eligible voter's decision to cast a ballot is influenced by one (or a handful of) other voter(s), we might think it less likely that it is influenced by many of the other voters. If we draw a random sample (i.e., choose the survey respondents at random), then the probability that any given respondent's decision to vote was influenced by another respondent's decision is effectively zero.[33]

We can use the Bernoulli distribution to describe the relative frequency distribution of the outcomes over "not vote, vote" in an Australian election. If, for example, 96% of the electorate submitted valid ballots, then the relative frequency distribution would look like Figure 10.4.

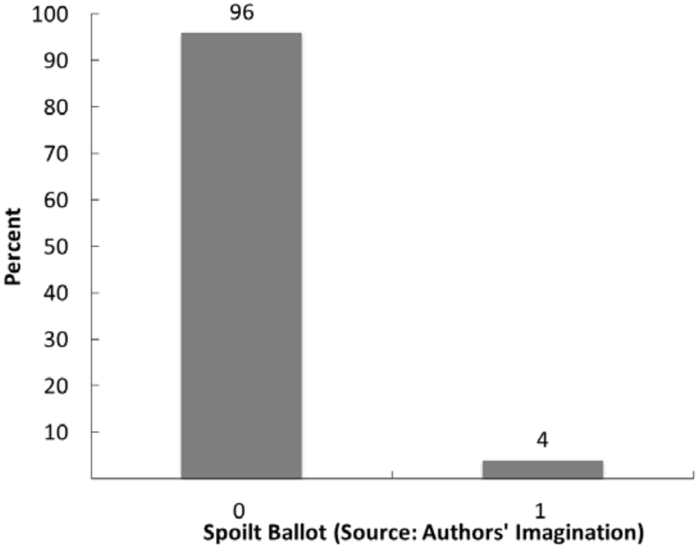

Figure 10.4: Bernoulli Distribution, $p = 0.04$

The Bernoulli distribution provides an important foundation for building more complex distributions, as we show below. It is useful for both statistical and theoretical models where one is interested in sequences of independent binary choices.

A more detailed overview of the Bernoulli distribution can be found online at `http://mathworld.wolfram.com/BernoulliDistribution.html`.

10.6.2 The Binomial Distribution

The PMF for the binomial distribution is defined by the equation:

$$Pr(Y = y|n, p) = \binom{n}{y} p^y (1-p)^{n-y} \tag{10.5}$$

[33] If this point does not make sense, please review a discussion of random sampling in a good research design text.

Table 10.8: Unanimous Court Decisions

Case 1	Case 2	Case 3	No. of Unanimous Cases
D	D	D	0
U	D	D	1
D	U	D	1
D	D	U	1
U	U	D	2
U	D	U	2
D	U	U	2
U	U	U	3

where $n \geq y$, n, and y are positive integers and $0 \leq p \leq 1$.[34] The variables n and y in equation (10.5) represent the number of cases (or observations) and the number of positive outcomes, respectively. We recognize that this is the sort of equation that makes many political scientists blanch and ask why they are messing with such exotica! So let's work it out via a concrete example, after first describing the assumptions that underlie it. You can also flip back to Figure 10.1 to recall what an example of the binomial distribution looks like.

The binomial distribution can describe any discrete distribution with three or more observations where (1) each observation is composed of a binary outcome, (2) the observations are independent, and (3) we have a record of the number of times one value was obtained (e.g., the sum of positive outcomes). As an example, a data source might record the number of unanimous votes by a court (e.g., Epstein, Segal, and Spaeth, 2001) but not provide us with the individual vote breakdown for each case. If we assume that the justices' votes are independent across cases, then the binomial distribution should be useful for describing the DGP.

To keep the example simple, we will assume that the court rules on only three cases per term. This is not terribly realistic, but one could extend it to twenty-five, thirty, or however many cases there actually are per term. Limiting the example to three keeps things tractable.

The first thing to do is to list the possible outcomes as ordered sets and count them. Since there are two possibilities (divided decision, unanimous decision) and three cases, there are $2^3 = 8$ possibilities, as listed in Table 10.8.

To develop the binomial distribution, we start with the Bernoulli distribution, which says that $Pr(Y = 1) = p$ and $Pr(Y = 0) = 1 - p$ (see equation (10.3)). We will assign a unanimous case (U) the value 1 and a divided case (D) the value 0. Since we have assumed that the three cases are independent, the probability that there are zero unanimous (i.e., three divided) cases is the product of the

[34] Recall from the previous chapter that $\binom{n}{y} = \left(\frac{n!}{y!(n-y)!}\right)$ is shorthand for choosing y from n, i.e., the number of combinations that involve choosing y elements of some type from n total elements.

marginal probabilities that each case is divided, or $Pr(Y = 0, 0, 0)$: $(1 - p) \times (1 - p) \times (1 - p) = (1 - p)^3$. This matches equation (10.5) when $n = 3$ and $y = 0$.

The probability that there is only one unanimous case is the sum of the products of the marginal probabilities over the three ordered sets that might produce that outcome. That's a mouthful, so let's break it down. In Table 10.8, D represents a divided decision and U represents a unanimous decision. The table indicates that there are three different ways we might end up with one unanimous decision. So we will need to sum the probabilities over those three ways. What is the probability that we will observe only one unanimous case in each manner it can be achieved? Again, we take the product of the marginal probabilities. In the first row in Table 10.8 with only one unanimous case, that product is $p \times (1 - p) \times (1 - p)$. In the second row with only one unanimous case, the joint probability is $(1 - p) \times p \times (1 - p)$. Finally, the third row produces the joint probability $(1 - p) \times (1 - p) \times p$. When we add those three together we get $3p(1 - p)^2$. This matches equation (10.5) when $n = 3$ and $y = 1$.

We determine the other outcomes the same way: we take the sum of the joint probabilities, each of which is the product of marginal probabilities. Table 10.8 indicates that there are again three ordered sets that yield two unanimous decisions. Thus, the probability that we observe two unanimous decisions is the sum of the joint probabilities of each of those combinations: $p \times p \times (1 - p)$ plus $p \times (1 - p) \times p$ plus $(1 - p) \times p \times p$, or $3p^2(1 - p)$. This matches equation (10.5) when $n = 3$ and $y = 2$.

Finally, there is only one ordered set that produces the outcome of three unanimous decisions. So the probability that there are three unanimous decisions is the joint probability $p \times p \times p = p^3$. This matches equation (10.5) when $n = 3$ and $y = 3$.

Equation (10.5) is simply a general representation of the sum of the joint probabilities that we discussed in the preceding paragraphs as individual equations. To get a graphical sense of what the binomial distribution looks like, please point your browser to Balasubramanian Narasimhan's "Binomial Probabilities" applet, available at `http://www-stat.stanford.edu/~naras/jsm/example5.html`.

There are some statistical routines that rely on the binomial distribution (e.g., `bitest` in Stata), and the binomial distribution can be assumed in generalized linear regression models. Though these tests are common in other fields, they are not used widely in political science.

Some readers might be interested in a more detailed presentation of the binomial distribution. Gill (2006, sections 1.4.3, 6.2, 7.1.3, 7.1.4), Lindsey (1995, pp. 13–14, 99–201), and King (1989, pp. 43–45) are great places to start. A thorough technical overview is available online at `http://mathworld.wolfram.com/BinomialDistribution.html`.

10.6.3 The Multinomial Distribution

The multinomial distribution is an extension of the binomial distribution to cases where more than two mutually exclusive (and collectively exhaustive) outcomes can occur. Whereas the binomial distribution describes the number of times $Y = 1$, where Y is a random variable described by a Bernoulli distribution, the multinomial distribution counts the number of times each one of k different outcomes happens, where each outcome happens with probability p_i, $i \in \{1, \ldots, k\}$. Since the outcomes are mutually exclusive and collectively exhaustive, all these probabilities sum to one. Let Y_i represent a random variable that counts the number of times outcome i occurs. If there are n independent events, then $Y_i \in \{0, 1, 2, \ldots, n\}$ for all i, and $\sum_{i=1}^{k} Y_i = n$. In this case we can write the multinomial PMF for non-negative integers $y_1, \ldots, y_k$ as

$$Pr((Y_1 = y_1) \cap \ldots \cap (Y_k = y_k)) = \begin{cases} \frac{n!}{y_1! \ldots y_k!} \prod_{i=1}^{k} p_i^{y_i} & \text{when } \sum_{i=1}^{k} y_i = n, \\ 0 & \text{otherwise.} \end{cases} \tag{10.6}$$

Though the multinomial distribution is not often invoked in applied statistical work in political science, it can be invoked as one of many possible distributions when using what is called the generalized linear model (GLM) to estimate a regression equation (i.e., a statistical model you will learn about). The GLM has not yet become popular in political science, but see Gill (2001) for an introduction by a political scientist.

Readers interested in more thorough and technical discussions should examine the MathWorld entry at `http://mathworld.wolfram.com/MultinomialDistribution.html` or Zelterman (2004, pp. 8–9).

10.6.4 Event Count Distributions

Many variables that political scientists have created are integer counts of events: the number of bills passed by a legislature, the number of wars in which a country has participated, the number of executive vetoes, etc. Event counts frequently exhibit frequency distributions consistent with those produced by a handful of well-known probability distributions.

10.6.4.1 The Poisson Distribution

The Poisson distribution is named after the French mathematician Siméon Denis Poisson. Its PMF can be written as

$$Pr(Y = y|\mu) = \frac{\mu^y}{y! \times e^{\mu}}, \tag{10.7}$$

where $\mu > 0$ is the expected number of events, y is a positive integer representing the number of events observed, and the variance, σ^2, is equal to the mean, μ.[35]

[35] You will also see this equation written as $Pr(Y = y|\mu) = e^{-\mu} \frac{\mu^y}{y!}$. Recall that $e^{-\mu} = \frac{1}{e^{\mu}}$.

The Poisson has a location parameter (μ) but it does not have a separate scale parameter.

The graph of the Poisson distribution, displayed in Figure 10.2 above, reveals an asymmetry: these distributions tend to have a long right tail. Note, however, that as the mean of the distribution rises, the asymmetry of the distribution declines.

Whence comes equation (10.7)? The goal is to produce a PMF that describes the number of times one observes zero events, one event, two events, 3 events, etc., over a fixed period of time (e.g., wars per century). The Poisson distribution describes event counts produced by a process that meets three criteria: integer count, independence, and a known mean. We discuss each in turn.

First, the individual events must be countable as whole numbers given a period of time, and it cannot be possible to count the non-events. The inability to count non-events may seem odd, but this is actually quite common. For example, we might want to observe the number of wars countries entered into during the twentieth century. We can easily count this using whole numbers. Note, however, that it is nonsensical to count the number of non-wars into which countries entered during the twentieth century.[36] Recall that the Bernoulli and binomial distributions involve events with binary countable outcomes: we can count the events *and* the non-events. When we can only count the events, and not the non-events, the Poisson distribution might be useful.

Second, the events must be produced independently from one another over the period of time one is counting them. Consequently, the probability that the count is, say, two, is computed independently of the probability that the count is, say, five.[37] Third, the average frequency of events in a given period (μ in equation (10.7)) must be known. When used in statistical analyses one can determine μ from one's data, but this requirement explains why we use the notation $Pr(Y = y|\mu)$ in equation (10.7).

A classic example of an event count generated by a Poisson process is the number of traffic accidents at a given intersection over time (e.g., the number of accidents per quarter year). Five years of quarterly data on the number of accidents at a given intersection will often prove to be Poisson distributed. Yet a large number of accidents in any three month period (say four or five) could lead people to conclude that the intersection is dangerous—which is to say that the accidents are *not* independent. The Poisson distribution—which assumes independence of events—shows that even when we assume that events

[36] Though one can, and scholars do, count the number of years that contain no wars between the countries in a particular pair of countries (a dyad).

[37] This implies that one adds all the probabilities of each count's occurring to get the probability that some number occurs (i.e., the probability that some count in the sample space occurs, which is 1). In fewer words, if S is the event that greater than or equal to 0 events occur, $1 = Pr(S) = \frac{\mu^0}{0!e^\mu} + \frac{\mu^1}{1!e^\mu} + \frac{\mu^2}{2!e^\mu} + \ldots = e^{-\mu}\sum_{i=0}^{\infty}\frac{\mu^i}{i!}$. Since the sum is the definition of (and Taylor series for) e^μ, we see that the RHS of this does equal 1. This is actually why the e^μ is present in the PMF: it is a normalization factor, to ensure that when one sums over all possible outcomes (i.e., all possible counts), one gets 1, as one must for a PMF.

are independent of one another we will still randomly get clusters of relatively large numbers of events. Such clustering of events will be unusual (i.e., have a low frequency), but we should be reluctant to accept a single large cluster as sufficient evidence to infer that the events were not independent. Thus, if one conducts a data analysis and finds that the data fit the Poisson distribution, one can conclude that the accidents were likely produced randomly. If they are not Poisson distributed, then perhaps the light at the intersection or the speed limit needs to be evaluated to determine what systematic factor is producing the accidents.

That said, many integer counts of events that interest political scientists are expected to be related to one another by theory. For example, it seems unlikely that bills passed in a legislature, unanimous court decisions, wars, or executive vetoes are independent of one another. And if we assume that the presence of one event either raises or lowers the probability of another event in a given period of time, then a variable measuring that event type would not be produced by a Poisson process.

For a lucid and detailed discussion of the Poisson distribution, visit Bruce Brooks's entry at his "Acquiring Statistics" site: `http://www.umass.edu/wsp/statistics/lessons/poisson/`.

10.6.4.2 The Negative Binomial Distribution

The Poisson distribution describes the distribution of event counts for rare random events. The negative binomial, on the other hand, provides one with the expected event count prior to the occurrence of a set number of non-events. Because it is built on the binomial distribution, the DGP is one where events have binary countable outcomes (i.e., once we know how many non-events occurred, we can determine the number of events by subtracting the number of non-events from the total number of trials).

The PMF for the negative binomial distribution can be written as

$$Pr(Y = y|r, p) = \binom{y + r - 1}{y} p^y (1 - p)^r, \tag{10.8}$$

where y is the number of observed events (typically called "successes"; e.g., presidential vetoes), r is the number of observed non-events (typically called "failures"; e.g., presidential signatures on bills) over $y + r$ opportunities (or Bernoulli trials), and p is the probability of any particular event ("success"; e.g., veto). The distribution describes the number of events (successes, vetoes), y, prior to observing the rth non-event (failures, signed bills).[38] We should note that what one calls an event (success) or non-event (failure) is arbitrary, and one can frame this distribution as describing the number of successes (vetoes) before a set number of failures (signed bills), as we have done, or the number

[38] Zelterman (2004, pp. 13–14) provides a proof that this PMF sums to 1.

of failures (vetoes) before a set number of successes (signed bills). To switch to the alternative formulation, swap p and $1 - p$ in equation (10.8).

The combination in the PMF, $\binom{y+r-1}{y}$, arises because the negative binomial distribution represents the probability of observing r observations of one outcome (call it "signs the bill") and y observations of the alternative outcome (call it "veto") in $y + r$ observations, given that the $(y + r)$th observation has the value "signs the bill." That is a mouthful, so let's break it down.

