2

Probability Theory and Classical Statistics

Statistical inference rests on probability theory, and so an in-depth understanding of the basics of probability theory is necessary for acquiring a conceptual foundation for mathematical statistics. First courses in statistics for social scientists, however, often divorce statistics and probability early with the emphasis placed on basic statistical modeling (e.g., linear regression) in the absence of a grounding of these models in probability theory and probability distributions. Thus, in the first part of this chapter, I review some basic concepts and build statistical modeling from probability theory. In the second part of the chapter, I review the classical approach to statistics as it is commonly applied in social science research.

2.1 Rules of probability

Defining "probability" is a difficult challenge, and there are several approaches for doing so. One approach to defining probability concerns itself with the frequency of events in a long, perhaps infinite, series of trials. From that perspective, the reason that the probability of achieving a heads on a coin flip is 1/2 is that, in an infinite series of trials, we would see heads 50% of the time. This perspective grounds the classical approach to statistical theory and modeling. Another perspective on probability defines probability as a subjective representation of uncertainty about events. When we say that the probability of observing heads on a single coin flip is 1/2, we are really making a series of assumptions, including that the coin is fair (i.e., heads and tails are in fact equally likely), and that in prior experience or learning we recognize that heads occurs 50% of the time. This latter understanding of probability grounds Bayesian statistical thinking. From that view, the language and mathematics of probability is the natural language for representing uncertainty, and there are subjective elements that play a role in shaping probabilistic statements.

Although these two approaches to understanding probability lead to different approaches to statistics, some fundamental axioms of probability are

important and agreed upon. We represent the probability that a particular event, E, will occur as $p(E)$. All possible events that can occur in a single trial or experiment constitute a sample space (S), and the sum of the probabilities of all possible events in the sample space is 1[1]:

$$\sum_{\forall E \in S} p(E) = 1. \tag{2.1}$$

As an example that highlights this terminology, a single coin flip is a trial/experiment with possible events "Heads" and "Tails," and therefore has a sample space of $S = \{\text{Heads, Tails}\}$. Assuming the coin is fair, the probabilities of each event are 1/2, and—as used in social science—the record of the outcome of the coin-flipping process can be considered a "random variable."

We can extend the idea of the probability of observing one event in one trial (e.g., one head in one coin toss) to multiple trials and events (e.g., two heads in two coin tosses). The probability assigned to multiple events, say A and B, is called a "joint" probability, and we denote joint probabilities using the disjunction symbol from set notation ($\cap$) or commas, so that the probability of observing events A and B is simply $p(A, B)$. When we are interested in the occurrence of event A *or* event B, we use the union symbol ($\cup$), or simply the word "or": $p(A \cup B) \equiv p(A \textit{ or } B)$.

The "or" in probability is somewhat different than the "or" in common usage. Typically, in English, when we use the word "or," we are referring to the occurrence of one or another event, but not both. In the language of logic and probability, when we say "or" we are referring to the occurrence of either event or both events. Using a Venn diagram clarifies this concept (see Figure 2.1).

In the diagram, the large rectangle denotes the sample space. Circles A and B denote events A and B, respectively. The overlap region denotes the joint probability $p(A, B)$. $p(A \textit{ or } B)$ is the region that is A only, B only, and the disjunction region. A simple rule follows:

$$p(A \textit{ or } B) = p(A) + p(B) - p(A, B). \tag{2.2}$$

$p(A, B)$ is subtracted, because it is added twice when summing $p(A)$ and $p(B)$.

There are two important rules for joint probabilities. First:

$$p(A, B) = p(A)p(B) \tag{2.3}$$

iff (if and only if) A and B are independent events. In probability theory, independence means that event A has no bearing on the occurrence of event B. For example, two coin flips are independent events, because the outcome of the first flip has no bearing on the outcome of the second flip. Second, if A and B are not independent, then:

[1] If the sample space is continuous, then integration, rather than summation, is used. We will discuss this issue in greater depth shortly.

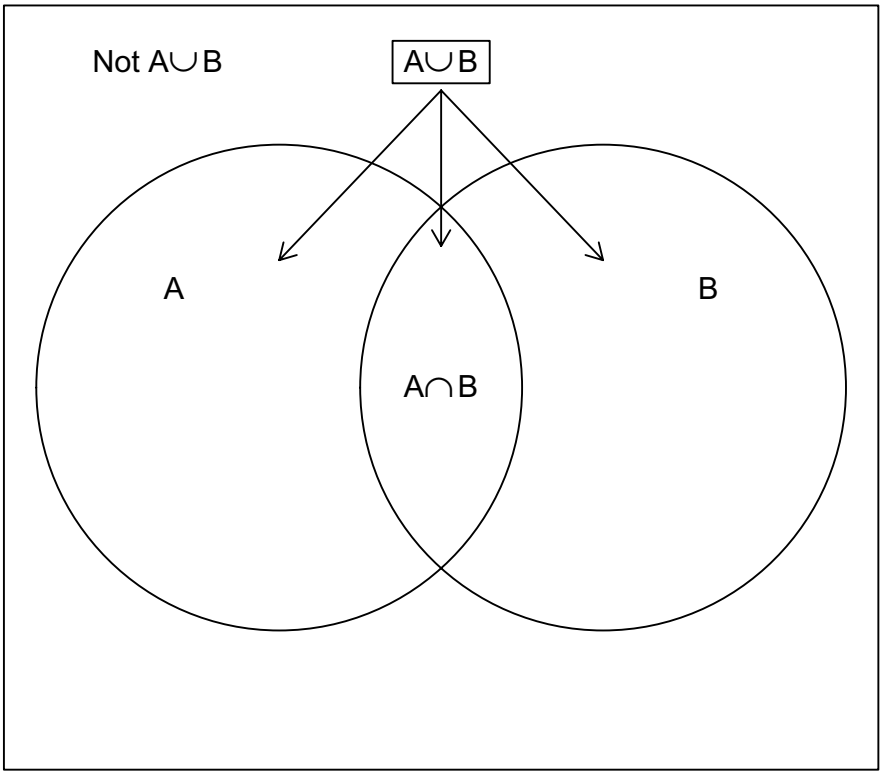

Fig. 2.1. Sample Venn diagram: Outer box is sample space; and circles are events A and B.

$$p(A, B) = p(A|B)p(B). \tag{2.4}$$

Expressed another way:

$$p(A|B) = \frac{p(A, B)}{p(B)}. \tag{2.5}$$

Here, the "|" represents a conditional and is read as "given." This last rule can be seen via Figure 2.1. $p(A|B)$ refers to the region that contains A, given that we know B is already true. Knowing that B is true implies a reduction in the total sample space from the entire rectangle to the circle B only. Thus, $p(A)$ is reduced to the (A, B) region, given the reduced space B, and $p(A|B)$ is the proportion of the new sample space, B, which includes A. Returning to the rule above, which states $p(A, B) = p(A)p(B)$ iff A and B are independent, if A and B are independent, then knowing B is true in that case does not reduce the sample space. In that case, then $p(A|B) = p(A)$, which leaves us with the first rule.

Although we have limited our discussion to two events, these rules generalize to more than two events. For example, the probability of observing three independent events A, B, and C, is $p(A, B, C) = p(A)p(B)p(C)$. More generally, the joint probability of n independent events, $E_1, E_2 \ldots E_n$, is $\prod_{i=1}^{n} p(E_i)$, where the $\prod$ symbol represents repeated multiplication. This result is very useful in statistics in constructing likelihood functions. See DeGroot (1986) for additional generalizations. Surprisingly, with basic generalizations, these basic probability rules are all that are needed to develop the most common probability models that are used in social science statistics.

2.2 Probability distributions in general

The sample space for a single coin flip is easy to represent using set notation as we did above, because the space consists of only two possible events (heads or tails). Larger sample spaces, like the sample space for 100 coin flips, or the sample space for drawing a random integer between 1 and 1,000,000, however, are more cumbersome to represent using set notation. Consequently, we often use functions to assign probabilities or relative frequencies to all events in a sample space, where these functions contain "parameters" that govern the shape and scale of the curve defined by the function, as well as expressions containing the random variable to which the function applies. These functions are called "probability density functions," if the events are continuously distributed, or "probability mass functions," if the events are discretely distributed. By continuous, I mean that all values of a random variable x are possible in some region (like $x = 1.2345$); by discrete, I mean that only some values of x are possible (like all integers between 1 and 10). These functions are called "density" and "mass" functions because they tell us where the most (and least) likely events are concentrated in a sample space. We often abbreviate both types of functions using "pdf," and we denote a random variable x that has a particular distribution $g(.)$ using the generic notation: $x \sim g(.)$, where the "$\sim$" is read "is distributed as," the g denotes a particular distribution, and the "." contains the parameters of the distribution g.

If $x \sim g(.)$, then the pdf itself is expressed as $f(x) = \ldots$, where the "..." is the particular algebraic function that returns the relative frequency/probability associated with each value of x. For example, one of the most common continuous pdfs in statistics is the normal distribution, which has two parameters—a mean (μ) and variance (σ^2). If a variable x has probabilities/relative frequencies that follow a normal distribution, then we say $x \sim N(\mu, \sigma^2)$, and the pdf is:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}.$$

We will discuss this particular distribution in considerable detail throughout the book; the point is that the pdf is simply an algebraic function that, given particular values for the parameters μ and σ^2, assigns relative frequencies for all events x in the sample space.

I use the term "relative frequencies" rather than "probabilities" in discussing continuous distributions, because in continuous distributions, technically, 0 probability is associated with any particular value of x. An infinite number of real numbers exist between any two numbers. Given that we commonly express the probability for an event E as the number of ways E can be realized divided by the number of possible equally likely events that can occur, when the sample space is continuous, the denominator is infinite. The result is that the probability for any particular event is 0. Therefore, instead of discussing the probability of a particular event, we may discuss the probability of observing an event within a specified range. For this reason, we need to define the cumulative distribution function.

Formally, we define a "distribution function" or "cumulative distribution function," often denoted "cdf," as the sum or integral of a mass or density function from the smallest possible value for x in the sample space to some value X, and we represent the cdf using the uppercase letter or symbol that we used to represent the corresponding pdf. For example, for a continuous pdf $f(x)$, in which x can take all real values ($x \in R$),

$$p(x < X) = F(x < X) = \int_{-\infty}^{X} f(x)\, dx. \tag{2.6}$$

For a discrete distribution, integration is replaced with summation and the "$<$" symbol is replaced with "$\leq$," because some probability is associated with every discrete value of x in the sample space.

Virtually *any* function can be considered a probability density function, so long as the function is real-valued and it integrates (or sums) to 1 over the sample space (the region of allowable values). The latter requirement is necessary in order to keep consistent with the rule stated in the previous section that the sum of all possible events in a sample space equals 1. It is often the case, however, that a given function will not integrate to 1, hence requiring the inclusion of a "normalizing constant" to bring the integral to 1. For example, the leading term outside the exponential expression in the normal density function $(1/\sqrt{2\pi\sigma^2})$ is a normalizing constant. A normalized density—one that integrates to 1—or a density that *can* integrate to 1 with an appropriate normalizing constant is called a "proper" density function. In contrast, a density that cannot integrate to 1 (or a finite value), is called "improper." In Bayesian statistics, the propriety of density functions is important, as we will discuss throughout the remainder of the book.

