Chapter 2

Mathematical Preliminaries

> A data scientist is someone who knows more statistics than a computer scientist and more computer science than a statistician.
>
> – Josh Blumenstock

You must walk before you can run. Similarly, there is a certain level of mathematical maturity which is necessary before you should be trusted to do anything meaningful with numerical data.

In writing this book, I have assumed that the reader has had some degree of exposure to probability and statistics, linear algebra, and continuous mathematics. I have also assumed that they have probably forgotten most of it, or perhaps didn't always see the forest (why things are important, and how to use them) for the trees (all the details of definitions, proofs, and operations).

This chapter will try to refresh your understanding of certain basic mathematical concepts. Follow along with me, and pull out your old textbooks if necessary for future reference. Deeper concepts will be introduced later in the book when we need them.

2.1 Probability

Probability theory provides a formal framework for reasoning about the likelihood of events. Because it is a formal discipline, there are a thicket of associated definitions to instantiate exactly what we are reasoning about:

- An *experiment* is a procedure which yields one of a set of possible outcomes. As our ongoing example, consider the experiment of tossing two six-sided dice, one red and one blue, with each face baring a distinct integer $\{1, \ldots, 6\}$.
- A *sample space* S is the set of possible outcomes of an experiment. In our

S.S. Skiena, *The Data Science Design Manual,*
Texts in Computer Science, https://doi.org/10.1007/978-3-319-55444-0_2

dice example, there are 36 possible outcomes, namely

$$S = \{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),(2,1),(2,2),(2,3),(2,4),(2,5),(2,6),$$
$$(3,1),(3,2),(3,3),(3,4),(3,5),(3,6),(4,1),(4,2),(4,3),(4,4),(4,5),(4,6),$$
$$(5,1),(5,2),(5,3),(5,4),(5,5),(5,6),(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\}.$$

- An *event* E is a specified subset of the outcomes of an experiment. The event that the sum of the dice equals 7 or 11 (the conditions to win at craps on the first roll) is the subset

$$E = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1),(5,6),(6,5)\}.$$

- The *probability of an outcome* s, denoted $p(s)$ is a number with the two properties:
  - For each outcome s in sample space S, $0 \leq p(s) \leq 1$.
  - The sum of probabilities of all outcomes adds to one: $\sum_{s \in S} p(s) = 1$.

  If we assume two distinct fair dice, the probability $p(s) = (1/6) \times (1/6) = 1/36$ for all outcomes $s \in S$.

- The *probability of an event* E is the sum of the probabilities of the outcomes of the experiment. Thus

$$p(E) = \sum_{s \in E} p(s).$$

  An alternate formulation is in terms of the *complement* of the event $\bar{E}$, the case when E does not occur. Then

$$P(E) = 1 - P(\bar{E}).$$

  This is useful, because often it is easier to analyze $P(\bar{E})$ than $P(E)$ directly.

- A *random variable* V is a numerical function on the outcomes of a probability space. The function "sum the values of two dice" ($V((a,b)) = a+b$) produces an integer result between 2 and 12. This implies a probability distribution of the values of the random variable. The probability $P(V(s) = 7) = 1/6$, as previously shown, while $P(V(s) = 12) = 1/36$.

- The *expected value* of a random variable V defined on a sample space S, $E(V)$ is defined

$$E(V) = \sum_{s \in S} p(s) \cdot V(s).$$

All this you have presumably seen before. But it provides the language we will use to connect between probability and statistics. The data we see usually comes from measuring properties of observed events. The theory of probability and statistics provides the tools to analyze this data.

2.1.1 Probability vs. Statistics

Probability and statistics are related areas of mathematics which concern themselves with analyzing the relative frequency of events. Still, there are fundamental differences in the way they see the world:

- *Probability* deals with predicting the likelihood of future events, while *statistics* involves the analysis of the frequency of past events.
- *Probability* is primarily a theoretical branch of mathematics, which studies the consequences of mathematical definitions. *Statistics* is primarily an applied branch of mathematics, which tries to make sense of observations in the real world.

Both subjects are important, relevant, and useful. But they are different, and understanding the distinction is crucial in properly interpreting the relevance of mathematical evidence. Many a gambler has gone to a cold and lonely grave for failing to make the proper distinction between probability and statistics.

This distinction will perhaps become clearer if we trace the thought process of a mathematician encountering her first craps game:

- If this mathematician were a probabilist, she would see the dice and think "Six-sided dice? Each side of the dice is presumably equally likely to land face up. Now *assuming* that each face comes up with probability 1/6, I can figure out what my chances are of crapping out."
- If instead a statistician wandered by, she would see the dice and think "How do I *know* that they are not loaded? I'll watch a while, and keep track of how often each number comes up. Then I can decide if my observations are consistent with the assumption of equal-probability faces. Once I'm confident enough that the dice are fair, I'll call a probabilist to tell me how to bet."

In summary, probability theory enables us to find the consequences of a given ideal world, while statistical theory enables us to measure the extent to which our world is ideal. This constant tension between theory and practice is why statisticians prove to be a tortured group of individuals compared with the happy-go-lucky probabilists.

Modern probability theory first emerged from the dice tables of France in 1654. Chevalier de Méré, a French nobleman, wondered whether the player or the house had the advantage in a particular betting game.[1] In the basic version, the player rolls four dice, and wins provided none of them are a 6. The house collects on the even money bet if at least one 6 appears.

De Méré brought this problem to the attention of the French mathematicians Blaise Pascal and Pierre de Fermat, most famous as the source of Fermat's Last Theorem. Together, these men worked out the basics of probability theory,

[1] He really shouldn't have wondered. The house *always* has the advantage.

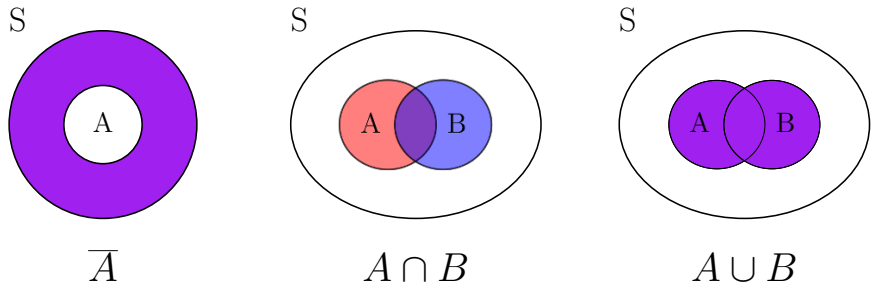

Figure 2.1: Venn diagrams illustrating set difference (left), intersection (middle), and union (right).

along the way establishing that the house wins this dice game with probability $p = 1 - (5/6)^4 \approx 0.517$, where the probability $p = 0.5$ would denote a fair game where the house wins exactly half the time.

2.1.2 Compound Events and Independence

We will be interested in complex events computed from simpler events A and B on the same set of outcomes. Perhaps event A is that at least one of two dice be an even number, while event B denotes rolling a total of either 7 or 11. Note that there exist certain outcomes of A which are not outcomes of B, specifically

$$A - B = \{(1,2),(1,4),(2,1),(2,2),(2,3),(2,4),(2,6),(3,2),(3,6),(4,1),$$
$$(4,2),(4,4),(4,5),(4,6),(5,4),(6,2),(6,3),(6,4),(6,6)\}.$$

This is the *set difference* operation. Observe that here $B - A = \{\}$, because every pair adding to 7 or 11 must contain one odd and one even number.

The outcomes in common between both events A and B are called the *intersection*, denoted $A \cap B$. This can be written as

$$A \cap B = A - (S - B).$$

Outcomes which appear in either A or B are called the *union*, denoted $A \cup B$. With the complement operation $\bar{A} = S - A$, we get a rich language for combining events, shown in Figure 2.1. We can readily compute the probability of any of these sets by summing the probabilities of the outcomes in the defined sets.

The events A and B are *independent* if and only if

$$P(A \cap B) = P(A) \times P(B).$$

This means that there is no special structure of outcomes shared between events A and B. Assuming that half of the students in my class are female, and half the students in my class are above average, we would expect that a quarter of my students are both female and above average if the events are independent.

Probability theorists love independent events, because it simplifies their calculations. But data scientists generally don't. When building models to predict the likelihood of some future event B, given knowledge of some previous event A, we want as strong a dependence of B on A as possible.

