Chapter Ten

An Introduction to (Discrete) Distributions

As discussed in Chapter 1, variables are indicators of concepts, and they take several values. If we look at a population (or sample), we often want to know how many people or states or other variables of interest hold each value. Put differently, we want to know the *distribution* of cases across the values of the variable. One can use mathematics to develop (1) an understanding of how the cases are distributed in a given population (sample) or (2) an expectation of how cases should be distributed given one's beliefs about the process that produces the variation in that concept (or variable). The first task is an empirically based exercise while the latter is conceptual, but we generally undertake the exercise as a precursor to conducting statistical analyses that compare our actual data against our theoretical expectations. The second task is also fundamental to game theory, as it relates to the utility an actor expects to receive in an uncertain scenario. In this chapter we introduce you to the mathematics involved in understanding (1) frequency distributions (which are empirical) and (2) probability distributions (which are theoretically constructed). The discussion of probability distributions is limited to concepts (variables) that take discrete values. In Chapter 11 we discuss the probability distributions of concepts (variables) that can take continuous values. We stress that the difference between discrete and continuous probability distributions is largely technical, and we separate them into different chapters because calculus is needed for continuous distributions but not for discrete ones.

The first section of this chapter covers distributions of one variable generally, and defines random variables, i.e., those variables that can take on multiple values with some likelihood. The second section details empirical sample distributions. The third discusses empirical joint and marginal distributions. The fourth details the theoretical probability mass function. The fifth presents the cumulative distribution function for discrete random variables. The sixth section presents examples of probability mass functions. Finally, the seventh section describes the concept of an expectation of a random variable, its relation to the moments of a distribution, and the notions of expected value and expected utility, which are fundamental to game theory.

10.1 THE DISTRIBUTION OF A SINGLE CONCEPT (VARIABLE)

There are several ways to characterize the distribution of a concept (variable). You are already familiar with the frequency and relative frequency distributions (though you may not know they have names) as they are widely used in media reports and textbooks. Before introducing frequency distributions, we would like to make a point about why you should study distributions. This is the headline: science involves generalization, and generalization involves the distributions of concepts (variables).

10.1.1 From Specifics to Generalizations

Science involves the identification of general causal processes. By causal we mean that one concept influences another such that change in the first consistently produces change in the latter.[1] By general we mean processes that apply to all or most members of a class or population: the causal process is not unique to individual members of the class or population. Most causal theory in political science is probabilistic: the hypothetical relationship between a causal variable X and a caused variable Y is not expected to hold in every case. In fields such as physics, theorists often posit laws that are expected to hold in all cases (e.g., the laws of thermodynamics). Political scientists rarely (if ever) posit laws. Instead, we posit probabilistic hypotheses about the impact of X on Y. That is, theories in political science posit causal relations that apply to most members of a class or population at most moments in time. For example, theories of voting that posit the hypothesis that party identification (e.g., Republican) is positively associated with vote choice (e.g., George W. Bush in 2000) are probabilistic: finding a registered Republican who voted for a Democratic candidate does not falsify the hypothesis.

Whether probabilistic or nomologic,[2] theories about political behavior and institutions are generalizations: they do not apply to a single specific individual. To be clearer, an article that discusses Charles de Gaulle's impact on French politics is descriptive; an article that posits a theory about the impact of the office of the president on French politics is general. An even more general theory would develop hypotheses about the impact of presidents on politics in democracies (and would require one to define the terms used in this sentence, as there is a great deal of variation among presidencies in democracies). Once we shift our attention away from specific individual members of a class or population toward most members of a class or population, we have already begun to think

[1] This is similar to the "inductive regularity" definition of causation found in Little (1991). As Little demonstrates, the "causal mechanism" definition of causation is superior to this one, but that distinction is not of immediate concern here. See Little (1991, pp. 13–29) for a detailed discussion.

[2] A nomologic theory proposition applies to all members of a class or population at all times.

about distributions. And the mathematical study of distributions provides some powerful tools to aid one in both developing theory and testing the hypotheses implied by theories. Put differently, if one is interested in theory (which, by definition, involves generalization) rather than description, then probability distributions are a useful tool at one's disposal. As such, it is important to develop a working familiarity with probability distributions.

Before moving on we would like to advance an important claim. While political scientists widely recognize the centrality of probability distributions to statistical analysis (an important tool for testing hypotheses), a considerably smaller group recognizes the usefulness of understanding them for developing theory. As such, instruction in distributions is almost exclusively restricted to the first statistics course and generally consists of a single lecture.[3] While recognizing that probability distributions are critical to the study of statistics, and thereby hypothesis testing, we argue that thinking about distributions is equally important for creating theories (and developing hypotheses). By showing how an understanding of probability distributions sharpens one's theorizing about politics, this chapter is intended to help correct an imbalance we have in our field. We discuss this specifically in the final section covering expected utility.

However, much as in the previous chapter, the formalism of distributions is the same whether or not we are referring to statistical or measurement error, uncertainty, errors in decision making, or any other source for probabilistic hypothesizing or theorizing. Along these lines, the same formalism is used whether we're discussing, for example, the distribution of individual characteristics or behavior across a population or the distribution of individual behavior across opportunities to act (e.g., vote sometimes and not others). In short, we reiterate that distributions are a remarkably powerful and useful tool in political science.

10.1.2 Random Variables

In a footnote we noted that a distribution defines the set of values that a random variable may take, but this notion is too important to leave to a footnote, as this connection implies that random variables are as fundamental to political science as distributions. For the same reason we noted that while hypotheses in political science (and many theories as well) are probabilistic, the components of hypotheses (and many theories) are random variables. So what are random variables?

Random variables are those that have their value determined, in part, by chance: the value they take in any given circumstance can be described probabilistically. More precisely, a **random variable** is a variable that can take some array of values, with the probability that it takes any particular value defined according to some random process. The set of values the variable might take is the **distribution** of the random variable and, as we show below, the function that defines the probability that each value occurs is known as the probability

[3] What we are calling distributions is also often discussed under the label "random variables," as distributions describe the set of values that random variables may take.

mass function or the probability distribution function, depending on whether or not the distribution of values is discrete or continuous.

