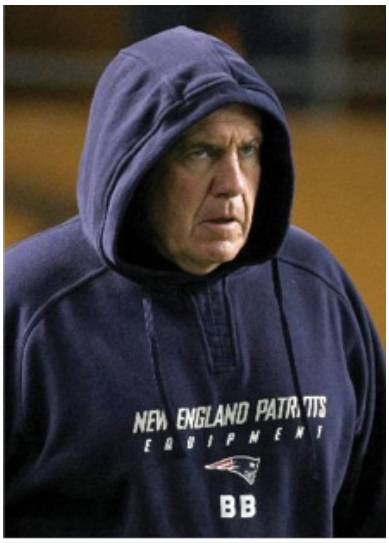
7
Testing hypotheses

Assessing the evidence for a hypothesis



Among professional football fans, the New England Patriots are a polarizing team. Their fan base is hugely devoted, probably due to their long run of success over more than a decade. Many others, however, dislike the Patriots for their highly publicized cheating episodes, whether for deflating footballs or clandestinely filming the practice sessions of their opponents. This feeling is so common among football fans that sports websites often run images like the one at right (of the Patriots' be-hoodied head coach, Bill Belichick), or articles with titles like "11 reasons why people hate the Patriots." Despite—or perhaps because of—their success, the Patriots always seems to be dogged by scandal and ill will.

But could even the Patriots cheat at the pre-game *coin toss*?

Believe it or not, many people think so! That's because, for a stretch of 25 games spanning the 2014-15 NFL seasons, the Patriots won 19 out of 25 coin tosses—that's a 76% winning percentage. Needless to say, the Patriots' detractors found this infuriating. As one TV commentator remarked when this unusual fact was brought to his attention: "This just proves that either God or the devil is a Patriots fan, and it sure can't be God."

But before turning to religion, let's take a closer look at the evidence. Just how likely is it that one team could win the pre-game coin toss at least 19 out of 25 times, assuming that there's no cheating going on?

This question is easy to answer using probability theory—specifically, something called the binomial distribution. But it's also very easy to answer using the Monte Carlo method, in which we write a computer program that simulates a random process. In Figure 7.1, we see the results of a Monte Carlo simulation for pre-game NFL coin tosses, where the Patriots ought to have a 50% chance of winning each toss. Specifically, we have repeated the

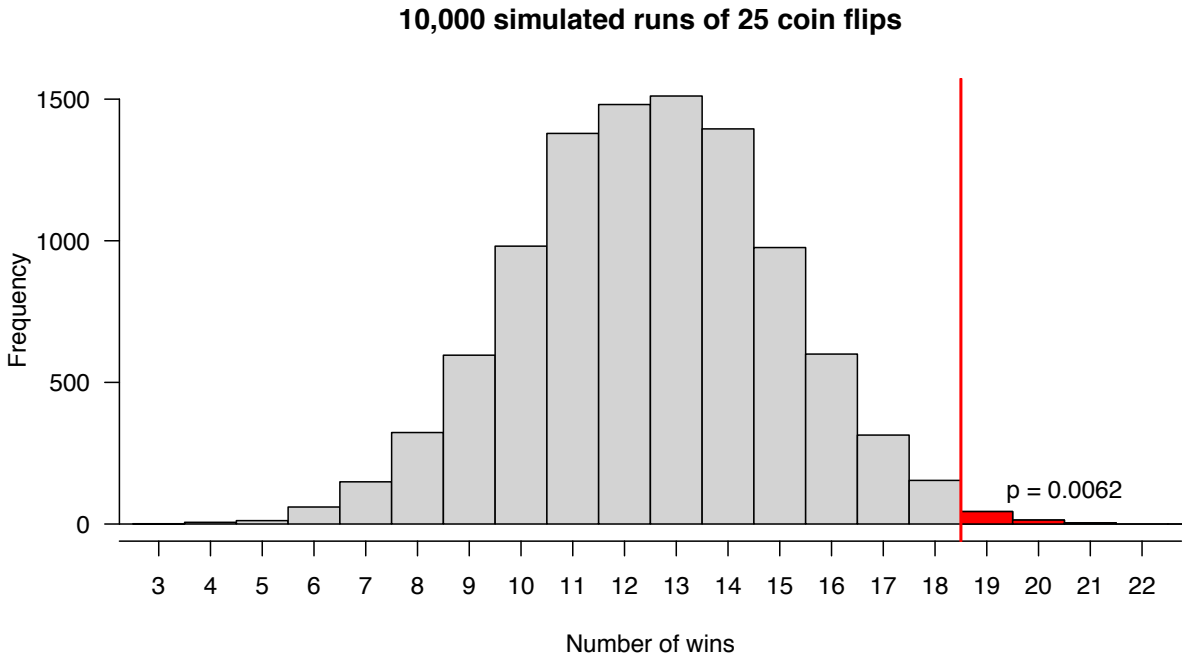


Figure 7.1: This histogram shows the results of a Monte Carlo simulation, in which we count the number of wins in 25 simulated coin flips over 10,000 different simulations. The red area (which has cumulative probability of 0.0062) approximates the probability of winning 19 or more flips, out of 25.

following simple process 10,000 times:

1. Simulate 25 coin tosses in which the Patriots have a 50% chance of winning each toss.
2. Count how many times out of 25 that the Patriots won the toss.

If you're counting, that's 250,000 coin tosses: 10,000 simulations of 25 tosses each.

Figure 7.1 shows a histogram of the number of coin tosses won by the Patriots across 10,000 simulations. Clearly 19 wins is an unusual, although not impossible, number under this distribution: in our simulation, the Patriots won at least 19 tosses only 62 of 10,000 times ($p = 0.0062$), shown as the red area in Figure 7.1.

So did the Patriots win 19 out of 25 coin tosses by chance? Well, nobody knows for sure—I report, you decide.[1] But unless you're a hard-core NFL conspiracy theorist, let me encourage you to forget the Patriots for a moment and focus instead on the process we've just gone through. This simple example has all the major elements of *hypothesis testing*, which is the subject of this chapter:

[1] Despite the small probability of such an extreme result, it's hard to believe that the Patriots cheated on the coin toss, for a few reasons. First, how could they? The coin toss would be extremely hard to manipulate, even if you were inclined to do so. Moreover, the Patriots are just one team, and this is just one 25-game stretch. There are 32 NFL teams, so the probability that *one* of them would go on an unusual coin-toss winning streak over *some* 25-game stretch over a long time period is a lot larger than the number we've calculated. Finally, after this 25-game stretch, the Patriots reverted back to a more typical coin-toss winning percentage, closer to 50%. The 25-game stretch was probably just luck.

(1) We have a *null hypothesis*, that the pre-game coin toss in the Patriots' games was truly random.

(2) We use a *test statistic*, number of Patriots' coin-toss wins, to measure the evidence against the null hypothesis.

(3) There is a way of calculating the probability distribution of the test statistic, assuming that the null hypothesis is true. Here, we just ran a Monte Carlo simulation of coin flips, assuming an unbiased coin.

(4) Finally, we used this probability distribution to assess whether the null hypothesis looked believable in light of the data.

All hypothesis testing problems have these same four elements. Usually the difficult part is Step 3: calculating the probability distribution of the test statistic, assuming that the null hypothesis is true. The essence of the problem is that, in most cases, we can't just run a simple simulation of coin flips. Luckily, there is a very general way of proceeding here, called the permutation test, which we will now learn about.

Permutation tests

Is gun violence correlated with gun policy?

Gun policy is an important and emotionally charged topic in 21st-century America, where gun violence occurs with far higher frequency than it does in other rich countries. Many people feel strongly that certain types of guns, like military-style assault weapons, should be banned, and that all gun purchases should be subject to stronger background checks. Others view gun ownership as both an important part of their cultural heritage and a basic right protected by the U.S. Constitution. Like with many issues, there seems to be little prospect of a national consensus.

Both gun laws, and the likelihood of dying violently as a result of gun crime, vary significantly from state to state. Figure 7.2 shows some of this variation in a *chloropleth map*, where discrete areas on the map are shaded according to the value of some numerical variable. Notice that the states are shown as a gridded tile of equal-sized hexagons, rather than as an actual map of the United States. This is common technique used to avoid the visual imbalances due to large differences in the states' total area.

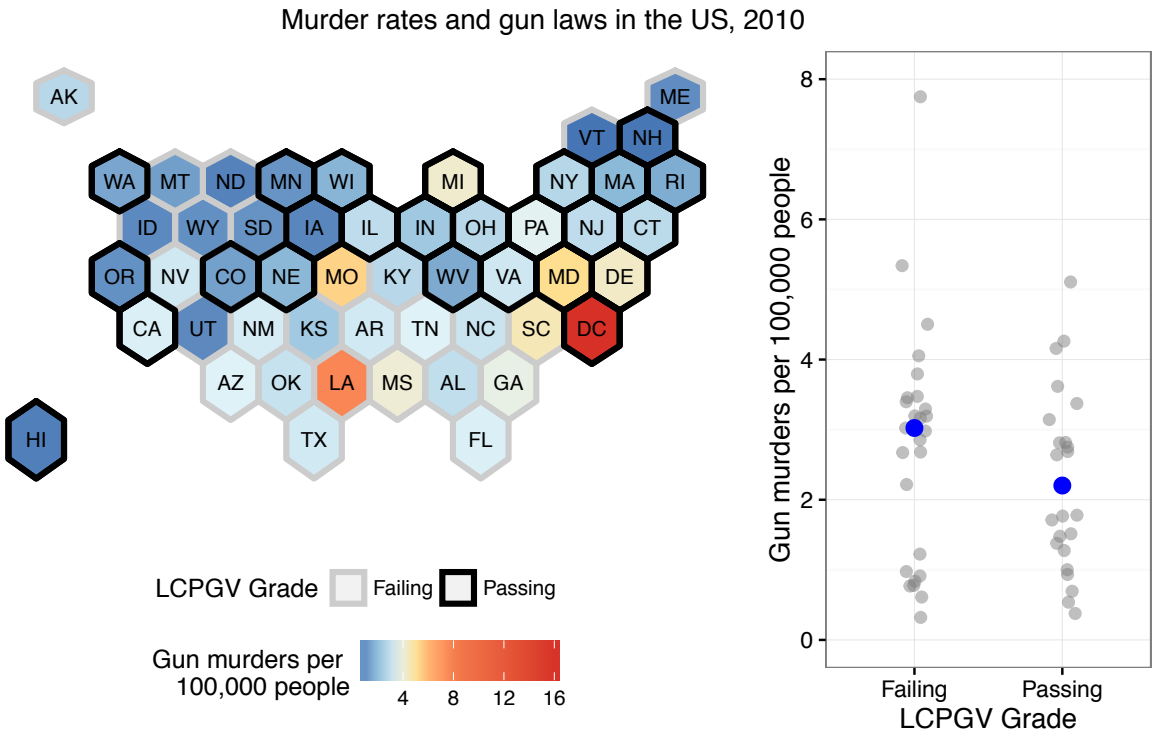


Figure 7.2: Left panel: a chloropleth map of murder rates versus gun laws across the U.S. states. The shaded color shows the state's gun-murder rate; blue is lower, and red is higher. The outline indicates whether a state's gun-control laws received a passing or a failing grade from the Law Center to Prevent Gun Violence (black for passing, grey for failing). The right panel shows a dot plot of the gun-murder rates across the two groups, together with the median for each group in blue. Washington (D.C.), at 16.2 gun murders per 100,000 people, is far off the top of the plot, but is still included in all calculations. According to its website, `http://smartgunlaws.org`, the LCPGV is "a national law center focused on providing comprehensive legal expertise in support of gun violence prevention and the promotion of smart gun laws that save lives." You can read a full description of the methodology used to grade states at this link.

In the chloropleth map in Figure 7.2, the fill color indicates each state's gun-murder rate in 2010: blue is lower, red is higher. The outline color indicates whether a state's gun-control laws received a passing or failing grade from the Law Center to Prevent Gun Violence (LCPGV). The center graded each state's gun laws on an A–F letter-grade scale; here "failing" means a grade of F. In the figure, a black outline means a passing grade, while a grey outline means a failing grade.

The right panel of Figure 7.2 summarizes the relationship between gun laws and gun violence via a dot plot, together with the median for each group in blue. We use the median rather than the mean to estimate the center of each group, because the median is more robust to outliers; a clear example of an outlier here is Washington (D.C.), which at 16.2 gun murders per 100,000 people has a drastically higher rate than everywhere else in the country.

This dotplot shows that the median murder rate of states with a failing gun-laws grade is 3 murders per 100,000 people, while the median murder rate of states with a passing grade is 2.2 per

100,000. On the face of it, it would seem as the states with stricter gun laws have lower murder rates.

Let's aside for a moment the fact that correlation does not establish causality. We will instead address the question: could this association have arisen due to chance? To make this idea more specific, imagine we took all 50 states and randomly divided them into two groups, arbitrarily labeled the "passing" states and the "failing" states. We would expect that the median murder rate would differ a little bit between the two groups, simply due to random variation (for the same reason that hands in a card game vary from deal to deal). But how big of a difference between these two groups could be explained by chance?

Null and alternative hypotheses

Thus there are two hypotheses that can explain Figure 7.2:

(1) There is no systematic relationship between murder rates and gun laws; the observed observed relationship between murder rates and gun laws is consistent with other unrelated sources of random variation.

(2) The observed relationship between murder rates and gun laws is too large to be consistent with random variation.

We call hypothesis 1 the *null hypothesis*, often denoted H_0. Loosely, it states that nothing special is going on in our data, and that any relationship we thought might have existed isn't really there at all.[2] Meanwhile, hypothesis 2 is *alternative hypothesis*. In some cases the alternative hypothesis may just be the logical negation of the null hypothesis, but it can also be more specific.

In the approach to hypothesis testing that we'll learn here, we don't focus a whole lot on the alternative hypothesis.[3] Instead, we set out to check whether the null hypothesis looks plausible in light of the data—just as we did when we tried to check whether randomness could explain the Patriots' impressive run of 19 out of 25 coin flips won.

A permutation test: shuffling the cards

In the Patriots' coin-flipping example, we could easily simulate data under the null hypothesis, by programming a computer to repeatedly flip a virtual coin and keep track of the winner. But of course, most real-life hypothesis-testing situations don't involve

[2] "Null hypothesis" is a term coined in the early twentieth century, back when "null" was a common synonym for "zero" or "lacking in distinctive qualities." So if the term sounds dated, that's because it is.

[3] Specifically, this approach is called the *Fisherian* approach, named after the English statistician Ronald Fisher. There are more nuanced approaches to hypothesis testing in which the alternative hypothesis plays a major role. These include the Neyman–Pearson framework and the Bayesian framework, both of which are widely used in the real world, but which are a lot more complicated to understand.

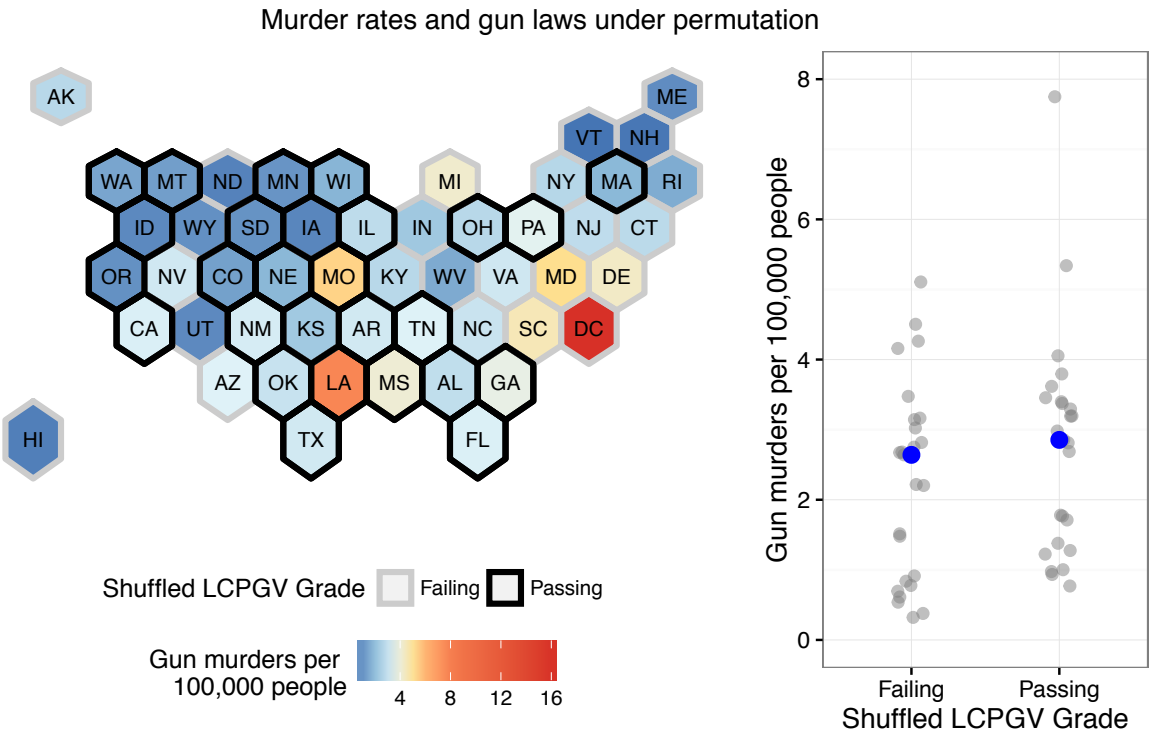


Figure 7.3: This map is almost identical to Figure 7.2, with one crucial difference: the identities of the states with passing and failing grades have been randomly permuted. There is still a small difference in the medians of the notionally passing and failing groups, due to random variation in the permutation process.

actual coin flips, which makes the virtual coin-flipping approach somewhat unhelpful as a general strategy.

It turns out, however, that in most situations, we can still harness the power of Monte Carlo simulation to understand what our data would look like if the null hypothesis were true. Rather than flipping virtual coins, we run something called a *permutation test*, which involves repeatedly permuting (or shuffling) the predictor variable and recalculating the statistic of interest.

To understand how this works, let's see an example. Figure 7.3 shows a map and dotplot very similar to those in Figure 7.2, with one crucial difference: in Figure 7.3, the identities of the states with notionally "passing" and "failing" gun laws have been randomly permuted. These grades bear no correspondence to reality. It's as though we took a deck of 51 cards, each card having some state's grade on it (treating D.C. as a state); shuffled the deck; and then dealt one card randomly to each state. The mathematical term for this is a *permutation* of the grades.

As expected, the median gun-murder rates of these two ran-

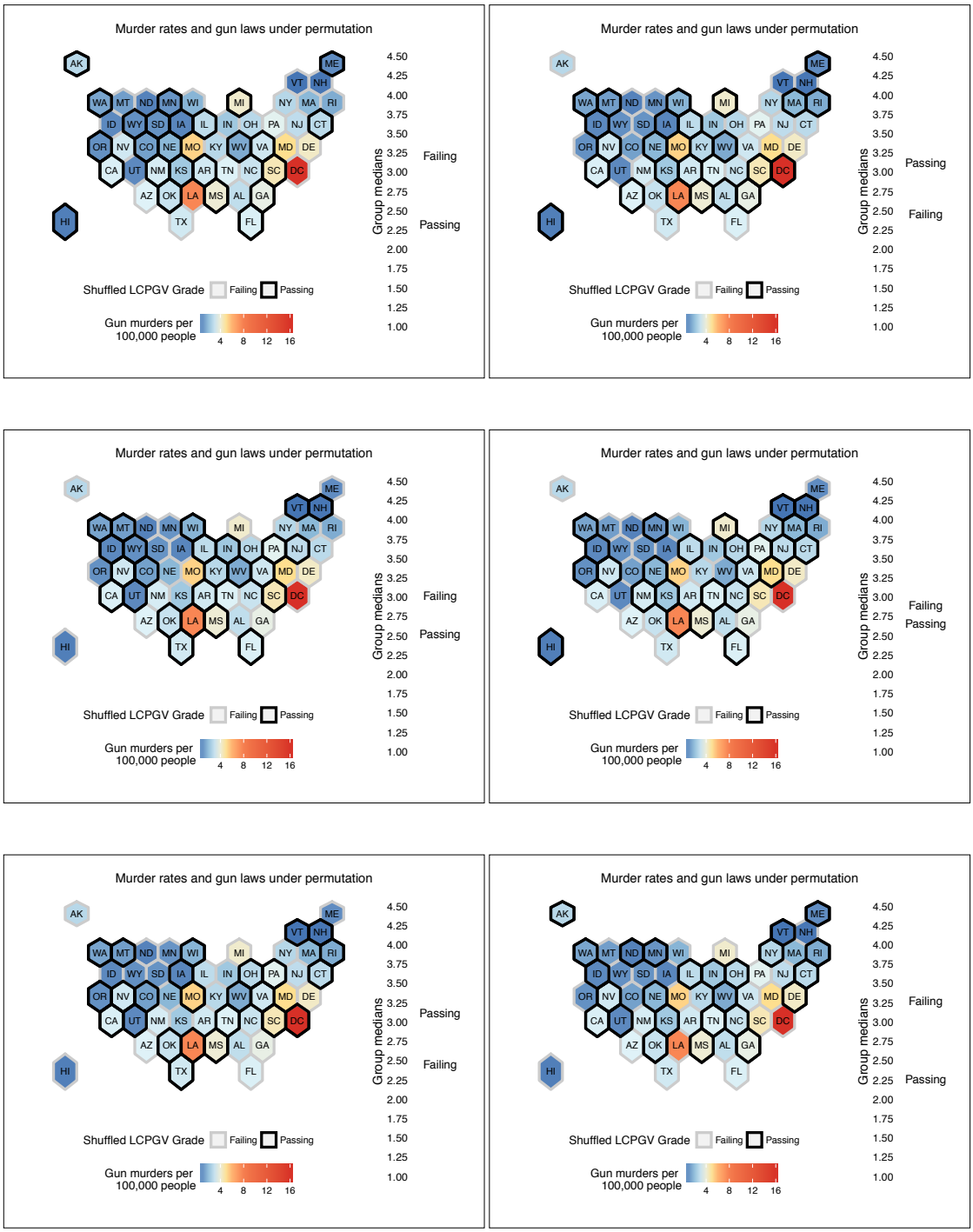


Figure 7.4: Six maps with permutated gun-law grades, with the medians for the passing and failing groups.

dom chosen "passing" and "failing" groups aren't identical (right panel). The randomly chosen "failing" states have a median of 2.6, while the randomly chosen "passing" states have a slightly larger median of 2.8. Clearly we can get a difference in medians of at least 0.2 quite easily, just by random chance—that is, when the null hypothesis is true by design.

But Figure 7.3 shows the difference in medians for only a single permutation of the states' gun-law grades. This permutation is random, and a different permutation would have given as a slightly different answer. Therefore, to assess whether could we get a difference in group medians as large as 0.8 just by random chance, we need to try several more permutations.

Figure 7.4 shows 6 more maps generated using the same permutation procedure. For each map, we shuffle the grade variables for all the states and recompute the median murder rates for the notionally "passing" and "failing" groups. Each map leads to its own difference in medians. In some maps, the difference is positive ("passing" states are higher), while in others it is negative ("failing" are states higher). In at least one of the 6 maps—the bottom right one—the median for the "failing" states exceeds the median for the "passing" states by more than 1 murder per 100,000 people, just by chance. This is a larger difference than we see for the real map, in Figure 7.2.

Six permutations give us some idea of how much a difference in the medians we could expect to see if the null hypothesis were true. But ideally we'd have many more than 6. Figure 7.5 addresses this need, showing the result of a much larger Monte Carlo simulation in which we generated 5,000 random maps, each one with its own random permutation of the states' gun-law grades. For each of these 5,000 maps, we computed the difference in medians between the notionally passing and failing groups. These 5,000 differences in group medians across the 5,000 maps are shown as a histogram in Figure 7.5.

Hypothesis testing: a four-step process

Let's review the vocabulary that describes what we've done here. First, we specified a null hypothesis: that the correlation between rates of gun violence and state-level gun policies could be explained by other unrelated sources of random variation. We decided to measure this correlation using a specific statistic: the difference in medians between the states with passing grades and

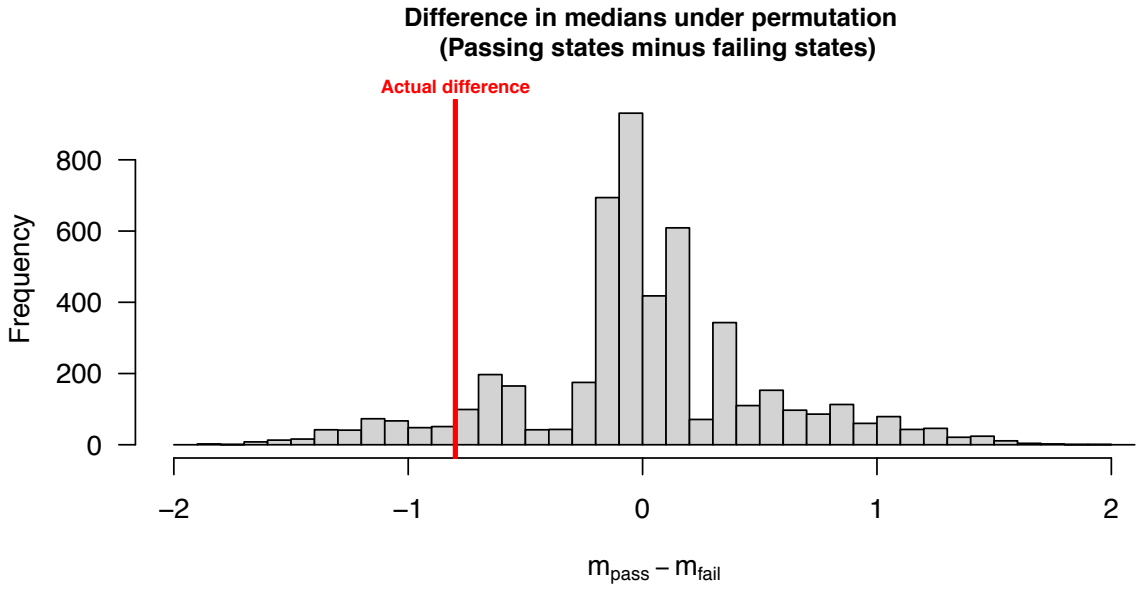


Figure 7.5: The histogram shows the difference in group medians for 5,000 simulated maps generated by the same permutation procedure as the 6 maps in Figure 7.4. Negative values indicate that the "failing" states had higher rates of gun violence than the "passing" states. The actual difference in medians for the real map in Figure 7.2 is shown as a vertical red line. This difference seems to be consistent with (although does not prove) the null hypothesis that other sources of random variation, and not necessarily state-level gun policy, explains the observed difference in murder rates.

those with failing grades. (Remember that a statistic is just some numerical summary of a data set.) To give this statistic a name, let's call it Δ (for difference in medians). It's intuitively clear that the larger Δ is, the less plausible the null hypothesis seems.

Figure 7.5 quantifies this intuition by giving us an idea of how much variation we can expect in the sampling distribution of our Δ statistic under the hypothesis that there is no systematic relationship between gun laws and rates of gun violence. As before, the sampling distribution is simply the probability distribution of the statistic under repeated sampling from the population—in this case, assuming that the null hypothesis is true.

There are two possibilities here, corresponding to the null and alternative hypotheses. First, suppose that we frequently get at least as extreme a value of Δ for a random map, like those in Figure 7.4, as we do in the real map from Figure 7.2. Then there's no reason to be especially impressed by the actual value of $\delta = -0.8$ we calculated from the real map.[4] It could have easily happened by chance. Hence we will be unable to reject the null hypothesis; it could have explained the data after all. (An important thing to remember is that *failing to reject* the null hypothesis is not the

[4] We use the lower-case δ to denote the value of the test statistic for your specific sample, to distinguish it from the Δ's simulated under permutation.

same thing as *accepting* the null hypothesis as truth. To use a relationship metaphor: failing to reject the null hypothesis is not like getting married. It's more like agreeing not to break up this time.)

On the other hand, suppose that we almost always get a smaller value of Δ in a random map than we do in the real map. Then we will probably find it difficult to believe that the correlation in the real map arose due to chance. We will instead be forced to reject the null hypothesis and conclude that it provides a poor description of the observable data.

Which of these two possibilities seems to apply in Figure 7.5? Here, the actual difference of -0.8 for the real map in Figure 7.2 is shown as a vertical red line. It's position on the histogram suggests possibility (1) here: $\delta = -0.8$ is consistent with (although does not prove) the null hypothesis that other sources of random variation unrelated to state-level gun policy can explain the observed difference in murder rates between the passing-grade and the failing-grade states.

To summarize, the four steps we followed above were:

(1) Choose a null hypothesis H_0, the hypothesis that there is no systematic relationship between the predictor and response variables.

(2) Choose a test statistic Δ that is sensitive to departures from the null hypothesis.

(3) Approximate $P(\Delta \mid H_0)$, the sampling distribution of the test statistic T under the assumption that H_0 is true.

(4) Assess whether the observed test statistic for your data, δ, is consistent with $P(\Delta \mid H_0)$.

For the gun-laws example, our test statistic in step (2) was the difference in medians between the "passing" states and the "failing" states. We then accomplished step (3) by randomly permuting the values of the predictor (gun laws) and recomputing the test statistic for the permuted data set. This shuffling procedure is called a permutation test when it's done in the context of this broader four-step process. There are other ways of accomplishing step (3)—for example, by appealing to probability theory and doing some math. But the permutation test is nice because it works for any test statistic (like the difference of medians in the previous example), and it doesn't require any strong assumptions.

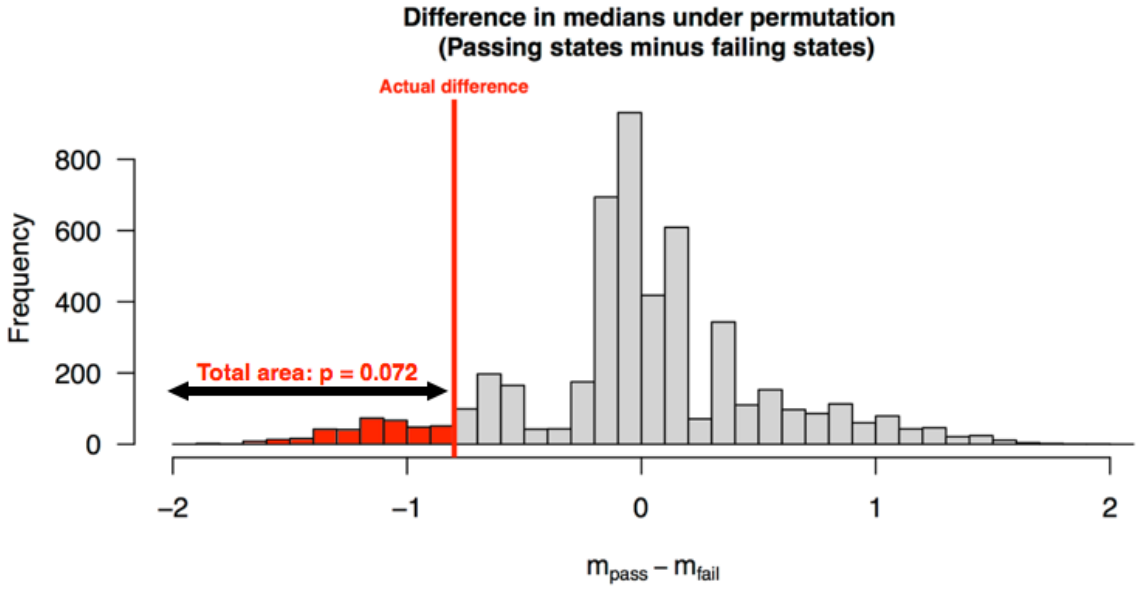


Figure 7.6: Assuming that the null hypothesis is true, the probability of observing a difference in medians at least as extreme as $\delta = -0.8$ is $p = 0.072$. This tail area to the left of $\delta = -0.8$ is the p-value of the test.

Using and interpreting p-values

There's one final question we haven't answered. How do we accomplish step (4) in the hypothesis test? That is, how can we measure whether the observed statistic for your data is consistent with the null hypothesis?

The typical approach here is to compute something called a *p-value*. Although we didn't call it by the name "p-value," this is exactly what we did for the Patriots' coin-flipping example at the beginning of the chapter.

Let's begin with a concise definition of a p-value, before we slowly unpack the definition (which is dense and non-intuitive). *A p-value is the probability of observing a test statistic as extreme as, or more extreme than, the test statistic actually observed, given that the null hypothesis is true.* The way to compute the p-value is to calculate a *tail area* indicating what proportion of the sampling distribution, $P(\Delta \mid H_0)$, lies beyond the observed test statistic δ.

This all sounds a bit abstract, but is much easier to understand by example. Let's go back to the gun-laws hypothesis test, where we observed a difference in the medians of $\delta = -0.8$. If the null hypothesis were true, the probability of getting $\delta = -0.8$ (or

something more extreme in the negative direction) would be $p = 0.072$. We calculate this by taking the tail area under the sampling distribution that to the left of our observed δ of -0.8. Figure 7.6 highlights this area in the left tail of the sampling distribution $P(\Delta \mid H_0)$. This is the p-value.

Using p-values has both advantages and disadvantages. The main advantage is that the p-value gives us a continuous measure of evidence against the null hypothesis. The smaller the p-value, the more unlikely it is that we would have seen our data under the null hypothesis, and therefore the greater the evidence the data provide that H_0 is false.

The main disadvantage is that the p-value is hard to interpret correctly. Just look at the definition—it's pretty counterintuitive! To avoid having to think too hard about what a p-value actually means, people often take $p \leq 0.05$ as a very important threshold that demarcates "significant" ($p \leq 0.05$) from "insignificant" ($p > 0.05$) results. While there are some legitimate reasons[5] for thinking in these terms, in practice, the $p \leq 0.05$ criterion can feel pretty silly. After all, there isn't some magical threshold at which a result becomes important: in all practical terms, $p = .049$ and $p = .051$ are nearly identical in terms of the amount of evidence they provide against a null hypothesis.

Because of how counterintuitive p-values are, people make mistakes with them all the time, even (perhaps especially) people with Ph.D's quoting p-values in original research papers. Here is some advice about a few common misinterpretations:

- The p-value is *not* the probability that the null hypothesis is true, given that we have observed our statistic.
- The p-value is *not* the probability of having observed our statistic, given that the null hypothesis is true. Rather, it is the probability of having observed our statistic, *or any more extreme statistic*, given that the null hypothesis is true.
- The p-value is *not* the probability that your procedure will falsely reject the null hypothesis, given that the null hypothesis is true.[6]

The moral of the story is: always be careful when quoting or interpreting p-values. In many circumstances, a better question to ask than "what is the p-value?" is "what is a plausible range for the size of the effect?" This question can be answered with a confidence interval.[7]

[5] If you are interested in these reasons, you should read up on the Neyman–Pearson school of hypothesis testing.

[6] To get a guarantee of this sort, you have to set up a pre-specified rejection region for your p-value (like 0.05), in which case the size of that rejection region—and not the observed p-value itself—can be interpreted as the probability that your procedure will reject the null hypothesis, given that the null hypothesis is true. As above: if you're interested, read about the Neyman–Pearson approach to testing.

[7] In this case, you could get a confidence interval by bootstrapping the difference in medians between the two groups of states.

Hypothesis testing in regression

To finish off this chapter, we will show how the permutation-testing framework can be used to answer questions about partial relationships in multiple regression modeling.

In a previous chapter, we asked the following question about houses in Saratoga, NY: what is the partial relationship between heating system type (gas, electric, or fuel oil) and sale price, once we adjust for the effect of living area, lot size, and the number of fireplaces? We fit a multiple regression model with these four predictors, which led to the following equation:

$$\begin{aligned}\text{Price} = \$29868 &+ 105.3 \cdot \text{SqFt} + 2705 \cdot \log(\text{Acres}) + 7546 \cdot \text{Fireplaces} \\ &- 14010 \cdot \mathbf{1}_{\{\text{fuel = electric}\}} - 15879 \cdot \mathbf{1}_{\{\text{fuel = oil}\}} + \text{Residual}\,.\end{aligned}$$

Remember that the baseline case here is gas heating, since it has no dummy variable. Our model estimated the premium associated with gas heating to be about $14,000 over electric heating, and about $16,000 over fuel-oil heating.

But are these differences due to heating-system type statistically significant, or could they be explained due to chance?

To answer this question, you could look at the confidence intervals for every coefficient associated with the heating-system variable, just as we learned to do in the chapter on multiple regression. The main difference is that before, we had one coefficient to look at, whereas now we have two: one dummy variable for fuel = electric, and one for fuel = oil. Two coefficients means two confidence intervals to look at.

Sometimes this strategy—that is, looking at the confidence intervals for all coefficients associated with a single variable—works just fine. For example, when the confidence intervals for all coefficients associated with a single variable are very far from zero, it's pretty obvious that the categorical variable in question is statistically significant.

But at other times, this strategy can lead to ambiguous results. In the context of the heating-system type variable, what if the 95% confidence interval for one dummy-variable coefficient contains zero, but the other doesn't? Or what if both confidence intervals contain zero, but just barely? Should we say that heating-system type is significant or not? This potential for ambiguous confidence intervals gets even worse when your categorical variable has more than just a few levels, because then there will be many more confi-

dence intervals to look at.

The core of the difficulty here is that we want to assess the significance of the heating-system variable itself, not the significance of any individual *level* of that variable. To assess the significance of the whole variable, with all of its levels, we'll use a permutation test. Specifically, we will compare two models:

- The *full model*, which contains variables for square footage, lot size, number of fireplaces, and heating system.
- The *reduced model*, which contains variables for square footage, lot size, and number of fireplaces, but not for heating system. We say that the reduced model is *nested* within the full model, since it contains a subset of the variables in the full model, but no additional variables.

As always, we must start by specifying H_0. Loosely speaking, our null hypothesis is that the reduced model provides an adequate description of house prices, and that the full model is needlessly complex. To be a bit more precise: the null hypothesis is that *there is no partial relationship* between heating system and house prices, once we adjust for square footage, lot size, and number of fireplaces. This implies that all of the *true* dummy variable coefficients for heating-system type are zero.

Next, we must pick a test statistic. A natural way to assess the evidence against the null hypothesis is to use improvement in R^2 under the full model, compared to the reduced model. This is the same quantity we look at when assessing the importance of a variable in an ANOVA table. The idea is simple: if we see a big jump in R^2 when moving from the reduced to the full model, then the variable we added (here, heating system) is important for predicting the outcome, and the null hypothesis of no partial relationship is probably wrong.

You might wonder here: why not use the coefficients on the dummy variables for heating-system type as test statistics? The reason is that there are two such coefficients (or in general, $K - 1$ coefficients for a categorical variable with K levels). But we need a single number to use as our test statistic in a permutation test. Therefore we use R^2: it is a single number that summarizes the predictive improvement of the full model over the reduced model.

Of course, even if we were to add a useless predictor to the reduced model, we would expect R^2 to go up, at least by a little bit, since the model would have more degrees of freedom (i.e. param-

Remember the four basic steps in a permutation test:

(1) Choose a null hypothesis H_0.

(2) Choose a test statistic Δ that is sensitive to departures from the null hypothesis.

(3) Repeatedly shuffle the predictor of interest and recalculate the test statistic after each shuffle, to approximate $P(\Delta \mid H_0)$, the sampling distribution of the test statistic T under the assumption that H_0 is true.

(4) Check whether the observed test statistic for your data, δ, is consistent with $P(\Delta \mid H_0)$.

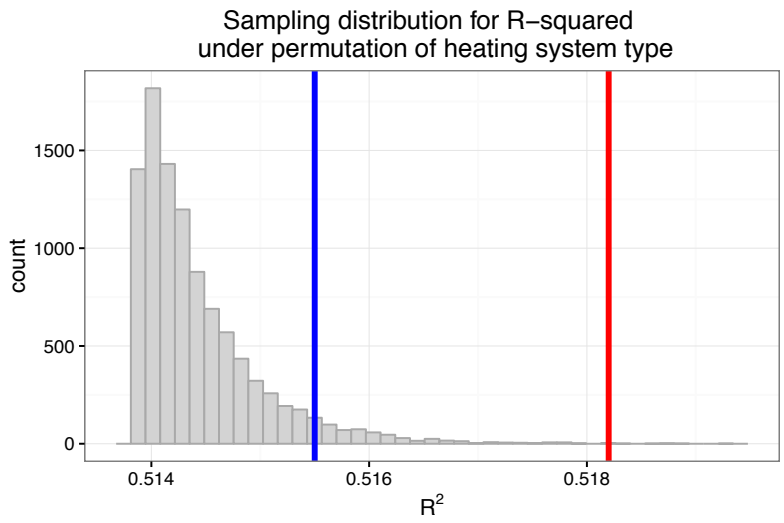


Figure 7.7: Sampling distribution of R^2 under the null hypothesis that there is no partial relationship between heating system and price after adjusting for effects due to square footage, lot size, and number of fireplaces. The blue vertical line marks the 95th percentile of the sampling distribution (and so corresponds to a rejection region at the 5% level). The red line marks the actual value of $R^2 = 0.518$ when we fit the full model by adding heating system to a model already containing the other three variables.

eters) that it can use to predict the observed outcome. Therefore, a more precise way of stating our null hypothesis is that, when we add heating system to a model already containing variables for square footage, lot size, and number of fireplaces, the improvement we see in R^2 could plausibly be explained by chance, even if this variable had no partial relationship with price.

To carry out a hypothesis test, we need to approximate the sampling distribution of R^2 under the null hypothesis. We will do so by repeatedly shuffling the heating system for every house (keeping all other variables the same), and re-fitting our model to each permuted data set. This breaks any partial relationship between heating system and price that may be present in our data. It tells us how big an improvement in R^2 we'd expect to see when fitting the full model, even if the null hypothesis were true.

This sampling distribution is shown in Figure 7.7, which was generating by fitting the model to 10,000 data sets in which the heating-system variable had been randomly shuffled, but where the response and the variables in the reduced model have been left alone. As expected, R^2 of the full model under permutation is always bigger than than the value of $R^2 = 0.513$ from the reduced model—but rarely by much. The blue line at $R^2 = 0.5155$ shows the 95th percentile of the sampling distribution (i.e. the critical value for a rejection region at the 5% level). The red line shows the actual value of $R^2 = 0.518$ from the full model fit the original

data set (i.e. with no shuffling). This test statistic falls far beyond the 5% rejection region. We therefore reject the null hypothesis and conclude that there is statistically significant evidence for an effect on price due to heating-system type.

One key point here is that we shuffled *only* heating-system type—or in general, whatever variable is being tested. We don't shuffle the response or any of the other variables. That's because we are interested in a partial relationship between heating-system type and price. Partial relationships are always defined with respect to a specific context of other control variables, and we have to leave these control variables as they are in order to provide the correct context for that partial relationship to be measured.

