A Probability and Statistics Refresher

Peter J. Haas

Probability theory provides a mathematical framework for modelling real-world situations characterized by "randomness," "uncertainty," or "unpredictability." The theory also provides a toolbox of techniques for computing probabilities of interest, as well as related quantities such as expected values. As with any mathematical tool, the usefulness of a probability model depends on how well the model mirrors the real-world situation of interest. The field of statistics is largely concerned with matching probabilistic models to data. A statistical analysis starts with a set of data and assumes that the data is generated according to a probability model. The data is then used to make inferences about this probability model. The goal can be to fit "the best" probability model to the data, to estimate the values of certain parameters of an assumed model, or to test hypotheses about the model. The results of the inference step can be used for purposes of prediction, analysis, and decisionmaking.

In the context of simulation, we often use statistical techniques to develop our simulation model of the system under study; a simulation model is essentially a complicated probability model. For such a model, the probabilities and expected values that are needed for prediction and decisionmaking typically cannot be computed analytically or numerically. We therefore simulate the model, i.e., use the model to generate data, and then once again apply statistical techniques to infer the model properties of interest based on the output of the simulation.

The following sections summarize some basic topics in probability and statistics. Our emphasis is on material that we will use extensively in the course. Our discussion glosses over some of the technical fine points—see Section 11 for some pointers to more careful discussions.

1 Probabilistic Experiments and Events

1.1 Probabilistic Experiments

We start with a set Ω, called the *sample space*, that represents the possible elementary outcomes of a probabilistic experiment. A classic example of a simple probabilistic experiment is the rolling of a pair of dice, one black and one white. Record the outcome of this experiment by a pair (n, m), where n (resp., m) is the number of spots showing on the black (resp., white) die. Then Ω is the set that contains the 36 possible elementary outcomes:

$$\Omega = \{\,(1,1),(1,2),\ldots,(1,6),(2,1),(2,2),\ldots,(2,6),\ldots,(6,1),(6,2),\ldots,(6,6)\,\}\,.$$

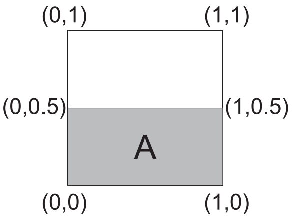

Figure 1: Definition of event for dartboard example.

As a more complicated example, consider an experiment in which we throw a dart at a square dartboard. (We are assuming here that our throw is so erratic that it can truly be considered unpredictable.) One way to describe the outcome of this experiment is to give the (x, y) coordinates of the dart's position, where we locate the origin $(0, 0)$ at the lower left corner. Note that Ω is an uncountably infinite set. As can be inferred from these examples, there is often some leeway in how Ω is defined when setting up a probability model.

1.2 Events

A subset $A \subseteq \Omega$ is called an *event*—if the outcome of the experiment is an element of A, then we say that "event A has occurred." For example, consider the two-dice example, together with the event

$$A = \{\, (1,6), (2,5), (3,4), (4,3), (5,2), (6,1) \,\}\,.$$

In words, A is the event in which "the sum of the spots showing on the two dice equals 7." If, for example, the outcome of our dice experiment is $(3, 4)$, then we say that event A has occurred since $(3, 4) \in A$. For our dart experiment, assuming that we measure coordinates in feet and the dartboard is $1' \times 1'$, we say that the event "the dart hits the lower half of the dartboard" occurs if the outcome is an element of the set

$$A = \{\, (x, y) \colon 0 \le x \le 1 \text{ and } 0 \le y \le 0.5 \,\}\,. \tag{1.1}$$

This event is illustrated in Figure 1. In general, the event Ω is the (certain) event that "an outcome occurs" when we run the probabilistic experiment. At the other extreme, the "empty" event $\varnothing$ contains no elements, and corresponds to the (impossible) situation in which no outcome occurs.

There is a close connection between set-theoretic notions and the algebra of events. The complimentary event $A^c = \Omega - A$ is the event in which A does not occur. The event $A \cap B$ is the event in which both A and B simultaneously occur, and the event $A \cup B$ is the event in which A occurs, or B occurs, or both. The event $B - A = B \cap A^c$ is the event in which B occurs and A does not occur. Events A and B are *disjoint* if they have no elements in common: $A \cap B = \varnothing$. Intuitively, disjoint events can never occur simultaneously. For an experiment in which we roll one die and record the number of spots, so that $\Omega = \{\, 1, 2, 3, 4, 5, 6 \,\}$, the events $A = \{\, 1, 3, 5 \,\}$ = "the die comes up odd" and $B = A^c = \{\, 2, 4, 6 \,\}$ = "the die comes up even" are disjoint. If $A \subseteq B$, then the occurrence of A implies the occurrence of B.

2 Probability Measures and Probability Spaces

2.1 Probability Measures

A *probability measure* P assigns to each[1] event a number between 0 and 1. The number $P(A)$ is interpreted as the likelihood that event A will occur when we run the probabilistic experiment: $P(A) = 0$ means that A will not occur, whereas $P(A) = 1$ means that A is certain to occur. For a given specification of P to be logically consistent, it must satisfy three basic ground rules: $P(\varnothing) = 0$, $P(\Omega) = 1$, and

$$P\Bigl(\bigcup_n A_n\Bigr) = \sum_n P(A_n)$$

whenever $A_1, A_2, \ldots$ form a finite or countably infinite collection of disjoint events. The first two requirements are that "with certainty, some outcome occurs" and "it is impossible that no outcome occurs" when we run the probabilistic experiment. The third condition guarantees, for example, that if the probability that a dart lands in the bottom third of the dartboard is 0.3 and the probability that the dart lands in the middle third of the dartboard is 0.4, then the probability that the dart lands in the bottom two-thirds of the dartboard is $0.3 + 0.4 = 0.7$.

For our two-dice example, one way of assigning probabilities is to set $P(A) = 1/36$ for each "singleton" event of the form $A_{n,m} = \{\,(n,m)\,\}$. In words, $A =$ "n spots are showing on the black die and m spots are showing on the white die." Observe that the singleton events are mutually disjoint and that any arbitrary event can be written as the union of (disjoint) singleton events. It follows that the probability of any event A is uniquely determined. For example, if $A = \{\,(1,2),(2,1)\,\} =$ "the sum of the spots showing on the two dice equals 3" then, from the ground rules, $P(A) = P(A_{1,2} \cup A_{2,1}) = P(A_{1,2}) + P(A_{2,1}) = 2/36$. This assignment of probabilities corresponds to a probability model in which each die is "fair" and the two dice behave independently (see Section 2.4).

For our dartboard example, we can model the situation in which the dart is equally likely to land anywhere on the board by defining $P(A) = \iint_A dx\,dy$ for each event A; an event A corresponds to a specified region on the dartboard and $P(A)$ is simply the area of the region. For the event A defined in (1.1), we have

$$P(A) = \iint_A dx\,dy = \int_0^{0.5}\int_0^1 dx\,dy = 1/2.$$

2.2 Basic Properties of Probability Measures

We now list some elementary properties of probability measures that follow from the three ground rules. All sums, unions, and intersections are taken over a finite or countably infinite collection of events.

(i) $0 \le P(A) \le 1$.

[1] Actually, when Ω is uncountably infinite, it turns out that there are some very weird events to which we can't assign probabilities in a consistent manner. You will never encounter such weird events in this course, however.

(ii) $P(A^c) = 1 - P(A)$.

(iii) $P(A) \leq P(B)$ whenever $A \subseteq B$.

(iv) (Boole's inequality) $P(\bigcup_n A_n) \leq \sum_n P(A_n)$.

(v) (Bonferroni's inequality) $P(\bigcap_n A_n) \geq 1 - \sum_n P(A_n^c)$.

Observe that Boole's inequality holds for *any* finite or countably infinite collection of events—the inequality becomes an equality if the events are mutually disjoint.

2.3 Probability Spaces

We refer to (Ω, P), a sample space together with a probability measure on events (i.e., on subsets of Ω) as a *probability space*. A probability space is the mathematical formalization of a probabilistic experiment.[2] Some useful advice: whenever anybody starts talking to you about probabilities, always make sure that you can clearly identify the underlying probabilistic experiment and probability space; if you can't, then the problem is most likely ill-defined.

2.4 Independent Events

Events A and B are *independent* if $P(A \cap B) = P(A)P(B)$. In our experiment with two dice, suppose that each elementary outcome is equally likely. Then the events $A =$ "the black die comes up even" and $B =$ "the white die comes up even" are independent, since $A \cap B$ comprises 9 outcomes and each of A and B comprises 18 outcomes, so that $P(A) = P(B) = 18/36 = 1/2$ and $P(A \cap B) = 9/36 = 1/4$.

Events $A_1, A_2, \ldots, A_n$ are *mutually independent* if

$$P(A_{n_1} \cap A_{n_2} \cap \cdots \cap A_{n_k}) = P(A_{n_1})P(A_{n_2}) \cdots P(A_{n_k})$$

for $2 \leq k \leq n$ and $1 \leq n_1 < n_2 < \cdots < n_k \leq n$. The events in a countably infinite collection are said to be mutually independent if the events in every finite subcollection are mutually independent.

3 Conditional Probability

3.1 Definition of Conditional Probability

Conditional probability formalizes the notion of "having information" about the outcome of a probabilistic experiment. Consider the two-dice example with equally likely outcomes. Let $A =$ "the black die has an even number of spots" and $B =$ "the sum of the spots showing on the two dice equals 4." In the absence of any information about the outcome of the experiment, an observer

[2] As mentioned previously, when Ω is uncountably infinite there are in general certain weird sets $A \subset \Omega$ on which probabilities cannot be defined. Therefore, advanced texts on probability define a probability space as a triple $(\Omega, P, \mathcal{F})$, where $\mathcal{F}$ is the "σ-field" of subsets on which P is defined.

would estimate the (unconditional) probability of event A as 0.5, since 18 of the 36 possible outcomes are elements of A. Given the additional information that event B has occurred, the observer knows that the outcome is one of $(1,3)$, $(2,2)$ or $(3,1)$. Of these, only the outcome $(2,2)$ corresponds to the occurrence of event A. Since the outcomes $(1,3)$, $(2,2)$, and $(3,1)$ are equally likely, the observer would estimate $P(A \mid B)$, the conditional probability that A has occurred given that B has occurred, as 1/3. In general, for events A and B such that $P(B) > 0$, we define

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}. \tag{3.1}$$

3.2 Independence, Law of Total Probability, Product Representation

Observe that our previous definition of independence for A and B is equivalent to requiring that $P(A \mid B) = P(A)$. I.e., knowing that B has occurred does not change our assessment of the probability that A has occurred.

