5. *Normal Distributions: The Bell Curve*

I couldn't claim that I was smarter than sixty-five other guys—but the average of sixty-five other guys, certainly.

—Richard Feynman

Distributions constitute part of the core knowledge base for any modeler. Later, we use distributions to construct and analyze models of path dependence, random walks, Markov processes, search, and learning. We also require a working knowledge of distributions to measure inequality in power, income, and wealth and to perform statistical tests. Our treatment of distributions unfolds over two short chapters—one each for normal and power law (long-tailed) distributions—in which we take the perspective of modelers rather than statisticians. As modelers, we are interested in two big questions: Why do we see the distributions we do, and why do distributions matter?

To address the first big question, we need to reacquaint ourselves with what distributions are. A *distribution* mathematically captures variation (differences within a type) and diversity (difference across types) by representing them as probability distributions defined over numerical values or categories. A *normal distribution* takes the familiar bell curve shape. Heights and weights of most species satisfy normal distributions. They are symmetric around their means and do not include particularly large or small events. We do not encounter many six-foot-long ants or four-pound elk. We can rely on the central limit theorem to explain the prevalence of normal distributions. It tells us that when we add up or average random variables, we can expect to obtain a normal distribution. Many empirical phenomena, in particular any aggregate like sales data or vote totals, can be written as sums of random events.

Not all event sizes are normal. Earthquakes, war deaths, and book sales exhibit *long-tailed distributions*: they consist mostly of tiny events but include the occasional whopper. Californians experience over 10,000 earthquakes each year. Unless you are staring at the quivering petals of a

jasmine blossom, you would not notice them. Occasionally, though, the earth opens up, highways collapse, and cities tremble.

Knowing whether a system produces a normal or long-tailed distribution matters for any number of reasons. We want to know whether a power grid will suffer massive outages, or whether a market system will produce a handful of billionaires and billions of poor people. With knowledge of distributions, we can predict the likelihood of floodwaters that exceed a levee's walls, the probability that Delta flight 238 arrives in Salt Lake City on time, and the odds that a transportation hub costs double its budgeted amount. Knowledge of distributions is also relevant in design. Normal distributions imply no large deviations, so airplane designers need not create leg space for the eighteen-foot human. An understanding of distributions can also guide actions. As we learn later, preventing riots depends less on reducing average levels of discontent than on appeasing people at the extreme.

In this chapter, we adopt a *structure-logic-function* organization. We define normal distributions, describe how they arise, and then ask why they matter. We apply our knowledge of distributions to explain why good things come in small samples, to test for significance of effects, and to explain Six Sigma process management. We then go back to the logic question and ask what happens if we multiply rather than add random variables. We learn that we obtain a *lognormal distribution*. Lognormal distributions include larger events and are not symmetric about their means. It follows that multiples of effects lead to more inequality, an insight that has implications for how policies for increasing salaries affect income distributions.

The Normal Distribution: Structure

A distribution assigns probabilities to events or values. The distribution of daily rainfall, test scores, or human height assigns a probability to every possible value of the outcomes. Statistical measures condense the information contained in a distribution into single numbers, such as the *mean*, the average value of the distribution. The mean height of a tree in Germany's Black Forest might be eighty feet, and the mean time spent in

the hospital following open-heart surgery might be five days. Social scientists rely on means to compare economic and social conditions across countries. In 2017, the United States per capita GDP of $57,000 exceeded that of France, which equaled $42,000, while mean life expectancy in France exceeds that of the United States by three years.

A second statistic, *variance,* measures a distribution's dispersion: the average of the squared distance of the data to the mean.[1] If every point in a distribution has the same value, the variance equals zero. If half of the data have value 4 and half have value 10, then, on average, each point lies distance 3 from the mean, and the variance equals 9. The *standard deviation* of a distribution, another common statistic, equals the square root of the variance.

The set of possible distributions is limitless. We could draw any line on a piece of graph paper and interpret it as a probability distribution. Fortunately, the distributions we encounter tend to belong to a few classes. The most common distribution, the normal distribution, or bell curve, is shown in figure 5.1.