Suppose I always use an umbrella if and only if it is raining. Assume that the probability it is raining here (event B) is, say, $p = 1/5$. This implies the probability that I am carrying my umbrella (event A) is $q = 1/5$. But even more, if you know the state of the rain you know exactly whether I have my umbrella. These two events are perfectly *correlated*.

By contrast, suppose the events were independent. Then

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A)$$

and whether it is raining has absolutely no impact on whether I carry my protective gear.

Correlations are the driving force behind predictive models, so we will discuss how to measure them and what they mean in Section 2.3.

2.1.3 Conditional Probability

When two events are correlated, there is a dependency between them which makes calculations more difficult. The *conditional probability* of A given B, $P(A|B)$ is defined:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Recall the dice rolling events from Section 2.1.2, namely:

- Event A is that at least one of two dice be an even number.
- Event B is the sum of the two dice is either a 7 or an 11.

Observe that $P(A|B) = 1$, because *any* roll summing to an odd value must consist of one even and one odd number. Thus $A \cap B = B$, analogous to the umbrella case above. For $P(B|A)$, note that $P(A \cap B) = 9/36$ and $P(A) = 25/36$, so $P(B|A) = 9/25$.

Conditional probability will be important to us, because we are interested in the likelihood of an event A (perhaps that a particular piece of email is spam) as a function of some evidence B (perhaps the distribution of words within the document). Classification problems generally reduce to computing conditional probabilities, in one way or another.

Our primary tool to compute conditional probabilities will be *Bayes theorem*, which reverses the direction of the dependencies:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

Often it proves easier to compute probabilities in one direction than another, as in this problem. By Bayes theorem $P(B|A) = (1 \cdot 9/36)/(25/36) = 9/25$, exactly

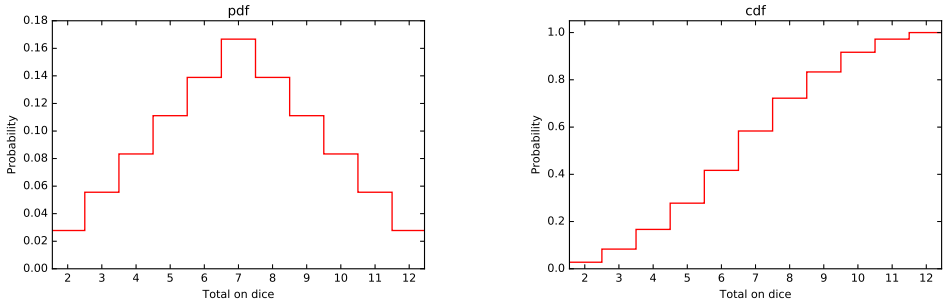

Figure 2.2: The probability density function (pdf) of the sum of two dice contains exactly the same information as the cumulative density function (cdf), but looks very different.

what we got before. We will revisit Bayes theorem in Section 5.6, where it will establish the foundations of computing probabilities in the face of evidence.

2.1.4 Probability Distributions

Random variables are numerical functions where the values are associated with probabilities of occurrence. In our example where $V(s)$ the sum of two tossed dice, the function produces an integer between 2 and 12. The probability of a particular value $V(s) = X$ is the sum of the probabilities of all the outcomes which add up to X.

Such random variables can be represented by their *probability density function*, or pdf. This is a graph where the x-axis represents the range of values the random variable can take on, and the y-axis denotes the probability of that given value. Figure 2.2 (left) presents the pdf of the sum of two fair dice. Observe that the peak at $X = 7$ corresponds to the most frequent dice total, with a probability of $1/6$.

Such pdf plots have a strong relationship to histograms of data frequency, where the x-axis again represents the range of value, but y now represents the observed frequency of exactly how many event occurrences were seen for each given value X. Converting a histogram to a pdf can be done by dividing each bucket by the total frequency over all buckets. The sum of the entries then becomes 1, so we get a probability distribution.

Histograms are statistical: they reflect actual observations of outcomes. In contrast, pdfs are probabilistic: they represent the underlying chance that the next observation will have value X. We often use the histogram of observations $h(x)$ in practice to estimate the probabilities[2] by normalizing counts by the total

[2] A technique called *discounting* offers a better way to estimate the frequency of rare events, and will be discussed in Section 11.1.2.

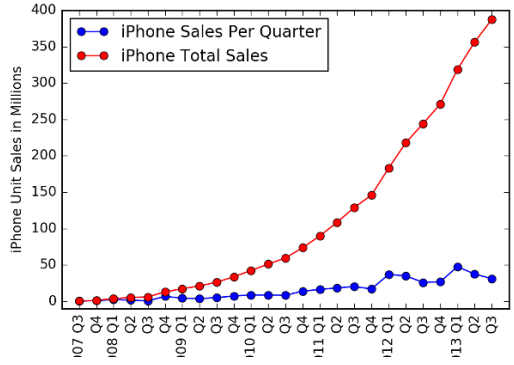

Figure 2.3: iPhone quarterly sales data presented as cumulative and incremental (quarterly) distributions. Which curve did Apple CEO Tim Cook choose to present?

number of observations:

$$P(k = X) = \frac{h(k = X)}{\sum_x h(x = X)}$$

There is another way to represent random variables which often proves useful, called a *cumulative density function* or cdf. The cdf is the running sum of the probabilities in the pdf; as a function of k, it reflects the probability that $X \leq k$ instead of the probability that $X = k$. Figure 2.2 (right) shows the cdf of the dice sum distribution. The values increase monotonically from left to right, because each term comes from adding a positive probability to the previous total. The rightmost value is 1, because all outcomes produce a value no greater than the maximum.

It is important to realize that the pdf $P(V)$ and cdf $C(V)$ of a given random variable V contain *exactly* the same information. We can move back and forth between them because:

$$P(k = X) = C(X \leq k + \delta) - C(X \leq k),$$

where $\delta = 1$ for integer distributions. The cdf is the running sum of the pdf, so

$$C(X \leq k) = \sum_{x \leq k} P(X = x).$$

Just be aware of which distribution you are looking at. Cumulative distributions always get higher as we move to the right, culminating with a probability of $C(X \leq \infty) = 1$. By contrast, the total area under the curve of a pdf equals 1, so the probability at any point in the distribution is generally substantially less.

An amusing example of the difference between cumulative and incremental distributions is shown in Figure 2.3. Both distributions show exactly the same data on Apple iPhone sales, but which curve did Apple CEO Tim Cook choose to present at a major shareholder event? The cumulative distribution (red) shows that sales are exploding, right? But it presents a misleading view of growth rate, because incremental change is the derivative of this function, and hard to visualize. Indeed, the sales-per-quarter plot (blue) shows that the rate of iPhone sales actually had declined for the last two periods before the presentation.

2.2 Descriptive Statistics

Descriptive statistics provide ways of capturing the properties of a given data set or sample. They summarize observed data, and provide a language to talk about it. Representing a group of elements by a new derived element, like mean, min, count, or sum reduces a large data set to a small summary statistic: aggregation as data reduction.

Such statistics can become features in their own right when taken over natural groups or clusters in the full data set. There are two main types of descriptive statistics:

- *Central tendency measures*, which capture the center around which the data is distributed.
- *Variation* or *variability measures*, which describe the data spread, i.e. how far the measurements lie from the center.

Together these statistics tell us an enormous amount about our distribution.

2.2.1 Centrality Measures

The first element of statistics we are exposed to in school are the basic centrality measures: mean, median, and mode. These are the right place to start when thinking of a single number to characterize a data set.

- *Mean*: You are probably quite comfortable with the use of the *arithmetic mean*, where we sum values and divide by the number of observations:

$$\mu_X = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  We can easily maintain the mean under a stream of insertions and deletions, by keeping the sum of values separate from the frequency count, and divide only on demand.

  The mean is very meaningful to characterize symmetric distributions without outliers, like height and weight. That it is symmetric means the number of items above the mean should be roughly the same as the number

below. That it is without outliers means that the range of values is reasonably tight. Note that a single MAXINT creeping into an otherwise sound set of observations throws the mean wildly off. The median is a centrality measure which proves more appropriate with such ill-behaved distributions.

