Chapter 3

Some basic methods in mathematics and probability

Simple methods from introductory mathematics and statistics have three important roles in regression modeling. First, linear algebra and simple probability distributions are the building blocks for elaborate models. Second, it is useful to understand the basic ideas of inference separately from the details of particular classes of model. Third, it is often useful in practice to construct quick estimates and comparisons for small parts of a problem—before fitting an elaborate model, or in understanding the output from such a model. This chapter provides a quick review of some of these basic ideas.

3.1 Weighted averages

In statistics it is common to reweight data or inferences so as to adapt to a target population.

Here is a simple example. In 2010 there were 456 million people living in North America: 310 million residents of the United States, 112 million Mexicans, and 34 million Canadians. The average age of people in each country in that year is displayed in Figure 3.1. The average age of all North Americans is a *weighted average*:

$$\begin{aligned}\text{average age} &= \frac{310\,000\,000 * 36.8 + 112\,000\,000 * 26.7 + 34\,000\,000 * 40.7}{310\,000\,000 + 112\,000\,000 + 34\,000\,000} \\ &= 34.6.\end{aligned}$$

This is a weighted average rather than a simple average because the numbers 36.8, 26.7, 40.7 are multiplied by "weights" proportional to the population of each country. The total population of North America was 310 + 112 + 34 = 456 million, and we can rewrite the above expression as

$$\begin{aligned}\text{average age} &= \frac{310\,000\,000}{456\,000\,000} * 36.8 + \frac{112\,000\,000}{456\,000\,000} * 26.7 + \frac{34\,000\,000}{456\,000\,000} * 40.7 \\ &= 0.6798 * 36.8 + 0.2456 * 26.7 + 0.0746 * 40.7 \\ &= 34.6.\end{aligned}$$

The above proportions 0.6798, 0.2456, and 0.0746 (which by necessity sum to 1) are the *weights* of the countries in this weighted average.

We can equivalently write a weighted average in summation notation:

$$\text{weighted average} = \frac{\sum_j N_j \bar{y}_j}{\sum_j N_j},$$

where j indexes countries and the sum adds over all the *strata* (in this case, the three countries).

The choice of weights depends on context. For example, 51% of Americans are women and 49% are men. The average age of American women and men is 38.1 and 35.5, respectively. The average age of all Americans is thus $0.51 * 38.1 + 0.49 * 35.5 = 36.8$ (which agrees with the U.S. average

Stratum, j	Label	Population, N_j	Average age, $\bar{y}_j$
1	United States	310 million	36.8
2	Mexico	112 million	26.7
3	Canada	34 million	40.7

Figure 3.1 *Populations and average ages of countries in North America. (Data from CIA World Factbook 2010.) The average age of all North Americans is a* weighted average *of the average ages within each country.*

in Figure 3.1). But now consider a slightly different problem: estimating the average salary of all teachers in the country. According to the Census in 2010, there were 5 700 000 female teachers and 1 500 000 male teachers (that is, the population of teachers was 79% female and 21% male) in the United States, with average incomes $45 865 and $49 207, respectively. The average income of all teachers was $0.79 * \$45\,865 + 0.21 * \$49\,207 = \$46\,567$, *not* $0.51 * \$45\,865 + 0.49 * \$49\,207 = \$47\,503$.

3.2 Vectors and matrices

A list of numbers is called a *vector*. A rectangular array of numbers is called a *matrix*. Vectors and matrices are useful in regression to represent predictions for many cases using a single model.

Example: Elections and the economy

In Section 1.2 we introduced a model for predicting the incumbent party's vote percentage in U.S. presidential elections from economic conditions in the years preceding the election:

$$\text{Predicted vote percentage} = 46.3 + 3.0 * (\text{growth rate of average personal income}),$$

which we shall write as

$$\hat{y} = 46.3 + 3.0x,$$

or, even more abstractly, as

$$\hat{y} = \hat{a} + \hat{b}x.$$

The expressions $\hat{a}$ and $\hat{b}$ denote estimates—the coefficients 46.3 and 3.0 were obtained by fitting a line to past data—and $\hat{y}$ denotes a predicted value. In this case, we would use y to represent an actual election result, and $\hat{y}$ is the prediction from the model. Here we are focusing on the linear prediction, and so we work with $\hat{y}$.

Let's apply this model to a few special cases:

1. $x = -1$. A rate of growth of -1% (that is, a 1% decline in the economy) translates into an incumbent party vote share of $46.3 + 3.0 * (-1) = 43.3$%.
2. $x = 0$. If there is zero economic growth in the year preceding the presidential election, the model predicts that the incumbent party's candidate will receive $46.3 + 3.0 * 0 = 46.3$% of the two-party vote; that is, he or she is predicted to lose the election.
3. $x = 3$. A 3% rate of economic growth translates to the incumbent party's candidate winning $46.3 + 3.0 * 3 = 55.3$% of the vote.

We can define x as the vector that comprises these three cases, that is $x = (-1, 0, 3)$.

We can put these three predictions together:

$$\begin{aligned}\hat{y}_1 &= 43.3 = 46.3 + 3.0 * (-1),\\ \hat{y}_2 &= 46.3 = 46.3 + 3.0 * 0,\\ \hat{y}_3 &= 55.3 = 46.3 + 3.0 * 3,\end{aligned}$$

which can be written as vectors:

$$\hat{y} = \begin{pmatrix} 43.3 \\ 46.3 \\ 55.3 \end{pmatrix} = \begin{pmatrix} 46.3 + 3.0 * (-1) \\ 46.3 + 3.0 * 0 \\ 46.3 + 3.0 * 3 \end{pmatrix},$$