Saying that the variables in our hypotheses (and many theories) are random variables amounts to saying that we expect the value of each variable in our hypotheses (and many theories) to be a draw from some associated distribution. A simple example of this would be the roll of a fair die. The random variable corresponding to the die's value would have an equal probability (one-sixth) of having each of the integer values from 1 to 6. This is called a uniform distribution. However, more complicated distributions are allowed, particularly when we're discussing dependent random variables. For example, we could say that one's decision to vote is a random variable Y that can take the values of 1 (for yes) and 0 (for no). The probability that it takes the value 1 will depend on many other factors, such as past voting behavior, education, income, etc. In other words, the probability distribution of values for Y is a conditional probability distribution: one for which the likely outcomes vary as one or more other variables vary in value (more on this below). Increasing the value of an independent variable such as education or income might shift the distribution of 1s and 0s toward more 1s, implying that a more educated or higher-earning person will be more likely to vote, and more likely to vote more often, given multiple opportunities to do so.[4] This is obviously weaker than a deterministic statement like "If you earn more than $100k a year you will vote," but political scientists generally believe such probabilistic statements are relevant to the real world.[5]

This is most of what one needs to know about random variables, but a little bit of terminology is helpful before moving on. As we've stated, a random variable can take many values.[6] The **realization** of a random variable is a particular value that it takes. When it's not confusing, we often use capital letters, such as Y, to correspond to random variables and lowercase letters, such as y, to correspond to particular realizations of a random variable. The statement $Pr(Y < y)$ then reads "the probability that Y is less than a particular value y." This will prove important later.

The probability that a random variable has *any* value is given by its probability distribution; the **support** of this distribution is the set of all values for which the probability that the random variable takes on that value is greater than zero. This terminology is commonly used for continuous probability distributions. Discrete random variables have discrete associated probability dis-

[4] The connection between these two statements relates to the expectation of a random variable, which we discuss in Section 7 of this chapter.

[5] Even theories that make deterministic claims typically highlight the way in which the claim varies with parameters of the model (i.e., independent variables). These *comparative statics* of theories, which we'll discuss in Chapter 17 of this book, are weaker statements than the theory's point predictions and more directly testable. Consequently, these are generally more believable.

[6] A slightly more complex way of thinking about this is that a random variable stochastically determines that some event happens, out of all the events that might have happened, and assigns a value (typically a real number) to that event.

Table 10.1: Top Problem Facing the United States, September 2001

Problem	% Listing it No. 1
Terrorism	60
Economy	8
No opinion	7
Noneconomic, other	6
Moral decline	6

tributions; continuous random variables have continuous associated probability distributions.

10.2 SAMPLE DISTRIBUTIONS

Sample distributions are empirical constructs. We distinguish them from probability distributions, which are constructed theoretically using the rules of classical probability. We discuss probability distributions below.[7]

Sample distributions are representations of the number of cases that take each of the values in a sample space for a given portion of the population of cases. Put differently, a sample distribution is the distribution of values of a variable resulting from the collection of actual data for a finite number of cases. It turns out that you have encountered many sample distributions in various textbooks, news articles, and other places.

10.2.1 The Frequency Distribution

The first sample distribution to consider is the frequency distribution. It is a count of the number of members that have a specific value on a variable. Public opinion scholars often ask a question about the issues that are most important to the country, thereby producing a rank order of several issues (e.g., crime, the environment, poverty, terrorism). The frequency distribution for that variable would be the number of respondents to the survey who listed the first issue as most important, the second issue as most important, etc. (Table 10.1 reports such a survey conducted in Georgia).[8]

Students of coalition governments are often interested in a variable that records the number of seats a party won in a given election. The frequency distribution for that variable simply lists the number of seats each party obtained (an example from Lithuania is given in Table 10.2).[9]

Studies in international relations often include a variable that records whether a given country initiated a militarized dispute with another country during a

[7] Many statistics textbooks refer to probability distributions as *population distributions*.

[8] Source: Peach State Poll, `http://www.cviog.uga.edu/peach-state-poll/2001-12-07.pdf`.

[9] Source: `http://www.electionworld.org/lithuania.htm`.

Table 10.2: Lithuanian Parliamentary Seats, 2000

Party Abbreviation	Seats Won
ABSK	51
LLS	33
NS	28
TS-LK	9
LVP	4
LKDP	2
LCS	2
LLRA	2
KDS	1
NKS	1
LLS	1
JL/PKS	1

Table 10.3: Militarized Interstate Dispute Initiators, 1816–2002

MID Initiator	No. of Countries	% of Countries
No	67	31
Yes	147	69

given period of time. In Table 10.3 the frequency distribution for this variable is the number of countries that initiated a dispute at some time between 1816 and 2002 and the number that did not.[10]

One can create a frequency distribution for any variable. The level of measurement is irrelevant: whether discrete (i.e., nominal [aka categorical], ordinal, or integer) or continuous, we can produce a frequency distribution if we collect data on a sample. That said, a frequency distribution is only of interest if all members of the population do **not** have unique values on the variable of interest. If each case has a unique value, then all of the values in the sample frequency distribution will be 1.

10.2.1.1 Why Should I Care?

The frequency distribution is widely used; you have encountered it countless times in textbooks, news articles, and elsewhere. Given that it is so simple and widely used, why dedicate space to it in a graduate-level text? The reason is that it is critical to think in terms of distributions over concepts when developing theory, and the frequency distribution is very convenient and useful for that purpose. When thinking about a political process one often thinks about a specific example of the process in question (e.g., one's own vote in an election, the formation of a specific coalition government, or a particular militarized dispute).

[10] Source: Militarized Interstate Dispute Data as available in the EUGene software, `http://www.eugenesoftware.org`.

This is natural and especially common for new graduate students. However, it is not a very good practice for developing general theories of politics, and failure to recognize this not only reduces the likelihood that one will develop a useful theory but can also lead one to errors in research design when it comes time to test hypotheses (e.g., selection on the dependent variable; see King, Keohane, and Verba, 1994, pp. 129–37). Thinking about specific examples is definitely a good place to *begin* theory development, but having established a potential causal relation from the example in question, one must move to thinking about all members of the class or population in question and ask whether the relationship might hold for most members most of the time. Simple thought experiments combined with rough knowledge of the distribution of the measure of a concept can lead one to reject ideas without having to formally test them, and also illuminate new puzzles that warrant explanation.

To be a political scientist (rather than a scribe of politics) one must shift one's thinking away from specific examples toward general patterns. The frequency distribution is the most intuitive, simple distribution and thus one with which you will want to become very comfortable not only with respect to actual sample data (i.e., the values of variables) but also in the abstract, when you are theorizing (i.e., concepts).

