$, then for every positive ϵ ,

$$P\left(\left\{\omega : \left|\frac{\sum_{i=1}^n X_i}{n} - \mu\right| > \epsilon\right\}\right) \rightarrow 0,$$

as $n \rightarrow \infty$.

Suppose, for example, that a random variable T measures the time between query requests arriving at a database server. This random variable is likely to exhibit an exponential distribution, as described in the previous section, with some rate parameter λ . The expected value and variance are $1/\lambda$ and $1/\lambda^2$ respectively. We take n observations of T and label them T_1, T_2, \dots, T_n . The weak law suggests that the number $\sum_{i=1}^n T_i/n$ will be close to $1/\lambda$. The precise meaning is more subtle. Since an exponential random variable can assume any nonnegative value, we can imagine a sequence of observations that are all larger than, say, twice the expected value. In that case, the average would also be

much larger than $1/\lambda$. It is then clear that not all sequences of observations will produce averages close to $1/\lambda$. The weak law states that the set of sequences that misbehave in this fashion is not large, when measured in terms of probability.

We envision an infinite sequence of independent database servers, each with its separate network of clients. Our probability space now consists of outcomes of the form $\omega = (t_1, t_2, \dots)$, which occurs when server 1 waits t_1 seconds for its next arrival, server 2 waits t_2 seconds, and so forth. Any event of the form $(t_1 \leq x_1, t_2 \leq x_2, \dots, t_p \leq x_p)$ has probability equal to the product of the factors $P(t_i \leq x_i)$, which are in turn determined by the common exponential distribution of the T_i . By taking unions, complements, and intersections of events of this type, we arrive at a σ -field that supports the probability measure. The random variables $\sum_{i=1}^n T_i/n$ are well defined on this new probability space, and the weak law asserts that, for large n , the set of sequences (t_1, t_2, \dots) with misbehaving prefixes (t_1, t_2, \dots, t_n) has small probability.

A given sequence can drift into and out of the misbehaving set as n increases. Suppose the average of the first 100 entries is close to $1/\lambda$, but the next 1000 entries are all larger than twice $1/\lambda$. The sequence is then excluded from the misbehaving set at $n = 100$ but enters that set before $n = 1100$. Subsequent patterns of good and bad behavior can migrate the sequence into and out of the exceptional set. With this additional insight, we can interpret more precisely the meaning of the weak law.

Suppose $1/\lambda = 0.4$. We can choose $\epsilon = 0.04$ and let $Y_n = \sum_{i=1}^n T_i/n$. The weak law asserts that $P(|Y_n - 0.4| > 0.04) \rightarrow 0$, which is the same as $P(0.36 \leq Y_n \leq 0.44) \rightarrow 1$. While the law does not inform us about the actual size of n required, it does say that eventually this latter probability exceeds 0.99. Intuitively, this means that if we choose a large n , there is a scant 1% chance that our average will fail to fall close to 0.4. Moreover, as we choose larger and larger values for n , that chance decreases.

The *Strong Law of Large Numbers* states that the average converges pointwise to the common expected value, except perhaps on a set of probability zero. Specifically, if X_1, X_2, \dots is an independent sequence with a common distribution such that $E[X_n] = \mu$ (possibly infinite), then

$$\frac{\sum_{i=1}^n X_i}{n} \rightarrow \mu \quad a.s.$$

as $n \rightarrow \infty$.

Applied in the above example, the strong law asserts that essentially all outcome sequences exhibit averages that draw closer and closer to the expected value as n increases. The issue of a given sequence forever drifting into and out of the misbehaving set is placed in a pleasant perspective. Such sequences must belong to the set with probability measure zero. This reassurance does not mean that the exceptional set is empty, because individual outcomes (t_1, t_2, \dots) have zero probability. It does mean that we can expect, with virtual certainty, that our average of n observations of the arrival time will draw ever closer to the expected $1/\lambda$.

