CHAPTER 3

The probability model of uncertainty

What we commonly call "probability" is really a formal model of uncertainty with only three axioms. These axioms are neither true nor false; they are assumed. By applying the rules of logic to these axioms, all of probability theory and distribution theory emerges. In what follows, I state these three axioms and provide a sense of a portion of the rich theory that developed from them. My general purpose is not to provide a complete treatment of any of these subsidiary topics, as that would require a library of additional volumes. Instead, I concentrate on a review of the fundamental rules and ideas of primary use in statistical modeling and likelihood inference.

My notation is slightly nontraditional from the perspective of probability theory. However, it closely follows the notation introduced in Chapter 1, emphasizing the important role of the probability model in specifying the complete statistical model. Since I assume most readers have seen some probability theory before, this chapter builds on the basics by demonstrating that it is not necessary to fit one's data to an existing, and perhaps inappropriate, stochastic model. Instead, we can state first principles at the level of political science theory and derive a stochastic model deductively. For the same reasons, the organization below does not follow the traditional approach in increasing order of complexity (discrete, continuous, compound, multivariate, etc.). Instead, this chapter is structured to emphasize how one builds distributions and stochastic models of political phenomena, grouping primarily by the similarity of substantive problems and data. Aside from that presented in this chapter, considerably more probability theory is described and derived in the course of the specific methods developed in Part II.

3.1 The probability model

The set of all possible outcomes of an experiment is the *sample space*, $\mathcal{S}$. The elements of this set are mutually exclusive and exhaustive outcomes of the experiment. Sample spaces may be *discrete and finite*, such as the outcome of presidential elections, $\mathcal{S} = \{\text{Democratic, Republican, Other}\}$. They may be *discrete, but countably infinite*, such as the number of Israeli raids into Southern Lebanon a year, $\mathcal{S} = \{0,1, \ldots\}$. Sample spaces may also be *continuous*, where listing events becomes impossible; for example, the duration of a parliamentary coalition is a continuous variable.

A *random variable* is a *function* that assigns a real number to every possible random output from a particular experiment, that is, to every element in the sample space. For example, the random variable for presidential elections might assign 1.0 to the event "Democrat wins," 2.0 to "Republican wins," and 3.0 to "other party wins." The word "random" in random variable refers to the stochastic variation in the outputs across experiments for each observation. "Variable" is used to emphasize that in actual runs of the experiment, the observed values of this process vary over observations, even for a single experiment. By contrast, since the parameter μ_i varies over observations but is constant in repeated experiments, it is called a *nonrandom variable*.[1]

Let Y_i be the random variable for observation i ($i = 1, \ldots, n$). In this chapter, only stochastic models of this single observation will be considered. Subsequent chapters will tie together all n observations into a single model by adding the systematic component. Many of the stochastic models developed below are defined as aggregates (sums, counts, means, etc.) of processes occurring over time that are observable only at the end of some period. Thus, even though stochastic *processes* occur through time, this entire stochastic model for Y_i still results in only a single observation. The set of all n observations may still vary over time, across space (as in cross-sectional data), or both (as in pooled time series-cross-sectional data).

I use y_i to denote a realization or hypothetical value of the random variable Y_i and y_{ji} to denote a real number. The subscript i still refers to the observation number ($i = 1, \ldots, n$), and j is a symbolic label for one possible outcome of the experiment. Furthermore, a set of outcomes is called an *event* and a particular event is labeled z_{ki}. An event of type k may include no outcomes (the null set, $z_{ki} = \{\phi\}$), a single outcome ($z = \{y_{ji}\}$), several outcomes ($z_{ki} = \{y_{1i}, y_{2i}, y_{5i}\}$), a range of outcomes [$z_{ki} \in (2,3)$], or the entire sample space ($z_{ki} = \{\mathcal{S}\}$).

The probability model for a particular experiment assigns a measure of uncertainty to every possible event. The three axioms of the model of probability are defined on the basis of these definitions:

1. For any event z_{ki}, $\Pr(z_{ki}) \geq 0$.
2. $\Pr(\mathcal{S}) = 1.0$.
3. If $z_{1i}, \ldots, z_{Ki}$ are K mutually exclusive events, then

$$\Pr(z_{1i} \cup z_{2i} \cup \cdots \cup z_{Ki}) = \Pr(z_{1i}) + \Pr(z_{2i}) + \cdots + \Pr(z_{Ki}).$$

[1] Strictly speaking, a random variable is neither random – because the probability model is extrinsic to this function – nor variable – since the function remains constant over observations. However, I adopt the somewhat less formal and more intuitive usage in the text.

Axiom 1 requires that for any event, all numerical values produced by the function $\Pr(\cdot)$ be greater than or equal to zero. If an event is not in the set of possible events $\mathcal{S}$, its probability is zero. Axiom 2 is the obvious requirement that something must happen in every run of the experiment (if you flip a coin, either heads or tails must appear). Since events outside the sample space $\mathcal{S}$ occur with probability zero, the event $\mathcal{S}$ must occur with probability one. Axioms 1 and 2 combine to require that any probability range only between zero and one, $0 \geq \Pr(z_{ki}) \geq 1$, for any event z_{ki}. Axiom 3 provides the basic additive rule for calculating the probabilities of mutually exclusive events.

Together, these three axioms define a probability model on the sample space $\mathcal{S}$, as the real valued function, $\Pr(\cdot)$. From these axioms one can derive all the rules of probability theory. Before moving on to complete stochastic models for a single observation, I mention three particularly important results here: First, probabilities of events z_{i1} and z_{k2} that are not necessarily mutually exclusive (i.e., may have one or more outcomes in common) may be calculated as:

$$\Pr(z_{1i} \cup z_{2i}) = \Pr(z_{1i}) + \Pr(z_{2i}) - \Pr(z_{1i} \cap z_{2i}),$$

where "$z_{1i} \cap z_{2i}$" represents the intersections of events z_{1i} and z_{2i}. This is derived from Axiom 3, since the overlapping part is subtracted out in the last term. If z_{1i} and z_{2i} are *mutually exclusive,* the last term drops out and this rule reduces to Axiom 3. Second, if two random variables, Y_{1i} and Y_{2i}, are *stochastically independent,* the probability law governing one has no influence on the probabilities governing the other. Accordingly, one can calculate their joint probability (of both occurring) from their marginal probabilities (of each occurring separately):

$$\Pr(Y_{1i} = y_{ji}, Y_{2i} = y_{li}) = \Pr(Y_{1i} = y_{ji}) \Pr(Y_{2i} = y_{li}).$$

Since this applies for all possible events, it may also be written more conveniently as

$$\Pr(Y_{1i}, Y_{2i}) = \Pr(Y_{1i}) \Pr(Y_{2i}). \tag{3.1}$$

For example, if we were to assume that elections to Congress and the Presidency are independent, we could calculate the probability of undivided Democratic government by multiplying the probability that the Democrats win control of the majority of both houses of Congress by the probability that they will win the Presidency.