10.2.2 The Relative Frequency Distribution

To this point, we have discussed distributions without concerning ourselves with probability. Let us remedy that. The probability that a specific case drawn at random[11] from a sample has a specific value, i, is the relative frequency of the value i.[12] The relative frequency of value i is the frequency (i.e., the number of cases with the value i) divided by the total number of cases. Put differently, the relative frequency of value i is the proportion of cases that have that value. As such, the relative frequency of a given value lies between 0 and 1, and the sum of all relative frequencies equals 1. Note that a probability of 0 indicates that there is no chance that a given value can be drawn from a sample, and a probability of 1 indicates that the value in question will be drawn with certainty. More generally, larger probability values indicate a greater likelihood that the value in question will be drawn.

The relative frequency distribution is a transformation of the frequency distribution (to reiterate and be specific, we divide the frequency by the total number of cases). It can also be represented in tabular or graphical form, is defined for all variables regardless of their level of measurement, and is uninteresting in samples where all cases have unique values. Finally, because most people are more familiar with percentages than with proportions, relative frequency distributions

[11] Note that this definition holds only when the case is *drawn at random*. This issue is discussed in more detail in most statistics texts or in a good research design text.

[12] We can define $i \in I$, where I is defined over the range from the minimum to maximum value in the sample.

are sometimes transformed to percentages (this transformation is conducted by multiplying the proportion by 100%).

10.2.2.1 Histograms

A histogram is a specific representation of the relative frequency distribution: it is a bar chart of the distribution of the relative frequencies in which the area under each bar is equal to the relative frequency for that value. In other words, the sum of the areas of each bar equals 1, and the area of each bar equals the probability that the value represented by the bar would be chosen at random from the sample depicted. The formula for the bars in a histogram can be represented by equation (10.1):

$$Pr(Y = y) = f_y \Delta_i, \tag{10.1}$$

where $Pr(Y = y)$ is the area covered by the bar (i.e., the probability that value y would be drawn at random), f_y is the height of the bar for value y, and Δ_i is the width of the bar.[13] By convention, one holds the width of each bar constant at 1 when dealing with a discrete distribution, thus making Δ_i known. Statistics packages that produce histograms (most of them do) need only solve for the unknown height, f_y, to produce the bar chart. Because the variation is only in the height of the bars (the widths are constant at 1), the histogram often provides an appealing graphical form to use to display a relative frequency distribution.

10.2.2.2 Why Should I Care?

The relative frequency distribution is relevant to political scientists for the same reason that the frequency distribution is of interest. As an example, Fearon and Laitin (1996) is a response to a spate of books and articles that sought to explain ethnic conflict.[14] These studies observed that an ethnic war occurred in one or more locations and offered explanations for the outbreak of such conflict. Fearon and Laitin begin by thinking about the distribution of ethnic conflict and observe that interethnic cooperation is far more common than interethnic conflict. They do not produce an actual relative frequency distribution to establish this point. Instead, they simply observe that socio-econo-political interaction across ethnic groups is extremely common: few human beings live their lives within ethnically homogeneous societies. Next they observe that violent conflict is rare. Again, they do not need to produce a relative frequency distribution over a specific spatial temporal domain to establish their point: most people most of the time are not engaged in violent conflict with other people. When one puts both

[13] Recall from geometry that the area of a rectangle equals the product of its height and its width.

[14] This interest was spurred by the collapse of the Soviet bloc, which (1) eliminated the Cold War as a topic of interest and (2) contributed to the collapse of Yugoslavia, which produced wars that demanded explanation.

of these points together, one observes that ethnic cooperation is common and ethnic conflict is rare.[15]

This claim causes problems for many of the theories that others have put forth: they focus their attention on explaining the presence of ethnic conflict without attending to the (relative) frequency distribution of ethnic conflict. Fearon and Laitin's attention to that distribution led them to recognize that a useful theory of ethnic conflict needed not only to account for the outbreak of such events but also to explain ethnic cooperation and the relative distribution of the two. That is, by thinking abstractly about the distribution of ethnic conflict and cooperation (note that they did not even have to collect any data or do any statistical tests—this was theoretical thinking linked to common knowledge about the rough empirical distribution) Fearon and Laitin cast dispersion on existing theories and provided themselves with a useful starting point for developing a new, better theory.

Fearon and Laitin (1996) show that by focusing on specific cases and theorizing only about the rare outbreak of such conflict, the entire field of ethnic conflict studies had erred.[16] Once one identifies this failure to think about the distribution of ethnic cooperation and conflict the weakness in such an approach seems obvious. Yet if that weakness was so obvious, then dozens of bright, talented political scientists would not have made the error, and the point raised by Fearon and Laitin would have been made long before 1996.

To summarize, the relative frequency distribution can be useful both as an abstract theoretical tool and as a concrete empirical tool. Becoming comfortable with it will prove useful to both the development and the testing of hypotheses.

10.3 EMPIRICAL JOINT AND MARGINAL DISTRIBUTIONS

The distribution of a single concept or variable can be of considerable theoretical and statistical interest, but causal theories of politics necessarily involve expected relationships among concepts or variables. As such, we want to study joint distributions; marginal distributions are a natural extension. For readability, we will focus on the case of two variables in this section.

10.3.1 The Contingency Table

A contingency table is the joint frequency distribution for two variables. While it can be created for both discrete and continuous variables, it is of considerably more value for discrete than for continuous variables.

With respect to construction, the contingency table is a matrix with one variable's values represented in the rows (typically the dependent or caused variable in empirical work) and the other variable's values represented in the

[15] This is so even if we attribute all violent conflict to ethnic cleavages.

[16] The study of war similarly suffered from this problem for decades (Most and Starr, 1989, pp. 57–58).

columns (typically the independent or explanatory variable). The cell entries record the number of cases that have the row value for the row variable and the column value for the column variable. The resulting matrix provides a quick summary of the joint distribution of the two variables.

Lest one think that contingency tables are only of value for empirical work, reconsider the discussion of Fearon and Laitin's (1996) article of interethnic cooperation and conflict. Though they did not produce a contingency table to illustrate their thought experiment, they could have done so. We have produced such a contingency table here.

Table 10.4: The Fearon and Laitin (1996) Contingency Table

	Ethnic Homogeneity	Ethnic Heterogeneity
Cooperation	Rare	Common
Conflict	Rare	Rare

One can readily see that ethnic heterogeneity is a poor explanation for the outbreak of ethnic conflict.[17] Of more importance, one can see that cooperation in ethnically heterogeneous communities is the modal[18] outcome, and thus it is important that theories of ethnic conflict be able to explain interethnic cooperation (i.e., theories need to account for the full range of phenomena, in this case both common and rare outcomes).