Although the above convergence results can be obtained with set-theoretic arguments, further progress is greatly facilitated with the concept of characteristic functions, which are essentially Fourier transforms in a probability space setting. For a random variable X , the *characteristic function* of X is the complex-valued function $\beta_X(u) = E[e^{iuX}]$. The exceptional utility of this device follows because there is a one-to-one relationship between characteristic functions and their generating random variables (distributions). For example, X is Gaussian with parameters $\mu = 0$ and $\sigma^2 = 1$ if and only if $\beta_X(u) = e^{-u^2/2}$. X is Poisson with parameter λ if and only if $\beta_X(u) = \exp(-\lambda(1 - e^{iu}))$.

If X has a density $f(t)$, the computation of β_X is a common integration: $\beta_X(u) = \int_{-\infty}^{\infty} e^{iut} f(t) dt$. Conversely, if β_X is absolutely integrable, then X has a density, which can be recovered by an inversion formula. That is, if $\int_{-\infty}^{\infty} |\beta(u)| du < \infty$, then the density of X is

$$f_X(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-iut} \beta(u) du.$$

These remarks have parallel versions if X is integer-valued. The calculation of β_X is a sum: $\beta_X(u) = \sum_{n=-\infty}^{\infty} e^{iun} P(X = n)$. Also, if β_X is periodic with period 2π , then the corresponding X is integer-valued and the point probabilities are recovered with a similar inversion formula:

$$P(X = n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-iun} \beta(u) du.$$

In more general cases, the β_X computation requires the measure-theoretic integral referenced earlier, and the recovery of the distribution of X requires more complex operations on β_X . Nevertheless, it is theoretically possible to translate in both directions between distributions and their characteristic functions.

Some useful properties of characteristic functions are as follows.

- (Linear combinations) If $Y = aX + b$, then $\beta_Y(u) = e^{iub} \beta_X(au)$.
- (Independent sums). If $Y = X_1 + X_2 + \dots + X_n$, where the X_i are independent, then $\beta_Y(u) = \prod_{i=1}^n \beta_{X_i}(u)$.
- (Continuity Theorem) A sequence of random variables X_1, X_2, \dots converges in distribution to a random variable X if and only if the corresponding characteristic functions converge pointwise to a function that is continuous at zero, in which case the limiting function is the characteristic function of X .
- (Moment Theorem) If $E[|X|^n] < \infty$, then β_X has derivatives through order n and $E[X^n] = (-i)^n \beta_X^{(n)}(0)$, where $\beta_X^{(n)}(u)$ is the n^{th} derivative of β_X .

These features allow us to study convergence in distribution of random variables by investigating the more tractable pointwise convergence of their characteristic

functions. In the case of independent, identically distributed random variables with finite variance, this method leads quickly to the Central Limit Theorem cited earlier.

For a nonnegative random variable X , the *moment generating function* $\phi_X(u)$ is less difficult to manipulate. Here $\phi_X(u) = E[e^{-uX}]$. For a random variable X that assumes only nonnegative integer values, the *probability generating function* $\rho_X(u)$ is another appropriate transform. It is defined by $\rho_X(u) = \sum_{n=0}^{\infty} P(X = n)u^n$. Both moment and probability generating functions admit versions of the moment theorem and the continuity theorem and are therefore useful for studying convergence in the special cases where they apply.

Computer Simulations

In various programming contexts, particularly with simulations, the need arises to generate samples from some particular distribution. For example, if we know that $P(X = 1) = 0.4$ and $P(X = 0) = 0.6$, we may want to *realize* this random variable as a sequence of numbers x_1, x_2, \dots . This sequence should exhibit the same variability as would the original X if outcomes were directly observed. That is, we expect a thorough mixing of ones and zeros, with about 40% ones. Notice that we can readily achieve this result if we have a method for generating samples from a uniform distribution U on $[0, 1]$. In particular, each time we need a new sample of X , we generate an observation U and report $X = 1$ if $U \leq 0.4$ and $X = 0$ otherwise.