The final result, on *conditional probability,* was already used in Chapter 1:

$$\Pr(Y_{1i} | Y_{2i}) = \frac{\Pr(Y_{1i}, Y_{2i})}{\Pr(Y_{2i})}. \tag{3.2}$$

By combining this result with Equation (3.1), one can see that independence also implies that Y_{2i} has no influence on the probability of Y_{1i},

$\Pr(Y_{1i}|Y_{2i}) = \Pr(Y_{1i})$, and Y_{1i} has no influence on the probability of Y_{2i}, $\Pr(Y_{2i}|Y_{1i}) = \Pr(Y_{2i})$.

3.2 Univariate probability distributions

The *probability distribution* of a random variable Y_i is a complete accounting of the probability of Y_i taking on any conceivable value y_i. Several different methods exist for representing these probabilities.

For example, the random variable Y_i is assigned 1 for heads and 0 for tails in a coin tossing experiment. To specify a probability distribution in this case, one only needs to write

$$\Pr(Y_i = 1) = \pi,$$
$$\Pr(Y_i = 0) = 1 - \pi,$$
$$\Pr(Y \neq 0,1) = 0.$$

π is the parameter of this distribution, where $\pi = 0.5$ when the coin is fair. The range of possible values of a parameter is called the *parameter space* and is usually denoted Θ. In this example, the parameter space for π is the range from 0 to 1 inclusive (i.e., $\pi \in \Theta = [0,1]$).

In this simple example, merely listing or graphing the probabilities of each outcome in the sample space is relatively convenient. For most interesting examples, however, the number of possible events is very large or infinite. For example, the random variables representing income (in dollars and cents), public approval of the president (in percentages) and occupation can each produce very large numbers of events with nonzero probability. Further analysis thus requires a more concise means of representation.

A mathematical formula is the solution. As most readers know, the usual method of presenting these formulas is merely as a list. However, although not often discussed in introductory books, each distribution was originally derived from a very specific set of theoretical assumptions. These assumptions may be stated in abstract mathematical form, but they may also be interpreted as political assumptions about the underlying process generating the data. The ultimate mathematical form for most distributions is usually not very intuitive, but the first principles from which they were originally derived represent models of interesting political science situations and are much closer to both data and theory. When a list of these first principles is known, understanding them is critical to the correct application of a particular probability distribution. If at any time in this chapter, the mathematics become too difficult, the reader should pay close attention to the first principles (usually appearing in italics) and final form of each distribution. The intervening steps of the derivation may be considered a black box.

Ultimately, political scientists will benefit from learning considerably more

about stochastic modeling. This will enable scholars to develop probability distributions that closely match whatever social science processes they desire to model. Fortunately, enormous numbers of distributions have already been developed by statisticians and others.[2] Although often developed for other purposes, analysts can marshal this wealth of raw material for modeling problems in political science. In so doing, researchers need not get bogged down in understanding precisely how every probability distribution is derived from first principles, but we must completely understand these initial substantive assumptions and be able to apply the final distributions. For each distribution in this chapter, I present the first principles and final distribution form. For some, I also present the full derivation. Chapter 4 demonstrates how to apply these distributions in problems of inference.

Bernoulli distribution

The simplest distribution represents the situation where a random variable has only two possible events with nonzero probability. A coin flip can be represented easily with the Bernoulli distribution. More generally, this distribution can represent one observation of any dichotomous random variable. Vote choice for Bush or Dukakis, employed or unemployed, and developing or industrialized nations are a few of the many possible examples.

The two first principles required to derive the Bernoulli probability distribution are quite simple: the random variable Y_i must have two *mutually exclusive* outcomes, $y_i = 0$ and $y_i = 1$, that are *exhaustive*. This distribution is usually written so that zero and one are the outcomes, but any other coding of a dichotomous variable could be mapped onto this representation. These are the only first principles required. Mutual exclusivity indicates that $\Pr(Y_i = 1 | Y_i = 0) = 0$. If voters were permitted to vote for both Bush and Dukakis, the Bernoulli distribution would not apply. Being exhaustive indicates that the probability of one event occurring is the complement of the probability of the other happening:

$$\Pr(Y_i = 1) = 1 - \Pr(Y_i = 0).$$

Hence, the Bernoulli distribution is a better approximation to the underlying process generating election outcomes, for example, when no popular third party candidates are in the race. Many binary variables fit these two simple assumptions.

From these first principles, deriving a single probability distribution is straightforward. To begin, define a parameter π, such that

[2] The four volumes written by Johnson and Kotz (1969, 1970a, 1970b, 1972) are the standard references on these matters. Rothschild and Logothetis (1987) is a more affordable paperback. See also Shapiro and Gross (1981).

$$\Pr(Y_i = 1) = \pi$$

A direct consequence of the events being mutually exclusive and exhaustive is that:

$$\Pr(Y_i = 0) = 1 - \pi.$$

The full probability distribution is then a way of putting together these two separate parts into a single equation:

$$Y_i \sim f_{\text{Bern}}(y_i|\pi) = \begin{cases} \pi^{y_i}(1-\pi)^{1-y_i} & \text{for } y_i = 0,\ 1, \\ 0 & \text{otherwise.} \end{cases} \tag{3.3}$$

In combination with an appropriate systematic component, this simple distribution will provide a useful model of dichotomous random variables.

