9
The bivariate normal distribution

Joint distribution for discrete variables

In this chapter, we study probability distributions for coupled sets of random variables in more detail. Recall that a *joint distribution* is a list of joint outcomes for two or more variables at once, together with the probabilities for each of these outcomes. We'll first work through an example of a joint distribution for two discrete random variables, before turning to more complex examples.

Suppose we wanted to build a probability model for 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 depicts an example of what the joint distribution $P(X_{ba}, X_{be})$ might look like; these numbers were derived using housing-market data from the fall of 2015.

	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	

You can check that the probabilities sum to 1, as they must. Notice that an extra row and column have been added in the bottom and right margins:

- The extra row at the bottom, obtained by summing the probabilities along each column, represents the marginal distribu-

tion over the number of bathrooms.

- The extra column at the right, obtained by summing the probabilities along each row, represents the marginal distribution over the number of bedrooms.

Using these marginal distributions alone, we can straightforwardly 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.

To quantify the strength of association between two variables, we will calculate their *covariance*. Recall the general definition of covariance. Suppose that there are N possible joint outcomes for X and Y. Then

$$\text{cov}(X, Y) = E\Big\{ [X - E(X)][Y - E(Y)] \Big\} = \sum_{i=1}^{n} p_i \big[x_i - E(X) \big] \big[y_i - E(Y) \big] .$$

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:

$$\begin{aligned} \text{cov}(X_{ba}, X_{be}) &= 0.003 \cdot (1 - 2.215)(1 - 3.149) \\ &\quad + 0.068 \cdot (1 - 2.215)(2 - 3.149) \\ &\quad + \cdots \\ &\quad + 0.185 \cdot (3 - 2.215)(4 - 3.149) \\ &\quad + \cdots \\ &\quad + 0.005 \cdot (4 - 2.215)(6 - 3.149) \\ &\quad \approx 0.285 . \end{aligned}$$

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.

We can also calculate the correlation between the two variables:

$$\text{cor}(X_{ba}, X_{be}) = \frac{0.285}{\sqrt{0.595} \cdot \sqrt{0.643}} \approx 0.745\,.$$

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 9.1.[2] As part of Galton's study of heredity, he collected data on the adult height of parent–child pairs.[3] 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 9.1. 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 9.1. These 20 men had a mean height

[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.

[3] The points in Figure 9.1 contain fathers and sons only, to avoid any confounding due to sex. In the figure, you'll see that we've 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.

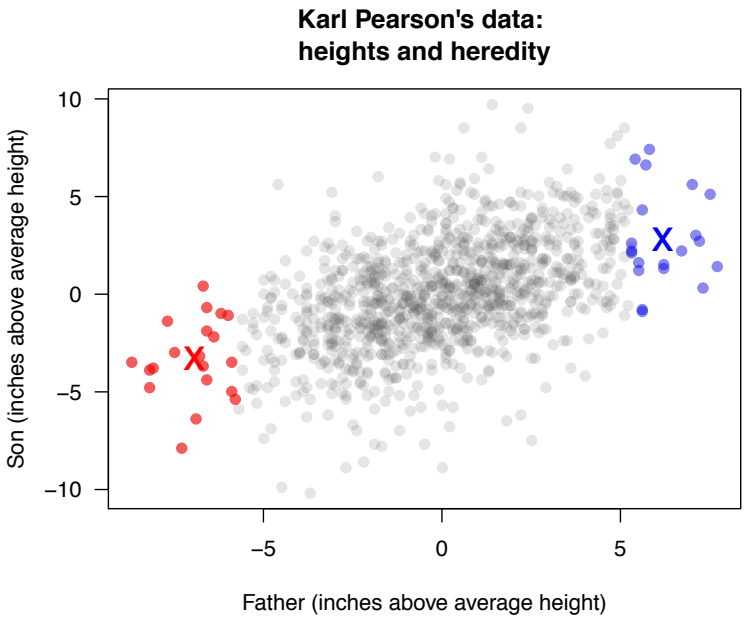

Figure 9.1: Karl Pearson's data on the height of fathers and their adult sons. The 20 tallest fathers (and their sons) are highlighted in blue, with the bivariate mean of this group shown as a blue X. Similarly, the 20 shortest fathers (and their sons) are highlighted in red, with the bivariate mean of this group shown as a red X.