That said, contingency tables are particularly useful for the analysis of empirical relationships between discrete variables. You will learn more about contingency tables in your introductory statistics course.

10.3.2 Marginal Probabilities

Marginal probability is a label that often confuses students, likely because it arises from a practice that is rarely used apart from in contingency tables. Simply put, as noted in the previous chapter, the marginal probability of an event A is the probability that A will occur unconditional on all the other events on which A may depend. To make this work, one must in general sum the conditional probabilities of A on all the other mutually exclusive, collectively exhaustive events B_i on which it may depend, each weighted by the chance that the particular B_i will occur. From the discussion of Bayes' rule in the previous chapter, we know this amounts to writing the marginal probability $Pr(A) = \sum_{i=1}^{n} Pr(A|B_i)Pr(B_i)$. In words, this means that one averages over other events and focuses on the one event, A, of interest.

For example, let's say we had computed the joint probability of rolling a 7 on two dice and drawing a king from a deck of cards. If all we cared about

[17] Some of the scholars Fearon and Laitin (1996) criticize recognize this point. See, for example, Posen (1993).

[18] The modal outcome is the most common outcome.

Table 10.5: Militarized Disputes, 1946–92

	Nonterritorial	Territorial	
State A or B Democracy < 10	43	31	74
State A or B Democracy $= 10$	21	2	23
	64	33	97

was the chance of rolling a 7—the marginal probability of rolling a 7—then we'd ignore the deck of cards entirely and just stick to the chance of rolling a 7, which is one-sixth. This example is artificially easy because the two parts of the joint event, rolling dice and drawing a card, are independent. But the argument holds for dependent events. For example, let's say we are interested in the probability of voting, but voting is conditional on whether or not it is raining. The conditional probability of voting given rain might be $Pr(V|R) = 0.4$. But what about the marginal probability of voting? To get this we need the conditional probability of voting given that it is not raining, as well as the probability of rain. Let's say the former is $Pr(V| \sim R) = 0.6$ and $Pr(R) = 0.3$. Then $Pr(\sim R) = 0.7$ and so the unconditional, marginal probability of voting is $Pr(V) = Pr(V|R)Pr(R) + Pr(V| \sim R)Pr(\sim R) = (0.4)(0.3) + (0.6)(0.7) = 0.54$.

However, when we are looking at empirical probability only, particularly when there are only two variables, the concept becomes much easier and the word "marginal" makes much more sense. To find the marginal probabilities, we sum the simple conditional probabilities of one variable across all values of the other variable.[19] The label "marginal" comes from the fact that the marginal probabilities are written in the margins of $n \times n$ (read "n by n") tables.

For example, Mitchell and Prins (1999) are interested in the frequency with which countries with strong democratic institutions get in militarized disputes over territory relative to countries without strong democratic institutions. Table 10.5 reproduces a portion of the evidence they report. With that information we can determine the empirical probabilities that (or the relative frequency with which) the compound events occurred. Tables 10.6 and 10.7 contain cell entries that represent the empirical probabilities. The marginal entries are the sum of the probabilities in the row or column, respectively, relative to all events. Each of these marginals sums to 1.

Let's walk through it. To calculate the cell entries in the center of Table 10.6 we need to determine the empirical probabilities across the rows. The upper left cell entry is the number of dyads (i.e., country pairs) without a strong democracy that became involved in nonterritorial disputes (43) relative to the total number of dyads without a strong democracy (74): $\frac{43}{74} \simeq 0.58$. The lower right cell is the number of dyads with a strong democracy that became involved in territorial disputes relative to the number of dyads with a strong democracy: $\frac{2}{23} \simeq 0.09$.

[19] A more accurate statement is that the above holds for discrete variables. For continuous variables one integrates over, rather than sums across, the values of the other variable. We discuss this issue in more detail in the next chapter.

Table 10.6: Row Probabilities

	Nonterritorial	Territorial	Row Marginals
State A or B democracy < 10	0.58	0.42	0.76
State A or B democracy $= 10$	0.91	0.09	0.24

Table 10.7: Column Probabilities

	Nonterritorial	Territorial
State A or B democracy < 10	0.67	0.94
State A or B democracy $= 10$	0.33	0.06
Column marginals	0.66	0.34

The same logic holds for the other two entries in the center of the table. The values of the row marginals at the right are the ratio of all the cases in each row to the total number of cases. The upper value is the number of dyads without a strong democracy (74) relative to all dyads (97): $\frac{74}{97} \simeq 0.76$. Note that the two row marginals in Table 10.6 sum to 1: $0.76 + 0.24 = 1.0$. The same is true of the values in each row in the center of the table.

We calculate the column probabilities in the same way except that we calculate the column totals, not the row totals. For example, the lower left cell in the center of Table 10.7 is the frequency of dyads with a strong democracy that became involved in a nonterritorial dispute (21) relative to all dyads that became involved in a nonterritorial dispute (64): $\frac{21}{64} \simeq 0.33$. Similarly, the upper right cell is the number of dyads without a strong democracy that engaged in territorial disputes (31) relative to the total number of dyads involved in territorial disputes (33): $\frac{31}{33} \simeq 0.94$. Finally, the cells listing the column marginals represent the relative frequency of each column to the total number of cases. Thus, there are 33 territorial disputes out of 97 total disputes: $\frac{33}{97} \simeq 0.34$. Observe that the marginals sum to 1: $0.66 + 0.34 = 1.0$.

Marginal probabilities are used in statistical inference. For example, the χ^2 (chi squared) statistic is used to evaluate the statistical significance of an association between two ordinal level variables.

10.4 THE PROBABILITY MASS FUNCTION

Above we discussed frequency distributions, which one can use to describe how the values of a variable for which we have collected data are distributed within a population or sample. It is also possible to use the laws of probability to develop expectations about how one *expects* the values of a concept (variable) to be distributed *given* one's beliefs about the process that generates the values that concept will take. Those of you who have had a course in statistics have already studied a number of these distributions, but even those of you who have never studied statistics have probably heard of the "bell curve." The bell, normal, or Gaussian curve is a probability distribution, though since it describes concepts (variables) that can take continuous values, we do not discuss it in this chapter.