Turning the definition in (3.1) around, we have the important relationship $P(A \cap B) = P(A \mid B)P(B)$. Observe that, by one of our basic ground rules, $P(A) = P(A \cap B) + P(A \cap B^c) = P(A \mid B)P(B) + P(A \mid B^c)P(B^c)$. An important generalization of this result is the *law of total probability*: if $B_1, B_2, \ldots, B_n$ are mutually disjoint events such that $B_1 \cup B_2 \cup \cdots \cup B_n = \Omega$, then

$$P(A) = P(A \mid B_1)P(B_1) + P(A \mid B_2)P(B_2) + \cdots + P(A \mid B_n)P(B_n). \tag{3.2}$$

Another important result that follows from the basic definition of conditional probability asserts that, for any events $A_1, A_2, \ldots, A_n$,

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1 \cap A_2) \cdots P(A_n \mid A_1 \cap A_2 \cap \cdots \cap A_{n-1}). \tag{3.3}$$

For example, suppose that we have n people in a room (where $2 \leq n \leq 365$), and each person is equally like to have been born on any of the 365 days of the year (ignore leap years and other anomalies). For $1 \leq i < n$, let A_i = "the birthday of person $(i+1)$ is different from persons 1 through i. Then $P(A_i \mid A_1 \cap \cdots \cap A_{i-1}) = (365 - i)/365$ for $1 \leq i < n$. It follows from (3.3) that the probability of the event B_n = "at least two people in the room share a birthday" is

$$P(B_n) = 1 - P(A_1 \cap \cdots \cap A_{n-1}) = 1 - \prod_{i=1}^{n-1} \left(1 - \frac{i}{365}\right).$$

This solution to the "birthday problem" is well known because it is somewhat counter-intuitive: for $n = 23$, the probability of at least one shared birthday exceeds 50%, and for $n = 50$ the probability exceeds 97%.

3.3 Bayes' Rule

Baye's Rule can be viewed as formalizing the process of learning from observations or data. Let $A_1, A_2, \ldots, A_n$ be a set of mutually exclusive and exhaustive events: $A_i \cap A_j = \varnothing$ for $i \neq j$ and

i	2	3	4	5	6	7	8	9	10	11	12
$p_X(i)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Table 1: PMF for random variable $X =$ "sum of spots" in two-dice example.

$\bigcup_{i=1}^{n} A_i = \Omega$. Then, starting with (3.1) and using the law of total probability, we can write

$$P(A_i \mid B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(A_i)P(B \mid A_i)}{\sum_{j=1}^{n} P(A_j)P(B \mid A_j)} \tag{3.4}$$

for $1 \le i \le n$, which is one form of Bayes' Rule. I.e., the "posterior" probability that A_i occurred, given that event B occurred, is the normalized product of the "prior" probability $P(A_i)$ that A_i occurred times the "likelihood" $P(B|A_i)$ of observing event B given that A_i occurred. Thus, Bayes' Rule tells us how to adjust our prior probability $P(A_i)$ to obtain the posterior probability $P(A_i \mid B)$, in light of our observation that B occurred.

4 Random Variables

4.1 Definition

Heuristically, a *random variable* X is a variable whose value is determined by the outcome of a probabilistic experiment. More formally, X is a real-valued function that is defined on the sample space Ω. For example, consider the two-dice experiment, and for each elementary outcome (n, m), set $X(n, m) = n + m$. Then the value of the random variable X is the sum of spots showing on the two dice. Observe that X can take on the possible values $2, 3, \dots, 11, 12$.

4.2 Indicator Random Variables

A particularly useful type of random variable is the *indicator* random variable for an event A. This random variable, denoted $I(A)$, equals 1 if event A occurs and equals 0 otherwise. Formally,

$$I(A)(\omega) = \begin{cases} 1 & \text{if } \omega \in A; \\ 0 & \text{if } \omega \notin A. \end{cases}$$

Observe that $P\{\, I(A) = 1 \,\} = P(A)$ and $P\{\, I(A) = 0 \,\} = 1 - P(A)$. Also observe that $I(A \cap B) = I(A)I(B)$ and $I(A \cup B) = I(A) + I(B) - I(A)I(B)$; if A and B are disjoint, then the latter expression simplifies to $I(A \cup B) = I(A) + I(B)$.

4.3 Distribution of a Random Variable

The *distribution* of a random variable X is the unique probability measure μ defined by

$$\mu(A) = P\{\, X \in A \,\} = P(\{\, \omega \in \Omega \colon X(\omega) \in A \,\})$$

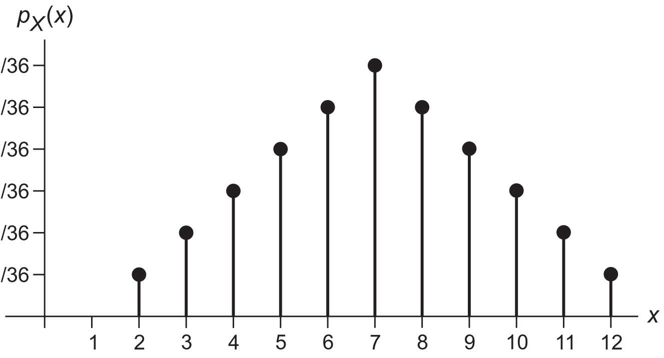

Figure 2: PMF for random variable X = "sum of spots" in two-dice example.

for each subset A of the real numbers. When X is *discrete*, i.e., takes values in a finite or countably infinite set $S = \{ a_1, a_2, \dots \}$, the distribution of X is usually described in terms of the *probability mass function* (PMF), sometimes denoted by p_X, where

$$p_X(a_i) = \mu(\{a_i\}) = P\{ X = a_i \}$$

for $a_i \in S$. Of course, a PMF p_X must satisfy $p_X(a_i) \geq 0$ for all i and

$$\sum_i p_X(a_i) = 1.$$

The set S is sometimes called the *support* of X. In the two-dice example with fair and independent dice, all of the elementary outcomes are equally likely, and we have, for example,

$$p_X(4) = P\{ X = 4 \} = P(A_4) = 3/36,$$

where

$$A_4 = \{ \omega \in \Omega \colon X(\omega) = 4 \} = \{ (1,3), (2,2), (3,1) \}.$$

The complete PMF for X is displayed in Table 1 and plotted in Figure 2.

The situation is usually more complicated for a *continuous* random variable X, i.e., a random variable taking values in an uncountably infinite set S such as an interval $[a, b]$ of the real line. In this case, we typically have $P\{ X = x \} = 0$ for any $x \in S$. The distribution of a continuous random variable X is often described in terms of the *probability density function* (PDF), sometimes denoted by f_X. Roughly speaking, for $x \in S$ and a small increment $\Delta x > 0$, we take the quantity $f_X(x)\Delta x$ as the approximate probability that X takes on a value in the interval $[x, x + \Delta x]$. More precisely, for a subset $A \subseteq S$, we have

$$\mu(A) = P\{ X \in A \} = \int_A f_X(x)\, dx.$$

In analogy with a PMF, a PDF must satisfy $f_X(x) \geq 0$ for all x and

$$\int_{-\infty}^{\infty} f_X(x)\, dx = 1.$$

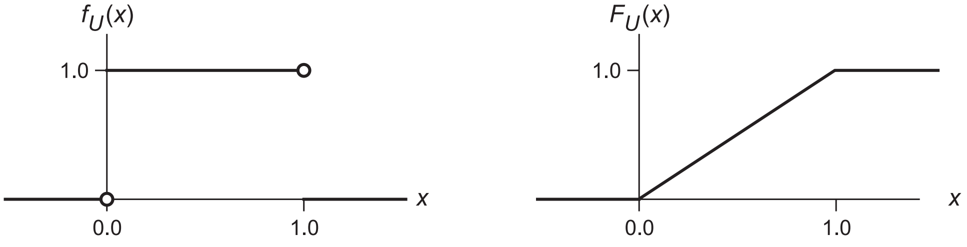

Figure 3: The PDF and CDF for a $U[0,1]$ random variable U.

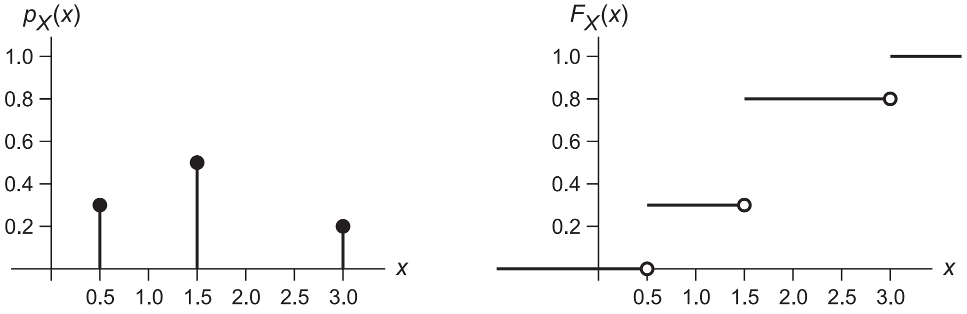

Figure 4: The PMF and CDF for a random variable X with $p_X(0.5) = 0.3$, $p_X(1.5) = 0.5$, and $p_X(3) = 0.2$.