To summarize: we can compare any two nested models using a permutation test based on R^2, regardless of whether the variable in question is categorical or numerical. To do so, we repeatedly shuffle the extra variable in the full model—without shuffling either the response or the control variables (i.e. those that also appear in the reduced model). We fit the full model to each shuffled data set, and we track the sampling distribution of R^2. We then compare this distribution with the R^2 we get when fitting the full model to the *actual* data set. If the actual R^2 is a lot bigger than what we'd expect under the sampling distribution for R^2 that we get under the permutation test, then we conclude that the extra variable in the full model is statistically significant.

F tests and the normal linear regression model. Most statistical software will produce an ANOVA table with an associated p-value for all variables. These p-values are approximations to the p-values that you'd get if you ran sequential permutation tests, adding and testing one variable at a time as you construct the ANOVA table. To be a bit more specific, they correspond to something called an F test under the normal linear regression model that we met awhile back:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + e_i \,, \quad e_i \sim N(0, \sigma^2) \,.$$

You might want to revisit the discussion of the normal linear regression model starting on page 120. But the upshot is that an F test is conceptually similar to a permutation test based on R^2—and if you're happy with the assumption of normally distributed residuals, you can treat the p-values from these two tests as virtually interchangeable.[8]

[8] If you're not happy with this assumption, then you're better off with the permutation test.

8
Building predictive models

Building predictive models

Suppose you have a house in Saratoga, NY that you're about to put up for sale. It's a 1900 square-foot house on a 0.7-acre lot. It has 3 bedrooms, 2.5 bathrooms,[1] 1 fireplace, gas heating, and central air conditioning. The house was built 16 years ago. How much would you expect it to sell for?

[1] A half-bathroom has a toilet but no bath or shower.

Although we've been focusing on only a few variables of interest so far, our house-price data set actually has information on all these variables, and a few more besides. A great way to assess the value of the house is to use the available data to fit a multiple regression model for its price, given its features. We can then use this model to make a best guess for the price of a house with some particular combination of features—and, optionally, to form a prediction interval that quantifies the uncertainty of our guess.

We refer to this as the process of *building a predictive model.* Although we will still use multiple regression, the goal here is slightly different than in the previous examples. Here, we don't care so much about isolating and interpreting one particular partial relationship (like that between fireplaces and price). Instead, we just want the most accurate predictions possible.

The key principle in building predictive models is *Occam's razor*, which is the broader philosophical idea that models should be only as complex as they need to be in order to explain reality well. The principle is named after a medieval English theologian called Willam of Occam. Since he wrote in Latin, he put it like this: *Frustra fit per plura quod potest fieri per pauciora* ("It is futile to do with more things that which can be done with fewer.") A more modern formulation of Occam's razor might be the KISS rule: keep it simple, stupid.

In regression modeling, this principle is especially relevant for *variable selection*—that is, deciding which possible predictor variables to add to a model, and which to leave out. In this context,

Occam's razor is about finding the right set of variables to include so that we fit the data, without overfitting the data. Another way of saying this is that we want to find the patterns in the data, without memorizing the noise.

In this chapter, we'll consider two main questions:

(1) How can we measure the predictive power of a model?

(2) How can we find a model with good predictive power?

Measuring generalization error

To understand how we measure the predictive power of a regression model, we first need a bit of notation. Specifically, let's say that we have estimated a multiple regression model with p predictors $(x_1, x_2, \ldots, x_p)$ to some data, giving us coefficients $(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$. Now we encounter a new case, not in our original data set. We'll let $x^\star = (x_1^\star, x_2^\star, \ldots, x_p^\star)$ be the predictor variables for this new case, and $y^\star$ denote the corresponding response. We will use the fitted regression model, together with $x^\star$, to make a prediction for $y^\star$:

$$\hat{y}^\star = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_j^\star \,.$$

Our goal is to make the *generalization error*—that is, the difference between $y^\star$ and $\hat{y}^\star$—as small as possible, on average.

A natural way to measure the generalization error of a regression model is using a quantity called the *mean-squared predictive error*, or MSPE. The mean-squared predictive error is a property of a fitted model, not an individual data point. It summarizes the magnitude of the errors we typically make when we use the model to make predictions $\hat{y}^\star$ on new data:

$$\text{MSPE} = \text{Average value of } (y^\star - \hat{y}^\star)^2 \text{ when sampling new data points}\,.$$

Here a "new" data point means one that hasn't been used to fit the model. You'll notice that, in calculating MSPE, we square the prediction error $y^\star - \hat{y}^\star$ so that both positive and negative errors count equally.

Low mean-squared predictive error means that $y^\star - \hat{y}^\star$ tends to be close to zero when we sample new data points. This gives us a simple principle for building a predictive model: find the model (i.e. the set of variables to include) with the lowest mean-squared predictive error.

Estimating the the mean-squared predictive error

Conceptually, the simplest way to estimate the mean-squared predictive error of a regression model is to actually collect new data and calculate the average predictive error made by our model. Specifically, suppose that, after having fit our model in the first place, we go out there and collect $n^\star$ brand new data points, with responses $y_i^\star$ and predictors $(x_{i1}^\star, \ldots, x_{ip}^\star)$. We can then estimate the mean-squared predictive error of our model in two simple steps:

1. Form the prediction for each new data point:

$$\hat{y_i}^\star = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{ij}^\star \,.$$

2. Calculate the average squared error of your predictions:

$$\widehat{\text{MSPE}}_{\text{out}} = \frac{1}{n^\star} \sum_{i=1}^{n^\star} (y_i^\star - \hat{y}_i^\star)^2 \,.$$

   Notice that we put a hat on MSPE, because the expression on the right-hand side is merely an *estimate* of the true mean-squared predictive error, calculated using a specific sample of new data points. (Calculating the *true* MSPE would require us, in principle, to average over all possible samples of new data points, which is obviously impractical.) We also use the subscript "out" to indicate that it is an *out-of-sample* measure—that is, calculated on new data, that falls outside of our original sample.

Conventionally, we report the square root of $\widehat{\text{MSPE}}_{\text{out}}$ (which is called *root mean-squared predictive error*, or RMSPE), because this has the same units as the original y variable. You can think of the RMSPE as the standard deviation of future forecasting errors made by your model.

Assuming your new sample size $n^\star$ isn't too small, these two steps are a nearly foolproof way to estimate the mean-squared predictive error of your model. The drawback, however, is obvious: you need a brand new data set, above and beyond the original data set that you used to fit the model in the first place. This new data set might be expensive or impractical to collect.

Thus we're usually left in the position of needing to estimate the mean-squared predictive error of a model, without having access to a "new" data set. For this reason, the usual practice is

make a *train/test split* of your data: that is, to randomly split your original data set into two subsets, called the *training* and *testing* sets.

- The training set is used only to fit ("train") the model—that is, to estimate the coefficients $(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$.
- The testing set is used only to estimate the mean-squared predictive error of the model. It is not used at all to fit the model. For this reason, the testing set is sometimes referred to as the "hold-out set," since it is held out of the model-fitting process.

From this description, it should be clear that the training set plays the role of the "old" data, while the testing set plays the role of the "new" data.

This gives us a simple three-step procedure for choosing between several candidate models (i.e. different possible sets of variables to include).

(1) Split your data into training and testing sets.

(2) For each candidate model:

   A. Fit the model using the training set.
   B. Calculate $\widehat{\text{MSPE}}_{\text{out}}$ for that model using the testing set.

(3) Choose the model with the lowest value of $\widehat{\text{MSPE}}_{\text{out}}$.

Choosing the training and testing sets. A key principle here is that you must *randomly* split your data into a training set and testing set. Splitting your data nonrandomly—for example, taking the first 800 rows of your data as a training set, and the last 200 rows as a testing set—may mean that your training and testing sets are systematically different from one another. If this happens, your estimate of the mean-squared prediction error can be way off.

How much of the data should you reserve for the testing set? There are no hard-and-fast rules here. A common rule of thumb is to use about 75% of the data to train the model, and 25% to test it. Thus, for example, if you had 100 data points, you would randomly sample 75 of them to use for model training, and the remaining 25 to estimate the mean-squared predictive error. But other ratios (like 50% training, or 90% training) are common, too.

My general guideline is that the more data I have, the larger the fraction of that data I will use for training the predictive model.

Thus with only 100 data points, I might use a 75/25 split between training and testing; but with 10,000 data points, I might use more like a 90/10 split between training and testing. That's because estimating the model itself is generally harder than estimating the mean-squared predictive error.[2] Therefore, as more data accumulates, I like to preferentially allocate more of that data towards the intrinsically harder task of model estimation, rather than MSPE estimation.

[2] By "harder" here, I mean "subject to more sources of statistical error," as opposed to computationally more difficult.

Averaging over different test sets. It's a good idea to average your estimate of the mean-squared predictive error over several different train/test splits of the data set. This reduces the dependence of $\widehat{\text{MSPE}}_{\text{out}}$ on the particular random split into training and testing sets that you happened to choose. One simple way to do this is average your estimate of MSPE over many different random splits of the data set into training and testing sets. Somewhere between 5 and 100 splits is typical, depending on the computational resources available (more is better, to reduce Monte Carlo variability).

Another classic way to estimate MSPE it is to divide your data set into K non-overlapping chunks, called *folds*. You then average your estimate of MPSE over K different testing sets, one corresponding to each fold of the data. This technique is called *cross validation*. A typical choice of K is five, which gives us five-fold cross validation. So when testing on the first fold, you use folds 2-5 to train the model; when testing on fold 2, you use folds 1 and 3-5 to train the model; and so forth.

Can we use the original data to estimate the MSPE?

A reasonable question is: why do even we need a new data set to estimate the mean-squared prediction error? After all, our fitted model has residuals, $e_i = y_i - \hat{y}_i$, which tell us how much our model has "missed" each data point in our sample. Why can't we just use the residual variance, s_e^2, to estimate the MSPE? This approach sounds great on the surface, in that we'd expect the past errors to provide a good guide to the likely magnitude of future errors. Thus you might be tempted to use the *in-sample* estimate of MSPE, denoted

$$\widehat{\text{MSPE}}_{\text{in}} = s_e^2 = \frac{1}{n-p}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\,,$$

where we recall that p is the number of parameters in the model.

Using $\widehat{\text{MSPE}}_{\text{in}}$ certainly removes the need to collect a new data set. This turns out, however, to be a false economy: $\widehat{\text{MSPE}}_{\text{in}}$ is usually too optimistic as an estimate of a model's generalization error. Practically speaking, this means the following. When we use $\widehat{\text{MSPE}}_{\text{in}}$ to quantify the *in-sample* error of a model, and then we actually go out and take new data to calculate the *out-of-sample* generalization error $\widehat{\text{MSPE}}_{\text{out}}$, we tend to discover that the out-of-sample error is larger—sometimes much larger! This is called overfitting, and it is especially likely to happen when the size of the data set is small, or when the model we're fitting is very complex (i.e. has lots of parameters).

An example

Let's see these ideas in practice, by comparing three predictive models for house prices in Saratoga, New York. Our models will draw from the following set of variables:

- lot size, in acres
- age of house, in years
- living area of house, in square feet
- percentage of residents in neighborhood with college degree
- number of bedrooms
- number of bathrooms
- number of total rooms
- number of fireplaces
- heating system type (hot air, hot water, electric)
- fuel system type (gas, fuel oil, electric)
- central air conditioning (yes or no)

We'll consider three possible models for price constructed from these 11 predictors.

Small model: price versus lot size, bedrooms, and bathrooms (4 total parameters, including the intercept).

Medium model: price versus all variables above, main effects only (14 total parameters, including the dummy variables).

Big model: price versus all variables listed above, together with all pairwise interactions between these variables (90 total parameters, include dummy variables and interactions).

Table 8.1 shows both $\widehat{\text{MSPE}}_{\text{in}}$ and $\widehat{\text{MSPE}}_{\text{out}}$ for these three models. To calculate $\widehat{\text{MSPE}}_{\text{out}}$, we used 80% of the data as a training

	In-sample RMSPE	Out-of-sample RMSPE	Difference
Small model: underfit	$76,144	$76,229	$85
Medium model: good fit	$65,315	$65,719	$403
Big model: overfit	$61,817	$71,426	$9,609

Table 8.1: In-sample versus out-of-sample estimates of the root mean-squared predictive error for three models of house prices in Saratoga, NY. The "difference" column shows the difference between the in-sample and out-of-sample estimates. The big model has a very large difference (over $9,000), indicating that the in-sample estimate is way too optimistic, and that the model is probably overfit to the data.

set, and the remaining 20% as a test set, and we averaged over 100 different random train/test splits of the data. The final column, labeled "difference," shows the difference between the in-sample and out-of-sample estimates of prediction error.

There are a few observations to take away from Table 8.1. The first is that the big model (with all the main effects and interactions) has the lowest in-sample error. With a residual standard deviation of $61,817, it seems nearly $3,500 more accurate than the medium model, which is next best. This is a special case of a very general phenomenon: a more complex model will always fit the data better, because it has more degrees of freedom to play with.

However, the *out-of-sample* measure of predictive error tells a different story. Here, the medium-sized model is clearly the winner. Its predictions on new data are off by about $65,719, on average, which is nearly $6,000 better than the big model.

Finally, notice how severely degraded the predictions of the big model become when moving from old (in-sample) data to new (out-of-sample) data: about $9,600 worse, on average. This kind of degradation is a telltale sign of overfitting. The medium model suffers only a mild degradation in performance on new data, while the small model suffers hardly any degradation at all—although it's still not competitive on the out-of-sample measure, because it wasn't that good to begin with. This is also a special case of a more general phenomenon: *some* degradation in predictive performance on out-of-sample versus in-sample data is inevitable, but simpler models tend to degrade a lot less.

Figure 8.1 demonstrates this point visually. Starting from a very simple model of price (using only lot size as a predictor), we've added one variable or interaction at a time[3] from the list on page 168. For each new variable or interaction, we recalculated both the in-sample ($\widehat{\text{MSPE}}_{\text{in}}$) and out-of-sample ($\widehat{\text{MSPE}}_{\text{out}}$) estimates of the generalization error. As we add variables, the out-of-sample error initially gets smaller, reflecting a better fitting model that still generalizes well to new data. But after 15 or 20 variables,

[3] To be specific here, at each stage we added the single variable or interaction that most improved the fit of the model. See the next section on stepwise selection.

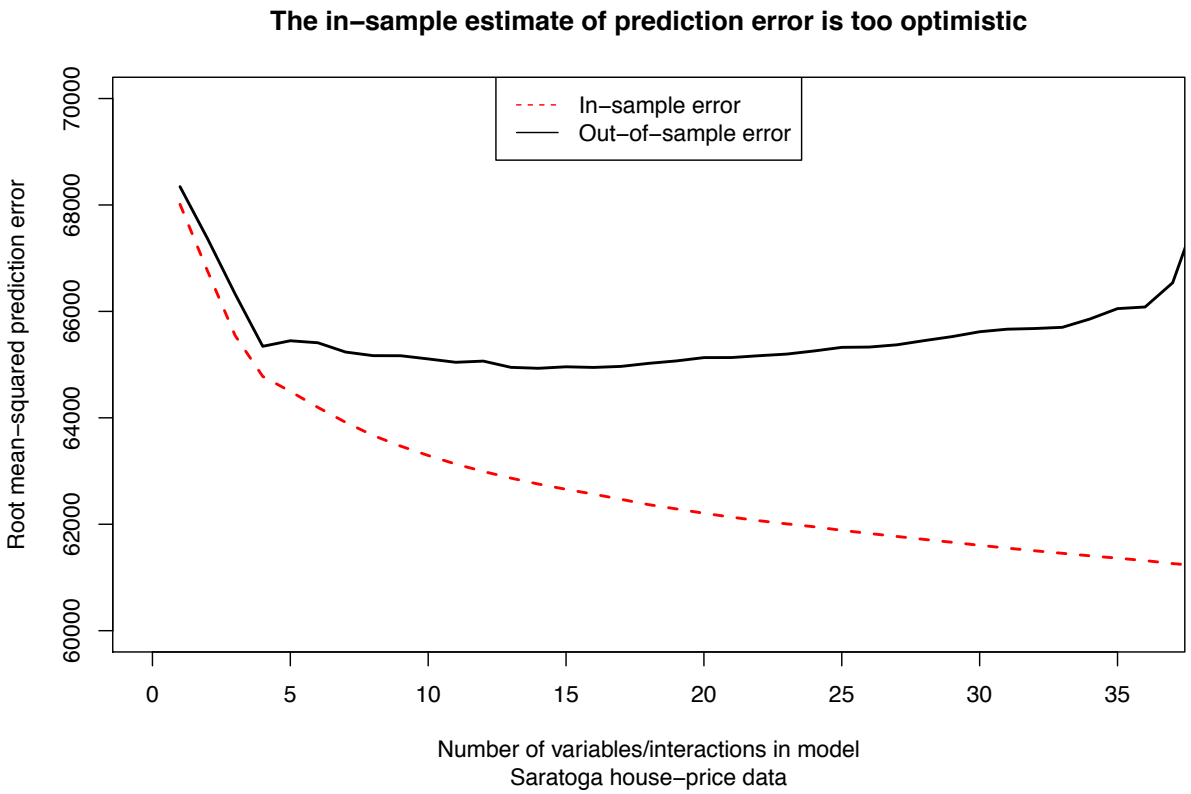


Figure 8.1: Starting from a small pricing model with just lot size as a predictor, we've added one variable or interaction at a time from the list on page 168. The red line shows the in-sample estimate of error, while the black line shows the out-of-sample estimate. After we add about 15 variables and interactions, the out-of-sample error starts to creep back up. Clearly the in-sample estimate is too optimistic, especially as the model gets more complex.

eventually the out-of-sample error starts creeping back up, due to overfitting. The in-sample estimate of error, however, keeps going down, falling even further out of line with the real out-of-sample error as we add more variables to the model.

In summary, you should remember the basic mantra of predictive model building: out-of-sample error is larger than in-sample error, especially for bigger models. If you care about minimizing out-of-sample error, you should always use an out-of-sample estimate of a model's MSPE, to make sure that you're not overfitting the original data. Our goal here should be obvious: to find the "turning point" in Figure 8.1, and to stop adding variables before we start overfitting.

Iterative model building via stepwise selection

Now that we know how to measure generalization error of a model, we're ready to introduce the overall steps in the process of building and using a predictive model from a set of candidate variables x_1, x_2, etc. We sometimes use the term *scope* to refer to this set of candidate variables.

The seemingly obvious approach is to fit all possible models under consideration to a training set, and to measure the generalization error of each one on a testing set. If you have only a few variables, this will work fine. For example, with only 2 variables, there are only $2^2 = 4$ possible models to consider: the first variable in, the second variable in, both variables in, or both variables out. You can fit and test those four models in no time. This is called *exhaustive enumeration*.

However, if there are lots of variables, exhaustive enumeration of all the models becomes a lot harder to do, for the simple fact that it's too exhausting—there are too many models to consider. For example, suppose we have 10 possible variables, each of which we could put in or leave out of the model. Then there are $2^{10} = 1024$ possible models to consider, since each variable could be in or out in any combination. That's painful enough. But if there are 100 possible variables, there are 2^{100} possible models to consider. That's *1 nonillion* models—about 10^{30}, or a thousand billion billion billion. This number is larger than the number of atoms in a human body.

You will quite obviously never be able to fit all these countless billions of models, much less compare their generalization errors on a testing set, even with the most powerful computer on earth. Moreover, that's for just 100 candidate variables *with main effects only*. Ideally, we'd like the capacity to build a model using many more candidate variables than that, or to include the possibility of interactions among the variables.

Thus a more practical approach to model-building is *iterative*: that is, to start somewhere reasonable, and to make small changes to the model, one variable at a time. Model-building in this iterative way is really a three-step process:

(1) Choose a baseline model, consisting of initial set of predictor variables to include in the model, including appropriate transformations, polynomial terms and interactions. Exploratory

data analysis (i.e. plotting your data) will generally help you get started here, in that it will reveal obvious relationships in the data. Then fit the model for y versus these initial predictors.

(2) Check the model. If necessary, change what variables are included, what transformations are used, etc.:

   (a) Are the assumptions of the model met? This is generally addressed using residual plots, of the kind shown in Figures 6.7 and 6.8. This allows you to assess whether the response varies linearly with the predictors, whether there are any drastic outliers, etc.

   (b) Are we missing any important variables or interactions? This is generally addressed by *adding* candidate variables or interactions to the model from step (1), to see how much each one improves the generalization error (MSPE).

   (c) Are there signs that the model might be overfitting the data? This is generally addressed by *deleting* variables or interactions that are already in the model, to see if doing so actually improves the model's generalization error.

   You may need to iterate these three questions a few times, going through many rounds of adding or deleting variables, before you're satisfied with your final model. Remember that the best way to measure generalization error is using an out-of-sample measure, like $\widehat{\text{MSPE}}_{\text{out}}$ derived from a train/test split of the data.

   Once you're happy with the model itself, then you can. . . .

(3) Use your fitted model to form predictions (and optionally, prediction intervals) for your new data points.

Can this process be automated?

In this three-step process, step 1 (start somewhere reasonable) and step 3 (use the final model) are usually pretty easy. The part where you'll spend the vast majority of your time and effort is step 2, when you consider many different possible variables to add or delete to the current model, and check how much they improve or degrade the generalization error of that model.

This is a lot easier than considering all possible combinations of variables in or out. But with lots of candidate variables, even this

iterative process can get super tedious. A natural question is, can it be automated?

The answer is: sort of. We can easily write a computer program that will automatically check for iterative improvements to some baseline ("working") model, using an algorithm called *stepwise selection*:

(1) From among a candidate set of variables (the scope), check all possible one-variable additions or deletions from the working model;

(2) Choose the single addition or deletion that yields the best improvement to the model's generalization error. This becomes the new "working model."

(3) Iteratively repeat steps (1) and (2) until no further improvement to the model is possible.

The algorithm terminates when it cannot find any one-variable additions or deletions that will improve the generalization error of the working model.

Some caveats. Stepwise selection tends to work tolerably well in practice. But it's far from perfect, and there are some important caveats. Here are three; the first one is minor, but the second two are pretty major.

First, if you run stepwise selection from two different baseline models, you will probably end up with two different final models. This tends not to be a huge deal in practice, however, because the two final models usually have similar mean-squared predictive errors. Remember, when we're using stepwise selection, we don't care too much about *which* combinations of variables we pick, as long as we get good generalization error. Especially if the predictors are correlated with each other, one set of variables might be just as good as another set of similar (correlated) variables.

Second, stepwise selection usually involves some approximation. Specifically, at each step of stepwise selection, we have to compare the generalization errors of many possible models. Most statistical software will perform this comparison *not* by actually calculating $\widehat{\text{MSPE}}_{\text{out}}$ on some test data, but rather using one of several possible heuristic approximations for MSPE. The most common one is called the AIC approximation:[4]

$$\widehat{\text{MSPE}}_{\text{AIC}} = \widehat{\text{MSPE}}_{\text{in}} \left(1 + \frac{p}{n}\right) = s_e^2 \left(1 + \frac{p}{n}\right) ,$$

[4] In case you're curious, AIC stands for "Akaike information criterion." If you find yourself reading about AIC on Wikipedia or somewhere similar, it will look absolutely nothing like the equation I've written here. The connection is via a related idea called "Mallows' C_p statistic," which you can read about here.

where n is the sample size and p is the number of parameters in the model.

The AIC estimate of mean-squared predictive error is not a true out-of-sample estimate, like $\widehat{\text{MSPE}}_{\text{out}}$. Rather, it is like an "inflated" or "penalized" version of the in-sample estimate, $\widehat{\text{MSPE}}_{\text{in}} = s_e^2$, which we know is too optimistic. The inflation factor of $(1 + p/n)$ is always larger than 1, and so $\widehat{\text{MSPE}}_{\text{AIC}}$ is always larger than $\widehat{\text{MSPE}}_{\text{in}}$. But the more parameters p you have relative to data points n, the larger the inflation factor gets. It's important to emphasize that $\widehat{\text{MSPE}}_{\text{AIC}}$ is just an approximation to $\widehat{\text{MSPE}}_{\text{out}}$. It's a better approximation than $\widehat{\text{MSPE}}_{\text{in}}$, but it still relies upon some pretty specific mathematical assumptions that can easily be wrong in practice.

The third and most important caveat is that, when using any kind of automatic variable-selection procedure like stepwise selection, we lose the ability to use our eyes and our brains each step of the way. We can't plot the residuals to check for outliers or violations of the model assumptions, and we can't ensure that the combination of variables visited by the algorithm make any sense, substantively speaking. It's worth keeping in mind that your eyes, your brain, and your computer are your three most powerful tools for statistical reasoning. In stepwise selection, you're taking two of these tools out of the process, for the sake of doing a lot of brute-force calculations very quickly.

None of these caveats are meant to imply that you *shouldn't* use stepwise selection—merely that you shouldn't view the algorithm as having God-like powers for discerning the single best model, or treat it as an excuse to be careless. You should instead proceed cautiously. Always verify that the stepwise-selected model makes sense and doesn't violate any crucial assumptions. It's also a good idea to perform a quick train/test split of your data and compute $\widehat{\text{MSPE}}_{\text{out}}$ for your final model, just as a sanity check, to make sure that you're actually improving the generalization error versus your baseline model.

9
Understanding cause and effect

Statistical questions versus causal questions

WHY have some nations become rich while others have remained poor? Do small class sizes improve student achievement? Does following a Mediterranean diet rich in vegetables and olive oil reduce your risk of a heart attack? Does a "green" certification (like LEED, for Leadership in Energy and Environmental Design) improve the value of a commercial property?

Questions of cause and effect like these are, fundamentally, questions about *counterfactual statements.* A counterfactual is an if–then statement about something that has not actually occurred. For example: "If Colt McCoy had not been injured early in the 2010 National Championship football game, then the Texas Longhorns would have beaten Alabama." If you judge this counterfactual statement to be true—and who but the most hopelessly blinkered Crimson Tide fan doesn't?—then you might say that Colt McCoy's injury caused the Longhorns' defeat.

Statistical questions, on the other hand, are about correlations. This makes them fundamentally different from causal questions.

- Causal: "If we invested more money in our school system, how much faster would our economy grow?" Statistical: "In looking at data on a lot of countries, how are education spending and economic growth related?"

- Causal: "If I ate more vegetables than I do now, how much longer would I live?" Statistical: "Do people who eat a lot of vegetables live longer, on average, than people who don't?"

- Causal: "If we hire extra teachers at our school and reduce our class sizes, will our students' test scores improve?" Statistical: "Do students in smaller classes tend to have higher test scores?"

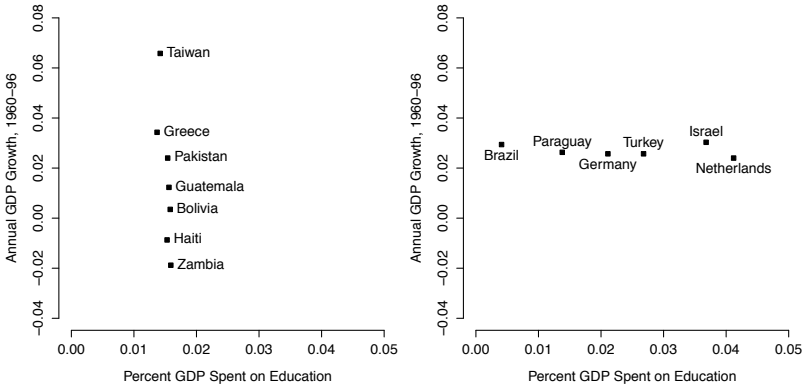


Figure 9.1: Two egregious examples of selective reporting.

Causal questions all invoke some kind of hypothetical intervention, where one thing is changed and everything else is held equal. In such a hypothetical intervention, there is no competing explanation for what might be causing the change we expect to see—in our economy, our lifespan, our students' test scores, a football game, or whatever outcome we're interested in.

Statistical questions, on the other hand, are about the patterns we observe in the real world. And the real world is rarely so simple as the hypothetical interventions we imagine. For example, people who eat more vegetables live longer—that's a clear pattern. But those same people also tend to exercise more, live in better housing, and have higher-status jobs. These other factors are *confounders*. A confounder is a competing explanation—some other factor correlated with both the "treatment" assignment (whether someone eats vegetables) and the response (lifespan). So in light of these confounders, how do we know it's the vegetables, rather than all that other stuff, that's making veggie-eaters live longer?

This is just a specific version of the general question we'll address in this chapter: under what circumstances can causal questions be answered using statistics?

Good evidence . . . and bad

Most of the cause-and-effect reasoning that you'll see out there in the real world is of depressingly poor quality. A common flaw is *cherry picking*: that is, pointing to data that seems to confirm some argument, while ignoring contradictory data.

Here's an example. In the left panel in Figure 9.1 we see a

group of seven countries that all spend around 1.5% of their GDP on education, but with very different rates of economic growth for the 37 years spanning 1960 to 1996. In the right panel, we see another group of six countries with very different levels of spending on education, but similar growth rates of 2–3%.

Both highly selective samples make it seem as though education and economic growth are barely related. If presented with the left panel alone, you'd be apt to conclude that the differences in growth rates must have been caused by something other than differences in education spending (of which there are none). Likewise, if presented with the right panel alone, you'd be apt to conclude that the large observed differences in education spending don't seem to have produced any difference in growth rates. The problem here isn't with the data—it's with the biased, highly selective *use* of that data.

This point seems almost obvious. Yet how tempting it is just to cherry pick and ignore the messy reality. Perhaps without even realizing it, we're all accustomed to seeing news stories that marshal highly selective evidence—usually even worse than that of Figure 9.1—on behalf of some plausible because-I-said-so story:

> [H]igher levels of education are critical to economic growth. . . . Boston, where there is a high proportion of college graduates, is the perfect example. Well-educated people can react more quickly to technological changes and learn new skills more readily. Even without the climate advantages of a city like San Jose, California, Boston evolved into what we now think of as an "information city." By comparison, Detroit, with lower levels of education, languished.[1]

[1] "Economic Scene." *New York Times* (Business section); August 5, 2004

And this from a reporter who presumably has no hidden agenda. Notice how the selective reporting of evidence—one causal hypothesis, two data points—lends an air of such graceful inevitability to what is a startlingly superficial analysis of the diverging economic fates of Boston and Detroit over the last half century.

Of course, most bad arguments are harder to detect than this howler from the New York Times. After all, using data to understand cause-and-effect relationships is hard. For example, consider the following summary of a recent neuroscience study:

> A study presented at the Society for Neuroscience meeting, in San Diego last week, shows people who start using marijuana at a young age have more cognitive shortfalls. Also, the more marijuana a person used in adolescence, the more trouble they had with focus and attention. "Early onset smokers

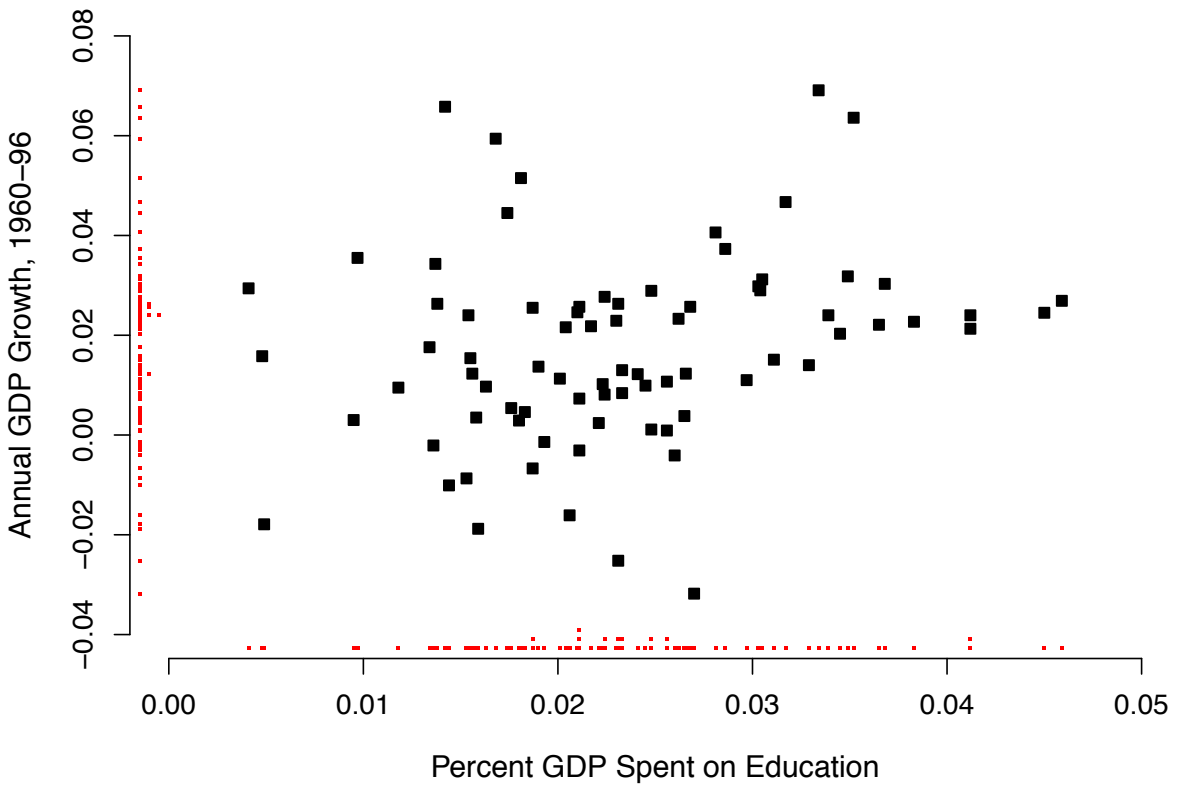


Figure 9.2: A scatter plot of GDP growth versus education spending for 79 countries. The tiny red dots clustered near the x and y axes are called *rug plots*. They are miniature histograms aligned with the axes of the predictor and the response.

have a different pattern of brain activity, plus got far fewer correct answers in a row and made way more errors on certain cognitive tests," says study author Staci Gruber.[2]

Did the marijuana smokers get less smart, or were the less-smart kids more likely to pick up a marijuana habit in the first place? It's an important question to consider in making drug policy, especially for states and countries where marijuana is legal. But can we know the answer on the basis of a study like this?

For another example, consider the bigger sample of countries in Figure 9.2, which provides a much more representative body of evidence on the GDP-versus-education story. This evidence takes the form of a scatter plot of GDP growth versus education spending for a sample of 79 countries worldwide. Notice the following two facts:

(1) Of the 29 countries that spent less than 2% of GDP on education, 18 fall below the median growth rate (1.58%).

[2] www.usatoday.com/yourlife/health/medical/pediatrics/2010-11-20-teendrugs22_ST_N.htm

(2) Of the 18 countries that spent more than 3% of GDP on education, 16 fall above the median growth rate.

These two facts, together with the upward trend in the scatter plot, suggest that economic growth and education spending are correlated. But this does not settle the causal question. For example, it might be that countries spend a lot on education because they are rich, rather the other way around.

The generic difficulty is that there are many different ways that two variables X and Y can appear correlated.

(1) One-way causality: the first domino falls, then the second; the rain falls, and the grass gets wet. (X causes Y directly.)

(2) Two-way causality: flowers and honey bees prosper together. (Both X and Y play a role in causing each other.)

(3) Common cause: People who go to college tend to get higher-paying jobs than those who don't. Does education directly lead to better economic outcomes? Or are a good education and a good job both just markers of a person's underlying qualities? (The role of X in causing Y is hard to distinguish from the role of C, which we may not have observed.)

(4) Common effect: either musical talent (X) or athletic talent (Y) will help you get into Harvard (Z). Among a population of Harvard freshmen, musical and athletic talent will thus appear negatively correlated, even if they are independent in the wider population. (X and Y both contribute to some common outcome C, inducing a correlation among a subset of the population defined by Z. This is often called Berkson's paradox; it is subtle, and we'll encounter it again.)

(5) Luck: the observed correlation is a coincidence.

This is the point where most books remind you that "correlation does not imply causation." Obviously. But if not to illuminate causes, what is the point of looking for correlations? Of course correlation does not imply causality, or else playing professional basketball would make you tall. But that hasn't stopped humans from learning that smoking causes cancer, or that lightning causes thunder, on the basis of observed correlations. The important question is: what distinguishes the good evidence-based arguments from the bad?

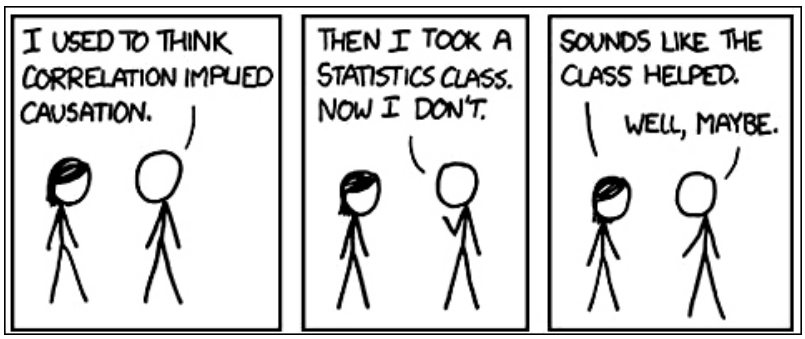


Figure 9.3: Originally published online at xkcd: http://xkcd.com/552/

Four common identification strategies

The key principle in using evidence to draw causal conclusions is that of a *balanced comparison*. To make things simple, we'll imagine that our predictor X is binary (i.e. has two groups), and we'll borrow the lingo of a clinical trial by referring to the two groups as the "treatment" and "control." To reach the conclusion that X causes Y, you must do two things: (1) *compare cases* in the treatment and control groups, to see how their Y values differ; and (2) *ensure balance*, by removing all other systematic differences between the cases in the treatment and control groups. Balance is crucial; it's what allows us to conclude that the differences in X (and not something else) cause the differences we observe in Y.

In general, there are four common ways to make a balanced comparison. These are often called *identification strategies*, in the sense that they are strategies for identifying a causal effect.

(1) Run a real experiment, randomizing subjects to the treatment and control groups. The randomization will ensure that, on average, there are no systematic differences between the two groups, other than the treatment.

(2) Find a natural experiment: that is, find a situation where the way that cases fall naturally into the treatment and control groups plausibly resembles a random assignment.

(3) Matching: artificially construct a balanced data set from an unbalanced data set, by explicitly matching treated cases with similar control cases, and discarding the cases without a good match. This will correct for lack of balance between control and treatment groups.

(4) Modeling: use multiple regression modeling to adjust for confounders and isolate a partial relationship between the response and the treatment of interest.

We'll take each of these four ideas in turn.

The power of experiment

THE idea of an experiment is simple. If you want to know what would happen if you intervened in some system, then you should intervene, and measure what happens. There is simply no better way to establish that one thing causes another.

Indeed, one kind of experiment—the randomized, controlled clinical trial—is one of the most important medical innovations in history. Suppose we want to establish whether a brand new cholesterol drug—we'll call it Zapaclot—works better than the old drug. Also suppose that we've successfully recruited a large cohort of patients with high cholesterol. We know that diet and genes play a role here, but that drugs can help, too. We express this as

$$\text{Cholesterol} \sim \text{Diet} + \text{Genes} + \text{Drugs}\,.$$

Interpret the plus sign as the word "and," not like formal addition: we're assuming that cholesterol depends upon diet, genes, and drugs, although we haven't said how. Of course, it's that third predictor in the model we care about; the first two, in addition to some others that we haven't listed, are potential confounders.

First, what not to do: don't proceed by giving Zapaclot to all the men and the old drug to all the women, or Zapaclot to all the marathon runners and the old drug to the couch potatoes. These highly non-random assignments would obviously bias any judgment about the relative effect of the new drug compared to the old one. We refer to this sort of thing as *selection bias*: that is, any bias in the selection of cases that receive the treatment. Moreover, you shouldn't just give the new drug to whomever wants it, or can afford it. The people with more engagement, more knowledge, more money, or more trust in the medical system would probably sign up in greater numbers—and if those people have systematic differences in diet or genes from the people who don't sign up, then you've just created a hidden selection bias.

Instead, you should two simple steps.

Randomize: randomly split the cohort into two groups, denoted the treatment group and the control group.

Intervene: allocate everyone in the treatment group to take the treatment (e.g. Zapaclot, the new drug), and everyone in the control group to take something else (e.g. the old drug or a placebo).[3]

[3] Everyone in the control group should be taking the *same* something else, whether it's the old drug or a placebo.

Randomize and intervene: a simple prescription, but the surest way to establish causality. The intervention allows you to pick up a difference between the new and old drug, if there's one to be found. The randomization ensures that other factors—even unknown factors, in addition to known ones like diet and lifestyle—do not lead us astray in our causal reasoning. The Latin phrase *ceteris paribus*, which translates roughly as "everything else being equal," is often used to describe such a situation. By randomizing and intervening, we have ensured that the only *systematic* difference between the groups is the treatment itself. The randomization gives us a balanced comparison.

This last point is crucial. It's not that diet, genes, and other lifestyle factors somehow stop affecting a patient's cholesterol level when we randomize and intervene. It's just that diet, genes, and lifestyle factors aren't correlated with the treatment assignment, and so they're balanced between the two groups, on average.

The need to avoid selection bias sounds obvious. But if selection bias in medical trials were not rigorously policed, then it would be easy for doctors to cherry pick healthy patients for newly proposed treatments. After all, a doctor who invents a new, seemingly effective form of treatment will almost surely become both rich and famous. As one physician reminisces:

> One day when I was a junior medical student, a very important Boston surgeon visited the school and delivered a great treatise on a large number of patients who had undergone successful operations for vascular reconstruction. At the end of the lecture, a young student at the back of the room timidly asked, "Do you have any controls?" Well, the great surgeon drew himself up to his full height, hit the desk, and said, "Do you mean did I not operate on half of the patients?" The hall grew very quiet then. The voice at the back of the room very hesitantly replied, "Yes, that's what I had in mind." Then the visitor's fist really came down as he thundered, "Of course not. That would have doomed half of them to their death." God, it was quiet then, and one could scarcely hear the small voice ask, "Which half?"[4]

[4] Dr. E. Peacock, University of Arizona. Originally quoted in *Medical World News* (September 1, 1972). Reprinted pg. 144 of *Beautiful Evidence*, Edward Tufte (Graphics Press, 2006).

These last two words—"Which half?"—should echo in your mind whenever you are asked to judge the quality of evidence offered in support of a causal hypothesis. There is simply no substitute for a controlled experiment: not a booming authoritative voice, not even fancy statistics.

In fact, government regulators are so fastidious in their attention to possible selection biases that, in most real clinical trials, neither the doctors nor the patients are allowed to know which drug each person receives. Such a "double-blind" experiment avoids the possibility that patients might simply imagine that the the latest miracle drug has made them feel better, in a feat of unconscious self-deception called the placebo effect.

A placebo, from the Latin *placere* ("to please"), is a fake treatment designed to simulate the real one.

Some history

The notion of a controlled experiment was certainly around in pre-Christian times. The first chapter of the book of Daniel relates the tale of one such experiment. Daniel and his three friends Hananiah, Mishael, and Azariah arrive in the court of Nebuchadnezzar, the King of Babylon. They enroll in a Babylonian school, and are offered a traditional Babylonian diet. But Daniel wishes not to "defile himself with the portion of the king's meat, nor with the wine which he drank." He goes to Melzar, the prince of the eunuchs, who is in charge of the school. Daniel asks not to be made to eat the meat or drink the wine. But Melzar responds that he fears for Daniel's health if he were to let them follow some crank new-age diet. More to the point, Melzar observes, if the new students were to fall ill, "then shall ye make me endanger my head to the king."