The negative binomial distribution is built from the binomial distribution, which was built on the Bernoulli distribution. As you know, the Bernoulli distribution concerns the probability of the outcomes for a binary variable, and in our example the binomial distribution describes the number of "veto" outcomes in a series of independent Bernoulli trials. The negative binomial describes a variable that counts the number of vetoes prior to the rth signed bill, which could be interpreted as the number of successes before the rth failure or the number of failures before the rth success. Lethen[39] offers the following succinct description, which employs the second of these interpretations:

> The negative binomial distribution is used when the number of successes is fixed and we're interested in the number of failures before reaching the fixed number of successes. An experiment which follows a negative binomial distribution will satisfy the following requirements:
>
> 1. The experiment consists of a sequence of independent trials.
> 2. Each trial has two possible outcomes, S or F.
> 3. The probability of success, $\Pi = P(S)$, is constant from one trial to another.
> 4. The experiment continues until a total of r successes are observed, where r is fixed in advance.

When would a political scientist suspect that a variable she is studying was produced by a negative binomial DGP? One possibility is the veto example considered above. Another possibility is a study of international conflict focused on the decision to use force in the presence of international disputes. Students of international politics often study event counts of international uses of force. Imagine that we know that the incidence of uses of force over the past two centuries is .01 (i.e., the probability that any given country uses force in any given year is .01). We can now use equation (10.8) to calculate the PMF for various counts of the use of force. That is, once we select a year in which to begin our observations we can use it to determine the probability that a given country will use force for the first, second, third, etc. time, in the first, second, third, etc. year of observation.

[39] "The Negative Binomial Distribution," available online at `http://stat.tamu.edu/stat30x/notes/node69.html`.

To be concrete, let's calculate the probability that the second use of force occurs in the sixth year, so that there are four years without force. If we call war a failure and peace a success, then equation (10.8) states

$$\begin{aligned} P(Y = 4|2, 0.01) &= \binom{4+2-1}{4} 0.01^2 (1-0.01)^4 \\ &= \binom{5}{4} \times .0001 \times 0.9606 \\ &= \frac{5!}{(4!)(1!)} \times 0.00009606 \\ &= 5 \times 0.00009606 \\ &= 0.00048. \end{aligned}$$

We could perform the same calculations for the probability that the third use of force occurs in the seventh year, etc., but that would get tedious very quickly. And since we have the PMF defined, there is no need to do such calculations as we can instruct a computer to do them if we ever need to calculate several.

The PMF for the negative binomial distribution looks similar to the PMF for the Poisson distribution, as we see in Figure 10.5.

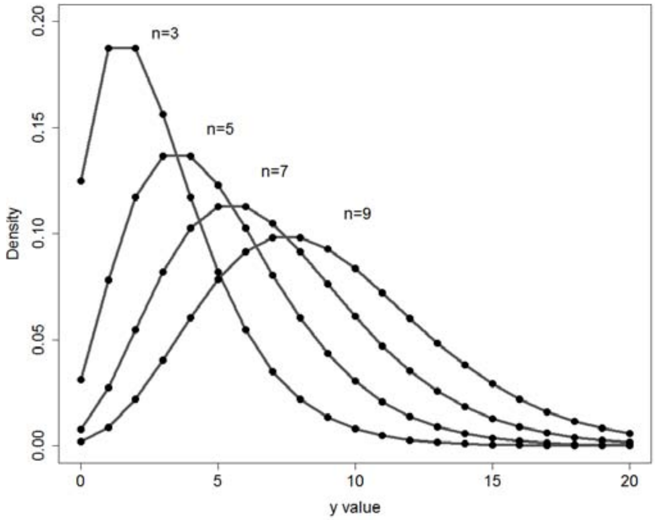

Figure 10.5: PMF of Negative Binomial Distribution, $n = 3, 5, 7, 9; p = 0.5$

One of the key features of the negative binomial distribution relative to the Poisson distribution is that the mean and variance are not constrained to equal one another. The variance is greater than the mean for the negative binomial distribution, and because the negative binomial distribution has two independent parameters, one can set both the mean and the variance separately.

King (1989) brought the negative binomial distribution to the widespread attention of political scientists, and it has been primarily used as a model of event counts when the mean and the variance of the sample data are not equal.[40] In this sense, Poisson regression models tend to be viewed as special cases of negative binomial statistical models.[41]

For a thorough technical overview of the negative binomial distribution, see `http://mathworld.wolfram.com/NegativeBinomialDistribution.html`.

10.6.5 Why Should I Care?

Political scientists are often interested in concepts that can be represented as binary outcomes, ordinal scores, or event counts. Even if one does not intend to use statistics to test hypotheses it is still useful to have an understanding of the difference between these types of distributions. In other words, thinking about distributions leads one to invest in theoretical speculation about what might lead a concept or variable to hold different values in different cases. Another way of saying the same thing is that theory building for the purpose of explaining why different outcomes occur in different cases is equivalent to speculating about a DGP. Further, if one does plan to use statistical hypothesis testing in one's empirical work, then knowledge of discrete distributions and their DGPs is critically important.

10.7 EXPECTATIONS OF RANDOM VARIABLES

We opened this chapter by discussing what a random variable is, but thus far we have mostly discussed how these variables are distributed. This, as we hope we have made clear, is undoubtedly important, but there are still many occasions when one desires more specific knowledge regarding a random variable. For instance, its expected value, its variation around its mean, and one's expected utility when it is a function of the variable are all useful to political scientists. To obtain this knowledge, we must deal with the expectation of the random variable.

The **expectation** of a random variable X, denoted $E_X[X]$ or simply $E[X]$ when no confusion (in the presence of more than one variable) is possible, is the weighted average value that the random variable can take, where the weights are given by the probability distribution.

Let's consider a common example one encounters in game theory and expected utility theory. As noted in the previous chapter, game theorists denote a

[40] Or when a dispersion parameter in Poisson regression models suggests that it is unlikely that the dependent variable was drawn from a Poisson distribution with equal mean and variance.

[41] The log-gamma distribution is an alternative to the negative binomial that is used widely in other fields.

lottery any outcome that is uncertain, including a lottery of the kind US states advertise, such as Powerball. A lottery consists of a set of values of outcomes and a corresponding set of probabilities that each outcome might occur. In other words, it is a probability distribution over values of outcomes, and the outcome of the lottery is a random variable.

We compute that variable's expectation by weighting (multiplying) each value by the chance that it occurs, and summing over all values. So, if the lottery has potential outcomes \$0, \$1,000, and \$1,000,000, and these occur with probabilities 0.9998999, 0.0001, and 0.0000001 respectively, then the expectation of the lottery's outcome is $(0.9998999 \cdot \$0) + (0.0001 \cdot \$1,000) + (0.0000001 \cdot \$1,000,000) = \$0 + \$0.1 + \$0.1 = \$0.20$, or twenty cents.

This is known as the **expected value** of the lottery. In general, if a discrete random variable X takes on values x_i, then the expected value is calculated for X according to the formula[42]

$$E_X[X] = \sum_i x_i(Pr(X = x_i)). \tag{10.9}$$

Note that the complex part of equation (10.9) is the $Pr(X = x_i)$ term. For the example we just did that term was provided for each value x_i. Let's try a slightly more complicated example before moving on, one in which the probabilities are dependent on, and specified for, the values. To do this, we'll use the Poisson distribution we introduced in the previous section.

Imagine that you are interested not in the distribution of event counts but rather in how many events one should expect to see. Recalling that $Pr(X = x_i|\mu) = \frac{\mu^{x_i}}{x_i! \cdot e^{\mu}}$, we can compute this via equation (10.9)

$$E_X[X] = \sum_i i(Pr(X = i)) = \sum_{i=0}^{\infty} i \frac{\mu^i}{i! \times e^{\mu}} = e^{-\mu} \sum_{i=0}^{\infty} \frac{\mu^i i}{i!}.$$

This doesn't give us an answer yet, but we can get there by expanding out the sum

$$\sum_{i=0}^{\infty} \frac{\mu^i i}{i!} = 0 + \sum_{i=1}^{\infty} \frac{\mu^i i}{i!} = \sum_{i=1}^{\infty} \mu \frac{\mu^{i-1}}{(i-1)!} = \mu \left(\sum_{i=0}^{\infty} \mu \frac{\mu^i}{i!} \right) = \mu e^{\mu}.$$

In the first step we pulled the $i = 0$ term in the sum out, which is zero. In the second step we divided top and bottom by i, recalling that $\frac{i!}{i} = (i-1)!$, and pulled a μ out, recalling that $\mu^i = \mu \times \mu^{i-1}$. In the third step we noted that the sum from one to infinity of $i - 1$ is the same as the sum from zero to infinity of i, since the first index of both is zero and they both go on forever. Finally, in the fourth step we used the definition of the exponential function. Plugging this back into the sum in the equation for the expected value produces our answer,

$$E_X[X] = e^{-\mu} \mu e^{\mu} = \mu.$$

[42] A very similar equation is true for continuous random variables, and we provide it in the next chapter. Basically, one replaces the sum with an integral and the PMF with a PDF.

Thus the expected value of the event count—which is also the mean—is equal to the parameter μ. Though we technically knew that, it's nice to be able to derive it ourselves, right? More important, this technique can be used for other PMFs, as well to find expected values of other distributions.

10.7.1 Expected Utility

Expected values are useful, but they are limited in that they consider only the (weighted) average value of the variable itself, and not the average value of more complex functions of that variable, which are what we typically care about in political science. In statistics, expectations of functions of the random variable allow the computation of moments of the distribution, which we consider in the next subsection. In this subsection we turn to the expectation of utility functions, or expected utility for short. **Expected utility**, typically denoted $EU(x)$, is much like expected value, except that rather than specifying a weighted average of the variable it specifies a weighted payoff, under a few assumptions on the utility function about which you will learn in your game theory class.[43] Its expression even looks much the same as that for the expected value:

$$EU(X) = \sum_i u(x_i)(Pr(X = x_i)). \tag{10.10}$$

In equation (10.10), the small u (a Bernoulli utility function) gives the payoffs for the *known* values that the random variable can take, while the $EU(x)$ (von Neumann–Morgenstern utility) provides the weighted average utility one can expect to get, given the probability distribution of the values of the random variable.

Let's start with a concrete example that has the character of the example of the lottery above. This relates to the game matching pennies we introduced in the previous chapter. We'll vary it a little and insert some payoff values to make the calculation clearer. Let two people each toss a penny. If the pennies turn up the same (both heads or both tails), then player 1 keeps player 2's penny. If they turn up mixed (one head and one tail), then player 2 keeps player 1's penny. The payoff (or utility) each player receives from a round of play is 1 cent if she wins and -1 cent if she loses.[44] The difficulty is that we do not know whether she will win or lose. That is, we are uncertain about the outcome. Of course, probabilities help us analyze uncertain situations, and an expected utility calculation is nothing more than a means of determining what utility a person should expect to receive in an uncertain situation.

[43] For a brief introduction see Shepsle and Bonchek (1997, pp. 15–35)

[44] Note that this game has the character of a lottery rather than of a strategic interaction. When matching pennies is introduced in game theory it typically involves the decision to play heads or tails, and is an example of a game in which there are no pure strategy equilibria (e.g., Osborne, 2004, pp. 19–20). That is to say, the optimal strategy is to play each of heads and tails half the time, using what is known as a mixed strategy, as noted in the previous chapter. The optimal strategy produces the lottery we analyze here.

We can use equation (10.10) to compute the expected utility for this game for each player. This is

$$\begin{aligned} EU(MP_{1,2}) &= (p_{HH} \times u_{1,2}(HH)) + (p_{HT} \times u_{1,2}(HT)) \\ &+ (p_{TH} \times u_{1,2}(TH)) + (p_{TT} \times u_{1,2}(TT)), \end{aligned} \tag{10.11}$$

where MP is the matching pennies game (or lottery); the subscripts 1 and 2 indicate the player; p denotes the probabilities of each joint outcome; H indicates a coin landing heads and T indicates a coin landing tails; and u indicates the utility (or payoff) associated with an outcome.

> One reads equation (10.11) as follows: *the expected utility of playing matching pennies for players 1 and 2 is the probability of heads-heads times the utility of heads-heads plus the probability of heads-tails times the utility of heads-tails plus the probability of tails-heads times the utility of tails-heads plus the probability of tails-tails times the utility of tails-tails.*

We can replace the variables with values and calculate the expected utility of this game (really lottery) for each player. We identified the utilities (or payoffs) to each player above (player 1: HH or TT is +1, HT or TH is −1; player 2: HH or TT is −1, HT or TH is +1), but where do the probabilities come from? The sample space has four outcomes that are equally likely: HH, HT, TH, or TT. Therefore, the probability of each outcome is $\frac{1}{4}$ or 0.25. Because the players have different payoffs we must calculate two expected utility equations, one for each player

$$\begin{aligned} EU(MP_1) &= 0.25(1) + (0.25)(-1) + (0.25)(-1) + 0.25(1) \\ &= 0.25 - 0.25 - 0.25 + 0.25 \\ &= 0. \\ EU(MP_2) &= 0.25(-1) + (0.25)(1) + (0.25)(1) + 0.25(-1) \\ &= -0.25 + 0.25 + 0.25 - 0.25 \\ &= 0. \end{aligned}$$

This demonstrates that the expected utility of playing this game is zero for each player. Perhaps that explains why this is not a very popular gambling game.

You may wonder what the point was of presenting this game, given that the utilities were each nothing more than values of the lottery, implying that an expected value computation would be entirely appropriate. Though true in this case, it needn't be: one could assume that both players were risk averse in the realm of gains but risk seeking in the realm of losses, as in prospect theory (Kahneman and Tversky, 1979). In this case, we might assign $u_1(TT) = u_1(HH) = u_2(HT) = u_2(TH) = 2$, and $u_2(TT) = u_2(HH) = u_1(HT) = u_1(TH) = -4$, so that winning is not as good for either player as losing is bad.

One can still use equation (10.10) even though the utilities do not equal the values of the lottery, and verify that $EU(MP_1) = EU(MP_2) = -1$ in this case. Risk-averse players not only get no benefit from playing the game, assuming their utility from doing nothing is zero, but actively prefer not to play the game at all.

We discuss risk preferences a bit more at the end of this section and consider more complex expected utility computations in the next chapter, but first we illustrate further with a more complex and more interesting example, taken from Stokes (2001), which we discussed in the context of Bayes' rule in the previous chapter.

Stokes asks us to consider a voter who places himself on the left side of a left-right ideological continuum (pp. 16–17). The election offers four candidates, none an incumbent, who are vying for the candidacy of two parties. The voter has beliefs about where on the ideological scale both parties sit and can thus identify the party whose policies are closest to (and farthest from) his own. However, he also believes that there are two types of politicians: *ideologues*, who will pursue the policies they campaign on, and *power seekers*, who will lie during the campaign when they know their preferred policy is unpopular, and then switch once in office. The voter's problem is trying to determine how to vote given that though he is confident about the policy the candidates for each party should adopt, he is uncertain whether the candidate for each party is a *power seeker* type or an *ideologue*. One can represent Stokes's voter's decision using the following expected utility equation:

$$EU(v_i) = (p_{iL} \cdot u(L)) + ((1 - p_{iL}) \cdot u(R)) \tag{10.12}$$

where v_i represents a vote for candidate i, p_{iL} represents the probability of a government under i enacting a set of leftist policies and $1 - p_{iL}$ a set of rightist policies,[45] respectively, u represents utility associated with a policy outcome, and L and R represent the leftist and rightist set of policies, respectively. Since i represents the candidate who wins and since several candidates are competing, i is drawn from the set of all candidates.