As an example, consider a random variable U having the "uniform distribution on $[0,1]$," abbreviated $U[0,1]$. Here U is equally likely to take on any value between 0 and 1. Formally, we set

$$f_U(x) = \begin{cases} 0 & \text{if } x < 0; \\ 1 & \text{if } 0 \le x < 1; \\ 0 & \text{if } x \ge 1 \end{cases}$$

for all real values of x.[3] Then, e.g., for $A = [0.25, 0.75]$, we have

$$P\{\, U \in A \,\} = P\{\, 0.25 \le U \le 0.75 \,\} = \mu(A) = \int_A f_U(x)\, dx = \int_{0.25}^{0.75} 1 \; dx = x \,\Big|_{0.25}^{0.75} = 0.5.$$

For either a discrete or continuous random variable X satisfying the regularity condition[4] $P\{\, -\infty < X < \infty \,\} = 1$, the right-continuous function F_X defined by

$$F_X(x) = P\{\, X \le x \,\} = \mu\big((-\infty, x]\big)$$

for real-valued x is the *cumulative distribution function* (CDF) of X. The function F_X is nondecreasing, with $F_X(-\infty) = 0$ and $F_X(\infty) = 1$. For a continuous random variable X, the CDF is

[3] We have defined $f_U(x)$ at the points $x = 0$ and $x = 1$ so that f_U is right-continuous; this is a standard convention. Changing the definition at these two points has no real effect, since $P\{\, U = x \,\} = 0$ for any single point x.

[4] Such a random variable is often called *proper*. We restrict attention throughout to proper random variables.

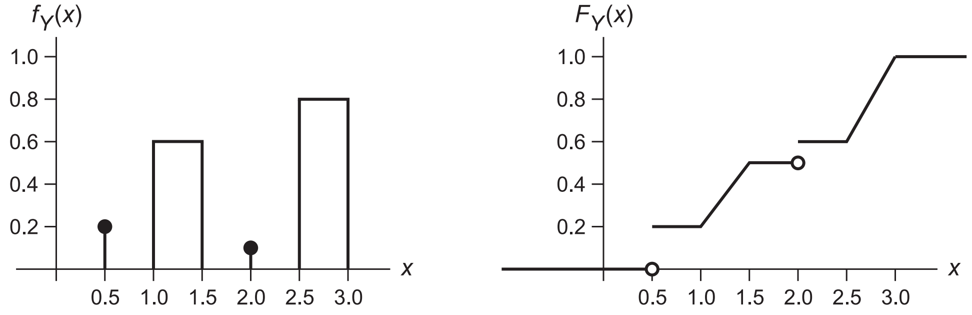

Figure 5: The PDF and CDF for the mixed random variable Y.

simply the indefinite integral of f_X:

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\, du.$$

For example, the CDF F_U for the $U[0,1]$ random variable U defined above is

$$F_U(x) = \begin{cases} 0 & \text{if } x < 0; \\ x & \text{if } 0 \le x \le 1; \\ 1 & \text{if } x > 1. \end{cases}$$

The PDF and CDF for a $U[0,1]$ random variable are plotted in Figure 3. On the other hand, the CDF F_X of a discrete random variable X is piecewise constant. The jumps occur at, and only at, the points in the support set S, and the magnitude of the jump at $x \in S$ equals $p_X(x)$; see Figure 4. A random variable can also have a mixed distribution, part discrete and part continuous. For example, Figure 5 shows the PDF and CDF for a mixed random variable Y such that with 20% probability Y equals 0.5, with 30% probability Y is uniformly distributed between 1.0 and 1.5, with 10% probability Y equals 2.0, and with 40% probability Y is uniformly distributed between 2.5 and 3.0. When displaying the PDF for a mixed random variable, we treat the PDF at the discrete points of support as a Dirac delta function or "impulse" function that places a specified "probability mass" at the point in question.

In general, it suffices to define the CDF F_X for a random variable X in order to completely specify its distribution. If X is discrete, then we can compute the PMF from the CDF using the relationship $p_X(x) = F_X(x) - F_X(x-)$, where $F_X(x-)$ is shorthand for $\lim_{\epsilon \to 0} F(x-\epsilon)$. It follows from the fact that F_X is piecewise constant that if X has support on the integers, then $p_X(i) = F_X(i) - F_X(i-1)$ for each integer i. If X is continuous, then we can compute the PDF from the CDF by differentiation. For mixed random variables, we can use a combination of these two techniques.

5 Multiple Random Variables

It is often the case that two or more random variables are defined on a given probability space (Ω, P). In our discussion, we will mostly focus on the case of two random variables X and Y, the generalization of our results to three or more random variables being obvious.

5.1 Joint, Marginal, and Conditional Distributions

We can describe the joint probability distribution of X and Y by means of the *joint* CDF $F_{X,Y}$, where $F_{X,Y}(x,y) = P\{\, X \le x, Y \le y \,\}$.[5] If X and Y are both discrete, we can also describe the joint distribution using the *joint* PMF $p_{X,Y}(x,y) = P\{\, X = x, Y = y \,\}$. If X and Y are both continuous, then (assuming differentiability) we can describe the joint distribution using the *joint* PDF defined by $f_{X,Y} = dF_{X,Y}/dx\,dy$. Given a joint distribution of X and Y, we might be interested in the *marginal distribution* of X alone. The *marginal* CDF of X is defined in terms of the joint CDF by

$$F_X(x) = P\{\, X \le x \,\} = P\{\, X \le x, Y < \infty \,\} = F_{X,Y}(x,\infty).$$

If X and Y are both discrete or both continuous, then the marginal PMF or marginal PDF of X can be computed from the corresponding joint distribution:

$$p_X(x) = P\{\, X = x \,\} = \sum_y p_{X,Y}(x,y) \qquad \text{or} \qquad f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy.$$

Of course, we can define quantities $F_Y(y)$, $p_Y(y)$, and $f_Y(y)$ in an analogous manner to obtain the marginal distribution of Y.

For discrete random variables X and Y, we can define the *conditional* PMF of X, given that $Y = y$, by adapting our general definition of conditional probability:

$$p_{X|Y}(x|y) = P\{\, X = x \mid Y = y \,\} = \frac{P\{\, X = x, Y = y \,\}}{P\{\, Y = y \,\}} = \frac{p_{X,Y}(x,y)}{p_Y(y)}.$$

If X and Y are both continuous, then a natural definition for a *conditional* PDF is

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}.$$

We conclude this section by giving a continuous analogue of Bayes' Rule (3.4):

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{\int_{-\infty}^{\infty} f_{Y|X}(y|x')f_X(x')\,dx'} \tag{5.1}$$

This rule shows how to update the prior density f_X of the random variable X to a posterior density $f_{X|Y}$ given the observation $Y = y$. The conditional density $f_{Y|X}$ is called the *likelihood* of Y, given the value of X. For example, if X has a Beta(α,β) distribution (see Sec. 7.5 below) and Y has a Binom(n,X) distribution (see Sec. 7.7), then

$$f_{X|Y}(x|y) = \frac{\binom{n}{y}x^{y+\alpha-1}(1-x)^{n-y+\beta-1}/B(\alpha,\beta)}{\int_0^1 \binom{n}{y}z^{y+\alpha-1}(1-z)^{n-y+\beta-1}\,dz/B(\alpha,\beta)} = \frac{x^{y+\alpha-1}(1-x)^{n-y+\beta-1}}{B(y+\alpha, n-y+\beta)}.$$

That is, $f_{X|Y}$ is the pdf of a Beta$(\alpha+Y, \beta+n-Y)$ distribution. Because the prior and posterior distributions belong to the same family, they are called *conjugate distributions*, and the beta distribution is called a *conjugate prior* for the binomial distribution.

[5] We use notation such as $P\{\, X \le x, Y \le y \,\}$ instead of the more cumbersome notation $P(\{X \le x\} \cap \{Y \le y\})$.

5.2 Independent Random Variables

The real-valued random variables $X_1, X_2, \ldots, X_n$ are *mutually independent* if

$$P\{\, X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n \,\} = P\{\, X_1 \in A_1 \,\}\, P\{\, X_2 \in A_2 \,\} \cdots P\{\, X_n \in A_n \,\}$$

for every collection of subsets $A_1, A_2, \ldots, A_n$ of the real line. To establish independence, it suffices to show that

$$P\{\, X_1 \le x_1, X_2 \le x_2, \ldots, X_n \le x_n \,\} = P\{\, X_1 \le x_1 \,\}\, P\{\, X_2 \le x_2 \,\} \cdots P\{\, X_n \le x_n \,\}$$

for every set of real numbers $x_1, x_2, \ldots, x_n$, i.e., it suffices to show that the joint CDF factors. If $X_1, X_2, \ldots, X_n$ are all discrete or all continuous, then it suffices to show that the joint PMF or joint PDF factors:

$$p_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = p_{X_1}(x_1) p_{X_2}(x_2) \cdots p_{X_n}(x_n)$$

or

$$f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_n}(x_n).$$

A countably infinite collection of random variables is said to be mutually independent if the random variables in each finite subcollection are independent. Observe that X and Y are independent if and only if the conditional PMF or PDF equals the marginal PMF or PDF for X.

6 Expectation

6.1 Definition

The *expected value* (aka the *mean*) of a random variable X, denoted $E[X]$, is one way of roughly indicating where the "center" of the distribution of X is located. It also has the interpretation of being the average value of X over many repetitions of the probabilistic experiment. The idea is to weight each value of X by the probability of its occurrence. If X is discrete, then we define

$$E[X] = \sum_{a_i \in S} a_i p_X(a_i),$$

and if X is continuous, then we define

$$E[X] = \int_{-\infty}^{\infty} x f_X(x)\, dx.$$

If X is a mixed random variable, then we compute the expected value by combining summation and integration. For example, consider the random variables U, X, and Y whose distributions are displayed in Figures 3 through 5. The expectations for these random variables are computed as

$$E[U] = \int_{-\infty}^{\infty} x f_U(x)\, dx = \int_0^1 x\ dx = \left. \frac{x^2}{2} \right|_0^1 = 1/2,$$

$$E\,[X] = \sum_{a_i \in S} a_i p_X(a_i) = 0.5\,p_X(0.5)+1.5\,p_X(1.5)+3.0\,p_X(3.0) = 0.5\,(0.3)+1.5\,(0.5)+3.0\,(0.2) = 1.5,$$

and

$$E\,[Y] = 0.5\,(0.2) + \int_{1.0}^{1.5} 0.6x\,dx + 2.0\,(0.1) + \int_{2.5}^{3.0} 0.8x\,dx = 1.775.$$

Observe that, for an indicator random variable $I(A)$, we have

$$E\,[I(A)] = 1 \cdot P\{\,I(A)=1\,\} + 0 \cdot P\{\,I(A)=0\,\} = P\{\,I(A)=1\,\} = P(A).$$

So that results about expectation can sometimes be used to obtain results about probability by judicious use of indicator functions.