So Daniel proposes a trial straight out of a statistics textbook:

> Prove thy servants, I beseech thee, ten days; and let them give us pulse to eat, and water to drink.
> Then let our countenances be looked upon before thee, and the countenance of the children that eat of the portion of the king's meat: and as thou seest, deal with thy servants.[5]

[5] King James Bible, Daniel 1:12–13.

The King agreed. When Daniel and his friends were inspected ten days later, "their countenances appeared fairer and fatter in flesh" than all those who had eaten meat and drank wine. Suitably impressed, Nebuchadnezzar brings Daniel and his friends in for an audience, and he finds that "in all matters of wisdom and understanding," they were "ten times better than all the magicians and astrologers that were in all his realm."

As for a placebo-controlled trial, in which some of the patients are intentionally given a useless treatment (the "placebo"): that came much later.[6] The first such trial seems to have taken place in 1784. It was directed by none other than Benjamin Franklin, the American ambassador to the court of King Louis XVI of France. A German doctor by the name of Franz Mesmer had gained some degree of notoriety in Europe for his claim to have discovered a new force of nature that he called "magnétisme animal," and which was said to have magical healing powers. The demand for Dr. Mesmer's services soon took off among the ladies of Parisian high society, whom he would "Mesmerize" using a wild contraption involving ropes and magnetized iron rods.

[6] See "The Power of Nothing" in the December 12, 2011 edition of *The New Yorker* (pp. 30–6).

Much to the king's dismay, his own wife, Marie Antoinette, was one of Mesmer's keenest followers. The king found the whole Mesmerizing thing frankly a bit dubious, and presumably wished for his wife to have nothing to do with the Herr Doctor's magnétisme animal. So he convened several members of the French Academy of Sciences to investigate whether Dr. Mesmer had indeed discovered a new force of nature. The panel included Antoine Lavoisier, the father of modern chemistry, along with Joseph Guillotin, whose own wild contraption was soon to put the King's difficulties with Mesmer into perspective. Under Ben Franklin's supervision, the scientists set up an experiment to replicate some of Dr. Mesmer's prescribed treatments, substituting non-magnetic materials—history's first placebo—for half of the patients. In many cases, even the patients in the control group would flail about and start talking in tongues anyway. The panel concluded that the doctor's method produced no effect other than in the patients' own minds. Mesmer was denounced as a charlatan, although he continues to exact his revenge via the dictionary.

A more recent and especially striking example of a placebo comes from Thomas Freeman, director of the neural reconstruction unit at Tampa General Hospital in Florida. Dr. Freeman performs placebo brain surgery. (You read that correctly.) According to the British Medical Journal,

> In the placebo surgery that he performs, Dr Freeman bores into a patient's skull, but does not implant any of the fetal nerve cells being studied as a treatment for Parkinson's disease. The theory is that such cells can regenerate brain cells in patients with the disease. Some colleagues decry the experimental method, however, saying that it is too risky and unethical, even though patients are told before the operation

that they may or may not receive the actual treatment.[7]

[7] BMJ. 1999 October 9; 319(7215): 942

"There has been a virtual taboo of putting a patient through an imitation surgery," Dr. Freeman said. (Imagine that.) "This is the way to start the discussion." Freeman has performed 106 real and placebo cell transplant operations since 1992. Dr. Freeman argues that the medical history is littered with examples of unsafe and ineffective surgical procedures—think of that small voice at the back of the room, asking "which half?"—that were not tested against a placebo and resulted in needless deaths, year after year, before doctors abandoned them.

Experimental evidence is the best kind of evidence

Let's practice here, by comparing two causal hypotheses arising from two different data sets. The first comes from a clinical trial in the 1980's on a then-new form of adjuvant chemotherapy for treating colorectal cancer, a dreadful disease that, as of 2015, has a five-year survival rate of only 60-70% in the developed world.

The trial followed a simple protocol. After surgical removal of their tumors, patients were randomly assigned to different treatment regimes. Some patients were treated with fluorouracil (the chemotherapy drug, also called 5-FU), while others received no follow-up therapy. The researchers followed the patients for many years afterwards and tracked which ones suffered from a recurrence of colorectal cancer.

The outcome of the trial are in Table 9.1, below. Among the patients who received chemotherapy, 39% (119/304) had relapsed by the end of the study period, compared with 57% of patients (177/315) in the group who received no therapy:

Chemotherapy?		Yes	No
Recurrence?	Yes	119	177
	No	185	138

Table 9.1: Data from: J. A. Laurie et. al. Surgical adjuvant therapy of large-bowel carcinoma: An evaluation of levamisole and the combination of levamisole and fluorouracil. J. Clinical Oncology, 7:1447–56, 1989. There was also a third treatment arm of the study in which patient received a drug called levamisole, which isn't discussed here. Survival statistics on colorectal cancer from Cunningham et. al (2010). "Colorectal cancer." Lancet 375 (9719): 1030–47.

The evidence strongly suggests that the chemotherapy reduced the risk of recurrence by a substantial amount: the relative risk of a relapse under the treatment group is 0.7, with a 95% confidence interval of $(0.59, 0.83)$.

We can be confident that this evidence reflects causality, and not merely correlation, because patients were randomly assigned

to the treatment and control groups. Randomization ensures *balance*: that is, it ensures that there are no systematic differences between the two groups with respect to any confounding factors that might be correlated with the patients' survival chances. This would obviously not be true if we had non-randomly assigned all the healthiest patients to the treatment group, and all the sickest patients to the control group.

It's worth emphasizing a key fact here. Randomization ensures balance both for the possible confounders that we can measure (like a patient's age or baseline health status), as well as for the ones we might *not* be able to measure (like a patient's will to live). This is what makes randomization so powerful, and randomized experiments so compelling. We don't even have to know what the possible confounding variables are in order for the experiment to give us reliable information about the causal effect of the treatment. *Randomization balances everything*, at least on average.

Next, let's examine data from a study from the 1990's conducted in sub-Saharan Africa about HIV, another dreadful disease which, at the time, was spreading across the continent with alarming speed. Several studies in Kenya had found that men who were uncircumcised seemed to contract HIV in greater numbers. This set off a debate among medical experts about the extent to which this apparent association had a plausible biological explanation.

Circumcised?		Yes	No
HIV positive?	Yes	105	85
	No	527	93

Table 9.2: Data from Tyndall et. al. Increased risk of infection with human immunodeficiency virus type 1 among uncircumcised men presenting with genital ulcer disease in Kenya. Clin. Infect. Dis. 1996 Sep; 23(3):449–53.

Table 9.2, above, shows some data from one of these studies, which found that among those recruited for the survey, 48% of uncircumcised men were HIV-positive, versus only 17% of circumcised men. The evidence seems to suggest that circumcision reduced a Kenyan man's chance of contracting HIV by a factor of 3.

Evaluating the evidence. If you suffer from colon cancer, should you get chemotherapy? Almost certainly: the researchers in the first study randomized and intervened, giving chemotherapy only to a random subset of patients. Unless you believe that the chemotherapy patients in this trial just happened to be much luckier than their peers, this result establishes that the reduction in recurrence must have been caused by the treatment.

But should all Kenyan men head straight to a surgeon? In this case we can't really be sure. The researchers in the second study neither randomized nor performed any snipping themselves. They merely asked whether each man was circumcised. It is therefore possible that they've been fooled by a confounder. To give one plausible example, a man's religious affiliation might affect both the likelihood that he is circumcised and the chances that he contracts HIV from unprotected sex. If that were true, the observed correlation between circumcisions and HIV rates might be simply a byproduct of an imbalanced, unfair comparison, rather than a causal relationship.[8]

[8] The authors of the study were obviously aware of these possible confounders. They used a technique called logistic regression to attempt to account for some them and isolate the putative effect of circumcision on HIV infection. This is like our fourth method for making balanced comparisons: use a model to adjust for confounders statistically. See the original paper for details.

Natural experiments

A randomized, controlled experiment is the gold standard of evidence for a causal hypothesis. Yet many times an experiment is impossible, impractical, unethical, or too expensive in time or money. In these situations, it often pays to look for something called a *natural experiment*, also called a *quasi-experiment*. A natural experiment is not something that you, as the investigator, design. Rather, it is an "experiment" where nature seems to have done the randomization and intervention for you, thereby giving you the same type of balance between treatment and control groups that you'd expect to get out of a real experiment.

This idea is best understood by example. Suppose you want to study the effect of class size on student achievement. You reason that, in smaller classes, students can get more individual attention from the instructor, and that intructors will feel a greater sense of personal connection to their students. All else being equal, you believe that smaller class sizes will help students learn better.

A cheap, naïve way to study this question would be to compare the test scores of students in small classes to those of students in larger classes. Any of these confounders, however, might render such a comparison highly unbalanced, and therefore dubious: (1) students in need of remediation are sometimes put in very small classes; (2) highly gifted students are also sometimes put in very small classes; (3) richer school districts can afford both smaller classes and many other potential sources of instructional advantage; or (4) better teachers successfully convince their bosses to let them teach the smaller classes themselves.

An expensive, intelligent way to study this question would

Question	Problem	Natural experiment	Lingering issues
Does being rich make people happy?	Even if richer people are happier on average, maybe happiness and success are the common effect of a third factor. Or maybe the rich grade on a different curve than the rest of us.	Compare a group of lottery winners with a similar group of people who played the lottery but didn't win.	Lottery winners may play the lottery far more often than people who played the lottery but didn't win, which might correlate with other important differences.
Does smoking increase a person's risk for Type-II diabetes?	People who smoke may also engage in other unhealthy behaviors at systematically different rates than non-smokers.	Compare before-and-after rates of diabetes in cities that recently enacted bans on smoking in public places.	Maybe the incidence of diabetes would have changed anyway.
Do bans on mobile phone use by drivers in school zones reduce the rate of traffic collisions?	Groups of citizens that enact such bans may differ systematically in their attitudes toward risk and behavior on the road.	Go to Texarkana, split by State Line Avenue. Observe what happens when Texas passes a ban and Arkansas doesn't.	There may still be systematic differences between the two halves of the city.

Table 9.3: Three hypothetical examples of natural experiments.

be to design an experiment, in conjunction with a scientifically inclined school district, that randomly assigned both teachers and students to classes of varying size. In fact, a few school systems have done exactly this. A notable experiment is Project STAR in Tennessee—an expensive, lengthy experiment that studied the effect of primary-school class sizes on high-school achievement, and showed that reduced class sizes have a long-term positive impact both on test scores and drop-out rates.[9]

But suppose you are neither naïve nor rich, and yet still want to study the question of whether small class sizes improve test scores. If you're in search of a third way—one that's better than merely looking at correlations, yet cheaper than a full-fledged experiment—you might be interested to know the following fact about the Israeli school system.

> [I]n Israel, class size is capped at 40. Therefore, a child in a fifth grade cohort of 40 students ends up in a class of 40 while a child in a fifth grade cohort of 41 students ends up in a class only half as large because the cohort is split. Since students in cohorts of size 40 and 41 are likely to be similar on other dimensions, such as ability and family background, we can think of the difference between 40 and 41 students enrolled as being "as good as randomly assigned."[10]

This is a lovely example of a natural experiment—something you didn't design yourself, but that is almost as good as if you

[9] The original study is described in Finn and Achilles (1990). "Answers and Questions about Class Size: a Statewide Experiment." *American Educational Research Journal* 28, pp. 557–77

[10] Angrist and Pischke (2009). *Mostly Harmless Econometrics*, Princeton University Press, p. 21

had. The researchers in this study compared the students in a group of 40 ("control group," in one large class) versus the students in a group of 41 ("treatment group," split into two smaller classes). This is a plausibly random assignment: the "randomization mechanism" is whether a student fell into a peer group of 40 versus a peer group of 41, and we would not expect this difference to be confounded by anything else that might predict test scores. Therefore, if we see a big difference in performance between the two groups, the most likely explanation is that class size caused the difference.

Some natural experiments, of course, are better than others. Consider the examples in Table 9.3, on page 188. For each one, ask yourself two questions. (1) What are the "treatment" and "control" groups? (2) How balanced are these two groups? (Said another way: how good is the quasi-randomization of cases to these groups?) Think carefully about each one, and you may begin to see "experiment" versus "non-experiment" as the black and white ends of a spectrum, with many shades of grey in between.

Matching

To estimate a causal effect by matching, we artificially construct a balanced data set out of an unbalanced one, by explicitly matching treated cases with similar control cases. We then compare the outcomes in treatment versus control groups, using only the balanced data set. This is most easily seen by example.

An example: the value of going green

For many years now, both investors and the general public have paid increasingly close attention to the benefits of environmentally conscious ("green") buildings. There are both ethical and economic forces at work here. To quote a recent report by Mercer, an investment-consulting firm, entitled "Energy efficiency and real estate: Opportunities for investors":

> Investing in energy efficiency has two intertwined virtues that make it particularly attractive in a world with a changing climate and a destabilized economy: It cuts global-warming greenhouse gas emissions and saves money by reducing energy consumption. Given that the built environment accounts for 39 percent of total energy use in the US and 38 percent of total indirect CO_2 emissions, real estate investment represents

> one of the most effective avenues for implementing energy efficiency.

This only scratches the surface. In commercial real estate, issues of eco-friendliness are intimately tied up with ordinary decisions about how to allocate capital. Every new project involves negotiating a trade-off between costs incurred and benefits realized over the lifetime of the building. In this context, the decision to invest in an eco-friendly building could pay off in at least four ways.

(1) Every building has the obvious list of recurring costs: water, climate control, lighting, waste disposal, and so forth. Almost by definition, these costs are lower in green buildings.

(2) Green buildings are often associated with indoor environments that are full of sunlight, natural materials, and various other humane touches. Such environments, in turn, might result in higher employee productivity and lower absenteeism, and might therefore be more coveted by potential tenants. The financial impact of this factor, however, is rather hard to quantify *ex ante*; you cannot simply ask an engineer in the same way that you could ask a question such as, "How much are these solar panels likely to save on the power bill?"

(3) Green buildings make for good PR. They send a signal about social responsibility and ecological awareness, and might therefore command a premium from potential tenants who want their customers to associate them with these values. It is widely believed that a good corporate image may enable a firm to charge premium prices, to hire better talent, and to attract socially conscious investors.

(4) Finally, sustainable buildings might have longer economically valuable lives. For one thing, they are expected to last longer, in a direct physical sense. (One of the core concepts of the green-building movement is "life-cycle analysis," which accounts for the high front-end environmental impact of acquiring materials and constructing a new building in the first place.) Moreover, green buildings may also be less susceptible to market risk—in particular, the risk that energy prices will spike, driving away tenants into the arms of bolder, greener investors.

Of course, much of this is mere conjecture. At the end of the day, tenants may or may not be willing to pay a premium for

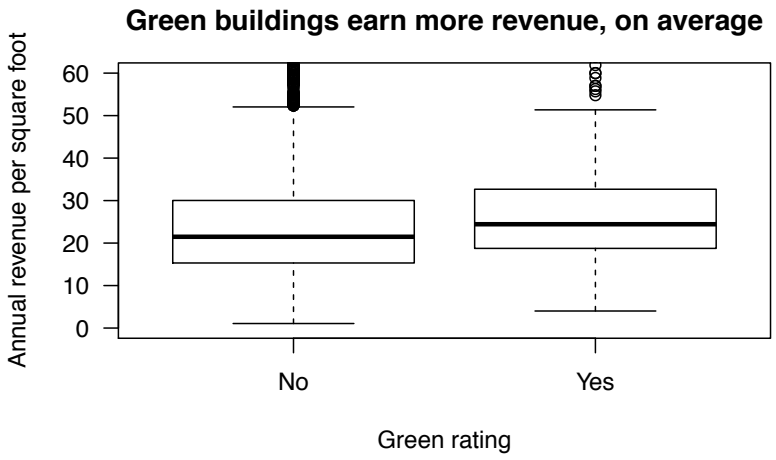


Figure 9.4: Green buildings seem to earn more revenue per square foot, on average, than non-green buildings.

rental space in green buildings. We can only find out by carefully examining data on the commercial real-estate market and comparing "green" versus "non-green" buildings. By "green," we mean that a commercial property has received some official certification, because its energy efficiency, carbon footprint, site selection, and building materials meet certain environmental benchmarks, as certified by outside engineers.[11]

Let's look at some data on 678 green-certified buildings in the United States, together with 6,298 non-green buildings in similar geographic areas. The boxplot above shows that, when we measure revenue by a building's rental rate per square foot per year, green buildings tend to earn noticeably higher revenue (mean = 26.97) than non-green buildings (mean = 24.51). That's a difference of $3.46 per square foot, or nearly a 15% market premium.

[11] The two most common certifications are LEED and EnergyStar; you can easily find out more about these rating systems on the web, e.g. at `www.usgbc.org`.

	Original data	
	Non-green buildings	Green buildings
Sample size	6928	678
Mean revenue/sq ft.	24.51	26.97
Mean age	49.2	23.9
Class A	37%	80%
Class B	48%	19%
Class C	15%	1%

Table 9.4: Covariate balance for the original data. Class A, B, and C are relative classifications within a specific real-estate market. Class A buildings are generally the highest-quality properties in a given market. Class B buildings are a notch down, but still of reasonable quality. Class C buildings are the least desirable properties in a given market.

However, there's a problem with this comparison. As Table 9.4 shows, the green buildings tend to be newer than the non-green buildings, and are more likely to be "Class A" buildings.

So the important question is: do green buildings command a market premium *because* they are green, or simply because they are newer, better buildings in the first place? We can't tell by simply computing the average revenue in each group, because the green ("treatment") and non-green ("control") groups are highly unbalanced with respect to some important confounders.

This is where matching comes in. Matching means constructing a balanced data set from an unbalanced one. It involves three steps:

(1) For each case in the treatment group, find the case in the control group that is the closest match in terms of confounding variables, and pair them up. Put these matched pairs into a new matched data set, and discard the cases in the original data set for which there are no close matches.

(2) Verify covariate balance for the matched data set, by checking that the confounders are well balanced between the treatment and control groups.

(3) Assuming that the confounders are approximately balanced, then compare the treatment-group outcomes with the control-group outcomes, using *only* the matched pairs.

Matching relies on a simple principle: compare like with like. In this example, that means if we have a 25-year-old, Class A building with a green rating, we try to find another 25-year old, Class A building without a green rating to compare it to.

In this particular example, once we've constructed the data set of matched pairs, the confounder variables are much more closely

	Matched data	
	Non-green buildings	Green buildings
Sample size	678	678
Mean revenue/sq ft.	25.94	26.97
Mean age	23.9	23.9
Class A	80%	80%
Class B	19%	19%
Class C	1%	1%

Table 9.5: Covariate balance for the matched data.

balanced between the treatment and control groups (see Table 9.5). A comparison of revenue rates for this matched data set makes the premium for green buildings look a lot smaller: $26.97 versus $25.94, or about a 4% premium. Compare that with the 15% green premium we estimated from the original, unmatched data.

How do we actually find matches? The nitty-gritty algorithmic details of actually finding good matched pairs of cases are best left to the experts who write the software for these things. The two most common types of matching are called *nearest-neighbor search* and *propensity-score matching*; follow the links if you'd like to know more. In R, the package `MatchIt` uses propensity-score matching as a default; this is a very commonly used algorithm in real-world data analysis. In addition, the paper linked here[12] has a much more detailed overview of different matching methods.

[12] "Matching Methods for Causal Inference: A Review and a Look Forward." Elizabeth A. Stuart, *Statistical Science*, 2010.

Matching isn't a silver bullet: a bigger example

If you've ever been admitted to the intensive-care unit at a hospital, you may have undergone a diagnostic procedure called right heart catheterization, or RHC. RHC is used to see how well a patient's heart is pumping, and to measure the pressures in that patient's heart and lungs. RHC is widely believed to be helpful, since it allows the doctor to directly measure what's going on inside a patient's heart. But it is an invasive procedure, since it involves inserting a small tube (the catheter) into the right side of your heart, and then passing that tube through into your pulmonary artery. It therefore poses some risks—for example, excessive bleeding, partial collapse of a lung, or infection.

A natural question is: do the diagnostic benefits of RHC outweigh the possible risks? But this turns out to be tricky to answer. The reason is that doctors would not consider it ethical to run a randomized, controlled trial to see if RHC improves patient outcomes. As the authors of one famous study from the 1990s pointed out:[13]

> Many cardiologistics and critical care physicians believe that the direct measurement of cardiac function provided by right heart catheterization (RHC) . . . is necessary to guide therapy for certain critically ill patients, and that such management leads to better patient outcomes. While the benefit of RHC has not been demonstrated in a randomized controlled trial (RCT), the popularity of this procedure, and the widespread

[13] "The effectiveness of right heart catheterization in the initial care of critically ill patients." Connors et. al. *Journal of the American Medical Association*. 1996 Sep 18; 276(11):889-97.

	Original data		Matched data	
	No RHC	RHC	No RHC	RHC
Sample size	3551	2184	2184	2184
180-day survival rate	0.370	0.320	0.354	0.320
mean APACHE score	50.934	60.739	57.643	60.739
Trauma	0.005	0.016	0.008	0.016
Heart attack	0.030	0.043	0.036	0.043
Congestive heart failure	0.168	0.195	0.209	0.195
Sepsis	0.148	0.321	0.24	0.321

Table 9.6: A before-and-after table of summary statistics showing covariate balance for the observational study on right-heart catheterization. The entries for trauma, heart attack, etc. show rates of these complications in the two groups. The left half of the table shows the original data set, while the right half shows the matched data set.

belief that it is beneficial, make the performance of an RCT difficult. Physicians cannot ethically participate in such a trial or encourage a patient to participate if convinced the procedure is truly beneficial.

We're therefore left with only observational data on the effectiveness of RHC—which, on the surface, doesn't look good! Here's the data from the study quoted above, showing that critically ill patients undergoing RHC actually have a *worse* 180-day survival rate (698/2184, or 32%) than patients not undergoing RHC (1315/3551, or 37%):

	No RHC	RHC
Survived 180 days	1315	698
Died within 180 days	2236	1486

What's going on here? Should we conclude that right heart catheterization is actually killing people, and that the doctors are all just plain wrong about its putative benefits?

Not so fast. The problem with this conclusion is that the treatment (RHC) and control (no RHC) groups are heavily unbalanced with respect to baseline measures of health. Put simply, the patients who received RHC were a lot sicker to begin with, so it's no surprise that they have a lower 6-month survival rate. To cite a few examples: the RHC patients were three times more likely to have suffered acute trauma, 50% more likely to have had a heart attack, and 16% more likely to be suffering from congestive heart failure. The RHC patients also had an average APACHE score that was 10 points higher than the non-RHC patients.[14] The left half of Table

[14] The APACHE score is a composite severity-of-disease score used by hospital ICUs to estimate which patients have a higher risk of death. Patients with higher numbers have a higher risk of death.

9.6 shows these rates of various complications for the two groups in the original data set. They're quite different, implying that the survival rates of these two groups cannot be fairly compared.

And what about after matching? Unfortunately, Table 9.6 shows that, even after matching treatment cases with controls having similar complications, the RHC group still seems to have a lower survival rate. The gap looks smaller than it did before, on the unmatched data—a 32% survival rate for RHC patients, versus a 35.4% survival rate for non-RHC patients—but it's still there.

Again we find ourselves asking: what's going on? Is the RHC procedure actually killing patients? Well, it might be, at least indirectly! The authors of the study speculate that one possible explanation for this finding is "that RHC is a marker for an aggressive or invasive style of care that may be responsible for a higher mortality rate." Given the prevalance of overtreatment within the American health-care system, this is certainly plausible.

But we can't immediately jump to that conclusion on the basis of the matched data. In fact, this example points to a couple of basic difficulties with using matching to estimate a causal effect.

The first (and most important) difficulty is that *we can't match on what we haven't measured.* If there is some confounder that we don't know about, then we'll never be able to make sure that it's balanced between the treatment and control groups within the matched data. This is why experiments are so much more persuasive: because they also ensure balance for unmeasured confounders. The authors of the study acknowledge as much, writing:

> A possible explanation is that RHC is actually beneficial and that we missed this relationship because we did not adequately adjust for some confounding variable that increased both the likelihood of RHC and the likelihood of death. As we found in this study, RHC is more likely to be used in sicker patients who are also more likely to die.

Another possible explanation is that we simply haven't been able to match treatment cases with control cases very effectively. The right half of Table 9.6 shows that covariate balance for the matched data is noticeably better than for the unmatched data, but it's not perfect. We still see some small differences in complication rates and APACHE scores between the treatment and control group. There are two main reasons for this.

(1) First, and most importantly, although finding a match on one or two variables is relatively easy, finding a match on several

variables is pretty hard. Think of this in terms of your own life experience—for example, in seeking a spouse or partner. It probably isn't too hard to find someone who's a good match for you in terms of your interests and your sense of humor. But if you require that this person *also* match you in terms of age, career, education, home town, height, weight, looks, and favorite sport, then you're a lot less likely to find a match. *Picky people are less likely to find a satisfying match in life.* For this same reason, it's unlikely that we'll be able to find an exact match for each treatment case in a matching problem, especially with lots of possible confounders.

(2) Second, finding matches for cases with rare confounders is especially hard—by definition, since the confounder is rare!

These two points underline a basic difficulty with matching: perfect matches usually don't exist, and we have no choice but to accept approximate matches. In practice, therefore, we give up on the requirement that every single pair of matched observations is similar in terms of all possible confounders, and settle for having matched groups that are similar in their confounders, *on average*. That's why it's so important to check the covariate balance after finding matched pairs, to make sure that there's nothing radically different between the two groups.

Model-based statistical adjustment

A fourth identification strategy for estimating a causal effect is to build a regression model. If some important (and quite strong) assumptions are met, then such a model is capable of isolating a causal relationship between predictor and response, by adjusting for the effects of confounders *statistically*, rather than experimentally. You may have heard this process described as "statistical control" or "statistical correction," both in the popular media and in scientific publications:

- "Schatz's numbers are unique in that they evaluate each play against the league average for plays of its type, adjust for the strength of the opponents' defense, and even try to divide credit for a given play among teammates."[15]
- "The committee concluded that a statistical adjustment of the 1990 census leads to an improvement of the counts."[16]

[15] "Pigskin Pythagoras: A guy from Framingham tries to remake the muddy field of football statistics." *Boston Globe*, February 1, 2004

[16] "Judge must decide on census adjustment." *Chicago Tribune*, 6/8/1992

- "Further adjustment for weight change and leukocyte count attenuated these risks substantially."[17]

[17] "Smoking, Smoking Cessation, and Risk for Type 2 Diabetes Mellitus: A Cohort Study." *Annals of Internal Medicine*, January 4, 2010

Estimating a causal effect using a regression model is, in principle, no different than estimating a partial relationship, which we've already learned how to do:

(1) Build a multiple regression model for the outcome (y) versus the predictor of interest (x) and other possible confounders;

(2) Interpret the coefficient on the x variable of interest as the partial linear relationship between y and x, holding confounders constant.

The key question is: under what circumstances can we interpret the partial relationship in a multiple regression model as the *causal* effect of x on y? By *causal effect*, you should think in terms of the counterfactuals we entertained at the beginning of the chapter: *if* I were to intervene and change x by one unit, holding all other variables constant, *then* how much would y change on average?

There are three important assumptions that must be met in order to give a causal interpretation to a regression coefficient. First, your model must include all confounding variables (that is, variables that have a causal effect on both the treatment assignment and the outcome). Second, the model must be correct. In this context, "correct" means that you have included the right interactions among confounding variables, and that you have specified the right functional form of the model (linear, polynomial, power law, etc.). Finally, you must *not* include any post-treatment effects as covariates in the model. A post-treatment effect is something causally "downstream" from the treatment variable, and that becomes known only as a result of receiving or not receiving the treatment. This is a subtle point, and we won't discuss it in detail. But the important thing is: include those confounders, and *only* those confounders, that affect the allocation of cases to the treatment and control groups.

If, and only if, these three assumptions about your model are true, then the regression coefficient of y on x has a causal interpretation. If, on the other hand, there are any unmeasured confounders affecting your x and y variable, then the coefficient of y on x measures association, not causation. This is called *omitted-variable bias*.[18]

[18] Or *lurking-variable bias.*

Another way of saying this is that *if* the possible confounders are all observed, then accurately estimating the causal effect of

x on y really just boils down to modeling the data well, and not using that model to extrapolate beyond the range of available data. However, the assumption that we've observed all relevant confounders, and can therefore adjust for them appropriately, is very strong. It's also unverifiable using the data; as with matching, you have to believe this assumption, and convince people of it, on extrinsic grounds.

Using regression analysis to estimate causal effects is a big, serious topic. Here are two full books about it:

- *Causality*, by Judea Pearl
- *Observational Studies*, by Paul Rosenbaum

For some additional, more easily digestible advice on choosing which covariates to include in a causal model, see Chapter 17 of Daniel Kaplan's book on statistical modeling.[19]

[19] Kaplan also has a good explanation for why it's not a good idea to include post-treatment effects (i.e. variables causally downstream of the treatment) as covariates in a regression model.

Matching versus regression, or matching and regression?

We've seen that it's easiest to infer causality if the cases in the treatment group are comparable to those in the control group. One way to do this is via matching: explicitly constructing a balanced data set from an unbalanced one. Another way to do this is via regression: adjust for confounders using a statistical model, so that we can evaluate the partial relationship between treatment and response, holding confounders constant.

This makes it sound as though regression and matching are competing identification strategies for causal inference. Sociologically speaking, there is certainly some truth to this, in that some people tend to use matching more often, and others tend to use regression more often. So which one should *you* use?

In the real world, if you're going to use only one strategy or the other, my advice is to use matching, mainly for three reasons:

(1) Matching is a lot easier for non-experts to understand, since you can point to the matched treatment and control groups and show that they are visibly balanced with respect to observed confounders. In other words, the nature of the "balanced comparison" being made via matching is much more transparent than the idea of a partial slope in a regression model. This will make it easier for you to convince others of your conclusions.

(2) Matching is a bit more robust than regression, at least in their "off the shelf" versions. The regression-based approach to causal inference relies on a whole bunch of hard-to-verify assumptions: linearity, all necessary interactions included, and so forth. By comparison, it's a lot easier to verify covariate balance using before-and-after tables of summary statistics. (Of course, neither method is robust to unmeasured confounders—only an experiment can fix that problem.)

(3) Unwarranted extrapolations are more apparent when matching than when using regression. Suppose that the treatment and control groups have highly nonoverlapping distributions of confounders—for example, that most the men are in the treatment group and most of the women in the control group. In such cases, the data are inherently limited in what they can tell us about the treatment–response relationship in this region of nonoverlap (i.e. how the treatment will work for women). This lack of overlap will be obvious if you use matching, because you'll still have drastic post-match covariate imbalances that will stick out like a sore thumb. But the lack of overlap will be less obvious if you throw all the confounders into a multiple regression model without plotting your data.

In summary, it's easier to convince others with matching, and easier to fool yourself with regression. These aren't intrinsic *statistical* advantages to matching; they are merely *practical* advantages worth keeping in mind.

It turns out, however, that there's no need to choose between matching and regression. Better still is to use both matching *and* regression, to get better estimates of causal effects than either technique is capable of getting on its own. In other words: first run matching to get an approximately balanced data set. Then run a regression model for the response versus the treatment variable and the confounders, to correct for minor imbalances in the matched data set. Under this approach, the primary role of matching is to correct for major covariate imbalances between the groups, while the primary role of regression is to model the treatment–response relationship in a way that adjusts for any minor confounding that remains in the matched data set.

There's one other major advantage of using matching and regression together. By fitting a regression model to a matched data set, you are able to search for interactions *between* the treatment

variable and possible confounders. For example, what if the treatment effect is different for men than for women? You can discover this kind of modulating effect much more easily using a regression model than you can with matching alone.

In summary, matching and regression make for an excellent pair. There's rarely a good reason to use just one or other!

10
Expected value and probability

Risky business

For most of us, life is full of worry. Some people worry about tornados or earthquakes; other people won't get on an airplane. Some people worry more about lightning; others, about terrorists. And then there are the everyday worries: about love, money, career, status, conflict, kids, and so on.

Jared Diamond worries a lot, too—about slipping in the shower.

Dr. Diamond is one of the most respected scientists in the world. Though he originally trained in physiology, Diamond left his most lasting mark on the popular imagination as the author of *Guns, Germs, and Steel: The Fates of Human Societies*. This Pulitzer-prize-winning book draws on ecology, anthropology, and geography to explain the major trends of human migration, conquest, and displacement over the last few thousand years.

Strangely enough, Diamond began to worry about slipping in the shower while conducting anthropological field research in the forests of New Guinea, 7,000 miles away from home, and a long day's walk from any shower. The seed of this worry was planted one day while he was out hiking in the wilds with some New Guineans. As night fell, Diamond suggested that they all make camp under the broad canopy of a nearby tree. But his companions reacted in horror, and refused. As Diamond tells it,

> They explained that the tree was dead and might fall on us. Yes, I had to agree, it was indeed dead. But I objected that it was so solid that it would be standing for many years. The New Guineans were unswayed, opting instead to sleep in the open without a tent.[1]

[1] Jared Diamond, "That Daily Shower Can Be a Killer." *New York Times*, January 29. 2013, page D1.

The New Guineans' fear initially struck Diamond as overblown. How likely could it possibly be that the tree would fall on them in the night? Surely they were being paranoid. For a famous professor like Diamond to get crushed by a tree while sleeping in the

forest would be the kind of freakish thing that made the newspaper, like getting struck by lightning at your own wedding, or being killed by a falling vending machine.

But in the months and years after this incident, it began to dawn on Diamond that the New Guineans' "paranoia" was well founded. A dead tree might stay standing for somewhere between 3 and 30 years, so that the daily risk of a toppling was somewhere between 1 in 1,000 and 1 in 10,000. This is small, but far from negligible. Here's Diamond again:

> [W]hen I did a frequency/risk calculation, I understood their point of view. Consider: If you're a New Guinean living in the forest, and if you adopt the bad habit of sleeping under dead trees whose odds of falling on you that particular night are only 1 in 1,000, you'll be dead within a few years.[2]

[2] *ibid.*

Having absorbed this attitude about the importance of everyday habits, Diamond began to apply it to his own life. He refers to it as a "hypervigilant attitude towards repeated low risks," or more memorably, "constructive paranoia."

Take the simple act of showering. If you're 75 years old, as Diamond was when he recounted this story, you can expect to live another 15 years. That's $15 \times 365 = 5{,}475$ more daily showers. So if your risk of a bad slip is "only" one in a thousand, you should expect to break your hip, or worse, about five times over that period. The implication is that, if you want a good chance of being around to blow out 90 candles, you must ensure that, by your own careful behavior, you reduce the risk of slipping in the shower to something much, much lower than one in a thousand.

And the same goes for all those other small risks we face day in, day out. Think about crossing a busy street, driving at night, touching the handle of a public toilet, or venturing out with the mad dogs and Englishmen into the mid-day sun. Each time the chance of a disaster is low. But most of us perform these actions again and again—and if we're slapdash about it, the expected number of disasters over several years can be alarmingly high. Diamond's conclusion? He needed to ensure that, for each repeated exposure to one of these risks, the chance of a disaster wasn't just low, but extremely low.

Expected value and the NP rule

Jared Diamond's philosophy of constructive paranoia arises from an understanding of *expected value*. This concept has a formal

mathematical definition, but the basic idea is simple. Think about risks like slipping in the shower, or having a dead tree fall on you in the night. These kinds of risks involve many repeated exposures to the same chance event. In the long run, the expected number of events is the frequency of encounters (N), times the probability of the event in a single encounter (P).

This is such a common scenario that we like to give it a name: the NP rule, where expected value = frequency times risk, or $N \times P$. For example, let's say that the risk of a dead tree falling down in the night is one chance in a thousand (so $P = 0.001$), and that you and 99 friends each sleep under your own dead tree every night for a year (so $N = 365 \times 100 = 36{,}500$ person-nights of exposure). In your cohort of 100, how many would you expect to get crushed by a tree? The math of the NP rule doesn't look good; you can expect about 36 of you to be crushed.

$$\begin{aligned}\text{Expected crushings} &= (\text{Risk of dead tree falling}) \times (\text{Number of exposures}) \\ &= \frac{1}{1000} \times (365 \times 100) \\ &= 36.5\,.\end{aligned}$$

What about some more familiar risks?

Of course, you probably don't live in a forest in New Guinea. How does the NP rule play out in thinking about risks for a typical 21st-century citizen of a western democracy?

To get specific, let's look at some expected values for an imaginary cohort of 100,000 Americans—about the size of a small city, like Boulder or Green Bay. Table 10.1 shows how many of these 100,000 people we would expect to die in any given year due to various causes.[3] This is exactly the kind of table that a public-health organization like the Gates Foundation might look at it in order to decide what kinds of initiatives would have the biggest return on investment, or that a life-insurance company would look at to set your premiums.

[3] Centers for Disease Control, http://www.cdc.gov/nchs/fastats/.

There are two take-away lessons from Table 10.1. First, the expected number of deaths due to the headline-grabbing causes in the bottom half of the table—from tornadoes to shark attacks to mass shootings—is tiny. Of course, tornadoes, sharks, and crazed gunmen are still very dangerous (P is high). But they're also rare (N is small). Remember: expected value = $N \times P$.

Cause	Expected deaths
Heart disease	203
Cancer	195
Respiratory disease	50
Stroke	42
Alzheimer's	28
Diabetes	25
Accidental poisoning	12
Car accident	11
Slips/falls	10
Homicide	5
Eating raw meat	2
Choking	1.5
Pregnancy	0.2
Dog bite	0.01
Falling vending machine	0.001
Hurricane	0.03
Tornado	0.02
Mass shooting	0.01
Lightning strike	0.01
Shark attack	0.0003
Plane crash	0.0001

(per 100,000)

Table 10.1: Expected deaths due to various causes over one year in an imaginary cohort of 100,000 Americans, of whom 99,200 are expected to survive.

Second, in light of these numbers, it might be wise to heed Jared Diamond's advice. While most people die of disease, cancer, or the depredations of age, a shockingly high number die in preventable accidents. Even unusual kinds of accidents are still far more common than the six sensational causes of death in the bottom half of the table. In fact, we'd expect ten times as many people to die from a falling vending machine as from a falling plane, and 20 times as many to die from choking as from all the bottom six causes put together. Of course, eating lunch or buying a granola bar are usually safe, so P is small. But people do these things every day, so N is huge.

Studies, however, repeatedly find that our concept of danger is woefully incomplete: we think only about P, and rarely about N. As a result, we overestimate the chance of dying in some dramatic event like those in the bottom half of Table 10.1, while simulta-

neously underestimating the chance of dying from one of the familiar causes in the top half.

To be fair, this has a lot to do with living in a world of mass media and near-instant communication. Thousands of people anonymously choke to death every year. But if someone gets attacked by a shark or blows himself up in a train station anywhere in the world, you will hear about it, no matter how unlikely the real risk. As folks in the statistics business put it: newspapers love numerators. While this brings a website plenty of clicks, it also short-circuits our natural cues for reasoning intuitively about risk.

But dwelling on the spectacular numerators isn't a smart way to stay alive. Many of life's mundane risks, from car accidents to skin cancer, do not strike out of the blue. Rather, they are direct results of our own day-to-day behavior. So follow your mother's advice. Look both ways, don't drive while tired, wear sunscreen, wash your hands—and don't sleep under dead trees.

The NP rule in health care and social policy

The concept of expected value is central to any cost/benefit analysis. For example, the same idea behind the NP rule is used routinely to evaluate medical procedures.

In a medical context, an expected-value calculation is usually phrased in terms of a number called the NNT: the number needed to treat. Here's the idea. Suppose you invent a perfect drug for some intractable disease. Anyone who takes the drug is cured, and it's the only cure. Here, we'd say that your drug has a "number needed to treat" of one: if you treat one person, you cure one person. You can't do any better than this.

Now let's say that the drug has only a 50% chance of curing someone ($P = 0.5$). In that case, if you treated $N = 2$ people, you would expect to cure one patient: $N \times P = 1$. Here, we'd say that the NNT is two: treat two, cure one.

An NNT of two is really good. But if you needed to treat 100 or 1000 people to cure just a single person, you might view the drug a bit more skeptically. More generally, suppose that a medical procedure has probability P of offering some specific health benefit to any one person—like curing a disease, or offering one extra year of life. If we treat N people, we would expect that $N \times P$ people would get the benefit. How many people do we need to treat so

Treatment	Benefit	NNT
Defibrillation for cardiac arrest	Prevents death	2.5
Corticosteroid injection for tennis elbow	Reduces pain	4
Zinc for the common cold	Reduces symptoms	5
Antibiotics for conjunctivitis	Full recovery within 5 days	7
Bone-marrow transplant after chemo for leukemia	Prevents relapse	9
Strength and balance programs for the elderly	Prevents falls	11
Warfarin for atrial fibrillation	Prevents stroke	25
Aspirin, for patients with known heart disease	Prevents heart attack or stroke	50
a Mediterrenean diet	Prevents heart disease	61
Magnesium sulfate for preeclampsia in pregnancy	Prevents seizures	90
Statins, for patients with no known heart disease	Prevents heart attack	104
CT scans of long-term smokers	Detects lung cancer	217
Aspirin, for patients with no known heart disease	Prevents heart attack or stroke	1667

Table 10.2: Numbers sourced from Cochrane reviews, as summarized by the NNT website: http://www.thennt.com.

that the expected number helped, $N \times P$, is one? That number is called the procedure's number needed to treat, or NNT.[4] Similarly, for a medical test like a mammogram or a prostate exam, there's the NNS: the "number needed to screen."

Table 10.2 has some estimates of the number needed to treat for some common medical interventions.

[4] This is a slight simplification. The NNT is more typically defined to be the number needed to treat in order to offer some benefit to one *additional* patient versus some baseline, like a placebo or the next-best drug.

Weighing medical harms and benefits

The number needed to treat is a big deal to doctors, health insurers, and governments that run national health services. A high NNT means a low expected value for the number of patients helped. Essentially, it's a measure of waste: if a treatment has an NNT of 100, then on average, it will fail to yield the stated benefit for 99 out of 100 patients.

Of course, we don't know who those 99 will be ahead of time. And if the treatment is cheap and mostly harmless, or if the possible benefit is extremely important, then a high NNT might be acceptable. For example, the use of aspirin to prevent a first heart attack has an NNT of over 1000, but plenty of doctors recommend it routinely, despite its side effects.[5]

But as you may have heard, modern health care is expensive. It already strains the budgets of most households and governments. Paying for one thing usually means not paying for something else, and knowing the NNT helps us to be clear-eyed about these

[5] Antithrombotic Trialists Collaboration. "Aspirin in the primary and secondary prevention of vascular disease: collaborative meta-analysis of individual participant data from randomised trials." *Lancet.* 2009; 373(9678); 1849-60.

opportunity costs.