Many of the most useful pdfs in social science statistics appear complicated, but as a simple first example, suppose we have some random variable x that can take any value in the interval (a, b) with equal probability. This

is called a uniform distribution and is commonly denoted as $U(a, b)$, where a and b are the lower and upper bounds of the interval in which x can fall. If $x \sim U(a, b)$, then

$$f(x) = \begin{cases} c & \text{if } a < x < b \\ 0 & \text{otherwise.} \end{cases} \tag{2.7}$$

What is c? c is a constant, which shows that any value in the interval (a, b) is equally likely to occur. In other words, regardless of which value of x one chooses, the height of the curve is the same. The constant must be determined so that the area under the curve/line is 1. A little calculus shows that this constant must be $1/(b - a)$. That is, if:

$$\int_a^b c \, dx = 1,$$

then

$$c \, x|_a^b = 1,$$

and so

$$c = \frac{1}{(b - a)}.$$

Because the uniform density function does not depend on x, it is a rectangle. Figure 2.2 shows two uniform densities: the $U(-1.5, .5)$ and the $U(0, 1)$ densities. Notice that the heights of the two densities differ; they differ because their widths vary, and the total area under the curve must be 1.

The uniform distribution is not explicitly used very often in social science research, largely because very few phenomena in the social sciences follow such a distribution. In order for something to follow this distribution, values at the extreme ends of the distribution must occur as often as values in the center, and such simply is not the case with most social science variables. However, the distribution is important in mathematical statistics generally, and Bayesian statistics more specifically, for a couple of reasons. First, random samples from other distributions are generally simulated from draws from uniform distributions—especially the standard uniform density $[U(0, 1)]$. Second, uniform distributions are commonly used in Bayesian statistics as priors on parameters when little or no information exists to construct a more informative prior (see subsequent chapters).

More often than not, variables of interest in the social sciences follow distributions that are either peaked in the center and taper at the extremes, or they are peaked at one end of the distribution and taper away from that end (i.e., they are skewed). As an example of a simple distribution that exhibits the latter pattern, consider a density in which larger values are linearly more (or less) likely than smaller ones on the interval (r, s):

$$f(x) = \begin{cases} c(mx + b) & \text{if } r < x < s \\ 0 & \text{otherwise.} \end{cases} \tag{2.8}$$

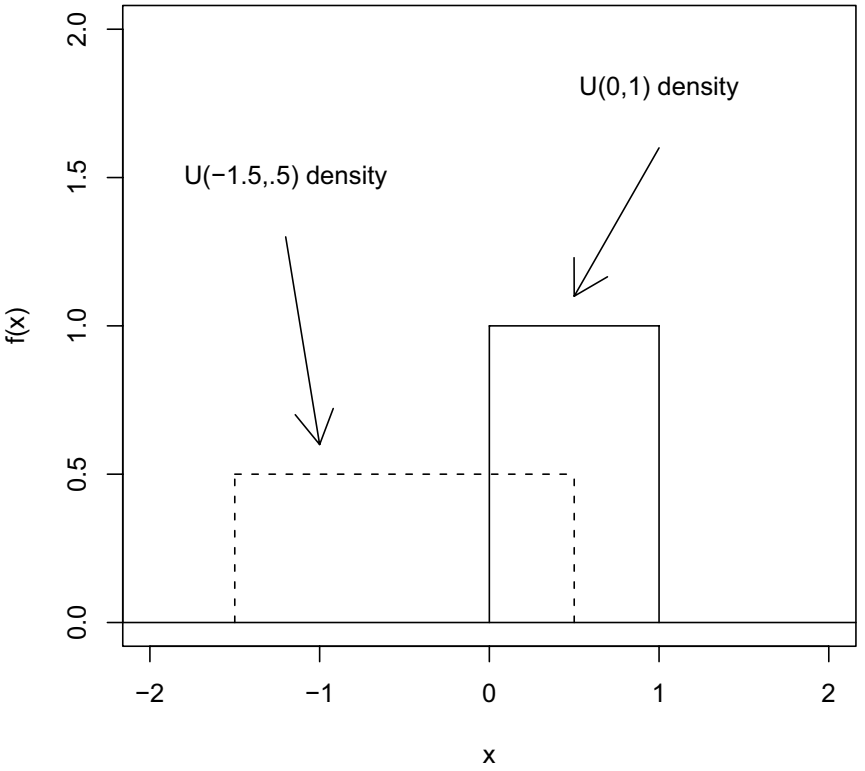

Fig. 2.2. Two uniform distributions.

This density function is a line, with r and s as the left and right boundaries, respectively. As with the uniform density, c is a constant—a normalizing constant—that must be determined in order for the density to integrate to 1. For this generic linear density, the normalizing constant is (see Exercises):

$$c = \frac{2}{(s-r)[m(s+r)+2b]}.$$

In this density, the relative frequency of any particular value of x depends on x, as well as on the parameters m and b. If m is positive, then larger values of x occur more frequently than smaller values. If m is negative, then smaller values of x occur more frequently than larger values.

What type of variable might follow a distribution like this in social science research? I would argue that many attitudinal items follow this sort of distribution, especially those with ceiling or floor effects. For example, in the 2000 General Social Survey (GSS) special topic module on freedom, a question was asked regarding the belief in the importance of being able to express unpopular views in a democracy. Figure 2.3 shows the histogram of responses for this item with a linear density superimposed. A linear density appears to

fit fairly well (of course, the data are discrete, whereas the density function is continuous).

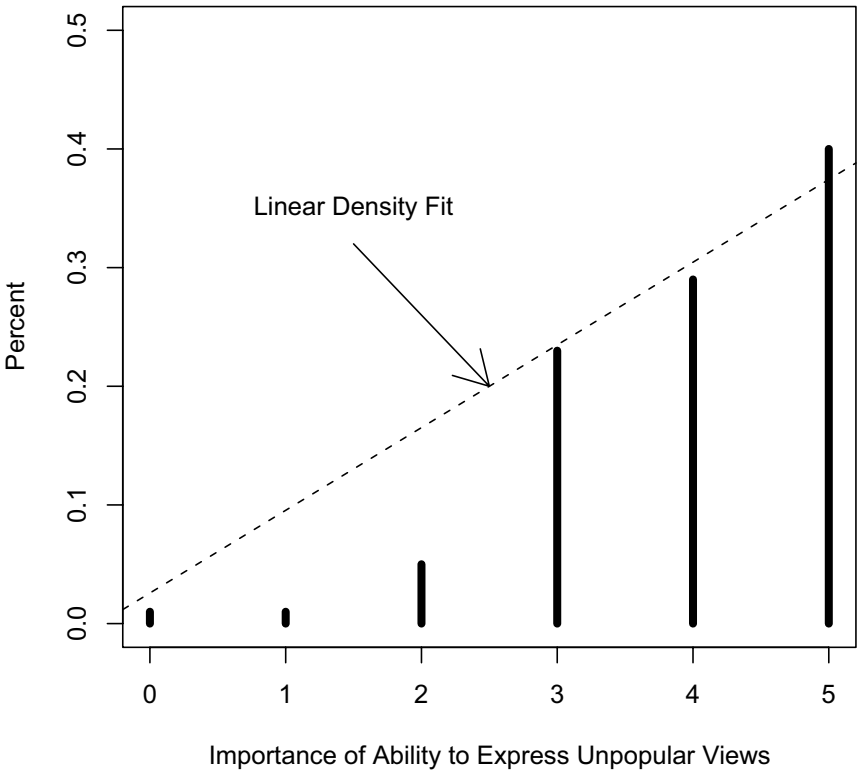

Fig. 2.3. Histogram of the importance of being able to express unpopular views in a free society (1 = Not very important...6 = One of the most important things).

To be sure, we commonly treat such attitudinal items as being normally distributed and model them accordingly, but they may follow a linear distribution as well as, or better than, a normal distribution. Ultimately, this is a question we will address later in the book under model evaluation.

Figure 2.4 shows a particular, arbitrary case of the linear density in which $m = 2$, $b = 3$; the density is bounded on the interval $(0, 5)$; and thus $c = 1/40$. So:

$$f(x) = \begin{cases} (1/40)(2x+3) & 0 < x < 5 \\ 0 & \text{otherwise.} \end{cases} \tag{2.9}$$

Notice that the inclusion of the normalizing constant ultimately alters the slope and intercept if it is distributed through: The slope becomes 1/20 and the intercept becomes 3/40. This change is not a problem, and it highlights

the notion of "relative frequency": The *relative* frequency of values of x are unaffected. For example, the ratio of the height of the original function at $x = 5$ and $x = 0$ is $13/3$, whereas the ratio of the new function at the same values is $\frac{13/40}{3/40} = 13/3$.

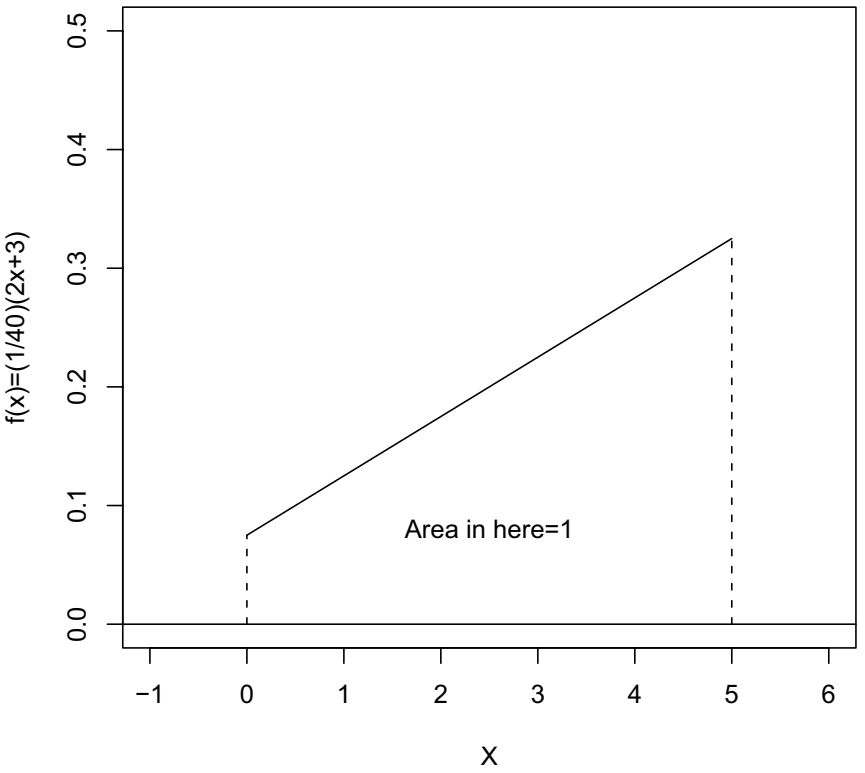

Fig. 2.4. Sample probability density function: A linear density.