The expected value of a random variable need not exist. A well known example of a distribution that is ill-behaved in this way is the *Cauchy distribution*, whose PDF is given by

$$f_X(x) = \frac{1}{\pi(1+x^2)}, \qquad -\infty < x < \infty,$$

and for which the expected value is undefined. A random variable having a well defined and finite expected value is said to be *integrable.*

We often are interested in the expected value of a function g of a random variable:

$$E\,[g(X)] = \sum_{a_i \in S} g(a_i) p_X(a_i) \quad \text{or} \quad \int_{-\infty}^{\infty} g(x) f_X(x)\,dx,$$

for X discrete or continuous, respectively. As with simple expectation, mixed random variables are handled using a combination of summation and integration. Sometimes computations of expectations can be simplified via a change of variable. E.g., if $y = h(x)$ for a monotone function h with inverse h_{inv}, then

$$\int_a^b g(x)\,dx = \int_{h(a)}^{h(b)} g\big(h_{\text{inv}}(y)\big) h'_{\text{inv}}(y)\,dy,$$

where h'_{inv} is the derivative of h_{inv}. For example, if $h(x) = x^3$, then $h_{\text{inv}}(y) = y^{1/3}$ and $h'_{\text{inv}}(y) = \frac{1}{3}y^{-2/3}$, so that, e.g.,

$$\int_0^a x^2\,dx = \int_0^{a^3} (y^{1/3})^2 \frac{1}{3} y^{-2/3}\,dy = \frac{1}{3}\int_0^{a^3} dy = a^3/3.$$

6.2 Basic Properties

We now state some basic properties of the expectation operator. Here and elsewhere, a property of a random variable is said to hold *almost surely* (a.s.) if it holds with probability 1.

If X and Y are random variables and a and b are real-valued constants, then

(i) X is integrable if and only if $E\,[|X|] < \infty$.

(ii) If $X = 0$ a.s., then $E\,[X] = 0$.

(iii) If X and Y are integrable with $X \le Y$ a.s., then $E\,[X] \le E\,[Y]$.

(iv) If X and Y are integrable, then $E[aX+bY] = aE[X] + bE[Y]$.

(v) $|E[X]| \le E[|X|]$.

(vi) If X and Y are independent, then $E[XY] = E[X]E[Y]$.

Note that, in general, $E[XY] = E[X]E[Y]$ does not imply that X and Y are independent.

6.3 Moments, Variance, Standard Deviation

The rth *moment* of a random variable X is $E[X^r]$ and the rth *central moment* is $E[(X-\mu)^r]$, where $\mu = E[X]$—the second central moment is called the *variance* of X and denoted $\mathrm{Var}[X]$. It is easy to show that

$$\mathrm{Var}[X] = E[X^2] - \mu^2.$$

The *standard deviation* of X is defined as the square root of the variance: $\mathrm{Std}[X] = \mathrm{Var}^{1/2}[X]$. For real numbers c and d, we have $\mathrm{Var}[cX+d] = c^2\,\mathrm{Var}[X]$ and $\mathrm{Std}[cX+d] = |c|\,\mathrm{Std}[X]$. The variance and standard deviation measure the degree to which the probability distribution of X is concentrated around its mean $\mu = E[X]$. When the variance or standard deviation equals 0, then $X = \mu$ with probability 1. The larger the variance or standard deviation, the greater the probability that X can take on a value far away from the mean.

We frequently talk about moments of distributions, rather than random variables. E.g., the mean μ and variance σ^2 of a random variable X having CDF F and PDF f are given by

$$\mu = \int_{-\infty}^{\infty} x f(x)\,dx \qquad \text{and} \qquad \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx.$$

We may refer to μ and σ^2 as the "mean and variance of X," the "mean and variance of f," or "mean and variance of F." Sometimes one may see the notation

$$\mu = \int_{-\infty}^{\infty} x\,dF \qquad \text{and} \qquad \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2\,dF$$

used for these quantities; this notation is mostly used to refer to situations in which X is a mixed random variable, so that F has some discontinuities. In this case, moments are computed by a combination of integration (over the intervals where F is differentiable) and summation (over the discontinuity points). E.g., for the mixed random variable Y in Figure 5, we have $E[Y] = (0.5)(0.2) + (2)(0.1) + \int_{1.0}^{1.5} 0.6y\,dy + \int_{2.5}^{3.0} 0.8y\,dy = 1.775$.

6.4 Identities and Inequalities

There are many identities and inequalities for moments of random variables—a very useful inequality for our purposes is *Hölder's inequality*: let X and Y be random variables, and let p and q be constants such that $1 < p < \infty$ and $1/p + 1/q = 1$. Then

$$E[|XY|] \le E^{1/p}[|X|^p]\,E^{1/q}[|Y|^q].$$

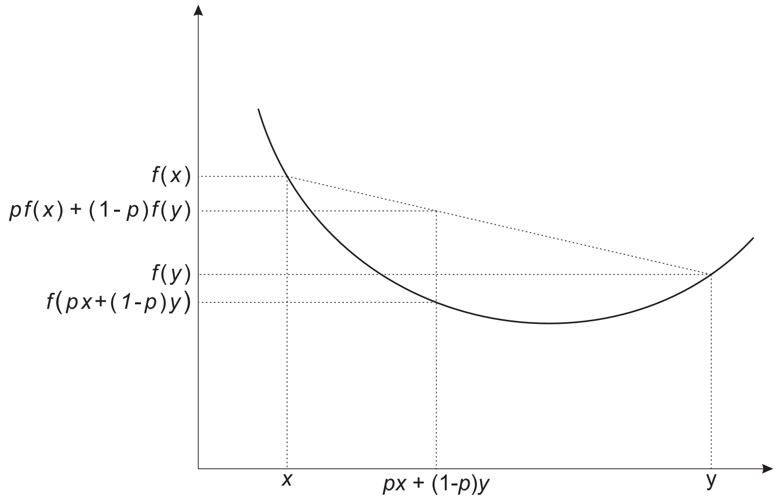

Figure 6: Definition of a convex function.

Take $p = q = 2$ to obtain the *Cauchy–Schwarz inequality*:

$$E[|XY|] \le E^{1/2}\left[X^2\right] E^{1/2}\left[Y^2\right].$$

In particular, $E^2[X] \le E\left[X^2\right]$—take $Y \equiv 1$ and use the fact that $E[X] \le E[|X|]$. Next, fix $0 < \alpha \le \beta$ and take $X = |Z|^\alpha$, $Y \equiv 1$, and $p = \beta/\alpha$ in Hölder's inequality to obtain

$$E^{1/\alpha}[|Z|^\alpha] \le E^{1/\beta}\left[|Z|^\beta\right],$$

which is *Lyapunov's inequality*. Observe that if a nonnegative random variable X has a finite rth moment for some $r > 0$, then Lyapunov's inequality implies that X has a finite qth moment for $q \in (0, r]$. We conclude our discussion of inequalities with *Jensen's* inequality. Recall that a function f is *convex* if for every real x, y, and p with $x < y$ and $p \in [0,1]$ we have $f\big(px+(1-p)y\big) \le pf(x)+(1-p)f(y)$; see Figure 6. If the foregoing inequality is reversed, then f is said to be *concave*. Jensen's inequality asserts that

$$f(E[X]) \le E[f(X)]$$

for any convex function f and random variable X. The inequality holds in the reverse direction for concave functions. An application of Jensen's inequality yields an alternative proof of the previous assertion that $E^2[X] \le E\left[X^2\right]$, since the function $f(x) = x^2$ is convex.

A useful representation of the rth moment ($r \ge 1$) of a nonnegative continuous random variable X is as follows:

$$E[X^r] = \int_0^\infty r x^{r-1} \bar{F}_X(x)\, dx,$$

where $\bar{F}_X = 1 - F_X$. In particular, taking $r = 1$, we find that

$$E[X] = \int_0^\infty \bar{F}_X(x)\, dx. \tag{6.1}$$

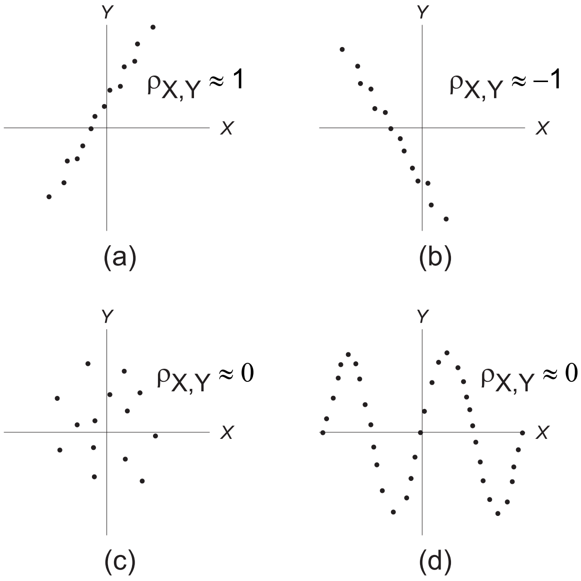

Figure 7: Samples of (X, Y) pairs from distributions with various correlation values.