> A conventional way to read equation (10.12) is: *the expected utility of voting for candidate i is equal to the product of the probability that candidate i adopts leftist policies and the utility derived from leftist policies plus the product of the probability that i does not adopt leftist policies and the utility derived from rightist policies.*

Stokes specifies values for the variables in the equation, thus making it possible to perform calculations and compare the candidates. For her left-leaning voter she assumes that the value of leftist policies is 10 and that the value of rightist policies is −10. If the politician is an ideologue then she will remain faithful to her announced platform with a probability of 1.[46] However, if the

[45] Note that $p_{iR} = 1 - p_{iL}$. One could rewrite the equation using p_{iR} instead of $1 - p_{iL}$.

[46] Since probabilities must sum to 1 and the politician will either remain faithful or switch, the probability that an ideologue switches is $1 - 1 = 0$.

politician is a power seeker, then he will switch policies after the election with probability 0.3.[47]

We can now consider different scenarios. Let us simplify and assume that there are only two candidates, l and r, standing for the left and right party, respectively. Assume further that our voter believes that both candidates are ideologues. To calculate the expected utility for voting for each candidate, place the relevant values from the paragraph above into the equation. In this case, i can take two values, l and r, for each of the two candidates. Because the payoffs to the voter are different for each candidate, we need to calculate the expected utility to the voter for each candidate

$$\begin{aligned}
EU(v_l) &= (p_{iL} \cdot u(L)) + ((1 - p_{iL}) \cdot u(R)) \\
&= 1.0(10) + (1 - 1.0)(-10) \\
&= 10. \\
EU(v_r) &= (p_{iL} \cdot u(L)) + ((1 - p_{iL}) \cdot u(R)) \\
&= 0(10) + (1 - 0)(-10) \\
&= -10.
\end{aligned}$$

Thus, under the specified assumption, $EU(v_l) > EU(v_r)$. In words, the expected utility of voting for an ideologue leftist candidate is greater than the expected utility of voting for an ideologue rightist candidate: the voter should cast a ballot for the leftist party. There is nothing surprising here.[48] Nonetheless, it illustrates how one can construct an expected utility model.

For practice, let us consider another scenario that Stokes does not evaluate. Let's assume that the voter believes that both candidates are power seekers

$$\begin{aligned}
EU(v_l) &= (p_{iL} \cdot u(L)) + ((1 - p_{iL}) \cdot u(R)) \\
&= 0.7(10) + (1 - 0.7)(-10) \\
&= 7 + 0.3(-10) \\
&= 4. \\
EU(v_r) &= (p_{iL} \cdot u(L)) + ((1 - p_{iL}) \cdot u(R)) \\
&= 0.3(10) + (1 - 0.3)(-10) \\
&= 3 + 0.7(-10) \\
&= -4.
\end{aligned}$$

Thus, $EU(v_l) > EU(v_r)$. In words, the expected utility of voting for a power seeking leftist is greater than the expected utility of voting for a power seeking rightist, so the voter will again vote for the left party candidate.

[47] Again, the probabilities must sum to 1 and there are only two options. Thus we can use the probability that a power seeker will switch (0.3) to determine the probability that the power seeker will remain faithful: $1 - 0.3 = 0.7$.

[48] If you are wondering why Stokes (2001) would analyze such a simple equation, it turns out that she does not. We made it up as an illustration based on her model, and explain below why she builds her model.

Since this example suggests that the voter's decision is not affected by his beliefs about whether the candidates are ideologues or power seekers, you might be wondering what use Stokes's model has. Had she constructed it for the purpose of determining vote choice it would not have been very interesting (at least not using these values).[49]

10.7.1.1 Expected Utility and Risk Preferences

When we discussed utility functions in Chapter 3 we were really discussing the little u in our expected utility equation. We talked a bit there about what different functional forms for the utility implied substantively, but didn't go into a lot of detail. We can say a little more by bringing expected utility into the mix. Recall our discussion of concave and convex functions in Chapter 8 (or flip there for a moment if you skipped that section or chapter). A function is concave if the secant line joining two points is below the curve, and convex if it is above it. Assume the curve is one's small-u utility function. Equation (10.10) is a linear combination of the utility function u evaluated at several points.

Let's consider two such points for clarity, so that our actor may realize one of two possible utility outcomes. This means that equation (10.10) specifies a point on the secant line joining these two utility outcomes. For a concave function, this secant is below the curve, implying that the expected utility for any lottery over utilities is less than the utility the actor would obtain by receiving with certainty the corresponding combination of the outcomes that produced these utilities. In other words, an actor with a concave utility function prefers the sure thing to the gamble. We call such actors **risk averse**. Conversely, should an actor have a convex utility function, then the secant is above the utility curve, and the actor prefers the gamble to the sure thing. Such actors are said to be **risk seeking**. Finally, if an actor has a linear utility function, then the secant is coincident with the utility function, and the actor is indifferent between the gamble and the sure thing. We call such actors **risk neutral**.

This is a bit complex, particularly in such a small space, but an example will help clarify. Consider the following gamble: you get 0 with probability one-half, and 4 with probability one-half. The expected value of this gamble is $\frac{1}{2}0 + \frac{1}{2}4 = 2$. We'll look at how three different types of people would treat this gamble. First, consider a risk-averse person with the concave utility $u(x) = \sqrt{x}$. Equation (10.10) states that the expected utility for this person is $\frac{1}{2}u(0) + \frac{1}{2}u(4) = \frac{1}{2}0 + \frac{1}{2}2 = 1$. If she were instead to receive with certainty the combination of outcomes that are possible in the lottery (0 and 4), weighted by the same chance that each occurs ($\frac{1}{2}$), then she would receive the expected value of the lottery, 2. Her utility for the expected value of the lottery is $u(2) = \sqrt{2}$,

[49] As an aside, a common exercise in such modeling is to set the expected utility of the options equal to each other and solve for the values of a given variable that make the actor indifferent between the choices. Among other things, this allows computation of what are known as mixed strategy equilibria. Those of you who go on to study formal models and game theory will learn how to do this.

which is *greater* than her expected utility for the lottery itself. Not only would she prefer to get the expected value of the lottery for certain, she'd actually take *less* then the expected value if she could be guaranteed that amount. This is what risk averse means: one is willing to give away potential gains or pay extra to avoid risk. As an example, Feddersen, Sened, and Wright (1990) offer a model of candidate entry that assumes risk aversion.[50]

Next consider a risk-seeking person with the convex utility $u(x) = x^2$. Equation (10.10) states that the expected utility for this person is $\frac{1}{2}u(0) + \frac{1}{2}u(4) = \frac{1}{2}0 + \frac{1}{2}16 = 8$. This is more than her valuation of the expected value of the lottery, $u(2) = 4$; she is so interested in the gamble itself that one would need to pay her to get her to accept the sure thing over the lottery, even though the sure thing here is what the lottery is expected to pay off.

Finally, consider a risk-neutral person with the linear utility $u(x) = x$. Such a person has expected utility equal to her valuation of expected value and so is indifferent between the gamble and the sure thing (e.g., Gradstein, 2006). Risk neutrality is the most common assumption seen in the game theoretic literature in political science for the simple reason that it is easier to deal with mathematically; however, risk aversion is more prevalent substantively, and many models account for this.

10.7.1.2 Why Should I Care?

Expected utility and expected value are central concepts not only in game theory but also in rational choice theory, expected utility theory, and many behavioral and boundedly rational models of politics. Even the theoretical portions of papers and books that are primarily empirical in focus will often use these concepts, and if you want to understand the theory, you will need to understand these concepts.

10.7.2 The Moments of a Distribution

If we replace $u(x)$ with more general functions, equation (10.10) can apply to the expected value of any function of the values of a random variable. One class of such expectations of particular use in statistics is the moments of a distribution. The **moments of a distribution** are an important set of parameters one can use to describe a distribution. They involve the expected values of particular functions of the random variable across the distribution, such that the kth moment of a variable X can be represented as $E[X^k]$, where $E[]$ indicates the expectation of the function of the variable inside the brackets.[51] The expected value of a variable is the sum of the possible values it might take weighted by the probability that each value will turn up, i.e., $E[X]$. The mean (or average)

[50] For a useful introduction to this model, see Gelbach (2013, pp. 16–20).

[51] More explicitly, the kth moment of a variable is the kth derivative of the moment-generating function evaluated at zero. See `http://mathworld.wolfram.com/Moment-GeneratingFunction.html` for more detail.

is the expected value of the variable X^1, and so it is also the first moment of a variable. The first moment is a location parameter and is also one measure of the central value of a distribution.[52]

One can define moments about zero and moments about the mean.[53] The equation for moments about zero of discrete variables is

$$\sum_i x_i^k (Pr(X = x_i)), \tag{10.13}$$

where k is the kth moment about zero. As noted, the first moment (i.e., where $k = 1$ in equation (10.13)) is the central tendency or mean. For a variable, X, that takes with equal probability the values x_i, $i = 1, 2, 3 \ldots N$,[54] the first moment of equation (10.13) is $\frac{1}{N}\sum_{i=1}^{N} x_i^1$. Note that this is the (unweighted) average of the values. Make sure it is apparent to you that when $k = 1$, equation (10.13) produces an average for variable X.

In statistical analyses we are often interested in the second, third, and fourth moments about the mean as they can provide useful information about the scale and shape of a distribution (and thus are known as scale and shape parameters). The second moment about the mean is of interest by itself, and the third and fourth moments about the mean are useful components of other indicators. Moments about the mean are defined by the equation

$$\sum_i (x_i - \mu)^k (Pr(X = x_i)). \tag{10.14}$$

The second moment about the mean (i.e., $k = 2$ in equation (10.14)) is the variance and it measures the variation of the distribution about its mean value.

Two other measures of interest are the skewness and kurtosis of a distribution. Skewness involves the third moment about the mean, and it is usually weighted by the standard deviation, though some people use the third moment without a denominator.[55] A common measure of skewness is the third moment divided

[52] You will learn in your statistics courses that there are three common measures of central tendency: the mean, median, and mode. We focus on the mean here.

[53] The equations are different for discrete and continuous variables, and we focus on discrete variables here. One takes the integral, rather than the sum, for continuous variables.

[54] This is known as a uniform distribution; it is more commonly observed in political science as a continuous distribution and so is covered in the next chapter.

[55] Some authors (e.g., Kmenta, 1986, p. 67) define skewness as the third moment alone. Further, most authors refer to *a* measure of skewness rather than *the* measure. The skewness entry at the MathWorld website offers this observation: “Several types of skewness are defined, the terminology and notation of which are unfortunately rather confusing” (`http://mathworld.wolfram.com/Skewness.html`).

by the standard deviation cubed:[56]

$$E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3}, \tag{10.15}$$

where μ_3 represents the third moment around the mean and σ represents the standard deviation. Skewness measures the symmetry of the distribution about its central value. When skewness is zero, the distribution is symmetric.

Like skewness, kurtosis is used as the label for a number of specific measures.[57] Kurtosis always involves the fourth moment about the mean, and it is often weighted by the standard deviation. A common measure of kurtosis is the fourth moment about the mean divided by the fourth power of the standard deviation:

$$E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{\mu_4}{\sigma^4}. \tag{10.16}$$

The kurtosis measures the flatness or peakedness of the distribution relative to a standard normal distribution.

10.7.2.1 Why Should I Care?

The moments of a distribution are often parameters in the functions one can use to describe them (e.g., the PMF/PDF, CDF, etc.). As such, they show up repeatedly in the study of statistics, and they are often of use when constructing formal models or otherwise trying to discipline one's thinking about a political process. You are likely quite familiar with thinking about the first and second moments of a distribution: in large undergraduate courses you probably paid close attention not only to your own score on a test but also to the average score (i.e., the first moment about zero) and, if your professor made it available, the dispersion or variance of the scores (i.e., the second moment about the mean). And regardless of the extent to which you use formal theory or statistics in your own research, you will need to be familiar with the moments of distributions to do competent grading in large lecture courses.

The third and fourth moments are frequently used to determine whether a given empirical sample deviates from a normal distribution. More generally, using the moments of a sample of data as estimates of the population moments is known as the method of moments in statistics.[58]

[56] The standardized moment is one that is divided by the standard deviation raised to the power of the moment. For example, the standardized second moment is the second moment divided by the standard deviation squared. This measure of skew, then, is the standardized moment. The standardized moments are of interest because they are the moments for a standardized normal distribution (i.e., a normal distribution with a mean of zero and a standard deviation of one). The standardized normal distribution is invoked for a number of hypothesis tests in statistics.

[57] The MathWorld entry observes: "There are several flavors of kurtosis commonly encountered, including the kurtosis proper" (`http://mathworld.wolfram.com/Kurtosis.html`).

[58] The method of moments is the most common way to teach statistics in political science.

10.8 SUMMARY

This chapter argues that political scientists need to think about the (likely) distributions of concepts when developing theories and that a working familiarity of specific distributions is important for (1) developing formal theories of politics and (2) applying statistical inference to hypothesis testing. We introduced several ways one can represent the distribution of a variable (e.g., frequency counts, relative frequency counts, and several functions) and then briefly described several commonly used distributions for discrete variables.

10.9 EXERCISES

1. Write down a research question that interests you. Try to state some assumptions, and then deduce one or more hypotheses from your assumptions. Write them down and bring them all to class.

2. How is the relative frequency distribution different from a frequency distribution?

3. Why can't one create a PDF by plotting the graph of the relative frequency distribution of each value in the sample?

4. What is the difference between a PMF and a CDF?

5. Write down an example where a contingency table would be useful for examining the joint distribution of two variables. Bring it to class.

6. Write down a political process that you think might be drawn from the following discrete distributions: Bernoulli or binomial, Poisson or negative binomial (you should have two political processes).

7. Visit the "Distributions" page of the Virtual Laboratory website at the University of Alabama, Huntsville (`http://www.math.uah.edu/stat/dist/index.xhtml`) and select the "Random Variable Experiment" link under "Applets." Go to the bottom of the Random Variable Experiment applet and select the "Applet" link. Under the label "Bernoulli Trials" you will find applets for the binomial and negative binomial distributions, and under the "Poisson Process" label you will find links to applets for the Poisson distribution (click on "Poisson Experiment"). Investigate the distributions covered in this chapter. More explicitly, select a distribution and note the scale and location of the density function. Adjust one of the parameters using the scroll bar. If there is more than one parameter, adjust it. Write down what happens when you adjust each parameter for the following distributions: Bernoulli, binomial, Poisson, negative binomial.

Wonnacott and Wonnacott (1977) is a good example, and in Chapters 18 and 19 they contrast the method of moments with two other techniques: maximum likelihood estimation and Bayesian inference.

8. Visit the Public Data site at Google (`http://www.google.com/publicdata/`). Select a dataset that is of interest to you (they have many from which to choose). Select the Explore the Data link and plot some univariate distributions (as of this writing, note the options for plotting to the upper right of the page; try different options—you won't break anything). Summarize two things of interest that you learn (or "confirm") by doing so. Now select some other variables that you believe might covary with the first one you selected, and plot some joint bivariate distributions. Again, summarize two or three things of interest that you learn from doing this.

9. If the mean number of wars is three per year, what is the probability that there will be four wars in any given year?

10. A person persuaded a friend to meet her at a concert. Her boss droned on forever at a meeting, and she is running late. To make matters worse, she accidentally dropped her cell phone down the elevator shaft and cannot recall what concert she had said they should attend. She recalls that an orchestra is playing Bach on the north side of town, but a Stravinsky concert is being performed on the west side. She prefers Bach to Stravinsky, such that seeing the former is worth ten units of utility and the latter only five. However, she prefers going with her friend to going alone such that being together yields eight units of utility and being alone yields minus two. We can depict her utility using the following matrix:

		Friend	
		Bach	Stravinsky
Woman	Bach	18	8
	Stravinsky	3	13

    Now assume that she recalls wanting to select the concert her friend would prefer, but she has no idea whether her friend likes Bach better than Stravinsky, and therefore assigns a probability of 0.5 that her friend is waiting for her at the Bach concert. Calculate the expected utility of going to the Bach concert and the expected utility of going to the Stravinsky concert. Which should she choose?