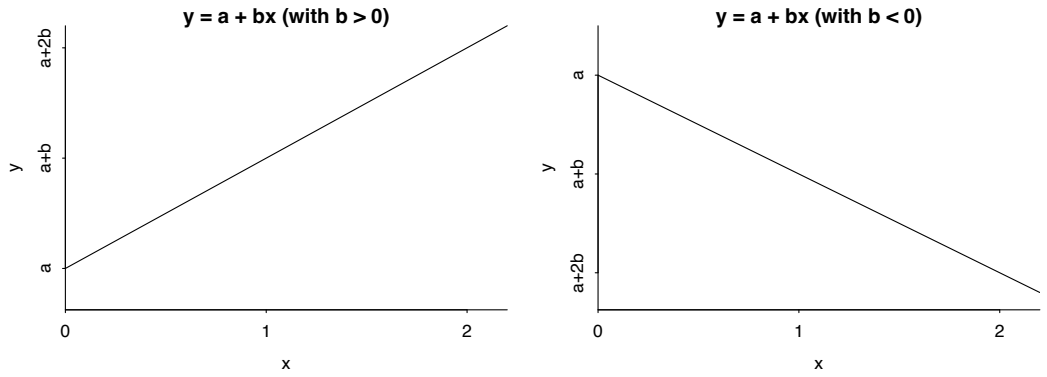

Figure 3.2: *Lines $y = a + bx$ with positive and negative slopes.*

or, in matrix form,

$$\hat{y} = \begin{pmatrix} 43.3 \\ 46.3 \\ 55.3 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} 46.3 \\ 3.0 \end{pmatrix},$$

or, more abstractly,

$$\hat{y} = X\hat{\beta}.$$

Here y and x are vectors of length 3, X is a 3×2 matrix with a column of ones and a column equal to the vector x, and $\hat{\beta} = (46.3, 3.0)$ is the vector of estimated coefficients.

3.3 Graphing a line

To use linear regression effectively, you need to understand the algebra and geometry of straight lines, which we review briefly here.

Figure 3.2 shows the line $y = a + bx$. The *intercept*, a, is the value of y when $x = 0$; the coefficient b is the *slope* of the line. The line slopes upward if $b > 0$ (as in Figure 3.2a), slopes downward if $b < 0$ (as in Figure 3.2b), and is horizontal if $b = 0$. The larger b is, in absolute value, the steeper the line will be.

Example: Mile run

Figure 3.3a shows a numerical example: $y = 1007 - 0.39x$. Thus, $y = 1007$ when $x = 0$, and y decreases by 0.39 for each unit increase in x. This line approximates the trajectory of the world record time (in seconds) for the mile run from 1900 to 2000 (see Figure A.1). Figure 3.3b shows the graph of the line alone on this scale. In R it is easy to draw this line:[1]

```
curve(1007 - 0.393*x, from=1900, to=2000, xlab="Year", ylab="Time (seconds)",
  main="Approximate trend of world record times\nfor the mile run")
```

How would we draw it by hand? We cannot simply start with the intercept at $x = 0$ and go from there, as then the entire range of interest, from 1900 to 2000, would be crammed into a small corner of the plot. The value $x = 0$ is well outside the range of the data. Instead, we use the equation to calculate the value of y at the two extremes of the plot:

$$\text{At } x = 1900,\ y = 1007 - 0.393 * 1900 = 260.$$
$$\text{At } x = 2000,\ y = 1007 - 0.393 * 2000 = 221.$$

These are the two endpoints in Figure 3.3, between which we can draw the straight line.

This example demonstrates the importance of location and scale. The lines in Figures 3.3a and

[1] Data and code for this example are in the folder `Mile`.

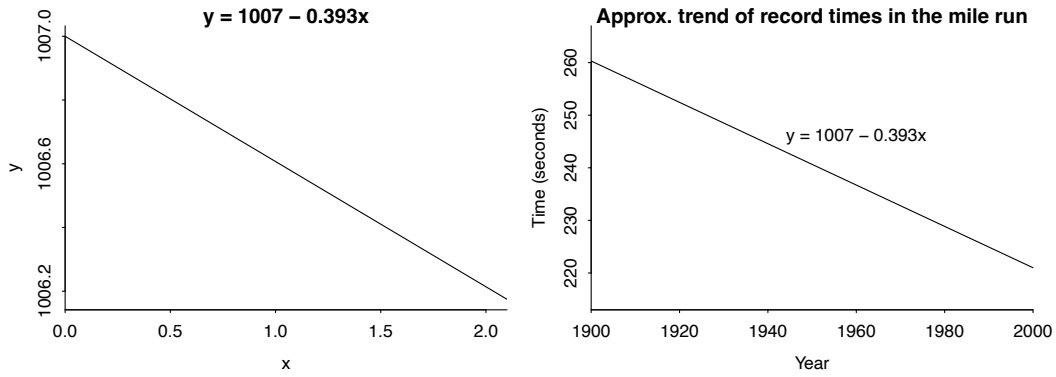

Figure 3.3 *(a) The line $y = 1007 - 0.393x$. (b) For x between 1900 and 2000, the line $y = 1007 - 0.393x$ approximates the trend of world record times in the mile run. Compare to Figure A.1.*

3.3b have the same algebraic equation but displayed at different ranges of x, and only the second graph serves any applied goal.

We also see the difficulty of interpreting the intercept (the a in $y = a + bx$) in this example. In the equation $y = 1007 - 0.393x$, the intercept of 1007 seconds (equivalently, 16.8 minutes) represents the predicted world record time in the mile run in the year 0, an obviously inappropriate extrapolation. It could be better to express this model as $y = 260 - 0.393\,(x - 1900)$, for example, or perhaps $y = 241 - 0.393\,(x - 1950)$.

Finally we revisit the interpretation of the slope. The usual shorthand is that an increase of one year leads to a decrease in the world record times for the mile run by 0.393 seconds on average. However it is not so helpful to give this sort of implicit causal role to the passage of time. It's more precise to describe the result in a purely descriptive way, saying that when comparing any two years, we see a world record time that is, on average, 0.393 seconds per year less for the more recent year.

3.4 Exponential and power-law growth and decline; logarithmic and log-log relationships

The line $y = a + bx$ can be used to express a more general class of relationships by allowing logarithmic transformations.