2.2.1 Important quantities in distributions

We generally want to summarize information concerning a probability distribution using summary statistics like the mean and variance, and these quantities can be computed from pdfs using integral calculus for continuous distributions and summation for discrete distributions. The mean is defined as:

$$\mu_x = \int_{x \in S} x \times f(x) dx, \tag{2.10}$$

if the distribution is continuous, and:

$$\mu_x = \sum_{x \in S} x \times p(x), \tag{2.11}$$

if the distribution is discrete. The mean is often called the "expectation" or expected value of x and is denoted as $E(x)$. The variance is defined as:

$$\sigma_x^2 = \int_{x \in S} (x - \mu_x)^2 \times f(x) dx, \tag{2.12}$$

if the distribution is continuous, and:

$$\sigma_x^2 = \sum_{x \in S} (x - \mu_x)^2 p(x), \tag{2.13}$$

if the distribution is discrete. Using the expectation notation introduced for the mean, the variance is sometimes referred to as $E((x - \mu_x)^2)$; in other words, the variance is the expected value of the squared deviation from the mean.[2]

Quantiles, including the median, can also be computed using integral calculus. The median of a continuous distribution, for example, is obtained by finding Q that satisfies:

$$.5 = \int_{-\infty}^{Q} f(x) dx. \tag{2.14}$$

Returning to the examples in the previous section, the mean of the $U(a, b)$ distribution is:

$$E(x) = \mu_x = \int_a^b x \times \left(\frac{1}{b-a} \right) \, dx = \frac{b+a}{2},$$

and the variance is:

$$E((x - \mu_x)^2) = \int_a^b \frac{1}{b-a} (x - \mu_x)^2 \, dx = \frac{(b-a)^2}{12}.$$

For the linear density with the arbitrary parameter values introduced in Equation 2.9 ($f(x) = (1/40)(2x + 3)$), the mean is:

$$\mu_x = \int_0^5 x \times (1/40)(2x + 3) dx = (1/240)(4x^3 + 9x^2) dx \Big|_0^5 = 3.02.$$

The variance is:

[2] The sample mean, unlike the population distribution mean shown here, is estimated with $(n - 1)$ in the denominator rather than with n. This is a correction factor for the known bias in estimating the population variance from sample data. It becomes less important asymptotically (as $n \to \infty$.)

$$\text{Var}(x) = \int_0^5 (x - 3.02)^2 \times (1/40)(2x + 3)dx = 1.81.$$

Finally, the median can be found by solving for Q in:

$$.5 = \int_0^Q (1/40)(2x + 3)dx.$$

This yields:

$$20 = Q^2 + 3Q,$$

which can be solved using the quadratic formula from algebra. The quadratic formula yields two real roots—3.22 and −6.22—only one of which is within the "support" of the distribution (3.22); that is, only one has a value that falls in the domain of the distribution.

In addition to finding particular quantiles of the distribution (like quartile cutpoints, deciles, etc.), we may also like to determine the probability associated with a given range of the variable. For example, in the U(0,1) distribution, what is the probability that a random value drawn from this distribution will fall between .2 and .6? Determining this probability also involves calculus[3]:

$$p(.2 < x < .6) = \int_{.2}^{.6} \frac{1}{1-0} dx = x\Big|_{.2}^{.6} = .4.$$

An alternative, but equivalent, way of conceptualizing probabilities for regions of a density is in terms of the cdf. That is, $p(.2 < x < .6) = F(x = .6) - F(x = .2)$, where F is $\int_0^X f(x)dx$ [the cumulative distribution function of $f(x)$].

2.2.2 Multivariate distributions

In social science research, we routinely need distributions that represent more than one variable simultaneously. For example, factor analysis, structural equation modeling with latent variables, simultaneous equation modeling, as well as other methods require the simultaneous analysis of variables that are thought to be related to one another. Densities that involve more than one random variable are called joint densities, or more commonly, multivariate distributions. For the sake of concreteness, a simple, arbitrary example of such a distribution might be:

$$f(x, y) = \begin{cases} c(2x + 3y + 2) \text{ if } & 0 < x < 2\ ,\ 0 < y < 2 \\ 0 & \text{otherwise.} \end{cases} \tag{2.15}$$

Here, the x and y are the two dimensions of the random variable, and $f(x, y)$ is the height of the density, given specific values for the two variables. Thus,

[3] With discrete distributions, calculus is not required, only summation of the relevant discrete probabilities.

$f(x, y)$ gives us the relative frequency/probability of particular values of x and y. Once again, c is the normalizing constant that ensures the function of x and y is a proper density function (that it integrates to 1). In this example, determining c involves solving a double integral:

$$c \int_0^2 \int_0^2 (2x + 3y + 2)\, dx\, dy \; = 1.$$

For this distribution, $c = 1/28$ (find this).

Figure 2.5 shows this density in three dimensions. The height of the density represents the relative frequencies of particular *pairs* of values for x and y. As the figure shows, the density is a partial plane (bounded at 0 and 2 in both x and y dimensions) that is tilted so that larger values of x and y occur more frequently than smaller values. Additionally, the plane inclines more steeply in the y dimension than the x dimension, given the larger slope in the density function.

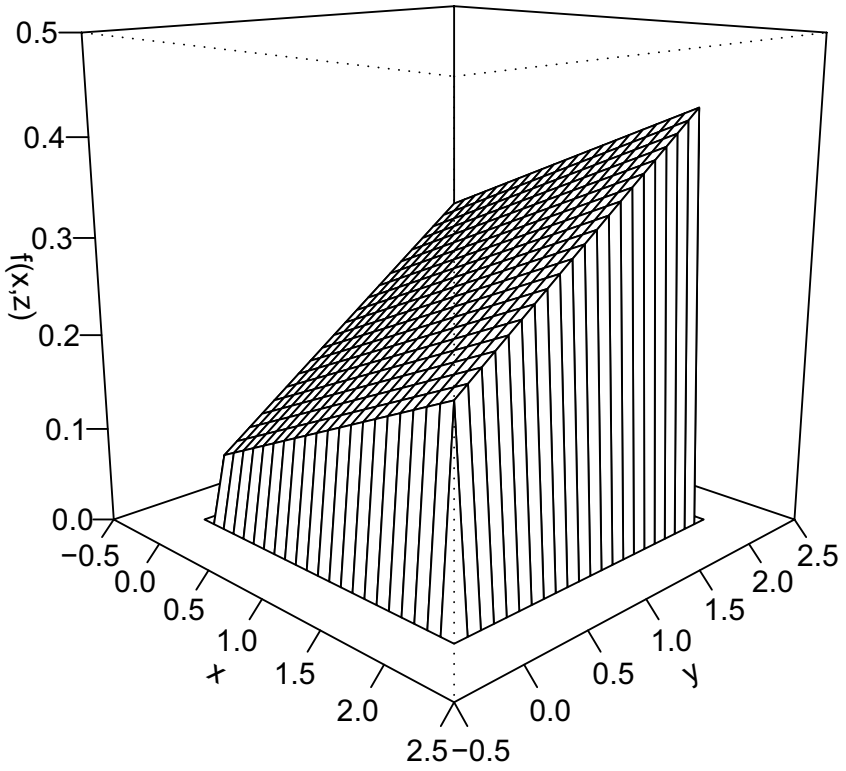

Fig. 2.5. Sample probability density function: A bivariate plane density.

What pair of variables might follow a distribution like this one (albeit with different parameters and domains)? Realistically, we probably would not use this distribution, but some variables might actually follow this sort of pattern. Consider two items from the 2000 GSS topic module on freedom: the one we previously discussed regarding the importance of the ability to express unpopular views in a free society, and another asking respondents to classify the importance of political participation to freedom. Table 2.1 is a cross-tabulation of these two variables. Considered separately, each variable follows a linear density such as discussed earlier. The proportion of individuals in the "Most Important" category for each variable is large, with the proportion diminishing across the remaining categories of the variable. Together, the variables appear to some extent to follow a planar density like the one above. Of course, there are some substantial deviations in places, with two noticeable 'humps' along the diagonal of the table.

Table 2.1. Cross-tabulation of importance of expressing unpopular views with importance of political participation.

Political Participation	Express Unpopular Views						
	1	2	3	4	5	6	
1	361	87	39	8	2	2	36%
2	109	193	51	13	2	3	27%
3	45	91	184	25	4	5	26%
4	15	17	35	17	4	2	7%
5	10	4	9	5	2	0	2%
6	11	9	4	3	1	5	2%
	40%	29%	23%	5%	1%	1%	100%

Note: Data are from the 2000 GSS special topic module on freedom (variables are expunpop and partpol). 1 = One of the most important parts of freedom ... 6 = Not so important to freedom.

Figure 2.6 presents a three-dimensional depiction of these data with an estimated planar density superimposed. The imposed density follows the general pattern of the data but fits poorly in several places. First, in several places the planar density substantially underestimates the true frequencies (three places along the diagonal). Second, the density tends to substantially overestimate frequencies in the middle of the distribution. Based on these problems, finding an alternative density is warranted. For example, a density with exponential or quadratic components may be desirable in order to allow more rapid declines in the expected relative frequencies at higher values of the variables. Furthermore, we may consider using a density that contains a parameter—like a correlation—that captures the relationship between the two variables, given their apparent lack of independence (the "humps" along the diagonal).

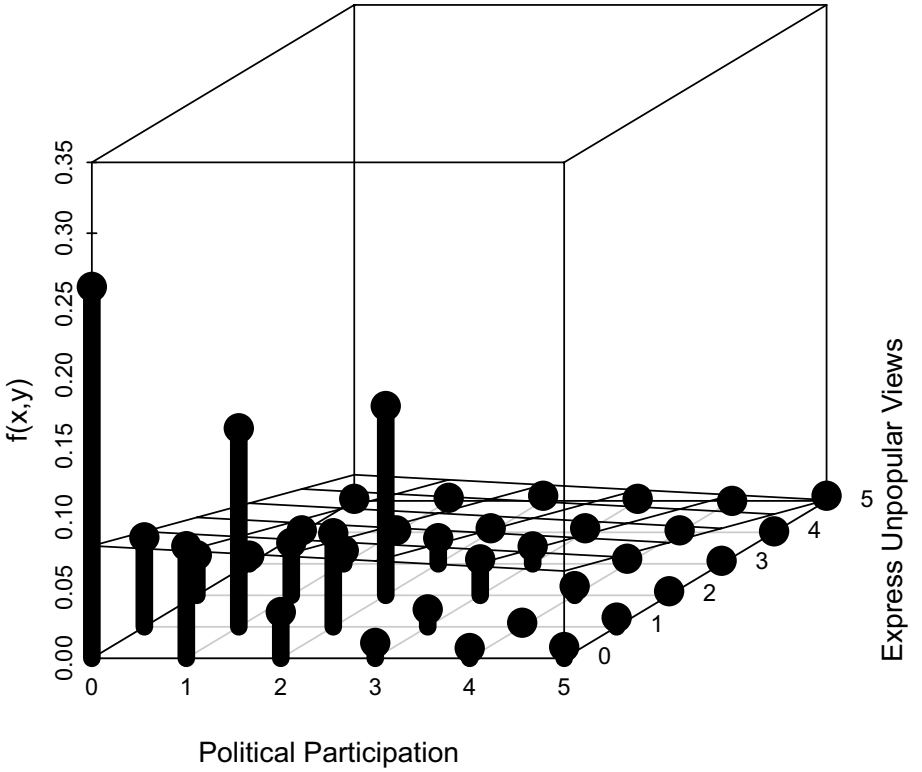

Fig. 2.6. Three-dimensional bar chart for GSS data with "best" planar density superimposed.

In multivariate continuous densities like this planar density, determining the probability that x and y fall in particular regions of the density is determined via integration, just as in univariate densities. That is, the concept of cumulative distribution functions extends to multivariate densities:

$$p(x < X\ ,\ y < Y) = F(x, y) = \int_{-\infty}^{X} \int_{-\infty}^{Y} f(x, y)\ dx\ dy. \tag{2.16}$$

Considering the planar density with parameters arbitrarily fixed at 2 and 3, for example, the probability that $x < 1$ and $y < 1$ is:

$$\int_0^1 \int_0^1 (1/28)(2x + 3y + 2)\ dx\ dy = \frac{9}{56}.$$

This region is presented in Figure 2.7, with the shadow of the omitted portion of the density shown on the $z = 0$ plane.