Figure 5.1: Normal Distribution with Standard Deviations

Normal distributions are *symmetric* about their mean. If the mean equals zero, the probability of a draw larger than 3 equals the probability of a draw less than -3. A normal distribution is characterized by its mean and standard deviation (or, equivalently, its variance). In other words, graphs of normal distribution all look identical, with approximately 68% of all outcomes within one standard deviation of the mean, 95% of all outcomes within two standard deviations, and more than 99% lying within three standard deviations. Normal distributions allow for any size outcome or event, though large events are rare. An event five standard deviations from the mean occurs about once in every 2 million draws.

We can exploit the regularity of normal distributions to assign probabilities to ranges of outcomes. If houses in Milwaukee, Wisconsin, have a mean square footage equal to 2,000 with a standard deviation of 500 square feet,

then 68% of houses have between 1,500 and 2,500 square feet and 95% have between 1,000 and 3,000 square feet. If the 2019 fleet of Ford Focuses can travel, on average, 40 miles per gallon with a standard deviation of 1 mile per gallon, then more than 99% of Focuses will get between 37 and 43 miles per gallon. As much as a consumer might hope, her new Focus will not run 80 miles on a gallon of gasoline.

The Central Limit Theorem: Logic

No end of phenomena exhibit normal distributions: physical sizes of flora and fauna, student test scores on exams, daily sales at convenience stores, and the life spans of sea urchins. The *central limit theorem*, which states that adding or averaging random variables produces a normal distribution, explains why (see box).

Central Limit Theorem

The sum of $N \geq 20$ random variables will be approximately a normal distribution provided that the random variables are independent, that each has finite variance, and that no small set of the variables contributes most of the variation.[2]

One remarkable aspect of this theorem is that the random variables themselves need not be normally distributed. They could have any distribution so long as each has finite variance and no small subset of them contributes most of the variance. Suppose that data on the purchasing behaviors of the people in a small town of population 500 shows that each person spends on average $100 a week. Some of those people might spend $50 one week and $150 the next. Others might spend $300 every third week, while others might spend random amounts between $20 and $180 each week. So long as each person's spending has finite variation and no small subset of people contribute most of the variation, the sum of the distributions will be normally distributed with a mean of $50,000. Aggregate weekly spending will also be symmetric: as likely to be above $55,000 as it is below $45,000. By the same logic, the number of bananas,

quarts of milk, or boxes of taco shells that people buy will also be normally distributed.

We can also apply the central limit theorem to explain the distribution of human heights. A person's height is determined by a combination of genetics, the environment, and interactions between the two. The genetic contribution could be as high as 80%, so we will assume that height depends only on genes. At least 180 genes contribute to human height.[3] One gene may contribute to having a longer neck or head and another to a longer tibia. Though genes interact, to a first approximation, we can assume that each contributes independently. If height equals the sum of the contributions of the 180 genes, then heights will be normally distributed. By the same logic, so too will the weights of wolves and the length of pandas' thumbs.

Applying Our Knowledge of Distributions: Function

Our first application of the normal distribution reveals why exceptional outcomes occur far more often in small populations, why the best schools are small, and why the counties with the highest cancer rates have small populations. Recall that in a normal distribution 95% of outcomes lie within two standard deviations and 99% lie within three standard deviations, and that by the central limit theorem, the mean of a collection of independent random variables will be normally distributed (with the caveats about variance). It follows that we can be pretty confident that population averages on test scores and the like will be normally distributed. The standard deviation of the average of the random variables, however, does not equal the average of the variables' standard deviations, nor does the standard deviation of the sum equal the sum of the standard deviations. Instead, those formulae depend on the square roots of the population sizes (see box).

The Square Root Rules

The standard deviations of the mean σ_μ and of the sum σ_Σ of N independent random variables each with standard deviation σ are given by the following formulae:[4]

The formula for the standard deviation of the mean implies that large populations have much lower standard deviations than small ones. From this, we can infer that we should see more good things and more bad things in small populations. And in fact we do. The safest places to live are small towns, as are the least safe. The counties with the highest rates of obesity and cancer have small populations. These facts can all be explained by differences in standard deviations.