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 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.)

Figure 9.2: Yao Ming, making J.J. Watt (6'5" tall, 290 pounds) look like a child.

- Therefore, our selected group of very tall people (the blue dots in Figure 9.1) 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 (because 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

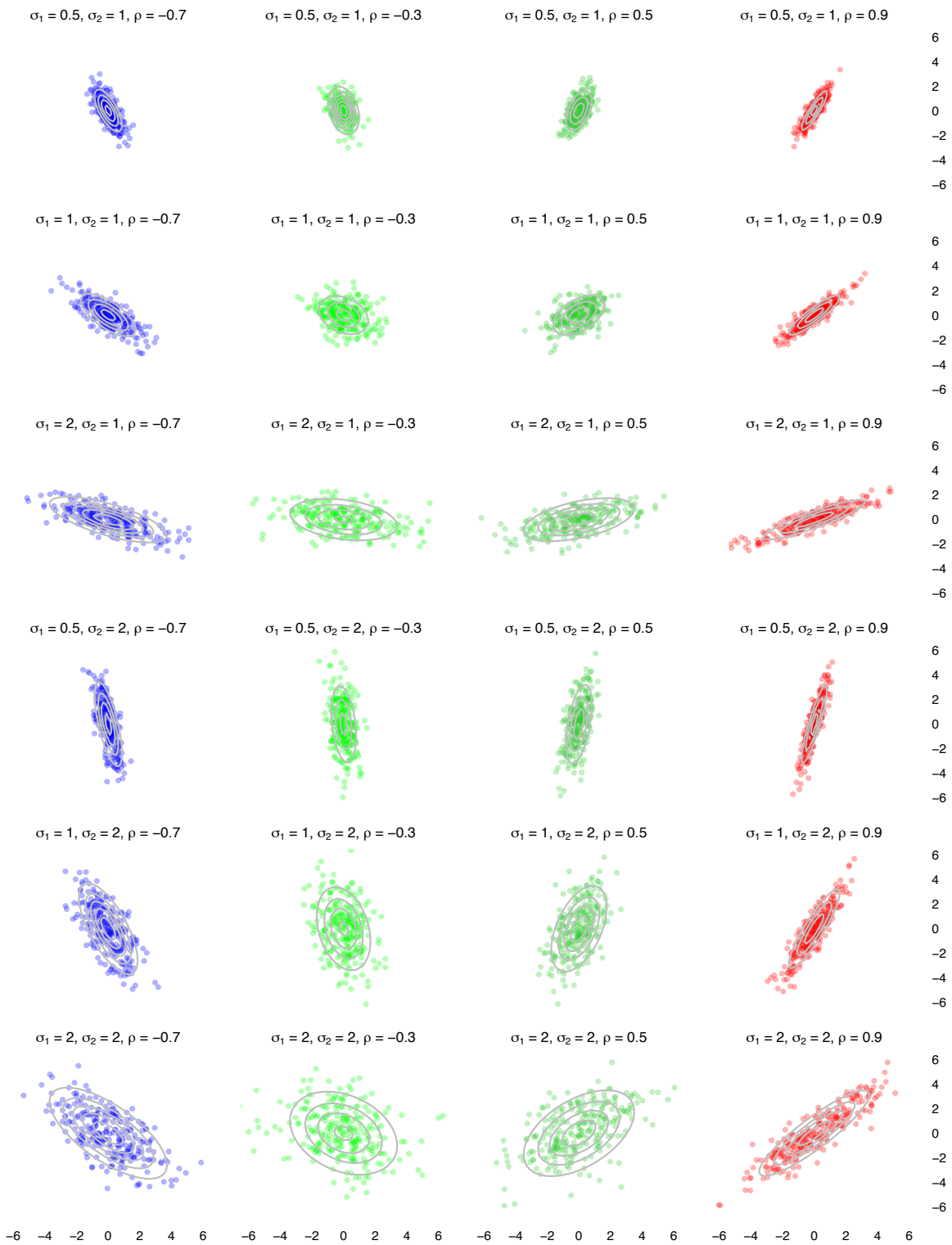

Figure 9.3: 24 examples of a bivariate normal distribution (250 samples in each plot).

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 $(X_1, X_2) \sim N(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$.

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 $X \sim \text{N}(\mu, \Sigma)$, 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_1, we find that $X_2 \sim N(\mu_2, \sigma_2^2)$.

Visualizing the bivariate normal distribution

Figure 9.3 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 9.3 also shows a *contour plot* of the probability density function for the corresponding bivariate normal

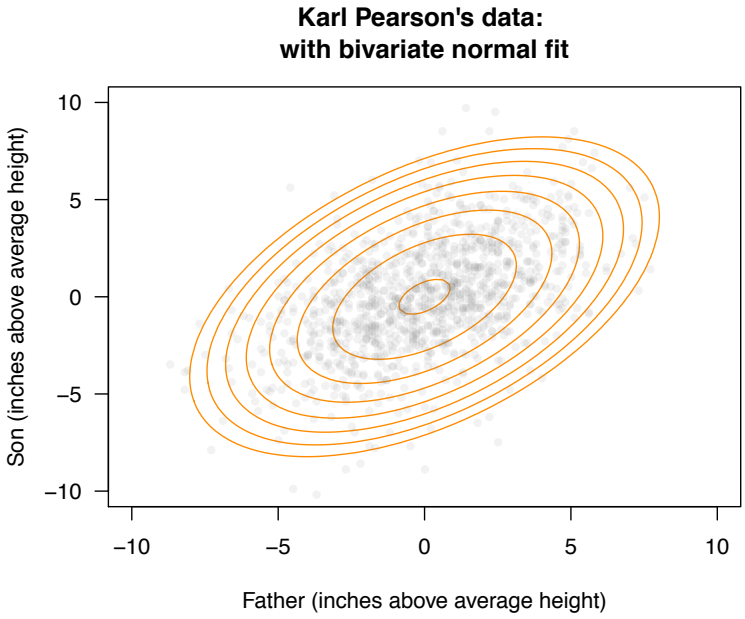