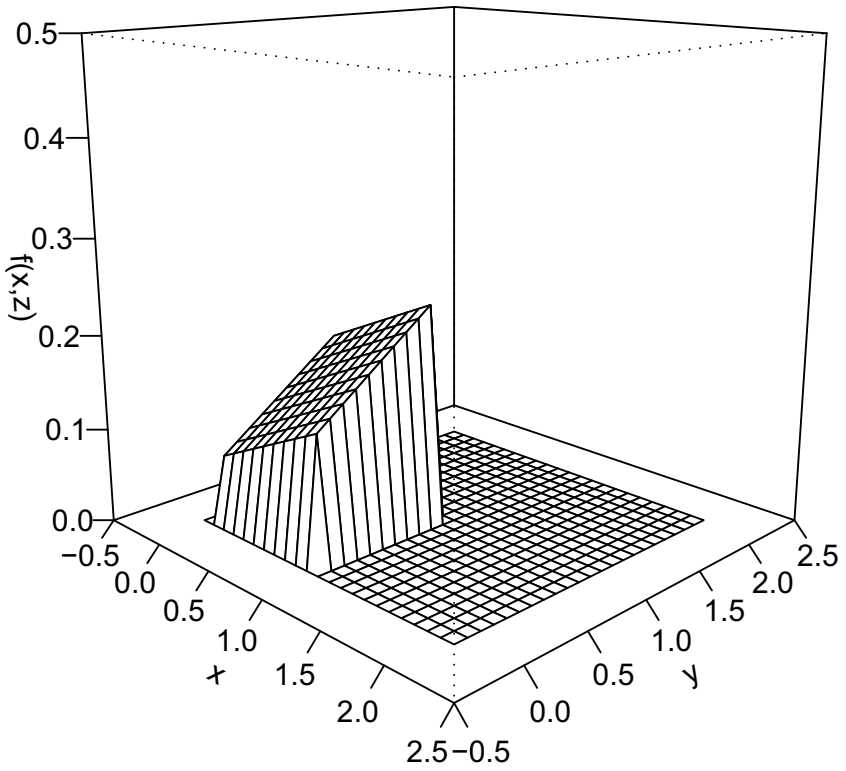

Fig. 2.7. Representation of bivariate cumulative distribution function: Area under bivariate plane density from 0 to 1 in both dimensions.

2.2.3 Marginal and conditional distributions

Although determining the probabilities for particular regions of multivariate densities is important, we may be interested in only a subset of the dimensions of a multivariate density. Two types of "subsets" are frequently needed: marginal distributions and conditional distributions. The data contained in Table 2.1 help differentiate these two types of distributions.

The marginal distribution for the "Express unpopular views" item is the row at the bottom of the table: It is the distribution of this variable summing across the categories of the other variable (or integrating, if the density were continuous). The conditional distribution of this item, on the other hand, is the row of the table corresponding to a particular value for the political participation variable. For example, the conditional distribution for expressing unpopular views, *given* the value of "1" for political participation, consists of the data in the first row of the table (361, 87, 39, 8, 2, and 2, or in renormalized percents: 72%, 17%, 8%, 2%, .4%, and .4%).

Thus, we can think of marginal distributions for a variable as being the original distribution "flattened" in one dimension, whereas the conditional distribution for a variable is a "slice" through one dimension.

Finding marginal and conditional distributions mathematically is conceptually straightforward, although often difficult in practice. Although Equation 2.5 was presented in terms of discrete probabilities, the rule also applies to density functions. From Equation 2.5, a conditional distribution can be computed as:

$$f(x|y) = \frac{f(x,y)}{f(y)} \tag{2.17}$$

This equation says that the conditional distribution for x given y is equal to the joint density of x and y divided by the *marginal* distribution for y, where a marginal distribution is the distribution of one variable, integrating/summing over the other variables in the joint density. Thus:

$$f(y) = \int_{x \in S} f(x,y)dx. \tag{2.18}$$

In terms of our bivariate distribution above ($f(x,y) = (1/28)(2x+3y+2)$), the marginal distributions for x and y can be found as:

$$f(x) = \int_{y=0}^{2} (1/28)(2x+3y+2)dy = (1/28)(4x+10)$$

and

$$f(y) = \int_{x=0}^{2} (1/28)(2x+3y+2)dx = (1/28)(6y+8).$$

The conditional distributions can then be found as:

$$f(x|y) = \frac{(1/28)(2x+3y+2)}{\int_{x=0}^{2}(2x+3y+2)dx} = \frac{(1/28)(2x+3y+2)}{(1/28)(6y+8)}$$

and

$$f(y|x) = \frac{(1/28)(2x+3y+2)}{\int_{y=0}^{2}(2x+3y+2)dy} = \frac{(1/28)(2x+3y+2)}{(1/28)(4x+10)}.$$

Observe how the marginal distributions for each variable exclude the other variable (as they should), whereas the conditional distributions do not. Once a specific value for x or y is chosen in the conditional distribution, however, the remaining function will only depend on the variable of interest. Once again, in other words, the conditional distribution is akin to taking a slice through one dimension of the bivariate distribution.

As a final example, take the conditional distribution $f(x|y)$, where $y = 0$, so that we are looking at the slice of the bivariate distribution that lies on the x axis. The conditional distribution for that slice is:

$$f(x|y=0) = \frac{2x + 3(y=0) + 2}{6(y=0) + 8} = (1/8)(2x+2).$$

With very little effort, it is easy to see that this result gives us the formula for the line that we observe in the x, z plane when we set $y = 0$ in the original unnormalized function and we exclude the constant 1/8. In other words:

$$(1/8)(2x+2) \propto (1/28)(2x+3y+2)$$

when $y = 0$. Thus, an important finding is that the conditional distribution $f(x|y)$ is proportional to the joint distribution for $f(x, y)$ evaluated at a particular value for y [expressed $f(x|y) \propto f(x, y)$], differing only by a normalizing constant. This fact will be useful when we discuss Gibbs sampling in Chapter 4.

2.3 Some important distributions in social science

Unlike the relatively simple distributions we developed in the previous section, the distributions that have been found to be most useful in social science research appear more complicated. However, it should be remembered that, despite their sometimes more complicated appearance, they are simply algebraic functions that describe the relative frequencies of occurence for particular values of a random variable. In this section, I discuss several of the most important distributions used in social science research. I limit the discussion at this point to distributions that are commonly applied to random variables as social scientists view them. In the next chapter, I discuss some additional distributions that are commonly used in Bayesian statistics as "prior distributions" for parameters (which, as we will see, are also treated as random variables by Bayesians). I recommend Evans, Hastings, and Peacock (2000) for learning more about these and other common probability distributions.

2.3.1 The binomial distribution

The binomial distribution is a common discrete distribution used in social science statistics. This distribution represents the probability for x successes in n trials, given a success probability p for each trial. If $x \sim Bin(n, p)$, then:

$$pr(x|n,p) = \binom{n}{x} p^x (1-p)^{n-x}. \tag{2.19}$$

Here, I change the notation on the left side of the mass function to "pr" to avoid confusion with the parameter p in the function. The combinatorial, $\binom{n}{x}$, at the front of the function, compensates for the fact that the x successes can come in any order in the n trials. For example, if we are interested

in the probability of obtaining exactly 10 heads in 50 flips of a fair coin [thus, $pr(x = 10 \mid n = 50,\ p = .5)$], the 10 heads could occur back-to-back, or several may appear in a row, followed by several tails, followed by more heads, etc. This constant is computed as $n!/(x!(n-x)!)$ and acts as a normalizing constant to ensure the mass under the curve sums to 1. The latter two terms in the function multiply the independent success and failure probabilities, based on the observed number of successes and failures. Once the parameters n and p are chosen, the probability of observing any number x of successes can be computed/deduced. For example, if we wanted to know the probability of *exactly* $x = 10$ heads out of $n = 50$ flips, then we would simply substitute those numbers into the right side of the equation, and the result would tell us the probability. If we wanted to determine the probability of obtaining *at least* 10 heads in 50 flips, we would need to sum the probabilities from 10 successes up to 50 successes. Obviously, in this example, the probability of obtaining more heads than 50 or fewer heads than 0 is 0. Hence, this sample space is bounded to counting integers between 0 and 50, and computing the probability of at least 10 heads would require summing 41 applications of the function (for $x = 10$, $x = 11$, ..., $x = 50$).

The mean of the binomial distribution is np, and the variance of the binomial distribution is $np(1-p)$. When $p = .5$, the distribution is symmetric around the mean. When $p > .5$, the distribution is skewed to the left; when $p < .5$, the distribution is skewed to the right. See Figure 2.8 for an example of the effect of p on the shape of the distribution ($n = 10$). Note that, although the figure is presented in a histogram format for the purpose of appearance (the densities are presented as lines), the distribution is discrete, and so 0 probability is associated with non-integer values of x.

A normal approximation to the binomial may be used when p is close to .5 and n is large, by setting $\mu_x = np$ and $\sigma_x = \sqrt{np(1-p)}$. For example, in the case mentioned above in which we were interested in computing the probability of obtaining 10 or more heads in a series of 50 coin flips, computing 41 probabilities with the function would be tedious. Instead, we could set $\mu_x = 25$, and $\sigma_x = \sqrt{50(.5)(1-.5)} = 3.54$, and compute a z-score as $z = (10-25)/(3.54) = -4.24$. Recalling from basic statistics that there is virtually 0 probability in the tail of the z distribution to the left of -4.24, we would conclude that the probability of obtaining at least 10 heads is practically 1, using this approximation. In fact, the actual probability of obtaining at least 10 heads is .999988.

When $n = 1$, the binomial distribution reduces to another important distribution called the Bernoulli distribution. The binomial distribution is often used in social science statistics as a building block for models for dichotomous outcome variables like whether a Republican or Democrat will win an upcoming election, whether an individual will die within a specified period of time, etc.

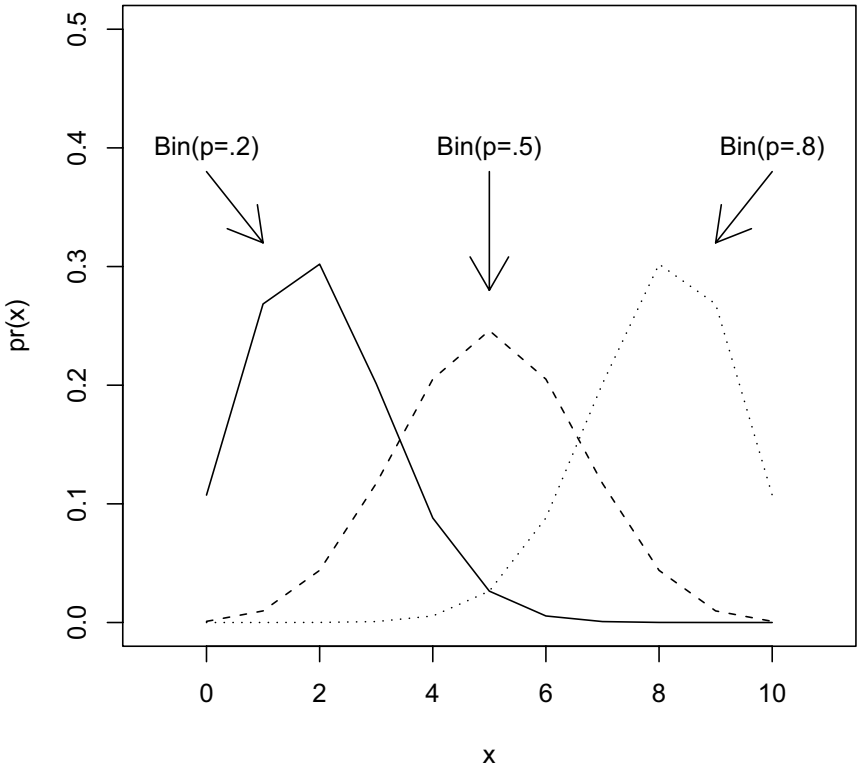

Fig. 2.8. Some binomial distributions (with parameter $n = 10$).