Failure to take sample size into account and inferring causality from outliers can lead to incorrect policy actions. For this reason, Howard Wainer refers to the formula for the standard deviation of the mean the "most dangerous equation in the world." For example, in the 1990s the Gates Foundation and other nonprofits advocated breaking up schools into smaller schools based on evidence that the best schools were small.[5] To see the flawed reasoning, imagine that schools come in two sizes—small schools with 100 students and large schools with 1,600 students—and that student scores at both types of schools are drawn from the same distribution with a mean score of 100 and a standard deviation of 80. At small schools, the standard deviation of the mean equals 8 (the standard deviation of the student scores, 80, divided by 10, the square root of the number of students). At large schools, the standard deviation of the mean equals 2.

If we assign the label "high-performing" to schools with means above 110 and the label "exceptional" to schools with means above 120, then only small schools will meet either threshold. For the small schools, an average score of 110 is 1.25 standard deviations above the mean; such events occur about 10% of the time. A mean score of 120 is 2.5 standard deviations above the mean; an event of that size should occur about once in 150 schools. When we do these same calculations for large schools, we find that the "high-performing" threshold lies five standard deviations above the mean and the "exceptional" threshold lies ten standard deviations above the

mean. Such events would, in practice, never occur. Thus, the fact that the very best schools are small is not evidence that smaller schools perform better. The very best schools will be small even if size has no effect solely because of the square root rules.

Testing Significance

We also use the regularity of the normal distribution to test for significant differences in mean values. If an empirical mean lies more than two standard deviations from a hypothesized mean, social scientists reject the hypothesis that the means are the same.[6] Suppose we advance a hypothesis that commute times in Baltimore equal those in Los Angeles. Suppose that our data show that commute times in Baltimore averaged 33 minutes, compared to 34 minutes in Los Angeles. If both data sets have standard deviations of the mean equal to 1 minute, then we could not reject the hypothesis that the commute times are the same. The means differ, but only by a single standard deviation. If instead commute times in Los Angeles averaged 37 minutes, then we would reject the hypothesis because the means differ by four standard deviations.

Physicists, though, might not reject the hypothesis, at least not if the data came from a physics experiment. Physicists impose stricter standards because they have larger data sets—there are a lot more atoms than people, and cleaner data. The evidence physicists relied on for the existence of the Higgs boson in 2012 would occur randomly less than once in 7 million trials were the Higgs boson not to exist.

The drug approval process used by the United States Food and Drug Administration (FDA) also uses tests of significance. If a pharmaceutical company claims that a new drug reduces the severity of eczema, that company must run two randomized controlled trials. To construct a randomized controlled trial the company would create two identical populations of eczema sufferers. One of the populations receives the drug. The other population receives a placebo. At the end of the trial, the average severity as well as average rates of negative side effects are compared. The company then runs statistical tests. If the drug significantly reduces eczema (measured in standard deviations) and does not significantly increase side

effects, the drug can be approved. The FDA does not use a hard-and-fast two-standard-deviation rule. The statistical bar will be lower for a drug that cures a fatal disease and exhibits only minor side effects than for a drug that cures toenail fungus but has a higher-than-expected incidence of bone cancer associated with its usage. The FDA also cares about the *power* of the statistical test—the probability that the test shows that the drug works.

Six Sigma Method

As our final application, we show how normal distributions inform quality control through the *Six Sigma method*. Developed in the mid-1980s by Motorola, the Six Sigma method reduces errors. The method models product attributes as drawn from a normal distribution. Imagine a company that produces bolts for door handles that must fit snugly into knobs made by another manufacturer. Specifications call for the bolts to be 14 millimeters in diameter, though any bolt between 13 and 15 millimeters in diameter will function properly. If the diameters of the bolts are normally distributed with a mean of 14 millimeters and a standard deviation of 0.5 millimeter, then any bolt that differs by more than two standard deviation fails. Two-standard-deviation events occur 5% of the time—far too high a rate for manufacturers.

The Six Sigma method involves working to reduce the size of a standard deviation to lower the probability of a failure. Companies can reduce error rates by tightening quality control. On February 26, 2008, Starbucks closed down over seven thousand shops for over three hours to retrain employees. Similarly, checklists used by airlines and now hospitals reduce variation.[7] Six Sigma reduces the standard deviation so that even a six-standard-deviation error avoids a malfunction. In our bolt example, that would require reducing the standard deviation of a bolt’s diameter to one-sixth of a millimeter. Six standard deviations implies an error rate of 2 per billion cases. The actual threshold used assumes an unavoidable rate of one and a half standard deviations. Thus, a six-sigma event actually corresponds to a four-and-a-half sigma event, and an allowable error rate of about 1 per 3 million.