This argument generalizes in various ways, but the gist of the extension is that essentially any random variable can be sampled by first sampling a uniform random variable and then resorting to some calculations on the observed value. While this reduction simplifies the problem, there remains the necessity of simulating observations from a uniform distribution on $[0, 1]$. Here we encounter two difficulties. First, the computer operates with a finite register length, say 32 bits, which means that the values returned are patterns from the 2^{32} possible arrangements of 32 bits. Second, a computer is a deterministic device.

To circumvent the first problem, we put a binary point at the left end of each such pattern, obtaining 2^{32} evenly spaced numbers in the range $[0, 1)$. The most uniform probability assignment allots probability $1/2^{32}$ to each such point. Let U be the random variable that operates on this probability space as the identity function. If we calculate $P(a < U < b)$ for subintervals (a, b) that are appreciably wider than $1/2^{32}$, we discover that these probabilities are nearly $b - a$, which is the value required for a true uniform random variable. The second difficulty is overcome by arranging that the values returned on successive calls exhaust, or very nearly exhaust, the full range of patterns before repeating. In this manner, any deterministic behavior is not observable under normal use. Some modern supercomputer computations may involve more than 2^{32} random samples, an escalation that has forced the use of 64 bit registers to maintain the appearance of nondeterminism.

After accepting an approximation based on 2^{32} (or more) closely spaced numbers in $[0, 1)$, we still face the problem of simulating a discrete probability

distribution on this finite set. This problem remains an area of active research today. One popular approach is the *linear congruential* method. We start with a seed sample x_0 , which is typically obtained in some nonreproducible manner, such as extracting a 32 bit string from the computer real-time clock. Subsequent samples are obtained with the recurrence $x_{n+1} = (ax_n + b) \bmod c$, where the parameters a, b, c are chosen to optimize the criteria of a long period before repetition and a fast computer implementation. For example, c is frequently chosen to be 2^{32} because the $(ax_n + b) \bmod 2^{32}$ operation involves retaining only the least significant 32 bits of $(ax_n + b)$. Knuth[6] discusses the mathematics involved in choosing these parameters.

On many systems, the resulting generator is called `rand()`. A program assignment statement, such as $x = \text{rand}()$, places a new sample in the variable x . From this point, we manipulate the returned value to simulate samples from other distributions. As noted above, if we wish to sample B , a Bernoulli random variable with parameter p , we continue by setting $B = 1$ if $x \leq p$ and $B = 0$ otherwise. If we need an random variable $U_{a,b}$, uniform on the interval $[a, b]$, we calculate $U_{a,b} = a + (b - a)x$.

If the desired distribution has a continuous cumulative distribution function, a general technique, called distribution inversion, provides a simple computation of samples. Suppose X is a random variable for which the cumulative distribution $F(t) = P(X \leq t)$ is continuous and strictly increasing. The inverse $F^{-1}(u)$ then exists for $0 < u < 1$, and it can be shown that the derived random variable $Y = F(X)$ has a uniform distribution on $(0, 1)$. It follows that the distribution of $F^{-1}(U)$ is the same as that of X , where U is the uniform random variable approximated by `rand()`. To obtain samples from X , we sample U instead and return the values $F^{-1}(U)$.

For example, the exponential random variable X with parameter λ has the cumulative distribution function $F(t) = 1 - e^{-\lambda t}$, for $t \geq 0$, which satisfies the required conditions. The inverse is $F^{-1}(u) = -[\log(1 - u)]/\lambda$. If U is uniformly distributed, so is $1 - U$. Therefore, the samples obtained from successive $-[\log(\text{rand}())]/\lambda$ values exhibit the desired exponential distribution.

A variation is necessary to accommodate discrete random variables, such as those that assume integer values. Suppose we have a random variable X that assumes nonnegative integer values n with probabilities p_n . Because the cumulative distribution now exhibits a discrete jump at each integer, it no longer possesses an inverse. Nevertheless, we can salvage the idea by acquiring a `rand()` sample, say x , and then summing the p_n until the accumulation exceeds x . We return the largest n such that $\sum_{i=0}^n p_i \leq x$. A moment's reflection will show that this is precisely the method we used to obtain samples from a Bernoulli random variable above.