The formula $\log y = a + bx$ represents exponential growth (if $b > 0$) or decline (if $b < 0$): $y = Ae^{bx}$, where $A = e^a$. The parameter A is the value of y when $x = 0$, and the parameter b determines the rate of growth or decline. A one-unit difference in x corresponds to an additive difference of b in $\log y$ and thus a multiplicative factor of e^b in y. Here are two examples:

- *Exponential growth.* Suppose that world population starts at 1.5 billion in the year 1900 and increases exponentially, doubling every 50 years (not an accurate description, just a crude approximation). We can write this as $y = A * 2^{(x-1900)/50}$, where $A = 1.5 * 10^9$. Equivalently, $y = A\,e^{(\log(2)/50))(x-1900)} = A\,e^{0.014(x-1900)}$. In statistics we use "log" to refer to the natural logarithm (log base e, not base 10) for reasons explained in Section 12.4.
  The model $y = A\,e^{0.014\,(x-1900)}$ is exponential growth with a rate of 0.014, which implies that y increases by a factor of $e^{0.014} = 1.014$ per year, or $e^{0.14} = 1.15$ per ten years, or $e^{1.4} = 4.0$ per hundred years. We can take the log of both sides of the equation to get $\log y = 21.1 + 0.014(x - 1900)$. Here, $\log A = \log(1.5 * 10^9) = 21.1$.
- *Exponential decline.* Consider an asset that is initially worth \$1000 and declines in value by 20% each year. Then its value at year x can be written as $y = 1000 * 0.8^x$ or, equivalently, $y = 1000\,e^{\log(0.8)x} = 1000\,e^{-0.22x}$. Logging both sides yields $\log y = \log 1000 - 0.22x = 6.9 - 0.22x$.

The formula $\log y = a + b \log x$ represents power-law growth (if $b > 0$) or decline (if $b < 0$):

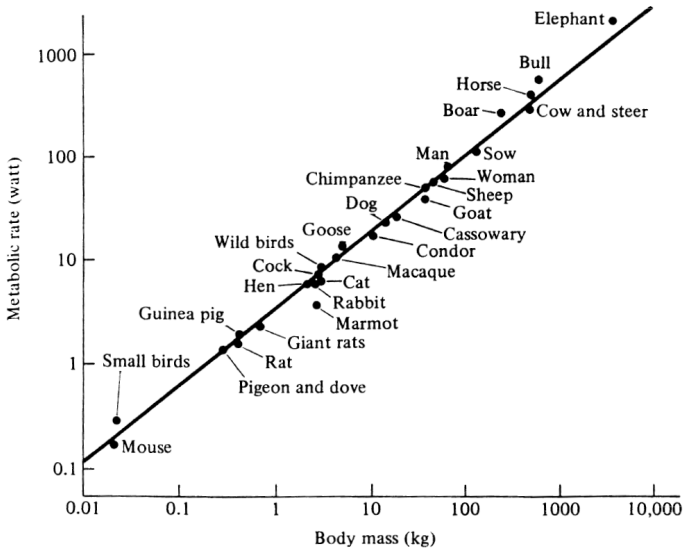

Figure 3.4 *Log metabolic rate vs. log body mass of animals, from Schmidt-Nielsen (1984). These data illustrate the log-log transformation. The fitted line has a slope of 0.74. See also Figure 3.5.*

$y = Ax^b$, where $A = e^a$. The parameter A is the value of y when $x = 1$, and the parameter b determines the rate of growth or decline. A one-unit difference in $\log x$ corresponds to an additive difference of b in $\log y$. Here are two examples:

- *Power law.* Let y be the area of a square and x be its perimeter. Then $y = (x/4)^2$, and we can take the log of both sides to get $\log y = 2\,(\log x - \log 4) = -2.8 + 2\log x$.
- *Non-integer power law.* Let y be the surface area of a cube and x be its volume. If L is the length of a side of the cube, then $y = 6L^2$ and $x = L^3$, hence the relation between x and y is $y = 6x^{2/3}$; thus, $\log y = \log 6 + \frac{2}{3}\log x = 1.8 + \frac{2}{3}\log x$.

Example: Metabolic rates of animals

Here is an example of how to interpret a power law or log-log regression.[2] Figure 3.4 displays data on log metabolic rate vs. log body mass indicating an approximate underlying linear relation. To give some context, the point labeled Man corresponds to a body mass of 70 kilogram and a metabolism of 80 watts; thus, a classroom with 100 men is the equivalent of a 8 000-watt space heater. By comparison, you could compute the amount of heat given off by a single elephant (which weighs about 3700 kilograms according to the graph) or 9 000 rats (which together also weigh about 3700 kilograms). The answer is that the elephant gives off less heat than the equivalent weight of men, and the rats give off more. This corresponds to a slope of less than 1 on the log-log scale.

What is the equation of the line in Figure 3.4? The question is not quite as simple as it looks, since the graph is on the log-log scale, but the axes are labeled on the original scale. We start by relabeling the axes on the logarithmic (base e) scale, as shown in Figure 3.5a. We can then determine the equation of the line by identifying two points that it goes through: for example, when $\log x = -3.8$, $\log y = -1.7$ (mouse), and when $\log x = 6.2$, $\log y = 5.7$ (cow and steer). So, comparing two animals where $\log x$ differs by 10, the average difference in $\log y$ is $5.7 - (-1.7) = 7.4$. The slope of the line is then $7.4/10.4 = 0.74$ (least squares fit would give slope 0.75). Since the line goes through the point $(0, 1.2)$, its equation can be written as,

$$\log y = 1.2 + 0.74 \log x.$$

We can exponentiate both sides of (3.1) to see the relation between metabolic rate and body mass on

[2] Code for this example is in the folder `Metabolic`.