Since this is a discrete probability distribution, $f(y_i|\pi) = \Pr(y_i|\pi)$; the same is obviously not true for continuous distributions. Once a probability distribution is completely specified, any feature of the random variable of interest may be calculated. For example, the mean of Y_i is calculated by taking the expected value:

$$\begin{aligned} E(Y_i) &= \sum_{\text{all } y_i} y_i f(y_i) \\ &= 0 f_i(0) + 1 f_i(1) \\ &= 0 + \pi \\ &= \pi. \end{aligned}$$

Hence, if heads is 1 and tails is 0, then one flip of the coin will yield, on average, the number π. If a vote for Dukakis is assigned a 1 and Bush a 0, then individual i will have a π probability of voting for Dukakis.

Binomial distribution

Suppose a series of N Bernoulli random variables exists, but *we only observe the sum* of these variables. A political scientist might still be interested in these unobserved binary random variables, but the only data available are based on this sum. For example, data might exist on the number of elections, out of the last five, in which a survey respondent recalled voting. The survey researcher might not have had the time or money to include separate questions about voting participation in each election. Indeed, this aggregate recall question may be subject to fewer recall problems since positive and negative errors in the individual (unobserved) binary questions would tend to cancel out.

By assuming that the unobserved binary variables (e.g., the five decisions to vote or not) are *independent* and *identically distributed,* one can derive a binomial distribution to model this situation. Thus, we suppose that the same

Table 3.1. *Deriving the binomial distribution*

Y_{1i}	Y_{2i}	Y_{3i}	$\Sigma_{j=1}^{3} Y_{ji} \equiv Y_i$	$\Pr(Y_i)$
0	0	0	0	$(1-\pi)^3$
1	0	0	1	$\pi(1-\pi)^2$
0	1	0	1	$\pi(1-\pi)^2$
0	0	1	1	$\pi(1-\pi)^2$
0	1	1	2	$\pi^2(1-\pi)$
1	0	1	2	$\pi^2(1-\pi)$
1	1	0	2	$\pi^2(1-\pi)$
1	1	1	3	π^3

Bernoulli distribution, with the same parameter π, describes each of the five constituent elections. A constant π implies that the probability of individual i voting is the same for each election. The independence assumption, as interpreted here, means that participation by individual i in one election is unrelated to participation by that individual in other elections, except that π is the same in each election.

To get a feel for how the binomial distribution is derived from these first principles, take the case of $N=3$ binary variables. Whereas a Bernoulli variable has two possible outcomes, a binomial with $N=3$ has $2^3=8$. Table 3.1 lists each of these eight outcomes with their associated probabilities, calculated using the three first principles. For example, the first line in the table records the outcome that all three binary random variables (Y_{1i},Y_{2i},Y_{3i}) were realized as zeros. Since the three variables are independent, their probabilities may be multiplied to derive the probability of the sum taking on the value zero [see Equation (3.1)]. Since they are identically distributed, the same probability in each case $(1-\pi)$ is multiplied. Probabilities for the other rows are calculated similarly.

Suppose one is interested in calculating the probability that the sum of the three variables, Y_i, is equal to one. The second, third, and fourth rows of Table 3.1 each have outcomes with a sum equal to one. Since these outcomes are mutually exclusive, they may be added:

$$\begin{aligned}\Pr(Y_i=2) &\equiv \Pr\left(\sum_{j=1}^{3} Y_{ji}=1\right) \\ &= \Pr(Y_{1i}=1,\, Y_{2i}=0,\, Y_{3i}=0) \\ &\quad + \Pr(Y_{1i}=0,\, Y_{2i}=1,\, Y_{3i}=0) \\ &\quad + \Pr(Y_{1i}=0,\, Y_{2i}=0,\, Y_{3i}=1) \\ &= \pi(1-\pi)^2 + \pi(1-\pi)^2 + \pi(1-\pi)^2 \\ &= 3\pi(1-\pi)^2.\end{aligned} \tag{3.4}$$

If $\pi = 0.5$, as in the case of a fair coin, the probability that only one of three coins would turn up heads is $3\pi(1-\pi)^2 = 3(0.5)(1-0.5)^2 = 0.375$.

In order to derive a single formula, instead of always listing all possible outcomes and associated probabilities, note that the last line in Equation (3.4) has two parts: the number of outcomes (3) and the probability of each of these outcomes $[\pi(1-\pi)^2]$. In this case, the probability of Y_i is calculated by taking the product of these two parts. Indeed, regardless of the number of possible events (N), the binomial probability always has these two parts. The first part may be generalized with a little knowledge of combinatorics as $\binom{N}{y_i}$. The second part is simply $\pi^{y_i}(1-\pi)^{N-y_i}$. The result is the familiar binomial distribution:

$$f(y_i|\pi) = \begin{cases} \binom{N}{y_i}\pi^{y_i}(1-\pi)^{N-y_i} & \text{for } y_i = 0, 1, \ldots, n, \\ 0 & \text{otherwise} \end{cases} \qquad (3.5)$$

$$= \begin{cases} \frac{N!}{y_i!(N-y_i)!}\pi^{y_i}(1-\pi)^{N-y_i} & \text{for } y_i = 0, 1, \ldots, n, \\ 0 & \text{otherwise,} \end{cases}$$

where $\binom{N}{y_i}$ is shorthand for $\frac{N!}{y_i!(N-y_i)!}$. If the Bernoulli assumptions of mutually exclusive and exhaustive binary events hold, and the additional assumptions that the N binary random variables are independent and identically distributed also hold, one can use this formula to calculate the probability that the random sum Y_i equals y_i, which may be between zero and N.

Extended beta-binomial distribution

Many situations exist in political science where data on the sums of random binary variables are available, but the binomial assumptions of independence and identical distributions are questionable. The number of members of the U.S. Senate (out of a total of 100) who vote for a particular bill, or the number of school districts in a state banning *The Catcher in the Rye,* are good examples. Although unobserved, one can reasonably hypothesize that the probability π of each senator voting for a particular bill is not identical. In addition, school districts, as well as senators, are likely to influence each other rather than be totally independent random variables.