Moreover, a lot of procedures present at least some probability Q of unwanted side effects—for example, the risk that a mammogram will lead to a false-positive finding. That means a medical cost/benefit analysis really has two expected values to contend with: the expected number of people helped, $N \times P$; and the expected number harmed, $N \times Q$. In this context, we speak of the "number needed to harm," or NNH: the number of people we'd need to treat in order to harm a single person in some specific way.

For these reasons, a high-NNT medical procedure usually provokes two questions.

For governments and insurers: Could we produce a greater good for a greater number of people by redirecting our limited resources to some other treatment?

For everyone: How bad are the side effects, and what's the number needed to harm (NNH)? Imagine a treatment that produces nasty side effects in every fifth patient (NNH = 5), but only cures every hundredth (NNT = 100). Depending on how bad the side effects are compared with the original condition, you might prefer no treatment at all.

Expected value and mammograms. Indeed, it was exactly this second question that spurred the American Cancer Society to revise its guidelines on screening mammograms for women with no family history of breast cancer. From the introduction, you may recall the Society's previous recommendation for these women: get a mammogram every year starting at age 40. This approach benefitted some people, and harmed others. Specifically, a systematic review of many earlier studies estimated that, under this approach, we'd need to regularly screen about 2,500 women aged 40-49 in order to save one life (NNS $\approx$ 2500). Of these 2,500 women, about 175 would end up experiencing a false-positive biopsy result. This imples an NNH of about 14: for every 14 women screened, someone got hurt.[6]

[6] Myers et. al., ibid.

However, if we were to apply the Society's new screening recommendations to these same 2,500 women, we'd still expect to save that one life, on average. But now we would expect only 120 false positives. That's 55 women out of every 2,500 who are spared from needless stress and medical intervention, with no detectable increase in the risk of someone dying. These expected-value cal-

culations were a big part of the reasoning behind the American Cancer Society's new recommendation: that women with no family history should get screened every year from 45-54, and every two years after that.

Expected value and PSA screening. Screening mammograms are not the only medical procedure to spark a debate about expected value. A common test for prostate cancer, called the prostate-specific antigen (PSA) test, has been at the center of a similar controversy for years. Prostate cancer kills over 300,000 men per year worldwide. However, it's also incredibly common for a prostate tumor to come late in life and grow slowly. In fact, autopsy records show that something like 2/3 of all elderly men die with asymptomatic tumors in their prostates.

Here's why the PSA test is controversial. The test detects elevated levels of prostate-specific antigen in the blood, which is a potential indicator of a prostate tumor. If a man's PSA levels are high enough, he's referred for a prostate biopsy to get a tissue sample. This has some small probability P of detecting a deadly tumor. But because asymptomatic prostate cancer is so common, the test also has some other probability Q of leading to unnecessarily aggressive courses of treatment for a tumor that never would have done much harm. Some of the men who undergo these treatments end up incontinent, impotent, or dead.

Is the life-saving potential of PSA screening for prostate cancer worth these harms? The U.S. Preventive Services Task Force says no: P is tiny and Q is large. Here's how their report describes PSA tests:

> The reduction in prostate cancer mortality after 10 to 14 years is, at most, very small, even for men in what seems to be the optimal age range of 55 to 69 years. There is no apparent reduction in all-cause mortality. In contrast, the harms associated with the diagnosis and treatment of screen-detected cancer are common, occur early, often persist, and include a small but real risk for premature death. Many more men in a screened population will experience the harms of screening and treatment of screen-detected disease than will experience the benefit. The inevitability of overdiagnosis and overtreatment of prostate cancer as a result of screening means that many men will experience the adverse effects of diagnosis and treatment of a disease that would have remained asymptomatic throughout their lives.

The Task Force concludes simply that "the benefits of PSA-based

screening for prostate cancer do not outweigh the harms."[7]

[7] Moyer et. al. "Screening for Prostate Cancer: U.S. Preventive Services Task Force Recommendation Statement." *Annals of Internal Medicine* 157(2), 2012.

Postscript

Now would be a good time to repeat our earlier disclaimer: we are not qualified to endorse or dispute the American Cancer Society's guidelines on mammograms, or the USPSTF's guidelines on PSA screening. We're merely trying to highlight the role of expected value in their thinking, and to emphasize two broader lessons to be found in these debates.

First, we appreciate that, if you're a patient thinking through your treatment options, what matters most are your own circumstances and preferences. While population-level quantities like an expected value or an NNT can guide your thinking, it's your own situation-specific, *conditional* probabilities that really ought to be decisive. However, those in the business of setting health policy—whether for a government, insurance company, or professional society—simply cannot avoid the principle of expected value. We ask these people to act like responsible utilitarians on behalf of a wider population. To do this, they must think about both N and P.

The second lesson is that cause and effect are both complicated and probabilistic. Most interventions produce the intended effect in any individual case only with some probability P. For many policies, P is very small, and the risk Q of unwanted side effects may be much higher. We should weigh the policy's costs and benefits in light of the expected values for both the good and the bad outcomes.

But it's all too easy to let ourselves fall into some counterfactual dream state, especially if can't shake the impression left by that one awesome example where the policy really *did* work. "If things turned out like that every time," we think to ourselves, "imagine how many lives/dollars/hours/puppies we could save." But that's a big "if." Controversial medical tests are great examples of this phenomenon. If you read up on the debates surrounding mammograms or PSA screening, you'll notice a striking rhetorical pattern. The medical societies and task forces recommending fewer screens always cite expected values based on peer-reviewed medical research. The doctors and patients who cry out in opposition often cite anecdotes or "clinical experience."

There are many other examples outside medicine. For example, in the 1990s, California passed its infamous "three-strikes" law, where someone with a third felony conviction automatically re-

ceived at least a 25-year prison sentence. These once-fashionable laws have now fallen out of favor, but it's easy to understand how they could have been passed in the first place. All it takes is for one judge to be a bit too lenient, and for a thrice-convicted felon to go on a headline-grabbing rampage after getting out of prison, for that single canonical example to become frozen in the public's mind. From there, the "obvious" policy solution is hardly a big leap: three-time felons must spend the rest of their lives in jail.

As it happens, while California's three-strikes laws may have prevented some crimes, many scholars have concluded that it was largely ineffective.[8] One thing the law did do, however, was create a sharp incentive for criminals to avoid that third arrest. As a result, the law may have caused more felonies than it prevented, by increasing the chance Q that a suspect with two strikes will assault or murder a police officer who's about to arrest them.[9] It also cost taxpayers a huge amount of money to prosecute, secure, feed, and clothe all those dangerous felons whose third strike consisted of an illegal left turn with three dimebags of marijuana in the passenger seat.

So if you ever get to make any kind of policy, keep expected value at the front of your thoughts, and mind your P and Q.

[8] Males et. al. "Striking Out: The Failure of California's 'Three Strikes and You're Out' Law." Stanford Law and Policy Review, Fall 1999.

[9] Johnson and Saint-Germain. "Officer Down: Implications of Three Strikes for Public Safety." *Criminal Justice Policy Review*, 16(4), 2005.

Probability: a language for uncertainty

All of these examples illustrate the concept of a *random variable*, which is a generic term for any uncertain outcome. For example:

- the number of trees that fall over tonight in a particular patch of New Guinean forest.
- the number of women aged 50-70, out of a group of 200, who will get breast cancer.
- how many users will click on a particular Google ad in the next hour.

These random variables all fall within the NP rule, where the expected value is found by mutiplying the risk times the exposure.

But here's where we run up against the limitations of thinking about randomness purely in terms of a simple risk/exposure calculation. One problem is a lack of generality. For example, it's not at all clear how we could use this approach to calculate an expected value for *these* uncertain outcomes:

- the rate of U.S. unemployment in 18 months.
- the value of your retirement portfolio in 30 years.
- your extra lifetime earnings from going to graduate school.

What does a risk/exposure calculation even look like here? And what about these uncertainties?

- the winner of next year's Tour de France.
- whether a defendant is guilty or innocent.
- whether you'll like the next person you're matched up with through a dating app.

Here the possible outcomes aren't even numbers.

A second, even bigger problem is that an expected value conveys nothing about *uncertainty*. We may expect that 11 people in Green Bay, Wisconsin (pop. 100,000) will die this year in a car accident. But it could be 5, or 20. It's a random variable; no one knows for sure.

To really understand risk deeply, we need a better language for helping us to communicate clearly about uncertainty. That language is probability.

Probability

Probability is a rich language for communicating about uncertainty. Up to now we've spoken in fairly loose terms about this concept. And while most of us have an intuitive notion of what it means, it pays to be a bit more specific.

A probability is just a number that measures how likely it is that some event, like rain, will occur. If A is an event, $P(A)$ is its probability: $P(\text{coin lands heads}) = 0.5$, $P(\text{rainy day in Ireland}) = 0.85$, $P(\text{cold day in Hell}) = 0.0000001$, and so forth.

Some probabilities are derived from data, like the knowledge that a coin comes up heads about 50% of the time in the long run, or that 11 people out of 100,000 die in a car accident. But it's also perfectly normal for a probability to reflect your subjective assessment or belief about something. Here, you should imagine a stock-market investor who has to decide whether to buy a stock or sell it. The performance of a stock over the subsequent months involves a bunch of one-off events that have never happened before, and will never be repeated. But that's OK. We can still talk

about a probability like $P(\text{Apple stock goes up next month})$. We just have to recognize that this probability reflects someone's subjective judgment, rather than a long-run frequency from some hypothetical coin-flipping experiment.

If you don't have any data, a great way to estimate the probability of some event is to get people to make bets on it. Let's take the example of the 2014 mens' final at Wimbledon, between Novak Djokovic and Roger Federer. This was one of the most anticipated tennis matches in years. Djokovic, at 27 years old, was the top-ranked player in the world and at the pinnacle of the sport. And Federer was—well, Federer! Even at 32 years old and a bit past his prime, he was ranked #3 in the world, and had been in vintage form leading up to the final.

How could you synthesize all this information to estimate a probability like $P(\text{Federer wins})$? Well, if you walked into any betting shop in Britain just before the match started, you would been quoted odds of 20/13 on a Federer victory.[10] To interpret odds in sports betting, think "losses over wins." That is, if Federer and Djokovic played 33 matches, Federer would be expected to win 13 of them and lose 20, meaning that

$$P(\text{Federer wins match}) = \frac{13}{13+20} \approx 0.4\,.$$

The markets had synthesized all the available information for you, and concluded that the pre-match probability of a Federer victory was just shy of 40%. (Djokovic ended up winning in five sets.)

[10] There are approximately 9,000 betting shops in the United Kingdom. In fact, it is estimated that approximately 4% of all retail storefronts in England are betting shops.

Conditional probability

Another very important concept is that of a *conditional probability.* A conditional probability is the chance that some event A happens, given that another event B happens. We write this as $P(A \mid B)$ for short, where the bar (|) means "given" or "conditional upon."

We're all accustomed to thinking about conditional probabilities in our everyday lives, even if we don't do so quantitatively. For example:

- $P(\text{rainy afternoon} \mid \text{cloudy morning})$,
- $P(\text{rough morning} \mid \text{out late last night})$,
- $P(\text{rough morning} \mid \text{out late last night, drank extra water})$,

and so forth. As the last example illustrates, it's perfectly valid to condition on more than one event.

A key fact about conditional probabilities is that they are not symmetric: $P(A \mid B) \neq P(B \mid A)$. In fact, these two numbers are sometimes very different. For example, just about everybody who plays professional basketball in the NBA practices very hard:

$$P(\text{practices hard} \mid \text{plays in NBA}) \approx 1\,.$$

But sadly, most people who practice hard with a dream of playing in the NBA will fall short:

$$P(\text{plays in NBA} \mid \text{practices hard}) \approx 0\,.$$

We'll see a few examples later where people get this wrong, and act as if $P(A \mid B)$ and $P(B \mid A)$ are the same. Don't do this.

Conditional probabilities are used to make statements about uncertain events in a way that reflects our assumptions and our partial knowledge of a situation. They satisfy all the same rules as ordinary probabilities, and we can compare them as such. For example, we all know that

$$\begin{aligned} P(\text{rainy afternoon} \mid \text{clouds}) &> P(\text{rainy afternoon} \mid \text{sun})\,, \\ P(\text{shark attack} \mid \text{swimming in ocean}) &> P(\text{shark attack} \mid \text{watching TV})\,, \\ P(\text{heart disease} \mid \text{swimmer}) &< P(\text{heart disease} \mid \text{couch potato})\,, \end{aligned}$$

and so forth, even if we don't know the exact numbers.

The rules of probability

Probability is an immensely useful language, and there are only a few basic rules. These are sometimes called Kolmogorov's rules, after a Russian mathematician. (Like chess and gymnastics, probability is a very Russian pursuit.)

(1) All probabilities are numbers between 0 and 1, with 0 meaning impossible and 1 meaning certain.

(2) Either an event occurs (A), or it doesn't (not A):

$$P(\text{not } A) = 1 - P(A)\,.$$

(3) If two events are mutually exclusive (i.e. they cannot both occur), then

$$P(A \text{ or } B) = P(A) + P(B)\,.$$

These are usually called Kolmogorov's rules. There's also a fourth, slightly more advanced rule for conditional probabilities:

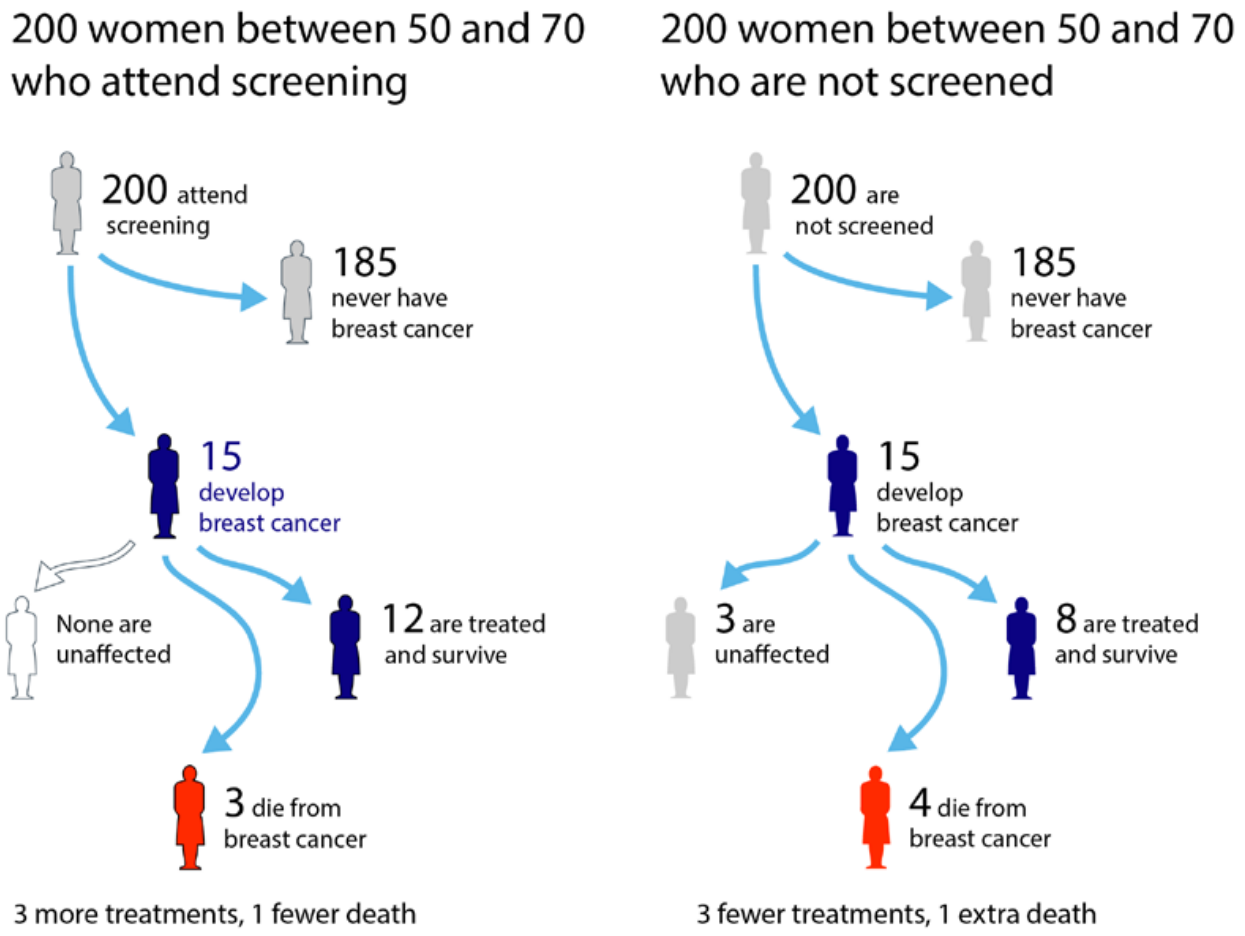


Figure 10.1: Two hypothetical cohorts of 200 women, ages 50-70. The 200 women on the left all go in for mammograms; the 200 on the right do not. The branches of the tree show how many women we would expect to experience various different outcomes. Figure from: "What can education learn from real-world communication of risk and uncertainty?" David Spiegelhalter and Jenny Gage, University of Cambridge. *Proceedings of the Ninth International Conference on Teaching Statistics* (ICOTS9, July, 2014). We're not the only fans of the picture: it won an award for excellence in scientific communication in 2014 from the UK Association of Medical Research Charities.

(4) Let $P(A, B)$ be the *joint probability* that both A and B happen. Then the conditional probability $P(A \mid B)$ is:

$$P(A \mid B) = \frac{P(A, B)}{P(B)}. \tag{10.1}$$

An equivalent way of expressing Rule 4 is to multiply both sides of the equation by $P(B)$, to yield

$$P(A, B) = P(A \mid B) \cdot P(B).$$

We can use these two versions interchangeably.

To illustrate these rules, we'll turn to Figure 10.1, which is is the brainchild of David Spiegelhalter and Jenny Gage of the University of Cambridge. These researchers asked themselves the question: how can we present the evidence on the benefits and risks of

screening in a way that doesn't make an explicit recommendation, but that helps people reach their own conclusion? The result of their efforts was a series of *probability trees* like Figure 10.1, each one depicting the likely experiences of women with and without screening.

This particular figure tracks what we'd expect to happen to two hypothetical cohorts of 200 women, aged 50 to 70. In the cohort of 200 on the left, all women are screened; while in the cohort of 200 on the right, none are screened. The expected results for each cohort are slightly different: on the right, we expect 1 fewer death, and 3 extra unnecessary screenings, versus the left.

Just about every major concept in probability is represented in this picture.

Expected value. In a group of 200 women, how many would we expect to get breast cancer? Our best guess, or expected value, is about 15, regardless of whether they get screened or not.

Probability. How likely is breast cancer for a typical woman? Fifteen cases of cancer in a cohort of 200 women means that an average woman aged 50-70 has a 7.5% chance of getting breast cancer ($15/200 = 0.075$). This is like the NP rule in reverse: if E is the expected value (here 15), then the probability is $P = E/N$.

Joint probability. Suppose that a typical woman does not go for a screening mammogram. How likely is she to get breast cancer and to die from it? In the cohort of 200 unscreened women on the right, 4 are expected to get breast cancer and die from it. Thus the risk for a typical woman is about $4/200 = 0.02$, or 2%.

Conditional probability. Suppose that a woman decides to forego screening. If she then goes on to develop breast cancer, how likely is she to die from that cancer? In the unscreened cohort, 15 women are expected to get breast cancer. Of these 15 women, 4 are expected to die from their cancer. Thus for an unscreened 50-70 year-old woman, the risk of dying from breast cancer, given that she develops breast cancer in the first place, is about 4/15, or about 27%. (Among screened women, this figure is 3/15, or 20%.)

Let's explicitly calculate this using the rule conditional probabil-

ity (Equation 10.1) instead. The rule says

$$P(\text{survives} \mid \text{gets cancer}) = \frac{P(\text{gets cancer and survives})}{P(\text{gets cancer})}.$$

We'll take this equation piece by piece.

- Out of 200 women, we expect that 15 will develop cancer. This is the denominator in our equation:

$$P(\text{gets cancer}) = \frac{15}{200}.$$

- Out of 200 women, we expect that 11 will develop cancer and survive. This is the numerator in our equation:

$$P(\text{gets cancer and survives}) = \frac{11}{200}.$$

- Therefore, using the rule for conditional probability,

$$P(\text{survives} \mid \text{cancer}) = \frac{11/200}{15/200} = 11/15.$$

Probability distributions

Now that we've learned the basic rules of probability, it's time to introduce the idea of a probability distribution.

Recall that a random variable is just a term for any uncertain outcome. We use the term *sample space* to mean the set of possible outcomes for a random variable. There are three common types of random variables, corresponding to three different types of sample spaces.

Categorical: the outcome will be one of many categories. For example, which party will win the next U.S. presidential election: Democrats, Republicans, or Other? Will your next interaction with customer service be Good, Fair, or Unrepeatable?

Discrete: the possible outcomes are whole numbers (1, 2, 3, etc.). Most of the examples we saw in our discussion of everyday risks—numbers of shark attacks, falls in the shower, and so forth—were discrete random variables.

Continuous: the random variable could be anything within a continuous range of numbers, like the price of Apple stock tomorrow, or the size of subsurface oil reservoir.

Discrete and continuous random variables are sometimes grouped together and called *numerical* random variables, since the possible outcomes are all numbers.

A *probability distribution* is just a list of probabilities for all the outcomes in the sample space of a random variable. Here's a silly example that will get the idea across. Imagine that you've just pulled up to your new house after a long cross-country drive, only to discover that the movers have buggered off and left all your furniture and boxes sitting in the front yard.[11] What a mess! You decide to ask your new neighbors for some help getting your stuff indoors. Assuming your neighbors are the kindly type, how many pairs of hands might come to your aid? Let's use the letter X to denote the (unknown) size of the household next door. The table at right shows a probability distribution for X, taken from U.S. census data in 2015; you might find this easier to visualize using the barplot in Figure 10.2.

[11] This actually happened to a friend of mine. -JS

Size of household, x	Probability, $P(X = x)$
1	0.280
2	0.336
3	0.155
4	0.132
5	0.060
6	0.023
7	0.011
8	0.003

Table 10.3: Probability distribution for household size in the U.S. in 2015. There is a vanishingly small probability for a household of size 9 or higher, which is just rounded off to zero here.

This probability distribution provides a complete representation of your uncertainty in this situation. It has all the key features of any probability distribution:

1. There is a random variable, or uncertain quantity—here, the size of the household next door (X).

2. There is a *sample space*, or set of possible outcomes for the random variable—here, the numbers 1 through 8.

3. Finally, there are probabilities for each outcome in the sample space—here provided via a simple look-up table. Notice that the table uses big X to denote the random variable itself, and little x to denote the elements of the sample space.

Most probability distributions won't be this simple, but they will all require specifying these three features.

Expected value: the mathematical definition

When you knock on your neighbors' door in the hopes of getting some help with your moving fiasco, how many people should you "expect" to be living there?

The *expected value* of a probability distribution for a numerical random variable is just an average of the items in the sample space—but a weighted average, rather than an ordinary average. If you take the 8 numbers in the sample space of Figure 10.2 and

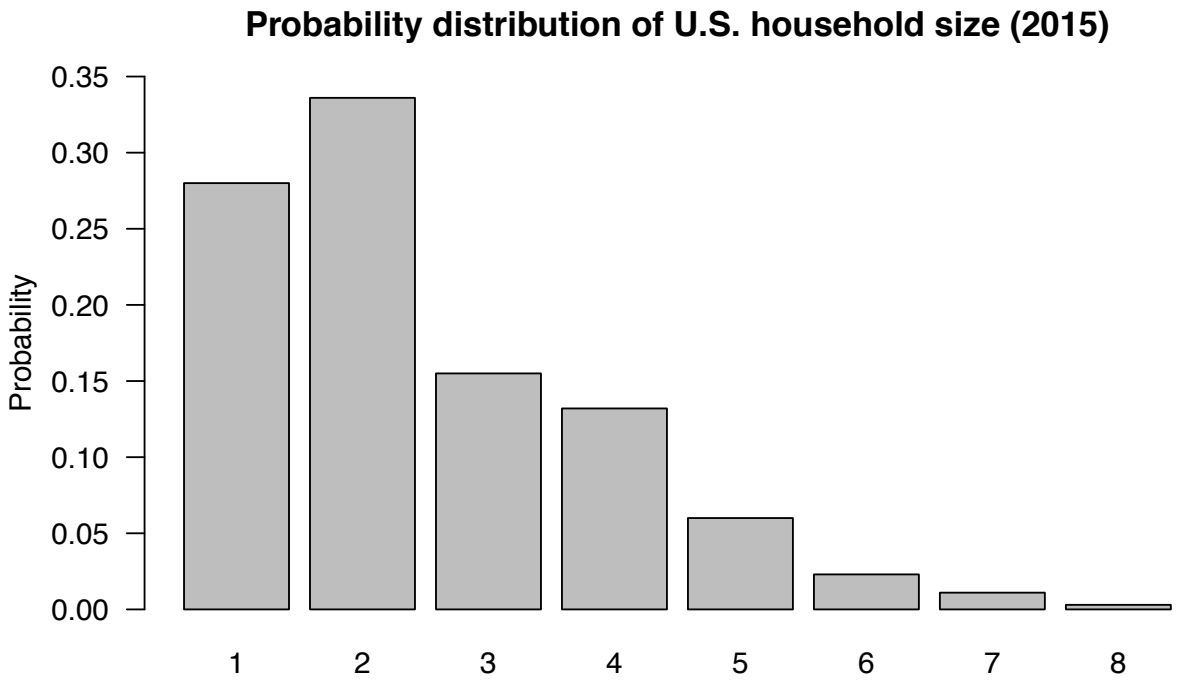


Figure 10.2: Probability distribution for the size of a random U.S. household in 2015. The elements of the sample space (the numbers $x = 1$ through $x = 8$) are shown along the horizontal axis. The probabilities $P(X = x)$ are shown on the vertical axis.

form their ordinary average, you get

$$\text{Ordinary average} = \frac{1}{8} \cdot 1 + \frac{1}{8} \cdot 2 + \cdots + \frac{1}{8} \cdot 7 + \frac{1}{8} \cdot 8 = 4.5\,.$$

Here, the weight on each number in the sample space is $1/8 = 0.125$, since there are 8 numbers. This is *not* the expected value; it give each number in the sample space an equal weight, ignoring the fact that these numbers have different probabilities.

To calculate an expected value, we instead form an average using *unequal* weights, given by the probabilities of each item in the sample space:

$$\text{Expected value} = (0.280) \cdot 1 + (0.336) \cdot 2 + \cdots + (0.011) \cdot 7 + (0.003) \cdot 8 \approx 2.5\,.$$

The more likely numbers (e.g. 1 and 2) get higher weights than $1/8$, while the unlikely numbers (e.g. 7 and 8) get lower weights.

This example conveys something important about expected values. Even if the world is black and white, an expected value is often grey. For example, the expected American household size is 2.5; a baseball player expects to get 0.25 hits per at bat; the typical person "expects" to be born with 1 testicle; and so forth.

As a general rule, suppose that the possible outcomes for a random variable X are the numbers $x_1, \ldots, x_N$. The formal definition for the expected value of X is

$$E(X) = \sum_{i=1}^{N} P(X = x_i) \cdot x_i \,. \tag{10.2}$$

This measures the "center" or mean of the probability distribution. Later, we'll learn how this more formal definition of expected value can be reconciled with the NP rule—that is, with our previous understanding of expected value as a risk/frequency calculation.

A related concept is the *variance*, which measures the dispersion or spread of a probability distribution. It is the expected (squared) deviation from the mean, or

$$\text{var}(X) = \text{E}\big(\{X - \text{E}(X)\}^2\big) \,.$$

The standard deviation of a probability distribution is $\sigma = \text{sd}(X) = \sqrt{\text{var}(X)}$. The standard deviation is more interpretable than the variance, because it has the same units (dollars, miles, etc.) as the random variable itself.

11
Conditional probability

In probability, as with many things in life, the real skill is in learning to ask the right question in the first place. As we'll discover, "asking the right question" usually means focusing on the right conditional probability.

Conditional probability: the art of asking the right question

During World War II, the size of the Allied air campaign over Europe was truly staggering. Every morning, huge squadrons of B-17 Flying Fortress bombers, each with a crew of 10 men, would take off from their air bases in the south of England, to make their way across the Channel and onwards to their targets in Germany. By 1943, they were dropping nearly 1 million pounds of bombs per week. At its peak strength, in 1944, the U.S. Army Air Forces (AAF) had 80,000 aircraft and 2.6 million people—4% of the U.S. male population—in service.

As the air campaign escalated, so too did the losses. In 1942, the AAF lost 1,727 planes; in 1943, 6,619; and in 1944, 20,394. And the bad days were very bad. In a single mission over Germany in August of 1943, 376 B-17 bombers were dispatched from 16 different air bases in the south of England, in a joint bombing raid on factories in Schweinfurt and Regensburg. Only 316 planes came back—a daily loss rate of 16%. Some units were devastated; the 381st Bomb Group, flying out of RAF Ridgewell, lost 9 of its 20 bombers that day.[1]

Like Yossarian in *Catch-22*, World War II airmen were painfully aware that each combat mission was a role of the dice. What's more, they had to complete 25 missions to be sent home. With such poor chances of returning from a *single* mission, they could be forgiven for thinking that they'd been sent to England to die.

But in the face of these bleak odds, the crews of the B-17s had at

[1] Numbers taken from *Statistical Abstract of the United States*, U.S. Census Bureau, (1944, 1947, 1950); and the Army Air Forces Statistical Digest (World War II), available at archive.org.

least three major defenses.

1. Their own tail and turret gunners, to defend the plane below and from the rear.

2. Their fighter escorts: the squadrons of P-47 Thunderbolts, RAF Spitfires, and P-51 Mustangs sent along to protect the bombers from the Luftwaffe.

3. A Hungarian-American statistician named Abraham Wald.

Abraham Wald never shot down a Messerschmitt or even saw the inside of a combat aircraft. Nonetheless, he made an outsized contribution to the Allied war effort, and no doubt saved the lives of many American bomber crews, using an equally potent weapon: conditional probability.

Figure 11.1: Abraham Wald.

Where should the military reinforce its planes?

Abraham Wald was born in 1902 in Austria-Hungary, where he went on to earn a Ph.D. in mathematics from the University of Vienna. Wald was Jewish, and when the Nazis invaded in 1938, he—like so many brilliant European mathematicians and scientists of that era—fled to America.

Wald soon went to work as part of the Applied Mathematics Panel, which had been convened by order of President Roosevelt to function as something of a mathematical tech-support hotline for the U.S. military. It was during these years of service to his adopted country that Wald prevented the military brass from making a major blunder, thereby saving many lives.

Here's the problem Wald analyzed.[2] While some airplanes came back from bombing missions in Germany unscathed, many others had visibly taken hits from enemy fire. In fact, someone examining the planes just after they landed would likely have found bullet holes and flak damage everywhere: on the fuselage, across the wings, on the engine block, and sometimes even near the cockpit.

At some point, a clever person, whose identity is lost to history, had the idea of analyzing the distribution of these hits over the surface of the returning planes. The thinking was that, if you could find patterns in where the B-17s were taking enemy fire, you could figure out where to reinforce them with extra armor, to improve survivability. (You couldn't reinforce them everywhere, or they would be too heavy to fly.)

[2] Distilled from: Mangel and Samaniego, "Abraham Wald's work on aircraft survivability." *Journal of the American Statistical Association* 79 (386): 259-67.

Researchers at the Center for Naval Analyses took this idea and ran with it. They examined data on hundreds of damaged airplanes that had returned from bombing runs in Germany. They found a very striking pattern[3] in where the planes had taken enemy fire. It looked something like this:

Location	Number of planes
Engine	53
Cockpit area	65
Fuel system	96
Wings, fuselage, etc.	434

[3] Alas, the actual data used in the original analyses cannot be located. But Wald wrote a report for the Navy on his methods, and we have attempted to simulate a data set that hews as closely as possible to the assumptions and (patchy) information that he provides in that report ("A Method of Estimating Plane Vulnerability Based on Damage of Survivors", from 1943). These and subsequent numbers are for hypothetical cohort of 800 airplanes, all taking damage.

If you turn those frequencies into probabilities, so that the numbers sum to 1, you get the following. (You should recall that a table of probabilities for the different possible outcomes of a random event, like this one, is called a probability distribution.)

Location	Probability of hit
Engine	0.08
Cockpit area	0.10
Fuel system	0.15
Wings, fuselage, etc.	0.67

Thus of all the planes that took hits and made it back to base, 67% of them had taken those hits on the wings and fuselage. In the aggregate, no other part of the plane had taken nearly as much damage. The Navy researchers reached a simple conclusion: put more armor on the wings and fuselage.

The tale of the missing bombers

But Wald pointed out that this recommendation suffered from a crucial flaw: *the Navy researchers only had data on the survivors.* The planes that had been shot down were missing from the analysis—and only the pattern of bullet holes on those missing planes could definitively tell the story of a B-17's vulnerabilities. In fact, Wald's thinking was nearly the opposite of the Navy researchers. If all those planes had safely made it home with damage to the wings and fuselage, then these areas probably weren't all that vulnerable!

Wald's argument was, essentially, that the Navy researchers were focusing on the wrong conditional probability. They had

looked at the data and concluded that the wings and fuselage were vulnerable, on the basis of the fact that

$$P(\text{hit on wings or fuselage} \mid \text{returns safely}) \approx 0.67\,.$$

But that's the right answer to the wrong question. Instead, the Navy researchers should have tried to calculate the *inverse* probability, namely

$$P(\text{returns safely} \mid \text{hit on wings or fuselage}) = ?$$

This might be a very different number. Remember: $P(\text{practices hard} \mid \text{plays in NBA}) \approx 1$, while $P(\text{plays in NBA} \mid \text{practices hard}) \approx 0$.

Of course, Wald had no data on the planes that had been shot down. Therefore, to actually calculate the probability $P(\text{returns safely} \mid \text{hit on wings or fuselage})$ required that Wald approach the data set like a forensic scientist. Essentially, he had to reconstruct the typical encounter of a B-17 with an enemy fighter, using only the mute testimony of the bullet holes on the planes that had made it back, coupled with some educated guessing.

So Wald went to work. He analyzed the likely attack angle of enemy fighters. He chatted with engineers. He studied the properties of a shrapnel cloud from a flak gun. He suggested to the army that they fire thousands of dummy bullets at a plane sitting on the tarmac. And yes, he did a lot of math.[4]

Remarkably, when all was said and done, Wald was able to reconstruct an estimate for the *joint probabilities* for the two distinct types of events that each airplane experienced: where it took a hit, and whether it returned home safely. In other words, although Wald couldn't bring the missing planes back into the air, he could bring their statistical signature back into the data set. For our hypothetical cohort of 800 bombers that took damage, Wald's best guess would have looked something like this:

[4] We don't go into detail on Wald's methods here, which were very complex. But later statisticians have taken a second look at those methods, with the hindsight provided by subsequent advances in the field. They have concluded, very simply: "Wald's treatment of these problems was definitive." (Mangel and Samaniego, *ibid.*)

	Returned	Shot down
Engine	53	57
Cockpit area	65	46
Fuel system	96	16
Wings, fuselage, etc.	434	33

Table 11.1: An example of how Abraham Wald could have reconstructed the joint frequency distribution over hit type and outcome for our hypothetical cohort of 800 planes taking enemy fire.

For example, Wald's method would have estimated that 53 of the 800 planes, or 6.6% overall, experienced the joint event (hit type

= engine, outcome = returned home safely). You'll notice that the numbers in the left column correspond exactly to the table given earlier: the pattern of hits to airplanes that made it back home. What's new is the right column: Wald's forensic reconstruction of the pattern of hits to planes that had been shot down.

This estimate for the joint frequencies for two random outcomes, hit type and outcome, now allowed Wald to answer the right question. Of the 467 planes that had taken hits to wings and fuselage, 434 of them had returned home, while 33 of them had not. Thus Wald estimated that the conditional probability of survival, given a hit to the wings and fuselage, was

$$P(\text{returns safely} \mid \text{hit on wings or fuselage}) = \frac{434}{434+33} \approx 0.93\,.$$

It turns out that B-17s were pretty robust to taking hits on the wings or fuselage.

On the other hand, of the 110 planes that had taken damage to the engine, only 53 only returned safely. Therefore

$$P(\text{returns safely} \mid \text{hit on engine}) = \frac{53}{53+57} \approx 0.48\,.$$

Similarly,

$$P(\text{returns safely} \mid \text{hit on cockpit area}) = \frac{65}{65+46} \approx 0.59\,.$$

The bombers were much more likely to get shot down if they took a hit to the engine or cockpit area.

Thus Wald's final, life-saving advice ran exactly counter to that of the experts at the Center for Naval Analyses: *put the armor where you don't see the bullet holes*.

Postscript. In the story of Abraham Wald and the missing B-17s, the path of counterintuitive facts eventually turns a full 360 degrees. Imagine asking any random person off the street: "Where should we put extra armor on airplanes to help them survive enemy fire?" We haven't done this survey, but we strongly suspect that most thoughtful people would answer: where the pilot and the engines are! But the data initially seem to suggest otherwise. This implies that we should turn 180 degrees away from our intuition: if the planes are taking damage on the wings and the fuselage, then let's put the armor there instead. It turns out that only a very careful analysis like Wald's can rehabilitate the initial, intuitive guess over the spurious attempt at data-based reasoning.

The moral of the story is that data alone isn't enough. You have to know enough about conditional probability to be able to pose the right question in the first place.

How Netflix knows your taste in movies so well

THE same math that Abraham Wald used to analyze bullet holes on B-17s also underpins the modern digital economy of films, television, music, and social media. To give one example: Netflix, Hulu, and other video-streaming services all use this same math to examine what shows their users are watching, and apply the results of their number-crunching to recommend new shows.

To see how this works, suppose that you're designing the movie-recommendation algorithm for Netflix, and you have access to the entire Netflix database, showing which customers have liked which films—for example, by assigning a film a five-star rating. Your goal is to leverage this vast data resource to make automated, personalized movie recommendations. The better these recommendations are, the more likely your customers are to keep their accounts on auto-pay.

You decide to start with an easy case: assessing how probable it is that a user will like the film *Saving Private Ryan* (event A), given that the same user has liked the HBO series *Band of Brothers* (event B). This is almost certainly a good bet: both are epic dramas about the Normandy invasion and its aftermath. Therefore, you might think: job done! Recommend away.

For this particular pair of shows, fine. But keep in mind that you want to be able to do this kind of thing automatically. It would not be cost effective to put a human in the loop here, laboriously tagging all possible pairs of movies for similar themes or content—to say nothing of all of the other stuff that might make two different films appeal to the same person.

As with Abraham Wald and the missing bombers, it's all about asking the right question. The key insight here is to frame the problem in terms of conditional probability. Suppose that, for some pair of films A and B, the probability $P(\text{likes } A \mid \text{likes } B)$ is high—say, 80%. Now we learn that Linda liked film B, but hasn't yet seen film A. Wouldn't A be a good recommendation? Based on her liking of A, there's an 80% chance she'll like it.

But how can we learn $P(\text{likes } A \mid \text{likes } B)$? This is where your

database, coupled with the rule for conditional probability, comes in handy. Suppose that there are 5 million people in your database who have seen both *Saving Private Ryan* and *Band of Brothers*, and that the ratings data on these 5 million users looks like this:

	Liked *Band of Brothers*	Didn't like
Liked *Saving Private Ryan*	2.8 million	0.3 million
Didn't like	0.7 million	1.2 million

Once again, we have a joint frequency distribution for two random outcomes: A = whether a user liked *Saving Private Ryan*, and B = whether the user liked *Band of Brothers*. From this joint distribution, we can easily work out the conditional probability that we need. Of the 5 million users in the database who have watched both programs, 2.8 + 0.7 = 3.5 million of them liked *Band of Brothers*. Of these 3.5 million people, 2.8 million (or 80%) also liked *Saving Private Ryan*. Therefore,

$$P(\text{liked } \textit{Saving Private Ryan} \mid \text{liked } \textit{Band of Brothers}) = \frac{2.8 \text{ million}}{3.5 \text{ million}} = 0.8\,.$$

Note that you could also jump straight to the math, which we outline in a few pages, and use the rule for conditional probabilities (Equation 10.1, on page 214), like this:

$$P(A \mid B) = \frac{P(A,B)}{P(B)} = \frac{2.8/5}{(2.8+0.7)/5} = 0.8\,.$$

You'd get the same answer in the end.

The key thing that makes this approach work so well is that it's automatic. Computers aren't very good (yet) at automatically scanning films for thematic content. But they're brilliant at calculating conditional probabilities from a vast database of users' movie-watching histories.

The same trick works for books, too. Suppose you examine the online book-purchase histories of two friends Albert and Pablo, and discovered the following items.

Albert: (1) Proof and Consequences. (2) A Body in Motion: Newton's Guide to Productivity. (3) Obscure Theorems of the 14th Century.

Pablo: (1) Your Face is Offside: Dora Maar at the Cubist Soccer Match. (2) A Short History of Non-representational Art. (3) Achtung, Maybe? Dali, Danger, and the Surreal.

What sorts of books are you likely to recommend to these friends for their birthdays? Amazon learned to use conditional probability to automate this process long ago, to the chagrin of independent bookstores everywhere. Similar math also underpins recommender systems for music (Spotify), ads (Google), and even friends (Facebook).

The digital economy truly is ruled by conditional probability.

The math of conditional probability

To understand the basic math behind joint, conditional, and marginal probabilities, we'll return to the story of Abraham Wald and the B-17s.

Joint probabilities

We start by turning Table 11.1, which contains counts of different joint event types for a cohort of 800 airplanes, into a table of probabilities:

	Returned	Shot down
Engine	0.066	0.071
Cockpit area	0.081	0.058
Fuel system	0.120	0.020
Wings, fuselage, etc.	0.542	0.042

This table gives summarizes the probabilities for two random outcomes: X = hit type, along the rows; and Y = outcome, along the columns. The entries in the table give the joint probabilities $P(X = x, Y = y)$. For example, 2% of all planes both took a hit in the fuel system and got shot down: $P(X = \text{fuel system}, Y = \text{shot down}) = 0.02$. Up to round-off error, these 8 probabilities all sum to 1.

We call a table like this a *joint probability distribution*: that is, a list of all possible joint events for multiple random variables, together with their joint probabilities.

Marginal probabilities

Next, we add an additional row and column of *marginal* (or overall) probabilities of the different event types and outcomes, like

in the table below. These are called the marginal probabilities because we calculate them by summing across the relevant margin (i.e. row or column) of the table.