2.3.2 The multinomial distribution

The multinomial distribution is a generalization of the binomial distribution in which there are more than two outcome categories, and thus, there are more than two "success" probabilities (one for each outcome category). If $x \sim Multinomial(n, p_1, p_2, \ldots, p_k)$, then:

$$pr(x_1 \ \ldots \ x_k \mid n, \ p_1 \ \ldots \ p_k) = \frac{n!}{x_1!\, x_2!\, \ldots\, x_k!}\, p_1^{x_1}\, p_2^{x_2}\, \ldots\, p_k^{x_k}, \qquad (2.20)$$

where the leading combinatorial expression is a normalizing constant, $\sum_{i=1}^{k} p_i = 1$, and $\sum_{i=1}^{k} x_i = n$. Whereas the binomial distribution allows us to compute the probability of obtaining a given number of successes (x) out of n trials, given a particular success probability (p), the multinomial distribution allows us to compute the probability of obtaining particular sets of successes, given n trials and given different success probabilities for each member of the set. To make this idea concrete, consider rolling a pair of dice. The sample space

for possible outcomes of a single roll is $S = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$, and we can consider the number of occurrences in multiple rolls of each of these outcomes to be represented by a particular x (so, x_1 represents the number of times a 2 is rolled, x_2 represents the number of times a 3 is rolled, etc.). The success probabilities for these possible outcomes vary, given the fact that there are more ways to obtain some sums than others. The vector of probabilities $p_1 \ldots p_{11}$ is $\{\frac{1}{36}, \frac{2}{36}, \frac{3}{36}, \frac{4}{36}, \frac{5}{36}, \frac{6}{36}, \frac{5}{36}, \frac{4}{36}, \frac{3}{36}, \frac{2}{36}, \frac{1}{36}\}$. Suppose we roll the pair of dice 36 times. Then, if we want to know the probability of obtaining one "2", two "3s", three "4s", etc., we would simply substitute $n = 36$, $p_1 = \frac{1}{36}, p_2 = \frac{2}{36}, \ldots, p_{11} = \frac{1}{36}$, and $x_1 = 1, x_2 = 2, x_3 = 3, \ldots$ into the function and compute the probability.

The multinomial distribution is often used in social science statistics to model variables with qualitatively different outcomes categories, like religious affiliation, political party affiliation, race, etc, and I will discuss this distribution in more depth in later chapters as a building block of some generalized linear models and some multivariate models.

2.3.3 The Poisson distribution

The Poisson distribution is another discrete distribution, like the binomial, but instead of providing the probabilities for a particular number of successes out of a *given* number of trials, it essentially provides the probabilities for a given number of successes in an infinite number of trials. Put another way, the Poisson distribution is a distribution for count variables. If $x \sim Poi(\lambda)$, then:

$$p(x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}. \tag{2.21}$$

Figure 2.9 shows three Poisson distributions, with different values for the λ parameter. When λ is small, the distribution is skewed to the right, with most of the mass concentrated close to 0. As λ increases, the distribution becomes more symmetric and shifts to the right. As with the figure for the binomial distribution above, I have plotted the densities as if they were continuous for the sake of appearance, but because the distribution is discrete, 0 probability is associated with non-integer values of x

The Poisson distribution is often used to model count outcome variables, (e.g., numbers of arrests, number of children, etc.), especially those with low expected counts, because the distributions of such variables are often skewed to the right with most values clustered close to 0. The mean and variance of the Poisson distribution are both λ, which is often found to be unrealistic for many count variables, however. Also problematic with the Poisson distribution is the fact that many count variables, such as the number of times an individual is arrested, have a greater frequency of 0 counts than the Poisson density predicts. In such cases, the negative binomial distribution (not discussed here) and mixture distributions (also not discussed) are often used (see Degroot 1986

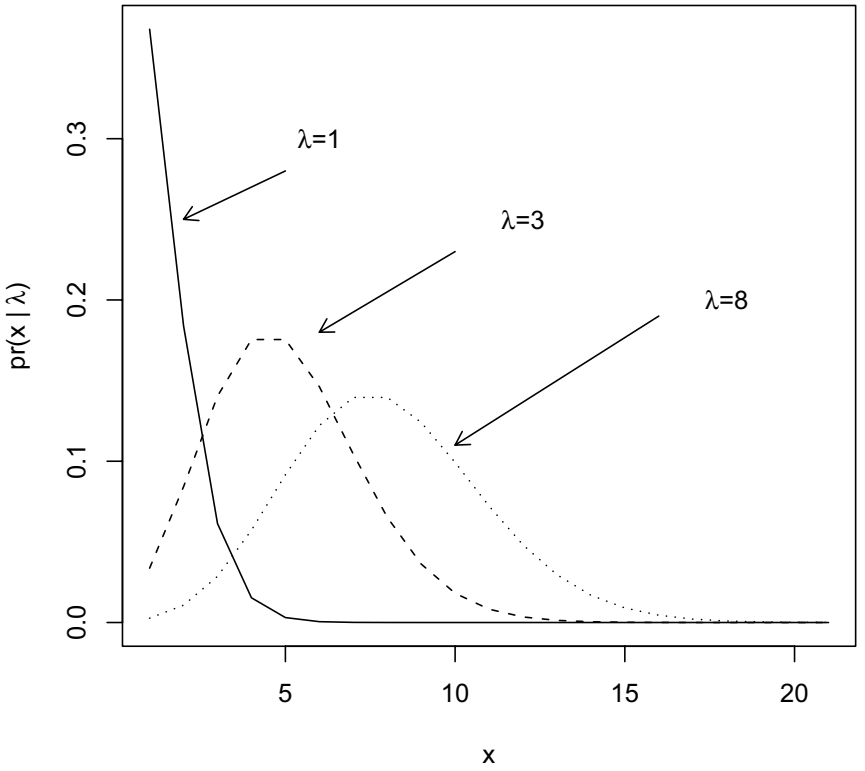

Fig. 2.9. Some Poisson distributions.

for the development of the negative binomial distribution; see Long 1997 for a discussion of negative binomial regression modeling; see Land, McCall, and Nagin 1996 for a discussion of the use of Poisson mixture models).

2.3.4 The normal distribution

The most commonly used distribution in social science statistics and statistics in general is the normal distribution. Many, if not most, variables of interest follow a bell-shaped distribution, and the normal distribution, with both a mean and variance parameter, fits such variables quite well. If $x \sim N(\mu, \sigma^2)$, then:

$$f(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}. \tag{2.22}$$

In this density, the preceding $(\sqrt{2\pi\sigma^2})^{-1}$ is included as a normalizing constant so that the area under the curve from $-\infty$ to $+\infty$ integrates to 1. The latter half of the pdf is the "kernel" of the density and gives the curve its

location and shape. Given a value for the parameters of the distribution, μ and σ^2, the curve shows the relative probabilities for every value of x. In this case, x can range over the entire real line, from $-\infty$ to $+\infty$. Technically, because an infinite number of values exist between any two other values of x (ironically making $p(x = X) = 0, \forall X$), the value returned by the function $f(x)$ does not reveal the probability of x, unlike with the binomial and Poisson distribution above (as well as other discrete distributions). Rather, when using continuous pdfs, one must consider the probability for regions under the curve. Just as above in the discussion of the binomial distribution, where we needed to sum all the probabilities between $x = 10$ and $x = 50$ to obtain the probability that $x \geq 10$, here we would need to integrate the continuous function from $x = a$ to $x = b$ to obtain the probability that $a < x < b$. Note that we did not say $a \leq x \leq b$; we did not for the same reason mentioned just above: The probability that x equals any number q is 0 (the area of a line is 0). Hence $a < x < b$ is equivalent to $a \leq x \leq b$.

The case in which $\mu = 0$ and $\sigma^2 = 1$ is called the "standard normal distribution," and often, the z distribution. In that case, the kernel of the density reduces to $\exp\left\{-x^2/2\right\}$, and the bell shape of the distribution can be easily seen. That is, where $x = 0$, the function value is 1, and as x moves away from 0 in either direction, the function value rapidly declines.

Figure 2.10 depicts three different normal distributions: The first has a mean of 0 and a standard deviation of 1; the second has the same mean but a standard deviation of 2; and the third has a standard deviation of 1 but a mean of 3.

The normal distribution is used as the foundation for ordinary least squares (OLS) regression, for some generalized linear models, and for many other models in social science statistics. Furthermore, it is an important distribution in statistical theory: The Central Limit Theorem used to justify most of classical statistical testing states that sampling distributions for statistics are, in the limit, normal. Thus, the z distribution is commonly used to assess statistical "significance" within a classical statistics framework. For these reasons, we will consider the normal distribution repeatedly throughout the remainder of the book.

2.3.5 The multivariate normal distribution

The normal distribution easily extends to more than one dimension. If $X \sim MVN(\mu, \Sigma)$, then:

$$f(X|\mu, \Sigma) = (2\pi)^{-\frac{k}{2}}|\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)\right\}, \qquad (2.23)$$

where X is a vector of random variables, k is the dimensionality of the vector, μ is the vector of means of X, and Σ is the covariance matrix of X. The multivariate normal distribution is an extension of the univariate normal in

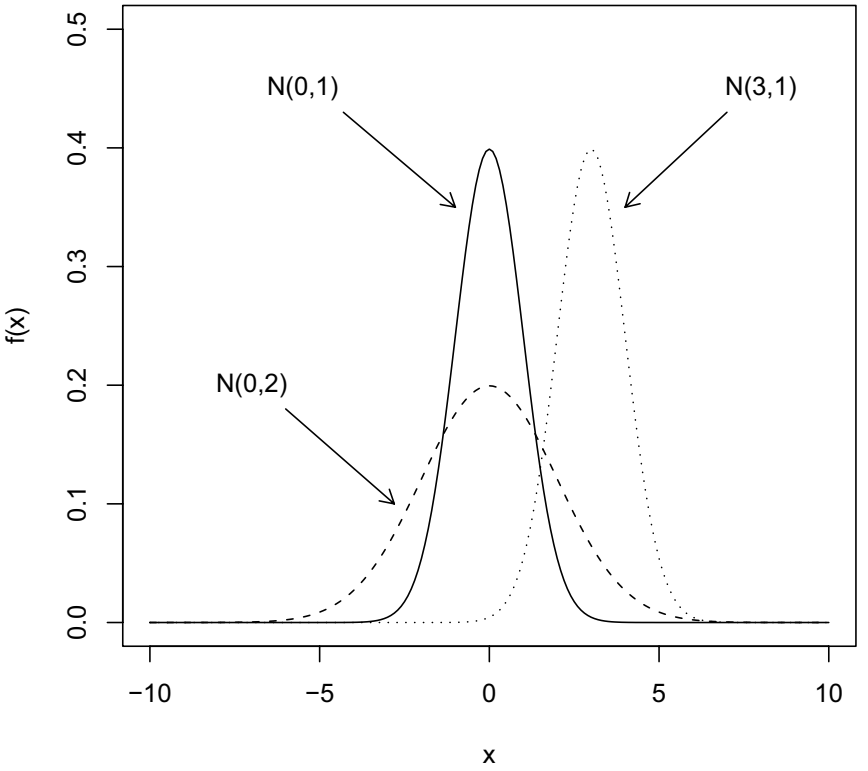

Fig. 2.10. Some normal distributions.

which x is expanded from a scalar to a k-dimensional vector of variables, $x_1, x_2, \ldots, x_k$, that are related to one another via the covariance matrix Σ. If X is multivariate normal, then each variable in the vector X is normal. If Σ is diagonal (all off-diagonal elements are 0), then the multivariate normal distribution is equivalent to k univariate normal densities.