- *Geometric mean*: The *geometric mean* is the nth root of the product of n values:

$$\left(\prod_{i=1}^{n} a_i\right)^{1/n} = \sqrt[n]{a_1 a_2 \dots a_n}$$

The geometric mean is always less than or equal to the arithmetic mean. For example, the geometric mean of the sums of 36 dice rolls is 6.5201, as opposed to the arithmetic mean of 7. It is very sensitive to values near zero. A single value of zero lays waste to the geometric mean: no matter what other values you have in your data, you end up with zero. This is somewhat analogous to having an outlier of ∞ in an arithmetic mean.

But geometric means prove their worth when averaging ratios. The geometric mean of 1/2 and 2/1 is 1, whereas the mean is 1.25. There is less available "room" for ratios to be less than 1 than there is for ratios above 1, creating an asymmetry that the arithmetic mean overstates. The geometric mean is more meaningful in these cases, as is the arithmetic mean of the *logarithms* of the ratios.

- *Median*: The *median* is the exact middle value among a data set; just as many elements lie above the median as below it. There is a quibble about what to take as the median when you have an even number of elements. You can take either one of the two central candidates: in any reasonable data set these two values should be about the same. Indeed in the dice example, both are 7.

A nice property of the median as so defined is that it must be a genuine value of the original data stream. There actually is someone of median height to you can point to as an example, but presumably no one in the world is of *exactly* average height. You lose this property when you average the two center elements.

Which centrality measure is best for applications? The median typically lies pretty close to the arithmetic mean in symmetrical distributions, but it is often interesting to see how far apart they are, and on which side of the mean the median lies.

The median generally proves to be a better statistic for skewed distributions or data with outliers: like wealth and income. Bill Gates adds $250 to the mean per capita wealth in the United States, but nothing to the median. If he makes you personally feel richer, then go ahead and use the mean. But the median is the more informative statistic here, as it will be for any power law distribution.

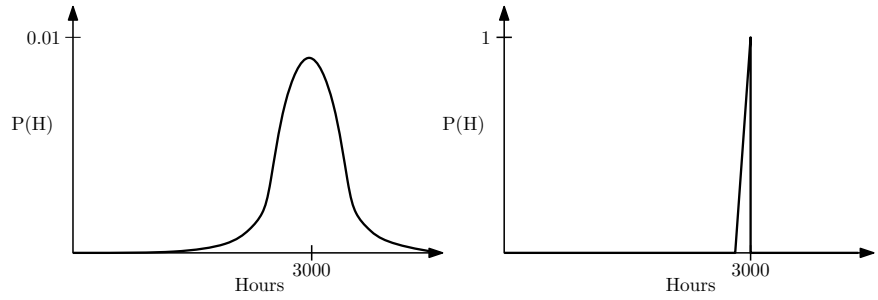

Figure 2.4: Two distinct probability distributions with $\mu = 3000$ for the lifespan of light bulbs: normal (left) and with zero variance (right).

- *Mode*: The *mode* is the most frequent element in the data set. This is 7 in our ongoing dice example, because it occurs six times out of thirty-six elements. Frankly, I've never seen the mode as providing much insight as centrality measure, because it often isn't close to the center. Samples measured over a large range should have very few repeated elements or collisions at any particular value. This makes the mode a matter of happenstance. Indeed, the most frequently occurring elements often reveal artifacts or anomalies in a data set, such as default values or error codes that do not really represent elements of the underlying distribution.

  The related concept of the peak in a frequency distribution (or histogram) is meaningful, but interesting peaks only get revealed through proper bucketing. The current peak of the annual salary distribution in the United States lies between $30,000 and $40,000 per year, although the mode presumably sits at zero.

2.2.2 Variability Measures

The most common measure of variability is the *standard deviation* σ, which measures sum of squares differences between the individual elements and the mean:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(a_i - \bar{a})^2}{n-1}}$$

A related statistic, the *variance* V, is the square of the standard deviation, i.e. $V = \sigma^2$. Sometimes it is more convenient to talk about variance than standard deviation, because the term is eight characters shorter. But they measure exactly the same thing.

As an example, consider the humble light bulb, which typically comes with an expected working life, say $\mu = 3000$ hours, derived from some underlying distribution shown in Figure 2.4. In a conventional bulb, the chance of it lasting longer than μ is presumably about the same as that of it burning out quicker, and this degree of uncertainty is measured by σ. Alternately, imagine a "printer

cartridge bulb," where the evil manufacturer builds very robust bulbs, but includes a counter so they can prevent it from ever glowing after 3000 hours of use. Here $\mu = 3000$ and $\sigma = 0$. Both distributions have the same mean, but substantially different variance.

The sum of squares penalty in the formula for σ means that one outlier value d units from the mean contributes as much to the variance as d^2 points each one unit from the mean, so the variance is very sensitive to outliers.

An often confusing matter concerns the denominator in the formula for standard deviation. Should we divide by n or $n-1$? The difference here is technical. The standard deviation of the full *population* divides by n, whereas the standard deviation of the *sample* divides by $n-1$. The issue is that sampling just one point tells us absolutely nothing about the underlying variance in any population, where it is perfectly reasonable to say there is zero variance in weight among the population of a one-person island. But for reasonable-sized data sets $n \approx (n-1)$, so it really doesn't matter.

2.2.3 Interpreting Variance

Repeated observations of the same phenomenon do not always produce the same results, due to random noise or error. *Sampling errors* result when our observations capture unrepresentative circumstances, like measuring rush hour traffic on weekends as well as during the work week. *Measurement errors* reflect the limits of precision inherent in any sensing device. The notion of *signal to noise ratio* captures the degree to which a series of observations reflects a quantity of interest as opposed to data variance. As data scientists, we care about changes in the signal instead of the noise, and such variance often makes this problem surprisingly difficult.

I think of variance as an inherent property of the universe, akin to the speed of light or the time-value of money. Each morning you weigh yourself on a scale you are guaranteed to get a different number, with changes reflecting when you last ate (sampling error), the flatness of the floor, or the age of the scale (both measurement error) as much as changes in your body mass (actual variation). So what is your real weight?

Every measured quantity is subject to some level of variance, but the phenomenon cuts much deeper than that. Much of what happens in the world is just random fluctuations or arbitrary happenstance causing variance even when the situation is unchanged. Data scientists seek to explain the world through data, but distressingly often there is no real phenomena to explain, only a ghost created by variance. Examples include:

- *The stock market*: Consider the problem of measuring the relative "skill" of different stock market investors. We know that Warren Buffet is much better at investing than we are. But very few professional investors prove consistently better than others. Certain investment vehicles wildly outperform the market in any given time period. However, the hot fund one

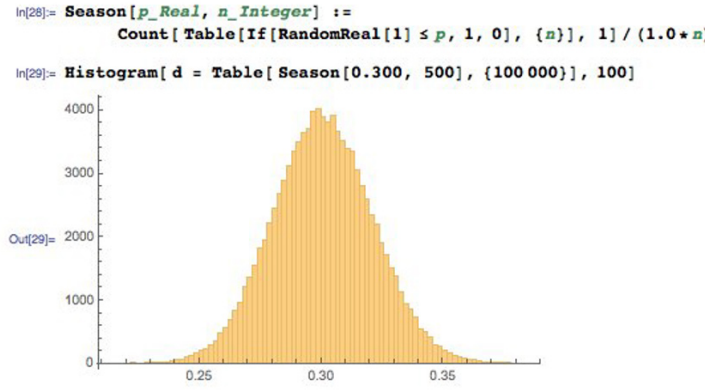

Figure 2.5: Sample variance on hitters with a real 30% success rate results in a wide range of observed performance even over 500 trials per season.

year usually underperforms the market the year after, which shouldn't happen if this outstanding performance was due to skill rather than luck.

The fund managers themselves are quick to credit profitable years to their own genius, but losses to unforeseeable circumstances. However, several studies have shown that the performance of professional investors is essentially random, meaning there is little real difference in skill. Most investors are paying managers for previously-used luck. So why do these entrail-readers get paid so much money?

- *Sports performance*: Students have good semesters and bad semesters, as reflected by their grade point average (GPA). Athletes have good and bad seasons, as reflected by their performance and statistics. Do such changes reflect genuine differences in effort and ability, or are they just variance?