	Returned	Shot down	Marginal
Engine	0.066	0.071	0.137
Cockpit area	0.081	0.058	0.139
Fuel system	0.120	0.020	0.140
Wings, fuselage, etc.	0.542	0.042	0.584
Marginal	0.809	0.191	1

The marginal probabilities we've calculated just reflect the fact that the probability of some event (like returning safely) is the sum of the probabilities for all the distinct ways that event can happen. For example, an airplane that takes a hit to the engine can do so in two ways: it can take the hit and return, or it can take the hit and not return. Therefore,

$$\begin{aligned} P(\text{hit to engine}) &= P(\text{returned}, \text{hit to engine}) + P(\text{shot down}, \text{hit to engine}) \\ &= 0.066 + 0.071 = 0.137\,. \end{aligned}$$

The rest of the marginal probabilities are calculated similarly, e.g.

$$\begin{aligned} P(\text{returned}) &= 0.066 + 0.081 + 0.120 + 0.542 \\ &= 0.809\,. \end{aligned}$$

Conditional probabilities

Finally, we are ready to understand the rule for conditional probabilities. You'll recall that this was the fourth of the basic rules of probability quoted earlier. It goes like this:

$$P(A \mid B) = \frac{P(A, B)}{P(B)}\,.$$

Remember how we used Table 11.1 to calculate $P(\text{returns} \mid \text{hit to engine})$? We looked at the total number of planes that had taken a hit to the engine. We then asked: of these planes, how many also returned home safely? As an equation, this gives us

$$\begin{aligned} P(\text{returns} \mid \text{engine hit}) &= \frac{\text{Number taking engine hit and returned safely}}{\text{Number taking engine hit}} \\ &= \frac{53}{110} \approx 0.48\,. \end{aligned}$$

You'll notice we get the exact same answer if we use the rule for conditional probabilities: $P(A \mid B) = P(A, B)/P(B)$. These probabilities are estimated using the relevant fractions from the data set:

$$\begin{aligned} P(\text{returns} \mid \text{engine hit}) &= \frac{\text{Fraction taking engine hit and returning safely}}{\text{Fraction taking engine hit}} \\ &= \frac{53/800}{110/800} \\ &= \frac{0.066}{0.137} \approx 0.48\,. \end{aligned}$$

While the rule for conditional probabilities may look a bit intimidating, it just codifies exactly the same intuition we used to calculate $P(\text{returns} \mid \text{engine hit})$ from the table of counts.

The rule of total probability

CONSIDER the following data on obstetricians delivering babies at a hospital in England. The table below shows the complication rates for both junior and senior doctors on the delivery ward, grouped by delivery type:

	Easier deliveries	Harder deliveries	Overall
Senior doctors	0.052 (231)	0.127 (102)	0.076 (315)
Junior doctors	0.067 (3169)	0.155 (206)	0.072 (3375)

The numbers in parentheses are the total deliveries of each type.

This table exhibits an aggregation paradox.[5] No matter what kind of delivery you have, whether easy or hard, you'd prefer to have a senior doctor. They have lower complication rates than junior doctors in both cases. Yet counterintuitively, the senior doctors have a higher overall complication rate: 7.6% versus 7.2%. Why? Because of a lurking variable: most of the deliveries performed by junior doctors are easier cases, where complication rates are lower overall. The senior doctors, meanwhile, work a much higher fraction of the harder cases. Their overall complication rate reflects this burden.

[5] Also called *Simpson's paradox*.

Here's another example. Jacoby Ellsbury and Mike Lowell were two baseball players for the Boston Red Sox during the 2007 and 2008 seasons. The table below shows their batting averages for

those two seasons, with their number of at-bats in parentheses. We see that Ellsbury had a higher batting average when he was a rookie, in 2007; a higher batting average a year later, in 2008; but a lower batting average overal!

	2007	2008	Overall
Lowell	.324 (589)	.274 (419)	.304 (1008)
Ellsbury	.353 (116)	.280 (554)	.293 (670)

Again we have an aggregation paradox, and again it is resolved by pointing to a lurking variable: in 2007, when both players had higher averages, Ellsbury had many fewer at-bats than Lowell.

It turns out the math of these aggregation paradoxes can be understood a lot more deeply in terms of something called the *rule of total probability*, or the *mixture rule*. This rule sounds impressive, but is actually quite simple. It says: the probability of any event is the sum of the probabilities for all the different ways in which the event can happen. In that sense, the law of total probability is really just Kolmogorov's third rule in disguise. The distinct ways in which some event A can happen are mutually exclusive. Therefore we just sum all their probabilities together to get $P(A)$.

Let's return to the example on obstetric complication rates on junior doctors at a hospital in England. In the table, there are two ways of having a complication: with an easy case, or with a hard case. Therefore, the total probability is the sum of two joint probabilities:

$$P(\text{complication}) = P(\text{easy and complication}) + P(\text{hard and complication}) .$$

If we now apply the rule for conditional probabilities (Equation 10.1) to each of the two joint probabilities on the right-hand side of this equation, we have this:

$$P(\text{complication}) = P(\text{easy}) \cdot P(\text{complication} \mid \text{easy}) + P(\text{hard}) \cdot P(\text{complication} \mid \text{hard}) .$$

Thus the rule of total probability says that overall probability is a weighted average—a mixture—of the two conditional probabilities.

So which probabilities should we report: the conditional probabilities, or the overall (total) probabilities? There's no one right answer; it depends on your conditioning variable, and your goals. In the obstetric data, the overall complication rates are clearly misleading. The distinction between easier and harder cases matters

a lot. Senior doctors work harder cases, on average, and therefore have higher overall complication rates. But what matters to the patient, and to anyone who assesses the doctors' performance, are the *conditional* rates. You have to account for the lurking variable.

The baseball data is different. Here the *conditional* probabilities for 2007 and 2008 are probably misleading. The distinction between 2007 and 2008 is nothing more than an arbitrary cutoff on the calendar. It's barely relevant from the standpoint of assessing baseball skill, and it needlessly splits one big sample of each player's history into two smaller, more variable samples. So in this case we'd probably go with the overall averages if we wanted to say which player was performing better.

A formal statement of the rule of total probability. Suppose that events $B_1, B_2, \ldots, B_N$ constitute an exhaustive partition of all possibilities in some situation. That is, the events themselves are mutually exclusive, but one of them must happen. This can be expressed mathematically as

$$P(B_i, B_j) = 0 \text{ for any } i \neq j, \quad \text{and} \quad \sum_{i=1}^{N} P(B_i) = 1\,. \tag{11.1}$$

Now consider any event A. If Equation 11.1 holds, then

$$\begin{aligned} P(A) &= \sum_{i=1}^{N} P(A, B_i) \\ &= \sum_{i=1}^{N} P(B_i) \cdot P(A \mid B_i)\,. \end{aligned} \tag{11.2}$$

Equation 11.2 is what is usually called the rule of total probability.

Surveys and the rule of total probability

One of the least surprising headlines of 2010 must surely have been the following, from the ABC News website:

Teens not always honest about drug use.[6]

[6] Kim Carollo, ABC News, Oct. 25, 2010. Link here.

In other news, dog bites man.

To be fair, the story itself was a bit more surprising than the headline. Yes, it's hardly news that teenagers would lie to their parents, teachers, coaches, and priests about drug use. But the ABC News story was actually reporting on a study showing that

teenagers also lie to researchers who conduct anonymous surveys about drug use—even when those teenagers know that their answers will be verified using a drug test.

Here's the gist of the study. Virginia Delaney-Black and her colleagues at Wayne State University, in Detroit, gave an anonymous survey to 432 teenagers, asking whether they had used various illegal drugs.[7] Of these 432 teens, 211 of them also agreed to give a hair sample. Therefore, for these 211 respondents, the researchers could compare people's answers with an actual drug test.

[7] V. Delaney–Black et. al. "Just Say 'I Don't': Lack of Concordance Between Teen Report and Biological Measures of Drug Use." *Pediatrics* 165:5, pp. 887-93 (2010).

The two sets of results were strikingly different. For example, of the 211 teens who provided a hair sample, only a tiny fraction of them (0.7%) admitted to having used cocaine. However, when the hair samples were analyzed in the lab, 69 of them (33.7%) came back positive for cocaine use.

And it wasn't just the teens who lied. The survey researchers also asked the *parents* of the teens whether they themselves had used cocaine. Only 6.1% said yes, but 28.3% of the hair samples came back positive.

Let's emphasize again that we're talking about a group of people who were guaranteed anonymity, who wouldn't be arrested or fired for saying yes, and who willingly agreed to provide a hair sample that they knew could be used to verify their survey answers. Yet a big fraction lied about their drug use anyway.

Surveys and lies

Drug abuse—whether it's crack cocaine in Detroit, or bathtub speed in rural Nebraska—is a huge social problem. It fills our jails, drains public finances, and perpetuates a trans-generational cycle of poverty. Getting good data on this problem is important. As it stands, pediatricians, schools, and governments all rely on self-reported measures of drug use to guide their thinking on this issue. Yet distressingly, the proportion of self-reported cocaine use in the Detroit study, 0.7%, was broadly similar to the findings in large, highly regarded surveys—for example, the federally funded National Survey on Drug Use and Health. The work of Dr. Delaney-Black and her colleagues would seem to imply that all of these self-reported figures might be way off the mark.

Moreover, theirs hasn't been the only study to uncover evidence that surveys cannot necessarily be taken at face value. Here are some other things that, according to research *on* surveys, people lie about *in* surveys.

- Churchgoers overstate the amount of money they give when the hat gets passed around during the service.
- Gang members embellish the number of violent encounters they have been in.
- Men exaggerate their salary, among other things.
- Ravers will "confess" to having gotten high on drugs that do not actually exist.

How to ask an embarrassing question: probability as an invisibility cloak

But there's actually some good news to be found here. It's this: when people lie in surveys, they tend to do so for predictable reasons (to impress someone or avoid embarrassment), and in predictable ways (higher salary, fewer warts). This opens the door for survey designers to use a bit of probability, and a bit of psychology, to get at the truth—even in a world of liars.

Let's go back to the example of drug-use surveys so that we can see this idea play out. Suppose that you want to learn about the prevalence of drug use among college students. You decide to conduct a survey at a large state university to find out how many of the students there have smoked marijuana in the last year. But as you now appreciate, if you ask people direct questions about drugs, you can't always trust their answers.

Here's a cute trick for alleviating this problem, in a way that uses probability theory to mitigate someone's psychological incentive to lie. Suppose that, instead of asking people point-blank about marijuana, you give them these instructions.

1. Flip a coin. Look at the result, but keep it private.
2. If the coin comes up heads, please use the space provided to write an answer to question Q1: "Is the last digit of your Social Security number odd?"
3. If the coin comes up tails, please use the space provided to write an answer to question Q2: "Have you smoked marijuana in the last year?"

The key fact here is that only the respondent knows which question he or she is answering. This gives people plausible deniability. Someone answering "yes" might have easily flipped heads and

answered the first, innocuous question rather than the second, embarrassing one, and the designer of the survey would never know the difference. This reduces the incentive to lie.

Moreover, despite the partial invisibility cloak we've provided to the marijuana users in our sample, we can still use the results of the survey to answer the question we care about: what fraction of students have used marijuana in the past year? We'll use the following notation:

- Let Y be the event "a randomly chosen student answers yes."
- Let Q_1 be the event "the student provided an answer to question 1, about their Social Security number."
- Let Q_2 be the event "the student provided an answer to question 2, about their marijuana use."

From the survey, we have an estimate of $P(Y)$, which is the overall fraction of survey respondents providing a "yes" answer. We really want to know $P(Y \mid Q_2)$, the probability that a randomly chosen student will answer "yes", given that he or she was answering the marijuana question. The problem is that we don't know which students were answering the marijuana question.

To understand the rule of total probability, let's return to our hypothetical survey in which we want to know the answer to the question: what fraction of students have used marijuana in the past year? Then we have each survey respondent privately flip a coin to determine whether they answer an innocurous question (Q1) or the question about marijuana use (Q2). We used the following notation:

- Let Y be the event "a randomly chosen student answers yes."
- Let Q_1 be the event "the student provided an answer to question 1, about their Social Security number."
- Let Q_2 be the event "the student provided an answer to question 2, about their marijuana use."

To solve this problem, we'll use rule of total probability. In the case of our drug-use survey, this means that

$$P(Y) = P(Y, Q_1) + P(Y, Q_2)\,. \tag{11.3}$$

In words, this equation says that there are two ways to get a yes answer: from someone answering the social-security-number question, and from someone answering the drugs question. The total

number of yes answers will be the sum of the yes answers from both types in this mixture.

Now let's re-write Equation 11.3 slightly, by applying the rule for conditional probabilities to each of the two joint probabilities on the right-hand side of this equation:

$$P(Y) = P(Q_1) \cdot P(Y \mid Q_1) + P(Q_2) \cdot P(Y \mid Q_2)\,. \tag{11.4}$$

This equation now says that the overall probability $P(Y)$ is a weighted average of two conditional probabilities:

- $P(Y \mid Q_1)$, the probability that a randomly chosen student will answer "yes", given that he or she was answering the social-security-number question.
- $P(Y \mid Q_2)$, the probability that a randomly chosen student will answer "yes", given that he or she was answering the marijuana question.

The weights in this average are the probabilities for each question: $P(Q_1)$ and $P(Q_2)$, respectively.

Now we're ready to use Equation 11.4 to calculate the probability that we care about: $P(Y \mid Q_2)$. We know that $P(Q_1) = P(Q_2) = 0.5$, since a coin flip was used to determine whether Q_1 or Q_2 was answered. Moreover, we also know that $P(Y \mid Q_1) = 0.5$, since it is equally likely that someone's Social Security number will end in an even or odd digit.[8]

We can use this information to simplify the equation above:

$$P(Y) = 0.5 \cdot 0.5 + 0.5 \cdot P(Y \mid Q_2)\,,$$

or equivalently,

$$P(Y \mid Q_2) = 2 \cdot \{P(Y) - 0.25\}\,.$$

Suppose, for example, that 35% of survey respondents answer yes, so that $P(Y) = 0.35$. This implies that

$$P(Y \mid Q_2) = 2 \cdot (0.35 - 0.25) = 0.2\,.$$

We would therefore estimate that about 20% of students have smoked marijuana in the last year.

[8] This survey design relies upon the fact that the survey designer doesn't know anyone's Social Security number. If you were running this survey in a large company, where people's SSNs were actually on file, you'd need to come up with some other innocuous question whose answer was unknown to the employer, but for which $P(Y \mid Q_1)$ was known.

12
Bayes' rule

Our conditional probabilities always depend on what we know. When our knowledge changes, these probabilities must also change. A 250-year-old mathematical principle called Bayes' Rule tells us how.

Bayesian search: finding the USS *Scorpion*

In February of 1968, the USS *Scorpion* set sail from the naval base in Norfolk, Virginia, under the command of Francis Slattery. The *Scorpion* was a Skipjack-class high-speed attack submarine, the fastest in the American fleet. Like other subs of her class, she played a major role in U.S. military strategy. Think *The Hunt for Red October* here: throughout the Cold War, both the Americans and the Soviets deployed large fleets of attack subs, whose mission was to locate, track, and—should the unthinkable happen—destroy the other side's ballistic-missile submarines.

On this deployment, the *Scorpion* sailed east, bound for the Mediterranean Sea, where for three months she participated in training exercises alongside the 6th Naval Fleet. Then in mid-May, the *Scorpion* was sent back west, past Gibraltar and out into the Atlantic. There she was ordered to observe Soviet naval vessels operating near the Azores—a remote island chain in the middle of the North Atlantic, about 850 miles off the coast of Portugal—and then to continue west, bound for home. The sub was due back in Norfolk at 1 PM on Monday, May 27th, 1968.

On the docks in Norfolk that day, the families of the *Scorpion*'s 99 crew members were gathered to welcome their loved ones back home. But as 1 PM came and went, the sub had not yet surfaced. Minutes stretched into hours; day gave way to night. Still the families waited. But there was no sign of the *Scorpion*.

With growing alarm, the Navy ordered a search. By 10 PM, the

operation involved 18 ships; by the next morning, 37 ships and 17 long-range patrol aircraft. But the odds of a good outcome were slim. The *Scorpion* had last made contact off the Azores, 6 days ago, and 2,670 miles away from Norfolk. She could have been anywhere along that strip of ocean between the Azores and the eastern seaboard. As the hours ticked by, the chances that the sub could be located, and that rescue gear could be deployed in time, were rapidly diminishing. At a tense news conference on May 28th, President Lyndon Johnson summarized the mood of a nation: "Nothing encouraging to report. . . . We are all quite distressed."

Day after day went by, but the search for the *Scorpion* turned up no results. Finally, after eight days, the Navy was forced to concede the obvious: the *Scorpion*'s crew of 99 men were declared lost at sea, presumed dead.

The Navy now turned to the grim task of locating the *Scorpion*'s final resting place—a tiny needle in a vast haystack stretching three-fourths of the way across the North Atlantic. Although hopes for saving the crew had been dashed, the stakes were still high, and not only for the families of those lost: the *Scorpion* had carried two nuclear-tipped torpedoes, each capable of sinking an aircraft carrier with a single hit. These dangerous warheads were now somewhere on the bottom of the sea.

John Craven, Bayesian search guru

To lead the search for the *Scorpion*, the Pentagon turned to Dr. John Craven, chief scientist in the Navy's Special Projects Office, and the leading guru on finding lost objects in deep water.

Remarkably, Dr. Craven had done this kind of thing before. Two years earlier, in 1966, an American B-52 bomber had collided in mid-air with a refueling tanker over the Spanish coast, near the seaside village of Palomares. Both planes crashed, and the B-52's four hydrogen bombs, each of them 50 times more powerful than the Hiroshima explosion, were scattered for miles. Luckily none of the warheads had detonated, and three of the bombs were found more or less immediately.[1] But the fourth bomb was missing, and was presumed to have fallen into the sea. John Craven was called upon to help find it.

Craven and his team had to ponder many unknown variables about the crash. Had the bomb remained in the plane, or had it fallen out? If the bomb had fallen out, had either or both of its parachutes deployed? If the parachutes had deployed, had the

[1] Albeit after one of them had contaminated a roughly one-square-mile area of tomato farms and woodland with radioactive plutonium. The clean-up operation in the wake of this incident continues 50 years later, with the latest negotiations between Spain and the U.S. taking place in 2015.

winds taken the bomb far out to sea? If so, in what direction, and exactly how far?

Bayesian search. To sort through this thicket of unknowns, Craven turned to his preferred strategy: *Bayesian search*. This search methodology had been pioneered during World War II, when the Allies used it to help locate German U-boats. But its origins stretched back much further, all the way to a mathematical principle called *Bayes' rule*, first worked out by an English reverend named Thomas Bayes, in the 1750s.

Bayesian search has three essential principles. First, you should combine the pre-search opinions of various experts about the plausibility of each possible scenario. In the case of the missing H-bomb, some of these experts would be familiar with mid-air crashes, some of them familiar with nuclear bombs, some with coastal winds and ocean currents, and so forth. These opinions should be synthesized to form a *prior probability* for each crash scenario—and, by extension, a prior probability that the bomb might be found in each possible search location. These probabilities are "prior," in the sense that they represent the best guess available, before anyone has any data.

Second, you must evaluate the capability of your search instruments to establish the likelihood that, if the object were in a given sector, you'd actually be able to find it there. This likelihood is combined with the prior probability to form a single search-effectiveness probability for each location. For example, let's say that the most plausible scenario puts the lost bomb at the bottom of a very deep ocean trench. Despite its high prior probability, this trench might still be a poor candidate location to begin your search, for the simple reason that the trench is so dark and remote that, even if the bomb were there, you'd be very unlikely to find it. To draw on a familiar metaphor, a Bayesian search has you start looking for your lost keys using a precise mathematical combination of two factors: where you think you lost them, and where the streetlight is shining brightest.

Third and finally, as new data comes in during the search process, you should use that new data to update your prior probability for each search location into a *posterior probability*. This Bayesian updating process is iterative, in the sense that today's posterior becomes tomorrow's prior. Suppose you search in today's region of high posterior probability, but find nothing. Then for tomorrow,

you reduce the probability in the region you just searched, reassess your beliefs about each scenario, and bump up the probability in the other regions accordingly. You keep doing this day after day, always concentrating on that day's new region of highest probability, until you find what you're looking for.

Craven is stymied. Unfortunately, military politics, and a clash of personalities, prevented Dr. Craven and his team from actually applying these Bayesian principles to the 1966 search for the missing H-bomb off the coast of Palomares. In a classic military move, the Pentagon had asked the right hand to do one thing, and then asked the left hand to put some handcuffs on the right one, to make its job more difficult. The commanding officer on the scene, Rear Admiral William S. Guest—nickname: Bull Dog—had a notably different view of the way the search should be conducted. Bull Dog was an excellent fast thinker. He had little patience for probabilities, and even less patience for the team of twentysomething-year-old math Ph.D's he now found himself commanding. His initial orders to Craven's team, perhaps only half sarcastic, were for them to prove that the bomb had fallen on land rather than in the sea, so that it would be someone else's job to find the damned thing.[2]

As a result, the search for the Palomares H-bomb was really two searches. There was Craven's Bayesian search, with its slide rules and probability maps, and with updated probabilities constantly chattering over the teletype machine as the mathematicians fed remote calculations to a mainframe computer back in Pennsylvania. But the insights arising from the Bayesian search were largely ignored in favor of Admiral Guest's "plan of squares," which guided the *real* search, and which was pretty much exactly what it sounds like. The frustrated Craven was like a high-school stock picker who records virtual trades in a ledger and watches his paper fortune grow, but never gets to buy and sell any shares.

Eventually, the bomb was found. It turns out that a fisherman named Francisco Orts had seen the bomb fall into the water under parachute, and he was able to guide the Navy to its exact point of entry. Thus while the search was a success, the Bayesian part of it had been a failure, for the simple reason that it had never been given a chance. Nonetheless, the Palomares incident taught John Craven some valuable lessons—both about the practicalities of conducting a Bayesian search, and about how to how to get the

[2] Sharon Bertsch McGrayne, *The Theory That Would Not Die*, Yale University Press, 2011 (pg. 190).

necessary support for that search from the military brass.

And two years later, when he was called upon to find the USS *Scorpion*, Craven was ready.

The search for the Scorpion continues

When the *Scorpion* disappeared in May of 1968, Craven and his Bayesian search team—including Frank Andrews, Daniel Wagner, Tony Richardson, and many others—were quickly reconvened. At first, the task seemed vastly more daunting than the search for the Palomares H-bomb had been. Back then, they had known to confine the search to a relatively small area off the coast of southern Spain. But here, the team had to find a submarine under 2 miles of water, somewhere between Virginia and the Azores, without so much as a single clue.

Luckily, they caught a break. Starting in the early 1960s, the U.S. military had spent $17 billion installing an enormous, highly classified network of underwater microphones throughout the North Atlantic. Essentially, they had wired the entire ocean for sound, so that they could track the movements of the Soviet navy. Highly trained technicians at secret listening posts were monitoring these microphones around the clock. The technicians could look at the output from these devices and immediately distinguish the acoustic signature of a submarine from that of a whale, an oil tanker, or hot magma under the seabed.

After sniffing around, Craven discovered that one of these secret listening posts in the Canary Islands had, one day in late May, recorded a very unusual series of 18 underwater sounds. Then he learned that two other listening posts—both of them thousands of miles away, off the coast of Newfoundland—had recorded those very same sounds around the same time. Craven's team compared these three readings and, by triangulation, worked out that the sounds must have emanated from a very deep part of the Atlantic Ocean, about 400 miles southwest of the Azores.

This location fell along the *Scorpion*'s expected route home. Moreover, the sounds themselves were highly suggestive: a muffled underwater explosion; then 91 seconds of silence; and then 17 further sonic events in rapid succession that, to Craven, sounded like the implosion of various compartments of a submarine as it sank beneath its hull-crush depth.[3]

This acoustic revelation dramatically narrowed the size of the search area. Still, the team had about 140 square miles of ocean

[3] PBS Nova documentary, "Submarines, Secrets, and Spies." Originally broadcast January 19, 1999. https://www.youtube.com/watch?v=NJWHiPSvzh8

floor to cover, all of it 10,000 feet below the surface, and therefore inaccessible to all but the most advanced submersibles.

The Bayesian search now kicked into high gear. Craven's first step was to take a map of the seabed and divide it up into a grid of little rectangles, each one a possible search location. Each rectangle got an alphanumeric code—B6, H3, and so on, just like in the board game *Battleship*. Craven and his team then interviewed expert submariners, and came up with nine possible scenarios—a fire on board, a torpedo exploding in its bay, a clandestine Russian attack, and so on—for how the submarine had sunk. They weighed the prior probability of each scenario, and ran Monte Carlo simulations to understand how the *Scorpion*'s likely movements might have unfolded under each one. They assessed the capabilities of the search fleet: its cameras, its magnetic-sensing instruments, its sonars, its submersible robots. They even blew up depth charges at precise locations, in order to calibrate their original acoustical data from the listening posts in the Canary Islands and Newfoundland.

Finally, they put all this information together to form a single search-effectiveness probability for each cell on the grid. This map crystallized thousands upon thousands of hours of interviews, calculations, experiments, and careful thinking. It would have looked something like this:

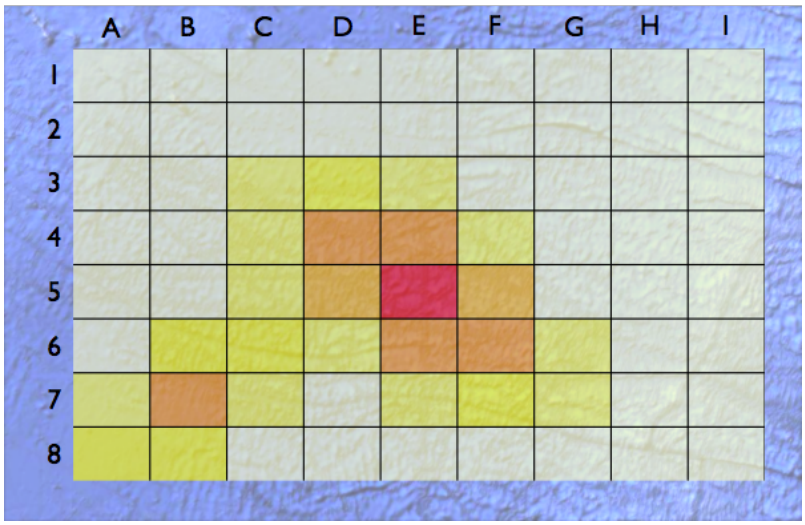


Figure 12.1: An attempted reconstruction of John Craven's probability map for the *Scorpion* search. Darker red rectangles indicate regions of relatively higher prior probability.

Mathematically speaking, this map represented the best chance for

finding the *Scorpion*.

Predictably, Craven encountered both logistical and bureacratic difficulties in getting the Pentagon to pay attention to his map of probabilities. Summer came and went. By this point, the search for the *Scorpion* had been going on since early June, to no avail.

But eventually his cajoling paid off, and the military brass ordered that Craven's map be used to guide the now months-old search. So starting in October, when commanders leading the search aboard the USS *Mizar* finally got ahold the map, the operation became truly Bayesian. Day by day, the team rigorously searched the region of highest probability, and crunched the numbers to update the map for tomorrow. And day by day, those numbers were slowly homing in on rectangle F6.

Found

On October 28th, Bayes finally paid off.

The *Mizar* was in the midst of its 5th cruise, and its 74th individual run over the ocean floor. All of a sudden the ship's magnetometer spiked, suggesting an anomaly on the sea floor. Cameras were hurriedly deployed to investigate—and sure enough, there she was:

Figure 12.2: A photo of the bow section of the USS *Scorpion*, taken in 1968 by the crew of the bathyscape *Trieste II*. USN photo #1136658.

Partially buried in the sand, 400 miles from the nearest landfall and two miles below the surface of a restive sea, the USS *Scorpion*

had been found at last.

To this day, nobody knows for sure what actually happened to the *Scorpion*—or if they do, they're not talking. The Navy's official version of events, though inconclusive, cites the accidental explosion of a torpedo or the malfunctioning of a garbage-disposal unit as two of the most likely possible causes of the tragedy. Many other explanations have been proposed over the years. And as with any famous mystery, conspiracy theories abound.

But there was at least one definitive conclusion to come out of the *Scorpion* incident: Bayesian search was a truly winning idea. As it turned out, the sub's final resting place lay a mere 260 yards away from rectangle E5, the initial region of highest promise on Craven's map of prior probabilities. The search team had actually passed over that location on a previous cruise, but had missed the telltale signs due to a broken sonar.[4]

[4] McGrayne, ibid.

Ponder that for a moment more. A lone submarine had been lost somewhere in a 2600-mile stretch of open ocean, and the Bayesian search had pinpointed her location to within 260 yards—only three lengths of the submarine itself. It was a remarkable triumph for Craven's team, and for Bayes' rule, the 250-year-old mathematical formula that had served as the search's guiding principle.

Today, Bayesian search is a small industry, with at least one college textbook[5] explaining the details, and with entire companies whose mission is to apply Bayesian principles to find what has been lost. To cite a recent example, many readers will remember the tragedy of Air France Flight 447, which crashed in the Atlantic Ocean on its way from Rio de Janeiro to Paris, in June of 2009. The search for the wreckage had been going on for two fruitless years; then in 2011, a Bayesian search firm was hired, a map of probabilities was drawn up—and the plane was found within one week of undersea search.[6]

[5] The Theory of Optimal Search (Operations Research Society of America, 1975), by Lawrence D. Stone.

[6] Stone et. al. "Search for the wreckage of Air France Flight AF 447." *Statistical Science* 2014, Vol. 29(1), pp. 69-80.

Moreover, the broader principle behind Bayesian search, Bayes' rule, is used almost everywhere: from courtrooms to doctor's offices, and from spam filters to self-driving cars. So if you want to learn more about the key equation that found the *Scorpion* and that helps power the modern world, then this chapter is for you.

Updating conditional probabilities

OUR probabilities are always contingent upon what we know.

The probability that a patient with chest pains has suffered a heart attack: Does the patient feel the pain radiating down his left side? What does his ECG look like? Does his blood test reveal elevated levels of myoglobin?

The probability of rain this afternoon in Milwaukee: What are the current temperature and barometric pressure? What does the radar show? Was it raining this morning in Chicago?

The probability that a person on trial is actually guilty: Did the accused have a motive? Means? Opportunity? Were any bloody gloves left at the scene that reveal a likely DNA match?

When our knowledge changes, our probabilities must change, too. Bayes' rule, the mathematical formula at the heart of the *Scorpion* story, tells us how to change them.

Imagine the person in charge of a Toyota factory who starts with a subjective probability assessment for some proposition A, like "our engine assembly robots are functioning properly." Just to put a number on it, let's say $P(A) = 0.95$; we might have arrived at this judgment, for example, based on the fact that the robots have been down for 5% of the time over the previous month. In the absence of any other information, this is as good a guess as any.

Now we learn something new, like information B: the last 5 engines off the assembly line all failed inspection. Before we believed there was a 95% chance that the assembly line was working fine. What about now?

Bayes's rule is an explicit equation that tells us how to incorporate this new information, turning our initial probability $P(A)$ into a new, updated probability:

$$P(A \mid B) = \frac{P(A) \cdot P(B \mid A)}{P(B)}. \tag{12.1}$$

Each piece of this equation has a name:

- $P(A)$ is the prior probability: how probable is A, before ever having seen data B?
- $P(A \mid B)$ is the posterior probability: how probable is A, now that we've seen data B?

Figure 12.3: Bayes' rule is named after Thomas Bayes (above), an English reverend of the 18th century who first derived the result. It was published posthumously in 1763 in "An Essay towards solving a Problem in the Doctrine of Chances."

- $P(B \mid A)$ is the likelihood: if A were true, how likely is it that we'd see data B?

- $P(B)$ is the marginal probability of B: how likely is it that we'd see data B anyway, regardless of whether A is true or not? This calculation is usually the tedious part of applying Bayes' rule. Usually, as we'll see in the examples, we use the rule of total probability, which we learned in the previous chapter.

Have you found the two-headed coin?

To get a feel for what's going on here, let's see an example of Bayes' rule in action.

Imagine a jar with 1024 normal quarters. Into this jar, a friend places a single two-headed quarter (i.e. with heads on both sides). Your friend then gives the jar a good shake to mix up the coins. You draw a single coin at random from the jar, and without examining it closely, flip the coin ten times. The coin comes up heads all ten times. Are you holding the two-headed quarter, or an ordinary quarter?

Now, you might be thinking that this example sounds pretty artificial. But it's not at all. In fact, in the real world, an awful lot of time and energy is spent looking for metaphorical two-headed coins—specifically, in any industry where companies compete strenuously for talented employees. To see why, let's change the story just a little bit.

Suppose you're in charge of a large trading desk at a major Wall Street bank. You have 1025 employees under you, and each one is responsible for managing a portfolio of stocks to make money for your firm and its clients.

One day, a young trader knocks on your door and confidently asks for a big raise. You ask her to make a case for why she deserves one. She replies:

> Look at my trading record. I've been with the company for ten months, and in each of those ten months, my portfolio returns have been in the top half of all the portfolios managed by my peers on the trading floor. If I were just an average trader, this would be very unlikely. In fact, the probability that an average trader would see above-average results for ten months in a row is only $(1/2)^{10}$, which is less than one chance in a thousand. Since it's unlikely I would be that lucky,

the implication is that I am a talented trader, and I should therefore get a raise.

The math of this scenario is exactly the same as the one involving the big jar of quarters. Metaphorically, the trader is claiming to be a two-headed coin (T), on the basis of some data D: that she performs above average, every single month without fail.

But from your perspective, things are not so clear. Is the trader lucky, or good? There are 1025 people in your office (i.e. 1025 coins). Now you're confronted with the data that one of them has had an above-average monthly return for ten months in a row (i.e. D = "flipped heads ten times in a row"). This is admittedly unlikely, and this person might therefore be an excellent performer, worth paying a great deal to retain. But excellent performers are probably also rare, so that the prior probability $P(T)$ is pretty small to begin with. To make an informed decision, you need to know $P(T \mid D)$: the posterior probability that the trader is an above-average performer, given the data.

Applying Bayes' rule. So our two-headed coin example definitely has real-world applications. Let's use it to see how a posterior probability is calculated using Bayes' rule:

$$P(T \mid D) = \frac{P(T) \cdot P(D \mid T)}{P(D)}.$$

We'll take this equation one piece at a time. First, what is $P(T)$, the prior probability that you are holding the two-headed quarter? Well, there are 1025 quarters in the jar: 1024 ordinary ones, and one two-headed quarter. Assuming that your friend mixed the coins in the jar well enough, then you are just as likely to draw one coin as another, and so $P(T)$ must be $1/1025$.

Next, what about $P(D \mid T)$, the likelihood of flipping ten heads in a row, given that you chose the two-headed quarter? Clearly this is 1—there is no possibility of seeing anything else.

Finally, what about $P(D)$, the marginal probability of flipping ten heads in a row? As is almost always the case when using Bayes' rule, $P(D)$ is the hard part to calculate. We will use the law of total probability to do so:

$$P(D) = P(T) \cdot P(D \mid T) + P(\text{not } T) \cdot P(D \mid \text{not } T).$$

Taking the pieces on the right-hand one by one:

- As we saw above, the prior probability of the two-headed coin, $P(T)$, is $1/1025$.
- This means that the prior probability of an ordinary coin, $P(\text{not } T)$, must be $1024/1025$.
- Also from above, we know that $P(D \mid T) = 1$.
- Finally, we can calculate $P(D \mid \text{not } T)$ quite easily. If the coin is an ordinary quarter, then there is a 50% chance of getting heads on any one coin flip. Each flip is independent. Therefore, by the compounding rule, the probability of a 10-head winning streak is

$$\begin{aligned} P(D \mid \text{not } T) &= \frac{1}{2} \times \frac{1}{2} \times \cdots \times \frac{1}{2} \quad \text{(10 times)} \\ &= \left(\frac{1}{2}\right)^{10} = \frac{1}{1024}. \end{aligned}$$

We can now put all these pieces together:

$$\begin{aligned} P(T \mid D) &= \frac{P(T) \cdot P(D \mid T)}{P(T) \cdot P(D \mid T) + P(\text{not } T) \cdot P(D \mid \text{not } T)} \\ &= \frac{\frac{1}{1025} \cdot 1}{\frac{1}{1025} \cdot 1 + \frac{1024}{1025} \cdot \frac{1}{1024}} = \frac{1/1025}{2/1025} \\ &= \frac{1}{2}. \end{aligned}$$

Perhaps surprisingly, there is only a 50% chance that you are holding the two-headed coin. Yes, flipping ten heads in a row with a normal coin is very unlikely. But so is drawing the one two-headed coin from a jar of 1024 normal coins! In fact, as the math shows, both explanations for the data are equally unlikely, which is why we're left with a posterior probability of 0.5.

Two-headed coins in the wild. Let's return to the scenario of the trader knocking at your door, asking for a rise on the basis of a 10-month winning streak. In light of what you know about Bayes' rule, which of the following replies is the most sensible?

(A) "You're right. Here's a giant raise."

(B) "Thank you for letting me know. While I need more data to give you a raise, you've had a good ten months. I'll review your case again in 6 months and will look closely at the facts you've showed me."

The best answer depends very strongly on your beliefs about whether excellent stock traders are common or rare. For example, suppose you believe that 10% of all stock traders are truly excellent, in the sense that they can reliably finish with above-average returns, month after month; and that the other 90% just muddle through and collect their thoroughly average bonus checks. Then $P(T) = 0.1$, and

$$P(T \mid D) = \frac{0.1 \cdot 1}{0.1 \cdot 1 + 0.9 \cdot \frac{1}{1024}} \approx 0.991\,,$$

so that there is better than a 99% chance that your employee is among those 10% of excellent performers. You should give her a raise, or risk letting some other bank save you the trouble.

What if, however, you believed that excellence were much rarer, say $P(T) = 1/10000$? In that case,

$$P(T \mid D) = \frac{0.0001 \cdot 1}{0.0001 \cdot 1 + 0.9999 \cdot \frac{1}{1024}} \approx 0.093\,.$$

In this case, even though the ten-month hot streak was unusual—$P(D \mid \text{not } T)$ is small, at $1/1024$—there is still more than a 90% chance that your employee got lucky.

The moral of the story is that the prior probability in Bayes' rule—in this case, the baseline rate of excellent stock traders, or two-headed coins—plays a very important role in correctly estimating conditional probabilities. Ignoring this prior probability is a big mistake, and such a common one that it gets its own name: the base-rate fallacy.[7]

[7] en.wikipedia.org/wiki/Base_rate_fallacy

So just how rare are two-headed coins? While it's very difficult to know the answer to this question in something like stock-trading, it is worth pointing out one fact: in the above example, a prior probability of 10% is almost surely too large. Remember the NP rule: if this prior probability were right, then out of your office of 1025 traders, you would expect there to be $0.1 \times 1025 \approx 100$ of them with 10-month winning streaks, all at your door at once clamoring for a raise. (Traders are not known for being shy about their winning streaks, or anything else.) Since this hasn't happened, the prior probability $P(T) = 0.1$ is too high to be consistent with all the data available, and should be revised downward.

On the flip side, we also know that two-headed coins in stock-picking do exist, or else there would be no explanation for Warren Buffett, known as the "Oracle of Omaha." Over the last 50 years, Warren Buffett has beaten the market so badly that it almost defies

belief: between 1964 and 2013, the share price of his holding company, Berkshire Hathaway, has risen by about 1 million percent, versus only 2300% for the S&P 500 stock index.

This line of reasoning demonstrates that, while the prior probability often reflects your own knowledge about the world, it can also be informed by data. Either way, it is very influential, and should not be ignored.

Understanding Bayes' rule using trees

Let's try a second example of Bayes' rule in action. You may have encountered the following brain-teaser, which is the frequent subject of "first caller with the right answer wins a prize"-style contests on the radio. Here's how one San Francisco radio station described the puzzle:

> There are two tribes in the jungle. The truth tellers always tell the truth and the liars always lie. A scientist comes to a fork in the road. He knows that the truth tellers' tribe is in one direction and the liars' tribe is in the other direction. But he does not know which direction is the truth tellers' tribe. There is a native sitting at the intersection of the roads. The scientist does not know whether this native is a truth teller or a liar. The scientist may ask the native one and only one question to determine the direction to the truth tellers' tribe. What question should he ask?

The scientist should ask "Which way to your village?" (Ponder for a minute why this is guaranteed to be informative.)

To illustrate Bayes' theorem, let's amend this example to include probabilities, and to adjust for the fact that it's really not cool to talk about lying tribes of jungle natives these days. Suppose you face the following situation:

> You are driving through unfamiliar territory in East Texas in your burnt-orange car sporting a bumper sticker from the University of Texas. You reach a fork in the road. In one direction lies College Station; in another direction, Austin. The road sign pointing to Austin has been stolen, but you see a man selling watermelons out of his pickup truck. You pull over and ask him for directions.
>
> You know that there are two kinds of people in this part of Texas, Longhorns and Aggies, with Aggies outnumbering Longhorns by a 60/40 margin. But you don't know which one this man is. If he's a Longhorn, he is sure to help you out to the best of his ability, and you judge that there is only a 5%

chance that he will get confused and point you in the wrong direction. But you believe that, if he is an Aggie and you ask him for directions, there is a 70% chance that he will see the bumper sticker on your car and send you the opposite way.

Having once won a pair of concert tickets by solving a similar brain teaser posed by a radio station, you decide to ask him "Which way is your university?" He stares for a moment at your bumper sticker, then smiles and points to the left. You go in the direction indicated, and two hours later you arrive in Austin.

Given that you ended up in Austin, what is the probability that the man you encountered was a Longhorn?

One way of solving this is to use Bayes' rule directly:

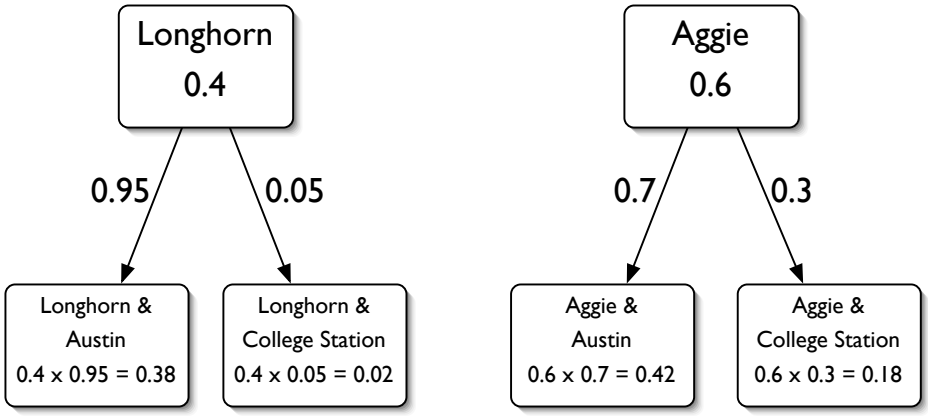
$$\begin{aligned} P(\text{Longhorn} \mid \text{pointed to Austin}) &= \frac{P(\text{Longhorn}) \cdot P(\text{pointed to Austin} \mid \text{Longhorn})}{P(\text{pointed to Austin})} \\ &= \frac{0.4 \cdot 0.95}{0.6 \cdot 0.7 + 0.4 \cdot 0.95} \\ &= 0.475\,. \end{aligned}$$

There is slightly better than an even chance that you were talking to an Aggie.

So Bayes' rule gets us the answer with little fuss, by simplying plugging in the appropriate terms to the formula. But an alternative, very intuitive way of solving this problem—and of understanding Bayes' theorem more generally—is to use a tree.