A simple proof of this result when X is continuous is as follows:

$$\begin{aligned} E\left[X^r\right] = \int_0^\infty u^r f_X(u)\,du &= \int_0^\infty \left(\int_0^u r x^{r-1}\,dx \right) f_X(u)\,du \\ &= \int_0^\infty \int_x^\infty r x^{r-1} f_X(u)\,du\,dx = \int_0^\infty r x^{r-1} \bar{F}_X(x)\,dx. \end{aligned}$$

Here the third equality is obtained by changing the order of integration. A similar argument using sums instead of integrals shows that the result in (6.1) holds for discrete random variables also.

6.5 Covariance and Correlation

The *covariance* of two random variables X and Y with respective means $\mu_X = E[X]$ and $\mu_Y = E[Y]$ is defined as

$$\mathrm{Cov}\left[X, Y\right] = E\left[(X - \mu_X)(Y - \mu_Y)\right]$$

and measures the degree to which a linear relationship holds between X and Y. A normalized version of the covariance that is independent of the units in which X and Y are expressed, i.e., that is *scale invariant*, is the *correlation coefficient* $\rho_{X,Y}$, defined by

$$\rho_{X,Y} = \frac{\mathrm{Cov}\left[X, Y\right]}{\sigma_X \sigma_Y},$$

where σ_X^2 and σ_Y^2 are the variances of X and Y. The quantity ρ is more formally known as the “Pearson linear correlation coefficient.” The Cauchy-Schwartz inequality implies that $-1 \leq \rho_{X,Y} \leq 1$. A value of $\rho_{X,Y}$ close to 1 (resp., -1) indicates a strong positive (resp., negative) linear relationship between X and Y, whereas a value of $\rho_{X,Y}$ close to 0 indicates the absence of a discernable linear relationship; see Figure 7, which plots some samples from the joint distribution of (X,Y) under several different scenarios. Note that $\rho_{X,Y}$ is close to 0 in Figure 7(d), even though there is a strong relationship between X and Y—the reason is that the relationship is nonlinear. It follows from basic property (vi) of expectation that if X and Y are independent, then $\mathrm{Cov}\,[X,Y]=0$; the converse assertion is not true in general.

Some simple algebra shows that

$$\mathrm{Var}\Biggl[\sum_{i=1}^{n} X_i\Biggr] = \sum_{i=1}^{n}\mathrm{Var}\,[X_i] + \sum_{i\neq j}\mathrm{Cov}\,[X_i,X_j]\,,$$

so that

$$\mathrm{Var}\Biggl[\sum_{i=1}^{n} X_i\Biggr] = \sum_{i=1}^{n}\mathrm{Var}\,[X_i]$$

if $X_1, X_2, \ldots, X_n$ are mutually independent.

7 Some Important Probability Distributions

In this section we discuss some probability distributions that play a particularly important role in our study of simulation; see Tables 6.3 and 6.4 in the textbook for more details. We start with several continuous distributions.

7.1 Uniform Distribution

We have already discussed the $U[0,1]$ distribution. In general, the distribution of a random variable that is equally likely to take on a value in a subinterval $[a,b]$ of the real line is called the *uniform distribution on* $[a,b]$, abbreviated $U[a,b]$. The PDF and CDF for a $U[a,b]$ random variable U are given by

$$f_U(x)=\begin{cases}0 & \text{if } x<a;\\ 1/(b-a) & \text{if } a\le x<b;\\ 0 & \text{if } x\ge b\end{cases} \qquad\text{and}\qquad F_U(x)=\begin{cases}0 & \text{if } x<a;\\ (x-a)/(b-a) & \text{if } a\le x\le b;\\ 1 & \text{if } x>b.\end{cases} \tag{7.1}$$

If U is a $U[0,1]$ random variable, then $V=a+(b-a)U$ is a $U[a,b]$ random variable. The easiest way to prove this assertion (and many others like it) is to work with the CDF of each random variable:

$$F_V(x)=P\{\,V\le x\,\}=P\{\,a+(b-a)U\le x\,\}=P\{\,U\le (x-a)/(b-a)\,\}=F_U\bigl((x-a)/(b-a)\bigr).$$

By inspection, $F_U\bigl((x-a)/(b-a)\bigr)$ coincides with the function F_U in (7.1). The PDF f_U is then obtained from F_U by differentiation. The mean and variance of the $U[a,b]$ distribution are $(a+b)/2$ and $(b-a)^2/12$.

7.2 Exponential Distribution

The PDF and CDF of an *exponential* distribution with intensity λ, abbreviated $\mathrm{Exp}(\lambda)$, are

$$f(x) = \begin{cases} 0 & \text{if } x < 0; \\ \lambda e^{-\lambda x} & \text{if } x \geq 0 \end{cases} \qquad \text{and} \qquad F(x) = \begin{cases} 0 & \text{if } x < 0; \\ 1 - e^{-\lambda x} & \text{if } x \geq 0 \end{cases}$$

This distribution is also called the "negative exponential distribution." The mean and variance are given by $1/\lambda$ and $1/\lambda^2$. A key feature of this distribution is the "memoryless" property: if X is an $\mathrm{Exp}(\lambda)$ random variable and u, v are nonnegative constants, then

$$P\{\, X > u + v \mid X > u \,\} = \frac{P\{\, X > u + v \,\}}{P\{\, X > u \,\}} = \frac{\bar{F}(u+v)}{\bar{F}(u)} = \frac{e^{-(u+v)}}{e^{-u}} = e^{-v} = P\{\, X > v \,\},$$

where, as before $\bar{F} = 1 - F$. E.g., suppose that X represents the waiting time until a specified event occurs. Then the probability that you have to wait at least v more time units is independent of the amount of time u that you have already waited; at time $t = u$, the "past" has been forgotten with respect to estimating the probability distribution of the remaining time until the event occurs.

Another important property of the exponential distribution concerns the distribution of $Z = \min(X, Y)$, where X is $\mathrm{Exp}(\lambda_1)$, Y is $\mathrm{Exp}(\lambda_2)$, and X and Y are independent. The distribution of Z can be computed as follows:

$$\begin{aligned} P\{\, Z > z \,\} &= P\{\, \min(X,Y) > z \,\} = P\{\, X > z, Y > z \,\} = P\{\, X > z \,\}\, P\{\, Y > z \,\} \\ &= e^{-\lambda_1 z} e^{-\lambda_2 z} = e^{-(\lambda_1 + \lambda_2) z}. \end{aligned}$$

That is, the distribution of Z is also exponential, but with parameter $\lambda = \lambda_1 + \lambda_2$. The probability that $X < Y$ can be computed using indicator functions together with a continuous version of the law of total probability:

$$\begin{aligned} P\{\, X < Y \,\} &= \int_0^\infty P\{\, X < Y \mid Y = y \,\}\, f_Y(y)\, dy = \int_0^\infty P\{\, X < y \mid Y = y \,\}\, f_Y(y)\, dy \\ &= \int_0^\infty P\{\, X < y \,\}\, f_Y(y)\, dy = \int_0^\infty F_X(y) f_Y(y)\, dy = \int_0^\infty (1 - e^{-\lambda_1 y}) \lambda_2 e^{-\lambda_2 y}\, dy \\ &= \int_0^\infty \lambda_2 e^{-\lambda_2 y}\, dy - \lambda_2 \int_0^\infty e^{-(\lambda_1 + \lambda_2) y}\, dy = 1 - \lambda_2/(\lambda_1 + \lambda_2) = \lambda_1/(\lambda_1 + \lambda_2). \end{aligned}$$

Here the third equality follows from the independence of X and Y, and the fourth equality follows from the fact that $P\{\, X = y \,\} = 0$ for any fixed y. An easy inductive argument generalizes these results to an arbitrary number of independent exponential random variables.

7.3 Normal and Related Distributions

The PDF ϕ of a *normal* (aka Gaussian) distribution with mean μ and variance σ^2 is

$$\phi(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{(x-\mu)^2/(2\sigma^2)}$$

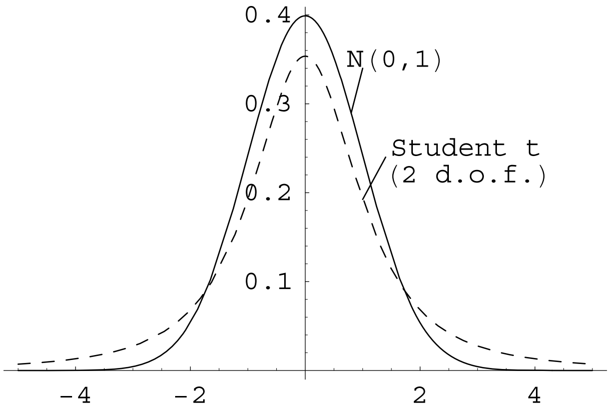

Figure 8: PDF for standard normal distribution snd Student t distribution with 2 degrees of freedom.

for $-\infty < x < \infty$. We abbreviate this distribution as $N(\mu, \sigma^2)$. There is no closed form expression for the CDF $\Phi(x; \mu, \sigma^2)$, and values of Φ must be computed numerically or looked up in tables. Two normal random variables X and Y are independent if and only if $\mathrm{Cov}\,[X, Y] = 0$. Let X be $N(\mu_X, \sigma_X^2)$ and Y be $N(\mu_Y, \sigma_Y^2)$ independent of X, and let c and d be real constants. Then $aX + b$ is $N(a\mu_X + b, a^2\sigma_X^2)$ and $X + Y$ is $N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$. In particular, if X is $N(\mu_X, \sigma_X^2)$, then $(X - \mu_X)/\sigma_X$ is $N(0, 1)$, that is, X has the *standard* normal distribution in which $\mu = 0$ and $\sigma^2 = 1$. We often denote the standard normal PDF and CDF simply as $\phi(x)$ and $\Phi(x)$.