    Now assume that she knows her friend prefers Stravinsky to Bach and assigns the probability 0.3 that her friend is at the Bach concert. Recalculate the expected utilities of her choices. Which concert should she go to now?

11. Two nations, A and B, face off at the brink of war. A knows B is either strong or weak, and that it would win any war with certainty if B were weak, but lose with certainty if B were strong. A has a prior belief of 40% that B is strong, and observes manuevers that a strong B would do 60% of the time but a weak B would do only 30% of the time. If A gets 1 for winning, −1 for losing, and 0 for not starting a war, should A start a war

after observing the maneuvers? (*Hint:* You'll have to use material from Chapter 9 as well.)

10.10 APPENDIX

Our presentation has been relatively informal, and one can find more formal treatments in Gill (2006) and online (e.g., the various MathWorld we entries we noted throughout). Those interested in studying methods as a subfield will want a more thorough treatment. Another place to look is King (1989, chaps. 2 and 3). The National Institute of Standards and Technology *Engineering Statistics Handbook*, section 1.3.6, "Probability Distributions" (`http://www.itl.nist.gov/div898/handbook/eda/section3/eda36.htm`) is also a good source. Finally, Zelterman (2004) provides a thorough discussion of discrete distributions.

Chapter Eleven

Continuous Distributions

In the previous chapter we covered the concepts of random variables and their distributions, but used only discrete distributions in our discussion and examples. We did this to keep a chapter very important for the development of both empirical and theoretical political science free of calculus, for those readers who might want to skip over Part II of the book. However, there is little in the previous chapter specific to discrete distributions. Indeed, as we show below, replacing sums with integrals gets you much of the way toward representing the distributions of continuous random variables.

In this chapter we make this replacement, as well as discuss the few other concepts necessary to get us all the way there. Section 1 tackles this job, and presents the changes to the conceptual edifice we built in the last chapter necessary for understanding the properties of continuous random variables. We also discuss joint distributions here, both empirically and theoretically. Section 2 makes explicit the comparison between discrete and continuous random variables via more complex examples of expected utility than were presented in the previous chapter. We also introduce the uniform distribution, probably the most common one used in applied game theory and one you've undoubtedly seen before in the discrete case. We discuss the notion of stochastic dominance here as well. Finally, Section 3 presents examples of continuous density functions useful for statistical analysis.

11.1 CONTINUOUS RANDOM VARIABLES

In the preceding chapter we limited the discussion to the probability distributions of discrete concepts or variables. In this chapter the focus is continuous concepts and variables. Though this is a bit loose, the difference can be thought of as similar to countability. If you can list each value the random variable can take and assign an integer and a probability to each, then you have a discrete random variable, represented by a discrete distribution. If you can't, you may have a continuous random variable, represented by a continuous distribution.

Most of the same concepts we introduced in the previous chapter for discrete variables and their distributions apply to continuous ones as well. In particular, like the PMF of a discrete variable, the **probability density function** (PDF) of a continuous variable is related to the relative frequency distribution of that variable. More specifically, the PDF is a function that describes the smooth curve that connects the various probabilities of specific (ranges of) values for a sample.

However, there is a difference between the PMF and the PDF, and the terminology we've used hints at it. Note that we said ranges of values, rather than values, and used the word density, rather than mass, in the name. A PDF differs from a PMF in that it does *not* describe the chance that any particular value of the random variable is drawn at random from the distribution. Rather, it describes the *relative likelihood* of drawing any specific value, and the *exact probability* of drawing a value within some range. This is why it is called a density function rather than a mass function: it describes the density of the probability within some range of values that the random variable may take, rather than the explicit "mass" of probability at a particular value.

We unpack this and make it a bit more formal below, but let's first consider why this difference exists. We'll start by assuming that some random variable X can take all the values between 0 and 1, inclusive. In other (fewer) words, $x \in [0, 1]$ for all values x that X might take.[1] Since all random variables have to take some value, the probability that $x \in [0, 1]$ is 1. Now consider the range $x \in [0, 0.5]$. The chance of being here is probably less than 1, so we've reduced the probability from 1. Now shrink it to $x \in [0, 0.25]$. Again, we've likely shrunk the probability of being in that region.[2] If we keep doing that over and over again we keep shrinking the probability. And because X is continuous, there is no point at which we can stop: X is defined over $[0, 0.001], [0, 0.00001]$, and so on, forever. The probability at a point would be the probability at the limit of this shrinking process, but that's ill-defined. So we don't define it, and don't in general speak about probabilities at specific points, even though the PDF will take non-zero values at these points.

This is likely confusing, and may remain so until we look at some examples. That's fine. But it might help to think about it another way. Specifically, another way to think about this is that the PDF is a function that allows one to sum a series of probability weights to produce the likelihood of drawing a value less than some value; the CDF described in the previous chapter gives this likelihood for all points. These weights do not directly correspond to true probabilities of drawing particular points, though; if they did, we would have to sum over an infinite number of finite probabilities to get the CDF at any point (because the range of values is continuous), and that would give an infinite probability. So the weights are instead the relative likelihoods that each value will be randomly drawn from the population. We show below that the PDF makes it possible to identify different probability distributions for continuous variables, and being able to do so turns out to be very important for developing statistical models that can produce valid hypothesis tests (more on why you care below).

[1] Recall that we're using capital letters for the random variables and lowercase letters for the values (realizations) of the random variables.

[2] As with most things in this book, this argument is loose, but the rough idea is what's important.

11.1.1 The PDF of a Continuous Variable

Now that we've discussed the main difference between a PDF and a PMF verbally, let's formalize it. Recall that for a PMF $f(x)$ describing the probability distribution of the random variable X, $Pr(X = x) = f(x)$. The equation is similar but a bit more complex for a PDF $f(x)$. In general, the probability that X takes values in some region B is $Pr(X \in B) = \int_B f(x)dx$. This might be too general, so we consider an application in which the values of the random variable are real numbers, as they typically are in statistical (and most formal) applications in one dimension:

$$Pr(X \in [a,b]) = \int_a^b f(x)dx. \tag{11.1}$$

Recalling Chapter 7, equation (11.1) states that the probability of the variable's being in the range $[a,b]$ is given by the definite integral of the PDF from a to b. As with the PMF, $f(x) \geq 0$ for all PDFs, so this definite integral is the area under the PDF curve between a and b. You might have heard the phrase linking probability and inference to "the area under the curve" in a research methods course; this is the origin of that phrase. Unlike a PMF, though, the PDF function is not limited to taking values no greater than 1, since it does not directly describe probability, only relative likelihood. So, for a PDF, $f(x) \in [0,\infty)$. A value of 0 for the PDF at some point x still means that we can't randomly draw that x, however.

If the probability of being in some interval $[a,b]$ is the integral of the PDF over that region, then we are saying that the probability is computed by summing lots of $f(x)dx$. If we replace the dx with a Δx, we're back to the area under a rectangle, or, in other words, the area under the relative frequency histogram. This is the connection between relative frequency and the PDF. A PDF in this sense is like a smoothed-out histogram.

This covers most of the properties of a PDF of importance for our purposes. But as with a PMF, the probability that some value in the sample space is drawn must be 1, so that $Pr(-\infty < X < \infty) = \int_{-\infty}^{\infty} f(x)dx = 1$.

To this point we have described the PDF only in generic terms. We do this so that you understand where it comes from (especially its connection to the [relative] frequency distribution). In Section 3 below we identify a number of specific PDFs commonly used in political science (and other fields). First, though, we discuss a few more topics: CDFs, parameters, joint distributions, and expectations.

11.1.1.1 Why Should I Care?

Many of the variables that interest political scientists have discrete, not continuous, distributions. Nevertheless, continuous distributions are important. First, some processes of interest—especially measures of time, such as how long a coalition government survives, the timing of sending legislation to the floor for a vote, the length of a war, etc.—are continuous. Second, it turns out that several continuous distributions are remarkably flexible and useful for modeling noncontinuous variables. Further, it is common to assume that various portions of statistical models (e.g., error terms) take a continuous distribution, and a number of statistical hypothesis tests are constructed using continuous distributions. Last, as we demonstrate below, continuous distributions are used in formal theory, and regardless of whether you seek literacy (i.e., the ability to read and understand formal models) or competence (i.e., the ability to create and use formal models), familiarity and comfort with continuous distributions is important. To that end, below we discuss several commonly used continuous distributions and identify their PDFs.

11.1.2 The CDF for a Continuous Variable

Not surprisingly the nettlesome difficulty of an infinite and uncountable number of potential values rears its head again when we think about the CDF of a continuous variable. As with the PDF, the solution lies in thinking about ranges of values instead of discrete ones.

Since we cannot write the CDF for a continuous variable as the sum of the probabilities of each discrete value below the specified value, we have to write it as the sum of all the value ranges below the specified value. Luckily, this is a more straightforward translation from the discrete case. In fact, we merely replace the sum with an integral to get the equation for the CDF:

$$P(X \leq x) = F(x) = \int_{-\infty}^{x} f(t)dt. \tag{11.2}$$

Equation (11.2) states that the probability that a variable drawn randomly from the sample has a value less than or equal to x is the sum of all of the probabilities of all ranges of values less than or equal to x. Because the CDF, unlike the PDF, is a probability, it is constrained to take values between 0 and 1.[3]

Equation (11.2) also introduces, or rather reintroduces, a piece of notation for the CDF: $F(x)$, if the PDF is $f(x)$. This notation is common in both game theory and statistics and arises from the relation between the PDF and the CDF: the latter is the antiderivative of the former.

11.1.3 The Parameters of Continuous Density Functions

Like the PMFs of discrete distributions, the PDFs of many continuous distributions have defined parameters. The most common are the location and scale (dispersion) parameters introduced in the preceding chapter, but some continuous distributions have a **shape parameter**. The shape parameter identifies a point of inflection in a PDF whose graph changes shape. Most distributions

[3] For this reason, CDFs are commonly used to model processes that are constrained between two values, or that involve binary choice.

we use do not have a shape parameter: their central location might change as a function of a parameter and the spread of their values might change as the function of a parameter, but the general shape of the PDF remains the same. Some distributions, however, can also change shape. This is easiest to see in a graph of a distribution that contains a shape parameter, like that of the beta distribution, seen in Figure 11.1.

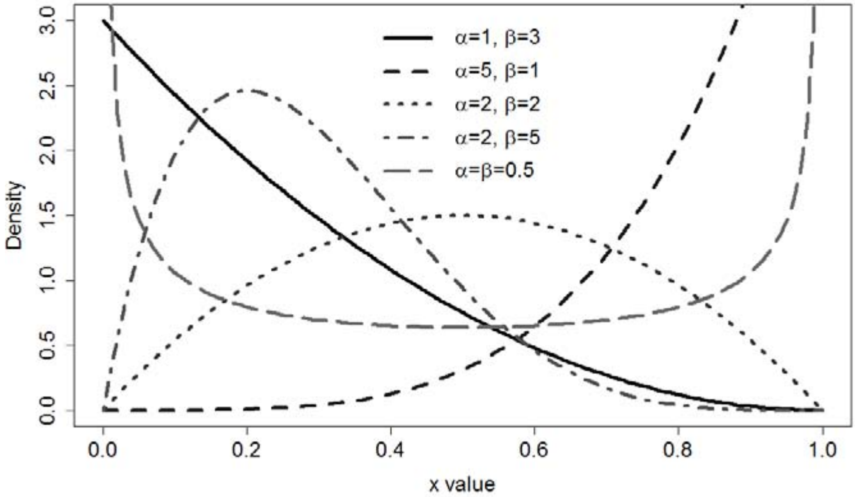

Figure 11.1: Beta PDF with Various Parameter Values

The beta distribution has two parameters, α and β, both of which are shape parameters. As the graph of the PDF demonstrates, the shape of the distribution changes as the parameters change values.

That said, we reiterate that the distributions most frequently used in our discipline do not have shape parameters. As such, you will encounter distributions with location and scale (dispersion) parameters in the political science literature with considerably greater frequency than distributions with a shape parameter.

11.1.4 Joint Distributions

As discussed in the preceding chapter on discrete distributions, we are often interested in the joint distribution of two variables. We discussed empirical joint distributions there. We extend that discussion here for continuous variables, and also introduce theoretical joint distributions of both discrete and theoretical variables.

11.1.4.1 Empirical Joint Distributions

Unlike for discrete variables, contingency tables are not useful for plotting the joint distribution of continuous variables. The tabular (or matrix) format of the contingency table limits its usefulness for looking at the joint distribution of continuous variables (or integer variables with more than a handful of values). The problem is that there are too many potential values that the variable may take. Thus, when we want to examine the joint distribution of continuous variables or one discrete and one continuous variable, rather than a contingency table we use a **scatter plot**.

A scatter plot is a graph with one variable's values listed on the vertical axis (typically the dependent or caused variable) and the other variable's values listed on the horizontal axis (typically the independent or causal variable). As examples, we have produced two scatter plots.

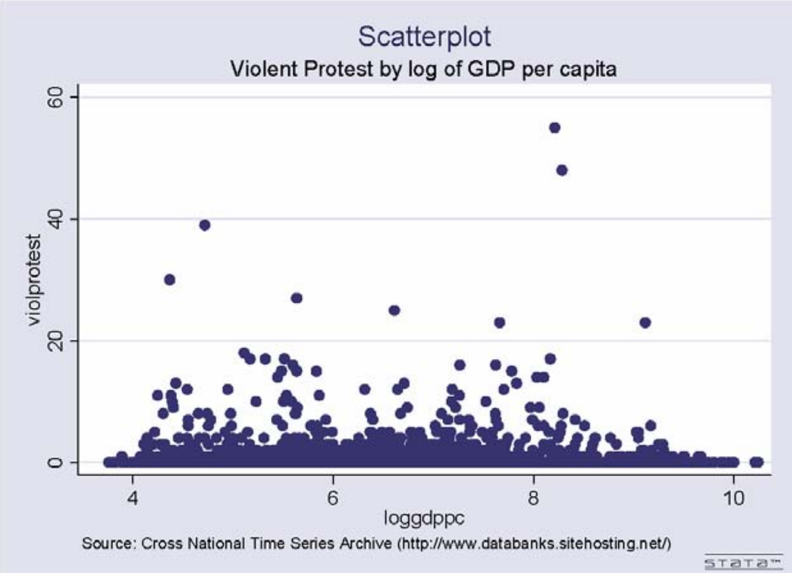

Figure 11.2: Is Violent Protest Related to Macroeconomic Output?

The first plot, in Figure 11.2, is composed of one discrete and one continuous variable. It seems to indicate a slight positive relationship between the size of the economy and the number of violent protest events.

The second plot, in Figure 11.3, depicts two continuous variables and suggests that the number of votes cast in national parliamentary elections is not strongly related to the size of government expenditures.