  In baseball, .300 hitters (players who hit with a 30% success rate) represent consistency over a full season. Batting .275 is not a noteworthy season, but hit .300 and you are a star. Hit .325 and you are likely to be the batting champion.

  Figure 2.5 shows the results of a simple simulation, where random numbers were used to decide the outcome of each at-bat over a 500 at-bats/season. Our synthetic player is a *real* .300 hitter, because we programmed it to report a hit with probability 300/1000 (0.3). The results show that a real .300 hitter has a 10% chance of hitting .275 or below, just by chance. Such a season will typically be explained away by injuries or maybe the inevitable effects of age on athletic performance. But it could just be natural variance. Smart teams try to acquire a good hitter after a lousy season, when the price is cheaper, trying to take advantage of this variance.

  Our .300 hitter also has a 10% chance of batting above .325, but you

can be pretty sure that they will ascribe such a breakout season to their improved conditioning or training methods instead of the fact they just got lucky. Good or bad season, or lucky/unlucky: it is hard to tell the signal from the noise.

- *Model performance*: As data scientists, we will typically develop and evaluate several models for each predictive challenge. The models may range from very simple to complex, and vary in their training conditions or parameters.

  Typically the model with the best accuracy on the training corpus will be paraded triumphantly before the world as the right one. But small differences in the performance between models is likely explained by simple variance rather than wisdom: which training/evaluation pairs were selected, how well parameters were optimized, etc.

  Remember this when it comes to training machine learning models. Indeed, when asked to choose between models with small performance differences between them, I am more likely to argue for the simplest model than the one with the highest score. Given a hundred people trying to predict heads and tails on a stream of coin tosses, one of them is guaranteed to end up with the most right answers. But there is no reason to believe that this fellow has any better predictive powers than the rest of us.

2.2.4 Characterizing Distributions

Distributions do not necessarily have much probability mass exactly at the mean. Consider what your wealth would look like after you borrow \$100 million, and then bet it all on an even money coin flip. Heads you are now \$100 million in clear, tails you are \$100 million in hock. Your expected wealth is zero, but this mean does not tell you much about the shape of your wealth distribution.

However, taken together the mean and standard deviation do a decent job of characterizing *any* distribution. Even a relatively small amount of mass positioned far from the mean would add a lot to the standard deviation, so a small value of σ implies the bulk of the mass must be near the mean.

To be precise, regardless of how your data is distributed, at least $(1 - (1/k^2))$th of the mass must lie within $\pm k$ standard deviations of the mean. This means that at least 75% of all the data must lie within 2σ of the mean, and almost 89% within 3σ for any distribution.

We will see that even tighter bounds hold when we know the distribution is well-behaved, like the Gaussian or normal distribution. But this is why it is a great practice to report both μ and σ whenever you talk about averages. The average height of adult women in the United States is 63.7 ± 2.7 inches, meaning $\mu = 63.7$ and $\sigma = 2.7$. The average temperature in Orlando, Fl is 60.3 degrees Fahrenheit. However, there have been many more 100 degree days at Disney World than 100 inch (8.33 foot) women visiting to enjoy them.

Take-Home Lesson: Report both the mean and standard deviation to characterize your distribution, written as $\mu \pm \sigma$.

2.3 Correlation Analysis

Suppose we are given two variables x and y, represented by a sample of n points of the form (x_i, y_i), for $1 \leq i \leq n$. We say that x and y are *correlated* when the value of x has some predictive power on the value of y.

The *correlation coefficient* $r(X, Y)$ is a statistic that measures the degree to which Y is a function of X, and vice versa. The value of the correlation coefficient ranges from -1 to 1, where 1 means fully correlated and 0 implies no relation, or independent variables. Negative correlations imply that the variables are *anti-correlated*, meaning that when X goes up, Y goes down.

Perfectly anti-correlated variables have a correlation of -1. Note that negative correlations are just as good for predictive purposes as positive ones. That you are less likely to be unemployed the more education you have is an example of a negative correlation, so the level of education can indeed help predict job status. Correlations around 0 are useless for forecasting.

Observed correlations drives many of the predictive models we build in data science. Representative strengths of correlations include:

- Are taller people more likely to remain lean? The observed correlation between height and BMI is $r = -0.711$, so height is indeed negatively correlated with body mass index (BMI).[3]

- Do standardized tests predict the performance of students in college? The observed correlation between SAT scores and freshmen GPA is $r = 0.47$, so yes, there is some degree of predictive power. But social economic status is just as strongly correlated with SAT scores ($r = 0.42$).[4]

- Does financial status affect health? The observed correlation between household income and the prevalence of coronary artery disease is $r = -0.717$, so there is a strong negative correlation. So yes, the wealthier you are, the lower your risk of having a heart attack.[5]

- Does smoking affect health? The observed correlation between a group's propensity to smoke and their mortality rate is $r = 0.716$, so for G-d's sake, don't smoke.[6]

[3] `https://onlinecourses.science.psu.edu/stat500/node/60`

[4] `https://research.collegeboard.org/sites/default/files/publications/2012/9/researchreport-2009-1-socioeconomic-status-sat-freshman-gpa-analysis-data.pdf`

[5] `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3457990/`.

[6] `http://lib.stat.cmu.edu/DASL/Stories/SmokingandCancer.html`.

- Do violent video games increase aggressive behavior? The observed correlation between play and violence is $r = 0.19$, so there is a weak but significant correlation.[7]

This section will introduce the primary measurements of correlation. Further, we study how to appropriately determine the strength and power of any observed correlation, to help us understand when the connections between variables are real.

2.3.1 Correlation Coefficients: Pearson and Spearman Rank

In fact, there are two primary statistics used to measure correlation. Mercifully, both operate on the same -1 to 1 scale, although they measure somewhat different things. These different statistics are appropriate in different situations, so you should be aware of both of them.

The Pearson Correlation Coefficient

The more prominent of the two statistics is *Pearson* correlation, defined as

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)}$$

Let's parse this equation. Suppose X and Y are strongly correlated. Then we would expect that when x_i is greater than the mean $\bar{X}$, then y_i should be bigger than its mean $\bar{Y}$. When x_i is lower than its mean, y_i should follow. Now look at the numerator. The sign of each term is positive when both values are above (1×1) or below (-1×-1) their respective means. The sign of each term is negative $((-1 \times 1)$ or $(1 \times -1))$ if they move in opposite directions, suggesting negative correlation. If X and Y were uncorrelated, then positive and negative terms should occur with equal frequency, offsetting each other and driving the value to zero.

The numerator's operation determining the sign of the correlation is so useful that we give it a name, *covariance*, computed:

$$Cov(X,Y) = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}).$$

Remember covariance: we will see it again in Section 8.2.3.

The denominator of the Pearson formula reflects the amount of variance in the two variables, as measured by their standard deviations. The covariance between X and Y potentially increases with the variance of these variables, and this denominator is the magic amount to divide it by to bring correlation to a -1 to 1 scale.

[7] `http://webspace.pugetsound.edu/facultypages/cjones/chidev/Paper/Articles/Anderson-Aggression.pdf`.

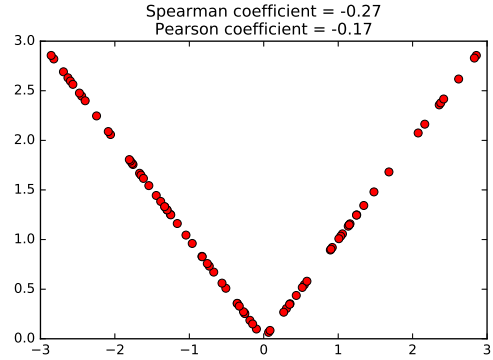

Figure 2.6: The function $y = |x|$ does not have a linear model, but seems like it should be easily fitted despite weak correlations.

The Spearman Rank Correlation Coefficient

The Pearson correlation coefficient defines the degree to which a linear predictor of the form $y = m \cdot x + b$ can fit the observed data. This generally does a good job measuring the similarity between the variables, but it is possible to construct pathological examples where the correlation coefficient between X and Y is zero, yet Y is completely dependent on (and hence perfectly predictable from) X.