The distribution of $Y = e^X$, where X is a normal random variable, is called the *lognormal* distribution; i.e., the logarithm of Y has a normal distribution. The distribution of $X_1^2 + X_2^2 + \cdots + X_k^2$, where $k \geq 1$ and $X_1, X_2, \ldots, X_k$ are independent standard normal random variables, is called the *chi-square distribution with k degrees of freedom*, abbreviated as χ_k^2. If X is $N(0, 1)$ and Y is χ_k^2, then the random variable $T = X/\sqrt{Y/k}$ has the *Student t distribution with k degrees of freedom*, abbreviated t_k. This distribution looks similar to the $N(0, 1)$ distribution but with fatter tails, i.e., higher variance—see Figure 8. As k increases, the Student t distribution becomes identical to the normal distribution.

7.4 Gamma Distribution

The pdf of the Gamma distribution with shape parameter α and rate parameter λ, abbreviated Gamma(α, λ), is given by

$$f(x; \alpha, \lambda) = \lambda e^{-\lambda x}(\lambda x)^{\alpha - 1}/\Gamma(\alpha)$$

for $x \geq 0$ and $f(x; \alpha, \lambda) = 0$ for $x < 0$. Here $\Gamma(\alpha)$ is the gamma function, defined by $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}\,dx$ and satisfying $\Gamma(\alpha) = (\alpha - 1)!$ whenever α is a postive integer. The mean and variance of the distribution are α/λ and α/λ^2.

7.5 Beta Distribution

The pdf of the Beta distribution with parameters α and β, abbreviated Beta(α, β), is given by

$$f(x; \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}$$

if $x \in [0, 1]$, and $f(x; \alpha, \beta) = 0$ otherwise. Here $B(\alpha, \beta)$ is the beta function, defined by $B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,dx$ and satisfying $B(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$. The mean and variance of the distribution are $\alpha/(\alpha+\beta)$ and $\alpha\beta(\alpha+\beta)^{-2}(\alpha+\beta+1)^{-1}$.

7.6 Discrete Uniform Distribution

We now discuss some important discrete distributions. A *discrete* uniform random variable X with range $[n, m]$, abbreviated $DU[n, m]$, is equally likely to take on the values $n, n+1, \ldots, m$, i.e., $p_X(k) = 1/(m-n+1)$ for $k = n, n+1, \ldots, m$. The mean and variance are given by $(n+m)/2$ and $[(m-n+1)^2 - 1]/12$. If U is a continuous $U[0, 1]$ random variable, then $V = \lfloor n + (m-n+1)U \rfloor$ is $DU[n, m]$, where $\lfloor x \rfloor$ is the largest integer less than or equal to x. Here's a proof: fix $k \in \{\, n, n+1, \ldots, m \,\}$ and observe that

$$\begin{aligned} P\{\, V = k \,\} &= P\{\, \lfloor n + (m-n+1)U \rfloor = k \,\} = P\{\, k \le n + (m-n+1)U < k+1 \,\} \\ &= P\left\{ \frac{k-n}{m-n+1} \le U < \frac{k-n+1}{m-n+1} \right\} \\ &= \frac{1}{m-n+1}. \end{aligned}$$

7.7 Bernoulli and Binomial Distributions

The *Bernoulli distribution with parameter* p, abbreviated Bern(p), has PMF given by $p(1) = 1 - p(0) = p$. That is, a Bern(p) random variable X equals 1 with probability p and 0 with probability $1-p$. Often, X is interpreted as an indicator variable for a "Bernoulli trial with success probability p." Here $X = 1$ if the trial is a "success" and $X = 0$ if the trial is a "failure." The mean and variance are given by p and $p(1-p)$.

The number of successes S_n in n independent Bernoulli trials, each with success probability p, can be represented as $S_n = X_1 + X_2 + \cdots + X_n$, where $X_1, X_2, \cdots, X_n$ are independent Bern(p) random variables. The random variable S_n has the *binomial distribution* with parameters n and p, abbreviated Binom(n, p). The PMF is given by

$$p(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

for $k = 0, 1, \ldots, n$. The mean and variance are given by np and $np(1-p)$.

7.8 Geometric Distribution

The *geometric distribution* with parameter p, abbreviated Geom(p), has support on the nonnegative integers and a PMF and CDF given by

$$p(k) = p(1-p)^k \qquad \text{and } F(k) = 1 - (1-p)^{k+1}$$

for $k = 0, 1, 2, \ldots$, where $p \in [0,1]$. The mean and variance are given by $(1-p)/p$ and $(1-p)/p^2$. Observe that, if X is Geom(p), then

$$\begin{aligned} P\{\, X \geq m+n \mid X \geq m \,\} &= \frac{P\{\, X \geq m+n \,\}}{P\{\, X \geq m \,\}} = \frac{\bar{F}(m+n-1)}{\bar{F}(m-1)} \\ &= \frac{(1-p)^{m+n}}{(1-p)^m} = (1-p)^n = P\{\, X \geq n \,\}, \end{aligned}$$

so that the Geom(p) distribution has a memoryless property analogous to that of the exponential distribution. Indeed, the geometric distribution can be viewed as the discrete analog of the exponential distribution.

7.9 Poisson Distribution

The *Poisson distribution* with parameter λ, abbreviated Poisson(λ), has support on the nonnegative integers and a PDF given by

$$p(k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

for $k = 0, 1, 2, \ldots$, where $\lambda > 0$. The mean and variance are both equal to λ.

8 Conditional Expectation

8.1 Definition

If X is a discrete random variable, then we define the conditional expectation of X, given that event B has occurred, as

$$E[X \mid B] = \sum_x x P\{\, X = x \mid B \,\}.$$

If X is an indicator random variable of the form $I(A)$ then this definition reduces to our definition of conditional probability in (3.1).

An important special case occurs when $B = \{\, Y = y \,\}$, where Y is a discrete random variable defined on the same probability space as X and y is a real number such that $P(B) = P\{\, Y = y \,\} > 0$. Then we can write

$$E[X \mid Y = y] = \sum_{x \in S} x\, p_{X|Y}(x|y).$$

For example, in the two-dice experiment with fair and independent dice, if Y is the number of spots on the black die, Z is the number of spots on the white die, and $X = Y + Z$, then, based on

y	1	2	3	4	5	6
$E[X \mid Y = y]$	27/6	33/6	39/6	45/6	51/6	57/6

Table 2: Conditional expectation of $X = Y + Z$, given Y, for the two-dice experiment.

Table 1, we have $E[X] = 7$. On the other hand, we have, for example,

$$p_{X|Y}(2|1) = P\{\, X = 2 \mid Y = 1 \,\} = \frac{P\{\, X = 2, Y = 1 \,\}}{P\{\, Y = 1 \,\}} = \frac{1/36}{6/36} = 1/6,$$

and similar calculations show[6] that $p_{X|Y}(x|1) = P\{\, X = x \mid Y = 1 \,\} = 1/6$ for $x = 3, 4, 5, 6, 7$. Thus

$$E[X \mid Y = 1] = \sum_{x=2}^{7} x\, p_{X|Y}(x|1) = \sum_{x=2}^{7} x\,(1/6) = 27/6.$$

Continuing in this manner, we obtain the results in Table 2.

If X and Y are continuous, then we define

$$E[X \mid Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)\, dx.$$

This definition has the same form as in the discrete case, but with the conditional PDF playing the role of the conditional PMF.

In both the discrete and continuous cases, we can interpret the quantity $E[X \mid Y = y]$ as an observer's computation of the expected value of X, given the additional information that the outcome of the probabilistic experiment was such that $Y = y$. Suppose that *before* we run the experiment, we decide that we will tell the observer the value of Y and then ask the observer to assess the conditional expected value of X, given Y. What is our *a priori* assessment of the observer's answer? First note that we can view the quantity $E[X \mid Y = y]$ as a deterministic function $g(y)$ of the variable y. E.g., in Table 2, we have $g(1) = 27/6$, $g(2) = 33/6$, and so forth. Prior to the experiment, we view Y as a random variable and our *a priori* assessment of the observer's answer is $g(Y)$, which is also a random variable. We write $E[X \mid Y]$ for the random variable $g(Y)$. In the two-dice experiment for example, we have $P\{\, Y = y \,\} = 1/6$ for $y = 1, 2, 3, 4, 5, 6$, so that, from Table 2, we see that $E[X \mid Y] = 27/6$ with probability 1/6, $E[X \mid Y] = 33/6$ with probability 1/6, and so forth.

We can extend our definition of conditional expectation in an obvious way to define quantities such as $E[X \mid Y_1, Y_2 \ldots, Y_n]$ for $n \geq 2$; sometimes we abbreviate such a conditional expectation using notation of the form $E[X \mid \mathcal{G}]$, where $\mathcal{G} = \{\, Y_1, Y_2 \ldots, Y_n \,\}$. Indeed, we can use this notation even when $\mathcal{G}$ is a singleton set—so that $\mathcal{G} = \{\, Y \,\}$ for some Y—by equating $E[X \mid \mathcal{G}]$ with $E[X \mid Y]$.

[6] Alternatively, note that, for $x = 2, 3, 4, 5, 6, 7$, we have $P\{\, X = x \mid Y = 1 \,\} = P\{\, Z = x - 1 \mid Y = 1 \,\} = P\{\, Z = x - 1 \,\} = 1/6$, where the second equality follows from the independence of Y and Z and the third equality follows from the fairness of the white die.

8.2 Basic Properties of Conditional Expectation

Some basic properties of conditional expectation are as follows. Let X, and Y be random variables and $\mathcal{G}$ a collection of one or more random variables, and let a, b, and c be real-valued constants. Then

(i) If $X = c$ a.s., then $E[X \mid \mathcal{G}] = c$ a.s..

(ii) If X and Y are integrable with $X \le Y$ a.s., then $E[X \mid \mathcal{G}] \le E[Y \mid \mathcal{G}]$ a.s..

(iii) $|E[X \mid \mathcal{G}]| \le E[|X| \mid \mathcal{G}]$ a.s..

(iv) If X and Y are integrable, then $E[aX + bY \mid \mathcal{G}] = aE[X \mid \mathcal{G}] + bE[Y \mid \mathcal{G}]$ a.s..