For certain cases, we can solve for the required n . For example, suppose X is a geometric random variable with parameter p . In this case, $p_n = p(1 - p)^n$. Therefore if x is the value obtained from `rand()`, we find

$$\max\left\{n : \sum_{k=0}^n p_k \leq x\right\} = \left\lfloor \frac{\log x}{\log(1 - p)} \right\rfloor.$$

For more irregular cases, we may need to perform the summation. Suppose we want to sample a Poisson random variable X with parameter λ . In this case, we have $p_n = e^{-\lambda}\lambda^n/n!$, and the following pseudocode illustrates the technique. We exploit the fact that $p_0 = e^{-\lambda}$ and $p_{n+1} = p_n\lambda/(n+1)$.

```

x = rand();
p = exp(-λ);
cum = p;
n = 0;
while (x > cum){
    n = n + 1;
    p = p * λ/(n + 1);
    cum = cum + p; }
return n;

```

Various enhancements are available to reduce the number of iterations necessary to locate the desired n to return. In the above example, we could start the search near $n = \lfloor \lambda \rfloor$, because values near this expected value are most frequently returned.

Another method for dealing with irregular discrete distributions is the *rejection filter*. If we have an algorithm to simulate distribution X , we can, under certain conditions, systematically withhold some of the returns to simulate a related distribution Y . Suppose X assumes nonnegative integer values with probabilities p_0, p_1, \dots , while Y assumes the same values but with different probabilities q_0, q_1, \dots . The required condition is that there exists a positive K such that $q_n \leq Kp_n$ for all n . The following pseudocode shows how to reject just the right number of X returns so as to correctly adjust the return distribution to that of Y . Here the routine $X()$ refers to the existing algorithm that returns nonnegative integers according to the X distribution. We also require that the p_n be nonzero.

```

while (true) {
    n = X();
    x = rand();
    if (x < q_n/(K * p_n))
        return n; }

```

Statistical Inference

Suppose we have several random variables X, Y, \dots of interest. For example, X might be the systolic blood pressure of a person who has taken a certain drug, while Y is the blood pressure of an individual who has not taken it. In this case, X and Y are defined on different probability spaces. Each probability space is a collection of persons who either have or have not used the drug in question. X and Y then have distributions in a certain range, say $[50, 250]$, but it is not feasible to measure X or Y at each outcome (person) to determine the detailed distributions. Consequently, we resort to samples. That is, we observe X for

various outcomes by measuring blood pressure for a subset of the X population. We call the observations X_1, X_2, \dots, X_n . We follow a similar procedure for Y if we are interested in comparing the two distributions. Here, we concentrate on samples from a single distribution.

A sample from a random variable X is actually another random variable. Of course, after taking the sample, we observe that it is a specific number, which hardly seems to merit the status of a random variable. However, we can envision that our choice is just one of many parallel observations that deliver a range of results. We can then speak of events such as $P(X_1 \leq t)$ as they relate to the disparate values obtained across the many parallel experiments as they make their first observations. We refer to the distribution of X as the *population* distribution and to that of X_n as the n^{th} *sample* distribution. Of course, $P(X_n \leq t) = P(X \leq t)$ for all n and t , but the term sample typically carries the implicit understanding that the various X_n are independent. That is, $P(X_1 \leq t_1, \dots, X_n \leq t_n) = \prod_{i=1}^n P(X \leq t_i)$. In this case, we say that the sample is a *random* sample.

With a random sample, the X_n are independent, identically distributed random variables. Indeed, each has the same distribution as the underlying population X . In practice, this property is assured by taking precautions to avoid any selection bias during the sampling. In the blood pressure application, for example, we attempt to choose persons in a manner that gives every individual the same chance of being observed.