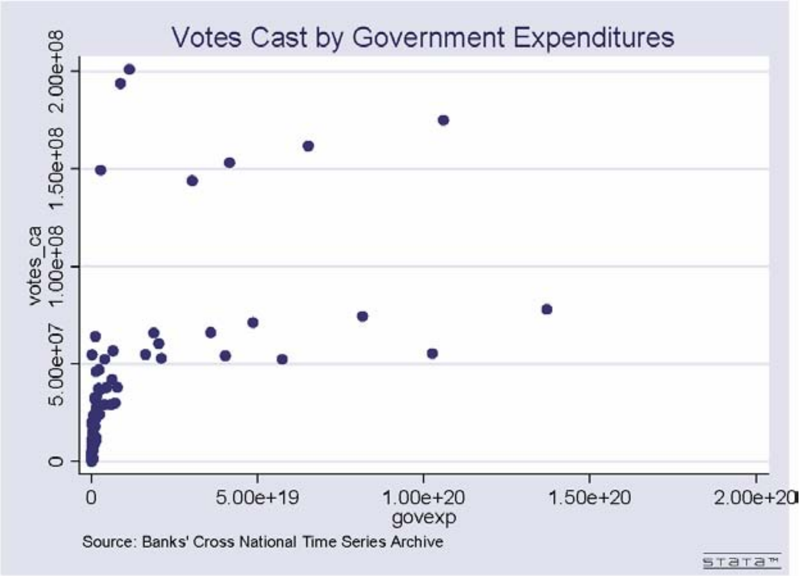

Figure 11.3: Is the Size of the Popular Vote Related to Government Expenditure?

11.1.4.2 Theoretical Joint Distributions

Whenever the values on the y-axes of these scatter plots are correlated with those on the x-axes, we might think there is a relationship between the two variables represented on the axes. Since we think of the variables on both axes as random variables, we can describe the joint variation of the variables, whether or not there is a conditional relationship between them, with joint probability distributions. These are theoretical constructs that describe the *simultaneous* realizations of more than one random variable. We stick to examining two random variables here, but everything we write here can be generalized. We also discuss joint distributions for discrete and continuous random variables at the same time.[4]

Joint distributions merely describe probabilities of more than one outcome at once. For two discrete random variables, we can write their joint PMF as $f(x, y) = Pr(X = x \cap Y = y)$. This is simply the chance that both x and y are simultaneously realized. If $x = y = 1$ and X and Y correspond to the values one might roll on two dice, the joint probability is the chance that two ones are rolled, or $\frac{1}{36}$.

For two continuous random variables, we can write their joint PDF the same

[4] We include discrete joint distributions in this chapter rather than in the previous chapter because the continuous version is more commonly observed, and the previous chapter is already relatively lengthy. Also, the logic behind joint discrete outcomes was provided in Chapter 9.

way: $f(x, y)$. "Summing" the small bits of probability $f(x, y)dxdy$ over some region $X \in A, Y \in B$ produces the probability $Pr(X \in A \cap Y \in B)$.

Thus, $f(x, y)$ is a way of writing the probability that two things occur simultaneously. As in Chapter 9, we can expand this "and" statement. For both discrete and continuous distributions, if the random variables X and Y are independent, then $f(x, y) = f(x)f(y)$. However, if X is conditional on Y (or vice versa), then $f(x, y) = f_{X|Y}(x|y)f_Y(y)$ (or with x and y and X and Y switched). Here the subscripts on the PDFs make clear the nature of the distribution. The function $f_Y(y)$ is the marginal distribution of the random variable Y, which averages over X. For the continuous distribution, this means that $f_Y(y) = \int f(x, y)dx$. The conditional distribution of X is given by $f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)}$, which is nothing more than the equation above rearranged.[5]

As with a single variable, something must happen, so that $\int\int f(x, y)dxdy = 1$ for the continuous case and $\sum_i \sum_j f(x_i, y_j) = 1$ for the discrete case. The double integrals (or sums) tell us to integrate (or sum) over first one variable and then the other; we discuss this more in Part V of the book. These same integrals and sums allow us to compute the joint CDFs: $Pr(X \leq x \cap Y \leq y)$ is $F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(x', y')dx'dy'$ for the continuous case[6] and $F(x, y) = \sum_{x_i<x} \sum_{y_i<y} f(x_i, y_i)$ for the discrete case.

This may have seemed complicated, but most of the complication is notational. Basically, joint probability distributions work in the same way as those for a single random variable: they tell you either the probability of two events happening at the same time (for two discrete random variables) or the relative likelihood of two events happening at the same time (for two continuous random variables). When the distributions for each variable are independent you can treat them entirely separately, but when one is conditional on the other you cannot do so. This is no different from how one treats any conditional probability, as we discussed in Chapter 9.

11.2 EXPECTATIONS OF CONTINUOUS RANDOM VARIABLES

In the previous chapter we discussed expectations of random variables. This discussion was general, but all our examples were of, and all our equations were for, discrete random variables. Here we discuss expectations of continuous random variables. Most of what we said in Chapter 10 holds here as well. In fact, all we're going to do is rewrite some of the equations used in the previous chapter for the continuous case, using the substitutions (integrals for sums, etc.) we introduced in the previous section. We'll start by presenting expectations

[5] This rearrangement should look familiar from our discussion of Bayes' rule in Chapter 9.

[6] You will often see integrals over x' and y' when the bounds of the integral contain x or y. Since it doesn't matter what letter or expression we use when we integrate over it (recall Chapter 7), using the "primed" versions of x and y enables us keep track of which integral corresponds to which bounds.

in general, then discuss expected utility and moments of distributions in subsequent subsections. The one new concept here is the uniform distribution, which is commonly used in game theory (and which serves as an uninformative prior in Bayesian statistics); we present this when discussing expected utility.

Recall from the previous chapter that we write the expectation of a random variable X as $E_X[X]$, or $E[X]$ when there is no confusion about other variables. In words, the expectation is the weighted average of the values that a variable can take, where the weights are given by the probability distribution of X. Also recall that the equation for an expectation is $E_X[X] = \sum_i x_i(Pr(X = x_i))$. So you multiply each value x_i by the weight on that value, and add all these up. Since the PMF $(f(x_i))$ of a discrete distribution provides the relevant probability weights $Pr(X = x_i)$, we can also write the expectation as $E_X[X] = \sum_i x_i f(x_i)$. This is for a discrete random variable; for a continuous one we translate the sum to an integral and the PMF to a PDF:

$$E_X[X] = \int_{-\infty}^{\infty} xf(x)dx. \tag{11.3}$$

The bounds on the integral ensure that the expectation includes all possible values of X that might be drawn. Equation (11.3) has the same interpretation as for the discrete case: it's a weighted average of the values of a continuous random variable, and so provides the mean of the distribution.

11.2.1 Expected Utility

As we noted in the previous chapter, we need not limit ourselves to the expectation of the variable itself; we can also consider functions of that variable. When we call these functions $u(x)$, we get expected utility $EU(X) = \sum_i u(x_i)(Pr(X = x_i))$ in the discrete case. Again, replacing the probability with the PMF produces the equation $EU(X) = \sum_i u(x_i)f(x_i)$. And changing to an integral and to the PDF provides the expected utility for the continuous case:

$$EU(X) = \int_{-\infty}^{\infty} u(x)f(x)dx. \tag{11.4}$$

Again, the bounds on the integral ensure that the expectation includes all possible values of X, and so all possible $u(x)$ that might be drawn. Equation (11.4) also has the same interpretation as for the discrete case: it's the weighted average utility one can expect to get, given the probability distribution of the random variable. Everything we said in the previous chapter about risk preferences and why one would need to understand and be able to compute expected utilities continues to be true in this case as well, and we won't repeat it. Instead we'll move right to an example. To do this, though, we'll first introduce the uniform distribution.

11.2.2 The Uniform Distribution

You've undoubtedly seen the uniform distribution before, even if you haven't heard it called that. It is the distribution that assigns equal probability or likelihood to all possible events in the sample space. The discrete case is so straightforward that we didn't even bother mentioning it in the previous chapter. If there are n possible outcomes, then the uniform distribution assigns probability $\frac{1}{n}$ to each outcome. For example, there are two outcomes in a coin flip, so the probability of getting either heads or tails is $\frac{1}{2}$, while there are six outcomes in the roll of a (fair) die, so the probability of getting any number between one and six is $\frac{1}{6}$.

Though this discrete distribution applies commonly outside the social sciences, it is not often used in political science. Few empirical scenarios place equal weight on every possible event. In game theory, whenever the probability is discrete, one typically wants to let it vary as a parameter, and so one assigns probabilities p_i to all events $i \in 1, \ldots, n$, as in the examples we used in the last chapter.[7] And even when the chances of all events are equal, the argument is more typically made in the context of classical probability, as in Chapter 9, than with a uniform distribution.

The continuous uniform distribution, however, is commonly used in game theory, and also as an uninformative prior in Bayesian statistics. We show below that the form of its PDF is somewhat more complicated than $\frac{1}{n}$, yet both the discrete and the continuous uniform distribution share the same fundamental property: the chance of drawing any value is the same.

To get a continuous PDF that satisfies this, let's begin with the easiest option and make the PDF constant at 1 throughout some range.[8] Let's call this range the interval $[\alpha, \beta]$. Then we could let the PDF be 1 from α to β, and 0 for all other values of X. It turns out this is almost good enough. The only problem is that when you integrate the PDF over all X, you need to get 1, and here you don't. Instead you get $\int_{-\infty}^{\infty} f(x) \cdot dx = \int_{\alpha}^{\beta} 1 \cdot dx = x|_{\alpha}^{\beta} = \beta - \alpha$.[9] But this is a constant, so we can readily fix this problem: we divide the PDF by $\frac{1}{\beta-\alpha}$ to cancel out the $\beta - \alpha$ and leave the integral as 1.[10] This yields the expression for the PDF of the uniform distribution:

$$f(x) = \begin{cases} \frac{1}{\beta-\alpha} & \text{if} \quad x \in [\alpha, \beta], \\ 0 & \text{otherwise.} \end{cases} \tag{11.5}$$

Note that α is a location parameter (i.e., it determines the center of the

[7] Some researchers will also assume a discrete uniform distribution as an uninformative prior in Bayesian statistical models.

[8] We need a finite range here, since otherwise we'd have to make an infinite number of subranges equally likely, so that the chance of getting any one of them (such as $[0, 1]$) would effectively be zero.

[9] The second step is true because the PDF is 0 outside $[\alpha, \beta]$, so the definite integral is 0 outside this range.

[10] The $\frac{1}{\beta-\alpha}$ is known as a normalization constant (or factor) for this reason.

distribution), and β is a scale parameter (i.e., it determines the dispersion of the distribution). If X is distributed according to a uniform distribution with these parameters we can write $X \sim U[\alpha, \beta]$. Figure 11.4 plots the PDF of the uniform distribution.

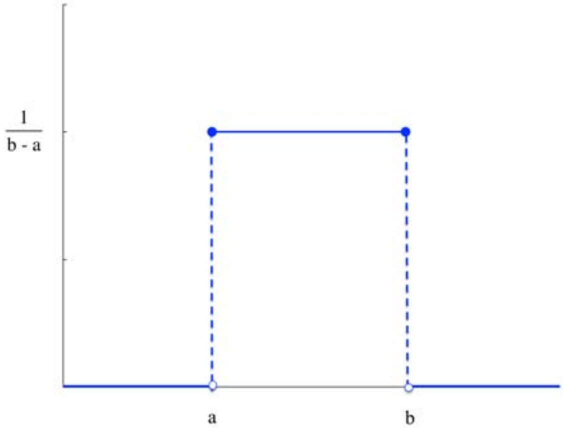

Figure 11.4: Uniform PDF

We are generally interested not in the uniform distribution's PDF but rather in its CDF. Recall that the CDF is the integral of the PDF from negative infinity up to some value x. For the uniform distribution, this function is $F(x) = \int_{-\infty}^{x} \frac{1}{\beta-\alpha} dt = \frac{1}{\beta-\alpha} \int_{\alpha}^{x} dt = \frac{1}{\beta-\alpha} t|_{\alpha}^{x} = \frac{x-\alpha}{\beta-\alpha}$ for any $x \in [\alpha, \beta]$. For smaller values of x the CDF is 0, and for larger values it is 1, since there is no chance of drawing an x less than α or more than β. Putting this together produces the CDF of the uniform distribution:

$$F(x) = \begin{cases} 0 & \text{if} \quad x < \alpha, \\ \frac{x-\alpha}{\beta-\alpha} & \text{if} \quad x \in [\alpha, \beta], \\ 1 & \text{if} \quad x > \beta. \end{cases} \tag{11.6}$$

The important thing to note about this CDF is that it is linear in x. Since the CDF represents the chance of drawing a value from the distribution less than or equal to x, the linearity of the CDF means that this chance increases proportionally with x. We illustrate this in Figure 11.5.

11.2.3 A Game Theoretic Example

In game theory, we often need to know the probability that a given variable has a value below some cutoff point. For instance, we might want to know what the chance is that one's expected utility for things like contesting an election or launching a military expedition would exceed the payoff from taking the alternative option (e.g., doing nothing).

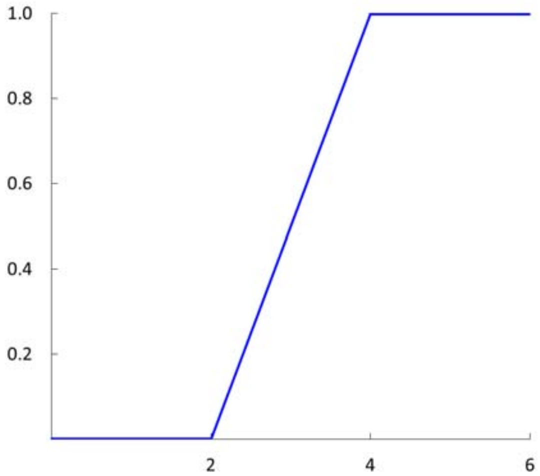

Figure 11.5: Uniform CDF

If one's utility is a function of a random variable, the CDF of the distribution of that variable determines the probability that any realization of the variable is less than some value. If the distribution is continuous, we need to integrate the PDF to get the CDF. In most cases this can be difficult, particularly if we care about obtaining closed-form solutions of the game (i.e., one can write down the equation for the answer). The simplicity of the uniform PDF allows us to readily compute the CDF, and its linearity implies that subsequent calculations will be easier to deal with than would be the case for nearly all alternative distributions.

Further, in game theory we often don't have strong beliefs about the distribution of some parameter of the model, and the uniform distribution allows one to assume no preference for any particular value in the distribution. For both these reasons, game theorists typically assume the uniform distribution when a particular distribution must be chosen.[11]

To see how this works, we'll consider an example of the type you may very well see in a game theory class in political science. Flip back to (or merely recall) Chapter 3, wherein we used as an example of a utility function the quadratic loss function $u(x) = -(x-z)^2$, where z is the ideal policy, x is the enacted policy, and the utility function indicates that z is the most preferred policy, with policies less preferred the further they are from z. We said there that this was a common form of utility for modeling voting behavior. Having now completed Chapter 8, we might recognize that the reason for this is that this utility function is

[11] Of course, it is always better *not* to have to choose a distribution, as then your model doesn't rely on an assumption (of the uniform distribution) that may be wrong, but this option isn't always optimal!

both differentiable and concave everywhere, and so has a maximum at the ideal policy $x = z$ that can be computed comparatively easily.

Assume there exists a pivotal voter who will determine the outcome of an election or the vote in a legislature,[12] and assume that she has a quadratic loss utility function with $z = m$. This means that the pivotal (aka median) voter prefers policies closer to m than those further away. However, while we can assume that the median voter knows her own most preferred option (aka ideal point) m, no one else does. In particular, neither of the two candidates contesting an election does. The candidates do realize, however, that in order to win, they must secure her vote.