Consider points of the form $(x, |x|)$, where x is uniformly (or symmetrically) sampled from the interval $[-1, 1]$ as shown in Figure 2.6. The correlation will be zero because for every point (x, x) there will be an offsetting point $(-x, x)$, yet $y = |x|$ is a perfect predictor. Pearson correlation measures how well the best *linear* predictors can work, but says nothing about weirder functions like absolute value.

The *Spearman rank correlation coefficient* essentially counts the number of pairs of input points which are out of order. Suppose that our data set contains points (x_1, y_1) and (x_2, y_2) where $x_1 < x_2$ and $y_1 < y_2$. This is a vote that the values are positively correlated, whereas the vote would be for a negative correlation if $y_2 < y_1$.

Summing up over all pairs of points and normalizing properly gives us Spearman rank correlation. Let $rank(x_i)$ be the rank position of x_i in sorted order among all x_i, so the rank of the smallest value is 1 and the largest value n. Then

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where $d_i = rank(x_i) - rank(y_i)$.

The relationship between our two coefficients is better delineated by the example in Figure 2.7. In addition to giving high scores to non-linear but monotonic functions, Spearman correlation is less sensitive to extreme outlier elements than Pearson. Let $p = (x_1, y_{\max})$ be the data point with largest value

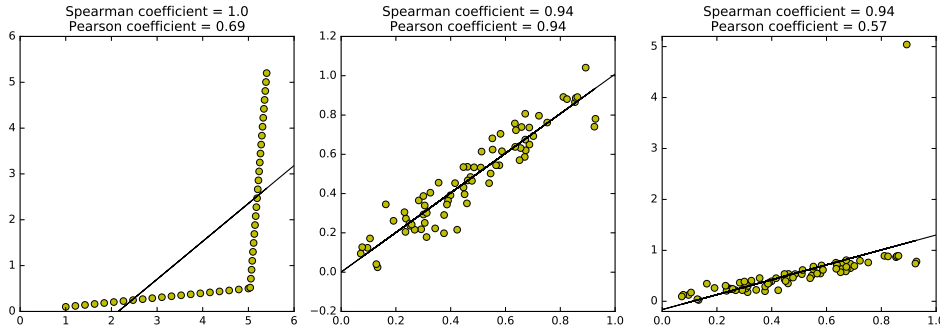

Figure 2.7: A monotonic but not linear point set has a Spearman coefficient $r = 1$ even though it has no good linear fit (left). Highly-correlated sequences are recognized by both coefficients (center), but the Pearson coefficient is much more sensitive to outliers (right).

of y in a given data set. Suppose we replace p with $p' = (x_1, \infty)$. The Pearson correlation will go crazy, since the best fit now becomes the vertical line $x = x_1$. But the Spearman correlation will be unchanged, since all the points were under p, just as they are now under p'.

2.3.2 The Power and Significance of Correlation

The correlation coefficient r reflects the degree to which x can be used to predict y in a given sample of points S. As $|r| \to 1$, these predictions get better and better.

But the real question is how this correlation will hold up in the real world, outside the sample. Stronger correlations have larger $|r|$, but also involve samples of enough points to be significant. There is a wry saying that if you want to fit your data by a straight line, it is best to sample it at only two points. Your correlation becomes more impressive the more points it is based on.

The statistical limits in interpreting correlations are presented in Figure 2.8, based on strength and size:

- *Strength of correlation:* R^2: The square of the sample correlation coefficient r^2 estimates the fraction of the variance in Y explained by X in a simple linear regression. The correlation between height and weight is approximately 0.8, meaning it explains about two thirds of the variance.

  Figure 2.8 (left) shows how rapidly r^2 decreases with r. There is a profound limit to how excited we should get about establishing a weak correlation. A correlation of 0.5 possesses only 25% of the maximum predictive power, and a correlation of $r = 0.1$ only 1%. Thus the predictive value of correlations decreases rapidly with r.

  What do we mean by "explaining the variance"? Let $f(x) = mx + c$ be the

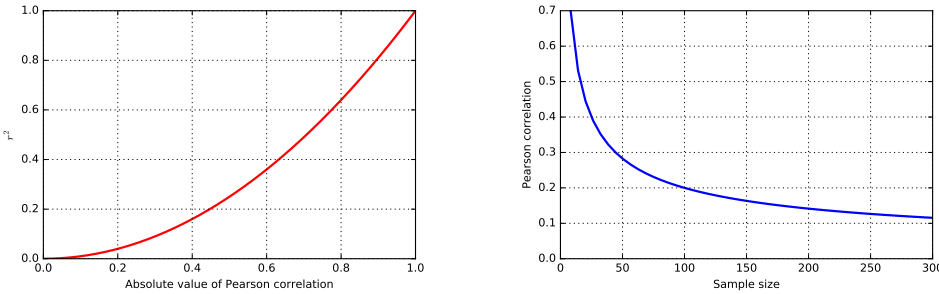

Figure 2.8: Limits in interpreting significance. The r^2 value shows that weak correlations explain only a small fraction of the variance (left). The level of correlation necessary to be statistically significance decreases rapidly with sample size n (right).

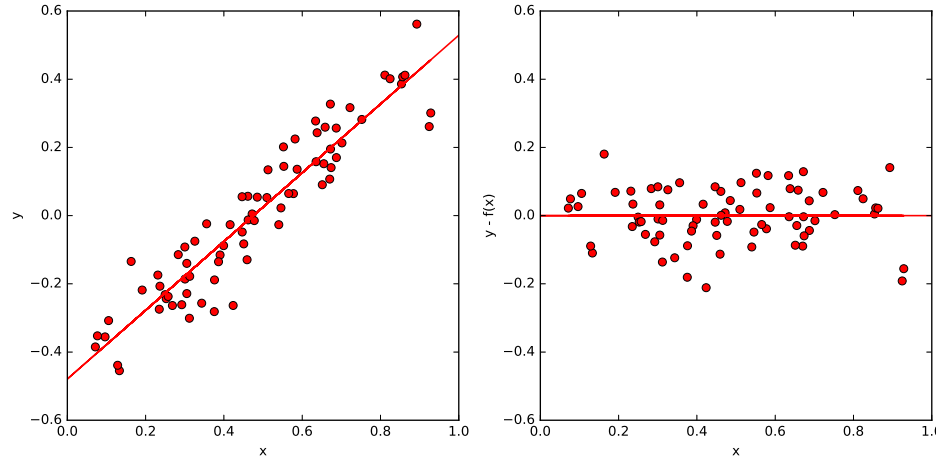

Figure 2.9: Plotting $r_i = y_i - f(x_i)$ shows that the residual values have lower variance and mean zero. The original data points are on the left, with the corresponding residuals on the right.

predictive value of y from x, with the parameters m and c corresponding to the best possible fit. The *residual* values $r_i = y_i - f(x_i)$ will have mean zero, as shown in Figure 2.9. Further, the variance of the full data set $V(Y)$ should be much larger than $V(r)$ if there is a good linear fit $f(x)$. If x and y are perfectly correlated, there should be no residual error, and $V(r) = 0$. If x and y are totally uncorrelated, the fit should contribute nothing, and $V(y) \approx V(r)$. Generally speaking, $1 - r^2 = V(r)/V(y)$.

Consider Figure 2.9, showing a set of points (left) admitting a good linear fit, with correlation $r = 0.94$. The corresponding residuals $r_i = y_i - f(x_i)$ are plotted on the right. The variance of the y values on the left $V(y) = 0.056$, substantially greater than the variance $V(r) = 0.0065$ on the right. Indeed,

$$1 - r^2 = 0.116 \longleftrightarrow V(r)/V(y) = 0.116.$$

- *Statistical significance*: The statistical significance of a correlation depends upon its sample size n as well as r. By tradition, we say that a correlation of n points is *significant* if there is an $\alpha \leq 1/20 = 0.05$ chance that we would observe a correlation as strong as r in any random set of n points.

  This is not a particularly strong standard. Even small correlations become significant at the 0.05 level with large enough sample sizes, as shown in Figure 2.8 (right). A correlation of $r = 0.1$ becomes significant at $\alpha = 0.05$ around $n = 300$, even though such a factor explains only 1% of the variance.

Weak but significant correlations can have value in big data models involving large numbers of features. Any single feature/correlation might explain/predict only small effects, but taken together a large number of weak but independent correlations may have strong predictive power. *Maybe.* We will discuss significance again in greater detail in Section 5.3.