Each of the above properties is asserted to hold almost surely, since the conditional expectations are random variables. On the other hand, these properties can often be expressed without reference to almost sure events. For example, consider the property in (i) when $\mathcal{G} = \{ Z \}$ and Z is discrete. For $E[X \mid Z] = c$ to hold almost surely, it must be the case that $E[X \mid Z = z] = c$ for each z such that $P\{ Z = z \} > 0$. Similarly, the conclusion in (ii) implies that $E[X \mid Z = z] \le E[Y \mid Z = z]$ for each z such that $P\{ Z = z \} > 0$. Each of the remaining properties can be specialized in this manner.

Another key property is the *law of total expectation*, which asserts that[7] $E[X] = E\big[E[X \mid \mathcal{G}]\big]$. When $\mathcal{G} = \{ Y \}$ with Y either discrete or continuous, the law of total expectation is often written as

$$E[X] = \sum_y E[X \mid Y = y] \, P\{ Y = y \}$$

or

$$E[X] = \int_{-\infty}^{\infty} E[X \mid Y = y] \, f_Y(y) \, dy.$$

If we take $X = I(A)$ and $Y = \sum_{k=1}^n kI(B_k)$, where $B_1, B_2, \ldots, B_n$ are mutually disjoint events such that $B_1 \cup B_2 \cup \cdots \cup B_n = \Omega$, then the law of total expectation reduces to the law of total probability given in (3.2).

Under certain conditions, a conditional expectation can be "factored" into two multiplicative terms. Specifically, suppose that the value of a random variable X can be determined exactly from the values of the random variables in the collection $\mathcal{G}$. If the random variables Y and XY are both integrable, then

$$E[XY \mid \mathcal{G}] = XE[Y \mid \mathcal{G}] \text{ a.s..}$$

It follows from this result, for example, that $E\left[X^2 Y \mid X\right] = X^2 E[Y \mid X]$ a.s..

The final property of conditional expectation that we consider concerns the effects of conditioning on different amounts of information. Let X be an integrable random variable, and let $\mathcal{G}_1$ and $\mathcal{G}_2$ be collections of random variables such that $\mathcal{G}_1 \subseteq \mathcal{G}_2$. Then

$$E\big[E[X \mid \mathcal{G}_1] \bigm| \mathcal{G}_2\big] = E\big[E[X \mid \mathcal{G}_2] \bigm| \mathcal{G}_1\big] = E[X \mid \mathcal{G}_1] \text{ a.s..}$$

[7] The phrase "almost surely" does not appear here, since we are asserting the equality of two numbers.

For example, it follows that $E\left[X \mid Y\right] = E\big[E\left[X \mid Y, Z\right] \bigm| Y\big]$ a.s.. If all of the random variables are discrete, this assertion implies that

$$E\left[X \mid Y = y\right] = \sum_z E\left[X \mid Y = y, Z = z\right] p_{Z|Y}(z|y)$$

for any y such that $P\left\{ Y = y \right\} > 0$.

8.3 Conditional Probability with Respect to Random Variables

Conditional probabilities with respect to a collection $\mathcal{G}$ of random variables can be defined by setting $P(A \mid \mathcal{G}) = E\left[I(A) \mid \mathcal{G}\right]$. Observe that $P(A \mid \mathcal{G})$ is a random variable. The properties of such conditional probabilities follow directly from the analogous properties of conditional expectations.

8.4 Conditional Moments, Variance Decomposition

We can define higher-order conditional moments and conditional central moments, such as

$$\mathrm{Var}\left[X \mid \mathcal{G}\right] = E[X^2 \mid \mathcal{G}] - E^2\left[X \mid \mathcal{G}\right].$$

Observe that

$$\begin{aligned} E\big[\mathrm{Var}\left[X \mid \mathcal{G}\right]\big] &= E\big[E[X^2 \mid \mathcal{G}] - E^2\left[X \mid \mathcal{G}\right]\big] = E\big[E[X^2 \mid \mathcal{G}]\big] - E\big[E^2\left[X \mid \mathcal{G}\right]\big] \\ &= E[X^2] - E\big[E^2\left[X \mid \mathcal{G}\right]\big] \end{aligned}$$

and

$$\mathrm{Var}\big[E\left[X \mid \mathcal{G}\right]\big] = E\big[E^2\left[X \mid \mathcal{G}\right]\big] - E^2\big[E\left[X \mid \mathcal{G}\right]\big] = E\big[E^2\left[X \mid \mathcal{G}\right]\big] - E^2\left[X\right],$$

where we have used the law of total expectation. Combining these results, we obtain the important relationship

$$\mathrm{Var}\left[X\right] = \mathrm{Var}\big[E\left[X \mid \mathcal{G}\right]\big] + E\big[\mathrm{Var}\left[X \mid \mathcal{G}\right]\big].$$

This decomposition of the variance is a key ingredient in a variety of variance-reduction techniques for simulation output analysis.

9 Stochastic Convergence and Basic Limit Theorems

The statistical methods used to analyze the output of a simulation rest on limit theorems for sequences of random variables. Such limit theorems involve several different modes of convergence.

9.1 Modes of Convergence

Let X and $\left\{ X_n \colon n \geq 1 \right\}$ be random variables defined on a common probability space. Then the sequence $\left\{ X_n \colon n \geq 1 \right\}$ *converges with probability 1* to X if

$$P\big\{ \lim_{n\to\infty} X_n = X \big\} = 1.$$

We also say that the sequence converges to X *almost surely* (a.s.), and we often write "$X_n \to X$ a.s. as $n \to \infty$" or "$\lim_{n\to\infty} X_n = X$ a.s.." A weaker form of convergence is "convergence in probability." We say that the sequence $\{\, X_n \colon n \geq 1 \,\}$ *converges in probability* to X if

$$\lim_{n\to\infty} P\{\, |X_n - X| \leq \epsilon \,\} = 1$$

for every $\epsilon > 0$, and we write $X_n \overset{\text{pr}}{\to} X$. A still weaker form of convergence is as follows. The sequence $\{\, X_n \colon n \geq 1 \,\}$ *converges in distribution* to X if

$$\lim_{n\to\infty} P\{\, X_n \leq x \,\} = P\{\, X \leq x \,\}$$

for each x at which the CDF of X is continuous, and we write $X_n \Rightarrow X$. Observe that the random variables involved in the foregoing definition need not be defined on the same probability space. Setting $F(x) = P\{\, X \leq x \,\}$ and $F_n(x) = P\{\, X_n \leq x \,\}$ for each n, convergence in distribution is sometimes expressed as "weak convergence" of the CDF sequence $\{\, F_n \colon n \geq 1 \,\}$ to F:

$$\lim_{n\to\infty} F_n(x) = F(x)$$

for each x at which F is continuous, and we write $F_n \Rightarrow F$. In our applications, the limiting distribution is typically continuous, e.g., the $N(0,1)$ distribution, so that convergence must occur for every point x. The following assertions summarize the relationships between the various modes of convergence:

(i) If $X_n \to X$ a.s., then $X_n \overset{\text{pr}}{\to} X$.

(ii) If $X_n \overset{\text{pr}}{\to} X$, then $X_n \Rightarrow X$.

(iii) If $X_n \Rightarrow c$ for some real constant c, then $X \overset{\text{pr}}{\to} c$.

By "$X_n \Rightarrow c$" in the foregoing result, we mean that the sequence $\{\, X_n \colon n \geq 1 \,\}$ converges to the degenerate random variable that equals c with probability 1.

An important result concerning convergence in distribution is the *continuous mapping theorem*, which states that if $X_n \Rightarrow X$, then $h(X_n) \Rightarrow h(X)$ for any continuous function h. (This same result holds trivially for the other two modes of convergence.) Another key result is *Slutsky's theorem*: If $X_n \Rightarrow X$ and if $Y_n \Rightarrow c$ for some real-valued constant c, then

(i) $X_n + Y_n \Rightarrow X + c$;

(ii) $Y_n X_n \Rightarrow cX$; and

(iii) $X_n/Y_n \Rightarrow X/c$ provided $c \neq 0$.

9.2 Limit Theorems

Now that we have defined various modes of stochastic convergence, we are ready to discuss limit theorems for random variables. The most basic limit theorems involve a sequence $\{\, X_n \colon n \geq 1 \,\}$ of mutually independent and identically distributed (i.i.d.) random variables. Denote by $\bar{X}_n$ the average of the first n random variables ($n \geq 1$):

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

The *strong law of large numbers* (SLLN) for i.i.d. random variables asserts that, if $\mu = E\left[X_1\right] < \infty$, then

$$\lim_{n \to \infty} \bar{X}_n = \mu \text{ a.s..}$$

That is, with probability 1 the sample average of the X_i's converges to the common expected value of these random variables. In the historically earliest version of the SLLN, each X_i is a Bern(p) random variable. In this setting, the SLLN asserts that $\bar{X}_n$, the fraction of successes in n trials, converges almost surely to p, the success probability, as the number of trials increases.

The *central limit theorem* (CLT) for i.i.d. random variables provides information about the approximate distribution of the random variable $\bar{X}_n$ as n becomes large, and illuminates the rate of convergence in the SLLN. As above, let $\{\, X_n \colon n \geq 0 \,\}$ be a sequence of i.i.d. random variables with common mean μ and common variance σ^2. The CLT asserts that if $0 < \sigma^2 < \infty$, then

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \Rightarrow N(0,1)$$

as $n \to \infty$, where $N(0,1)$ is a standard (mean 0, variance 1) normal random variable. Intuitively, $\bar{X}_n$ is distributed approximately as $\mu + (\sigma/\sqrt{n})N(0,1)$, and hence as $N(\mu, \sigma^2/n)$, when the sample size n is large. Note that the asymptotic variance σ^2/n converges to 0 as n becomes large, so that the probability distribution of $\bar{X}_n$ becomes increasingly concentrated around μ; this behavior is consistent with the SLLN. (The SLLN, of course, makes an even stronger assertion about the convergence of $\bar{X}_n$ to μ.)