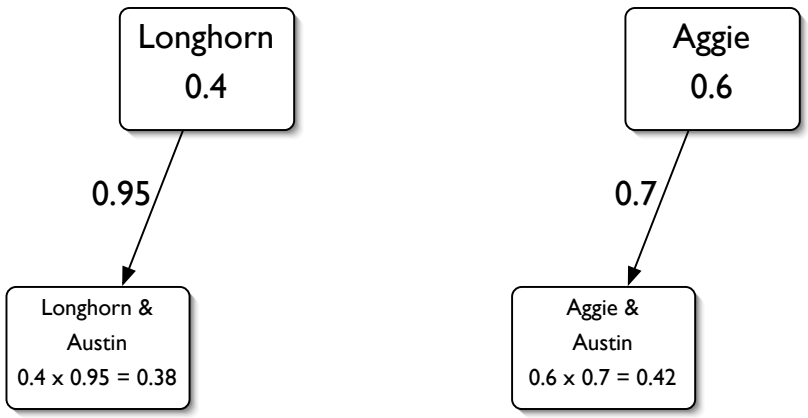
Let's see how this works. First, start by listing the possible states of the world, along with their probabilities. I like to put these in boxes:

Next, draw arrows from each state of the world to the possible observational consequences. Along the arrows, put the conditional probabilities that you will observe each data point, given the corresponding state of the world:



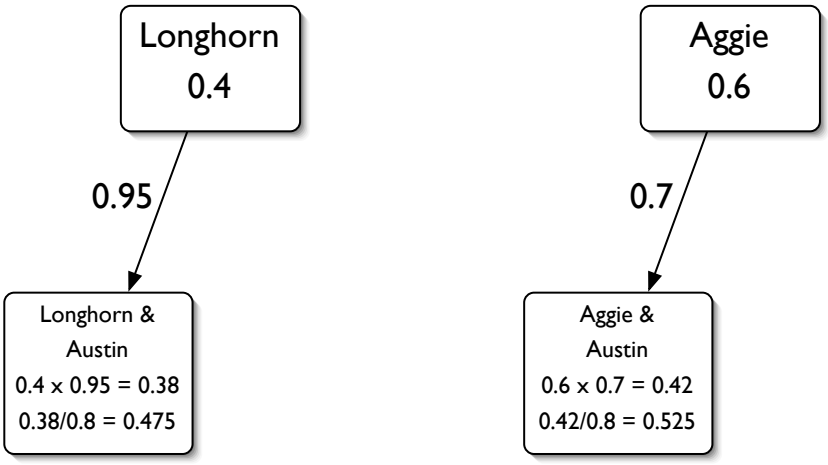
At the terminal leaves of the tree, multiply out the probabilities according to the multiplication rule: $P(A, B) = P(A) \cdot P(B \mid A)$. So, for example, the probability that the man is an Aggie and that he points you to Austin is $0.7 \times 0.6 = 0.42$. The sum of all the probabilities in the leaves of the tree must be 1, since they exhaust all possibilities.

But now that you've arrived back home, you know that the man was pointing to Austin. To use the tree to compute the probability that he was a Longhorn, simply cut off all the branches corresponding to data that wasn't actually observed, leaving only the actual data:



The remaining leaf probabilities are proportional to the posterior probabilities of the corresponding states of the world. These prob-

abilities, of course, do not sum to 1 anymore. But this is easily fixed: simply divide each probability by the sum of all remaining probabilities (in this case, $0.38 + 0.42 = 0.8$):



As before, we find that there is a probability of 0.475 that the man at the watermelon stand was a Longhorn. Just so you can keep things straight, the sum of 0.8, by which you divided the terminal-leaf probabilities in the final step to ensure that they summed to 1, corresponds exactly to the denominator, P(points to Austin), in Bayes' rule.

Bayes' rule and the law

SUPPOSE you're serving on a jury in the city of New York, with a population of roughly 10 million people. A man stands before you accused of murder, and you are asked to judge whether he is guilty (G) or not guilty ($\sim G$). In his opening remarks, the prosecutor tells you that the defendant has been arrested on the strength of a single, overwhelming piece of evidence: that his DNA matched a sample of DNA taken from the scene of the crime. Let's call denote this evidence by the letter D. To convince you of the strength of this evidence, the prosecutor calls a forensic scientist to the stand, who testifies that the probability that an innocent person's DNA would match the sample found at the crime scene is only one in a million. The prosecution then rests its case.

Would you vote to convict this man?

If you answered "yes," you might want to reconsider! You are charged with assessing $P(G \mid D)$—that is, the probability that the defendant is guilty, given the information that his DNA matched the sample taken from the scene. Bayes' rule tells us that

$$P(G \mid D) = \frac{P(G) \cdot P(D \mid G)}{P(D)} = \frac{P(G) \cdot P(D \mid G)}{P(D \mid G) \cdot P(G) + P(D \mid\sim G)P(\sim G)}.$$

We know the following quantities:

- The prior probability of guilt, $P(G)$, is about one in 10 million. New York City has 10 million people, and one of them committed the crime.
- The probability of a false match, $P(D \mid\sim G)$, is one in a million, because the forensic scientist testied to this fact.

To use Bayes' rule, let's make one additional assumption: that the likelihood, $P(D \mid G)$, is equal to 1. This means we're assuming that, if the accused were guilty, there is a 100% chance of seeing a positive result from the DNA test.

Let's plug these numbers into Bayes' rule and see what we get:

$$\begin{aligned} P(G \mid D) &= \frac{\frac{1}{10{,}000{,}000} \cdot 1}{1 \cdot \frac{1}{10{,}000{,}000} + \frac{1}{1{,}000{,}000} \cdot \frac{9{,}999{,}999}{10{,}000{,}000}} \\ &\approx 0.09\,. \end{aligned}$$

The probability of guilt looks to be only 9%! This result seems shocking in light of the forensic scientist's claim that $P(D \mid\sim G)$ is so small: a "one in a million chance" of a positive match for an innocent person. Yet the prior probability of guilt is very low—$P(G)$ is a mere one in 10 million—and so even very strong evidence still only gets us up to $P(G \mid D) = 0.09$.

Conflating $P(\sim G \mid D)$ with $P(D \mid\sim G)$ is a serious error in probabilistic reasoning. These two numbers are typically very different from one another, because conditional probabilities aren't symmetric. As we've said more than once, $P(\text{practices hard} \mid \text{plays in NBA}) \approx 1$, while $P(\text{plays in NBA} \mid \text{practices hard}) \approx 0$. Getting this wrong—that is, conflating $P(A \mid B)$ with $P(B \mid A)$—is so common that it has its own name: the prosecutor's fallacy.[8]

[8] en.wikipedia.org/wiki/Prosecutor's_fallacy

An alternate way of thinking about this result is the following. Of the 10 million innocent people in New York, ten would have DNA matches merely by chance. The one guilty person would also have a DNA match. Hence there are 11 people with a DNA match, only one of whom is guilty, and so $P(G \mid D) \approx 1/11$. Your intuition may mislead, but Bayes' rule never does!

13
Parametric probability models

Describing randomness

THE MAJOR ideas of the last few chapters all boil down to a simple idea: even random outcomes exhibit structure and obey certain rules. In this chapter, we'll learn to use these rules to build probability models, which employ the language of probability theory to provide mathematical descriptions of random phenomena. Probability models can be used to answer interesting questions about real-world systems. For example:

- American Airlines oversells a flight from Dallas to New York, issuing 140 tickets for 134 seats, because they expect at least 6 no-shows (i.e. passengers who bought a ticket but fail to show up for the flight). How likely is it that the airline will have to bump someone to the next flight?
- Arsenal scores 1.6 goals per game; Manchester United scores 1.3 goals per game. How likely it is that Arsenal beats Man U when they play each other?
- Since 1900, stocks have returned about 6.5% per year on average, net of inflation, but with a lot of variability around this mean. How does this variability affect the likely growth of your investment portfolio? How likely it is that you won't meet your retirement goals with your current investment strategy?

Building a probability model involves three steps.

(1) Identify the *random variables* in your system, A random variable is just a term for any uncertain quantity or source of randomness. In the airline example, there is just one uncertain quantity: X = the number of no-shows on the Dallas–NYC flight. In the soccer game between Arsenal and Man U, there

are two uncertain quantities: X_1 = the number of goals scored by Arsenal, and X_2 = the number of goals scored by Man U.

(2) Describe the possible outcomes for the random variables. These possible outcomes are called *events*, and the set of all possible events is referred to as the *sample space* of the probability model. In the airline example, our random variable X, the number of no-shows, could be any number between 0 and 140 (the number of tickets sold). Thus the sample space is the set of integers 0 to 140.

In the soccer-game example, the sample space is a bit more complicated: it is the set of all possible scores (1-0, 2-3, 7-0, etc.) in a soccer game.

(3) Finally, provide a rule for calculating probabilities associated with each event in the sample space. This rule is called a *probability distribution*. In the airline example, this distribution might be described using a simple lookup table based on historical data, e.g. 1% of all flights have 1 no-show, 1.2% have 2 no-shows, 1.7% have 3 no-shows, and so forth.

Parametric models for discrete outcomes

Of the three steps required to build a probability model, the third—provide a rule that can be used to calculate probabilities for each event in the sample space—is usually the hardest one. In fact, for most scenarios, if we had to build such a rule from scratch, we'd be in for an awful lot of careful, tedious work. Imagine trying to list, one by one, the probabilities for all possible outcomes of a soccer game, or all possible outcomes for the performance of a portfolio containings a mix of stocks and bonds over 40 years.

Thus instead of building probability distributions from scratch, we will rely on a simplification called a *parametric probability model.* A parametric probability model involves a probability distribution that can be completely described using a relatively small set of numbers, far smaller than the sample space itself. These numbers are called the parameters of the distribution. There are lots of commonly used parametric models—you might have heard of the normal, binomial, Poisson, and so forth—that have been invented for specific purposes. A large part of getting better at probability modeling is to learn about these existing parametric models, and to gain an appreciation for the typical kinds of real-world prob-

lems where each one is appropriate.

Before we start, we need two quick definitions. First, by a *discrete random variable*, we mean one whose sample space consists of events that you can count on your fingers and toes. Examples here include the number of no-shows on a flight, the number of goals scored by Man U in a soccer game, or the number of gamma rays emitted by a gram of radioactive uranium over the next second. (In a later section, we'll discuss continuous random variables, which can take on any value within a given range, like the price of a stock or the speed of a tennis player's serve.)

Second, suppose that the sample space for a discrete random variable X consists of events x_1, x_2, and so forth. You'll recall that, to specify a probability model, we must provide a rule that can be used to calculate $P(X = x_k)$ for each event. When building parametric probability models, this rule takes the form of a *probability mass function*, or PMF:

$$P(X = x_k) = f(x_k \mid \theta) .$$

In words, this equation says that the probability that X takes on the value x_k is a function of x_k. The probability mass function depends a number (or set of numbers) θ, called the parameter(s) of the model.

To specify a parametric model for a discrete random variable, we must both provide both the probability mass function f and the parameter θ. This is best illustrated by example. We'll consider two: the binomial and Poisson distributions.

The binomial distribution

One of the simplest parametric models in all of probability theory is called the binomial distribution, which generalizes the idea of flipping a coin many times and counting the number of heads that come up. The binomial distribution is a useful parametric model for any situation with the following properties:

(1) We observe N different random events, each of which can be either a "yes" or a "no."

(2) The probability of any individual event being "yes" is equal to P, a number between 0 and 1.

(3) Each event is independent of the others.

(4) The random variable X of interest is total number of "yes" events. Thus the sample space is the set $\{0, 1, \ldots, N-1, N\}$.

The meaning of "yes" events and "no" events will be context-dependent. For example, in the airline no-show example, we might say that a "yes" event corresponds to a single passenger failing to show up for his or her flight (which is probably not good for the passenger, but definitely a success in the eyes of an airline that's overbooked a flight). Another example: in the PREDIMED study of the Mediterranean diet, a "yes" event might correspond to single study participant experiencing a heart attack.

If a random variable X satisfies the above four criteria, then it follows a binomial distribution, and the PMF of X is

$$P(X = k) = f(k \mid N, P) = \binom{N}{k} P^k \, (1-P)^{N-k}, \tag{13.1}$$

where N and P are the parameters of the model. The notation $\binom{N}{k}$, which we read aloud as "N choose k," is shorthand for the following expression in terms of factorials:

$$\binom{N}{k} = \frac{N!}{k!(N-k)!}.$$

This term, called a binomial coefficient, counts the number of possible ways there are to achieve k "yes" events out of N total events. (You'll see how this is derived in a bit.)

Example: airline no-shows Let's use the binomial distribution as a probability model for our earlier example on airline no-shows. The airline sold tickets 140 people, each of which will either show up to fly that day (a "yes" event) or not (a "no" event). Let's make two simplifying assumptions: (1) that each person decides to show up or not independently of the other people, and (2) that the probability of any individual person failing to show up for the flight is 9%.[1] These assumptions make it possible to apply the binomial distribution. Thus the distribution for X, the number of ticketed passengers who fail to show up for the flight, has PMF

$$P(X = k) = \binom{140}{k} (0.09)^k \, (1-0.09)^{140-k}.$$

This function of k, the number of no-shows, is plotted in Figure 13.1. The horizontal axis shows k; the vertical axis shows $P(X = k)$ under the binomial model with parameters $N = 140, p = 0.09$.

[1] This is the industry average, quoted in "Passenger-Based Predictive Modeling of Airline No-show Rates," by Lawrence, Hong, and Cherrier (SIGKDD 2003 August 24-27, 2003).

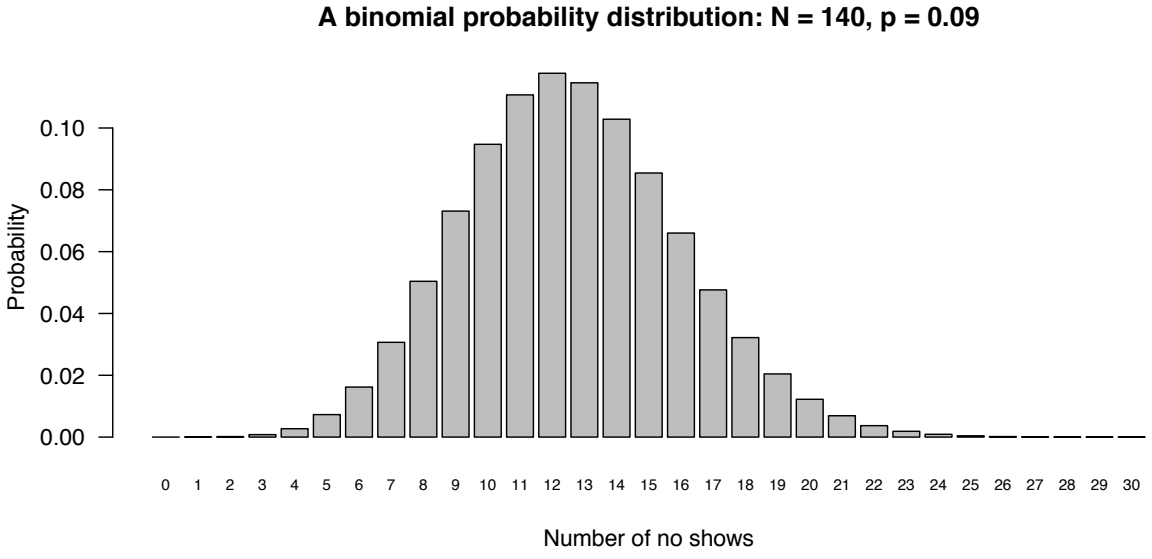


Figure 13.1: A barplot showing the probability distribution for the number of no-shows on an overbooked airline flight with 140 tickets sold, assuming a no-show rate of 9% and that individual no-shows are independent. The horizontal axis has been truncated at $k = 30$ because the probability of more than 30 no-shows is vanishly small under the binomial model.

According to this model, the airline should expect to see around $E(X) = Np = 140 \cdot 0.09 = 12.6$ no shows, with a standard deviation of $\text{sd}(X) = \sqrt{140 \cdot 0.09 \cdot (1 - 0.09)} \approx 3.4$. But remember that the question of interest is: what is the probability of fewer than 6 no-shows? If this happens, the airline will have to compensate the passengers they bump to the next flight. We can calculate this as

$$P(X < 6) = P(X = 0) + P(X = 1) + \cdots + P(X = 5) \approx 0.011\,,$$

i.e. by adding up the probabilities for 0 no-shows through 5 no-shows. The airline should anticipate a 1.1% chance that more people will show up than can fit on the plane.

The trade-offs of the binomial model. It's worth noting that real airlines use much more complicated models than we've just built here. These models might take into account, for example, the fact that passengers on a late connecting flight will fail to show up together non-independently, and that business travelers are more likely no-shows than families on a vacation.

The binomial model—like all parametric probability models—cannot incorporate these (very real) effects. It's just an approximation. This approximation trades away flexibility for simplicity: instead of having to specify the probability of all possible outcomes between 0 and 140, we only have to specify two numbers:

$N = 140$ and $p = 0.09$, the parameters of the binomial distribution. These parameters then determine the probabilities for all events in the sample space.

In light of this trade-off, any attempt to draw conclusions from a parametric probability model should also involve the answer to two important questions. First, what unrealistic simplifications have we made in building the model? Second, have these assumptions made our model *too* simple? This second answer will always be context dependent, and it's hard to provide general guidelines about what "too simple" means. Often this boils down to the question of what might go wrong if we use a simplified model, rather than invest the extra work required to build a more complicated model. This is similar to the trade-off that engineers face when they build simplified physical models of something like a suspension bridge or a new fighter jet. Like many things in statistics and probability modeling, this is a case where there is just no substitute for experience and subject-area knowledge.

The connection with the NP rule

The binomial distribution brings us back to our discussion of the NP rule, and establishes a connection between the two definitions we've seen so far of *expected value*:

The simple definition. Suppose we are in a situation with many repeated exposures (N) to the same chance event that has probability P of happening. In the long run, the expected number of events is the frequency of encounters (N), times the probability of the event in a single encounter (P). Thus expected value = $N \times P$.

The formal definition. Suppose that the possible outcomes for a random variable X are the numbers $x_1, \ldots, x_N$. Back in Equation 10.2 on page 219, we learned that the formal definition for the expected value of X is

$$E(X) = \sum_{K=1}^{N} P(X = x_i) \cdot x_i \,.$$

Thus the expected value is the probability-weighted average of possible outcomes.

To see the connection between these two definitions, let's suppose that X is a binomial random variable: $X \sim \text{Binomial}(N, P)$.

If we apply the formal definition of expected value and churn through the math, we find that

$$\begin{aligned} E(X) &= \sum_{k=0}^{k=N} \binom{N}{k} P^k \, (1-P)^{N-k} \cdot k \\ &= NP \, . \end{aligned}$$

We've skipped a lot of algebra steps here, but the punchline is a lot more important than the derivation: a random variable with a binomial distribution has expected value $\mathrm{E}(X) = NP$.

This gives us a richer understanding of NP rule for expected value. The NP rule is a valid way of calculating an expected value precisely for those random events that can be described by a binomial distribution—that is, those events satisfying criteria (1)-(3) on page 257. For random events that *don't* meet these criteria, you'll need to use the formal definition from Equation 10.2 on page 219.

Note: a similar calculation shows that a random variable with a binomial distribution has standard deviation $\mathrm{sd}(X) = \sqrt{NP(1-P)}$.

Advanced topic: a derivation of the binomial distribution

To motivate the idea of the binomial distribution, suppose we flip a fair coin only twice.[2] Let our random variable X be the number of times we see "heads" in two coin flips. Thus our sample space for X has three possible outcomes—zero, one, or two. Since the coin flips are independent, all four possible sequences for the two flips (HH, HT, TH, TT) are equally likely, and the probability distribution for X is given by the following table:

[2] By fair, we mean that coin is equally likely to come up heads or tails when flipped.

x_k	$P(X = k)$	Cases
0	0.25	0 heads (TT)
1	0.50	1 head (HT or TH)
2	0.25	2 heads (HH)

The logic of this simple two-flip case can be extended to the general case of N flips: by accounting for every possible sequence of heads and tails that could arise from N flips of a fair coin. Since successive flips are independent, every sequence of heads and tails has the same probability: $1/2^N$. Therefore,

$$P(X = k \text{ heads}) = \frac{\text{Number of sequences with } k \text{ heads}}{\text{Total number of possible sequences}} \, . \qquad (13.2)$$

There are 2^N possible sequences, which gives us the denominator. To compute the numerator, we must count the number of these sequences where we see exactly k heads.

How many such sequences are there? To count them, imagine distributing the k heads among the N flips, like putting k items in N boxes, or handing out k cupcakes among N people who want one. Clearly there are N people to which we can assign the first cupcake. Once we've assigned the first, there are $N-1$ people to which we could assign the second cupcake. Then there are $N-2$ choices for the third, and so forth for each successive cupcake. Finally for the kth and final cupcake, there are $N-k+1$ choices. Hence we count

$$N \times (N-1) \times (N-2) \times \cdots \times (N-k+1) = \frac{N!}{(N-k)!}$$

possible sequences, where $N!$ is the factorial function. For example, if $m = 10$ and $k = 7$, this gives 604,800 sequences.

But this is far too many sequences. We have violated an important principle of counting here: don't count the same sequence more than once. The problem is that have actually counted all the ordered sequences, even though we were trying to count unordered sequences. For example, in the $N = 10, k = 7$ case, we have counted "Heads on flips $\{1,2,3,4,5,6,7\}$" and "Heads on flips $\{7,6,5,4,3,2,1\}$" as two different sequences. But they clearly both correspond to the same sequence: HHHHHHHTTT.

So how many times have we overcounted each unordered sequence in our tally of the ordered ones? The way to compute this is to count the number of ways we could order k objects. Given a group of k numbers which will be assigned to the "heads" category, we could have chosen from k of the objects to be first in line, from $k-1$ of them to be second in line, from $k-2$ of them to be third in line, and so forth. This means we have counted each unordered sequence $k!$ times. Thus the correct number of ways we could choose k objects out of N possiblities is

$$\frac{N!}{k!(N-k)!} = \binom{N}{k}.$$

For $N = 10$ and $k = 7$, this is 120 sequences—the right answer, and a far cry from the 604,800 we counted above.

Putting all these pieces together, we find that the probability of getting k heads in N flips of a fair coin is

$$P(k \text{ heads}) = \frac{N!}{k!(N-k)!}\,\frac{1}{2^N} = \binom{N}{k}\frac{1}{2^N}. \tag{13.3}$$

The general case. The above derivation assumes that "yes" (success) and "no" (failure) events are equally likely. Let's now relax this assumption to see where the general definition of the binomial distribution comes from, when the probability of any individual success is not 0.5, but some rather some generic probability p.

Let's take a sequence of N trials where we observed k successes. Each success happens with probability p, and there are k of them. Each failure happens with probability $1 - p$, and there are $m - k$ of them. Because each trial is independent, we multiply all of these probabilities together to get the probability of the whole sequence: $p^k \ (1-p)^{m-k}$. Moreover, our analysis above shows that there are precisely $\binom{N}{k}$ such sequences (i.e. unique ways of getting exactly k successes and $N - k$ failures).

So if we let X denote the (random) number of successes in N trials, then for any value of k from 0 to N,

$$P(X = k) = \binom{N}{k} \ p^k \ (1-p)^{N-k},$$

which is the probability mass function given in Equation 13.1.

The Poisson distribution

Our second example of a parametric probability model is the Poisson distribution, named after the French mathematician Siméon Denis Poisson.[3] The Poisson distribution is used to model the number of times than some event occurs in a pre-specified interval of time. For example:

(1) How many goals will Arsenal score in their game against Man U? (The event is a goal, and the interval is a 90-minute game.)

(2) How many couples will arrive for dinner at a hip new restaurant between 7 and 8 PM on a Friday night? (The event is the arrival of a couple asking to sit at a table for two, and the interval is one hour).

(3) How many irate customers will call the 1-800 number for AT&T customer service in the next minute? (The event is a phone call that must be answered by someone on staff, and the interval is one minute.)

In each case, we identify the random variable X as the total number of events that occur in the given interval. The Poisson dis-

[3] The French speakers among you, or at least the fans of Disney movies, might recognize the word poisson from a different context. Run, Sebastian!

tribution will provide an appropriate description for this random variable if the following criteria are met:

(1) The events occur independently; seeing one event neither increases nor decreases the probability that a subsequent event will occur.

(2) Events occur the same average rate throughout the time interval. That is, there is no specific sub-interval where events are more likely to happen than in other sub-intervals. For example, this would mean that if the probability of Arsenal scoring a goal in a given 1-minute stretch of the game is 2%, then the probability of a goal during *any* 1-minute stretch is 2%.

(3) The chance of an event occuring in some sub-interval is proportional to the length of that sub-interval. For example, this would mean that if the probability of Arsenal scoring a goal in any given 1-minute stretch of the game is 2%, then the probability that they score during a 2-minute stretch is 4%.

A random variable X meeting these criteria is said to follow a Poisson distribution. The sample space of a Poisson distribution is the set of non-negative integers $0, 1, 2$, etc. This is one important way in which the Poisson differs from the binomial. A binomial random variable cannot exceed N, the number of trials. But there is no fixed upper bound to a Poisson random variable.

The probability mass function of Poisson distribution takes the following form:

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda},$$

with a single parameter λ (called the rate). This parameter governs the average number of events in the interval: $E(X) = \lambda$. It also governs the standard deviation: $\text{sd}(X) = \sqrt{\lambda}$.

Example: modeling the score in a soccer game. Let's return to our soccer game example. Across all games in the 2015-16 English Premiere League (widely considered to be the best professional soccer league in the world), Arsenal scored 1.6 goals per game, while Manchester United scored 1.3 goals per game. How likely is it that Arsenal beats Man U? How likely is a scoreless draw at 0-0? To answer these questions, let's make some simplifying assumptions.

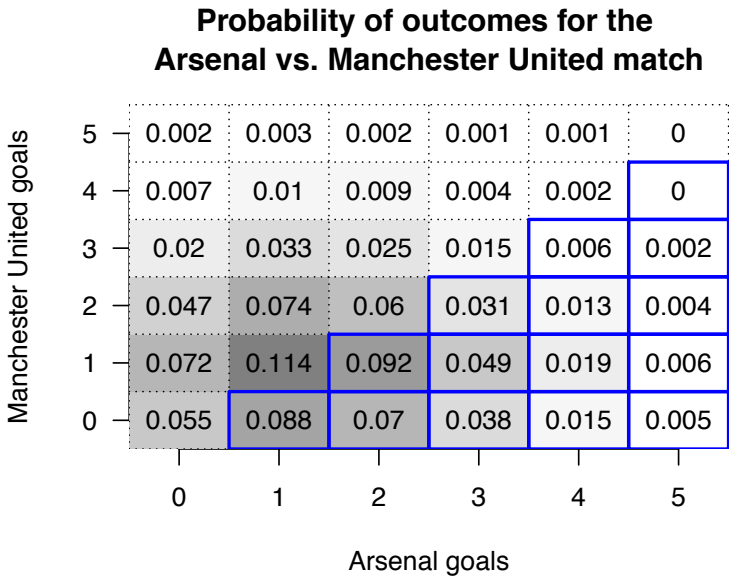


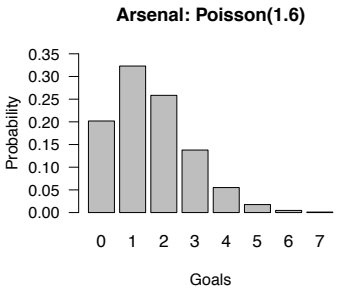
Figure 13.2: A matrix of probabilities associated with various match scores under the independent Poisson model of an Arsenal vs. Man U match, based on scoring statistics from 2015-16 Premiere League season. Each entry in the matrix is the probability with the corresponding score (darker grey = higher probability). The cells outlined in blue correspond to an Arsenal win, which happens with probability 44% (versus 25% for a draw and 31% for a Manchester United win.

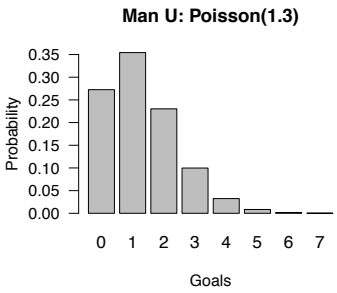
(1) Let X_A be the number of goals scored in a game by Arsenal. We will assume that X_A can be a described by a Poisson distribution with rate parameter 1.6: that is, $X_A \sim \text{Poisson}(\lambda = 1.6)$.

(2) Let X_M be the number of goals scored in a game by Manchester United. We will assume that $X_M \sim \text{Poisson}(\lambda = 1.3)$.

(3) Finally, we will assume that X_A and X_M are independent of one another.

Our model sets the rate parameters for each team's Poisson distribution to match their average scoring rates across the season. The corresponding PMFs are shown at right.





Under these simplifying assumptions, we can calculate the probability of any possible score—for example, Arsenal 2–0 Manchester United. Because we have assumed that X_A and X_M are independent, we can multiply together the two probabilities we get from each random variable's Poisson distribution:

$$P(X_A = 2, X_M = 0) = \left(\frac{1.6^2}{2!}e^{-1.6}\right) \cdot \left(\frac{1.3^0}{0!}e^{-1.3}\right) \approx 0.07\,.$$

Figure 13.2 shows a similar calculation for all scores ranging from 0–0 to 5–5 (according to the model, the chance of a score

larger than this is only 0.6%). By summing up the probabilities for the various score combinations, we find that:

- Arsenal wins with probability 44%.
- Man U wins with probability 31%.
- The game ends in a draw with probability 25%. In particular, a scintillating 0–0 draw happens with probability 5.5%.

The normal distribution

This chapter's third and final example of a parametric probability model is the normal distribution—the most famous and widely used distribution in the world.

Some history

Historically, the normal distribution arose an an approximation to the binomial distribution. In 1711, a Frenchman named Abraham de Moivre published a book called *The Doctrine of Chances*. The book was reportedly was prized by gamblers of the day for its many useful calculations that arose in dice and card games. In the course of writing about these games, de Moivre found it necessary to perform computations using the binomial distribution for very large values of N, the number of independent trial in a binomial distribution. (Imagine flipping a large number of coins and making bets on the outcomes, and you too will see the necessity of this seemingly esoteric piece of mathematics.)

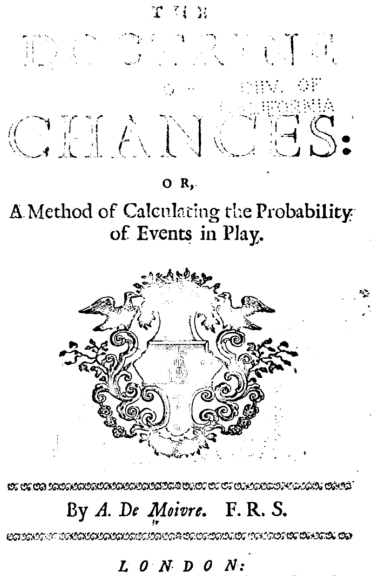

THE
DOCTRINE
OF
CHANCES:
OR,
A Method of Calculating the Probability
of Events in Play.
By *A. De Moivre.* F. R. S.
LONDON:

Figure 13.3: The title page of de Moivre's "The Doctrine of Chances" (1711), from an early edition owned by the University of California, Berkeley. One interesting thing about the history of statistics is the extent to which beautiful mathematical results came out of the study of seemingly trivial gambling and parlor games.

As you recall the previous section, these calculations require computing binomial coefficients $\binom{N}{k}$ for very large values of N. But because these computations involve the factorial function, they were far too time-consuming without modern computers, which de Moivre didn't have. So he derived an approximation based on the number $e \approx 2.7183$, the base of the natural logarithm. He discovered that, if a random variable X has a binomial distribution with parameters N and p, which we recall is written $X \sim \text{Binomial}(N, p)$, then the approximate probability that $X = k$ is

$$P(X = k) \approx \frac{1}{\sqrt{2\pi\sigma^2}} \, e^{-\frac{(k-\mu)^2}{2\sigma^2}}, \tag{13.4}$$

where $\mu = mp$ and $\sigma^2 = Np(1-p)$ are the expected value and variance, respectively, of the binomial distribution. When consid-

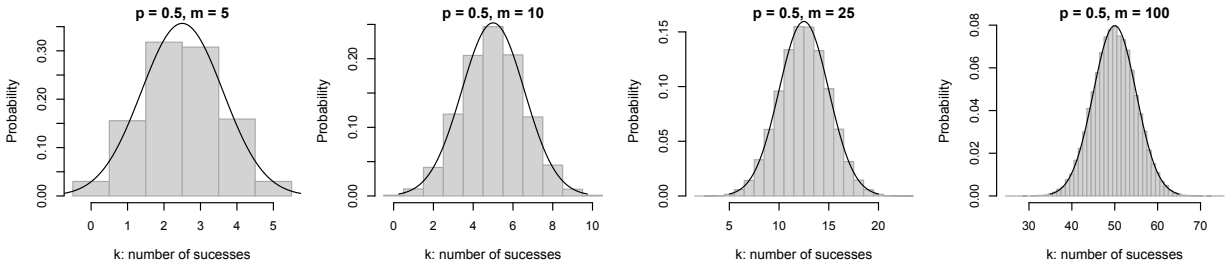


Figure 13.4: The binomial distribution for $p = 0.5$ and an increasingly large number of trials, together with de Moivre's normal approximation.

ered as a function k, this results in the familiar bell-shaped curve plotted in Figure 13.4—the famous *normal distribution*.

We can usually (though not always) avoid working with this expression directly, since every piece of statistical software out there can compute probabilities under the normal distribution. The important thing to notice is how the binomial samples in Figure 13.4 start to look more normal as the number of trials N gets progressively larger: first 5, then 10, 25, and finally 100. The histograms show the binomial distribution itself, while the black curves show de Moivre's approximation. Clearly he was on to something. This famous result of de Moivre's is usually thought of as the first *central limit theorem* in the history of statistics, where the word "central" should be understood to mean "fundamental."

The normal distribution: a modern understanding

The other term for the normal distribution is the Gaussian distribution, named after the German mathematician Carl Gauss. This raises a puzzling question. If de Moivre invented the normal approximation to the binomial distribution in 1711, and Gauss (1777–1855) did his work on statistics almost a century after de Moivre, why then is the normal distribution also named after Gauss and not de Moivre? This quirk of eponymy arises because de Moivre only viewed his approximation as a narrow mathematical tool for performing calculations using the binomial distribution. He gave no indication that he saw it as a more widely applicable probability distribution for describing random phenomena. But Gauss—together with another mathematician around the same time, named Laplace—did see this, and much more.

If we want to use the normal distribution to describe our uncertainty about some random variable X, we write $X \sim \text{N}(\mu, \sigma^2)$. The numbers μ and σ^2 are parameters of the distribution. The first

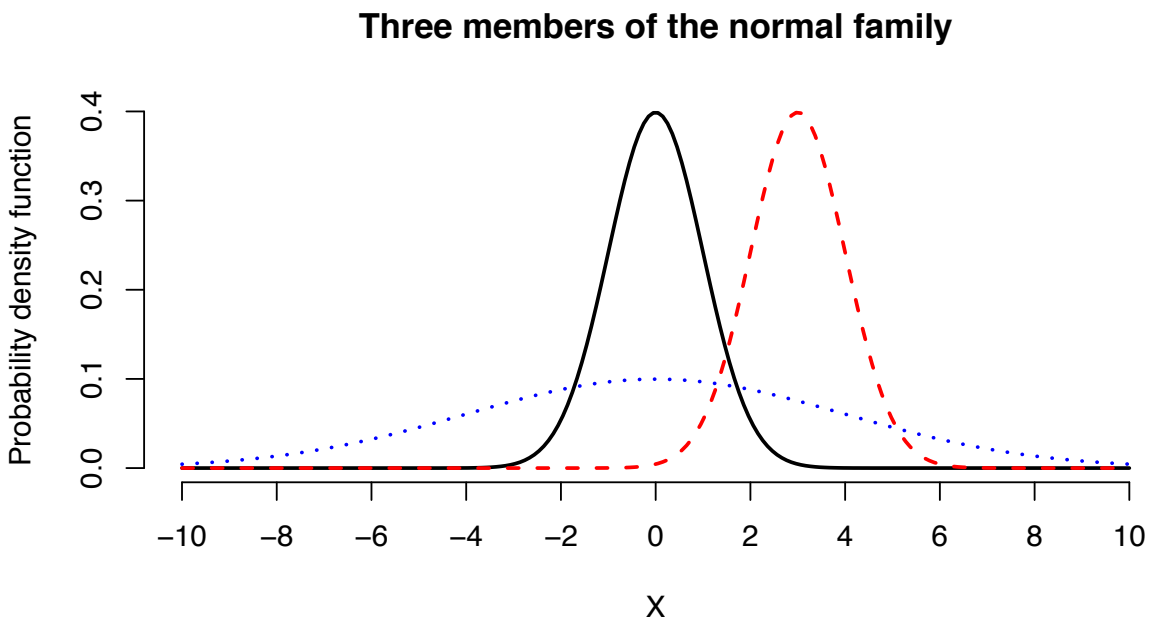


Figure 13.5: Three members of the normal family: $N(0, 1^2)$, $N(0, 4^2)$, and $N(3, 1^2)$. See if you can identify which is which using the guideline that 95% of the probability will be within two standard deviations σ of the mean. Remember, the second parameter is the variance σ^2, not the standard deviation. So $\sigma^2 = 4^2$ means a variance of 16 and a standard deviation of 4.

parameter, μ, describes where X tends to be centered; it also happens to be the expected value of the random variable. The second parameter, σ^2, describes how spread out X tends to be around its expected value; it also happens to be the variance of the random variable. Together, μ and σ^2 completely describe the distribution, and therefore completely characterize our uncertainty about X.

The normal distribution is described mathematically by its probability density function, or PDF:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{13.5}$$

If you plot this as a function of x, you get the famous bell curve (Figure 13.5). How can you interpret a "density function" like this one? If you the take the area under this curve between two values z_1 and z_2, you will get the probability that the random variable X will end up falling between z_1 and z_2 (see Figure 13.6). The height of the curve itself is a little more difficult to interpret, and we won't worry about doing so—just focus on the "area under the curve" interpretation.

Here are two useful facts about normal random variables

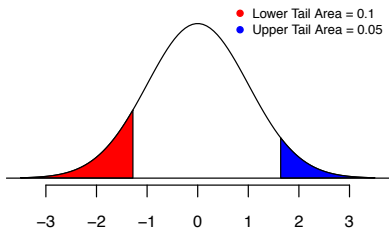


Figure 13.6: Examples of upper and lower tail areas. The lower tail area of 0.1 is at $z = -1.28$. The upper tail area of 0.05 is at $z = 1.64$

areas—or more specifically, about the central areas under the curve, between the tails. If $X \sim \mathrm{N}(\mu, \sigma^2)$, then the chance that X will be within 1σ of its mean is about 68%, and the chance that it will be within 2σ of its mean is about 95%. Said in equations:

$$\begin{aligned} P(\mu - 1\sigma < X < \mu + 1\sigma) &\approx 0.68 \\ P(\mu - 2\sigma < X < \mu + 2\sigma) &\approx 0.95\,. \end{aligned}$$

Actually, it's more like 1.96σ rather than 2σ for the second part. So if your problem requires a level of precision to an order of 0.04σ or less, then don't use this rule of thumb, and instead go with the true multiple of 1.96.

When is the normal distribution an appropriate model?

The normal distribution is now used as a probability model in situations far more diverse than de Moivre, Gauss, or Laplace ever would have envisioned. But it still bears the unmistakeable traces of its genesis as a large-sample approximation to the binomial distribution. That is, it tends to work best for describing situations where each normally distributed random variable can be thought of as the sum of many tiny, independent effects of about the same size, some positive and some negative. In cases where this description doesn't apply, the normal distribution may be a poor model of reality. Said another way: the normal distribution describes an aggregation of nudges: some up, some down, but all pretty small.

As a result, the normal distribution shares the property of the binomial distribution that huge deviations from the mean are unlikely. It has, in statistical parlance, "thin tails." Using our rule of thumb above, a normally distributed random variable has only a 5% chance of being more than two standard deviations away from the mean. It also has less than a 0.3% chance of being more than three standard deviations away from the mean. Large outliers are vanishingly rare.

For example, in the histogram of daily returns for Microsoft stock in the left panel Figure 13.7, notice the huge outliers in the lower tail. These returns would be wildly implausible if the returns really followed a normal distribution. A daily return tends to be dominated by one or two major pieces of information. It does not resemble an aggregation of many independent up-or-down nudges, and so from first principles alone, we should probably expect the normal distribution to provide a poor fit. As we

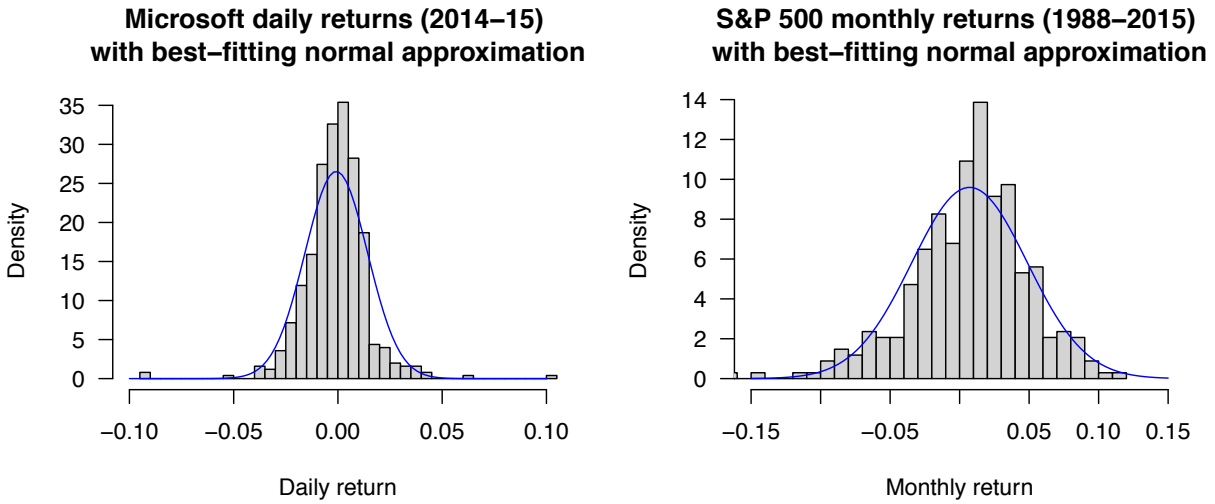


Figure 13.7: Daily stock returns for Microsoft (left) and the S&P 500 (right), together with the best-fitting normal approximations. The approximation on the right is not bad, while the approximation on the left drastically underestimates the probability of extreme results.

would expect, the best-fitting normal approximation (i.e. the one that matches the sample mean and standard deviation of the data) does not fit especially well.