I begin by weakening the binomial assumption that the unobserved binary random variables making up Y_i have constant π. To derive an alternative distribution, one cannot let π vary haphazardly. Instead, we must choose a specific distribution that governs the variation in π across these individual binary variables within the single observation, Y_i. Although π is a fixed parameter in the binomial distribution, it becomes a random variable in this derivation of a new probability distribution. The choice of possible distributions is limited to those where the random variable is bounded as is the param-

eter space for π, between zero and one. A variety have been proposed, some leading to intractable expressions, others somewhat less intuitive. The most commonly used distribution for π is the *beta* density. Thus, in addition to the assumptions required to derive the binomial distribution, the key first principle required to derive the extended beta-binomial distribution is that π *varies according to a beta density.*

The beta distribution has π as a random variable and two parameters ρ and γ (in my parameterization) and may be written as follows:

$$f_\beta(\pi|\rho,\gamma)=\frac{\Gamma(\rho\gamma^{-1}+(1-\rho)\gamma^{-1})}{\Gamma(\rho\gamma^{-1})\Gamma[(1-\rho)\gamma^{-1}]}\pi^{\rho\gamma^{-1}-1}(1-\pi)^{(1-\rho)\gamma^{-1}-1} \quad (3.6)$$

for $0<\pi<1$, and zero otherwise, and where $\Gamma(x)$ is the gamma function:

$$\Gamma(x)=\int_0^\infty z^{x-1}e^{-z}dz. \quad (3.7)$$

One can calculate values of the gamma function for particular values of x either directly from this integral (with z as the integration dummy) or from tables designed for this purpose (see Johnson and Kotz, 1970a: Chapter 17). For integer values of x, $\Gamma(x+1)=x!=x(x-1)(x-2)\cdots 1$. Noninteger values of x produce a continuous interpolation.

The beta distribution is relatively flexible, and, depending on different values of ρ and γ, it can be unimodal, bimodal, or skewed. This distribution is assumed here not because it was derived from some set of first principles about how π varies over the individual binary variables, but rather because it is a first principle. The benefit of assuming this distribution is its flexibility in handling a variety of interesting cases. It is also mathematically simple, particularly in combination with the binomial distribution.[3] Substantively, $\rho\equiv E(\pi)$ is the mean of the distribution of π, and γ is an index for how much variation exists in π across the binary random variables.

The goal now is to derive the extended beta-binomial probability distribution. The procedure will be to modify the binomial distribution $[f_b(y_i|\pi, N)]$ by letting π vary according to a beta distribution $[f_\beta(\pi|\rho,\gamma)]$. This procedure is generally called *compounding* a probability distribution with another distribution. Two steps are required.

First, the joint distribution (f_j) of Y_i and π is derived by using the basic equation for conditional probability. Thus, Equation (3.2) may also be presented as this:

[3] See Sheps and Menken (1973) on the beta density's mathematical properties and Heckman and Willis (1977: Appendix) for an interesting justification of the application of the beta density to the probability of labor force participation by married women.

$$\Pr(AB) = \Pr(A|B)\Pr(B).$$

Thus, I take the product of the binomial and beta distributions:

$$f_j(y_i, \pi|\rho,\gamma) = f_b(y_i|\pi) f_\beta(\pi|\rho,\gamma), \tag{3.8}$$

where γ is just carried through as a conditioning parameter, unaffected by the procedure.[4]

Second, the extended beta-binomial distribution (f_{ebb}) is derived by collapsing this joint distribution over π:

$$f_{ebb}(y_i|\rho,\gamma) = \int_{-\infty}^{\infty} f_j(y_i, \pi|\rho,\gamma)\, d\pi.$$

Since a distribution was substituted for it, π no longer appears on the left hand side of this equation. In the extended beta-binomial distribution ρ is now the mean probability of the unobserved binary variables, as was π in the binomial distribution. To make clearer the relationship between the two distributions, I reparameterize the extended beta-binomial by substituting each occurrence of ρ with π. For $y_i = 0, \ldots, N$, the extended beta-binomial distribution is thus defined as

$$\begin{aligned} f_{ebb}(y_i|\pi,\gamma) &= \Pr(Y_i = y_i|\pi,\gamma,N) \\ &= \frac{N!}{y_i!(N-y_i)!} \prod_{j=0}^{y_i-1} (\pi + \gamma j) \prod_{j=0}^{N-y_i-1} (1-\pi+\gamma j) / \prod_{j=0}^{N-1} (1+\gamma j), \end{aligned} \tag{3.9}$$

where one adopts the convention that $\Pi_{i=0}^{y} c_i = 1$ for any $x < 0$.[5] This equation may appear more complicated than the ultimate form of the binomial and other distributions, but it is only algebraically complicated. Conceptually, it is just like any other traditional probability. One sets N, and the parameters π and γ, at specific values; then one can easily use arithmetic to determine the probability that Y_i takes on any particular value y_i.

In the extended beta-binomial distribution, π is an average probability of a binary variable equaling 1.0. In the binomial, π is the same for each binary variable and is thus also the same average. However, this distribution has an additional unknown parameter, γ, which governs the degree to which π varies across the unobserved binary variables making up each observation. When $\gamma = 0$, this distribution reduces to the binomial and all the πs are constant. Larger amounts of variation in π lead to larger values of γ.

[4] If f_β were reconceptualized as the degree of belief about a constant parameter π, and ρ and γ were fixed to specific numbers representing the prior expectations about π, then this procedure is equivalent to Bayesian analysis, with f_β as the prior and the left hand side of Equation (3.8) as the posterior distribution.

[5] The symbol $\Pi_{i=1}^{n}$ means "the product from $i = 1$ to n." It is an analog to the summation symbol, $\Sigma_{i=1}^{n}$.

Although this derivation weakened only the binomial assumption of identical distributions of the individual binary variables, one can prove that certain types of *dependence* among the individual binary variables lead to exactly the same extended beta-binomial distribution. Thus, heterogeneity (in π) and dependence are both modeled in this new distribution by γ. The different ranges of γ have implications for both the underlying binary variables and the aggregate observed count (Y_i).