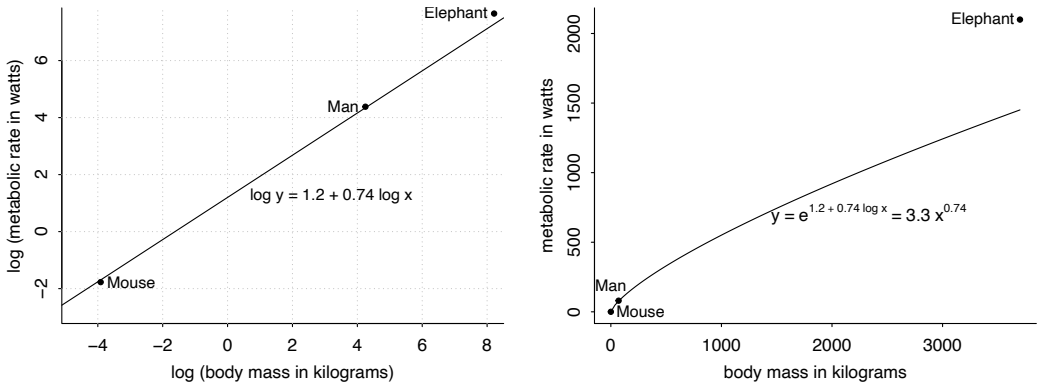

Figure 3.5 *Fitted curve (from data in Figure 3.4) of metabolic rate vs. body mass of animals, on the log-log and untransformed scales. The difference from the elephant's metabolic rate from its predictive value is relatively small on the logarithmic scale but large on the absolute scale.*

the untransformed scales:

$$\begin{aligned} e^{\log y} &= e^{1.2+0.74 \log x} \\ y &= 3.3\, x^{0.74}. \end{aligned} \tag{3.1}$$

This curve is plotted in Figure 3.5b. For example, when increasing body mass on this curve by a factor of 2, metabolic rate is multiplied by $2^{0.74} = 1.7$. Multiplying body mass by 10 corresponds to multiplying metabolic rate by $10^{0.74} = 5.5$, and so forth.

Now we return to the rats and the elephant. The relation between metabolic rate and body mass is less than linear (that is, the exponent 0.74 is less than 1.0, and the line in Fig. 3.5b is concave, not convex), which implies that the equivalent mass of rats gives off more heat, and the equivalent mass of elephant gives off less heat, than the men. This seems related to the general geometrical relation that surface area and volume are proportional to linear dimension to the second and third power, respectively, and thus surface area should be proportional to volume to the $\frac{2}{3}$ power. Heat produced by an animal is emitted from its surface, and it would thus be reasonable to suspect metabolic rate to be proportional to the $\frac{2}{3}$ power of body mass. Biologists have considered why the empirical slope is closer to $\frac{3}{4}$ than to $\frac{2}{3}$; our point here is not to discuss these issues but rather to show the way in which the coefficient in a log-log regression (equivalently, the exponent in power-law relation) can be interpreted.

See Section 12.4 for further discussion and examples of logarithmic transformations.

3.5 Probability distributions

In Section 3.3, we reviewed straight-line prediction, which is the deterministic part of linear regression and the key building block for regression modeling in general. Here we introduce probability distributions and random variables, which we need because our models do not fit our data exactly. Probability distributions represent the unmodeled aspects of reality—the *error term* ϵ in the expression $y = a + bx + \epsilon$—and it is randomness that greases the wheels of inference.

A probability distribution corresponds to an urn with a potentially infinite number of balls inside. When a ball is drawn at random, the "random variable" is what is written on this ball. Our treatment is not formal or axiomatic; rather, we mix conceptual definitions with mathematical formulas where we think these will be useful for using these distributions in practice.

Areas of application of probability distributions include:

- Distributions of data (for example, heights of men, incomes of women, political party preference), for which we use the notation $y_i,\ i = 1, \ldots, n$.

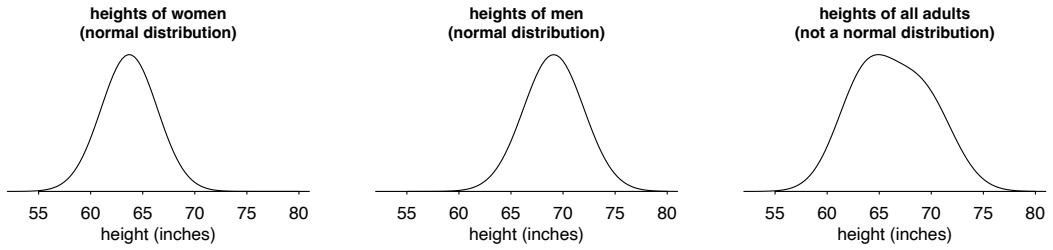

Figure 3.6 *(a) Heights of women, which approximately follow a normal distribution, as predicted from the Central Limit Theorem. The distribution has mean 63.7 and standard deviation 2.7, so about 68% of women have heights in the range* 63.7 ± 2.7. *(b) Heights of men, approximately following a normal distribution with mean 69.1 and standard deviation 2.9. (c) Heights of all adults in the United States, which have the form of a mixture of two normal distributions, one for each sex.*

- Distributions of error terms, which we write as $\epsilon_i,\ i = 1, \ldots, n$.

A key component of regression modeling is to describe the typical range of values of the outcome variable, given the predictors. This is done in two steps that are conceptually separate but which in practice are performed at the same time. The first step is to predict the average value of the outcome given the predictors, and the second step is to summarize the variation in this prediction. Probabilistic distributions are used in regression modeling to help us characterize the variation that remains *after* predicting the average. These distributions allow us to get a handle on how uncertain our predictions are and, additionally, our uncertainty in the estimated parameters of the model.

Mean and standard deviation of a probability distribution

A probability distribution of a random variable z takes on some range of values (the numbers written on the balls in the urn). The *mean* of this distribution is the average of all these numbers or, equivalently, the value that would be obtained on average from a random sample from the distribution. The mean is also called the expectation or expected value and is written as $\mathrm{E}(z)$ or μ_z. For example, Figure 3.6a shows the (approximate) distribution of heights of women in the United States. The mean of this distribution is 63.7 inches: this is the average height of all the women in the country and it is also the average value we would expect to see from sampling one woman at random.