All functions are a mapping of the values in one set to another, and a probability function describes the likelihood of each of the values a concept (variable) might take given a description of the process that generates it. It turns out that it is possible to develop a number of different probability functions to describe the distribution of a concept (variable), but we limit the discussion to two of them: the probability mass (or distribution) function and the cumulative density function. You should know, however, that there are other functions: the hazard, cumulative hazard, survival, inverse survival, and percent point functions can also be used to characterize different distributions. While the probability mass (or distribution) function and the cumulative density function are the most widely used in statistics texts popular in political science, a full understanding requires additional coursework or self-study. But this is just an introduction, and a sound understanding of the material contained here will nicely pave the route for additional study for those of you who become interested in pursuing it.

In the chapter on probability we looked only at the classical probability of a single value being drawn randomly from a sample. Political scientists are often more interested in being able to say something about a range of values. The probability mass function (PMF) allows us to do this. As important, it also allows the chance of drawing any particular value to vary by value. The PMF is a function that specifies the probabilities of drawing discrete values (we discuss the density function of continuous variables in the following chapter).[20] As we show below, the PMF of a discrete variable is related to the relative frequency distribution.[21] More specifically, the PMF is a function that connects the various classical probabilities of specific values for a sample.

Another way to think about this is that the PMF is a function that allows one to sum a series of weights. More specifically, the weights are the probabilities that each value will be randomly drawn from the sample. The PMF makes it possible to identify different probability distributions, 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).

The PMF of a discrete (i.e., nominal, ordinal, or integer) variable assigns probabilities to each value being drawn randomly from a population. The relative frequency distribution of a discrete variable is a tabular or graphical representation of the empirical probabilities that each value is drawn at random from the sample. In other words, the PMF is the function that describes the expected relative frequency distribution.

More formally, the PMF of a discrete variable, Y, may be written as $p(y_i) = Pr(Y = y_i)$ where $0 \leq p(y_i) \leq 1$ and $\sum p(y_i) = 1$, and Y is the variable and

[20] More formally, let X be a discrete random variable that maps elements in a sample space to the real numbers. Then a probability mass function is a map from the set of real numbers to probabilities in $[0, 1]$: $f_X(x) : \mathbb{R} \rightarrow [0, 1]$.

[21] These functions are sometimes referred to more generically as the probability distribution.

y_i is a specific value of Y. This function describes the height of the bars of a histogram, as depicted in Figure 10.1.

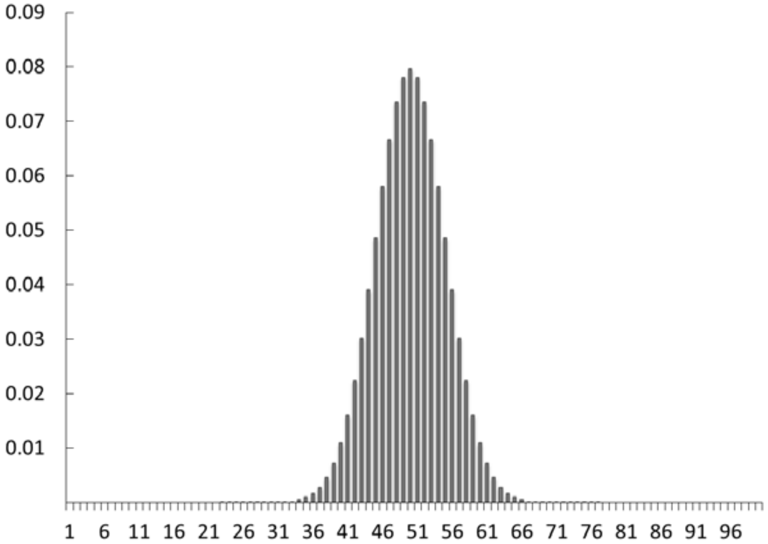

Figure 10.1: PMF of a Binomial Distribution, $n = 100, p = 0.5$

You might be wondering why we discuss discrete variables here but leave a discussion of continuous variables for the next chapter. The reason is that continuous variables have an infinite number of possible values, and the probability that we will randomly select any given value from a set with infinite support is zero.[22] As such, the probability function of a continuous variable is a bit more complex.

10.4.1 Where Does a Specific PMF Come From?

The representation of a PMF above is general—we did not identify a specific functional form. Many people who look at a given PMF find it far from obvious where it came from or what it actually means. Here we briefly describe the mathematical process used to create a PMF. Below we introduce a number of specific PMFs and verbally describe the data-generating process that the PMF describes.

In Chapter 3 we observed that one does not want to memorize a set of functions and then build theories of politics by applying each until one finds one that fits both one's conjectures and the relevant data. Rather, one wants to become familiar with functions so that one can discipline one's speculations and

[22] We see in the next chapter that the way around this is to ask instead what is the probability of selecting any value within a small region. This will be non-zero, but answering this question requires calculus. Hence we discuss it in the next chapter.

conjectures and make them more explicit. King (1989, p. 38) makes the same point with respect to the distributions that underlie our statistical models: "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."[23] King is arguing that rather than simply employ a statistical model off the shelf because someone put it there, we want to think about the political process that produced the concept or variable that interests us and then use the rules of probability theory to write down a function that describes that process. A PMF is precisely that: a function that uses probability theory to describe the process that generated the expected frequency of a discrete concept (variable) across cases. King (1989, p. 41) explains it this way:

> each distribution [presented in statistics textbooks] was originally derived from a very specific set of theoretical assumptions. These assumptions may be stated in abstract mathematical form [i.e., as a function], 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 derived represent models of interesting political science situations and are much closer to both data and theory. When a list of these principles is known, understanding them is critical to the correct application of a particular probability distribution.

We reviewed the principles of probability in the previous chapter. While few political scientists actually write down their own probability functions in empirical work, it is important to gain a familiarity with how one can go about it so that one can understand where the different probability functions that are commonly used come from and make informed choices about which is most appropriate.[24] Further, those who study game theory will come to use several probability functions more directly, particularly when computing expected values and expected utilities, the topic of the last section of this chapter.

10.4.1.1 Why Should I Care?

It is possible to assign qualitative (i.e., nominal and ordinal) and integer values to a wide array of political phenomena. We can measure political attitudes, record vote choice, count the number of seats a party holds in the legislature or

[23] A stochastic model is just a model that employs random variables. The most common way we see this in political science is via statistical models, and statistics courses tend to talk most about stochastic models. However, many formal models are also stochastic, including game theoretic models that involve lotteries over payoffs or opponents' strategies or uncertainty, and many models of bounded rationality.