Thus, when $\gamma = 0$, the binary variables are independent and identically distributed and the extended beta-binomial distribution reduces exactly to the binomial. When $\gamma > 0$, either positive correlation among the binary variables or heterogeneity among the πs causes *overdispersion* in the aggregate variable Y_i. When Y_i is overdispersed, its variance is larger than one would expect under the binomial distribution's assumption of binary variate independence. Finally, $\gamma < 0$ indicates negative correlations among the binary variables and results in *underdispersion*.[6,7]

Since one observes only the total count, and not the individual binary variables, determining which of the possible causes is responsible for observed over- or underdispersion is not possible: dependence among the binary variables and heterogeneity among their expected values have the same observed effect on Y_i. Indeed, this is a classic problem of identification. The data contain insufficient information at this aggregated level to distinguish between these two substantively different cases. If one collected data on the individual binary variables, this distinction would be possible. An estimated value of γ different from zero should nevertheless alert a researcher to potentially interesting information. On this basis, one could decide whether collecting more disaggregated data is worthwhile.

Poisson distribution

Another important discrete probability distribution is for a count with no upper bound. The Poisson distribution is theoretically appropriate when the occurrence of one event has no influence on the expected number of subsequent

[6] A somewhat more intuitive interpretation of γ is that the correlation among the binary variables is a direct function of this parameter: $\delta = \gamma/(1+\gamma)$. Hence, one could reparameterize the extended beta-binomial model by solving for $\gamma = \delta/(1-\delta)$ and substituting $\delta/(1-\delta)$ into the distribution for every occurrence of γ. Due to the invariance property of ML, estimating γ and transforming afterwards, if desired, is as easy.

[7] Because the possible binary variables are always constrained in the extent to which they can be negatively correlated, γ is constrained such that

$$\gamma \geq \max[-\pi(N-1)^{-1}, -(1-\pi)(N-1)^{-1}].$$

See Prentice (1986) for details.

events, λ. With three other first principles, the full Poisson probability distribution may be derived.[8]

Consider the time interval for observation i in which events are occurring. Although the random variable Y_i is a count of the total number of events that have occurred at the end of period i, the assumptions required to derive the Poisson distribution are about the generation of events during this unobserved period. As usual, the underlying data generation process is not observed, only its consequences (the total count).

To begin, denote the random variable representing the number of events that have occurred up to time t during observation period i as Y_{ti}. Then write the probability of an addition, and of no addition, respectively, to the total count during the interval from t to $t+\Delta t$ as:

$$\Pr(Y_{(t+\Delta t)i} = y_{ti} + 1 | Y_{ti} = y_{ti}) = \lambda \Delta t + o(\Delta t) \tag{3.10}$$

and

$$\Pr(Y_{(t+\Delta t)i} = y_{ti} | Y_{ti} = y_{ti}) = 1 - [\lambda \Delta t + o(\Delta t)], \tag{3.11}$$

where $o(\Delta t)$ is the probability that more than one event occurs during Δt and which, when divided by Δt, tends to zero as Δt gets smaller. We can then write the unconditional probability $\Pr(Y_{(t+\Delta t)i} = y_{ti} + 1)$ as the sum of two mutually exclusive situations: (1) y_{ti} events have occurred by time t and one additional event occurs over the next Δt interval, and (2) $y_{ti} + 1$ events have occurred at time t and no new events occur from t to $t+\Delta t$:

$$\Pr(Y_{(t+\Delta t)i} = y_{ti} + 1) = \Pr(Y_{ti} = y_{ti})\lambda \Delta t + \Pr(Y_{ti} = y_{ti} + 1)(1 - \lambda \Delta t). \tag{3.12}$$

Although Equations (3.10) and (3.11) are axiomatic, deriving Equation (3.12) from these equations requires two first principles. First, assuming that *two events cannot occur at precisely the same instant,* one can drop the $o(\Delta t)$ terms. In addition, by assuming that *the probability of an event occurring during the period from t to t + Δt is independent of any events occurring prior to time t,* each of the two terms in Equation (3.12) may be written as the product of their respective marginal probabilities.

From Equation (3.12), observe how $\Pr(Y_{ti} = y_{ti} + 1)$ changes with respect to time as Δt gets smaller and smaller:

$$\begin{aligned} \frac{\partial \Pr(Y_{ti} = y_{ti} + 1)}{\partial t} &\equiv \lim_{\Delta t \to 0} \frac{\Pr(Y_{(t+\Delta t)} = y_{ti} + 1) - \Pr(Y_{ti} = y_{ti} + 1)}{\Delta t} \\ &= \begin{cases} \lambda[\Pr(Y_{ti} = y_{ti}) - \Pr(Y_{ti} = y_{ti} + 1)] & \text{for } y_{ti} + 1 > 1, \\ -\lambda \Pr(Y_{ti} = 0) & \text{for } y_{ti} + 1 = 1. \end{cases} \end{aligned} \tag{3.13}$$

[8] The following proof relies on insights into the continuous time, discrete space Markov process outlined by Feller (1968: Chapter 17). See also King (1988a).

If we make a third assumption that *zero events have occurred at the start of the period,* $t=0$, then $\Pr(y_{0i}=0)=1$. As such, the probability distribution of the count emerging from this underlying process can begin to be built. First, solve the last part of Equation (3.13) as:

$$\begin{aligned}\Pr(Y_{ti}=0) &= -\left(\lambda^{-1}\right)\frac{\partial \Pr(Y_{ti}=0)}{\partial t}\\ &= e^{-\lambda t},\end{aligned}$$

where the last line uses the fact that the exponential function is the only function that is equal to its derivative.[9] Then, substituting into the other part of Equation (3.13) yields the probability of a single event happening between time 0 and t:

$$\Pr(Y_{ti}=1)=\lambda t e^{-\lambda t}.$$

Finally, by successively substituting and solving, one can derive the formula for the probability that Y_{ti} takes on zero, one, and all other nonnegative integers. The general formula for the Poisson distribution may be written as follows:

$$f(y_i|\lambda,t)=\begin{cases}\frac{e^{-\lambda t}(\lambda t)^{y_i}}{y_i!} & \text{for } t>0,\ \lambda>0, \text{ and } y_i=0,\ 1,\ \ldots,\\ 0 & \text{otherwise.}\end{cases} \tag{3.14}$$

The more usual form of the Poisson distribution emerges from the fourth and final assumption that *all observation periods* $(0,t)$ *are of the same length.* By convention, we let $t=1$, for all observations, which yields:

$$f(y_i|\lambda)=\begin{cases}\frac{e^{-\lambda}\lambda^{y_i}}{y_i!} & \text{for } \lambda>0 \text{ and } y_i=0,\ 1,\ \ldots,\\ 0 & \text{otherwise.}\end{cases} \tag{3.15}$$