Figure 9.5: Best-fitting bivariate normal distribution for Karl Pearson's height data based on the sample standard deviations and sample correlation.

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? 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 9.4). 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.

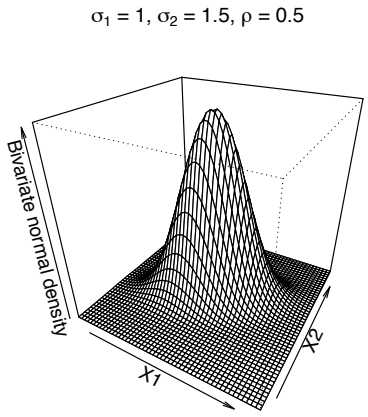

Figure 9.4: A three-dimensional wire-frame plot of a bivariate normal density function.

Figure 9.5 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 distribution 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

$$E(X_2 \mid X_1 = x_1) = \mu_2 + \rho \cdot \frac{\sigma_2}{\sigma_1} \cdot (x_1 - \mu_1) \tag{9.1}$$

$$\text{var}(X_2 \mid X_1 = x_1) = \sigma_2^2 \cdot (1 - \rho^2)\,, \tag{9.2}$$

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.[4]

This fact brings us straight back to the concept of regression to the mean. Let's re-arrange Equation 9.1 to re-express the conditional mean in a slightly different way:

$$\frac{E(X_2 \mid X_1 = x_1) - \mu_2}{\sigma_2} = \rho \cdot \left(\frac{x_1 - \mu_1}{\sigma_1} \right)\,. \tag{9.3}$$

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 9.3 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.