Let's say that the policy space—a line over which all policies can be aligned in order, such as a left–right continuum—is $[0, 1]$, which means these are all the available policies from which to choose. Let's also say that neither candidate has any clue where the median voter's ideal point might be within this range. We represent this cluelessness formally by saying $M \sim U[0, 1]$, where M is the random variable corresponding to the median voter's ideal point.

Game theorists are often interested in where the candidate faced with such circumstances will place policy along the continuum. If the candidates knew where m was located, they would select m, as doing so is the best way to win the election. However, they do not know the location of m. This is often referred to as a location game, and it has two or more candidates each choosing a position in the policy space for their platforms. They run on these platforms, then voters (in this case, the median voter only) vote for the candidate they like best, and the winning candidate (the one with the most votes) typically must enact her platform.[13]

In the real game the candidates either enter simultaneously or in some sequence. For our purposes, we'll simplify things. Let's say that there is already an incumbent in office—Sunhee—who has cleverly staked out the position $x_A = \frac{1}{2}$. Sunhee reasons, quite correctly, that occupying the mean of the distribution of M gives her the best chance of winning, which is what her primary interest is in the election. Ryan seeks to challenge Sunhee, but unlike Sunhee he is extreme in his views and only cares about enacting a far left policy of 0, or as close to that as he can get. Specifically, his utility is $u_B = -(x - 0)^2$. If he wins the election with platform x_B, Ryan gets utility $-x_B^2$, and if he loses the election he gets utility $-(\frac{1}{2} - 1)^2 = -\frac{1}{4}$, as Sunhee's platform of $\frac{1}{2}$ gets enacted.

To figure out where Ryan should locate his platform, we need to maximize

[12] In rational choice theories one often appeals to what is called the median voter theorem, which, speaking loosely, says that the voter whose ideal policy in one dimension is in the middle (the median) of all voters' ideal policies is a pivotal voter who determines the outcome of the vote. See Shepsle and Bonchek (1997) for an introduction.

[13] This ignores what are called credible commitment problems (e.g., the candidate once in office can do what she wants), but we need not concern ourselves with this here. There is, however, quite a broad literature on this topic (e.g., McCarty and Rothenberg, 1996), and our discussion of Stokes (2001) in previous chapters is one example.

the expected utility

$$EU(x_B) = Pr(win|x_B)(-x_B^2) + Pr(lose|x_B)(-\frac{1}{4}). \tag{11.7}$$

In Chapter 8 we learned how to maximize this, but before doing so we need to know $Pr(win|x_B)$. (Note that $Pr(lose|x_B) = 1 - Pr(win|x_B)$ since there are only two candidates.) But how does one go about finding this?

Consider the median voter. Her utility function implies she will always vote for the candidate closer to her. So when is she closer to Sunhee's platform of one-half, and when is she closer to Ryan's platform of x_B? Well, the midpoint of x_B and one-half is $\frac{x_B+\frac{1}{2}}{2} = \frac{x_B}{2} + \frac{1}{4}$. Assuming (safely) that Ryan locates to the left of Sunhee, whenever the realized m is less than this midpoint, Ryan wins, because the median voter's ideal point is closer to his position than to Sunhee's, and whenever the realized m is greater than this, Sunhee wins. Stated mathematically, $Pr(win|x_B) = Pr\left(m \leq \frac{x_B}{2} + \frac{1}{4}\right)$.

This probability is the CDF of M evaluated at $\frac{x_B}{2} + \frac{1}{4} \in [0, 1]$. For $U[0, 1]$, $\alpha = 0$ and $\beta = 1$, so the CDF is $F(m) = m$. Thus, $F(\frac{x_B}{2} + \frac{1}{4}) = \frac{x_B}{2} + \frac{1}{4}$. The probability of the median voter's ideal point being less than $\frac{x_B}{2} + \frac{1}{4}$ is therefore equal to $\frac{x_B}{2} + \frac{1}{4}$, and this is the probability that Ryan wins. Plugging this into equation (11.7) yields $EU(x_B) = -x_B^2\left(\frac{x_B}{2} + \frac{1}{4}\right) - \frac{1}{4}\left(1 - \frac{x_B}{2} - \frac{1}{4}\right)$. Simplifying gives $EU(x_B) = -\frac{x_B^3}{2} - \frac{x_B^2}{4} + \frac{x_B}{8} - \frac{3}{16}$.

Maximizing this utility entails first taking the first-order condition (see Chapter 8), which is $-\frac{3x_B^2}{2} - \frac{x_B}{2} + \frac{1}{8} = 0$. We can multiply through by -8 to get rid of the fractions, which yields $12x_B^2 + 4x_B - 1 = 0$. Then we use the tools of Chapter 2 to solve this; e.g., the quadratic equation gives us $\frac{-4\pm\sqrt{16+48}}{24} = \frac{-1\pm 2}{6}$. Only one of these is in the range $[0, 1]$, so we'll choose that one. This gives the candidate extremum $x_B^* = \frac{1}{6}$.

We next check whether this is a local maximum by computing the second-order condition at that point. This is $-3x_B^* - \frac{1}{2} < 0$, so x_B^* is a local maximum. Finally, we compare the utility at this point to that at the bounds. At $x_B = 0$, we get $EU(0) = -\frac{3}{16}$. At $x_B = \frac{1}{2}$, which is the furthest Ryan can go and still be to the left of Sunhee, we get $EU(\frac{1}{2}) = -\frac{4}{16}$. Finally, at $x_B^* = \frac{1}{6}$, we get $EU(x_B^*) = -\frac{1}{432} - \frac{1}{144} + \frac{1}{48} - \frac{3}{16} = -\frac{19}{108}$, which is the biggest of the three values. So the global maximum occurs at $x_B^* = \frac{1}{6}$.

Ryan thus locates at a position considerably to the left of one-half, and so accepts that he will lose more often than Sunhee; he's just willing to take the extra risk of losing so as to enact his ideal policy when he wins.

This was a pretty involved example (though it is a simplification of Calvert (1985)), but it illustrated, we hope, the way in which the CDF can be used in game theory. For those interested, we discuss this at a bit higher level in the last part of this section, which introduces the notion of stochastic dominance. Before getting there, though, we briefly illustrate the moments of continuous distributions, again using the uniform distribution as an example.

11.2.4 Moments of Continuous Distributions

In the preceding chapter we discussed moments of distributions and why they are important; since that discussion continues to apply to the case of continuous distributions we do not repeat it. Rather, we will merely present the definitions for the kth moment in the continuous case. The kth moment about zero is

$$\int_{-\infty}^{\infty} x^k f(x)dx, \tag{11.8}$$

and the kth moment about the mean is

$$\int_{-\infty}^{\infty} (x-\mu)^k f(x)dx. \tag{11.9}$$

We can see how this works with the uniform distribution. First we compute its first moment, the mean μ. This is

$$\begin{aligned} \mu &= \int_{-\infty}^{\infty} xf(x)dx \\ &= \frac{1}{\beta-\alpha}\int_{\alpha}^{\beta} xdx \\ &= \frac{1}{\beta-\alpha}\frac{1}{2}x^2|_{\alpha}^{\beta} \\ &= \frac{\beta^2-\alpha^2}{2(\beta-\alpha)} \\ &= \frac{\beta+\alpha}{2}. \end{aligned} \tag{11.10}$$

You may recognize this as the midpoint of the line segment $[\alpha, \beta]$. When $\alpha = 0$ and $\beta = 1$, $\mu = \frac{1}{2}$.

We can also compute the variance, which is the second moment about the mean. This is

$$\begin{aligned} \sigma^2 &= \int_{-\infty}^{\infty}(x-\mu)^2 f(x)dx \\ &= \frac{1}{\beta-\alpha}\int_{\alpha}^{\beta}(x^2-2\mu x+\mu^2)dx \\ &= \frac{1}{\beta-\alpha}\left(\frac{1}{3}x^3-\mu x^2+\mu^2 x\right)|_{\alpha}^{\beta} \\ &= \frac{(\beta^3-\alpha^3)/3-\mu(\beta^2-\alpha^2)+\mu^2(\beta-\alpha)}{\beta-\alpha} \\ &= (\beta^2+\alpha\beta+\alpha^2)/3-(\beta^2+\alpha^2+2\alpha\beta)/2+(\beta^2+\alpha^2+2\alpha\beta)/4 \\ &= (\beta^2+\alpha\beta+\alpha^2)/3-(\beta^2+\alpha^2+2\alpha\beta)/4 \\ &= \frac{\beta^2+\alpha^2-2\alpha\beta}{12} \\ &= \frac{(\beta-\alpha)^2}{12}. \end{aligned} \tag{11.11}$$

When $\alpha = 0$ and $\beta = 1$, $\sigma^2 = \frac{1}{12}$. Higher moments may be computed similarly. The same procedure may be followed for other, more complex distributions, though their integrals are likely to be more difficult.

We can also use our earlier discussion of theoretical joint distributions to compute a different sort of second moment that will prove convenient for work in statistics. When there are two variables distributed jointly we can discuss their **covariance**. Like the variance, the covariance considers variation around the mean, but now around the means of two variables. It is computed according to the equation $\sigma_{xy} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_x)(y - \mu_y) f(x, y) dx dy$, with μ_x and μ_y the means of the random variables X and Y, respectively.

The form of the covariance implies that when both variables exceed their means or both are below their means the integrand is positive, while if one is above and one is below its mean the integrand is negative. Thus, the covariance measures the degree to which two random variables "move together" in their joint distribution. If one often tends to be large (small) when the other is large (small), then their covariance will be positive, while if one tends to be large while the other is small, then their covariance will be negative. A covariance of zero implies that the two variables are not correlated in this fashion; this often happens when the two variables are drawn from independent distributions, but this is not necessary for a covariance of zero.

Like the variance, the covariance can become large in magnitude. When interested in the relative degree of correlation between two variables, we can instead form a **correlation coefficient**. This is computed from the covariance and the variances of each variable and varies between -1 and 1. A 0 means no correlation, and a 1 (-1) means perfect positive (negative) correlation. It has the form $\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$, where the two components of the denominator are the standard deviations (the square roots of the variances) of X and Y, respectively.

11.2.5 Stochastic Dominance

In game theory and expected utility theory, expected utilities for actions are compared together and the action that produces the highest expected utility is chosen. It is natural to compare payoffs for individual outcomes, but when actions produce stochastic payoffs one can also compare the distributions directly. This leads to the important notion of stochastic dominance.

Before giving the formal definitions, let's consider an intuitive example. Consider a type of lottery in which you can receive either nothing or a million dollars. Now compare two specific lotteries, one in which you have a 0.0000000001 chance of getting the million dollars and one in which you have a 0.3 chance of getting the million dollars. Which one would you prefer?

The answer is obvious, of course, but it does illustrate the concept of explicitly comparing probability distributions to each other. The notion of stochastic dominance formalizes this comparison. If we let $f(x)$ and $g(x)$ be two different PDFs, then we say $f(x)$ first-order stochastically dominates (FOSD) $g(x)$ if their CDFs obey this relation: $F(x) \leq G(x)$ for all x.

This is very abstract, so we break it down. The CDF tells you the chance of drawing a value from that distribution below a certain value x. If the CDF for $g(x)$ is always greater than that for $f(x)$ for all x, then the chance of drawing a lower value from the distribution is always greater in $g(x)$ than in $f(x)$. This implies, since $1 - F(x) \geq 1 - G(x)$, that the chance of drawing a value higher than x is always greater in $f(x)$ than in $g(x)$ for all x. So, in a sense, $f(x)$ can be expected to produce higher values than $g(x)$, and we say the former dominates the latter. For our example this is true: the first lottery is more likely to produce the lesser value than the second, and less likely to produce the higher value. So the second FOSD the first.

How does this relate to preference? Another way of writing $f(x)$ FOSD $g(x)$ is $\int_{-\infty}^{\infty} u(x)f(x)dx \geq \int_{-\infty}^{\infty} u(x)g(x)dx$ for all increasing functions $u(x)$.[14] In other words, if you place higher value on obtaining greater levels of some random variable X, perhaps because it corresponds to revenue, shares in a government's cabinet, or your piece of the division of land in a cease-fire bargain, then you always prefer that the distribution $f(x)$ be the one that determines levels of revenue, cabinet shares, or land distributions, as opposed to $g(x)$. And this is for the reason we stated above: it is more likely to produce higher values of these things.

FOSD is thus a useful concept because it lets you state preference over distributions without having to go to the trouble of figuring out expected utilities; you can just compare CDFs. Further, as it works for any increasing utility function, one need not even specify a particular function. This is particularly useful when employing techniques in game theory such as monotone comparative statics (e.g., Ashworth and Bueno de Mesquita, 2005).

It does, however, require a pretty strong assumption on the distributions that may be hard to justify substantively in some cases. We can also define a lesser form of dominance, second-order stochastic dominance (SOSD). It has a very similar definition: $f(x)$ SOSD $g(x)$ if $\int_{-\infty}^{\infty} u(x)f(x)dx \geq \int_{-\infty}^{\infty} u(x)g(x)dx$ for all increasing *concave* functions $u(x)$. Recalling from the previous chapter that concave utility functions represent risk-averse actors, if $f(x)$ SOSD $g(x)$ then it is preferred by all risk-averse individuals.

We can compare two uniform distributions to illustrate these concepts. Let $f(x) \sim U[1,3]$ and $g(x) \sim U[0,2]$. Then $f(x)$ FOSD $g(x)$. *Anyone* who wants higher values of x prefers $f(x)$, because it yields consistently higher values of x. Now let $f(x) \sim U[1,3]$ and $g(x) \sim U[0,4]$. Then $f(x)$ SOSD $g(x)$, but not FOSD. A risk-neutral person (i.e., one with a linear utility function) is indifferent between two lotteries represented by these distributions; they each have the same mean value, and the potential for higher values in $g(x)$ is balanced out by the potential for lower values. However, a risk-averse person prefers the first lottery to the second because she is less likely to draw a really low value.

11.3 IMPORTANT CONTINUOUS DISTRIBUTIONS FOR STATISTICAL MODELING

In this section we introduce several commonly used continuous distributions: the Gaussian family, the logistic distribution, some duration distributions, and three distributions used frequently in statistical hypothesis tests. The Gaussian family includes the normal distribution and the power transformed normal

[14] An equivalent equation holds for discrete distributions, using sums instead of integrals: $\sum_i u(x_i)f(x_i) \geq \sum_i u(x_i)g(x_i)$. Note that if we were to call $u(x)$ utility, then both formulations express expected utility under different probability distributions.

distribution. Duration distributions include the exponential distribution, the Pareto distribution, the gamma distribution, and the Weibull distribution. Finally, we review the chi squared (χ^2), the F, and the (Student's) t distributions.

11.3.1 The Gaussian Family

Families of distributions have the same basic parameter structure. The Gaussian distribution is named after one of the first scholars to use it, Johann Carl Friedrich Gauss.

11.3.1.1 The Normal Distribution

The normal distribution is the best known of all continuous distributions. It may be written as $N(\mu, \sigma^2)$, so that if X is distributed normally, $X \sim N(\mu, \sigma^2)$. As seen in this notation, the distribution admits two parameters, the mean (or average) value, represented by μ, and the variance (or dispersion) of values around the mean, represented by σ^2. The PDF of the normal distribution is given by

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}. \tag{11.12}$$

Note that while π is often used as a symbol to indicate the probability of observing an event, in equation (11.12) it represents the value 3.14159.... The normal distribution is so commonly used, particularly the standard normal, in which $\mu = 0$ and $\sigma^2 = 1$, that the standard normal PDF has its own symbol: $\phi(x)$. The standard normal CDF is denoted $\Phi(x)$. We can use equation (11.12) to graph some normal distributions, and we have done so in Figure 11.6.