To summarize, three first principles about the dynamics within each observation are required to derive the first form of the Poisson distribution in Equation (3.14): (1) more than one event cannot occur at the same instant; (2) the probability of an event occurring in one time is constant and independent of all previous events; and (3) zero events have occurred at the start of each period. The key substantive assumption here is (2), the other two being more technical requirements. To derive the more usual form of the Poisson distribution in Equation (3.15), one also must assume (4) that the length of each observation period i is identical. An implicit assumption of both distributions is that the rate of event occurrence λ remains constant, or at least unresponsive to y_i, over the observation period. This rate during the observation is also the expected count in the complete distribution. Indeed, the variance of the

[9] This equation gives a version of the exponential distribution, which is useful for modeling the time between independent events.

random count is also λ, indicating that the variation around the expected value increases as λ grows. This is easy to conceptualize when λ is very small: since the count cannot be negative, the variance must also be small. Social science examples include the number of patents a firm is awarded, the number of presidential vetoes a year, and the number of news stories about a candidate per year. In each case, the actual number of events has no effective maximum.

Negative binomial distribution

Two key substantive assumptions required to derive the Poisson distribution are that events accumulating during observation period i are independent and have a constant (or unresponsive) rate of occurrence, λ. If either of these first principles does not hold, a different distribution for the total count, Y_i, is produced. For example, suppose one were counting publications in a cross-sectional survey of new professors. Assuming the rate of publication λ is constant across these individuals is implausible (Allison, 1987). For another example, the number of political kidnappings a year is unlikely to meet the independence assumption because successful attempts are likely to breed other attempts.

To derive a new, more appropriate compound distribution (what is called the negative binomial), I proceed as in the derivation of the extended beta-binomial distribution from the binomial. Thus, the proof proceeds in two steps. First, instead of requiring that λ be constant over the observation period as in the Poisson distribution, I must now assume a probability distribution for λ. The choice is not obvious, but the fact that λ is restricted to positive values limits the possibilities some. Using mathematical tractability and substantive flexibility as criteria, the gamma distribution is the usual choice (Greenwood and Yule, 1920). The gamma distribution (in my parameterization) has $E(\lambda)=\phi$ and $v(\lambda)=\phi(\sigma^2-1)$. Note that as $\sigma^2\rightarrow 1$, the gamma distribution collapses to a spike over ϕ, leaving λ constant. Hence, to derive the negative binomial distribution, I make all the assumptions of the Poisson distribution with one exception: λ *is assumed to vary within an observation according to the gamma distribution.* The gamma density, for $\lambda>0$, $\phi>0$, and $\sigma^2>0$, is written as:

$$f_\gamma(\lambda|\phi,\sigma^2)=\frac{\lambda^{\phi(\sigma^2-1)^{-1}-1}e^{-\lambda(\sigma^2-1)^{-1}}}{\Gamma[\phi(\sigma^2-1)^{-1}](\sigma^2-1)^{\phi(\sigma^2-1)^{-1}}}, \tag{3.16}$$

where $\Gamma(\cdot)$ is the gamma function in Equation (3.7). With this distribution and the Poisson, we can now derive the joint distribution (f_j) of Y_i and λ again using the basic equation for conditional probability in Equation (3.2):

$$f_j(y_i,\lambda|\phi,\sigma^2)=f_p(y_i|\lambda)f_\gamma(\lambda|\phi,\sigma^2).$$

Second, one can derive the negative binomial distribution (f_{nb}) by collapsing this joint distribution over λ:

$$f_{nb}(y_i|\phi,\sigma^2)=\int_0^\infty f_j(y_i,\lambda|\phi,\sigma^2)d\lambda.$$

The parameter ϕ plays the same role of the mean rate of event occurrence as λ does in the Poisson distribution. Thus, to maintain comparability, I reparameterize by substituting λ for each occurrence of ϕ:

$$f_{nb}(y_i|\lambda,\sigma^2)=\frac{\Gamma\left(\frac{\lambda}{\sigma^2-1}+y_i\right)}{y_i!\Gamma\left(\frac{\lambda}{\sigma^2-1}\right)}\left(\frac{\sigma^2-1}{\sigma^2}\right)^{y_i}(\sigma^2)^{\frac{-\lambda}{\sigma^2-1}}, \tag{3.17}$$

where $\lambda>0$ and $\sigma^2>1$. In a manner analogous to the extended beta-binomial distribution, this distribution was derived by allowing λ to vary according to a gamma distribution. However, as Thompson (1954) first showed, the same negative binomial distribution also results from assuming a particular form of contagion among the individual events making up Y_i.[10]

In the negative binomial distribution, λ is still the expected number of events. The more events within observation i that either have heterogeneous λ or are positively correlated, the larger the parameter σ^2 will be. Although σ^2 cannot equal one in this distribution, the smaller σ^2 is, the closer the negative binomial distribution is to the Poisson. In Chapter 5, I use a generalization of the negative binomial and Poisson distributions, called the generalized event count distribution, to derive a more universal statistical model for data of this sort.

The beta and gamma distributions were used in this section only to derive the extended beta-binomial and negative binomial compound distributions, respectively, but they can also be used directly to model certain continuous processes. For example, the gamma distribution can be used to model the time between events as a generalization of the exponential distribution (see Footnote 9) rather than counts of events. Gamma distributions, being non-negative everywhere and skewed, might also be of use to model variation across individuals without a group. The beta distribution might be appropriate for modeling a proportion, since it varies between zero and one.

[10] More specifically, Thompson (1954) showed that a limiting form of the contagious Polya-Eggenberger distribution and Neyman's contagious distributions is the negative binomial. See also Neyman (1965).

Normal distribution

The most familiar continuous probability distribution is the *Normal* or *Gaussian* distribution:

$$f_N(y_i|\mu, \sigma) = (2\pi\sigma^2)^{-1/2} e^{-(y_i-\mu)^2/2\sigma^2} \quad (3.18)$$
$$\text{for } \sigma > 0, \ -\infty < \mu < \infty, \ -\infty < y_i < \infty.$$

The Normal distribution has two parameters, the mean μ and the variance σ^2. π is the mathematical constant 3.14159 ⋯ and is not a parameter here. Many different axiomatic developments of this distribution have been derived, so a variety of different assumptions can be made in order to apply it appropriately in a substantive example.