[24] Those who wish to respond to King's call and develop their own probability functions will need to pursue additional studies in probability and statistics, most likely in their statistics department.

the number of militarized disputes in which a country has participated, etc. It should not surprise you that if one graphed the relative frequency distributions of the many and varied variables that political scientists have created, these graphs would not all look the same. It might surprise you, however, to learn that each and every one of them (roughly) approximates a PMF (for discrete variables) that statisticians have studied and named. Further, statisticians have gone to the trouble of developing and naming distributions for the express purpose of developing statistical models that can be used to test causal hypotheses. So if you are interested in applying the considerable power of inferential statistics to your own research, it is critical that you gain a working familiarity with PMFs. Doing so will put you in a position to choose statistical models appropriately and thus ensure that your statistical hypothesis testing is valid. Further, as King urges, some of you may want to focus on political methodology in which case you may end up writing down new distributions that are more appropriate for given theories in political science than any of the probability distributions developed to date. And, as we noted, if you read or employ game theory in your research, you will use probability distributions in computing expectations.

10.4.2 Parameters of a PMF

We introduced the concept of a sample space in the previous chapter. Here we need to define the terms **parameter** and **parameter space**. In discussions of probability functions, the term "parameter" refers to a term of known or unknown value in the function that specifies the precise mathematical relationship among the variables. Further, parameters are independent of the values of the sample space. Parameters can take multiple values, and the parameter space is the set of all values the parameters can take.[25] More specifically, "the functional form of [a probability distribution] and the value of the parameters . . . together determine the shape, location, and spread of the distribution" (Hendry, 1995, p. 34).

That description is jargon-laden, and we offer some examples momentarily. But first, note that the general representation above did not contain any parameters: it is not specific about the mathematical relationships among the variables. The specific PMFs that we introduce below contain parameters.

To illustrate, let's consider the case of voter turnout where we ask, "Which registered voters cast ballots?" There are two outcomes for each voter: (0) *did not cast a ballot* and (1) *cast a ballot.* We can write the following PMF:

$$\begin{aligned} p(y_i = 0) &= \pi, \\ p(y_i = 1) &= 1 - \pi. \end{aligned}$$

[25] We referenced the parameter space way back in Chapter 1, as an example of a space of interest in game theory. As you can see, it is of interest in statistics as well. It means the same thing in both cases: parameters help to dictate the dependence of functions on variables, each parameter can take a range of values, and the space of all values all parameters can take is known as the parameter space.

Here, π is the parameter of this PMF.[26] Because probabilities range between 0 and 1, $\pi \in \Theta = [0, 1]$. And Θ is the parameter space, as it is the set of all values the parameter π can take. We do not know what value to assign π, but we could turn to voter turnout data from previous elections to develop an expectation about π.

To further illustrate, let's consider the example of tossing a fair coin. It turns out that the PMF is precisely the same as above (e.g., [0] *heads* and [1] *tails*), except that in this example we know the value of π: it equals 0.5. If the coin is biased and turns up "heads" twice as often as "tails," then $\pi \simeq 0.67$ and $1 - \pi \simeq 0.33$. If the coin is biased and produces "tails" three-fourths of the time, then $\pi = 0.25$ and $1 - \pi = 0.75$.

Note that in these examples, when we change the value of the parameter the distribution of outcomes change. That is precisely what a PMF should tell us: the distribution of outcomes across the values of the variable given parameter values. The parameter space is the set of values the parameter might possibly take, and it thus dictates the set of possible distributions of outcomes.

Two important parameters are the location and scale (dispersion) parameters, and we introduce those below. It is important to note that not all probability distributions have parameters. When we discuss a specific distribution below we note whether it has any parameters, and if so, which ones.

10.4.2.1 Location and Scale (Dispersion) Parameters

Many distributions have a location and a scale (dispersion) parameter. Some have only a location parameter, and still others do not have any parameters.

The **location parameter** specifies the location of the center of the distribution. Thus, as one changes the value of the location parameter, one shifts the graph of the distribution's PMF to the left or right along the horizontal axis. For some distributions (especially the normal distribution, a continuous distribution we discuss in the following chapter), the location parameter has an empirical referent known as the **mean**. The location parameter (mean) is often represented in classical (empirical) probability by the Greek letter μ (mu).

For example, Figure 10.2 displays graphs of the PMF of a Poisson distribution (we introduce the PMF of the Poisson and other distributions below). It has only one parameter, a location parameter, μ. We have set μ equal to 1, 3, and 5. As the location parameter gets larger (i.e., changes from 1 to 3 to 5), the center of the distribution (i.e., its highest point) moves to the right. For all distributions with a location parameter, larger values will move the center of the PMF to the right and lower values will move the center to the left.

Second, the **scale parameter** provides information about the spread (or scale) of the distribution around its central location. As such, changing the scale parameter stretches or squeezes the graph of the PMF. Compared with a scale parameter equal to one, values greater than one increase the width of the graph

[26] It turns out that this is the PMF for the Bernoulli distribution. We discuss it in more detail below.

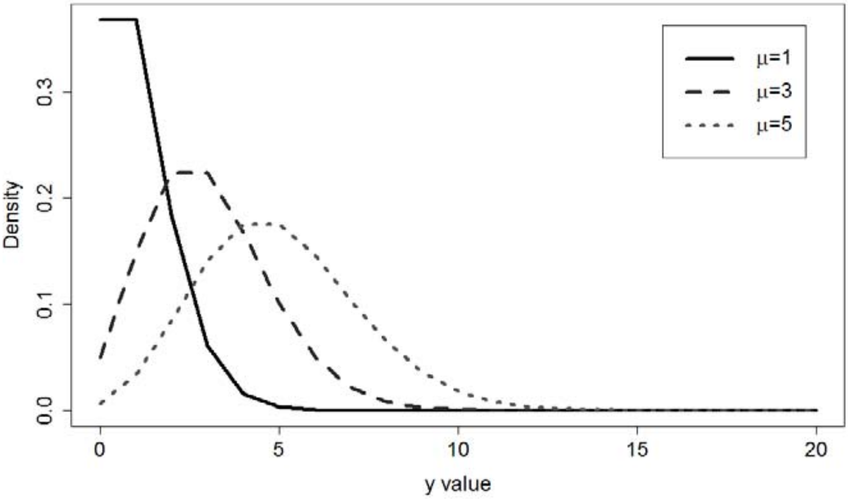

Figure 10.2: PMF of a Poisson Distribution, $\mu = 1, 3, 5$

(i.e., indicate both a lower minimum and higher maximum value than a PMF with a scale value of one). Similarly, scale parameters between zero[27] and one squeeze the PMF relative to one with a scale value of one: the minimum value is greater and the maximum value lower than they are for a PMF with a scale value of one. The scale parameter has an empirical referent known as the **standard deviation**, which is a measure of the distance of the distribution's values from its mean (or average) value. Both the scale parameter (classical probability) and the standard deviation (empirical probability) are usually represented with the Greek letter σ (sigma).