The *variance* of the distribution of z is $\mathrm{E}((z - \mu_z)^2)$, that is, the mean of the squared difference from the mean. To understand this expression, first consider the special case in which z takes on only a single value. In that case, this single value is the mean, so $z - \mu_z = 0$ for all z in the distribution, and the variance is 0. To the extent that the distribution has variation, so that sampled values of z from the "urn" can be different, this will show up as values of z that are higher and lower than μ_z, and the variance of z is nonzero.

The *standard deviation* is the square root of the variance. We typically work with the standard deviation rather than the variance because it is on the original scale of the distribution. Returning to Figure 3.6a, the standard deviation of women's heights is 2.7 inches: that is, if you randomly sample a woman from the population, observe her height z, and compute $(z - 63.7)^2$, then the average value you will get is 7.3; this is the variance, and the standard deviation is $\sqrt{7.3} = 2.7$ inches. The variance of 7.3 is on the uninterpretable scale of inches squared.

Normal distribution; mean and standard deviation

The Central Limit Theorem of probability states that the sum of many small, independent random variables will be a random variable that approximates what is called a *normal distribution*. If we write this summation of independent components as $z = \sum_{i=1}^{n} z_i$, then the mean and variance of z are the sums of the means and variances of the z_i's: $\mu_z = \sum_{i=1}^{n} \mu_{z_i}$ and $\sigma_z = \sqrt{\sum_{i=1}^{n} \sigma_{z_i}^2}$. In statistical

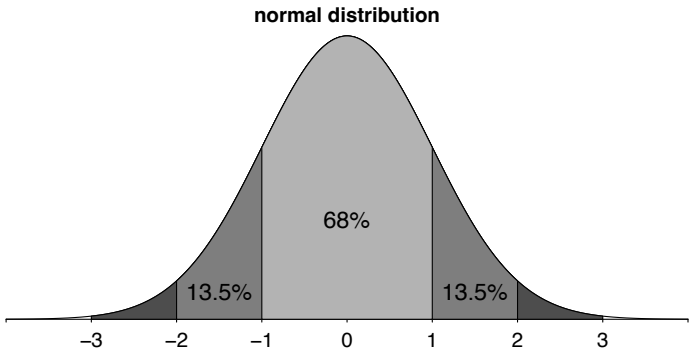

Figure 3.7 *Approximately 50% of the mass of the normal distribution falls within 0.67 standard deviations from the mean, 68% of the mass falls within 1 standard deviation from the mean, 95% within 2 standard deviations of the mean, and 99.7% within 3 standard deviations.*

notation, the normal distribution is written as $z \sim \text{normal}(\mu_z, \sigma_z)$. The Central Limit Theorem holds in practice—that is, $\sum_{i=1}^{n} z_i$ actually follows an approximate normal distribution—if the individual σ_{z_i}'s are small compared to the standard deviation σ_z of the sum.

We write the normal distribution with mean μ and standard deviation σ as $\text{normal}(\mu, \sigma)$. Approximately 50% of the mass of this distribution falls in the range $\mu \pm 0.67\sigma$, 68% in the range $\mu \pm \sigma$, 95% in the range $\mu \pm 2\sigma$, and 99.7% in the range $\mu \pm 3\sigma$. The 1 and 2 standard deviation ranges are shown in Figure 3.7. To put it another way, if you take a random draw from a normal distribution, there is a 50% chance it will fall within 0.67 standard deviations from the mean, a 68% chance it will fall within 1 standard deviation from the mean, and so forth.

Example: Heights of men and women

There's no reason to expect that a random variable representing direct measurements in the world will be normally distributed. Some examples exist, though. For example, the heights of women in the United States follow an approximate normal distribution. We can posit that the Central Limit Theorem applies here because women's height is affected by many small additive factors. In contrast, the distribution of heights of *all* adults in the United States is not so close to the normal curve. The Central Limit Theorem does not apply in this case because there is a single large factor—sex—that represents much of the total variation. See Figure 3.6c.[3]

The normal distribution is useful in part because summaries such as sums, differences, and estimated regression coefficients can be expressed mathematically as averages or weighted averages of data. There are many situations in which we can use the normal distribution to summarize uncertainty in estimated averages, differences, and regression coefficients, even when the underlying data do not follow a normal distribution.

Linear transformations

Linearly transformed normal distributions are still normal. If y is a variable representing men's heights in inches, with mean 69.1 and standard deviation 2.9, then 2.54 y is height in centimeters, with mean $2.54 * 69 = 175$ and standard deviation $2.54 * 2.9 = 7.4$.

For a slightly more complicated example, suppose we take independent samples of 100 men and 100 women and compute the difference between the average heights of the men and the women. This difference between the average heights will be approximately normally distributed with mean $69.1 - 63.7 = 5.4$ and standard deviation $\sqrt{2.9^2/100 + 2.7^2/100} = 0.4$; see Exercise 3.6.

[3]Code for this example is in the folder `CentralLimitTheorem`.

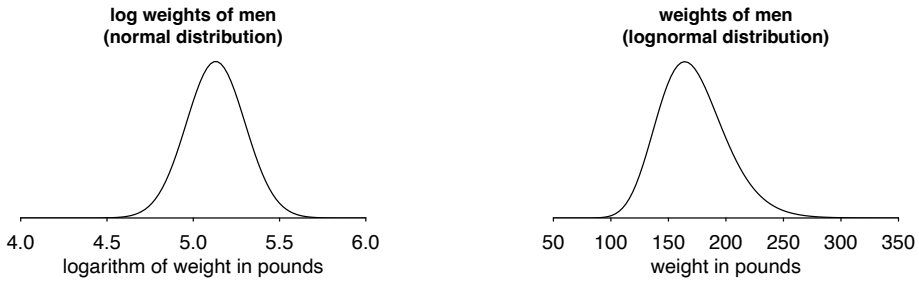

Figure 3.8 *Weights of men (which approximately follow a lognormal distribution, as predicted from the Central Limit Theorem from combining many small multiplicative factors), plotted on the logarithmic and original scales.*

Mean and standard deviation of the sum of correlated random variables

If two random variables u and v have mean μ_u, μ_v and standard deviations σ_u, σ_v, then their *correlation* is defined as $\rho_{uv} = \mathrm{E}((u - \mu_u)(v - \mu_v))/(\sigma_u \sigma_v)$. It can be shown mathematically that the correlation must be in the range $[-1, 1]$, attaining the extremes only when u and v are linear functions of each other.