When the dimensionality of the MVN distribution is equal to two, the distribution is called the "bivariate normal distribution." Its density function, although equivalent to the one presented above, is often expressed in scalar form as:

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}(Q - R + S)\right], \tag{2.24}$$

where

$$Q = \frac{(x_1 - \mu_1)^2}{\sigma_1^2}, \tag{2.25}$$

$$R = \frac{2\rho(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_1\sigma_2}, \tag{2.26}$$

and

$$S = \frac{(x_2 - \mu_2)^2}{\sigma_2^2}. \tag{2.27}$$

The bivariate normal distribution, when the correlation parameter ρ is 0, looks like a three-dimensional bell. As ρ becomes larger (in either positive or negative directions), the bell flattens, as shown in Figure 2.11. The upper part of the figure shows a three-dimensional view and a (top-down) contour plot of the bivariate normal density when $\rho = 0$. The lower part of the figure shows the density when $\rho = .8$.

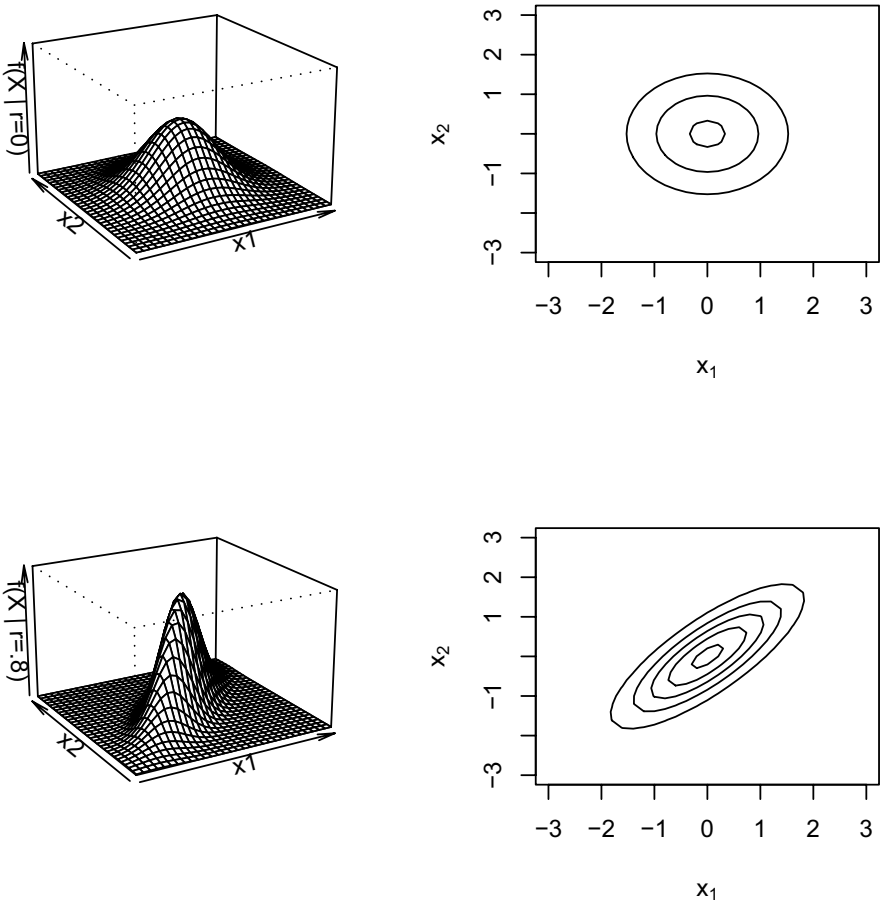

Fig. 2.11. Two bivariate normal distributions.

The multivariate normal distribution is used fairly frequently in social science statistics. Specifically, the bivariate normal distribution is used to model simultaneous equations for two outcome variables that are known to be related, and structural equation models rely on the full multivariate normal distribution. I will discuss this distribution in more depth in later chapters describing multivariate models.

2.3.6 *t* and multivariate *t* distributions

The t (Student's t) and multivariate t distributions are quite commonly used in modern social science statistics. For example, when the variance is unknown in a model that assumes a normal distribution for the data, with the variance following an inverse gamma distribution (see subsequent chapters), the marginal distribution for the mean follows a t distribution (consider tests of coefficients in a regression model). Also, when the sample size is small, the t is used as a robust alternative to the normal distribution in order to compensate for heavier tails in the distribution of the data. As the sample size increases, uncertainty about σ decreases, and the t distribution converges on a normal distribution (see Figure 2.12). The density functions for the t distribution appears much more complicated than the normal. If $x \sim t(\mu, \sigma, v)$, then:

$$f(x) = \frac{\Gamma((v+1)/2)}{\Gamma(v/2)\sigma\sqrt{v\pi}} \left(1 + v^{-1}\left(\frac{x-\mu}{\sigma}\right)^2\right)^{-(v+1)/2}, \tag{2.28}$$

where μ is the mean, σ is the standard deviation, and v is the "degrees of freedom." If X is a k-dimensional vector of variables $(x_1 \ldots x_k)$, and $X \sim mvt(\mu, \Sigma, v)$, then:

$$f(X) = \frac{\Gamma((v+d)/2)}{\Gamma(v/2)v^{k/2}\pi^{k/2}} \mid \Sigma \mid^{-1/2} \left(1 + v^{-1}(X-\mu)^T\Sigma^{-1}(X-\mu)\right)^{-(v+k)/2}, \tag{2.29}$$

where μ is a vector of means, and Σ is the variance-covariance matrix of X.

We will not explicitly use the t and multivariate t distributions in this book, although a number of marginal distributions we will be working with will be implicitly t distributions.

2.4 Classical statistics in social science

Throughout the fall of 2004, CNN/USAToday/Gallup conducted a number of polls attempting to predict whether George W. Bush or John F. Kerry would win the U.S. presidential election. One of the key battleground states was Ohio, which ultimately George Bush won, but all the polls leading up

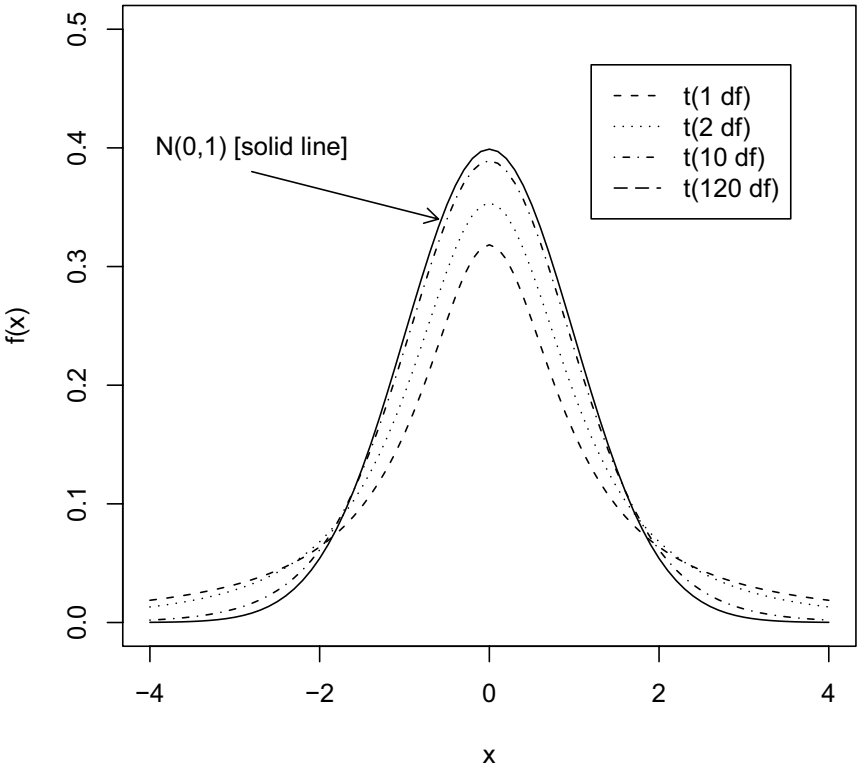

Fig. 2.12. The $t(0, 1, 1)$, $t(0, 1, 10)$, and $t(0, 1, 120)$ distributions (with an $N(0, 1)$ distribution superimposed).

to the election showed the two candidates claiming proportions of the votes that were statistically indistinguishable in the state. The last poll in Ohio consisted of 1,111 likely voters, 46% of whom stated that they would vote for Bush, and 50% of whom stated that they would vote for Kerry, but the poll had a margin of error of $\pm 3\%$.[4]

In the previous sections, we discussed probability theory, and I stated that statistics is essentially the inverse of probability. In probability, once we are given a distribution and its parameters, we can deduce the probabilities for events. In statistics, we have a collection of events and are interested in

[4] see http://www.cnn.com/ELECTION/2004/special/president/showdown/OH/polls.html for the data reported in this and the next chapter. Additional polls are displayed on the website, but I use only the CNN/USAToday/Gallup polls, given that they are most likely similar in sample design. Unfortunately, the proportions are rounded, and so my results from here on are approximate. For example, in the last poll, 50% planned to vote for Kerry, and 50% of 1,111 is 556. However, the actual number could range from 550 to 561 given the rounding.

determining the values of the parameters that produced them. Returning to the polling data, determining who would win the election is tantamount to determining the population parameter (the proportion of actual voters who will vote for a certain candidate) given a collection of events (a sample of potential votes) thought to arise from this parameter and the probability distribution to which it belongs.

Classical statistics provides one recipe for estimating this population parameter; in the remainder of this chapter, I demonstrate how. In the next chapter, I tackle the problem from a Bayesian perspective. Throughout this section, by "classical statistics" I mean the approach that is most commonly used among academic researchers in the social sciences. To be sure, the classical approach to statistics in use is a combination of several approaches, involving the use of theorems and perspectives of a number of statisticians. For example, the most common approach to model estimation is maximum likelihood estimation, which has its roots in the works of Fisher, whereas the common approach to hypothesis testing using p-values has its roots in the works of both Neyman and Pearson and Fisher—each of whom in fact developed somewhat differing views of hypothesis testing using p-values (again, see Hubbard and Bayarri 2003 or see Gill 2002 for an even more detailed history).

2.5 Maximum likelihood estimation

The classical approach to statistics taught in social science statistics courses involves two basic steps: (1) model estimation and (2) inference. The first step involves first determining an appropriate probability distribution/model for the data at hand and then estimating its parameters. Maximum likelihood (ML) is the most commonly used method of estimating parameters and determining the extent of error in the estimation (steps 1 and 2, respectively) in social science statistics (see Edwards 1992 for a detailed, theoretical discussion of likelihood analysis; see Eliason 1993 for a more detailed discussion of the mechanics of ML estimation).