The application of the central limit theorem (and therefore an implicit model of additive error) in the Six Sigma method is so subtle as to almost go unnoticed. The bolt manufacturer likely does not perform a precise measurement of the diameter of every bolt. It may sample a few hundred. From that sample, it estimates a mean and a standard deviation. Then, by assuming that variations in diameter result from the sum of random effects such as machine vibrations, variation in the quality of metals, and fluctuations in the temperature and speed of a press, they can invoke the central limit theorem and infer a normal distribution of diameters. The manufacturer then has a benchmark standard deviation that it can seek to reduce.

Lognormal Distributions: Multiplying Shocks

The central limit theorem requires that we add or average independent random variables in order to get a normal distribution. If the random variables are not added but interact in some way, or if they fail to be independent, then the resulting distribution need not be normal. In fact, generally it will not be. For example, random variables that are the product of independent random variables produce *lognormal* rather than normal distributions.[8] Lognormal distributions lack symmetry because products of numbers larger than 1 grow faster than sums ($4 + 4 + 4 + 4 = 16$, but $4 \times 4 \times 4 \times 4 = 256$) and multiples of numbers less than 1 decrease faster than sums (image, but image). If we multiply sets of twenty random variables with values uniformly distributed between zero and 10, their product will consist of many outcomes near zero and some large outcomes, creating the skewed distribution shown in 5.2.

image

Figure 5.2: A Lognormal Distribution

The length of the tail in a lognormal distribution depends on the variance of the random variables multiplied together. If they have low variance, the tail will be short. If they have high variance, the tail can be quite long because, as noted, multiplying together a sequence of large numbers produces a very large number. Lognormal distributions arise in a wide range of examples,

including the sizes of British farms, the concentration of minerals in the earth, and the time from infection with a disease to the appearance of symptoms.[9] Income distributions within many countries approximate lognormal distributions, though many deviate from lognormal at the upper end by having too many people with high incomes.

A simple model that can explain why income distributions are closer to lognormal than normal links policies about salary increases to their implied distributions. Most organizations assign raises by percentages. People who perform above average receive high-percentage raises. People who perform below average receive low-percentage raises. Instead, organizations could assign raises by absolute amounts. The average employee could receive a $1,000 raise. Those who perform better could receive more, and those who perform worse could receive less. The distinction between percentages and absolute amounts may appear semantic, but it is not.[10] Allocating raises by percentages based on employee performance when performances from year to year are independent and random produces a lognormal distribution. Differences in income become exacerbated in future years even with identical subsequent performance. An employee who has performed well in the past and earns $80,000 will receive $4,000 from a 5% raise. Another employee, who earns only $60,000, receives only $3,000 from the same 5% raise. Inequality begets more inequality even with identical performance. Had the organization allocated raises by absolute amounts, the two employees would receive the same raise and the resulting distribution of incomes would be closer to a normal distribution.

Summary

In this chapter, we covered the structure, logic, and function of normal distributions. We saw that normal distributions can be characterized by a mean and a standard deviation. We described the central limit theorem, which shows how normal distributions arise whenever we add up or average independent random variables with finite variance. And we described formulae for the standard deviations of the mean and sum of random variables. We then showed the consequences of those properties. We learned that small populations will be far more likely to produce

exceptional events and how when we lack that insight we make improper inferences and take unwise actions. We learned how assumption of normally distributed random variables allows scientists to make claims about the significance and power of statistical tests, and how process management can predict the likelihood of failure using an assumption of normality.

Not every quantity can be written as the sum, or the average, of independent random variables. Thus, not every distribution will be normal. Some quantities are products of independent random variables and will be lognormally distributed. Log-normal distributions only take on positive values. They also have longer tails, which means more large events and many more very small events. Those tails become long when random variables multiplied together have high variance. Long-tailed distributions imply less predictability, whereas normal distributions imply regularity. As a rule, we prefer regularity to the potential for large events. Therefore, we benefit from knowing the logic that creates the various distributions. In general, we would prefer that we add random shocks rather than multiply them together so as to reduce the likelihood of large events.