One interesting (and unusual) property of the normal distribution is that the parameters (μ and σ^2) are independent of one another.[15] The mean (μ) determines the central location of the distribution and the variance (σ^2) determines the scale of the distribution. A second property of interest is the symmetry of the normal distribution: the graph of the function to the right of the mean is the mirror image of the graph of the function to the left of the mean.[16]

You have likely heard that the normal distribution has a bell-shaped curve. This is true for many (but as Figure 11.6 demonstrates, not all) values of μ and σ^2, but it is not very meaningful. We have already seen a discrete distribution (the binomial) that frequently produces a bell-shaped curve, and we introduce many others below. It follows that one cannot draw the inference that sample data with a bell-shaped relative frequency distribution were drawn from a normal distribution. In fact, it is a good idea to unlearn the habit of referring to distributions as bell-shaped, as that observation provides precious little information about the distribution. In your statistics courses you will learn some

[15] Note that we could have replaced "μ and σ^2" with "the first and second moments." You will sometimes encounter that usage.

[16] Its skewness is zero.

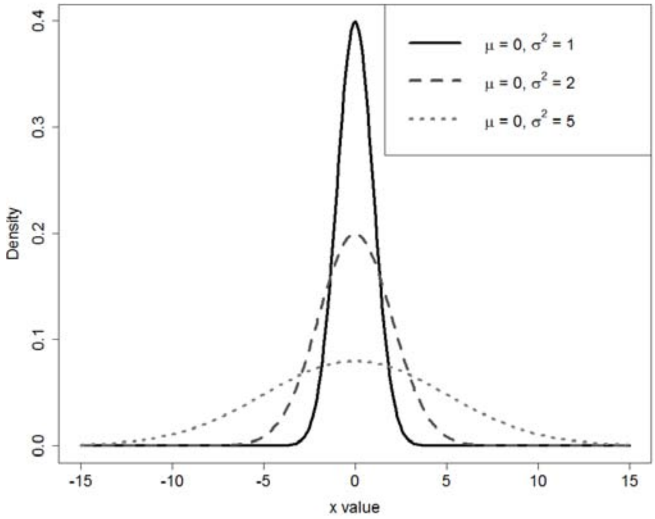

Figure 11.6: Three Normal PDFs, μ, $\sigma^2 = 0, 1; 0, 3; 0, 10$

formal tests one can conduct to determine the probability that a given sample of data was drawn from a normal distribution (e.g., the Jarque-Bera test).

What kinds of processes are likely to produce a normal distribution? Lindsey (1995, p. 113) observes that the "normal distribution describes a continuous response variable, taking any real value, positive or negative, which is the result of a large number of small accumulating, unknown, additive factors." We suspect that description does not strike you as likely to represent the process that produces a majority of the variables political scientists want to explain.

To better understand the limits of the usefulness of the normal distribution in empirical political science, try this as an exercise. Compile a list of variables that come to mind that political scientists use to measure concepts in their theories. How many of them are continuous measures (especially those with both negative and positive values)? Our expectation is that your list will be dominated by integer variables (e.g., the number of seats in parliament or the number of militarized disputes) and ordinal or nominal variables (e.g., an attitudinal scale or party identification).

This does not mean that the normal distribution is useless in statistics. Far from it! In fact, the central limit theorem states that in sufficiently large samples, sampling distributions approximate the normal distribution. We note that the central limit theorem does not imply that the normal distribution is the most appropriate distribution available. However, it does suggest that if one does not have a positive case to make for why the concept for which one has developed hypotheses fits a particular distribution, then the normal distribution is the best choice (though a better decision would be to go back to develop a stronger theory and develop a positive case for the likely distribution of one's concept).

That said, as Lindsey (1995, p. 113) observes, "the normal distribution is primarily important for its nice mathematical properties, and is much overused in many areas of research for this reason." To elaborate, owing to the nice mathematical properties of the normal distribution, it was relatively easier to develop techniques (and, later, software) for inferential statistics assuming a normal distribution than it was to assume other distributions. As such, the practice and teaching of applied statistical work focused on models that invoked the normal distribution. However, the past forty years have witnessed a dramatic increase in computing power, and software that can implement models that invoke different distributional assumptions has become commonplace. Thus, while it is important to use models that invoke a normal distribution when using a dependent variable that is normally distributed, the general point is that *it is important to use a model that invokes the appropriate distribution.* Political scientists became widely aware of this in the 1990s, and the overuse of models that assume a normal distribution has been declining ever since (Krueger and Lewis-Beck, 2008). Nevertheless, the appeal to the central limit theorem remains an important counterargument, but proper consideration will have to wait for your statistics courses.

Interestingly, though the normal distribution may be overused in empirical political science, it is perhaps underused in formal theoretical political science. Parameters that might in fact be distributed normally are rarely modeled as such. The reason is that, as we have noted, CDFs of distributions are important in formal theory, and the CDF of the normal distribution admits no closed-form expression. In other words, to compute the CDF at some value x, one must numerically approximate the integral defining the CDF. This is not an issue in statistics, as statistical computing software can use numerical approximation to do so. You either have already seen or will soon see tables of z-scores;[17] these are computations of the standard normal CDF $\Phi(x)$ or transformations of these. However, in game theory one often wants a closed-form expression that one can maximize, and the normal distribution does not admit this. Computational modeling, lacking this constraint, is more likely to take up use of the normal distribution when appropriate (e.g., Siegel, 2009).

Before turning our attention to the distributions of other variables, we briefly consider some variations of the normal distribution.

11.3.1.2 The Power-Transformed and Log-Normal

Power transformations can be used to make variables whose distributions deviate from the normal more closely approximate the normal. We can write a power-

[17] A z-score is obtained by transforming x to $z = \frac{x-\mu}{\sigma}$, and such tables list the corresponding values of $\Phi(z)$ or transforms thereof.

transformed normal distribution as follows:

$$f(x;\mu,\sigma^2,\lambda) = \frac{\lambda x^{\lambda-1}}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x^\lambda-\mu)^2}. \tag{11.13}$$

Like all Gaussian distributions, it has the location and scale parameters μ and σ^2, but it also has a shape parameter λ. Equation (11.13) is quite similar to equation (11.12); to see how this transformation operates, observe that when $\lambda = 1$, equation (11.13) reduces to equation (11.12). However, when $\lambda \leq 1$, the right side of the distribution will be longer than the left side, and when $\lambda \geq 1$, the left side will be longer than the right. An asymmetry where one side (or tail) of the graph of the function is longer than the other is called skewness.[18] Figure 11.7 is a plot of the power-normal.[19]

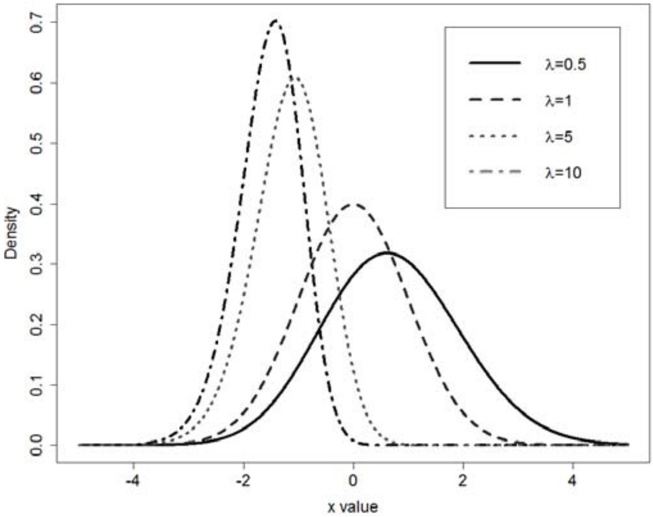

Figure 11.7: Power-Transformed Normal PDF

In the 1970s and 1980s, when political scientists recognized that much of their sample data had a skewed distribution, they frequently sought to transform the variable to make it more closely approximate the normal distribution. The most common transformations are power transformations, and the log transformation is the most widely used of the power transformations. We can write the log-normal distribution as

$$f(x;\mu,\sigma^2) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\ln(x)-\mu)^2}. \tag{11.14}$$

[18] Recall from the previous chapter that skewness is the third moment about the mean.

[19] This figure is from the National Institute of Standards and Technology's online *Engineering and Statistics Handbook*. Note that their p is our λ. To see their full entry (including the equation that they use), see `http://www.itl.nist.gov/div898/handbook/eda/section3/eda366d.htm`.

The graph of this function looks like Figure 11.8.

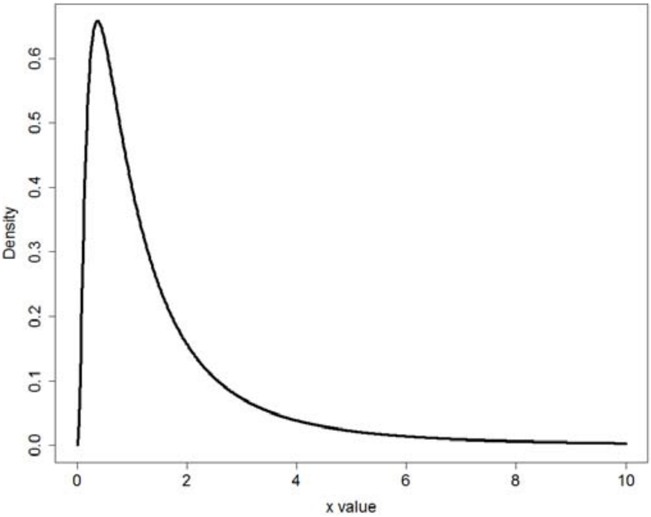

Figure 11.8: Log-Normal PDF, $\mu = 0$, $\sigma^2 = 1$

11.3.1.3 Why Should I Care?

As noted, it used to be fairly common practice to transform skewed distributions. When theory suggests that a given variable has a log (or other power) normal distribution, then these transformations are entirely appropriate. However, in the past ten to fifteen years, increasing numbers of political scientists have come to recognize that skewed sample data might well imply a non-Gaussian distribution. And the popularity of these transformations has declined as computers have made it easier to estimate models that assume non-Gaussian distributions. However, it is critical to keep in mind that some political processes may well produce log-normal or other power variants of the normal distribution, and that it is entirely appropriate to perform such transformations in such cases. For further reading, Mills (1991, pp. 40–50) provides a useful discussion of transforming sample data so that they approximate the normal distribution. For a more thorough overview of the normal distribution, please see the webpage at `http://mathworld.wolfram.com/NormalDistribution.html`.

11.3.2 The Logistic Distribution

The PDF for the logistic distribution can be written as

$$f(x;\mu,\sigma^2) = \frac{\pi e^{-\frac{\pi(x-\mu)}{\sigma\sqrt{3}}}}{\sigma\sqrt{3}\left(1+e^{-\frac{\pi(x-\mu)}{\sigma\sqrt{3}}}\right)^2}. \tag{11.15}$$

Like the normal distribution, the PDF for the logistic distribution is defined by two moments, the mean and the variance. Further, as Figure 11.9 indicates, the logistic distribution's PDF is symmetric. However, the logistic distribution is not part of the Gaussian family.

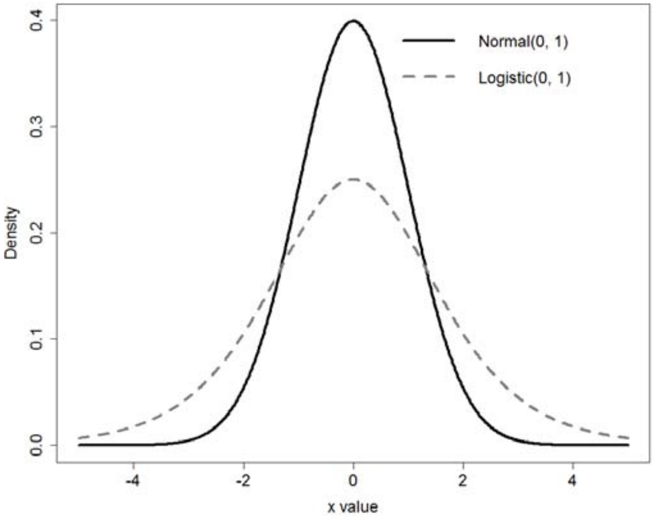

Figure 11.9: Logistic and Normal PDFs, $\mu = 0$, $\sigma^2 = 1$ (the logistic distribution has a lower peak and wider tails)

Figure 11.9 depicts two distributions with the same mean and variance. One is the logistic distribution and the other is the normal. Their similarity is remarkable, yet they have a clear difference: the logistic distribution displays considerably thicker tails, implying that values further from the mean are more likely to be drawn from a logistic distribution than a normal one. If you recall our discussion of moments in the last chapter, the fourth moment, or kurtosis, describes the thickness of the tails of distributions. The logistic has a non-zero kurtosis, while the normal distribution's is zero.

The logistic distribution is primarily used by political scientists modeling *binary outcomes* (e.g., voted/didn't vote, participated in a militarized dispute/did not participate). You will learn more about this in your statistics classes. You can find a more thorough technical overview of the logistic distribution at `http://mathworld.wolfram.com/LogisticDistribution.html`.

11.3.3 Duration Distributions

How long do legislators typically serve in a lower (e.g., provincial or state) house before seeking election to a higher (e.g., federal) legislature? How long do different types of coalition governments (e.g., majority vs. minority) survive? How long do different types of polities (e.g., democracies vs. autocracies) persist? How long do different types of military alliances (e.g., defensive vs. offensive) last? Political scientists are increasingly interested in concepts that are measured in units of time. Variables that measure such concepts can usually be modeled as if they were drawn from a duration distribution.

If you take advanced statistics courses you will likely encounter these models. A widely used model, the Cox proportional-hazards model, makes no distributional assumptions. That sounds too good to be true: we can use this model to employ statistical inference regardless of the distribution of our measure of duration. Yet, as Box-Steffensmeier and Zorn (2001) show, the Cox model can produce misleading inferences when hazards are not proportional. The articles by Zorn (2000) and Box-Steffensmeier and Zorn (2002) may also be consulted to learn more about the statistical models available for duration analysis.

11.3.3.1 The Exponential Distribution

The PDF for the exponential distribution is

$$f(x;\mu) = \frac{1}{\mu}e^{-\frac{x}{\mu}}, \tag{11.16}$$

where $\mu > 0$ is the mean duration between events. If we define $\lambda = \frac{1}{\mu}$, then we can rewrite equation (11.16) as $f(x;\lambda) = \lambda e^{-\lambda x}$, and this is common notation that you may encounter. We provide some representations of the exponential distribution's PDF in Figure 11.10.

The exponential distribution describes events produced by a process with a constant risk to failure. That is, the probability of failure does not change over time. "Failure" is a generic term that indicates the presence of a new state and should not be taken literally. That is, we can use the exponential distribution to study processes where it would be awkward to speak of "failure."