The Normal distribution has been applied in the social sciences countless numbers of times. The primary reason for its initial adoption is that its use leads to very simple mathematical and statistical calculations (least squares, linear relationships, etc.). Of course, if computational issues are treated as transparent, as they should now be because of dramatically decreased cost, this justification is no longer adequate. Another reason for using the Normal distribution is the unquestioning application of a version of the *central limit theorem* (the theorem establishing the first principles required to derive Normality): scholars often argue that their disturbance term is the sum of a large number of independent but unobserved factors. If this is the process actually driving the data and a number of other conditions hold, then the Normal distribution is appropriate, but this case must be *explicitly* made. If the process being modeled is the sum of a number of unobserved variables, then researchers ought to speculate what these might be and state exactly how the central limit theorem applies. The specific limit theorems that enable one to derive the Normal distribution from this type of situation are often no more plausible than similar limit theorems that lead to other distributions (Bartels, 1977; Koenker, 1982). The simplest proof of the central limit theorem I am aware of requires only three-quarters of a page, but still relies on concepts beyond of the scope of this book (see Tardiff, 1981; see also Spanos, 1986, for an interesting historical presentation).

When some version of the central limit theorem or other set of first principles leads one to adopt the Normal distribution, the data have certain recognizable characteristics. For example, the distribution is continuous. Hence, *a discrete random variable cannot be directly generated by the Normal distribution.* This is a critical point, since so many researchers have wasted significant amounts of information by assuming a Normal distribution when more specific information exists. Furthermore, the distribution is symmetric about μ. This means that a skewed random variable also could not be generated by the Normal distribution. Finally, a random variable that is Normally distrib-

uted has events with nonzero probabilities occurring everywhere on the real number line. This has particular consequences if a variable is bounded (both theoretically and empirically). For example, most measures of income are both bounded below at zero and positively skewed (with fewer people making very large amounts of money).

The idea that the Normal, or indeed any, distribution can be applied to every statistical problem is a distinctly 18th century idea. For back then, statisticians were searching for the "curve of errors" that would apply in all or almost all cases (see Stigler, 1986). Due to the unfortunate application of statistical tests with very low power, many mistakenly believed for some time that the Normal distribution applied to most naturally occurring situations. More modern tests and sophisticated understanding of the processes that drive observed data revealed the fallacy in this assumption. Indeed, with the enormous number of probability distributions that have been developed (Johnson and Kotz, 1969, 1970a, 1970b, 1972), this early notion now seems almost bizarre. Unfortunately, in too many cases, social scientists have yet to get beyond this mode of thinking.[11] "Everyone believes in the [Normal] law of errors, the experimenters because they think it is a mathematical theorem, the mathematicians because they think it is an empirical fact" [Poincaré, c.f. Harvey (1981a: 112)].

Log-Normal distribution

Normally distributed random variables extend theoretically over the entire number line. One class of continuous processes for which this does not apply involves those that are never negative. Income, population figures, crime statistics, and budget totals are a few examples of positive value random variables.

Many continuous distributions exist that are defined only over the positive number line. The gamma distribution in Equation (3.16) is one example that might be useful in some circumstances. Another distribution, much more closely related to the very popular Normal distribution, is the log-Normal. Suppose Z_i is a Normally distributed random variable, with mean 0 and variance 1, such that

$$Z_i = \frac{\ln Y_i - \mu}{\sigma}. \tag{3.19}$$

Then, Y_i has a log-Normal distribution with mean μ and variance σ. Since Y_i is the random variable of interest in this case, and

[11] For example, in a tour de force of econometric analysis, Judge et al. (1985: 201) write, "In practice, normality is assumed most often and we will only consider this case." Cramer (1986: xiii), in a book on maximum likelihood, writes, "the only probability distribution that occurs at all frequently [in this book] is the normal."

$$Y_i = \exp(\sigma Z_i + \mu), \tag{3.20}$$

the distribution should probably be called the exponential (a name already taken), but I will adopt the conventional name.

A general rule exists in stochastic modeling for deriving a distribution of a variable that is a function of another variable with a known distribution. Suppose X is a continuous random variable with distribution $f_x(X)$ and $Y = u(X)$ defines a one-to-one transformation of X onto Y. Let the inverse transformation be denoted as $X = u^{-1}(Y)$. For example, $Y = 2X$ and $Y/2 = X$ are the transformation and inverse transformation, respectively. The goal is to determine the distribution of Y from the distribution of X and the function $u(\cdot)$; the rule is as follows:

$$f_Y(y) = f_x\left(u^{-1}(y)\right) \left| \frac{\partial u^{-1}(y)}{\partial y} \right|. \tag{3.21}$$

The last term, the absolute value of the derivative of the inverse function, is called the *Jacobian,* and this procedure for deriving a probability distribution is sometimes called the *Jacobian method.* This procedure also works if the function $u(\cdot)$ is only piecewise invertible, instead of a one-to-one function, like the transformation $Y = X^2$ and inverse transformation $\sqrt{Y} = X$.

In the case of the log-Normal derivation, the transformation $u(\cdot)$ is in Equation (3.20) and the inverse transformation is Equation (3.19). Thus, the log-Normal distribution may be derived as follows. The standard Normal distribution of Z_i is simply the Normal [Equation (3.18)] with $\mu = 0$ and $\sigma = 1$:

$$f_{sn}(z_i) = (2\pi)^{-1/2} e^{-z_i^2/2}.$$

The Jacobian is calculated as:

$$\left| \frac{\partial Z_i}{\partial Y_i} \right| = \left| \frac{1}{\sigma Y_i} \right| = \frac{1}{\sigma Y_i}, \quad Y_i > 0.$$

Then, the full log-Normal distribution is written by combining these two results:

$$\begin{aligned} f_{ln}(y_i|\mu,\sigma) &= f_{sn}\left[\frac{\ln y_i - \mu}{\sigma}\right] \left| \frac{\partial z_i}{\partial y_i} \right| \\ &= (y_i \sigma \sqrt{2\pi})^{-1} \exp\left[\frac{-[\ln(y_i) - \mu]^2}{2\sigma^2}\right]. \end{aligned}$$

Now Y_i is a positive random variable characterized by the log-Normal distribution. It has mean $E(Y_i) \equiv \exp(\mu + \frac{1}{2}\sigma^2)$, variance $V(Y_i) \equiv e^{\sigma^2}(e^{\sigma^2} - 1)e^{2\mu}$, $E(\ln Y_i) = \mu$, and $V(\ln Y_i) = \sigma^2$.