[4] 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.

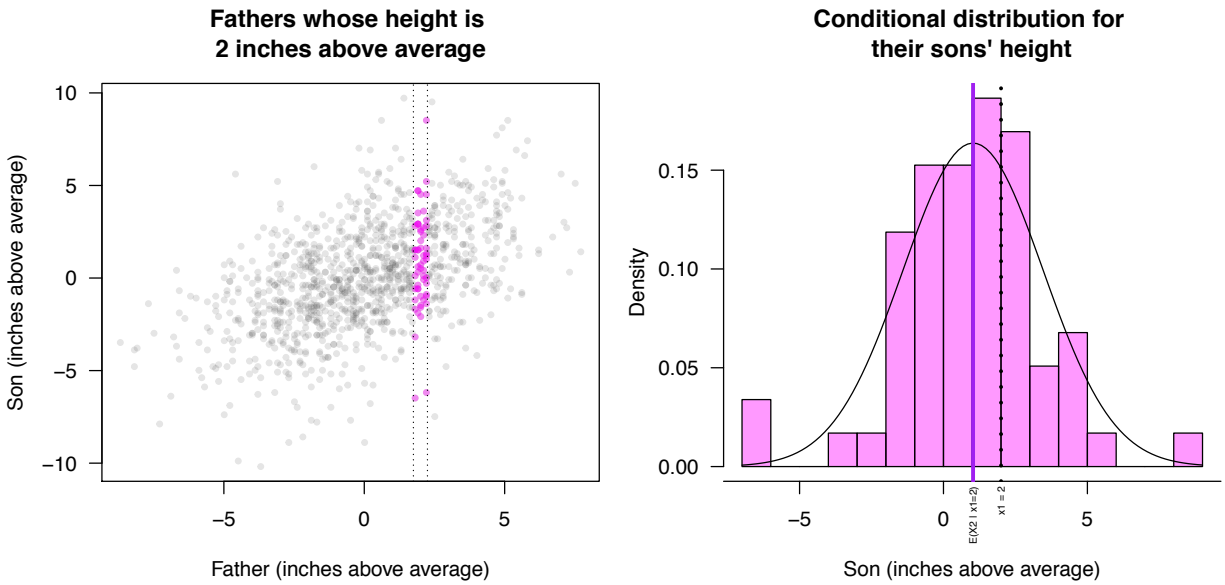

Figure 9.6: Left: father–son pairs where the father's height is about 2 inches above average are highlighted in purple. Right: the histogram of the sons' height, together with the conditional distribution $P(X_2 \mid X_1 = 2)$ predicted by the bivariate normal fit to the joint distribution for (X_1, 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.

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 9.1 together with the parameters of the best-fitting bivariate normal distribution from Figure 9.5, 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 9.6 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 9.1 and 9.2. 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 examples 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.[5]

[5] 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."[6]