Armed with a random sample, we now attempt to infer features of the unknown distribution for the population X . Ideally, we want the cumulative distribution of $F_X(t)$, which announces the fraction of the population with blood pressures less than or equal to t . Less complete, but still valuable, information lies with certain summary features, such as the expected value and variance of X .

A *statistic* is simply a function of a sample. Given the sample X_1, X_2, \dots, X_n , the new random variables

$$\begin{aligned}\bar{X} &= \frac{1}{n} \sum_{k=1}^n X_k \\ S^2 &= \frac{1}{n-1} \sum_{k=1}^n (X_k - \bar{X})^2\end{aligned}$$

are statistics known as the *sample mean* and *sample variance* respectively. If the population has $E[X] = \mu$ and $\text{Var}[X] = \sigma^2$, then $E[\bar{X}] = \mu$ and $E[S^2] = \sigma^2$. The expected value and variance are called *parameters* of the population, and a central problem in statistical inference is to estimate such unknown parameters through calculations on samples. At any point we can declare a particular statistic to be an *estimator* of some parameter. Typically we only do so when the value realized through samples is indeed an accurate estimate.

Suppose θ is some parameter of a population distribution X . We say that a statistic Y is an *unbiased* estimator of θ if $E[Y] = \theta$. We then have that the

sample mean and sample variance are unbiased estimators of the population mean and variance. The quantity

$$\hat{S}^2 = \frac{1}{n} \sum_{k=1}^n (X_k - \bar{X})^2$$

is also called the sample variance, but it is a *biased* estimator of the population variance σ^2 . If context is not clear, we need to refer to the biased or unbiased sample variance. In particular $E[\hat{S}^2] = \sigma^2(1 - 1/n)$, which introduces a bias of $b = -\sigma^2/n$. Evidently, the bias decays to zero with increasing sample size n . A sequence of biased estimators with this property is termed *asymptotically unbiased*.

A statistic can be a vector-valued quantity. Consequently, the entire sample (X_1, X_2, \dots, X_n) is a statistic. For any given t , we can compute the fraction of the sample values that are less than or equal to t . For a given set of t values, these computation produce a *sample distribution function*:

$$F_n(t) = \frac{\#\{k : X_k \leq t\}}{n}.$$

Here we use $\#\{\dots\}$ to denote the size of a set. For each t , the Glivenko-Cantelli Theorem states that the $F_n(t)$ constitute an asymptotically unbiased sequence of estimators for $F(t) = P(X \leq t)$.

Suppose X_1, X_2, \dots, X_n is a random sample of the population random variable X , which has $E[X] = \mu$ and $\text{Var}[X] = \sigma^2 < \infty$. The Central Limit Theorem gives the limiting distribution for $\sqrt{n}(\bar{X} - \mu)/\sigma$ as the standard Gaussian $Z_{0,1}$. Let us assume (unrealistically) for the moment that we know σ^2 . Then, we can announce \bar{X} as our estimate of μ , and we can provide some credibility for this estimate in the form of a *confidence interval*. Suppose we want a 90% confidence interval. From tables for the standard Gaussian, we discover that $P(|Z_{0,1}| \leq 1.645) = 0.9$. For large n , we have

$$\begin{aligned} 0.9 &= P(|Z_{0,1}| \leq 1.645) \approx P\left(\left|\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma}\right| \leq 1.645\right) \\ &= P\left(|\bar{X} - \mu| \leq \frac{1.645\sigma}{\sqrt{n}}\right). \end{aligned}$$

If we let $\delta = 1.645\sigma/\sqrt{n}$, we can assert that, for large n , there is a 90% chance that the estimate \bar{X} will lie within δ of the population parameter μ . We can further manipulate the equation above to obtain $P(\bar{X} - \delta \leq \mu \leq \bar{X} + \delta) \approx 0.9$. The specific interval obtained by substituting the observed value of \bar{X} into the generic form $[\bar{X} - \delta, \bar{X} + \delta]$ is known as the (90%) confidence interval. It must be properly interpreted. The parameter μ is an unknown constant, not a random variable. Consequently, either μ lies in the specified confidence interval or it does not. The random variable is the interval itself, which changes endpoints when new values of \bar{X} are observed. The width of the interval remains constant

at δ . The proper interpretation is that 90% of these nondeterministic intervals will bracket the parameter μ .