The example of Microsoft stock recalls the earlier discussion on the trustworthiness of the simplifying assumptions that must go into building a probability model. To recap:

> Have these assumptions made our model too simple? This . . . answer will always be context dependent, and it's hard to provide general guidelines about what "too simple" means. Often this boils down to the questin of what might go wrong if we use a simplified model, rather than invest the extra work required to build a more complicated model.

What might go wrong if we use a normal probability model for Microsoft returns? In light of what we've seen here, the answer is: we might be very unpleasantly surprised by monetary losses that are far more extreme than envisioned under our model. This sounds very bad, and is probably a sufficient reason not to use the normal model in the first place. To make this precise, observe that the 2 most extreme daily returns for Microsoft stock were both 6 standard deviations below the mean. According to the normal model, we should only expect to see such an extreme result once every billion trading days, since

$$P(X < \mu - 6\sigma) \approx 10^{-9}\,.$$

This is a wildly overtoptimistic assessment, given that we actually saw two such results in the 503 trading days from 2014-15.

On the other hand, the normal distribution works a lot better for stock indices than it does for individual stocks, especially if we aggregate those returns over a month rather than only a day, so that the daily swings tend to average out a bit more. Take, for example, the best-fitting normal approximation for the monthly returns of the S&P 500 stock index from 1988 to 2015, in the right panel of Figure 13.7. Here the best-fitting normal distribution, though imperfect, looks a lot better than the corresponding fit for an individual stock on the left. Here, the most extreme monthly return was 4 standard deviations below the mean (which happened in October 2008, during the financial crisis of that year that augured the Great Recession). According to the normal model, we would expect such an extreme event to happen with about 2% probability in any given 27-year stretch. Thus our model looks a tad optimistic, but not wildly so.

Example: modeling a retirement portfolio

From 1900–2015, the average annual return[4] of the S&P 500 stock index is 6.5%, with a standard deviation of 19.6%. Let's use these facts to build a probability model for the future 40-year performance of a $10,000 investment in a diversified portfolio of U.S. stocks (i.e. an index fund). While there's no guarantee that past returns are a reliable guide to future returns, they're the only data we have. After all, as Mark Twain is reputed to have said, "History doesn't repeat itself, but it does rhyme."

[4] Real returns net of infation and dividends. Remember that a return is simply the implied interest rate from holding an asset for a specified period. If you buy a stock at $100 and sell a year later at $110, then your return is $(110 - 100)/100 = 0.1$, or 10%. If inflation over that year was 3%, then your real return was 7%.

Let's say that your initial investment is $W_0 = \$10,000$, and that X_t is the return of your portfolio in year t expressed as a decimal fraction (e.g. a 10% return in year 1 would mean that $X_t = 0.1$). Here t will run from 1 to 40, since we want to track your portfolio over 40 years. If we knew the returns $X_1, X_2, \ldots, X_{40}$ all the way into the future, we could calculate your terminal wealth as

$$W_{40} = W_0 \cdot \prod_{t=1}^{40} (1 + X_t)\,,$$

by simply compounding the interest year after year.[5] This formula follows from the fact that W_{t+1}, your wealth in year t, is given by the simple interest formula: $W_{t+1} = W_t \cdot (1 + X_t)$. Accumulating returns year after year then gives us the above formula.

[5] Here the symbol $\prod$ means we take the running product of all the terms, from $t = 1$ to $t = 40$, just like Σ means we take a running sum.

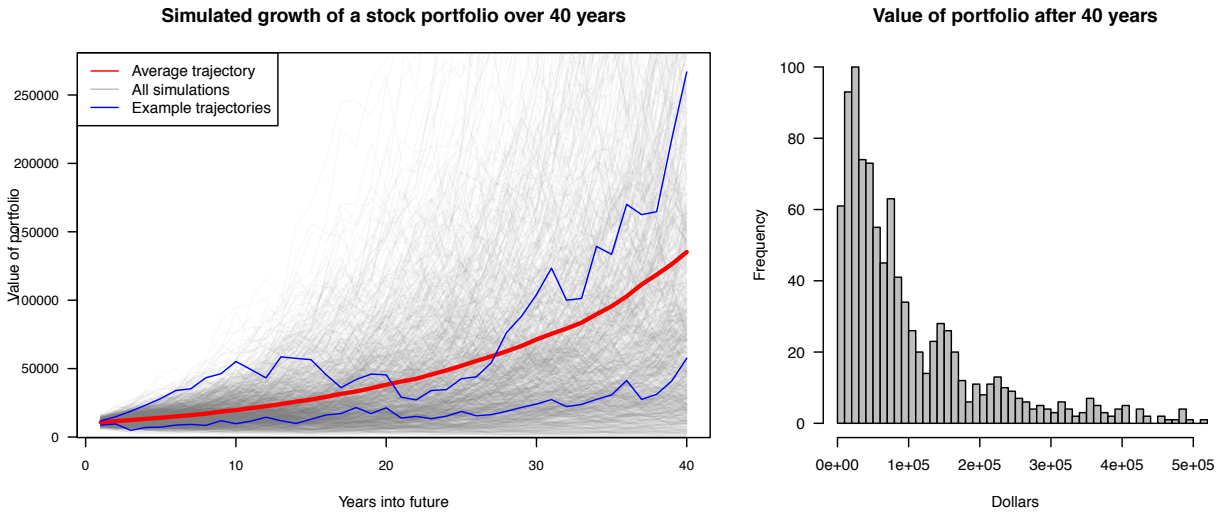


Figure 13.8: Left panel: 1000 simulated trajectories for the growth of a $10,000 stock investment over 40 years, assuming that year stock returns are normally distributed with a mean of 6.5% and a standard deviation of 19.6%. Two individual trajectories (leading to very different outcomes) are highlighted in blue; the average trajectory is shown in red. The right panel shows the simulated probability distribution for W_{40}, the final value of the portfolio after 40 years of random returns.

Of course, we don't know these interest rates. But we do have a probability model for them, whose parameters have been chosen to match the historical record: $X_t \sim N(\mu = 0.065, \sigma^2 = 0.196^2)$. Thus to estimate the probability distribution of the random variable W_{40}, your terminal wealth after 40 years, we will use a Monte Carlo simulation, in which we repeat the following steps many thousands of times:

(1) Simulate random returns from the normal probability model: $X_t \sim N(0.065, 0.196^2)$ for $t = 1, \ldots, 40$.

(2) Starting with year $t = 1$ and ending with year $t = 40$, chain these simulated interest rates together using the simple-interest formula

$$W_{t+1} = W_t \cdot (1 + X_t)$$

to form a single simulated trajectory $W_1, W_2, \ldots, W_{40}$ of wealth.

As a byproduct of this, we get a simulated probability distribution of W_t for all values of t from 1 up to 40.

Figure 13.8 shows 1000 trajectories simulated according to this algorithm, along with the histogram of the 1000 different values of W_{40}, your wealth in 40 years. There are several interesting things to point out about the result:

(1) The *average* trajectory in Figure 13.8 results in a final value of $W_{40} \approx$ $135,000 from your initial $10,000 investment.[6]

[6] Remember that our assumed rates of return are adjusted for inflation, so this corresponds to the purchasing power of $135,000 in today's money. The actual dollar value of this portfolio, as measured in the currency of the future, would be a good deal higher.

(2) But there is tremendous variability about this average trajectory, both over time for a single trajectory, and across all trajectories. To illustrate this point, two simulated trajectories are shown in blue in Figure 13.8: one resulting in a final portfolio of about $250,000, and another resulting in less than $50,000.

(3) The simulated probability distribution of final wealth (right panel of Figure 13.8) was constructed using nothing but normally distributed random variables as inputs. But this distribution is itself highly non-normal.[7] This provides a good example of using Monte Carlo simulation to simulate a complex probability distribution by breaking down into a function of many smaller, simpler parts (in this case, the yearly returns).

(4) The estimated probability that your $10,000 investment will have lost money (net of inflation) after 10 years is about 19%; after 20 years, about 13%; after 40 years, about 6%.

(5) The estimated probability that your investment will grow to $1 million or more after 40 years is about 1%.

[7] In particular it has a long right tail, reflecting the small probability of explosive growth in your investment.

The moral of the story is that the stock market is probably a good way to get rich over time. But there's a nonzero chance of losing money—and the riches come only in the long run, and with a lot of uncertainty about how things will unfold along the way.

Postscript

We've now seen three examples of parametric probability models: a binomial model for airline no-shows, a Poisson model for scoring in a soccer game, and a normal model for annual returns of the stock market. In each case, we chose the parameters of the probability model from real-world data, using simple and obvious criteria (e.g. the overall no-show rate for commercial flights, or the mean return of stocks over the last century).[8] In essence, we performed a naïve form of statistical inference for the parameters of our probability models. This intersection where probability modeling meets data is an exciting place where the big themes of the book all come together.

[8] Technically what we did here was called *moment matching*, wherein we match sample moments (e.g. mean, variance) of the data to the corresponding moments of the probability distribution.

14
Correlated random variables

Joint distributions for discrete variables

In this chapter, we study probability distributions for coupled sets of random variables. We'll first work through a simple example involving two discrete random variables. This will allow us to introduce some basic concepts before turning to more complex examples.

A simple example

The key concept in this chapter is that of a *joint distribution*. We recall that a joint distribution is a list of joint outcomes for two or more variables at once, together with the joint probabilities for each of these outcomes.

Let's look at a simple example, regarding the number of bedrooms and bathrooms for houses and condos currently up for sale in Austin, Texas. Let X_{be} be the number of bedrooms that a house has, and let X_{ba} be the number of bathrooms. The following matrix of joint probabilities specifies a joint probability distribution $P(X_{ba}, X_{be})$:

	Bathrooms				
Bedrooms	1	2	3	4	Marginal
1	0.003	0.001	0.000	0.000	0.004
2	0.068	0.113	0.020	0.000	0.201
3	0.098	0.249	0.126	0.004	0.477
4	0.015	0.068	0.185	0.015	0.283
5	0.002	0.005	0.017	0.006	0.030
6	0.001	0.001	0.002	0.001	0.005
Marginal	0.187	0.437	0.350	0.026	

Using the marginal probabilities alone, we can straightfor-

wardly calculate the expected value and variance for the number of bedrooms and bathrooms. We'll explicitly show the calculation for the expected number of bathrooms, and leave the rest as an exercise to be verified:

$$\begin{aligned} E(X_{ba}) &= 0.187 \cdot 1 + 0.437 \cdot 2 + 0.350 \cdot 3 + 0.026 \cdot 4 \\ &= 2.215 \\ \text{var}(X_{ba}) &= 0.595 \\ E(X_{be}) &= 3.149 \\ \text{var}(X_{be}) &= 0.643 \end{aligned}$$

Covariance

But these moments only tell us about the two variables in isolation, rather than the way they vary together. When two or more variables are in play, the mean and the variance of each one are no longer sufficient to understand what's going on. In this sense, a quantitative relationship is much like a human relationship: you can't describe one by simply listing off facts about the characters involved. You may know that Homer likes donuts, works at the Springfield Nuclear Power Plant, and is fundamentally decent despite being crude, obese, and incompetent. Likewise, you may know that Marge wears her hair in a beehive, despises the *Itchy and Scratchy Show*, and takes an active interest in the local schools. Yet these facts alone tell you little about their marriage. A quantitative relationship is the same way: if you ignore the interactions of the "characters," or individual variables involved, then you will miss the best part of the story.

To quantify the strength of association between two variables, we will calculate their *covariance*. The general definition of covariance is as follows. Suppose that there are N possible joint outcomes for X and Y. Then

$$\text{cov}(X, Y) = E\Big\{ [X - E(X)][Y - E(Y)] \Big\} = \sum_{i=1}^{n} p_i \left[x_i - E(X) \right] \left[y_i - E(Y) \right].$$

This sum is over all possible combinations of joint outcomes for X and Y. In our example about houses for sale, there are 24 terms in the sum, because there are 24 unique combinations for X_{be} and X_{ba}. In the following calculation, a handful of these terms are

shown explicitly, with most shown as ellipses:

$$\begin{aligned}\operatorname{cov}(X_{ba}, X_{be}) &= 0.003 \cdot (1 - 2.215)(1 - 3.149) \\ &\quad + 0.068 \cdot (1 - 2.215)(2 - 3.149) \\ &\quad + \cdots \\ &\quad + 0.185 \cdot (3 - 2.215)(4 - 3.149) \\ &\quad + \cdots \\ &\quad + 0.005 \cdot (4 - 2.215)(6 - 3.149) \\ &\quad \approx 0.285\,.\end{aligned}$$

In this summation, some of the terms are positive and sum of the terms are negative. The positive terms correspond to joint outcomes when the number of bedrooms and bathrooms are on the *same side* of their respective means—that is, both above the mean, or both below it. The negative terms, on the other hand, correspond to outcomes where the two quantities are on *opposite sides* of their respective means. In this case, the "same side" outcomes are more likely than the "opposite side" outcomes, and therefore the covariance is positive.

Correlation as standardized covariance

One difficulty that arises in interpreting covariance is that it depends upon the scale of measurement for the two sets of observations. This isn't so objectionable in the above example (it's hard to imagine what other units we would use). Nonetheless, it's nice to have a unit-free measure of association—especially for a variable like distance, which we could measure in miles or millimeters.

One such scale-invariant measure is the *correlation* between two random variables, which is analogous to the concept of sample correlation between two variables in a data set. The correlation coefficient for two random variables X and Y is just their covariance, rescaled by their respective standard deviations:

$$\operatorname{cor}(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{var}(X)} \cdot \sqrt{\operatorname{var}(Y)}}\,.$$

It runs from -1 (perfect negative correlation) to +1 (perfect positive correlation).

Let's apply this definition to calculation the correlation between the number of bedrooms (X_{be}) and number of bathrooms (X_{ba})

under the joint distribution given earlier:

$$\text{cor}(X_{ba}, X_{be}) = \frac{0.285}{\sqrt{0.595} \cdot \sqrt{0.643}} \approx 0.745\,.$$

The bivariate normal distribution

Heredity and regression to the mean

The history of statistics is intertwined with the history of how scientists came to understand heredity. How strongly do the features of one generation manifest themselves in the next generation? What governs this process, and how can we quantify it mathematically? These questions fascinated scientists of the late 19th and early 20th centuries. As they grappled with them, they also invented a lot of new statistical tools.[1]

One famous study of heredity, by Francis Galton in the 1880's, resulted in the data similar to what you see in the left panel of Figure 14.1.[2] As part of Galton's study of heredity, he collected data on the adult height of parent–child pairs. He wanted to quantify mathematically the extent to which height was inherited from one generation to the next. In looking into this question, Galton noticed some interesting facts about his data.

- Consider the 20 tallest fathers in the data set, highlighted in blue in Figure 14.1. These 20 men had a mean height that was about 6.2 inches above their generation's average height. But the sons of these 20 men had an average height that was only 2.8 inches above their generation's average height. Thus the sons of very tall men were taller than average, but not by as much as their fathers were.

- Now consider the 20 shortest fathers in the data set, highlighted in red in Figure 14.1. These 20 men had a mean height that was about 6.9 inches below their generation's average height. But the sons of these 20 men had an average height that was only 3.3 inches below their generation's average height. Thus the sons of very short men were shorter than average, but not by as much as their fathers were.

Galton called this phenomenon "regression towards mediocrity," where "mediocre" should be understood in the sense of "average." Galton's proposed explanation for this phenomenon

[1] It's important to mention that many these developments were pursued at least partially in the name of the eugenics movement. While the mathematical tools left to us as a result of these studies remain valuable, their history is not something to be unreservedly proud of. If you're interested in reading more about this, try the following article: "Sir Francis Galton and the birth of eugenics," by N.W. Gilham. Annual Review of Genetics, 2001, 35:83-101.

[2] This data was actually collected an analyzed by Galton's protégé, Karl Pearson. But Galton worked with very similar data, so we'll pretend for the purposes of exposition that this was Galton's data, since he was the first one to follow this line of thought.

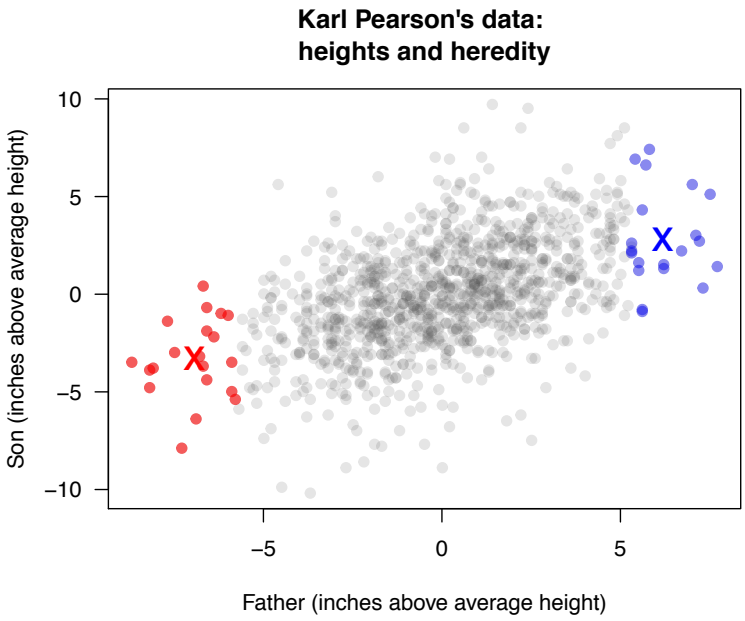


Figure 14.1: Karl Pearson's data on the height of fathers and their adult sons. The 20 tallest fathers (and their sons) are highlighted in blue, with the bivariate mean of this group shown as a blue X. Similarly, the 20 shortest fathers (and their sons) are highlighted in red, with the bivariate mean of this group shown as a red X. The points show fathers and sons only, to avoid any confounding due to sex. We've also mean-centered the data, by subtracting the average height of all fathers from each father's height, and the average height of all sons from each son's height. This doesn't change the shape of the point cloud; it merely re-centers it at $(0, 0)$. This accounts for the fact that the sons' generation, on average, was about an inch taller than the fathers' generation—possibly due to improving standards of health and nutrition.

turned out to be incorrect, but today we understand it as a product of genetics. It's hard to explain exactly why this happens without getting deep into the weeds on multifactorial inheritance, but the rough idea is the following. (We'll focus on the tallest fathers in the data set, but the same line of reasoning works for the shortest fathers, too.)

- Very tall people, like Yao Ming at right, turn out that way for a combination of two reasons: height genes and height luck. (Here "luck" is used to encompass both environmental forces as well as some details of multifactorial inheritance not worth going into here.)
- Therefore, our selected group of very tall people (the blue dots in Figure 14.1) is biased in two ways: extreme height genes *and* extreme height luck.
- These very tall people pass on their height genes to their children, but not their height luck.
- Height luck will average out in the next generation. Therefore, the children of very tall parents will still be tall (be-

Figure 14.2: Yao Ming makes J.J. Watt (6'5" tall, 290 pounds) look like a child.

cause of genes), but not as tell as their parents (because they weren't as lucky, on average).

Notice that this isn't a claim about causality. It is not true that the children of very tall people are likely to have less extreme "height luck" *because* their parents had a lot of it. Rather, these children are likely to have less luck than their parents because extreme luck is, by definition, rare—and they are no more likely to experience this luck than any randomly selected group of people.

This phenomenon that we've observed about height and heredity is actually quite general. Take any pair of correlated measurements. If one measurement is extreme, then the other measurement will tend to be closer to the average. Today we call this *regression to the mean*. Just as Galton did in 1889, we can make this idea mathematically precise using a probability model called the bivariate normal distribution. This requires a short detour.

Notation for the bivariate normal

The *bivariate normal distribution* a parametric probability model for the joint distribution of two correlated random variables X_1 and X_2. You'll recall that the ordinary normal distribution is a distribution for one variable with two parameters: a mean and a variance. The bivariate normal distribution is for two variables (X_1 and X_2), and it has five parameters:

- The mean and variance of the first random variable: $\mu_1 = E(X_1)$ and $\sigma_1^2 = \text{var}(X_1)$.
- The mean and variance of the second random variable: $\mu_2 = E(X_2)$ and $\sigma_2^2 = \text{var}(X_2)$.
- The covariance between X_1 and X_2, which we denote as σ_{12}.

Equivalently, we can specify the correlation instead of the covariance. We recall that the correlation is just the covariance rescaled by both standard deviations:

$$\rho = \frac{\text{cov}(X_1, X_2)}{\text{sd}(X_1) \cdot \text{sd}(X_2)} = \frac{\sigma_{12}}{\sigma_1 \cdot \sigma_2}.$$

In practice will usually instead refer to the standard deviations σ_1 and σ_2 and correlation ρ rather than the variances and covariances, and use the shorthand $(X_1, X_2) \sim N(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$.

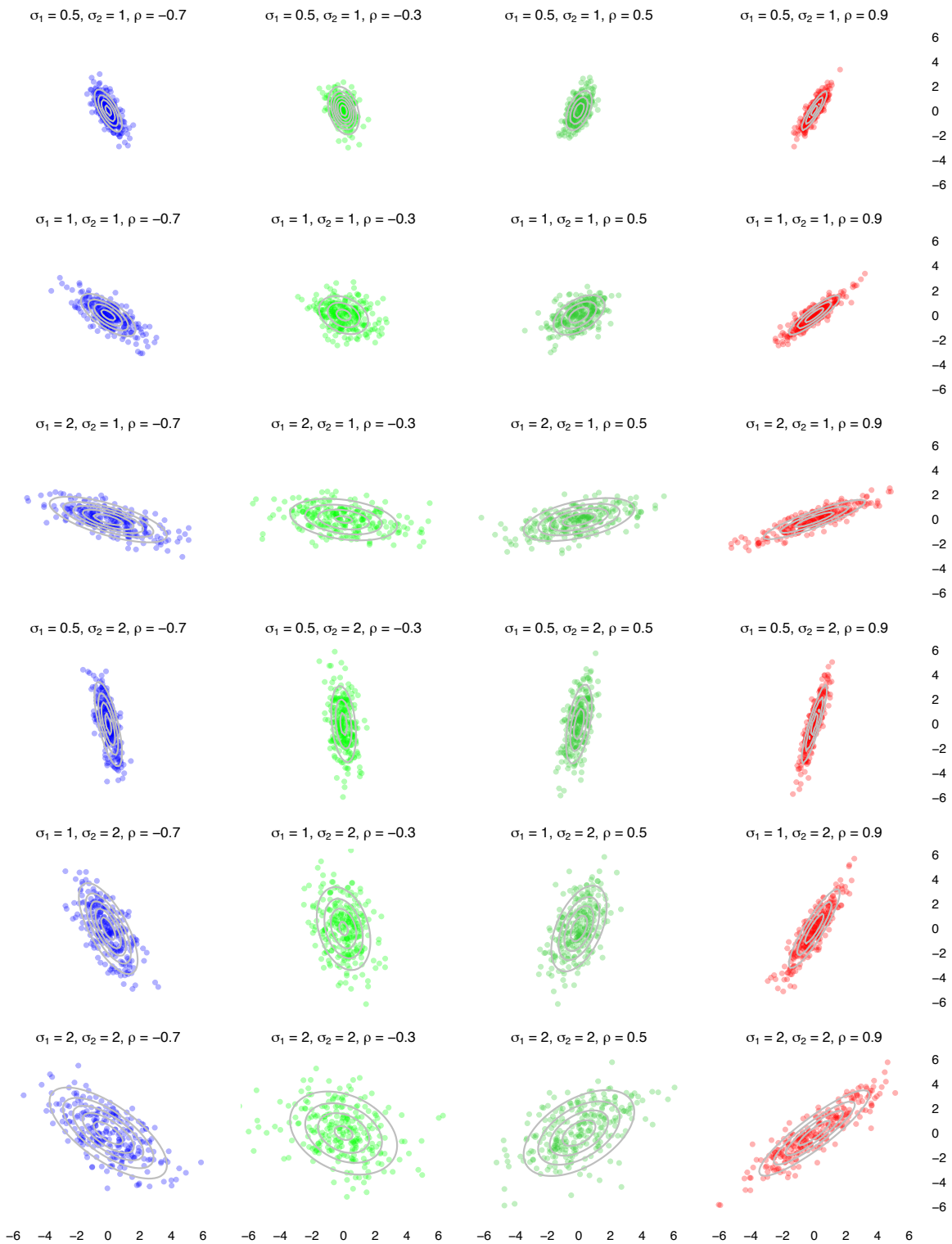


Figure 14.3: 24 examples of a bivariate normal distribution (250 samples in each plot).

We can also write a bivariate normal distribution using matrix–vector notation, to emphasize the fact that $X = (X_1, X_2)$ is a random vector:

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \right),$$

or simply $X \sim \mathrm{N}(\mu, \Sigma)$, where μ is the mean vector and Σ is called the covariance matrix.

The bivariate normal distribution has the nice property that each of its two marginal distributions are ordinary normal distributions. That is, if we ignore X_2 and look only at X_1, we find that $X_1 \sim N(\mu_1, \sigma_1^2)$. Similarly, if we ignore X_1 and look only at X_2, we find that $X_2 \sim N(\mu_2, \sigma_2^2)$.

Visualizing the bivariate normal distribution

Figure 14.3 provides some intuition for how the various parameters of the bivariate normal distribution affect its shape. Here we see 24 examples of a bivariate normal distribution with different combinations of standard deviations and correlations. In each panel, 250 random samples of (X_1, X_2) from the corresponding bivariate normal distribution are shown:

- Moving down the rows from top to bottom, the standard deviations of the two variables change, while the correlation remains constant within a column.
- Moving across the columns from left to right, the correlation changes from negative to positive, while the standard deviations of the two variables remain the same within a row.

The mean of both variables is 0 in all 24 panels. Changing either mean would translate the point cloud so that it was centered somewhere else, but would not change the shape of the cloud.

Each panel of Figure 14.3 also shows a *contour plot* of the probability density function for the corresponding bivariate normal distribution, overlaid in grey. We read these contours in a manner similar to how we would on an ordinary contour map: they tell us how high we are on the three-dimensional surface of the bivariate normal density function, like the one shown at right.

To interpret this density function, imagine specifying two intervals, one for X_1 and another for X_2, and asking: what is the probability that both X_1 and X_2 fall in their respective intervals?

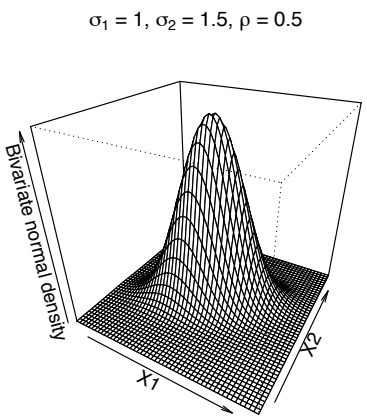


Figure 14.4: A three-dimensional wireframe plot of a bivariate normal density function.

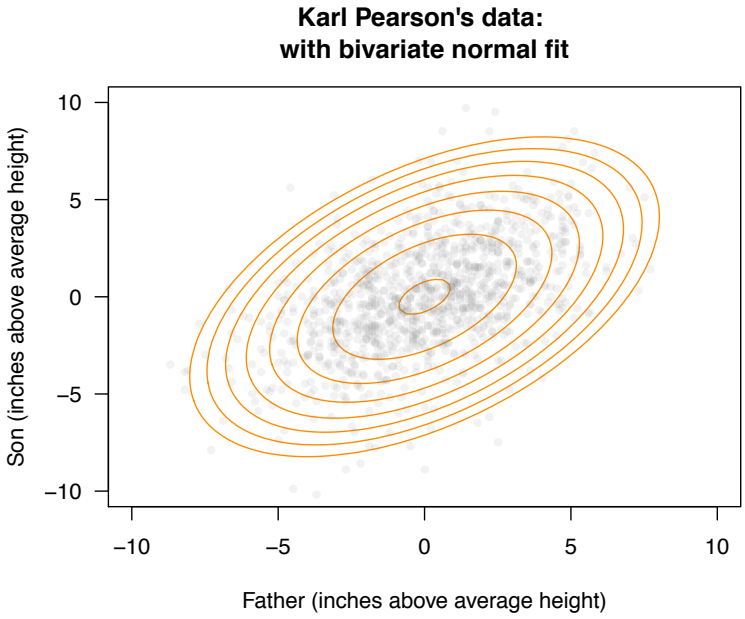


Figure 14.5: Best-fitting bivariate normal distribution for Karl Pearson's height data based on the sample standard deviations and sample correlation.

Written mathematically, we want to know the joint probability $P[X_1 \in (a,b), X_2 \in (c,d)]$. The two intervals (a,b) and (c,d) define a rectangle in the (X_1, X_2) plane (i.e. the "floor" of the 3D plot in Figure 14.4). To calculate this joint probability, we ask: what is the volume under the density function that sits above this rectangle? This generalizes the "area under the curve" interpretation of a density function for a single random variable.

Figure 14.5 shows the best fitting bivariate normal distribution to the heights data:

$$(X_1, X_2) \sim N(\mu_1 = 0,\ \mu_2 = 0,\ \sigma_1 = 2.75,\ \sigma_2 = 2.82,\ \rho = 0.5)\,.$$

Remember that both means are zero because we centered the data.

Conditional distributions for the bivariate normal

Take any pair of correlated random variables X_1 and X_2. Because they are correlated, the value of one variable gives us information about the value of the second variable. To make this precise, say we fix the value of X_1 at some known value x_1. What is the conditional probability distribution of X_2, given that $X_1 = x_1$? In our heights example, this would be like asking: what is the dis-

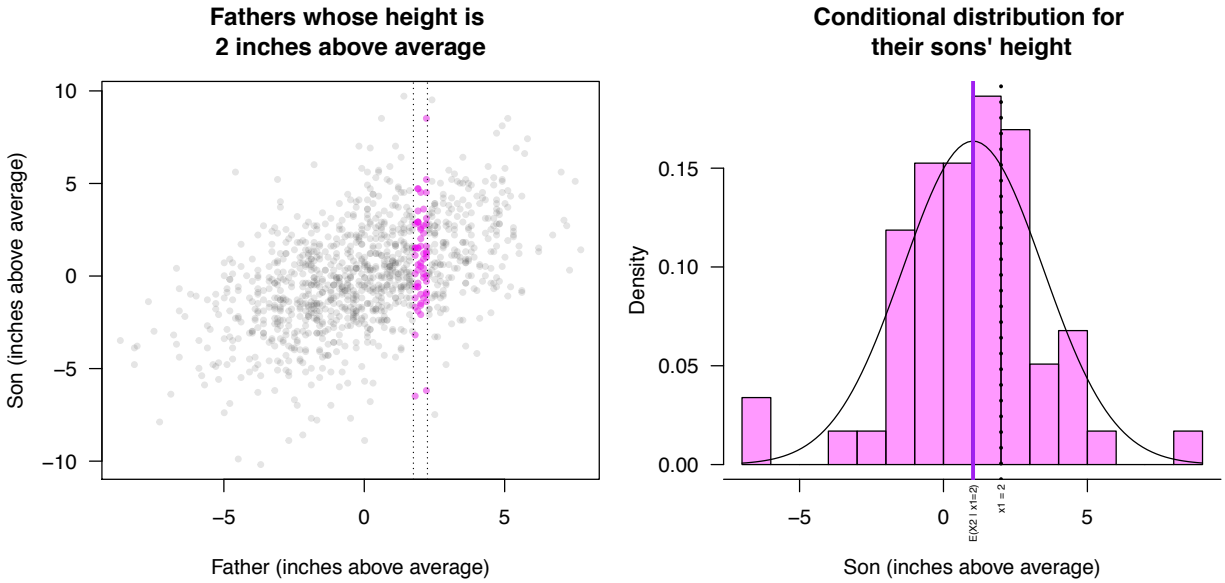


Figure 14.6: Left: father–son pairs where the father's height is about 2 inches above average are highlighted in purple. Right: the histogram of the sons' height, together with the conditional distribution $P(X_2 \mid X_1 = 2)$ predicted by the bivariate normal fit to the joint distribution for (X_1, X_2). The sons' average height, $E(X_2 \mid X_1 = 2)$ (purple line) is shrunk back towards 0 compared to the fathers' height of 2 inches above average (black dotted line). This illustrates regression to the mean.

tribution for the heights of sons (X_2) for fathers whose height is 2 inches above the mean ($X_1 = 2$)?

If X_1 and X_2 follow a bivariate normal distribution, i.e.

$$(X_1, X_2) \sim N(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)\,,$$

then this question is easy to answer. It turns out that the conditional probability distribution $P(X_2 \mid X_1 = x_1)$ is an ordinary normal distribution, with mean and variance

$$E(X_2 \mid X_1 = x_1) = \mu_2 + \rho \cdot \frac{\sigma_2}{\sigma_1} \cdot (x_1 - \mu_1) \tag{14.1}$$

$$\text{var}(X_2 \mid X_1 = x_1) = \sigma_2^2 \cdot (1 - \rho^2)\,, \tag{14.2}$$

where σ_1, σ_2, and ρ are the standard deviations of the two variables and their correlation, respectively. You'll notice that the conditional mean $E(X_2 \mid X_1 = x_1)$ is a linear function of x_1, the assumed value for X_1. Galton called this the regression line—that is, the line that describes where we should expect to find X_2 for a given value of X_1.[3]

This fact brings us straight back to the concept of regression to the mean. Let's re-arrange Equation 14.1 to re-express the condi-

[3] This use of the term "regression" is the origin of the phrase "linear regression" to describe the process of fitting lines to data. But keep in mind that linear regression (in the sense of fitting equations to data) actually predates Galton's use of the term by almost 100 years. So while Galton's reasoning using the bivariate normal distribution does provide the historical underpinnings for the *term* regression in the sense that we used it earlier in the book, it is not the origin for the idea of curve fitting.

tional mean in a slightly different way:

$$\frac{E(X_2 \mid X_1 = x_1) - \mu_2}{\sigma_2} = \rho \cdot \left(\frac{x_1 - \mu_1}{\sigma_1} \right) . \tag{14.3}$$

The left-hand side asks: how many standard deviations is X_2 expected to be above (or below) its mean, given that $X_1 = x_1$? The right-hand side answers: the number of standard deviations that x_1 was above (or below) its mean, *discounted by a factor of* ρ. Because ρ can never exceed 1, we expect that X_2 will be "shrunk" a bit closer to its mean than x_1 was—and the weaker the correlation between the two variables, the stronger this shrinkage effect is. Equation 14.3 therefore provides a formal mathematical description of regression to the mean. In the extreme case of $\rho = 1$, there is no regression to the mean at all.

Let's return to the data on the heights of fathers and sons and use this result to measure the magnitude of the regression-to-mean effect. Specifically, let's consider fathers whose heights are about 2 inches above average ($X_1 = 2$). Using Equation 14.1 together with the parameters of the best-fitting bivariate normal distribution from Figure 14.5, we find that:

$$E(X_2 \mid X_1 = 2) = \rho \cdot \frac{\sigma_2}{\sigma_1} \cdot 2 = 0.5 \cdot \frac{2.81}{2.75} \cdot 2 \approx 1.03 .$$

That is, the sons should be about 1 inch taller than average for their generation (rather than 2 inches taller, as their fathers were).

Sure enough, as Figure 14.6 shows, this prediction is borne out. We have highlight all the fathers in the data set who are approximately inches above average (purple dots, left panel). On the right, we see a histogram for the height of their sons. This histogram shows us the conditional distribution $P(X_2 \mid X_1 = 2)$, together with the normal distribution whose mean and variance are calculated using the formulas for the conditional mean and variance in Equations 14.1 and 14.2. Given the small sample size ($n = 59$), the normal distribution looks like a good fit—in particularly, it captures the regression-to-the-mean effect, correctly predicting that the conditional distribution will be centered around $X_2 = 1$.

Further applications of the bivariate normal

Example 1: regression to the mean in baseball

Regression to the mean is ubiquitous in professional sports. If you're a baseball fan, you may have heard of the "sophomore jinx":

> A sophomore jinx is the popularly held belief that after a successful rookie season, a player in his second year will be jinxed and not have the same success. Most players suffer the "sophomore jinx" as scouting reports on the former rookie are now available and his weaknesses are known around the league.[4]

[4] http://www.baseball-reference.com/bullpen/Sophomore_jinx

This idea comes up all the time in discussion among baseball players, coaches, and journalists:

> Fresh off one of their best seasons in decades, the Cubs look primed to compete for a division title and more in 2016. As rookies in 2015, Kris Bryant, Addison Russell, Jorge Soler and Kyle Schwarber had significant roles in the success and next year, Cubs manager Joe Maddon is looking to help them avoid the dreaded sophomore jinx. "I think the sophomore jinx is all about the other team adjusting to you and then you don't adjust back," Maddon said Tuesday at the Winter Meetings. "So the point would be that we need to be prepared to adjust back. I think that's my definition of the sophomore jinx."[5]

[5] "Focus for Joe Maddon: Avoiding 'sophomore jinx' with young Cubs." Matt Snyder, CBSsports.com, December 8, 2015.

The sophomore jinx—that outstanding rookies tend not to do quite as well in their second seasons—is indeed real. But it can be explained in terms of regression to the mean! Recall our definition of this phenomenon, from several pages ago: "Take any pair of correlated measurements. If one measurement is extreme, then the other measurement will tend to be closer to the average."

Let's apply this idea to baseball data. Say that X_1 is batting average of a baseball player last season, and that X_2 is that same player's batting average this season. Surely these variables are correlated, because more skillful players will have higher averages overall. But the correlation will be imperfect (less than one), because luck plays a role in a player's batting average, too.

Now focus on the players with the very best batting averages last year—that is, those where X_1 is the most extreme. Among players in this group, we should expect that X_2 will be less extreme overall than X_1. Again, this isn't a claim about good performance last year *causing* worse performance this year. It's just that

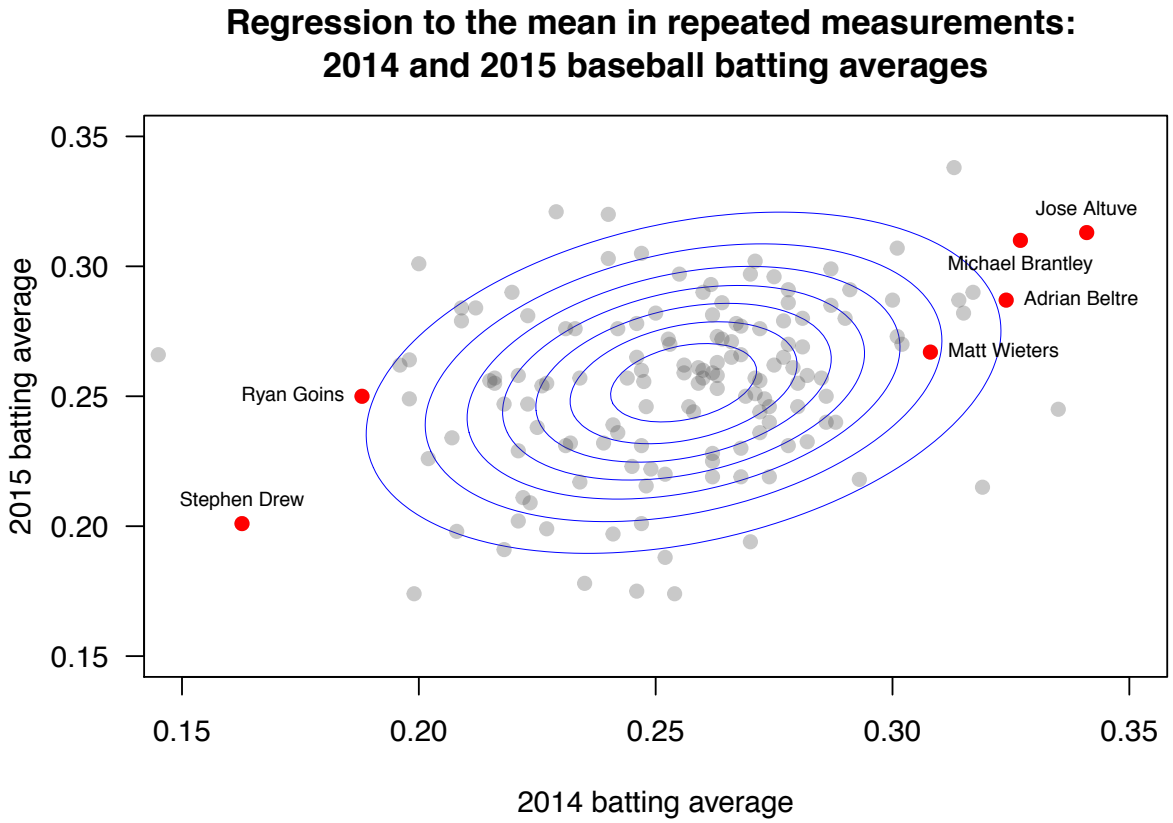


Figure 14.7: Baseball batting averages in the 2014 and 2015 seasons for all players with at least 100 at-bats in both years.

last year's very best performers were both lucky and good—and while they might still be good this year, they are no more likely to be lucky than any other group of baseball players.[6]

Figure 14.7 shows this phenomenon in action. Here we see the batting averages across the 2014 and 2015 baseball seasons for all players with at least 100 at-bats in both seasons. The figure highlights some of the very best and very worst performers in 2014. Sure enough, although 2014's best were still good in 2015, they weren't *as good* as they had been the previous year. Similarly, the very worst performers in 2014 were still not very good in 2015, but they weren't as bad as they'd been the previous year. This is another great example of regression to the mean.

[6] Although it's possible Joe Maddon's theory of "not adjusting back" might be partially true, too, the mere existence of the "sophomore jinx" phenomenon certainly doesn't prove it.

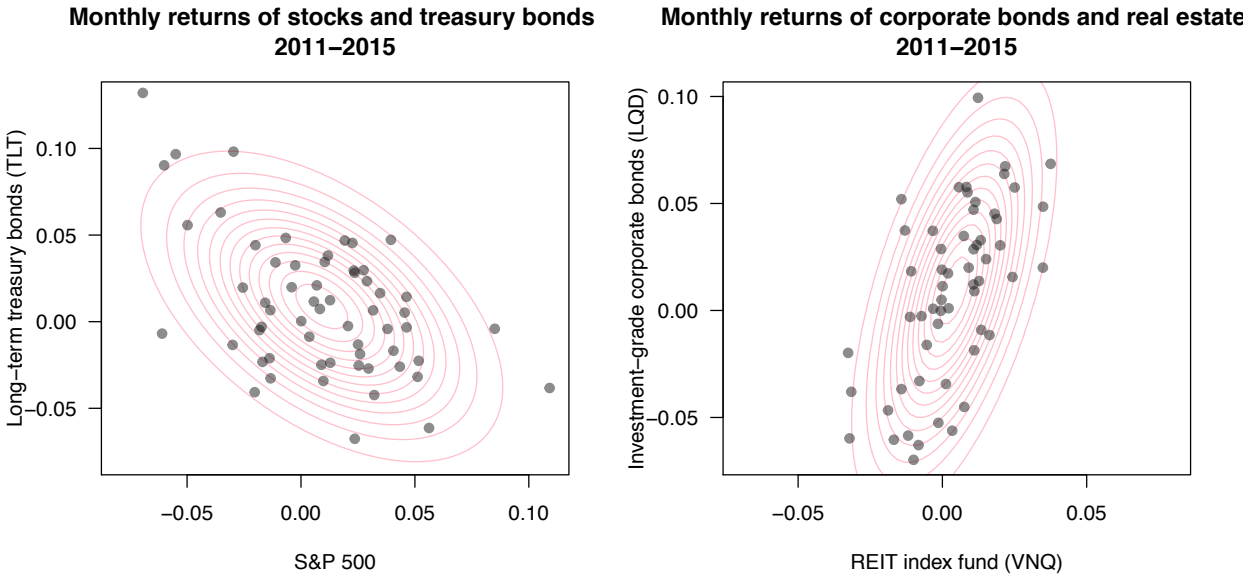


Figure 14.8: Correlation between stocks and government bonds (left); correlation between corporate bonds and real estate (right).