The **dispersion parameter** is the square of the scale parameter. As such, it also describes the spread about the central location of the distribution, except it emphasizes extreme values. Most treatments consider the decision to identify the scale or the dispersion parameter when defining the parameters of a classical probability distribution an arbitrary choice. In statistics, the dispersion parameter corresponds to the **variance** of an empirical distribution, and both are typically identified as σ^2 (sigma squared).

10.4.2.2 The Standard Form

The standard form of a distribution (or standard form PMF) is one in which the location parameter is set to zero and the scale parameter is set to one.

[27] Scale parameters cannot take negative values.

10.4.2.3 Why Should I Care?

The parameters of a classical probability distribution are important because they help us define the probability function (or PMF, for discrete distributions). We can also use the parameters of a distribution to help us determine whether data we have collected closely match the expected distribution given our beliefs about the process that produced the data. For example, we might think that the number of appointments a US president makes to the Supreme Court is produced randomly by what is known as a Poisson process (described below). The Poisson distribution has one parameter (location). Ulmer (1982) proffers just that hypothesis. He collected data on the number of Supreme Court appointments over the period from 1790 to 1980 and used it to estimate the location parameter to see whether it is likely that those data were produced by a Poisson process (you will learn about parameter estimation in your statistics courses).[28] He finds that the data were likely produced by a random Poisson process. Similarly, Midlarsky, Crenshaw, and Yoshida (1980) use the Poisson distribution to study contagion among transnational terror events data.

10.5 THE CUMULATIVE DISTRIBUTION FUNCTION

When conducting hypothesis tests it is often useful to determine the probability that a value drawn at random from a sample is above or below a specific value. The cumulative distribution function (CDF) describes the function that covers a range of values below a specific value and is defined for both discrete and continuous random variables.[29]

The CDF for a discrete random variable is, hopefully, intuitive: if we want to know the cumulative (or total) probability that a random draw from a population produces a value less than some quantity, then we need to add together the individual probabilities of each of the values below that number. We sum the individual probabilities because the values are mutually exclusive, and the joint probability of mutually exclusive events is the sum of the probabilities of the individual events. We can write the CDF for discrete variables as

$$Pr(Y \leq y) = \sum_{i \leq y} p(i). \tag{10.2}$$

Equation (10.2) states that we sum the probabilities of each value for all values less than or equal to y.[30] Sometimes you will see the notation $f(x)$ for a probability distribution function (PDF or PMF) and $F(x)$ for a CDF; using that notation makes clearer the connection between the two functions. We discuss this

[28] The term *parameter* can be used somewhat differently in probability theory and statistics, though at times they mean the same thing. See the wikipedia entry at `http://en.wikipedia.org/wiki/Parameter`.

[29] The CDF is also sometimes called the distribution function (see, e.g., the MathWorld entry at `http://mathworld.wolfram.com/DistributionFunction.html`).

[30] To add some precision, we can note that $0 \leq Pr(Y \leq y) \leq 1$, and that $Pr(Y \leq y)$ is increasing in y (i.e., $Pr(Y \leq y)$ gets larger as y gets larger).

more in the next chapter. Note that, since the values are mutually exclusive and all the values together are collectively exhaustive, $Pr(Y \leq y) + Pr(Y > y) = 1$, which implies that $Pr(Y > y) = 1 - Pr(Y \leq y)$. In words, the probability that a random draw exceeds some quantity is equal to one minus the probability that it does not exceed that quantity. Further, if y is the highest value that Y can take, then $Pr(Y \leq y) = 1$, since in this case we are adding the probability of all outcomes in the sample space. So all CDFs plateau at one.

Let's consider a concrete example. We might be interested in knowing the probability that a potential voter in the United States is partisan (e.g., self-identifies with either the Democratic or Republican Party). Imagine that we have a randomly drawn sample of survey data that records whether the respondents identify with the Democratic Party (value = 1), the Republican Party (value = 2), or a third party/no party at all (value = 3). Our party identification variable is discrete (in this case it is nominal), and we want to know the probability that a respondent drawn at random from the sample is partisan (i.e., has a value less than 3). Assume that the frequencies for the sample are

1. Democratic: 330,
2. Republican: 240,
3. Other: 180.

Given that there are 750 people in our fictitious survey, the relative frequencies are

1. Democratic: 0.44,
2. Republican: 0.32,
3. Other: 0.24.

To determine the probability that a respondent drawn at random from our sample is partisan, we add the probability that she self-identifies with the Democratic Party to the probability that she self-identifies with the Republican Party: $0.44 + 0.32 = 0.76$.

That is simple enough, but it is only one value, and the CDF is a function, so we need to specify the value of the CDF at all values the variable might take. For discrete variables this is straightforward (if tedious for variables with a large number of values). In the present example we need to know the probability that a randomly drawn respondent has a value less than or equal to 1 (0.44), the probability that the value is less than or equal to 2 (0.76), and the probability that it is less than or equal to 3 (1.0). The function that traces the graph of these values is the CDF for our example, as depicted in Figure 10.3.

10.5.1 Why Should I Care?

The CDF is widely used in the construction of hypothesis tests in inferential statistics. For example, we often want to know whether a given value is likely to

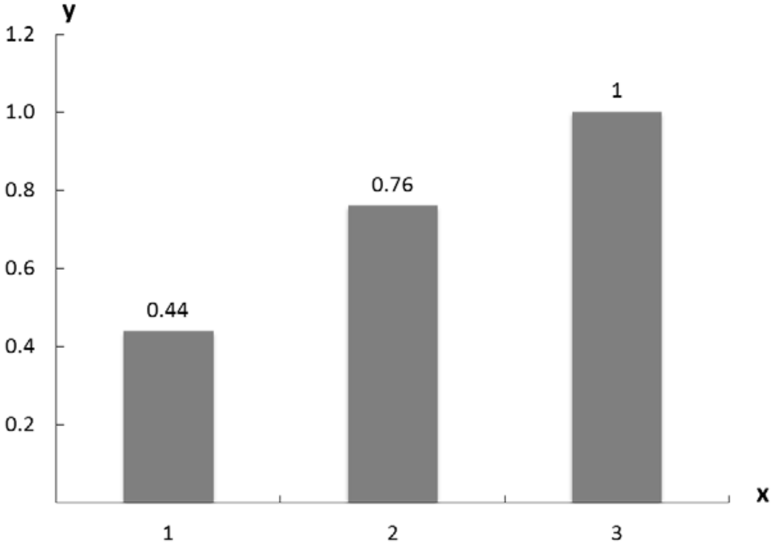

Figure 10.3: CDF of Party ID

have been drawn from a specific portion of a distribution (i.e., the chance that we would observe an outcome greater or less than some specific value). The CDF helps us answer that question. Pragmatically, you will find a nontrivial portion of the material in your introductory statistics course considerably easier to master if you have a working familiarity with the CDF.