2.3.3 Correlation Does Not Imply Causation!

You have heard this before: correlation does not imply causation:

- The number of police active in a precinct correlate strongly with the local crime rate, but the police do not cause the crime.
- The amount of medicine people take correlates with the probability they are sick, but the medicine does not cause the illness.

At best, the implication works only one way. But many observed correlations are completely spurious, with neither variable having any real impact on the other.

Still, *correlation implies causation* is a common error in thinking, even among those who understand logical reasoning. Generally speaking, few statistical tools are available to tease out whether A really causes B. We can conduct controlled experiments, if we can manipulate one of the variables and watch the effect on

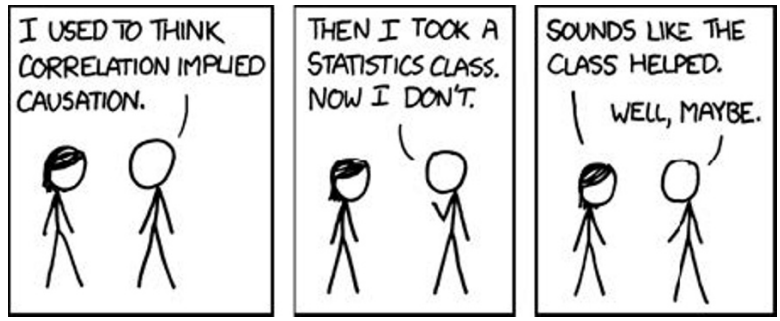

Figure 2.10: Correlation does not imply causation. (Source `https://www.xkcd.com/552`.)

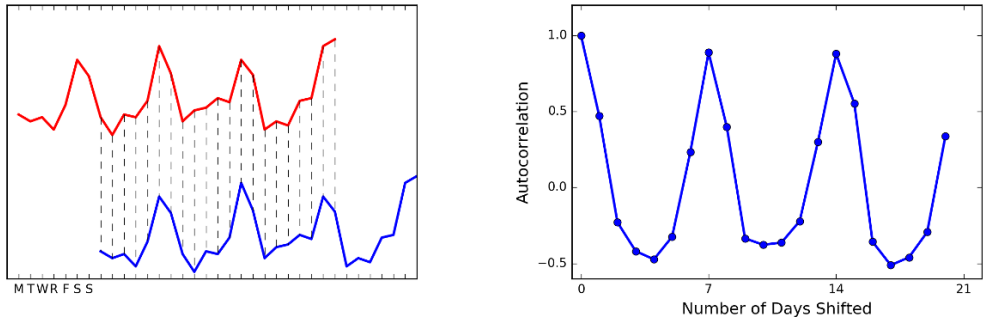

Figure 2.11: Cyclic trends in a time series (left) are revealed through correlating it against shifts of itself (right) .

the other. For example, the fact that we can put people on a diet that makes them lose weight without getting shorter is convincing evidence that weight does not *cause* height. But it is often harder to do these experiments the other way, e.g. there is no reasonable way to make people shorter other than by hacking off limbs.

2.3.4 Detecting Periodicities by Autocorrelation

Suppose a space alien was hired to analyze U.S. sales at a toy company. Instead of a nice smooth function showing a consistent trend, they would be astonished to see a giant bump every twelfth month, every year. This alien would have discovered the phenomenon of Christmas.

Seasonal trends reflect cycles of a fixed duration, rising and falling in a regular pattern. Many human activities proceed with a seven-day cycle associated with the work week. Large populations of a type of insect called a *cicada* emerge on a 13-year or 17-year cycle, in an effort to prevent predators from learning to

eat them.

How can we recognize such cyclic patterns in a sequence S? Suppose we correlate the values of S_i with S_{i+p}, for all $1 \leq i \leq n-p$. If the values are in sync for a particular period length p, then this correlation with itself will be unusually high relative to other possible lag values. Comparing a sequence to itself is called an *autocorrelation*, and the series of correlations for all $1 \leq k \leq n-1$ is called the *autocorrelation function*. Figure 2.11 presents a time series of daily sales, and the associated autocorrelation function for this data. The peak at a shift of seven days (and every multiple of seven days) establishes that there is a weekly periodicity in sales: more stuff gets sold on weekends.

Autocorrelation is an important concept in predicting future events, because it means we can use previous observations as features in a model. The heuristic that tomorrow's weather will be similar to today's is based on autocorrelation, with a lag of $p = 1$ days. Certainly we would expect such a model to be more accurate than predictions made on weather data from six months ago (lag $p = 180$ days).

Generally speaking, the autocorrelation function for many quantities tends to be highest for very short lags. This is why long-term predictions are less accurate than short-term forecasts: the autocorrelations are generally much weaker. But periodic cycles do sometimes stretch much longer. Indeed, a weather forecast based on a lag of $p = 365$ days will be much better than one of $p = 180$, because of seasonal effects.

Computing the full autocorrelation function requires calculating $n-1$ different correlations on points of the time series, which can get expensive for large n. Fortunately, there is an efficient algorithm based on the *fast Fourier transform* (FFT), which makes it possible to construct the autocorrelation function even for very long sequences.

2.4 Logarithms

The *logarithm* is the inverse exponential function $y = b^x$, an equation that can be rewritten as $x = \log_b y$. This definition is the same as saying that

$$b^{\log_b y} = y.$$

Exponential functions grow at a very fast rate: consider $b = \{2^1, 2^2, 2^3, 2^4, \ldots\}$. In contrast, logarithms grow a very slow rate: these are just the exponents of the previous series $\{1, 2, 3, 4, \ldots\}$. They are associated with any process where we are repeatedly multiplying by some value of b, or repeatedly dividing by b. Just remember the definition:

$$y = \log_b x \longleftrightarrow b^y = x.$$

Logarithms are very useful things, and arise often in data analysis. Here I detail three important roles logarithms play in data science. Surprisingly, only one of them is related to the seven algorithmic applications of logarithms

I present in *The Algorithm Design Manual* [Ski08]. Logarithms are indeed very useful things.

2.4.1 Logarithms and Multiplying Probabilities

Logarithms were first invented as an aide to computation, by reducing the problem of multiplication to that of addition. In particular, to compute the product $p = x \cdot y$, we could compute the sum of the logarithms $s = \log_b x + \log_b y$ and then take the inverse of the logarithm (i.e. raising b to the sth power) to get p, because:

$$p = x \cdot y = b^{(\log_b x + \log_b y)}.$$

This is the trick that powered the mechanical slide rules that geeks used in the days before pocket calculators.

However, this idea remains important today, particularly when multiplying long chains of probabilities. Probabilities are small numbers. Thus multiplying long chains of probability yield *very* small numbers that govern the chances of very rare events. There are serious numerical stability problems with floating point multiplication on real computers. Numerical errors will creep in, and will eventually overwhelm the true value of small-enough numbers.

Summing the logarithms of probabilities is much more numerically stable than multiplying them, but yields an equivalent result because:

$$\prod_{i=1}^{n} p_i = b^P, \text{ where } P = \sum_{i=1}^{n} \log_b(p_i).$$

We can raise our sum to an exponential if we need the real probability, but usually this is not necessary. When we just need to compare two probabilities to decide which one is larger we can safely stay in log world, because bigger logarithms correspond to bigger probabilities.

There is one quirk to be aware of. Recall that the $\log_2(\frac{1}{2}) = -1$. The logarithms of probabilities are all negative numbers except for $\log(1) = 0$. This is the reason why equations with logs of probabilities often feature negative signs in strange places. Be on the lookout for them.

2.4.2 Logarithms and Ratios

Ratios are quantities of the form a/b. They occur often in data sets either as elementary features or values derived from feature pairs. Ratios naturally occur in normalizing data for conditions (i.e. weight after some treatment over the initial weight) or time (i.e. today's price over yesterday's price).

But ratios behave differently when reflecting increases than decreases. The ratio 200/100 is 200% above baseline, but 100/200 is only 50% below despite being a similar magnitude change. Thus doing things like averaging ratios is committing a statistical sin. Do you really want a doubling followed by a halving to average out as an increase, as opposed to a neutral change?