10 Estimating Means and Variances

In this section, we consider the problem of estimating the mean μ and variance σ^2 (both assumed unknown to us) of a probability distribution, given a set of n independent and identically distributed (i.i.d.) observations $X_1, X_2, \ldots, X_n$ drawn from this distribution.

We first define two desirable properties for a general estimator Z_n of an unknown parameter θ, where Z_n is computed from i.i.d. observations $X_1, X_2, \ldots, X_n$. The estimator Z_n is said to be *unbiased* for θ if $E\left[Z_n\right] = \theta$ for each n, and is *strongly consistent* for θ if $Z_n \to \mu$ with probability 1 as $n \to \infty$. An unbiased estimator Z_n is equal "on average" to the quantity that it is trying to estimate—there are no systematic errors that cause Z_n to consistently be too high or too low over a sequence of probabilistic experiments. The value of a strongly consistent estimator becomes (with probability 1) closer and closer to the quantity that it is trying to estimate as the sample size becomes larger and larger. I.e., the more observations, the better the estimate.

10.1 Point Estimates of the Mean and Variance

A natural point estimator of μ is the sample average $\bar{X}_n = (1/n)\sum_{i=1}^n X_i$. Observe that

$$E[\bar{X}_n] = E\big[\frac{1}{n}(X_1 + X_2 + \cdots + X_n)\big] = \frac{1}{n}\big(E[X_1] + E[X_2] + \cdots + E[X_n]\big) = \frac{1}{n}(n\mu) = \mu,$$

Thus $\bar{X}_n$ is unbiased for μ. Moreover, the law of large numbers implies that $\bar{X}_n$ is strongly consistent for μ: $\bar{X}_n \to \mu$ a.s. as $n \to \infty$.

A perhaps unnatural point estimator of σ^2 is the sample variance

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}_n^2\right).$$

To compute the expected value of S_n^2, we can assume without loss of generality that $\mu = 0$, so that $\mathrm{Var}\,[X] = E[X^2]$, since adding a constant c to each X_i does not affect the value of S_n^2. (Here X denotes a generic sample from the distribution of interest.) Then

$$\begin{aligned} E\left[\frac{n-1}{n}S_n^2\right] &= E\left[\frac{1}{n}\sum_i X_i^2 - \frac{1}{n^2}\sum_i\sum_j X_iX_j\right] = E\left[\frac{n-1}{n^2}\sum_{i=1}^n X_i^2 - \frac{1}{n^2}\sum_{i\neq j} X_iX_j\right] \\ &= \frac{n-1}{n^2}\sum_{i=1}^n E[X_i^2] - \frac{1}{n^2}\sum_{i\neq j} E\left[X_iX_j\right] = \frac{n-1}{n}E[X^2] = \frac{n-1}{n}\mathrm{Var}\,[X], \end{aligned}$$

where we have used the fact that, by independence, $E\left[X_iX_j\right] = E\left[X_i\right]E\left[X_j\right] = \mu^2 = 0$ for $i \neq j$. Thus $E[S_n^2] = \sigma^2$, and S_n^2 is unbiased for σ^2. This result motivates the use of the factor $1/(n-1)$ rather than $1/n$ in the definition of S_n^2. It follows from the SLLN that, provided $\mathrm{Var}\,[X] < \infty$, $\lim_{n\to\infty}(1/n)\sum_{i=1}^n X_i^2 = E[X^2]$ a.s. and $\lim_{n\to\infty}\bar{X}_n^2 = E^2[X]$ a.s., so that $S_n^2 \to \sigma^2$ a.s. as $n \to \infty$. Thus S_n^2 is strongly consistent for σ^2.

10.2 Confidence Intervals: Normally-Distributed Observations

In order to meaningfully interpret a point estimate, it is essential to assess the precision of the estimate, i.e., to quantify the uncertainty associated with the estimate. A standard approach to this problem is to provide a $100(1-\delta)\%$ *confidence interval* for the point estimate, where δ is a small number such as 0.01 or 0.05. In the case of an unknown mean μ, this means that we compute endpoint estimators L and U from the data, such that with probability $1-\delta$ these numbers bracket μ. That is, $P\left\{\, L \le \mu \le U \,\right\} = 1-\delta$. Roughly speaking, if we were to repeatedly draw samples of size n from the distribution under study and compute a confidence interval from each sample, then the confidence interval $[L, U]$ would contain the unknown number μ about $100(1-\delta)\%$ of the time.

Suppose that the distribution of interest is $N(\mu, \sigma^2)$, and that we wish to estimate the unknown mean μ based on an i.i.d. sample $X_1, X_2, \ldots, X_n$. This situation may seem a bit artificial, but

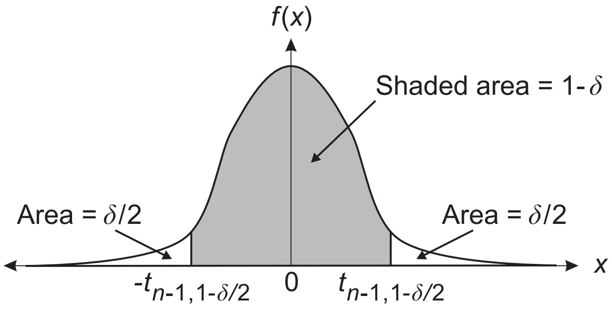

Figure 9: Critical points for t distribution.

it turns out to be highly relevant to a number of estimation problems that arise when analyzing simulation output. From basic properties of the normal distribution, we know that the sample mean $\bar{X}_n$ is $N(\mu, \sigma^2/n)$. Moreover, it is not too hard to show that the normalized sample variance nS_n^2/σ^2 has a χ^2_{n-1} distribution and is independent of $\bar{X}_n$. It follows that the random variable $\sqrt{n}(\bar{X}_n - \mu)/S_n$ has a t_{n-1} distribution, i.e., a Student t distribution with $n-1$ degrees of freedom. Let t be the unique number such that $P\{-t \le T_{n-1} \le t\} = 1-\delta$, where T_{n-1} is a random variable having the t_{n-1} distribution. Then

$$P\left\{ \bar{X}_n - \frac{tS_n}{\sqrt{n}} \le \mu \le \bar{X}_n + \frac{tS_n}{\sqrt{n}} \right\} = P\left\{ -t \le \frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} \le t \right\} = 1 - \delta, \tag{10.1}$$

where the first equality follows from the equivalence of the respective events and the second equality follows from the definition of t. Note that, because Student t distributions are symmetric, t is also the unique number such that $P\{T_{n-1} \le t\} = 1 - (\delta/2)$, that is, t is the $1 - (\delta/2)$ quantile of the Student t distribution with $n-1$ degrees of freedom. We therefore write $t_{n-1,1-\delta/2}$ rather than just t to denote this quantity; see Figure 9. Using this notation, we see from (10.1) that a $100(1-\delta)\%$ confidence interval for μ is given by $[L, U] = [\bar{X}_n - H_n, \bar{X}_n + H_n]$, where $H_n = t_{n-1,1-\delta/2}S_n/\sqrt{n}$. The quantity H_n is known as the *half-width* of the confidence interval, and the quantity $S_n/\sqrt{n}$, which is the standard deviation of $\bar{X}_n$, is sometimes called the *standard error* of the estimator $\bar{X}_n$.

10.3 Confidence Intervals: Large Samples

In many (if not most) cases, we have no reason to believe that the distribution of interest is normal, so the preceding results do not apply. If the sample size n is large,[8] however, then we can obtain an approximate $100(1-\delta)\%$ confidence interval by using the central limit theorem. Specifically, the central limit theorem given in Section 9.2 asserts that

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \Rightarrow N(0,1)$$

[8] When the true data distribution is reasonably symmetric about the mean and the tails of the distribution are not too heavy, then values of $n \ge 50$ can be considered large

as $n \to \infty$. It follows from the SLLN that $\sigma/S_n \to 1$ a.s., and hence $\sigma/S_n \Rightarrow 1$, as $n \to \infty$. Therefore, by Slutsky's theorem,

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma_n} \frac{\sigma}{S_n} \Rightarrow N(0,1)$$

as $n \to \infty$. Thus the random variable $\sqrt{n}(\bar{X}_n - \mu)/S_n$ has approximately a $N(0,1)$ distribution for large n. We can now proceed almost exactly as in Section 10.2 and obtain an approximate $100(1-\delta)\%$ confidence interval $[\bar{X}_n - H_n, \bar{X}_n + H_n]$, where $H_n = z_{1-\delta/2} S_n/\sqrt{n}$ and $z_{1-\delta/2}$ is the $1-(\delta/2)$ quantile of the standard normal distribution. We sometimes call such a confidence interval an *asymptotic* confidence interval because the probability that the interval contains μ does not equal $1-\delta$ exactly, but rather converges to $1-\delta$ as $n \to \infty$.

11 Further Resources

Chapter 4 in [5] also reviews some basic topics in probability and statistics. Some classic introductory treatments of probability are given in [2, 4] and some highly regarded recent textbooks include [1, 6]. A good starting point for web resources (including a free downloadable introductory textbook by Grinstead and Snell) can be found at:

`http://www.dartmouth.edu/~chance/teaching_aids/books_articles/probability_book/book.html`

The Appendix in [3] gives a somewhat higher-level introduction to the topics covered in this document, and provides pointers to several advanced references.

References

[1] D. P. Bertsekas and J. N. Tsitsiklis. *Introduction to Probability*, Athena Scientific, 2002.

[2] W. Feller. *An Introduction to Probability Theory and Its Applications, Volume 1*, 3rd ed., Wiley, 1968.

[3] P. J. Haas. *Stochastic Petri Nets: Modelling, Stability, Simulation*, Springer-Verlag, 2002.

[4] P. G. Hoel, S. C. Port, and C. J. Stone. *Introduction to Probability Theory*, Houghton-Mifflin, 1971.

[5] A. M Law. *Simulation Modeling and Analysis*, 4th ed., McGraw-Hill, 2007.

[6] S. Ross. *A First Course in Probability*, 6th ed., Prentice Hall, 2001.