In addition, the CDF is also used in some areas of game theory. Often one is interested in whether or not one thing that exhibits an element of randomness exceeds another—for example, whether the outcome from a policy is greater than a cutoff, or, in a Bayesian game, whether the type of opposing player has preferences sufficiently aligned so as to make a bargain possible. In cases like these, we need to know the probability that the random variable corresponding to outcome or type exceeds some value, which is just one minus the CDF of the distribution function at that value, as noted above.

10.6 PROBABILITY DISTRIBUTIONS AND STATISTICAL MODELING

Hendry (1995, p. 21) explains that to begin statistical modeling, "we first conjecture a potentially relevant probability process as the mechanism that generated the data" we are studying. Put differently, one begins by specifying a probability function that one suspects generated the data one is trying to explain. Thus, one reason we are interested in probability distributions is that they are formal statements of our conjecture about the **data-generating process** (DGP).

As an example, we might be interested in studying the probability that an American voter voted for the Democratic or Republican Party, the duration of a government in office in a parliamentary democracy, or the number of wars in which a country participated. Following King (1989, p. 38), the DGP is a formal statement of our beliefs about the probability process that produced the party the voter voted for, the length of time the government held office, or the number of wars in which a country was involved. We can use the rules of probability to specify a PMF or CDF of a given DGP.

Probability distributions are of interest to political scientists in large part because an adequate understanding of them makes it possible to use statistical inference to test the hypotheses implied by theories of politics. Let us be more specific. A person with no understanding whatsoever of distributions can read about (or develop) a theory of politics, identify one or more hypotheses implied by the theory, assemble a relevant dataset, read the data into a statistical software package (or even spreadsheet software), click some buttons, evaluate the output, and thereby invoke statistical inference as a test of the hypothesis that some X has an association with some Y. However, there is a wide (and growing) variety of statistical models available to test a given hypothesis, and the vast majority of them are *not* appropriate for any given hypothesis. The appropriateness of a given statistical model for a given hypothesis depends in large part (though not exclusively) on the distributional assumptions of the statistical model and the distribution of the dependent variable that measures the concept that is hypothesized to be caused by various factors. In other words, if one wants to draw valid inferences (and there is little reason to be interested in drawing an invalid inference), then one must match the distributional assumptions of the statistical model to the distribution of one's dependent variable. To do that, one must have a working knowledge of probability distributions. As noted above, PMFs identify different probability distributions of discrete variables (we discuss the distributions of continuous variables in the following chapter).

We present in this section the distributions most commonly used by political scientists by first writing an equation that identifies the PMF. We then describe the types of processes or events that most often produce variables with that distribution, and provide examples. It is important to understand that these are theoretical distributions or—if you will—the distributions of populations. Few samples of data will fit these distributions perfectly: a sample is a specific realization of a population, and any given sample will likely differ from its population. In your statistics courses you will learn more about the difference between population and sample distributions, including some formal tests one can invoke to determine the probability that a given sample was drawn from a specific distribution.

10.6.1 The Bernoulli Distribution

The first PMF we will consider applies to binary variables only and can be written as

$$Pr(Y = y|p) = \begin{cases} 1-p & \text{for } y = 0, \\ p & \text{for } y = 1. \end{cases} \tag{10.3}$$

Equation (10.3) states that the probability that $Y = 0$ is $1-p$ and the probability that $Y = 1$ is p, where $0 \leq p \leq 1$ (or $p \in [0, 1]$). Put differently, this says that if the probability that $Y = 1$ is 0.4, then the probability that $Y = 0$ is $1 - 0.4$, or 0.6.

We can also write the PMF for the Bernoulli distribution as

$$Pr(Y = y|p) = p^y(1-p)^{1-y}, \tag{10.4}$$

where $y = 0$ or $y = 1$. If we solve equation (10.4) for $y = 0$ and $y = 1$, we get the information provided in equation (10.3): $Pr(Y = 0) = p^0(1-p)^{1-0} = 1-p$, and $Pr(Y = 1) = p^1(1-p)^{1-1} = p$.[31]

The Bernoulli distribution is the building block for other discrete distributions (e.g., the binomial and negative binomial), and we will use the representation in equation (10.4) when we introduce other distributions.

The Bernoulli distribution describes randomly produced binary variables and is generally introduced using the example of flipping coins. But we can also think of political science events. For instance, we might use the Bernoulli distribution to model the expected frequency of valid versus spoilt ballots in an Australian national election (e.g., Mackerras and McAllister, 1999). Voting is compulsory in Australia, so those who wish to protest often mangle or otherwise spoil their ballot rather than cast a valid one.[32] Why might the Bernoulli distribution be useful for describing this process? The Bernoulli distribution describes the frequency of two outcomes over repeated observations. Each voter is an observation in this example. If the process that determines whether a given voter casts a valid versus a spoilt ballot is random (i.e., is not deterministic), then we can use the Bernoulli distribution as long as we are willing to assume that one voter's decision does not influence another's.

That is because the Bernoulli distribution is built on an assumption that the individual events are *independent* of one another (e.g., the outcome of one flip of a fair coin does not influence the outcome of the subsequent flip of a fair coin). So we need to assume that the probability that a given eligible voter casts a ballot in an election is independent of other eligible voters' decisions to cast a ballot. The assumption of independence often (but not always!) underlies a given distribution, and political scientists have to make judgments about whether they can assume independence. While it is certainly likely that a given

[31] Recall that any number to the zero power is equal to one: $p^0 = 1$.

[32] To be sure, some invalid ballots are due to error rather than intention, but the percentage of invalid ballots in compulsory voting systems considerably exceeds that in voluntary systems, and it is widely understood among the Australian electorate that a spoilt ballot is a protest vote for "none of the above."