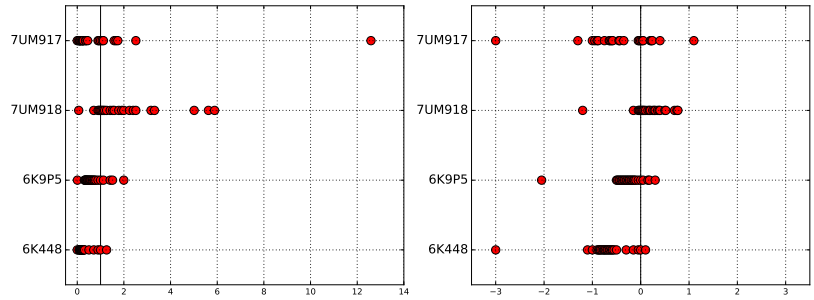

Figure 2.12: Plotting ratios on a scale cramps the space allocated to small ratios relative to large ratios (left). Plotting the logarithms of ratios better represents the underlying data (right).

One solution here would have been to use the geometric mean. But better is taking the logarithm of these ratios, so that they yield equal displacement, since $\log_2 2 = 1$ and $\log_2(1/2) = -1$. We get the extra bonus that a unit ratio maps to zero, so positive and negative numbers correspond to improper and proper ratios, respectively.

A rookie mistake my students often make involves plotting the value of ratios instead of their logarithms. Figure 2.12 (left) is a graph from a student paper, showing the ratio of new score over old score on data over 24 hours (each red dot is the measurement for one hour) on four different data sets (each given a row). The solid black line shows the ratio of one, where both scores give the same result. Now try to read this graph: it isn't easy because the points on the left side of the line are cramped together in a narrow strip. What jumps out at you are the outliers. Certainly the new algorithm does terrible on 7UM917 in the top row: that point all the way to the right is a real outlier.

Except that it isn't. Now look at Figure 2.12 (right), where we plot the logarithms of the ratios. The space devoted to left and right of the black line can now be equal. And it shows that this point wasn't *really* such an outlier at all. The magnitude of improvement of the leftmost points is much greater than that of the rightmost points. This plot reveals that new algorithm generally makes things better, only because we are showing logs of ratios instead of the ratios themselves.

2.4.3 Logarithms and Normalizing Skewed Distributions

Variables which follow symmetric, bell-shaped distributions tend to be nice as features in models. They show substantial variation, so they can be used to discriminate between things, but not over such a wide range that outliers are overwhelming.

But not every distribution is symmetric. Consider the one in Figure 2.13

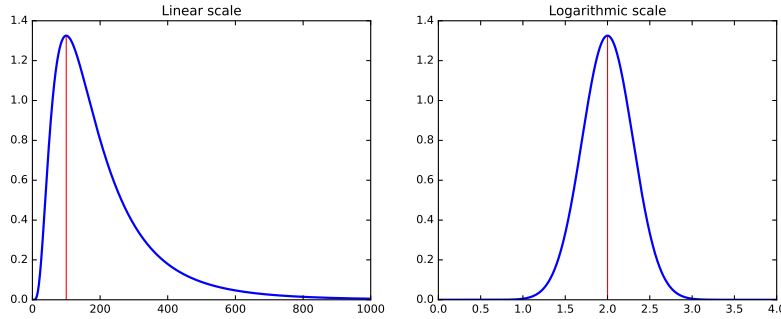

Figure 2.13: Hitting a skewed data distribution (left) with a log often yields a more bell-shaped distribution (right).

(left). The tail on the right goes much further than the tail on the left. And we are destined to see far more lopsided distributions when we discuss power laws, in Section 5.1.5. Wealth is representative of such a distribution, where the poorest human has zero or perhaps negative wealth, the average person (optimistically) is in the thousands of dollars, and Bill Gates is pushing $100 billion as of this writing.

We need a normalization to convert such distributions into something easier to deal with. To ring the bell of a power law distribution we need something non-linear, that reduces large values to a disproportionate degree compared to more modest values.

The logarithm is the transformation of choice for power law variables. Hit your long-tailed distribution with a log and often good things happen. The distribution in Figure 2.13 happened to be the *log normal* distribution, so taking the logarithm yielded a perfect bell-curve on right. Taking the logarithm of variables with a power law distribution brings them more in line with traditional distributions. For example, as an upper-middle class professional, my wealth is roughly the same number of logs from my starving students as I am from Bill Gates!

Sometimes taking the logarithm proves too drastic a hit, and a less dramatic non-linear transformation like the square root works better to normalize a distribution. The acid test is to plot a frequency distribution of the transformed values and see if it looks bell-shaped: grossly-symmetric, with a bulge in the middle. That is when you know you have the right function.

2.5 War Story: Fitting Designer Genes

The word *bioinformatician* is life science speak for "data scientist," the practitioner of an emerging discipline which studies massive collections of DNA sequence data looking for patterns. Sequence data is very interesting to work

with, and I have played bioinformatician in research projects since the very beginnings of the human genome project.

DNA sequences are strings on the four letter alphabet $\{A, C, G, T\}$. Proteins form the stuff that we are physically constructed from, and are composed of strings of 20 different types of molecular units, called amino acids. *Genes* are the DNA sequences which describe exactly how to make specific proteins, with the units each described by a triplet of $\{A, C, G, T\}$s called *codons*.

For our purposes, it suffices to know that there are a huge number of possible DNA sequences describing genes which *could* code for any particular desired protein sequence. But only one of them *is* used. My biologist collaborators and I wanted to know why.

Originally, it was assumed that all of these different synonymous encodings were essentially identical, but statistics performed on sequence data made it clear that certain codons are used more often than others. The biological conclusion is that "codons matter," and there are good biological reasons why this should be.

We became interested in whether "neighboring pairs of codon matter." Perhaps certain pairs of triples are like oil and water, and hate to mix. Certain letter pairs in English have order preferences: you see the bigram gh far more often than hg. Maybe this is true of DNA as well? If so, there would be pairs of triples which should be underrepresented in DNA sequence data.

To test this, we needed a score comparing the number of times we actually see a particular triple (say $x = CAT$) next to another particular triple (say $y = GAG$) to what we would expect by chance. Let $F(xy)$ be the frequency of xy, number of times we actually see codon x followed by codon y in the DNA sequence database. These codons code for specific amino acids, say a and b respectively. For amino acid a, the probability that it will be coded by x is $P(x) = F(x)/F(a)$, and similarly $P(y) = F(y)/F(b)$. Then the expected number of times of seeing xy is

$$Expected(xy) = \left(\frac{F(x)}{F(a)}\right)\left(\frac{F(y)}{F(b)}\right)F(ab)$$

Based on this, we can compute a codon pair score for any given hexamer xy as follows:

$$CPS(xy) = \ln\left(\frac{Observed(xy)}{Expected(xy)}\right) = \ln\left(\frac{F(xy)}{\frac{F(x)F(y)}{F(a)F(b)}F(ab)}\right)$$

Taking the logarithm of this ratio produced very nice properties. Most importantly, the sign of the score distinguished over-represented pairs from under-represented pairs. Because the magnitudes were symmetric ($+1$ was just as impressive as -1) we could add or average these scores in a sensible way to give a score for each gene. We used these scores to design genes that should be bad for viruses, which gave an exciting new technology for making vaccines. See the chapter notes (Section 2.6) for more details.

Fr. Dep.	Score		Fr. Ind.	Score
CA**TAG**G	-1.74		GGGGGG	-1.01
TC**TAG**C	-1.61		CCCCCC	-0.95
GT**TAG**G	-1.58		GGCGCC	-0.66
GC**TAG**T	-1.48		GGGGGT	-0.63
CC**TAG**T	-1.44		CGGGGG	-0.59
GG**TAG**G	-1.41		AGGGGG	-0.58
CT**TAG**G	-1.40		CACGTG	-0.58
AC**TAG**C	-1.38		ACCCCC	-0.56
GC**TAG**C	-1.37		GGGCCC	-0.56
GC**TAG**A	-1.36		CCCCCT	-0.53
CC**TAG**C	-1.35		CGCCCC	-0.52
GA**TAG**G	-1.35		CCCCCG	-0.51

Figure 2.14: Patterns in DNA sequences with the lowest codon pair scores become obvious on inspection. When interpreted in-frame, the stop symbol TAG is substantially depleted (left). When interpreted in the other two frames, the most avoided patterns are all very low complexity, like runs of a single base (right)

Knowing that certain pairs of codons were bad did not explain *why* they were bad. But by computing two related scores (details unimportant) and sorting the triplets based on them, as shown in Figure 2.14, certain patterns popped out. Do you notice the patterns? All the bad sequences on the left contain TAG, which turns out to be a special codon that tells the gene to stop. And all the bad sequences on the right consist of C and G in very simple repetitive sequences. These explain biologically why patterns are avoided by evolution, meaning we discovered something very meaningful about life.