[6] "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

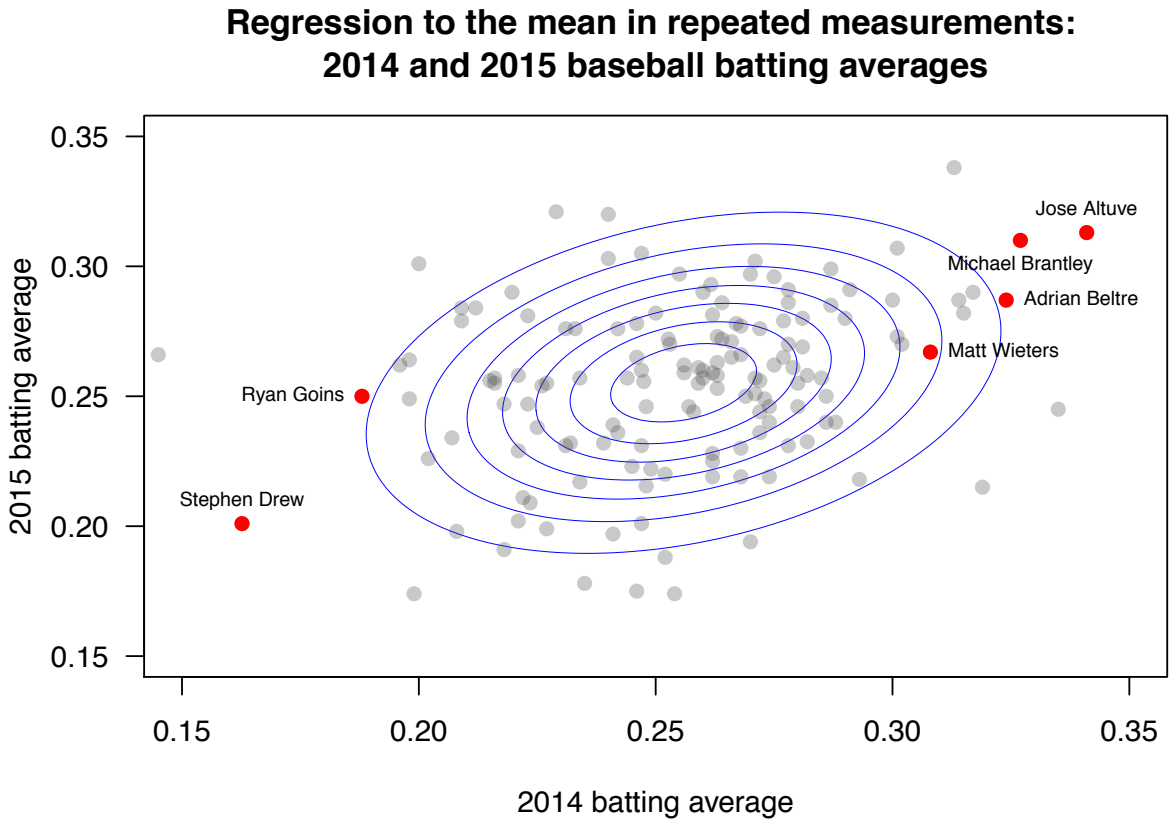

Figure 9.7: Baseball batting averages in the 2014 and 2015 seasons for all players with at least 100 at-bats in both years.

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.[7]

Figure 9.7 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.

[7] 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.