Under more realistic conditions, neither the mean μ nor the variance σ^2 of the population is known. In this case, we can make further progress if we assume that the individual X_i samples are normal random variables. There are various devices, such as composing each X_i as a sum of a subset of the samples, that render this assumption more viable. In any case, under this constraint, we can show that $n(\bar{X} - \mu)^2/\sigma^2$ and $(n - 1)S^2/\sigma^2$ are independent random variables with known distributions. These random variables have chi-squared distributions.

A *chi-squared* random variable with m degrees of freedom is the sum of the squares of m independent standard normal random variables. It is actually a special case of the gamma distributions discussed previously; it occurs when the parameters are $\gamma = m$ and $\lambda = 1/2$. If Y_1 is chi-squared with m_1 degrees of freedom and Y_2 is chi-squared with m_2 degrees of freedom, then the ratio $m_2 Y_1 / (m_1 Y_2)$ has an F distribution with (m_1, m_2) degree of freedom. A symmetric random variable is said to follow a *t* distribution with m_2 degrees of freedom if its square has an F distribution with $(1, m_2)$ degrees of freedom. For a given random variable R and a given value p in the range $(0, 1)$, the point r_p for which $P(R \leq r_p) = p$ is called the p^{th} percentile of the random variable. Percentiles for *F* and *t* distributions are available in tables.

Returning to our sample X_1, X_2, \dots, X_n , we find that under the normal inference constraint, the two statistics mentioned above have independent chi-squared distributions with 1 and $n - 1$ degrees of freedom respectively. Therefore the quantity $\sqrt{n}|\bar{X} - \mu|/\sqrt{S^2}$ has a *t* distribution with $n - 1$ degrees of freedom. Given a confidence level, say 90%, we consult a table of percentiles for the *t* distribution with $n - 1$ degrees of freedom. We obtain a symmetric interval $[-r, r]$ such that

$$0.9 = P\left(\frac{\sqrt{n}|\bar{X} - \mu|}{\sqrt{S^2}} \leq r\right) = P\left(|\bar{X} - \mu| \leq \frac{r\sqrt{S^2}}{\sqrt{n}}\right).$$

Letting $\delta = r\sqrt{S^2}/\sqrt{n}$, we obtain the 90% confidence interval $[\bar{X} - \delta, \bar{X} + \delta]$ for our estimate \bar{X} of the population parameter μ . The interpretation of this interval remains as discussed above.

This discussion above is an exceedingly abbreviated introduction to a vast literature on statistical inference. The references below provide a starting point for further study.

Reading List

1. Fristedt, Bert; Gray, Lawrence. A Modern Approach to Probability Theory, Birkhauser, 1997.
2. Gut, Allan. Probability: A Graduate Course, Springer, 2006.

3. Hacking, Ian. *The Emergence of Probability*, Cambridge University Press, 1975.
4. Hacking, Ian. *The Taming of Chance*, Cambridge University Press, 1990.
5. Johnson, James L. *Probability and Statistics for Computer Science*, Wiley, 2003.
6. Knuth, Donald E. *The Art of Computer Programming, Vol. 2 (Third Edition)*, Addison-Wesley, 1998.
7. Ore, Oystein. *Cardano, the Gambling Scholar*, Princeton, 1953.
8. Ore, Oystein, *Pascal and the Invention of Probability Theory*, *American Mathematical Monthly* 67, 1960.
9. Pickover, Clifford A. *Computers and the Imagination*, St. Martin's Press, 1991.
10. Ross, Sheldon M. *Probability Models for Computer Science*, Academic Press, 2002.
11. Royden, Halsley. *Real Analysis (Third Edition)*, Prentice Hall, 1988.