The fundamental idea behind maximum likelihood estimation is that a good choice for the estimate of a parameter of interest is the value of the parameter that makes the observed data most likely to have occurred. To do this, we need to establish some sort of function that gives us the probability for the data, and we need to find the value of the parameter that maximizes this probability. This function is called the "likelihood function" in classical statistics, and it is essentially the product of sampling densities—probability distributions—for each observation in the sample. The process of estimation thus involves the following steps:

1. Construct a likelihood function for the parameter(s) of interest.
2. Simplify the likelihood function and take its logarithm.
3. Take the partial derivative of the log-likelihood function with respect to each parameter, and set the resulting equation(s) equal to 0.

4. Solve the system of equations to find the parameters.

This process seems complicated, and indeed it can be. Step 4 can be quite difficult when there are lots of parameters. Generally, some sort of iterative method is required to find the maximum. Below I detail the process of ML estimation.

2.5.1 Constructing a likelihood function

If $x_1, x_2 \ldots x_n$ are independent observations of a random variable, x, in a data set of size n, then we know from the multiplication rule in probability theory that the joint probability for the vector X is:

$$f(X|\theta) \equiv L(\theta \mid x) = \prod_{i=1}^{n} f(x_i \mid \theta). \tag{2.30}$$

This equation is the likelihood function for the model. Notice how the parameter and the data switch places in the $L(.)$ notation versus the $f(.)$ notation. We denote this as $L(.)$, because from a classical standpoint, the parameter is assumed to be fixed. However, we are interested in estimating the parameter θ, given the data we have observed, so we use this notation. The primary point of constructing a likelihood function is that, given the data at hand, we would like to solve for the value of the parameter that makes the occurence of the data most probable, or most "likely" to have actually occurred.

As the right-hand side of the equation shows, the construction of the likelihood function first relies on determining an appropriate probability distribution $f(.)$ thought to generate the observed data. In our election polling example, the data consist of 1,111 potential votes, the vast majority of which were either for Bush or for Kerry. If we assume that candidates other than these two are unimportant—that is, the election will come down to whom among these two receives more votes—then the data ultimately reduce to 556 potential votes for Kerry and 511 potential votes for Bush. An appropriate distribution for such data is the binomial distribution. If we are interested in whether Kerry will win the election, we can consider a vote for Kerry a "success," and its opposite, a vote for Bush, a "failure," and we can set up our likelihood function with the goal of determining the success probability p. The likelihood function in this case looks like:

$$L(p|X) = \binom{1067}{556} p^{556}(1-p)^{511}.$$

As an alternative view that ultimately produces the same results, we can consider that, at the individual level, each of our votes arises from a Bernoulli distribution, and so our likelihood function is the product of $n = 1,067$ Bernoulli distributions. In that case:

$$L(p|X) = \prod_{i=1}^{n=1067} p^{x_i}(1-p)^{1-x_i}. \tag{2.31}$$

Given that we know nothing about our potential voters beyond for whom they plan to vote, we can consider the individuals "exchangeable," and after carrying out the multiplication across individuals, this version of the likelihood function is proportional to the first one based on the binomial distribution, only differing by a combinatorial expression. This expression simply scales the curve, and so it is ultimately unimportant in affecting our estimate. Figure 2.13 shows this result: The upper figure shows the likelihood function based on the binomial distribution; the lower figure shows the likelihood function based on the Bernoulli distribution. The only difference between the two functions can be found in the scale of the y axis.

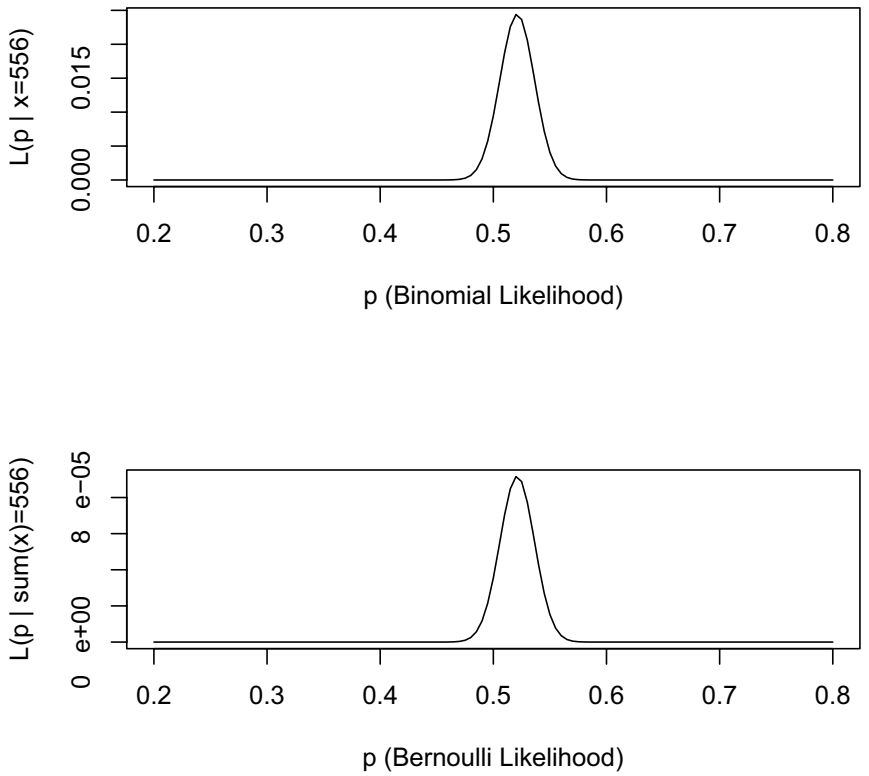

Fig. 2.13. Binomial (top) and Bernoulli (bottom) likelihood functions for the OH presidential poll data.

2.5.2 Maximizing a likelihood function

How do we obtain the estimates for the parameters after we set up the likelihood function? Just as many pdfs are unimodal and slope away from the mode of the distribution, we expect the likelihood function to look about the same. So, what we need to find is the peak of this curve. From calculus we know that the slope of the curve should be 0 at its peak. Thus, we should take the derivative of the likelihood function with respect to the parameter, set it equal to 0, and find the x coordinate (the parameter value) for which the curve reaches a maximum.

We generally take the logarithm of the likelihood function before we differentiate, because the log function converts the repeated multiplication to repeated addition, and repeated addition is much easier to work with. The log-likelihood reaches a maximum at the same point as the original function. Generically:

$$\text{Log-Likelihood} \equiv LL(\theta \mid X) = \sum_{i=1}^{n} \ln\left(f(x_i \mid \theta)\right). \tag{2.32}$$

For our specific problem:

$$LL(p|x) \propto 556 \ln p + 511 \ln(1-p).$$

To find the value of p where this log-likelihood function reaches a maximum, we need to take the derivative of the function with respect to p, set it equal to 0, and solve for p. Generically, the derivative of a binomial log-likelihood function is:

$$\frac{dLL}{dp} = \frac{\sum x_i}{p} - \frac{n - \sum x_i}{1-p}. \tag{2.33}$$

If we set this derivative equal to 0 and solve for p, we obtain:

$$\frac{n - \sum x_i}{1-p} = \frac{\sum x_i}{p}.$$

Simplifying yields:

$$\hat{p} = \frac{\sum x_i}{n}. \tag{2.34}$$

This result shows that the maximum likelihood estimate for p is simply the observed proportion of successes. In our example, this is the proportion of potential votes for Kerry, out of those who opted for either Kerry or Bush (here, $556/1067 = .521$). Given that this value for p is an estimate, we typically denote it $\hat{p}$, rather than p.

Figure 2.14 displays this process of estimation graphically. The figure shows that both the likelihood function and the log-likelihood functions peak at the same point. The horizontal lines are the tangent lines to the curve

where the slopes of these lines are 0; they are at the maximum of the functions. The corresponding x coordinate where the curves reach their maximum is the maximum likelihood estimate (MLE).

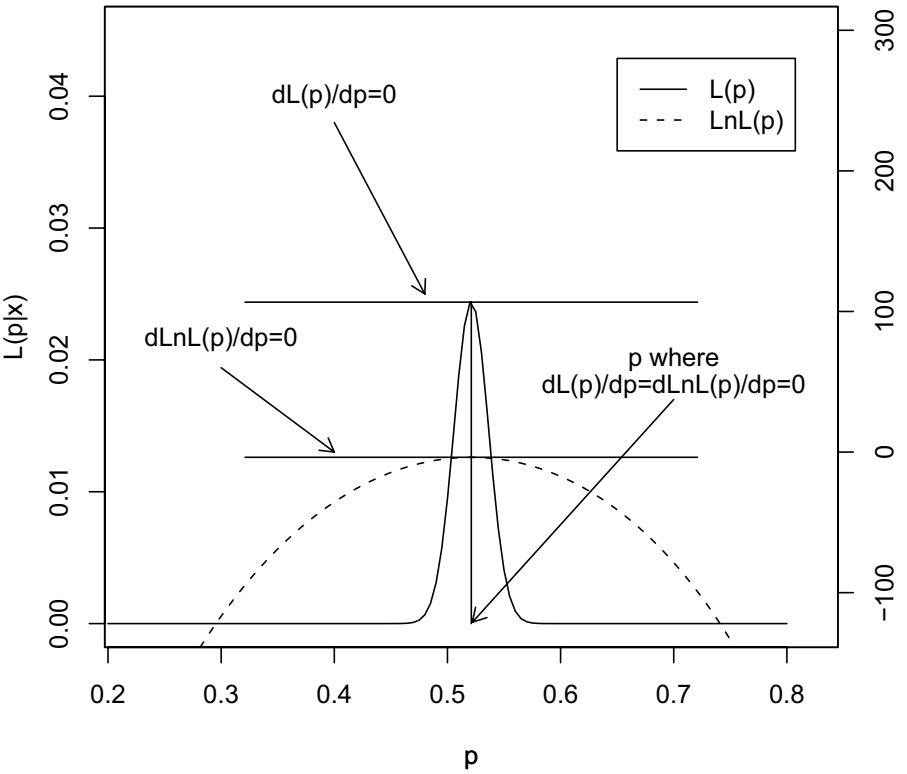

Fig. 2.14. Finding the MLE: Likelihood and log-likelihood functions for the OH presidential poll data.

2.5.3 Obtaining standard errors

$\hat{p}$ is an estimate and is not guaranteed to equal the population parameter p in any particular sample. Thus, we need some way to quantify our uncertainty in estimating p with $\hat{p}$ from our sample. A nice additional feature of the log-likelihood is that a function of the second derivative of the log-likelihood function can be used to estimate the variance of the sampling distribution (the square root of which is called the "standard error").[5] Specifically, we

[5] See Appendix B for a discussion of the Central Limit Theorem and the basis for classical inference.

must take the inverse of the negative expected value of the second derivative of the log-likelihood function. Mathematically:

$$I(\theta)^{-1} = \left(-E\left(\frac{\partial^2 LL}{\partial\theta\partial\theta^T}\right)\right)^{-1}, \tag{2.35}$$

where θ is our parameter or vector of parameters and $I(\theta)$ is called the "information matrix." The square root of the diagonal elements of this matrix are the parameter standard errors. $I(\theta)^{-1}$ can be computed using the following steps:

1. Take the second partial derivatives of the log-likelihood. In multiparameter models, this produces a matrix of partial derivatives (called the Hessian matrix).
2. Take the negative of the expectation of this matrix to obtain the "information matrix" $I(\theta)$.
3. Invert this matrix to obtain estimates of the variances and covariances of parameters (get standard errors by square-rooting the diagonal elements of the matrix).