Knowing the correlation gives information about linear combinations of u and v. Their sum $u + v$ has mean $\mu_u + \mu_v$ and standard deviation $\sqrt{\sigma_u^2 + \sigma_v^2 + 2\rho\sigma_u\sigma_v}$. More generally, the weighted sum $au + bv$ has mean $a\mu_u + b\mu_v$, and its standard deviation is $\sqrt{a^2\sigma_u^2 + b^2\sigma_v^2 + 2ab\rho\sigma_u\sigma_v}$. From this we can derive, for example, that $u-v$ has mean $\mu_u - \mu_v$ and standard deviation $\sqrt{\sigma_u^2 + \sigma_v^2 - 2\rho\sigma_u\sigma_v}$.

Lognormal distribution

Example: Weights of men

It is often helpful to model all-positive random variables on the logarithmic scale because it does not allow for values that are 0 or negative. For example, the logarithms of men's weights (in pounds) have an approximate normal distribution with mean 5.13 and standard deviation 0.17. Figure 3.8 shows the distributions of log weights and weights among men in the United States. The exponential of the mean and standard deviations of log weights are called the *geometric mean* and *geometric standard deviation* of the weights; in this example, they are 169 pounds and 1.18, respectively.

The logarithmic transformation is nonlinear and, as illustrated in Figure 3.8, it pulls in the values at the high end, compressing the scale of the distribution. But in general, the reason we perform logarithmic transformations is *not* to get distributions closer to normal, but rather to transform multiplicative models into additive models, a point we discuss further in Section 12.4.

Binomial distribution

If you take 20 shots in basketball, and each has 0.3 probability of succeeding, and if these shots are independent of each other (that is, success in one shot is not associated with an increase or decrease in the probability of success for any other shot), then the number of shots that succeed is said to have a *binomial distribution* with $n = 20$ and $p = 0.3$, for which we use the notation $y \sim \text{binomial}(n, p)$. Even in this simple example, the binomial model is typically only an approximation. In real data with multiple measurements (for example, repeated basketball shots), the probability p of success can vary, and outcomes can be correlated. Nonetheless, the binomial model is a useful starting point for modeling such data. And in some settings—most notably, independent sampling with Yes/No responses—the binomial model generally is appropriate, or very close to appropriate. The binomial distribution with parameters n and p has mean np and standard deviation $\sqrt{np(1-p)}$.

Poisson distribution

The *Poisson distribution* is used for count data such as the number of cases of cancer in a county, or the number of hits to a website during a particular hour, or the number of people named Michael whom you know:

- If a county has a population of 100 000, and the average rate of a particular cancer is 45.2 per million people per year, then the number of cancers in this county could be modeled as Poisson with expectation 4.52.
- If hits are coming at random, with an average rate of 380 per hour, then the number of hits in any particular hour could be modeled as Poisson with expectation 380.
- If you know approximately 750 people, and 1% of all people in the population are named Michael, and you are as likely to know Michaels as anyone else, then the number of Michaels you know could be modeled as Poisson with expectation 7.5.

As with the binomial distribution, the Poisson model is almost always an idealization, with the first example ignoring systematic differences among counties, the second ignoring clustering or burstiness of the hits, and the third ignoring factors such as sex and age that distinguish Michaels, on average, from the general population.

Again, however, the Poisson distribution is a starting point—as long as its fit to data is checked. We generally recommend the Poisson model to be expanded to account for "overdispersion" in data, as we discuss in Section 15.2.

Unclassified probability distributions

Real data will not in general correspond to any named probability distribution. For example, the distribution shown in Figure 3.6c of heights of all adults is not normal, or lognormal, or any other tabulated entry. Similarly, the distribution of incomes of all Americans has no standard form. For a discrete example, the distribution of the number of children of all U.S. adults does not follow the Poisson or binomial or any other named distribution. That is fine. Catalogs of named distributions are a starting point in understanding probability but they should not bound our thinking.

Probability distributions for error

Above we have considered distributions for raw data. But in regression modeling we typically model as much of the data variation as possible with a *deterministic model*, with a probability distribution included to capture the *error*, or unexplained variation. A simple and often useful model for continuous data is y = deterministic_part + error. A similar model for continuous positive data is y = deterministic_part $*$ error.

For discrete data, the error cannot be so easily mathematically separated from the rest of the model. For example, with binary data, the deterministic model will give predicted probabilities, with the error model corresponding to the mapping of these probabilities to 0 and 1. Better models yield predicted probabilities closer to 0 or 1. The probability that a U.S. professional basketball team wins when playing at home is about 60%; thus we can model the game outcome ($y = 1$ if the home team wins, 0 otherwise) as a binomial distribution with $n = 1$ and $p = 0.6$. However, a better job of prediction can be done using available information on the matchups before each game. Given a reasonably good prediction model, the probability that the home team wins, as assessed for any particular game, might be some number between 0.3 and 0.8. In that case, each outcome y_i is modeled as binomially distributed with $n = 1$ and a probability p_i that is computed based on a fitted model.