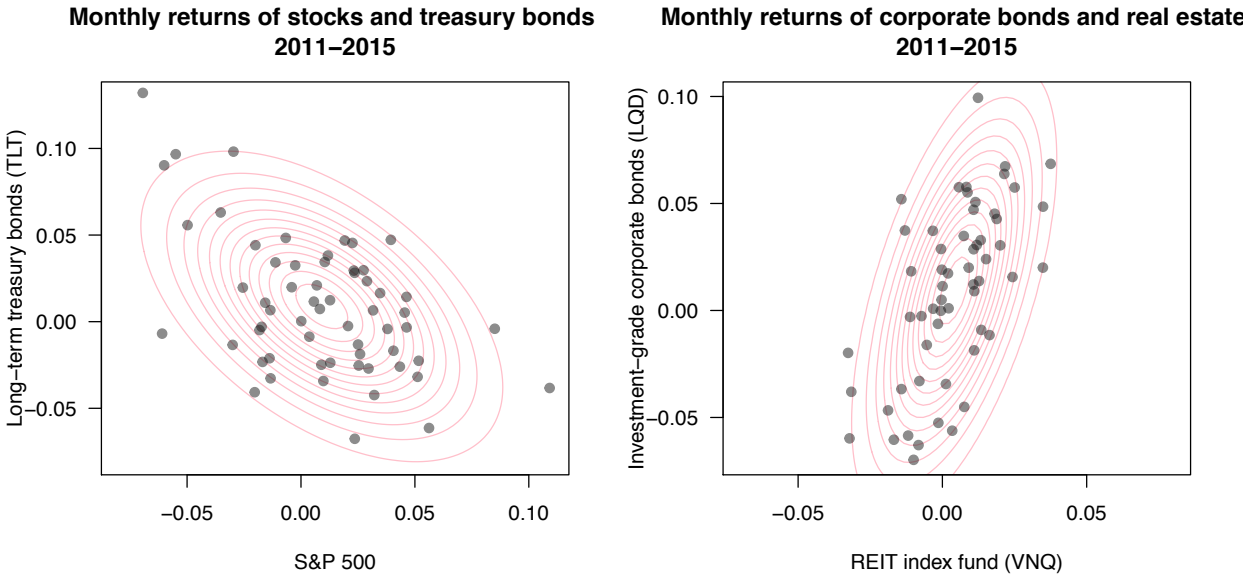

Figure 9.8: Correlation between stocks and government bonds (left); correlation between corporate bonds and real estate (right).

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 correlation between two correlated random variables. As a final example, let's look at correlation between pairs of financial assets. We'll consider two pairs of 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.[8] 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.[9] The left panel of Figure 9.8 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 9.8, which shows returns for real-estate investment trusts (X_1) and corporate bonds

[8] Recall that a Treasury bond entailed lending money to the U.S. federal government and collecting interest in return.

[9] 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.

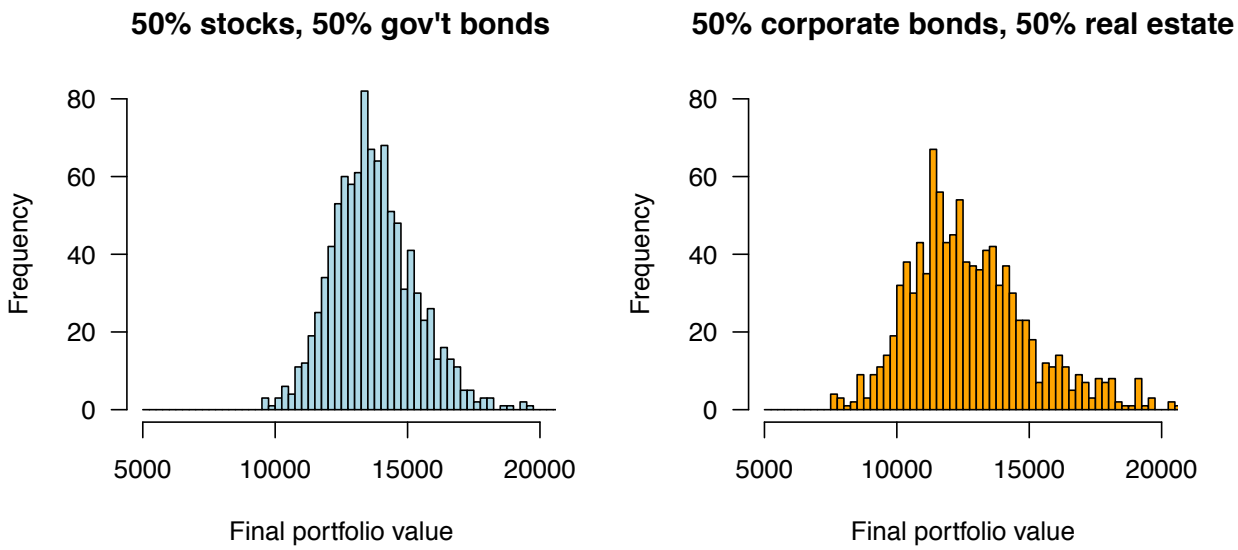

Figure 9.9: Final value of 36-month investments in 50/50 mixes of: (1) stocks and government bonds (left), and (2) corporate bonds and real estate (right).

(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 9.9 shows the results of this simulation, assuming that returns following the bivariate normal distributions fit to the data in Figure 9.8. 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.