Example 2: stocks and bonds.

The bivariate normal distribution is useful for more than simply describing regression to the mean. We can also use it as a building block for describing the joint probability distribution for two correlated random variables. As a final example, let's look at correlation between different pairs of financial assets.

First, say that X_1 is the return on the S&P 500 index next month, while X_2 is the return on 30-year treasury bond next month.[7] These two variables are almost sure to be correlated, although the magnitude and even the direction of this correlation has changed a lot over the last century. The conventional explanation for this is the so-called "flight to quality" effect: when stock prices plummet, investors get scared and pile their money into safer assets (like bonds), thereby driving up the price of those safer assets. This effect will typically produce a negative correlation between the returns of stocks and bonds held over a similar period.[8] The left panel of Figure 14.8 shows the 2011-2015 monthly returns for long-term U.S. Treasury bonds versus the S&P 500 stock index, together with the best-fitting bivariate normal approximation.

Next, consider the right panel of Figure 14.8, which shows returns for real-estate investment trusts (X_1) and corporate bonds

[7] Recall that a Treasury bond entailed lending money to the U.S. federal government and collecting interest in return.

[8] This need not happen. In fact, a "flight to quality" effect can also produce a positive correlation between U.S. stocks and bonds. If you're interested in more detail, see this short article written by two economists at the Reserve Bank of Australia.

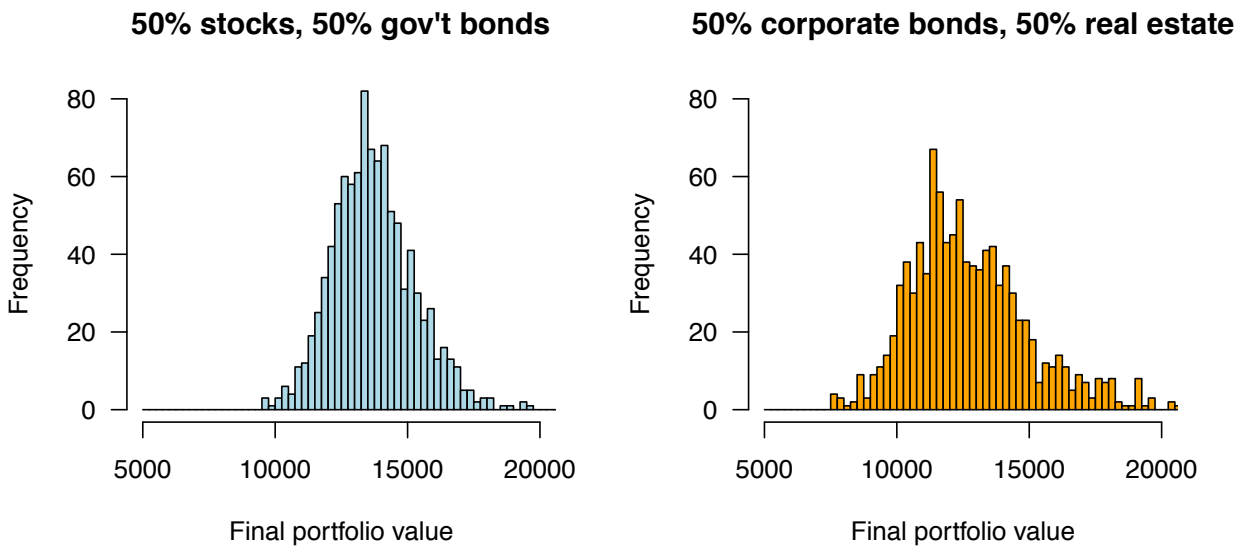


Figure 14.9: Final value of 36-month investments in 50/50 mixes of: (1) stocks and government bonds (left), and (2) corporate bonds and real estate (right).

(X_2). These assets' monthly returns were positively correlated, presumably because they both respond in similar ways to underlying macroeconomic forces.

How do these patterns of correlation affect the medium-term growth of a portfolio of mixed assets? To understand this, we'll run a Monte Carlo simulation where we chain together the results of 36 months (3 years) of investment. We'll compare two portfolios with an initial value of $W_0 = \$10,000$: a mix of stocks ($X_1$) and government bonds ($X_2$), versus a mix of real-estate (X_1) and corporate bonds (X_2). We'll let $W_{t,1}$ and $W_{t,2}$ denote the amount of money you have at step t in assets 1 and 2, respectively. Each 36-month period will be simulated as follows, starting with month $t = 1$ and ending with month $t = 36$.

(1) Simulate a random return for month t from the bivariate normal probability model: $(X_{t1}, X_{t2)} \sim N(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$.

(2) Update the value of your investment to account for the period-t returns in each asset:

$$W_{t+1,i} = W_{t,i} \cdot (1 + X_{t,i})$$

for $i = 1, 2$.

At every step, your current total wealth is $W_t = W_{t,1} + W_{t,2}$. For the sake of illustration, we'll assume that the initial allocation is a 50/50 mix, so that $W_{0,1} = W_{0,2} = \$5,000$.

Figure 14.9 shows the results of this simulation, assuming that returns following the bivariate normal distributions fit to the data in Figure 14.8. Clearly the 50/50 mix of stocks and government bonds is preferred under this scenario: it has both a higher return and a lower variance than the mix of corporate bonds and real-estate. In particular, in the second portfolio, the positive correlation between corporate bonds and real estate is especially troublesome. This results in a portfolio with far higher variance than necessary, because the ups and the downs tend to occur together.

Two major caveats here are: (1) the assumption that future returns will be statistically similar to past returns, and (2) that we can describe correlation among pairs of asset returns using a bivariate normal. Both of these assumptions can be challenged. Therefore, it's better to think of simulations like these as a way of building scenarios under various assumptions about future performance, rather than as a firm guide to what it is likely to happen.

Functions of random variables (advanced topic)

A VERY important set of equations in probability theory describes what happens when you construct a new random variable as a linear combination of other random variables—that is, when

$$W = aX + bY + c$$

for some random variables X and Y and some constants a, b, and c.

The fundamental question here is: how does *joint* variation in X and Y (that is, correlation) influence the behavior a random variable formed by adding X and Y together? To jump straight to the point, it turns out that

$$E(W) = aE(X) + bE(Y) + c \tag{14.4}$$

$$\text{var}(W) = a^2 \text{ var}(X) + b^2 \text{ var}(Y) + 2ab \text{ cov}(X, Y)\,. \tag{14.5}$$

Why would you care about a linear combination of random variables? Consider a few examples:

- You know the distribution for X, the number of points a basketball team will score in one quarter of play. Then the

random variable describing the points the team will score in four quarters of play is $W = 4x$.

- A weather forecaster specifies a probability distribution for tomorrow's temperature in Celsius (a random variable, C). You can compute the moments of C, but you want to convert to Fahrenheit (another random variable, F). Then F is also a random variable, and is a linear combination of the one you already know: $F = (9/5)C + 32$.

- You know the joint distribution describing your uncertainty as to the future prices of two stocks X and Y. A portfolio of stocks is a linear combination of the two; if you buy 100 shares of the first and 200 of the second, then

$$W = 100X + 200Y$$

  is a random variable describing the value of your portfolio.

- Your future grade on the statistics midterm is X_1, and your future grade on the final is X_2. You describe your uncertainty for these two random variables with some joint distribution. If the midterm counts 40% and the final 60%, then your final course grade is the random variable

$$C = 0.4X_1 + 0.6X_2\,,$$

  a linear combination of your midterm and final grades.

- The speed of Rafael Nadal's slice serve is a random variable S_1. The speed on his flat serve is S_2. If Rafa hits 70% slice serves, his opponent should anticipate a random service speed equal to $0.7S_1 + 0.3S_2$.

In all five cases, it is useful to express the moments of the new random variable in terms of the moments of the original ones. This saves you a lot of calculational headaches! We'll now go through the mathematics of deriving Equations (14.4) and (14.5).

Multiplying a random variable by a constant

Let's first examine what happens when you make a new random variable W by multiplying some other random variable X by a constant:

$$W = aX\,.$$

This expression means that, whenever $X = x$, we have $W = ax$. Therefore, if X takes on values $x_1, \ldots, x_n$ with probability $p_1, \ldots, p_n$, then we know that

$$E(X) = \sum_{i=1}^{n} x_i p_i \,,$$

and so

$$E(W) = \sum_{i=1}^{n} a x_i p_i = a \sum_{i=1}^{n} x_i p_i = aE(X) \,.$$

The constant a simply comes out in front of the original expected value. Mathematically speaking, this means that the expectation is a linear function of a random variable.

The variance of W can be calculated in the same way. By definition,

$$\text{var}(X) = \sum_{i=1}^{n} p_i \{x_i - E(X)\}^2 \,.$$

Therefore,

$$\begin{aligned}
\text{var}(W) &= \sum_{i=1}^{n} p_i \{a x_i - E(W)\}^2 \\
&= \sum_{i=1}^{n} p_i \{a x_i - aE(X)\}^2 \\
&= \sum_{i=1}^{n} p_i a^2 \{x_i - E(X)\}^2 \\
&= a^2 \sum_{i=1}^{n} p_i \{x_i - E(X)\}^2 \\
&= a^2 \, \text{var}(X)
\end{aligned}$$

Now we have a factor of a^2 out front.

What if, in addition to multiplying X by a constant a, we also add another constant c to the result? This would give us

$$W = aX + c \,.$$

To calculate the moments of this random variable, revisit the above derivations on your own, adding in a constant term of c where appropriate. You'll soon convince yourself that

$$\begin{aligned}
E(W) &= aE(X) + c \\
\text{var}(W) &= a^2 \text{var}(X) \,.
\end{aligned}$$

The constant simply gets added to the expected value, but doesn't change the variance at all.

A linear combination of two random variables

Suppose X and Y are two random variables, and we define a new random variable as $W = aX + bY$ for real numbers a and b. Then

$$\begin{aligned} E(W) &= \sum_{i=1}^{n} p_i\{ax_i + by_i\} \\ &= \sum_{i=1}^{n} p_i a x_i + \sum_{i=1}^{n} p_i b y_i \\ &= a\sum_{i=1}^{n} p_i x_i + b\sum_{i=1}^{n} p_i y_i \\ &= aE(X) + b(E(Y)\,. \end{aligned}$$

Again, the expectation operator is linear.

The variance of W, however, takes a bit more algebra:

$$\begin{aligned} \text{var}(W) &= \sum_{i=1}^{n} p_i\Big\{[ax_i + by_i] - [aE(X) + bE(Y)]\Big\}^2 \\ &= \sum_{i=1}^{n} p_i\Big\{[ax_i - aE(X)] + [by_i - bE(Y)]\Big\}^2 \\ &= \sum_{i=1}^{n} p_i\Big\{[ax_i - aE(X)]^2 + [by_i - bE(Y)]^2 + 2ab[x_i - E(X)][y_i - E(Y)]\Big\} \\ &= \sum_{i=1}^{n} p_i[ax_i - aE(X)]^2 + \sum_{i=1}^{n} p_i[by_i - bE(Y)]^2 + \sum_{i=1}^{n} p_i 2ab[x_i - E(X)][y_i - E(Y)] \\ &= \text{var}(aX) + \text{var}(bY) + 2ab\text{cov}(X,Y) \\ &= a^2\text{var}(X) + b^2\text{var}(Y) + 2ab\text{cov}(X,Y) \end{aligned}$$

The covariance of X and Y strongly influences the variance of their linear combination. If the covariance is positive, then the variance of the linear combination is *more than* the sum of the two individual variances. If the covariance is negative, then the variance of the linear combination is *less than* the sum of the two individual variances.

An example: portfolio choice under risk aversion

Let's revisit the portfolio-choice problem posed above. Say you plan to allocate half your money to one asset X, and the other half to some different asset Y. Look at Equations (14.4) and (14.5), which specify the expected value and variance of your portfolio in terms of the moments of the joint distribution for X and Y. If you are a risk-averse investor, would you prefer to hold two assets with a positive covariance or a negative covariance?

To make things concrete, let's imagine that the joint distribution for X and Y is given in the table at right. Each row is a possible joint outcome for X and Y: the first column lists the possible values of X; the second, the possible values of Y; and the third, the probabilities for each joint outcome. You should interpret the numbers in the X and Y columns as the value of \$1 at the end of the investment period—for example, after one year. If $X = 1.1$ after a year, then your holdings of that stock gained 10% in value.

x	y	$P(x,y)$
1.0	1.0	0.15
1.0	1.1	0.10
1.0	1.2	0.05
1.1	1.0	0.10
1.1	1.1	0.20
1.1	1.2	0.10
1.2	1.0	0.05
1.2	1.1	0.10
1.2	1.2	0.15

Table 14.1: Positive covariance.

Under this joint distribution, a single dollar invested in a portfolio with a 50/50 allocation between X and Y is a random variable W. This random variable has an expected value of 1.1 and variance

$$\begin{aligned}\text{var}(W) &= 0.5^2\text{var}(X) + 0.5^2\text{var}(Y) + 2 \cdot 0.5^2 \cdot \text{cov}(X,Y) \\ &= 0.5^2 \cdot 0.006 + 0.5^2 \cdot 0.006 + 2 \cdot 0.5^2 \cdot (0.002) \\ &= 0.004\,,\end{aligned}$$

for a standard deviation of $\sqrt{0.004}$, or about 6.3%.

What if, on the other hand, the asset returns were negatively correlated, as they are in the table at right? (Notice which entries have been switched, compared to the previous distribution.)

x	y	$P(x,y)$
1.0	1.0	0.05
1.0	1.1	0.10
1.0	1.2	0.15
1.1	1.0	0.10
1.1	1.1	0.20
1.1	1.2	0.10
1.2	1.0	0.15
1.2	1.1	0.10
1.2	1.2	0.05

Table 14.2: Negative covariance.

Under this new joint distribution, the expected value of \$1 invested in a 50/50 portfolio is still 1.1. But since the covariance between X and Y is now negative, the variance of the portfolio changes:

$$\begin{aligned}\text{var}(W) &= 0.5^2\text{var}(X) + 0.5^2\text{var}(Y) + 2 \cdot 0.5^2 \cdot \text{cov}(X,Y) \\ &= 0.5^2 \cdot 0.006 + 0.5^2 \cdot 0.006 + 2 \cdot 0.5^2 \cdot (-0.002) \\ &= 0.002\,,\end{aligned}$$

for a standard deviation of $\sqrt{0.002}$, or about 4.5%. Same expected return, but lower variance, and therefore more attractive to a risk-averse investor!

What's going on here? Intuitively, under the first portfolio, where X and Y are positively correlated, the bad days for X and Y tend to occur together. So do the good days. (When it rains, it pours; when it's sunny, it's 100 degrees.) But under the second portfolio, where X and Y are negatively correlated, the bad days and good days tend to cancel each other out. This results in a lower overall level of risk.

The morals of the story are:

1. Correlation creates extra variance.
2. Diversify! (Extra variance hurts your compounded rate of return.)

15
Generalized linear models

Binary responses

In many situations, we would like to predict the outcome of a binary event, given some relevant information:

- Given the pattern of word usage and punctuation in an e-mail, is it likely to be spam?
- Given the temperature, pressure, and cloud cover on Christmas Eve, is it likely to snow on Christmas Day?
- Given a person's credit history and income, is he or she likely to default on a mortgage loan?

In all of these cases, the y variable is the answer to a yes-or-no question. This is a bit different to the kinds of problems we've become used to seeing, where the response is a real number.

Nonetheless, we can still use regression for these problems. Let's suppose, for simplicity's sake, that we have only one predictor x, and that we let $y_i = 1$ for a "yes" and $y_i = 0$ for a "no." One naïve way of forecasting y is simply to plunge ahead with the basic, one-variable regression equation:

$$\hat{y}_i = \mathrm{E}(y_i \mid x_i) = \beta_0 + \beta_1 x_i \,.$$

Since y_i can only take the values 0 or 1, the expected value of y_i is simply a weighted average of these two cases:

$$\begin{aligned} \mathrm{E}(y_i \mid x_i) &= 1 \cdot P(y_i = 1 \mid x_i) + 0 \cdot P(y_i = 0 \mid x_i) \\ &= P(y_i = 1 \mid x_i) \end{aligned}$$

Therefore, the regression equation is just a linear model for the conditional probability that $y_i = 1$, given the predictor x_i:

$$P(y_i = 1 \mid x_i) = \beta_0 + \beta_1 x_i \,.$$

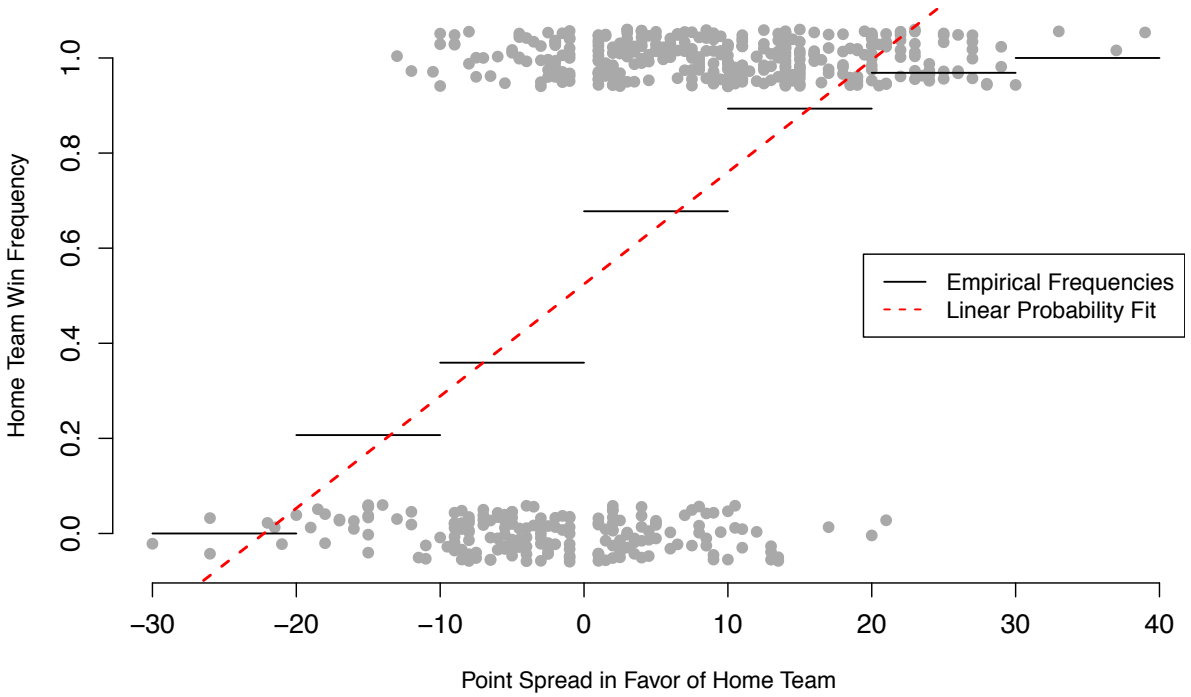


Figure 15.1: Win frequency versus point spread for 553 NCAA basketball games. Actual wins are plotted as 1's and actual losses as zeros. Some artificial vertical jitter has been added to the 1's and 0's to allow the dots to be distinguished from one another.

This model allows us to plug in some value of x_i and read off the forecasted probability of a "yes" answer to whatever yes-or-no question is being posed. It is often called the linear probability model, since the probability of a "yes" varies linearly with x.

Let's try fitting it to some example data to understand how this kind of model behaves. In Table 15.1 on page 297, we see an excerpt of a data set on 553 men's college-basketball games. Our y variable is whether the home team won ($y_i = 1$) or lost ($y_i = 0$). Our x variable is the Las Vegas "point spread" in favor of the home team. The spread indicates the betting market's collective opinion about the home team's expected margin of victory—or defeat, if the spread is negative. Large spreads indicate that one team is heavily favored to win. It is therefore natural to use the Vegas spread to predict the probability of a home-team victory in any particular game.

Figure 15.1 shows each of the 553 results in the data set. The

home-team point spread is plotted on the x-axis, while the result of the game is plotted on the y-axis. A home-team win is plotted as a 1, and a loss as a 0. A bit of artificial vertical jitter has been added to the 1's and 0's, just so you can distinguish the individual dots.

The horizontal black lines indicate empirical win frequencies for point spreads in the given range. For example, home teams won about 65% of the time when they were favored by more than 0 points, but less than 10. Similarly, when home teams were 10–20 point underdogs, they won only about 20% of the time.

Finally, the dotted red line is the linear probability fit:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.524435   0.019040   27.54   <2e-16 ***
spread      0.023566   0.001577   14.94   <2e-16 ***
---
Residual standard error: 0.4038 on 551 degrees of freedom
Multiple R-squared: 0.2884
```

This is the result of having regressed the binary y_i's on the point spreads, simply treating the 1's and 0's as if they were real numbers. Under this model, our estimated regression equation is

$$\mathrm{E}(y_i \mid x_i) = P(y_i = 1 \mid x_i) = 0.524 + 0.024 \cdot x_i \,.$$

Plug in an x, and read off the probability of a home-team victory. Here, we would expect the intercept to be 0.5, meaning that the home team should win exactly 50% of the time when the point spread is 0. Of course, because of sampling variability, the estimated intercept $\widehat{\beta}_0$ isn't exactly 0.5. But it's certainly close—about 1 standard error away.

The linear probability model, however, has a serious flaw. Try plugging in $x_i = 21$ and see what happens:

$$P(y_i = 1 \mid x_i = 21) = 0.524 + 0.024 \cdot 21 = 1.028 \,.$$

We get a probability larger than 1, which is clearly nonsensical. We could also get a probability less than zero by plugging in $x_1 = -23$:

$$P(y_i = 1 \mid x_i = -23) = 0.524 - 0.024 \cdot 23 = -.028 \,.$$

The problem is that the straight-line fit does not respect the rule that probabilities must be numbers between 0 and 1. For many values of x_i, it gives results that aren't even mathematically legal.

Game	Win	Spread
1	0	-7
2	1	7
3	1	17
4	0	9
5	1	-2.5
6	0	-9
7	1	10
8	1	18
9	1	-7.5
10	0	-8
⋮		
552	1	-4.5
553	1	-3

Table 15.1: An excerpt from a data set on 553 NCAA basketball games. "Win" is coded 1 if the home team won the game, and 0 otherwise. "Spread" is the Las Vegas point spread in favor of the home team (at tipoff). Negative point spreads indicate where the visiting team was favored.

Link functions and generalized linear models

THE PROBLEM can be summarized as follows. The right-hand side of the regression equation, $\beta_0 + \beta_1 x_i$, can be any real number between $-\infty$ and ∞. But the left-hand side, $P(y_i = 1 \mid x_i)$, must be between 0 and 1. Therefore, we need some transformation g that takes an unconstrained number from the right-hand side, and maps it to a constrained number on the left-hand side:

$$P(y_i \mid x_i) = g(\beta_0 + \beta_1 x_i) \,.$$

Such a function g is called a *link function*; a model that incorporates such a link function is called a *generalized linear model*, or GLM. The part inside the parentheses $(\beta_0 + \beta_1 x_i)$ is called the *linear predictor*.

We use link functions and generalized linear models in most situations where we are trying to predict a number that is, for whatever reason, constrained. Here, we're dealing with probabilities, which are constrained to be no smaller than 0 and no larger than 1. Therefore, the function g must map real numbers on $(-\infty, \infty)$ to numbers on $(0, 1)$. It must therefore be shaped a bit like a flattened letter "S," approaching zero for large negative values of the linear predictor, and approaching 1 for large positive values.

Figure 15.2 contains the most common example of such a link function. This is called the *logistic* link, which gives rise to the *logistic regression model*:

$$P(y_i = 1 \mid x_i) = g(\beta_0 + \beta_1 x_i) = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \,.$$

Think of this as just one more transformation, like the logarithm or powers of some predictor x. The only difference is that, in this case, the transformation gets applied to the whole linear predictor at once. The logistic regression model is often called the logit model for short.[1]

[1] The "g" in "logit" is pronounced softly, like in "gentle" or "magic."

With a little bit of algebra, it is also possible to isolate the linear predictor $\beta_0 + \beta_1 x_i$ on one side of the equation. If we let p_i denote

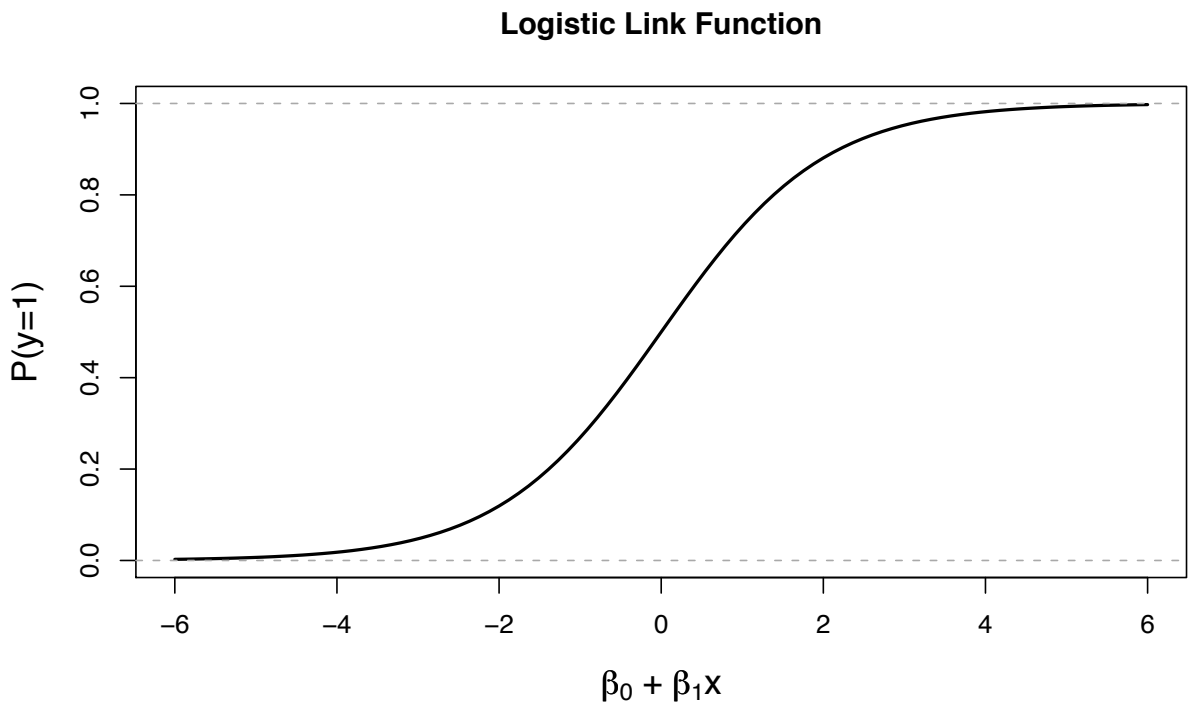


Figure 15.2: The logistic link function.

the probability that $y_i = 1$, given x_i, then

$$
\begin{aligned}
p_i &= \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} \\
p_i + p_i e^{\beta_0+\beta_1 x_i} &= e^{\beta_0+\beta_1 x_i} \\
p_i &= (1-p_i)e^{\beta_0+\beta_1 x_i} \\
\log\left(\frac{p_i}{1-p_i}\right) &= \beta_0 + \beta_1 x_i
\end{aligned}
$$

Since $p_i = P(y_i = 1 \mid x_i$, we know that $1 - p_i = P(y_i = 0 \mid x_i)$. Therefore, the ratio $p_i/(1-p_i)$ is the odds in favor of the event $y_i = 1$, given the predictor x_i. Thus the linear predictor $\beta_0 + \beta_1 x_i$ (on the right-hand side of the last equation) gives us the logarithm of the odds in favor of success ($y_i = 1$), on the left-hand side of the last equation.

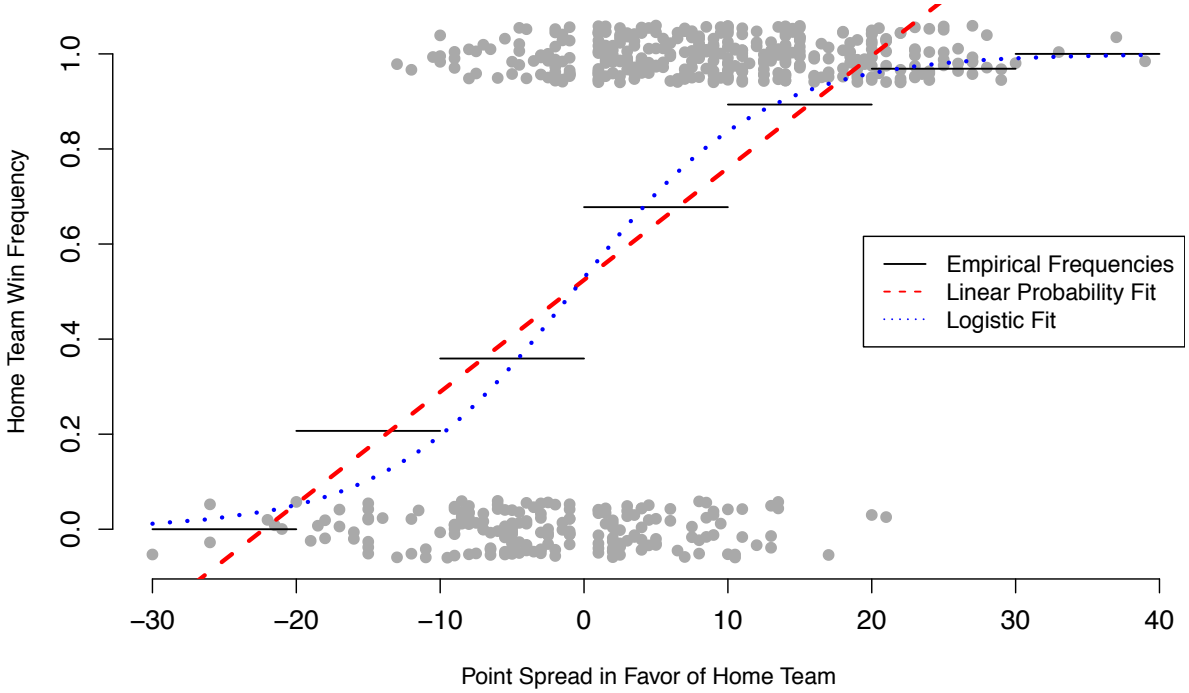


Figure 15.3: Win frequency versus point spread for 553 NCAA basketball games. Actual wins are plotted as 1's and actual losses as zeros. Some artificial vertical jitter has been added to the 1's and 0's to allow the dots to be distinguished from one another.

The logistic regression fit for the point-spread data

Let's return briefly to the data on point spreads in NCAA basketball games. The figure above compares the logistic model to the linear-probability model. The logistic regression fit ($\widehat{\beta}_0 = 0.117$, $\widehat{\beta}_1 = 0.152$) eliminates the undesirable behavior of the linear model, and ensures that all forecasted probabilities are between 0 and 1. Note the clearly non-linear behavior of the dotted blue curve. Instead of fitting a straight line to the empirical success frequencies, we have fit an S-shape.

Interpreting the coefficients

Interpreting the coefficients in a logistic regression requires a bit of algebra. For the sake of simplicity, imagine a data set with only a single regressor x_i that can take the values 0 or 1 (a dummy variable). Perhaps, for example, x_i denotes whether someone received

the new treatment (as opposed to the control) in a clinical trial.

For this hypothetical case, let's consider the ratio of two quantities: the odds of success for person i with $x_i = 1$, versus the odds of success for person j with $x_j = 0$. Denote this ratio by R_{ij}. We can write this as

$$\begin{aligned} R_{ij} &= \frac{O_i}{O_j} \\ &= \frac{\exp\{\log(O_i)\}}{\exp\{\log(O_j)\}} \\ &= \frac{\exp\{\beta_0 + \beta_1 \cdot 1\}}{\exp\{\beta_0 + \beta_1 \cdot 0\}} \\ &= \exp\{\beta_0 + \beta_1 - \beta_0 - 0\} \\ &= \exp(\beta_1)\,. \end{aligned}$$

Therefore, we can interpret the quantity e^{β_1} as an *odds ratio*. Since $R_{ij} = O_i/O_j$, we can also write this as:

$$O_i = e^{\beta_1} \cdot O_j\,.$$

In words: if we start with $x = 0$ and move to $x = 1$, our odds of success ($y = 1$) will change by a multiplicative factor of e^{β_1}.

For this reason, we usually refer to the exponentiated coefficient e^{β_j} as the odds ratio associated with predictor j.

Advanced topic: estimating the parameters of the logistic regression model

In previous chapters we learned how to estimate the parameters of a linear regression model using the least-squares criterion. This involved choosing values of the regression parameters to minimize the quantity

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\,,$$

where $\hat{y}_i$ is the value for y_i predicted by the regression equation.

In logistic regression, the analogue of least-squares is Gauss's principle of maximum likelihood, which we introduced when discussing the normal linear regression model. The idea here is to choose values for β_0 and β_1 that make the observed patterns of 1's and 0's look as likely as possible.

To understand how this works, observe the following two facts:

- If $y_i = 1$, then we have observed an event that occurred with probability $P(y_i = 1 \mid x_i)$. Under the logistic-regression

model, we can write this probability as

$$P(y_i = 1 \mid x_i) = \frac{e^{\beta_0+\beta_1 x_i}}{1 + e^{\beta_0+\beta_1 x_i}}$$

- If $y_i = 0$, then we have observed an event that occurred with probability $P(y_i = 0 \mid x_i) = 1 - P(y_i = 1 \mid x_i)$. Under the logistic regression model, we can write this probability as

$$1 - P(y_i = 1 \mid x_i) = 1 - \frac{e^{\beta_0+\beta_1 x_i}}{1 + e^{\beta_0+\beta_1 x_i}}$$

Since all of the individual 1's and 0's are independent, given the parameters β_0 and β_1, the joint probability of all the 1's and 0's is the product of their individual probabilities. We can write this as:

$$P(y_1, \ldots, y_n) = \prod_{i:y_i=1} \left(\frac{e^{\beta_0+\beta_1 x_i}}{1 + e^{\beta_0+\beta_1 x_i}} \right) \cdot \prod_{i:y_i=0} \left(1 - \frac{e^{\beta_0+\beta_1 x_i}}{1 + e^{\beta_0+\beta_1 x_i}} \right).$$

This expression is our *likelihood*: the joint probability of all our data points, given some particular choice of the model parameters.[2] The logic of maximum likelihood is to choose values for β_0 and β_1 such that $P(y_1, \ldots, y_n)$ is as large as possible. We denote these choices by $\widehat{\beta}_0$ and $\widehat{\beta}_1$. These are called the *maximum-likelihood estimates* (MLE's) for the logistic regression model.

[2] Remember that the big $\prod$ signs mean "product," just like $\sum$ means "sum." The first product is for the observations where y_i was a 1, and the second product is for the observations where y_i was a 0.

This likelihood is a difficult expression to maximize by hand (i.e. using calculus and algebra). Luckily, most major statistical software packages have built-in routines for fitting logistic-regression models, absolving you of the need to do any difficult analytical work.

The same is true when we move to multiple regression, when we have p predictors rather than just one. In this case, the logistic-regression model says that

$$P(y_i = 1 \mid x_{i1}, \ldots, x_{i,p} = g(\beta_0 + \beta_1 x_i) = \frac{e^{\psi_{ij}}}{1 + e^{e^{\psi_{ij}}}}, \quad , \psi_{ij} = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

where ψ_{ij} is the linear predictor for observation i.

Extensions to the basic logit model

The ordinal logit model

We can modify the logistic regression model to handle ordinal responses. The hallmark of ordinal variables is that they are measured on a scale that can't easily be associated with a numerical

magnitude, but that does imply an ordering: employee evaluations, survey responses, bond ratings, and so forth.

There are several varieties of ordinal logit model. Here we consider the *proportional-odds* model, which is most easily understood as a family of related logistic regression models. Label the categories as $1, \ldots, K$, ordered in the obvious way. Consider the probability $c_{ik} = P(y_i \leq k)$: the probability that the outcome for the ith case falls in category k *or any lower category.* (We call it c_{ik} because it is a cumulative probability of events at least as "low" as k.) The proportional-odds logit model assumes that the logit transform of c_{ik} is a linear function of predictors:

$$\text{logit}(c_{ik}) = \log\left(\frac{c_{ik}}{1 - c_{ik}}\right) = \eta_k + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}\,.$$

Crucially, this relationship is assumed to hold for all categories at once. Because $c_{iK} = 1$ for the highest category K, we have specified $K - 1$ separate binary logit models that all share the same predictors x_j and the same coefficients β_j. The only thing that differs among the models are the intercepts η_k; these are commonly referred to as the *cutpoints*. Since the log odds differ only by an additive constant for different categories, the odds differ by a multiplicative factor—thus the term "proportional odds."

To interpret the ordinal-logit model, I find it easiest to re-express individual fitted values in terms of covariate-specific category probabilities $w_{ik} = P(y_i = k)$:

$$w_{ik} = P(y_i \leq k) - P(y_i \leq k - 1) = c_{ik} - c_{i,k-1}\,,$$

with the convention that $c_{i0} = 0$. Good software makes it fairly painless to do this.

The multinomial logit model

Another generalization of the binary logit model is the multinomial logit model. This is intended for describing *unordered* categorical responses: PC/Mac/Linux, Ford/Toyota/Chevy, plane/train/automobile, and so forth. Without a natural ordering to the categories, the quantity $P(y_i \leq k)$ ceases to be meaningful, and we must take a different approach.

Suppose there are K possible outcomes ("choices"), again labeled as $1, \ldots, K$ (but without the implied ordering). As before, let $w_{ik} = P(y_i = k)$. For every observation, and for each of the K

choices, we imagine that there is a linear predictor ψ_{ik} that measures the preference of subject i for choice k. Intuitively, the higher ψ_{ik}, the more likely that $y_i = k$.

The specific mathematical relationship between the linear predictors and the probabilities w_{ik} is given the multinomial logit transform:[3]

$$\begin{aligned} w_{ik} &= \frac{\exp(\psi_{ik})}{\sum_{l=1}^{K} \exp(\psi_{il})} \\ \psi_{ik} &= \beta_0^{(k)} + \beta_1^{(k)} x_{i1} + \cdots \beta_p^{(k)} x_{ip} \,. \end{aligned}$$

[3] Some people, usually computer scientists, will refer to this as the *softmax* function.

Each category gets its own set of coefficients, but the same set of predictors x_1 through x_p.

There is one minor issue here. With a bit of algebra, you could convince yourself that adding a constant factor to each ψ_{ik} would not change the resulting probabilities w_{ik}, as this factor would cancel from both the numerator and denominator of the above expression. To fix this indeterminacy, we choose one of the categories (usually the first or last) to be the reference category, and set its coefficients equal to zero.

Models for count outcomes

The Poisson model. For modeling event-count data (photons, mortgage defaults in a ZIP code, heart attacks in a town), a useful place to start is the Poisson distribution. The key feature of counts is that they must be non-negative integers. Like the case of logistic regression, where probabilities had to live between 0 and 1, this restriction creates some challenges that take us beyond ordinary least squares.

The Poisson distribution is parametrized by a rate parameter, often written as λ. Let k denote an integer, and y_i denote the event count for subject i. In a Poisson model, we assume that

$$P(y_i = k) = \frac{\lambda_i^k}{k!} e^{-\lambda_i} \,,$$

and we wish to model λ_i in terms of covariates. Because the rate parameter of the Poisson cannot be negative, we must employ the same device of a link function to relate λ_i to covariates. By far the most common is the (natural) log link:

$$\log \lambda_i = \beta_0 + \beta_1 x_{i1} + \cdots \beta_p x_{ip} \,,$$

or equivalently,

$$\lambda_i = \exp\{\beta_0 + \beta_1 x_{i1} + \cdots \beta_p x_{ip}\}\,.$$

As with the case of logistic regression, the model is fit via maximum-likelihood.

Interpreting the coefficients. Because we are fitting a model on the log-rate scale, additive changes to an x variable are associated with multiplicative changes in the y variable. As before, let's consider the ratio of two quantities: the rate of events for person i with $x_1 = x^\star + 1$, versus the rate of events for person j with $x_1 = x^\star$. Let's further imagine that all other covariates are held constant at values x_2 to x_p, respectively. This implies that the only difference between subjects i and j is a one-unit difference in the first predictor, x_1.

We can write their ratio of rates as

$$\begin{aligned} R_{ij} &= \frac{\lambda_i}{\lambda_j} \\ &= \frac{\exp\{\beta_0 + \beta_1 \cdot (x^\star + 1) + \beta_2 x_2 + \cdots \beta_p x_p\}}{\exp\{\beta_0 + \beta_1 \cdot x^\star + \beta_2 x_2 + \cdots \beta_p x_p\}} \\ &= \exp\{\beta_1 (x^\star + 1 - x^\star)\} \\ &= \exp(\beta_1)\,. \end{aligned}$$

Thus person i experiences events events e^{β_1} times as frequently as person j.

Overdispersion. For most data sets outside of particle physics, the Poisson assumption is usually one of convenience. Like the normal distribution, it is familiar and easy to work with. It also has teeth, and may bite if used improperly. One crucial feature of the Poisson is that its mean and variance are equal: that is, if $y_i \sim \text{Pois}(\lambda_i)$, then the expected value of y_i is λ_i, and the standard deviation of y_i is $\sqrt{\lambda_i}$. (Since λ_i depends on covariates, we should really be calling these the *conditional* expected value and standard deviation.)

As a practical matter, this means that if your data satisfy the Poisson assumption, then roughly 95% of observations should fall within $\pm 2\sqrt{\lambda_i}$ of their conditional mean λ_i. This is quite narrow, and many (if not most) data sets exhibit significantly more variability about their mean. If the conditional variance exceeds the

conditional mean, the data exhibits *overdispersion with respect to the Poisson*, or just *overdispersion* for short.

Overdispersion can really mess with your standard errors. In other words, if you use (i.e. let your software use) the Poisson assumption to calculate error bars, but your data are overdispersed, then you will end up overstating your confidence in the model coefficients. Sometimes the effect is dramatic, meaning that the blind use of the Poisson assumption is a recipe for trouble.

There are three common strategies for handling overdispersion:

(1) Use a quasi-likelihood approach ("family=quasipoisson" in R's glm function);

(2) Fit a different count-data model, such as the negative binomial or Poisson-lognormal, that can accommodate overdispersion;

(3) Fit a hierarchical model.

Alas, these topics are for a more advanced treatment of generalized linear models.