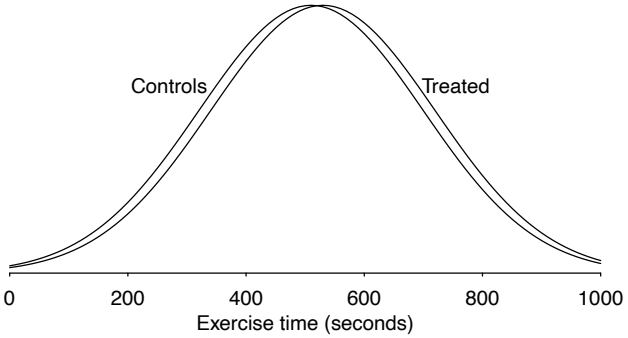

Figure 3.9 *Distributions of potential outcomes for patients given placebo or heart stents, using a normal approximation and assuming a treatment effect in which stents improve exercise time by 20 seconds, a shift which corresponds to taking a patient from the 50th to the 54th percentile of the distribution under the placebo.*

Comparing distributions

We typically compare distributions using summaries such as the mean, but it can also make sense to look at shifts in quantiles.

Example: Stents

We demonstrate with data from a clinical trial of 200 heart patients in which half received percutaneous coronary intervention (stents) and half received placebo, and at follow-up various health measures were recorded. Here, we consider a particular outcome, the amount of time that a patient was able to exercise on a treadmill. They show a pre-randomization distribution (averaging the treatment and control groups) with a mean of 510 seconds and a standard deviation of 190 seconds.[4]

The treatment effect was estimated as 20 seconds. How does this map to the distribution? One way to make the connection is to suppose that the responses under the control group follow a normal distribution, and suppose that, under the treatment, the responses would increase by 20 seconds; see Figure 3.9. Take a person at the median of the distribution, with an exercise time of 510 seconds under the control and an expected 530 seconds under the treatment. This corresponds to a shift from the 50th to the 54th percentile of the distribution. We compute these probabilities in R like this: `pnorm(c(510, 530), 510, 190)`. Comparing to the quantiles gives some context to an estimated treatment effect of 20 seconds, a number that could otherwise be hard to interpret on its own.

3.6 Probability modeling

Example: Probability of a decisive vote

What is the probability that your vote is decisive in an election? No dice, coins, or other randomization devices appear here, so any answer to this question will require assumptions. We use this example to demonstrate the challenges of probability modeling, how it can work, and how it can go wrong.

Consider an election with two candidates and n voters. An additional vote will be potentially decisive if all the others are divided equally (if n is even) or if the preferred candidate is otherwise one vote short of a tie (if n is odd).

We consider two ways of estimating this probability: an empirical forecasting approach that we recommend, and a binomial probability model that has serious problems.

Using an empirical forecast

Suppose an election is forecast to be close. Let y be the proportion of the vote that will be received by one of the candidates, and suppose we can summarize our uncertainty in y by a normal distribution with mean 0.49 and standard deviation 0.04: that is, the candidate is predicted to lose, but there is enough uncertainty that the vote could go either way. The probability that the n votes are exactly split

[4] Data and code for this example are in the folder `Stents`.

if n is even, or that they are one vote less than being split if n is odd, can then be approximated as $1/n$ times the forecast vote share density evaluated at 0.5. For example, in an election with 200 000 voters, we can compute this probability as `dnorm(0.5, 0.49, 0.04)/2e5`, which comes to `4.8e-5`, approximately 1 in 21 000.

This is a low probability for an individual voter; however it is not so low for a campaign. For example, if an additional 1000 voters could be persuaded to turn out for your candidate, the probability of *this* being decisive is the probability of the election being less than 1000 votes short of a tie, which in this case is approximately `1000*dnorm(0.5, 0.49, 0.04)/2e5`, or 1/21. A similar probability of decisiveness arises from convincing 10 000 people to vote in an election with 10 times as many other voters, and so on.

The key step in the above calculation is the probabilistic election forecast. The probability of a decisive vote can be substantial if the probability density is high at the 50% point, or very low if the forecast distribution is far from that midpoint. If no election-specific forecast is available, one can use a more generic probability model based on the distribution of vote shares in candidates in some large set of past elections. For example, the normal distribution with mean 0.5 and standard deviation 0.2 corresponds to a forecast that each candidate is likely to get somewhere between 30% and 70% of the vote. Under this very general model, we can calculate the probability of a tie from 200 000 voters as `dnorm(0.5, 0.5, 0.2)/2e5`, which comes to 1 in 100 000. This makes sense: if the election is not forecast to be particularly close, there is a smaller chance of it being exactly tied.

Using an reasonable-seeming but inappropriate probability model

We next consider an alternative calculation that we have seen but which we do not think makes sense. Suppose there are n voters, each with probability p of voting for a particular candidate. Then the probability of an exact tie, or one vote short of a tie, can be computed using the binomial distribution. For example, with $n = 200\,000$ and $p = 0.5$, the probability of a tie election can be computed in R as `dbinom(1e5, 2e5, 0.5)`, which comes to 0.0018, or 1/560.

If we shift to $p = 0.49$, the probability of a tie—that is, exactly 100 000 votes for each candidate—becomes `dbinom(1e5, 2e5, 0.49)`, which is approximately 10^{-20}, a number that is much too close to zero to make sense: unusual things happen in elections, and it is inappropriate to make a claim with such a high level of certainty about a future result

What went wrong? Most directly, the binomial model does not apply here. This distribution represents the number of successes in n independent trials, each with probability p. But voters are not making their decisions independently—they are being affected by common factors such as advertising, candidates' statements, and other news. In addition, voters do not have a shared probability p. Some voters are almost certainly going to choose one candidate, others are almost certainly on the other side, and voters in the middle have varying leanings.

But the evident wrongness of the model is not enough, by itself, to dismiss its implications. After all, the normal distribution used earlier is only an approximation for actual forecast uncertainty. And other examples of the binomial distribution we have used, such as successes in basketball shots, can work well for many purposes even while being just approximate.

The key problem in the election example is that the binomial model does not do a good job at capturing uncertainty. Suppose we were to provisionally consider the n voters as making independent decisions, and further assume (unrealistically) that they each have a common probability p of voting for a certain candidate, or perhaps p can be interpreted as an average probability across the electorate. Then where does p come from? In any real election, we would not know this probability; we might have a sense that the election is close (so p might be near 0.5), or that one side is dominant (so that p will likely be far from 0.5), or maybe we have no election-specific information at all (so that p could be in some wide range, corresponding to some prior distribution or reference set of historical elections). In any of these cases, p is not known, and any reasonable probability model must average over a distribution for p, which returns us to the forecasting problem used in the earlier calculation.