There are two take-home lessons from this story. First, developing numerical scoring functions which highlight specific aspects of items can be very useful to reveal patterns. Indeed, Chapter 4 will focus on the development of such systems. Second, hitting such quantities with a logarithm can make them even more useful, enabling us to see the forest for the trees.

2.6 Chapter Notes

There are many excellent introductions to probability theory available, including [Tij12, BT08]. The same goes for elementary statistics, with good introductory texts including [JWHT13, Whe13]. The brief history of probability theory in this chapter is based on Weaver [Wea82].

In its strongest form, the efficient market hypothesis states that the stock market is essentially unpredictable using public information. My personal advice is that you should invest in index funds that do not actively try to predict the direction of the market. Malkiel's *A Random Walk Down Wall Street* [Mal99]

is an excellent introduction to such investment thinking.

The Fast Fourier Transform (FFT) provides an $O(n \log n)$ time algorithm to compute the full autocorrelation function of an n-element sequence, where the straightforward computation of n correlations takes $O(n^2)$. Bracewell [Bra99] and Brigham [Bri88] are excellent introductions to Fourier transforms and the FFT. See also the exposition in Press et.al. [PFTV07].

The comic strip in Figure 2.10 comes from Randall Munroe's webcomic *xkcd*, specifically `https://xkcd.com/552`, and is reprinted with permission.

The war story of Section 2.5 revolves around our work on how the phenomenon of codon pair bias affects gene translation. Figure 2.14 comes from my collaborator Justin Gardin. See [CPS+08, MCP+10, Ski12] for discussions of how we exploited codon pair bias to design vaccines for viral diseases like polio and the flu.

2.7 Exercises

Probability

2-1. *[3]* Suppose that 80% of people like peanut butter, 89% like jelly, and 78% like both. Given that a randomly sampled person likes peanut butter, what is the probability that she also likes jelly?

2-2. *[3]* Suppose that $P(A) = 0.3$ and $P(B) = 0.7$.

(a) Can you compute $P(A \text{ and } B)$ if you only know $P(A)$ and $P(B)$?

(b) Assuming that events A and B arise from independent random processes:

- What is $P(A \text{ and } B)$?
- What is $P(A \text{ or } B)$?
- What is $P(A|B)$?

2-3. *[3]* Consider a game where your score is the maximum value from two dice. Compute the probability of each event from $\{1, \ldots, 6\}$.

2-4. *[8]* Prove that the cumulative distribution function of the maximum of a pair of values drawn from random variable X is the square of the original cumulative distribution function of X.

2-5. *[5]* If two binary random variables X and Y are independent, are $\bar{X}$ (the complement of X) and Y also independent? Give a proof or a counterexample.

Statistics

2-6. *[3]* Compare each pair of distributions to decide which one has the greater mean and the greater standard deviation. You do not need to calculate the actual values of μ and σ, just how they compare with each other.

(a) i. $3, 5, 5, 5, 8, 11, 11, 11, 13$.
ii. $3, 5, 5, 5, 8, 11, 11, 11, 20$.

(b) i. $-20, 0, 0, 0, 15, 25, 30, 30$.
ii. $-40, 0, 0, 0, 15, 25, 30, 30$.

(c) i. $0, 2, 4, 6, 8, 10$.
ii. $20, 22, 24, 26, 28, 30$.

(d) i. $100, 200, 300, 400, 500$.
ii. $0, 50, 300, 550, 600$.

2-7. *[3]* Construct a probability distribution where none of the mass lies within one σ of the mean.

2-8. *[3]* How does the arithmetic and geometric mean compare on random integers?

2-9. *[3]* Show that the arithmetic mean equals the geometric mean when all terms are the same.

Correlation Analysis

2-10. *[3]* True or false: a correlation coefficient of -0.9 indicates a stronger linear relationship than a correlation coefficient of 0.5. Explain why.

2-11. *[3]* What would be the correlation coefficient between the annual salaries of college and high school graduates at a given company, if for each possible job title the college graduates always made:

(a) $5,000 more than high school grads?

(b) 25% more than high school grads?

(c) 15% less than high school grads?

2-12. *[3]* What would be the correlation between the ages of husbands and wives if men always married woman who were:

(a) Three years younger than themselves?

(b) Two years older than themselves?

(c) Half as old as themselves?

2-13. *[5]* Use data or literature found in a Google search to estimate/measure the strength of the correlation between:

(a) Hits and walks scored for hitters in baseball.

(b) Hits and walks allowed by pitchers in baseball.

2-14. *[5]* Compute the Pearson and Spearman Rank correlations for uniformly drawn samples of points (x, x^k). How do these values change as a function of increasing k?

Logarithms

2-15. *[3]* Show that the logarithm of any number less than 1 is negative.

2-16. *[3]* Show that the logarithm of zero is undefined.

2-17. *[5]* Prove that

$$x \cdot y = b^{(\log_b x + \log_b y)}$$

2-18. *[5]* Prove the correctness of the formula for changing a base-b logarithm to base-a, that

$$\log_a(x) = \log_b(x) / \log_b(a).$$

Implementation Projects

2-19. *[3]* Find some interesting data sets, and compare how similar their means and medians are. What are the distributions where the mean and median differ on the most?

2-20. *[3]* Find some interesting data sets and search all pairs for interesting correlations. Perhaps start with what is available at `http://www.data-manual.com/data`. What do you find?

Interview Questions

2-21. *[3]* What is the probability of getting exactly k heads on n tosses, where the coin has a probability of p in coming up heads on each toss? What about k or more heads?

2-22. *[5]* Suppose that the probability of getting a head on the ith toss of an ever-changing coin is $f(i)$. How would you efficiently compute the probability of getting exactly k heads in n tosses?

2-23. *[5]* At halftime of a basketball game you are offered two possible challenges:

(a) Take three shots, and make at least two of them.

(b) Take eight shots, and make at least five of them.

Which challenge should you pick to have a better chance of winning the game?

2-24. *[3]* Tossing a coin ten times resulted in eight heads and two tails. How would you analyze whether a coin is fair? What is the p-value?

2-25. *[5]* Given a stream of n numbers, show how to select one uniformly at random using only constant storage. What if you don't know n in advance?

2-26. *[5]* A k-streak starts at toss i in a sequence of n coin flips when the outcome of the ith flip and the next $k-1$ flips are identical. For example, sequence HTTTHH contains 2-streaks starting at the second, third, and fifth tosses. What are the expected number of k-streaks that you will see in n tosses of a fair coin ?

2-27. *[5]* A person randomly types an eight-digit number into a pocket calculator. What is the probability that the number looks the same even if the calculator is turned upside down?

2-28. *[3]* You play a dice rolling game where you have two choices:

(a) Roll the dice once and get rewarded with a prize equal to the outcome number (e.g, \$3 for number "3") and then stop the game.

(b) You can reject the first reward according to its outcome and roll the dice a second time, and get rewarded in the same way.

Which strategy should you choose to maximize your reward? That is, for what outcomes of the first roll should you chose to play the second game? What is the statistical expectation of reward if you choose the second strategy?

2-29. *[3]* What is A/B testing and how does it work?

2-30. *[3]* What is the difference between statistical independence and correlation?

2-31. *[3]* We often say that correlation does not imply causation. What does this mean?

2-32. *[5]* What is the difference between a skewed distribution and a uniform one?

Kaggle Challenges

2-33. Cause–effect pairs: correlation vs. causation.
`https://www.kaggle.com/c/cause-effect-pairs`

2-34. Predict the next "random number" in a sequence.
`https://www.kaggle.com/c/random-number-grand-challenge`

2-35. Predict the fate of animals at a pet shelter.
`https://www.kaggle.com/c/shelter-animal-outcomes`