The fact that $I(\theta)^{-1}$ contains the standard errors is not intuitive. But, if you recall that the first derivative is a measure of the slope of a function at a point (the rate of change in the function at that point), and the second derivative is a measure of the rate of change in the slope, we can think of the second derivative as indicating the rate of curvature of the curve. A very steep curve, then, has a very high rate of curvature, which makes its second derivative large. Thus, when we invert it, it makes the standard deviation small. On the other hand, a very shallow curve has a very low rate of curvature, which makes its second derivative small. When we invert a small number, it makes the standard deviation large. Note that, when we evaluate the second derivative, we substitute the MLE estimate for the parameter into the result to obtain the standard error at the estimate.

Returning to our data at hand, the second partial derivative of the generic binomial log-likelihood function with respect to p is:

$$\frac{\partial^2 LL}{\partial p^2} = \frac{\sum x}{p^2} - \frac{n - \sum x}{(1-p)^2}. \tag{2.36}$$

Taking expectations yields:

$$E\left(\frac{\partial^2 LL}{\partial p^2}\right) = E\left[-\frac{\sum x}{p^2} - \frac{n - \sum x}{(1-p)^2}\right].$$

The expectation of these expressions can be computed by realizing that the expectation of $\sum x/n$ is p (put another way: $E(\hat{p}) = p$). Thus:

$$E\left(\frac{\partial^2 LL}{\partial p^2}\right) = -\frac{np}{p^2} - \frac{n - np}{(1-p)^2}.$$

Some simplification yields:

$$E\left(\frac{\partial^2 LL}{\partial p^2}\right) = -\frac{n}{p(1-p)}.$$

At this point, we can negate the expectation, invert it, and evaluate it at the MLE ($\hat{p}$) to obtain:

$$I(p)^{-1} = \frac{\hat{p}(1-\hat{p})}{n}. \tag{2.37}$$

Taking the square root of this yields the estimated standard error. In our polling data case, the standard error is $\sqrt{(.521 \times .479)/1067} = .015$.

Recall that our question is whether Kerry would win the vote in Ohio. Our estimate for the Ohio population proportion to vote for Kerry (versus Bush) was .521, which suggests he would win the popular vote in Ohio (discounting third party candidates). However, the standard error of this estimate was .015. We can construct our usual confidence interval around the maximum likelihood estimate to obtain a 95% interval for the MLE. If we do this, we obtain an interval of $[.492, .550]$ ($CI = \hat{p} \pm 1.96 \times s.e.(\hat{p})$). Given that the lower bound on this interval is below .5, we can conclude that we cannot rule out the possibility that Kerry would not win the popular vote in Ohio.

An alternative to the confidence interval approach to answering this question is to construct a t test, with a null hypothesis $H_0 : p < .5$. Following that approach:

$$t = \frac{(.521 - .5)}{.015} = 1.4.$$

This t value is not large enough to reject the null hypothesis (that Kerry's proportion of the vote is less than .5), and thus, the conclusion we would reach is the same: We do not have enough evidence to conclude that Kerry will win (see Appendix B for more discussion of null hypotheses, confidence intervals, and t tests).

Note that this result is consistent with the result I stated at the beginning of this section: The results of the original poll suggested that the vote was too close to call, given a ±3% margin of error. Here, I have shown essentially from where that margin of error arose. We ended up with a margin of error of .0294, which is approximately equal to the margin of error in the original poll.

2.5.4 A normal likelihood example

Because the normal distribution is used repeatedly throughout this book and throughout the social sciences, I conclude this chapter by deriving parameter estimates and standard errors for a normal distribution problem. I keep this example at a general level; in subsequent chapters, we will return to this likelihood function with specific problems and data.

Suppose you have n observations $x_1, x_2, \ldots, x_n$ that you assume are normally distributed. Once again, if the observations are assumed to be independent, a likelihood function can be constructed as the multiple of independent normal density functions:

$$L(\mu, \sigma \mid X) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x_i - \mu)^2}{2\sigma^2}\right\}. \tag{2.38}$$

We can simplify the likelihood as:

$$L(\mu, \sigma \mid X) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right\}.$$

The log of the likelihood is:

$$LL(\mu, \sigma \mid X) \propto -n\ln(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2. \tag{2.39}$$

In the above equation, I have eliminated the $-\frac{n}{2}\ln(2\pi)$, an irrelevant constant. It is irrelevant, because it does not depend on either parameter and will therefore drop once the partial derivatives are taken. In this example, we have two parameters, μ and σ, and hence the first partial derivative must be taken with respect to each parameter. This will leave us with two equations (one for each parameter). After taking the partial derivatives with respect to each parameter, we obtain the following:

$$\frac{\partial LL}{\partial \mu} = \frac{n(\bar{x} - \mu)}{\sigma^2}$$

and

$$\frac{\partial LL}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Setting these partial derivatives each equal to 0 and doing a little algebra yields:

$$\hat{\mu} = \bar{x} \tag{2.40}$$

and

$$\hat{\sigma^2} = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}. \tag{2.41}$$

These estimators should look familiar: The MLE for the population mean is the sample mean; the MLE for the population variance is the sample variance.[6]

Estimates of the variability in the estimates for the mean and standard deviation can be obtained as we did in the binomial example. However, as

[6] The MLE is known to be biased, and hence, a correction is added, so that the denominator is $n - 1$ rather than n.

noted above, given that we have two parameters, our second derivate matrix will, in fact, be a matrix. For the purposes of avoiding taking square roots until the end, let $\tau = \sigma^2$, and we'll construct the Hessian matrix in terms of τ. Also, let θ be a vector containing both μ and τ. Thus, we must compute:

$$\frac{\partial^2 LL}{\partial\theta\partial\theta^T} = \begin{bmatrix} \frac{\partial^2 LL}{\partial\mu^2} & \frac{\partial^2 LL}{\partial\mu\partial\tau} \\ \frac{\partial^2 LL}{\partial\tau\partial\mu} & \frac{\partial^2 LL}{\partial\tau^2} \end{bmatrix}. \tag{2.42}$$

Without showing all the derivatives (see Exercises), the elements of the Hessian matrix are then:

$$\frac{\partial^2 LL}{\partial\theta\partial\theta^T} = \begin{bmatrix} \frac{-n}{\tau} & -\frac{n(\bar{x}-\mu)}{\tau^2} \\ -\frac{n(\bar{x}-\mu)}{\tau^2} & \frac{n}{2\tau^2} - \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\tau^3} \end{bmatrix}.$$

In order to obtain the information matrix, which can be used to compute the standard errors, we must take the expectation of this Hessian matrix and take its negative. Let's take the expectation of the off-diagonal elements first. The expectation of $\bar{x} - \mu$ is 0 (given that the MLE is unbiased), which makes the off-diagonal elements of the information matrix equal to 0. This result should be somewhat intuitive: There need be no relationship between the mean and variance in a normal distribution.

The first element, $(-n/\tau)$, is unchanged under expectation. Thus, after substituting σ^2 back in for τ and negating the result, we obtain n/σ^2 for this element of the information matrix.

The last element, $(n/2\tau^2) - (\sum_{i=1}^{n}(x_i - \mu)^2)/\tau^3$, requires a little consideration. The only part of this expression that changes under expectation is $\sum_{i=1}^{n}(x_i - \mu)^2$. The expectation of this expression is $n\tau$. That is, $E(x_i - \mu)^2 = \tau$, and we are taking this value n times (notice the summation). Thus, this element, after a little algebraic manipulation, negation, and substitution of σ^2 for τ, becomes: $n/2\sigma^4$. So, our information matrix appears as:

$$I(\theta) = \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}. \tag{2.43}$$

To obtain standard errors, we need to (1) invert this matrix, and (2) take the square root of the diagonal elements (variances) to obtain the standard errors. Matrix inversion in this case is quite simple, given that the off-diagonal elements are equal to 0. In this case, the inverse of the matrix is simply the inverse of the diagonal elements.

Once we invert and square root the elements of the information matrix, we find that the estimate for the standard error for our estimate $\hat{\mu}$ is $\hat{\sigma}/\sqrt{n}$, and our estimate for the standard error for $\hat{\sigma}^2$ is $\hat{\sigma}^2\sqrt{2/n}$. The estimate for the standard error for $\hat{\mu}$ should look familiar: It is the standard deviation of

the sampling distribution for a mean based on the Central Limit Theorem (see Appendix B).

2.6 Conclusions

In this chapter, we have reviewed the basics of probability theory. Importantly, we have developed the concept of probability distributions in general, and we have discussed a number of actual probability distributions. In addition, we have discussed how important quantities like the mean and variance can be derived analytically from probability distributions. Finally, we reviewed the most common approach to estimating such quantities in a classical setting—maximum likelihood estimation—given a collection of data thought to arise from a particular distribution. As stated earlier, I recommend reading DeGroot (1986) for a more thorough introduction to probability theory, and I recommend Billingsley (1995) and Chung and AitSahlia (2003) for more advanced and detailed expositions. For a condensed exposition, I suggest Rudas 2004. Finally, I recommend Edwards (1992) and Gill (2002) for detailed discussions of the history and practice of maximum likelihood (ML) estimation, and I suggest Eliason (1993) for a highly applied perspective on ML estimation. In the next chapter, we will discuss the Bayesian approach to statistics as an alternative to this classical approach to model building and estimation.

2.7 Exercises

2.7.1 Probability exercises

1. Find the normalizing constant for the linear density in Equation 2.8.
2. Using the binomial mass function, find the probability of obtaining 3 heads in a row with a fair coin.
3. Find the probability of obtaining 3 heads in a row with a coin weighted so that the probability of obtaining a head is .7.
4. What is the probability of obtaining 3 heads OR 3 tails in a row with a fair coin?
5. What is the probability of obtaining 3 heads and 1 tail (order irrelevant) on four flips of a fair coin?
6. Using a normal approximation to the binomial distribution, find the probability of obtaining 130 or more heads in 200 flips of a fair coin.
7. Plot a normal distribution with parameters $\mu = 5$ and $\sigma = 2$.
8. Plot a normal distribution with parameters $\mu = 2$ and $\sigma = 5$.
9. Plot the $t(0, 1, df = 1)$ and $t(0, 1, df = 10)$ distributions. Note: Γ is a function. The function is: $\Gamma(n) = \int_0^\infty e^{-u} u^{n-1} du$. For integers, $\Gamma(n) = (n-1)!$ Thus, $\Gamma(4) = (4-1)! = 6$. However, when the argument to the function is not an integer, this formula will not work. Instead, it is easier to use a software package to compute the function value for you.

10. Show that the multivariate normal density function reduces to the univariate normal density function when the dimensionality of the distribution is 1.

2.7.2 Classical inference exercises

1. Find the MLE for p in a binomial distribution representing a sample in which 20 successes were obtained out of 30 trials.
2. Based on the binomial sample in the previous question, if the trials involved coin flips, would having 20 heads be sufficient to question the fairness of the coin? Why or why not?
3. Suppose a sample of students at a major university were given an IQ test, which resulted in a mean of 120 and a standard deviation of 10. If we know that IQs are normally distributed in the population with a mean of 100 and a standard deviation of 16, is there sufficient evidence to suggest that the college students are more intelligent than average?
4. Suppose a single college student were given an IQ test and scored 120. Is there sufficient evidence to indicate that college students are more intelligent than average based on this sample?
5. What is the difference (if any) between the responses to the previous two questions?
6. Derive the Hessian matrix for the normal distribution example at the end of the chapter.
7. Derive the MLE for λ from a sample of n observations from a Poisson distribution.
8. Derive the standard error estimate for λ from the previous question.