To put it another way, if we were to apply the model that the number of votes received by one

candidate in an election is binomial with $n = 200\,000$ and $p = 0.5$, this corresponds to *knowing* that the election is virtually tied, and this is not knowledge that would be available before any real election. Similarly, modeling the vote count as binomial with $n = 200\,000$ and $p = 0.49$ would imply an almost certain knowledge that the election outcome would be right around 0.49—not 0.485 and not 0.495—and that would not make sense for a real election. Those computed probabilities of 1/560 or 10^{-20} came from models that make very strong assumptions, as can be seen in this case by the extreme sensitivity of the output to the input.

General lessons for probability modeling

In some settings, the binomial probability model can make sense. A casino can use this sort of model to estimate its distribution of winnings from a slot machine or roulette wheel, under the assumption that there is no cheating going on and the machinery has been built to tolerance and tested sufficiently that the probability assumptions are reasonable. Casino management can also compute alternatives by perturbing the probability model, for example, seeing what might happen if the probability of winning in some game is changed from 0.48 to 0.49.

Ultimately we should check our probability models by their empirical implications. When considering the probability of a tied election, we can compare to past data. Elections for the U.S. House of Representatives typically have about 200 000 votes and they are occasionally very close to tied. For example, during the twentieth century there were 6 elections decided by fewer than 10 votes and 49 elections decided by fewer than 100 votes, out of about 20 000 elections total, which suggests a probability on the order of $1/40\,000$ that a randomly chosen election will be tied.

When a probability model makes a nonsensical prediction (such as that 10^{-20} probability of a tied election), we can take that as an opportunity to interrogate the model and figure out which assumptions led us astray. For another example, suppose we were to mistakenly apply a normal distribution to the heights of all adults. A glance at Figure 3.6c reveals the inappropriateness of the model, which in turn leads us to realize that the normal distribution, which applies to the sum of many small factors, does not necessarily fit when there is one factor—in this case, sex—that is very predictive of the outcome. Probability modeling is a powerful tool in part because when it goes wrong we can use this failure to improve our understanding.

3.7 Bibliographic note

Ramsey and Schafer (2001) and Snedecor and Cochran (1989) are good sources for classical statistical methods. A quick summary of probability distributions appears in Appendix A of Gelman et al. (2013).

The data on teachers come from Tables 603 and 246 of the 2010 edition of the *Statistical Abstract of the United States*. The example of animals' body mass and metabolism comes from Gelman and Nolan (2017, section 3.8.2); for further background, see Schmidt-Nielsen (1984). The data on heights and weights of Americans come from Brainard and Burmaster (1992). See Swartz and Arce (2014) and J. F. (2015) for more on the home-court advantage in basketball. The stents example comes from Gelman, Carlin, and Nallamothu (2019). The probability of a tie vote in an election is discussed by Gelman, King, and Boscardin (1998), Mulligan and Hunter (2003), Gelman, Katz, and Bafumi (2004), and Gelman, Silver, and Edlin (2012).

3.8 Exercises

3.1 *Weighted averages*: A survey is conducted in a certain city regarding support for increased property taxes to fund schools. In this survey, higher taxes are supported by 50% of respondents aged 18–29, 60% of respondents aged 30–44, 40% of respondents aged 45–64, and 30% of respondents aged 65 and up. Assume there is no nonresponse.

Suppose the sample includes 200 respondents aged 18–29, 250 aged 30–44, 300 aged 45–64, and 250 aged 65+. Use the weighted average formula to compute the proportion of respondents in the *sample* who support higher taxes.

3.2 *Weighted averages*: Continuing the previous exercise, suppose you would like to estimate the proportion of all adults in the *population* who support higher taxes, so you take a weighted average as in Section 3.1. Give a set of weights for the four age categories so that the estimated proportion who support higher taxes for all adults in the city is 40%.

3.3 *Probability distributions*: Using R, graph probability densities for the normal distribution, plotting several different curves corresponding to different choices of mean and standard deviation parameters.

3.4 *Probability distributions*: Using a bar plot in R, graph the Poisson distribution with parameter 3.5.

3.5 *Probability distributions*: Using a bar plot in R, graph the binomial distribution with $n = 20$ and $p = 0.3$.

3.6 *Linear transformations*: A test is graded from 0 to 50, with an average score of 35 and a standard deviation of 10. For comparison to other tests, it would be convenient to rescale to a mean of 100 and standard deviation of 15.

(a) Labeling the original test scores as x and the desired rescaled test score as y, come up with a linear transformation, that is, values of a and b so that the rescaled scores $y = a + bx$ have a mean of 100 and a standard deviation of 15.

(b) What is the range of possible values of this rescaled score y?

(c) Plot the line showing y vs. x.

3.7 *Linear transformations*: Continuing the previous exercise, there is another linear transformation that also rescales the scores to have mean 100 and standard deviation 15. What is it, and why would you *not* want to use it for this purpose?

3.8 *Correlated random variables*: Suppose that the heights of husbands and wives have a correlation of 0.3, husbands' heights have a distribution with mean 69.1 and standard deviation 2.9 inches, and wives' heights have mean 63.7 and standard deviation 2.7 inches. Let x and y be the heights of a married couple chosen at random. What are the mean and standard deviation of the average height, $(x + y)/2$?

3.9 *Comparison of distributions*: Find an example in the scientific literature of the effect of treatment on some continuous outcome, and make a graph similar to Figure 3.9 showing the estimated population shift in the potential outcomes under a constant treatment effect.

3.10 *Working through your own example*: Continuing the example from Exercises 1.10 and 2.10, consider a deterministic model on the linear or logarithmic scale that would arise in this topic. Graph the model and discuss its relevance to your problem.