What sort of political process might produce a variable with an exponential distribution? The duration of cabinet governments seems an unlikely candidate since if the government persists long enough, elections are required after a given period of time (e.g., Cioffi-Revilla, 1984; King et al., 1990). So, if elections are required after five years in office and the government survives 1,824 days (i.e., one day less than five years), then we know with certainty that the government will dissolve the next day. So the duration of cabinet governments is not constant over time. Further, one likely expects a government's prospects for failure to be low in the first months in office, then rise some, etc. And lots of political processes seem likely to have risks of failure that change over time. For example, would you be willing to assume that the probability that a person votes is

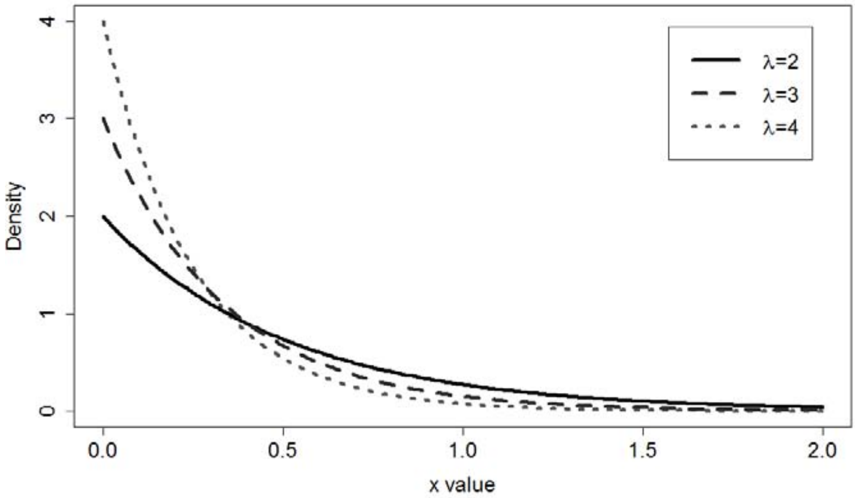

Figure 11.10: Exponential PDF, $\lambda = 2, 3, 4$

independent of her age? Do you suspect that the probability that a country goes to war is independent of the time that has passed since it last went to war?

That we can think of variables that interest political scientists which have failure risks that probably vary over time does not, however, suggest that the exponential distribution is useless. In fact, it is quite useful for processes with a constant risk of failure. Whether there are many duration processes of interest to political scientists that have a constant risk of failure over time is something for scholars to determine as the literatures that explore durations continue to grow.

You can find a thorough technical overview of the exponential distribution at `http://mathworld.wolfram.com/ExponentialDistribution.html`.

11.3.3.2 The Pareto Distribution

The PDF for the Pareto distribution can be written

$$f(x;\kappa,\beta) = \begin{cases} \frac{\kappa\beta^{\kappa}}{x^{\kappa+1}} & \text{for} \quad x \geq \beta, \\ 0 & \text{for} \quad x < \beta, \end{cases} \tag{11.17}$$

where $\kappa > 0$ is a shape parameter and β is a scale parameter such that $x \geq \beta > 0$.

Midlarsky (1988) uses the Pareto distribution to both theoretically and empirically model land inequality in Latin America. He argues that land was settled sequentially over time such that those who first claimed rights to land were able to secure larger properties than those who came later. Unlike many probability density functions which assume independence among observations,

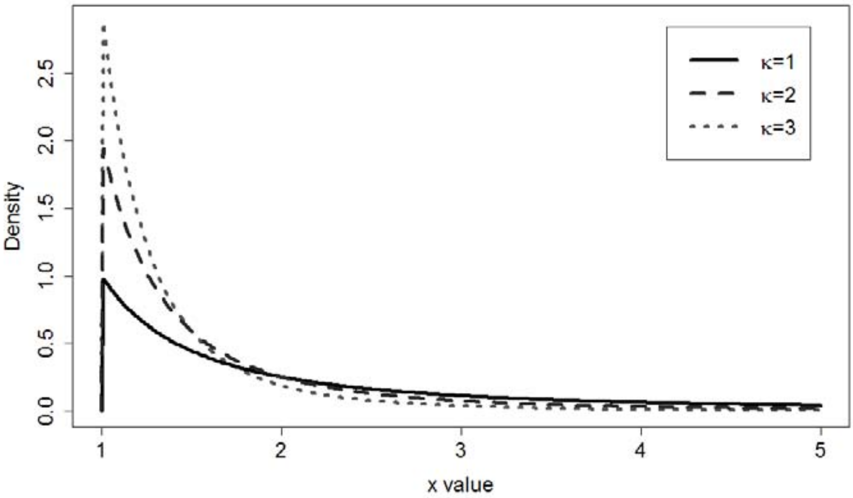

Figure 11.11: Pareto PDF, $\beta = 1, \kappa = 1, 2, 3$

the Pareto distribution assumes that the initial values have an effect on the size of subsequent values. As Midlarsky (1988, p. 494) puts it, "the assumption of a progressive sequential inequality leads to the Pareto distribution."

A thorough technical overview of the Pareto distribution is available at `http://mathworld.wolfram.com/ParetoDistribution.html`.

11.3.3.3 The Gamma Distribution

The PDF for the gamma distribution can be represented for $x \geq 0$ and $\alpha, \beta > 0$ as

$$f(x; \alpha, \beta) = \frac{x^{\alpha-1} e^{-\frac{x}{\beta}}}{\beta^{\alpha}\Gamma(\alpha)}, \tag{11.18}$$

where α is a shape parameter, β is a scale parameter (the mean is $\alpha\beta$), and $\Gamma(\alpha)$ is a named integral function that is equal to $(\alpha - 1)!$ if α is a positive integer.[20] As Figure 11.12 demonstrates, the gamma distribution produces rather different PDFs, depending on the values of α and β.

When $\alpha = 1$, the gamma distribution reduces to the exponential distribution, with $\mu = \beta$. That is because we can think of α as the number of distinct periods of constant risk. The exponential assumes that risk is constant over the entire time, so there is only one period (i.e., $\alpha = 1$). However, imagine that a political scientist were to argue that there are four distinct periods that cabinet

[20] You can find a definition of the gamma function at `http://mathworld.wolfram.com/GammaFunction.html`.

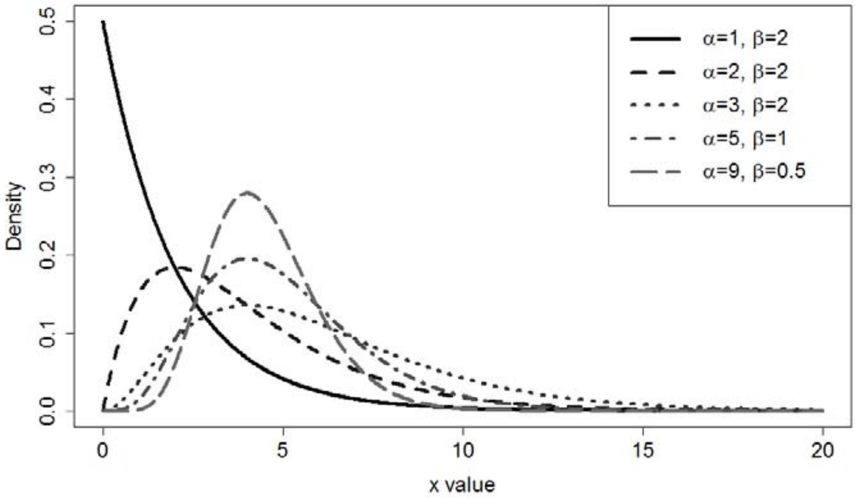

Figure 11.12: Gamma PDFs

governments experience with respect to risk of failure: (1) a honeymoon period with low risk, (2) a period of risk (when the honeymoon is over), (3) a mature period with reduced risk (for those governments that survive), and (4) a high-risk period (because the government is approaching the constitutional limit of its life). If that is a reasonable theory, then one would expect that if a variable measuring the life of a government is produced by a gamma distribution, then $\alpha = 4$.

You can find a thorough technical overview of the gamma distribution at `http://mathworld.wolfram.com/GammaDistribution.html`.

11.3.3.4 The Weibull Distribution

One can represent the PDF for the Weibull distribution as follows:

$$f(x;\alpha,\beta) = \begin{cases} \frac{\alpha x^{\alpha-1}}{\beta^{\alpha}} e^{\left(-\frac{x}{\beta}\right)^{\alpha}} & \text{for} \quad x \geq 0, \\ 0 & \text{for} \quad x < 0, \end{cases} \tag{11.19}$$

where α is a shape parameter and β is the scale parameter. The Weibull PDF is very similar to the gamma PDF; the differences are that Γ is not in the denominator, α is in the numerator, and the exponent of the e is raised to the power α (compare equations (11.18) and (11.19)). As such, note that, as with the gamma distribution, when $\alpha = 1$ the Weibull reduces to the exponential distribution, again with $\mu = \beta$.

We provide some graphs of the Weibull distribution PDF using different values of the parameters in Figure 11.13. You can see that the risk changes over time, and further, that the changes depend on the values of the parameters α and β.

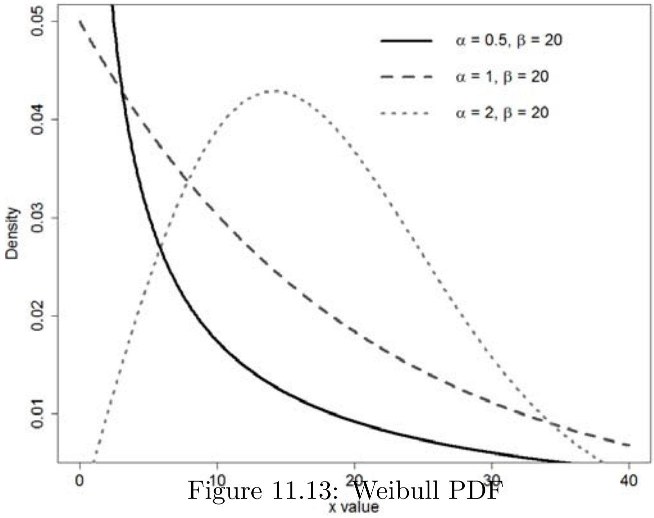

Figure 11.13: Weibull PDF

Lindsey (1995, p. 133) offers the following explanation of the processes that produce a Weibull distribution:

> It can be interpreted as if several processes are running in parallel, with the first to stop ending the duration. This is a weakest link mechanism, as when the failure of some part causes a machine to break down and the total operating time of the machine is the duration.

In other words, imagine that a number of "things" were required for a cabinet government to persist, such that if any one of those "things" were no longer present, the government would fall. The Weibull distribution should be useful when we want to study a variable measuring duration and we believe that the political processes that affect the duration are each necessary conditions to continuation. Though the hypotheses that are frequently tested using models built on the Weibull distribution are not often stated as sets of necessary conditions, it is probably the most widely used duration distribution in political science.

As a concrete example, we may consider Bennett (1997), who studies the duration of international military alliances. He estimates a statistical model that assumes that the duration of alliances have a Weibull probability distribution. In other words, the probability that an alliance is broken at any given moment in time, t, depends on (1) how long the alliance has lasted and (2) a number of other factors that Bennett specifies (e.g., changes in the power of the allies, regime change, etc.; see the article for details).

Those interested in a thorough technical overview of the Weibull distribution should visit `http://mathworld.wolfram.com/WeibullDistribution.html`.

11.3.4 Distributions Used Frequently in Statistical Hypothesis Tests

These distributions are of interest primarily in testing statistical hypotheses and are not much used to structure theoretical expectations. They will also be discussed at length—or at least the tests based on them will be—in your statistics classes. Accordingly, we will introduce them only briefly here.

11.3.4.1 Chi-squared (χ^2) Distribution

The sum of the squares of n independent variables each distributed according to a standard normal distribution is distributed according to a chi-squared (χ^2) distribution. We write a variable distributed in this way as $Q \sim \chi^2(n)$, where n is the number of degrees of freedom. Its PDF is

$$f(x;n) = \begin{cases} \frac{x^{n/2-1}e^{-x/2}}{2^{n/2}\Gamma(n/2)} & \text{for} \quad x \geq 0, \\ 0 & \text{for} \quad x < 0. \end{cases} \tag{11.20}$$

The chi-squared distribution is actually a special case of the gamma distribution, as one can see by using the parameters $\alpha = \frac{n}{2}$ and $\beta = 2$ in equation (11.18).

11.3.4.2 The (Student's) t Distribution

The Student's t distribution is the distribution of a random variable that is proportional to the ratio of a variable that is distributed according to the standard normal distribution and the square root of a variable that is distributed according to a chi-squared distribution. Such ratios arise when normalizing differences in means by the standard deviation. The distribution looks much like a normal distribution but with thicker tails, is particularly useful for small sample sizes, and approaches the standard normal distribution as the sample size approaches infinity. Its PDF, where n is the number of degrees of freedom, is

$$f(x;n) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}\left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}. \tag{11.21}$$

11.3.4.3 The F Distribution

The F distribution is the distribution of a random variable that is equal to the ratio of two random variables, each distributed according to a chi-squared distribution and each scaled according to its number of degrees of freedom. It is used commonly in the analysis of variance and in testing the hypothesis that several parameters are not jointly null. If its two degrees of freedom are n_1 and n_2 and if $x \geq 0$, then the F distribution's PDF is

$$f(x;n_1,n_2) = \frac{\sqrt{\frac{(n_1x)^{n_1}n_2^{n_2}}{(n_1x+n_2)^{n_1+n_2}}}}{xB\left(\frac{n_1}{2},\frac{n_2}{2}\right)}, \tag{11.22}$$

where $B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$ is the beta function.

11.4 EXERCISES

1. Why can't one create a PDF by plotting the graph of the relative frequency distribution of each value in the sample?

2. What is the difference between a PMF and a PDF?

3. What is the difference between a PDF and a CDF?

4. Write down an example where a scatter plot would be useful for examining the joint distribution of two variables.

5. Why can't we eyeball a probability distribution and determine whether it is normal?

6. What is the difference between a relative frequency distribution of a sample and a PDF?

7. Why does a PDF require calculating an integral?

8. Show that $\text{Var}(X) = E[(X - \mu)^2] = E[X^2] - \mu^2$.

9. An individual benefits from an action whenever $X > 10$. If X is a random variable distributed uniformly on $[0, 25]$, what is the probability that the individual will benefit?

10. Annual country budget deficits (surpluses) are distributed normally, with a mean of −\$100 million and a standard deviation of \$300 million. What do both of these parameters tell us substantively about this distribution? Explain.

11. Write down a political process that you think might be drawn from the following distributions: normal or log-normal; logistic; or exponential, Pareto, gamma, or Weibull (you should have three political processes).

12. Visit the "Distributions" page of the Virtual Laboratory Website at the University of Alabama, Huntsville (`http://www.math.uah.edu/stat/dist/index.xhtml`). Click on the "Random Variable Experiment" link under "Applets." You can do experiments changing the parameters of a number of distributions (the normal, gamma, chi-squared, Student's *t*, F, beta, Weibull, Pareto, logistic, and log-normal are available). Investigate the distributions covered in this chapter. More explicitly, select a distribution and note the shape and location of the density function. Adjust one of the parameters using the scroll bar. If there is more than one parameter, adjust it. Write down what happens when you adjust each parameter for the following distributions: normal, log-normal, logistic, beta, gamma,

Pareto, and Weibull. Note the distributions that can be made to have a bell shape given some parameter values.

13. Using the random variable experiment applet you used in the previous exercise, run the simulation 1,000 times (set Stop to 1,000) with an update frequency of 10 (use the Update tab), and note the apparent convergence of the empirical density to the true density. What does this imply, in your opinion, for the shape of the distribution of real data relative to the shape of a theoretically derived PDF?

11.5 APPENDIX

Our presentation has been relatively informal, and one can find more formal treatments in Gill (2006) and online (e.g., the various MathWorld entries we noted throughout). Those interested in studying methods as a subfield will want a more thorough treatment. Another place to look is King (1989, chaps. 2 and 3). The National Institute of Standards and Technology Engineering Statistics Handbook, section 1.3.6, "Probability Distributions," available at `http://www.itl.nist.gov/div898/handbook/eda/section3/eda36.htm`, is also a good source.