This distribution was derived from two first principles: (1) *Z_i is a standard Normal variable* and (2) *Y_i is a convenient mathematical function of Z_i in*

Equation (3.20). If this function is not motivated completely from substantive arguments in some applications, at least the function is familiar and the resulting distribution more appropriately describes some characteristics of the aggregate level random variables.

Where derivation from first principles is difficult or indeterminate

The distributions presented above can all be neatly derived from different assumptions about a social system. In general, this sort of derivation from first principles is the best means of choosing a probability distribution for use in the stochastic component of a statistical model. However, in many research situations, such a derivation is very difficult, requires assumptions that are unrealistic simplifications of the true stochastic process, leads to multiple possible distributions, or is just analytically intractable. In these instances, one can choose an existing distribution, or create one, that seems to cover most interesting cases. This procedure is often an adequate compromise, permitting further analysis where otherwise none would be possible.

For example, let Y_i be the proportion of citizens in a legislative district who would cast their ballots for the Democratic candidate in a two-party system. Y_i varies between zero (no Democratic votes) and one (all Democratic votes) *across* legislative districts *within* a state. Since Y_i is an unobserved random variable, one needs a probability distribution, which I call the *mean voter preference distribution,* from which an election in each district is randomly drawn. This distribution must be flexible enough to include cases where it is unimodal, to allow for competitive systems with most districts in the state having proportions near 0.5, bimodal, to allow for uncompetitive party systems with many successful incumbents in both parties, skewed, in the case of bias toward one of the parties, and combinations of these. A histogram of the proportion voting Democratic in each district across an actual state is an empirical version of this distribution, but a model requires an explicit density abstracting the key features of this histogram.

Deriving a distribution by making assumptions about individual voters or their geographical arrangement turns out to narrow the range of possible distributions only negligibly as Quandt (1988) and Gudgin and Taylor (1979) demonstrate. Alternatively, one can choose an existing distribution which is flexible enough to handle most interesting aggregate forms. Unfortunately, none exist in the literature.[12] Thus, as part of a stochastic model and an em-

[12] Obvious choices include the Beta distribution or one of those developed for correlation coefficients. Only the Beta distribution allows for bimodality, but it is not flexible enough for present purposes (see Johnson and Kotz, 1970b: Chapter 24).

pirical analysis of legislative redistricting (King, in press-d), I derived a new tractable probability distribution specifically designed to handle most of these forms. This distribution is defined on the interval (0,1); special cases of it are unimodal, bimodal, peaked, uniform, skewed, and various combinations of these features.

This mean voter preference distribution may be defined as follows:

$$f_{mvp}(y_i|\rho,\lambda) = \rho e^{\lambda}\left[e^{\lambda} + \left(\frac{y_i}{1-y_i}\right)^{\rho}\right]^{-2} y_i^{-(1-\rho)}(1-y_i)^{-(1-\rho)}. \quad (3.22)$$

This distribution has two parameters: λ, which indexes direction and degree of skewness, and ρ, which indexes peakedness (ranging from a single spike to extreme bimodality).

3.3 Multivariate probability distributions

A univariate probability distribution provides a model of the stochastic component for a single random observation, Y_i. Suppose instead that Y_i were a vector of N random variables. To model this vector, a multivariate distribution must be used.

For example, the multivariate Normal distribution for the N random observations Y_i can be written as a function of an $N \times 1$ vector μ and $N \times N$ symmetric variance-covariance matrix Σ:

$$f(y_i|\mu,\Sigma) = (2\pi)^{-N/2}|\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(y_i-\mu)'\Sigma^{-1}(y_i-\mu)\right\}. \quad (3.23)$$

Several features of this distribution are worthy of note. First, if Y_i includes only one random variable, then $N=1$, $\Sigma=\sigma$, and this multivariate Normal distribution reduces to the univariate Normal distribution in Equation (3.18). Second, suppose Σ is a diagonal matrix (i.e., Σ has variances on the diagonal and zeros for all the off-diagonal covariances). In this case, it is easy to show that the random variables are independent. Hence, in this special case, the product of univariate Normal distributions $f_n(y_{1i}|\mu_1,\sigma_1^2)f_n(y_{2i}|\mu_2,\sigma_2^2)\cdots f_n(y_{Ni}|\mu_N,\sigma_N^2)$ can be shown to be equal to the multivariate Normal distribution. The proof of this assertion comes from Equation (3.1).

In the general statistical model [Equations (1.3) and (1.4)], an important special case is when $Y_1, \ldots, Y_n$ are independent random variables. By assuming independence, one can derive a multivariate distribution from only the product of univariate distributions. This situation is commonly referred to as the absence of autocorrelation (Chapter 7 demonstrates how to model processes without assuming independence). In the even more special case where every random observation has the same probability distribution, except for a

parameter vector, modeling the stochastic component only requires one to specify a single univariate probability distribution.

3.4 Concluding remarks

The univariate probability distributions given above represent an extremely small proportion of known distributions. A variety of others are introduced and derived throughout the remainder of this book as needed. Literally thousands of others have been invented. Yet, many interesting data sets still exist for which no adequate probabilistic models have been developed. For many problems, political scientists can afford to be merely consumers of "pure" developments in probability theory and distribution theory. However, since many situations remain for which we cannot rely on statisticians to come to the rescue, political scientists must begin to learn more about probability distributions and, more generally, about stochastic processes. To the extent that political processes differ from natural and physical ones – areas which statisticians seem to pay most attention to – political scientists will be responsible for the development of their own stochastic models.

Probability theory has two roles in scientific inference. First, it is the primary means by which the stochastic components of statistical models are built. Indeed, the systematic component of statistical models uses the probability distribution's parameters to model the systematic portions of statistical relationships. Second, probability theory is the critical tool in likelihood inference. Chapter 4 uses probability for this latter purpose. The remaining chapters tap both uses of probability theory.